TOPOLOGICAL CAUSAL EFFECTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Estimating causal effects becomes particularly challenging when outcomes possess complex, non-Euclidean structures, where conventional approaches often fail to capture meaningful structural variation. We introduce a novel framework for topological causal inference, defining treatment effects through changes in the underlying topological structure of outcomes. In our framework, intervention-driven topological shifts across homology are summarized via power-weighted silhouettes. We propose a doubly robust estimator, derive its asymptotic properties, and develop a formal test for the null hypothesis of no topological effect. Empirical studies demonstrate that our approach reliably quantifies treatment effects and remains robust across diverse, complex outcome spaces.

1 INTRODUCTION

Causal inference has recently emerged as a core tool across diverse disciplines, providing a statistical framework for understanding the effects of interventions beyond associations. Central to this framework is the notion of potential outcomes (Imbens & Rubin, 2015), which conceptualize causal effects by imagining counterfactual scenarios, i.e., what would have occurred under alternative treatment conditions. As scientific data grow increasingly complex, existing methods often fail to capture intervention-induced changes in the structural properties of such outcomes. Surprisingly, relatively little work has addressed causal inference for outcomes that lie outside vector-valued Euclidean spaces, particularly in settings involving high-dimensional, unstructured data.

In this work, we focus on settings where changes in underlying topological characteristics, rather than simple numerical summaries, encode the scientifically relevant causal effects of interest. Such settings arise naturally in numerous applications: in the biomedical sciences, where interventions may alter molecular conformations or induce protein folding (Kovacev-Nikolic et al., 2016; Cang & Wei, 2018; Axelrod & Gomez-Bombarelli, 2022); in neuroscience, where brain connectivity networks evolve in response to stimuli (Sizemore et al., 2019); and in signal processing or medical imaging, where detecting structural changes in dynamical systems or CT scans is of primary importance (Kim et al., 2018; Gholizadeh & Zadrozny, 2018).

Specifically, we study a new class of causal effects designed to capture shifts in the underlying topological structure of complex outcomes leveraging tools from Topological Data Analysis (TDA). TDA is an emerging area that applies techniques from algebraic topology to extract robust, multi-scale features from various forms of complex data structure (Carlsson, 2020). We specifically utilize persistent homology, which captures the birth and death of topological features (e.g., connected components, holes, voids) across scales (Chazal & Michel, 2021). Recent advances have demonstrated the utility of TDA in enhancing robustness of predictive models under covariate shift and data perturbations (e.g., Carrière et al., 2020; 2021; Kim et al., 2020). However, to the best of our knowledge, the integration of TDA into causal inference, particularly within the potential-outcome framework, has not been formally investigated, highlighting a clear gap in the literature.

Expository Example. We illustrate our proposed methodology using a macromolecule dataset (Axelrod & Gomez-Bombarelli, 2022), depicted in Figure 1, which will be revisited in the experimental section. The molecular data are represented as simplified graphs of different sizes. We posit a hypothetical chemical treatment that induces additional loop-like connected components in the molecular structure. Such structural shifts are challenging to capture with conventional methods; however, our approach uncovers clear differences in the first-dimensional homology features, as reflected in the corresponding persistence diagrams. These topological features are then transformed

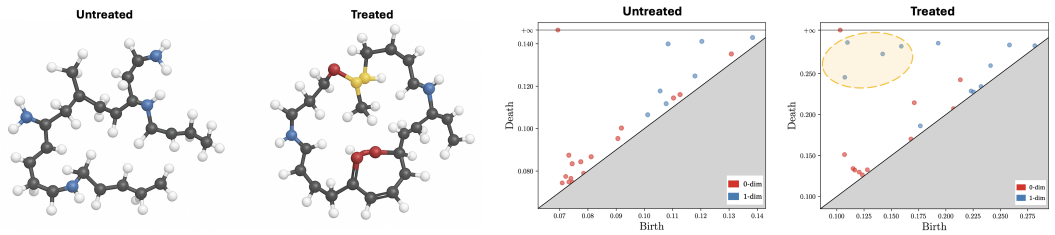


Figure 1: Left: Example of untreated vs. treated macromolecule structures. Right: Corresponding persistence diagrams, highlighting treatment-induced changes in the 1st-order homology features.

into functional summaries that are suitable for downstream machine learning tasks (Chazal et al., 2014; Bubenik et al., 2015). Consequently, our framework is closely related to recent work on functional causal inference (Ecker et al., 2024; Testa et al., 2025), an area still in the early stages of methodological development.

Contributions. We propose and analyze a novel class of *topological causal effects*, defined as the expected difference between the silhouette functions of the persistence diagrams under the potential outcomes before and after intervention. Our work is the first to formally address learning causal effects in topological spaces, bridging the methodological gap between statistical causal inference and algebraic topology. This offers several key advantages: it is invariant to smooth perturbations, captures both local and global structural features, and facilitates causal inference in settings with non-scalar, non-Euclidean, and structurally complex outcomes. We propose doubly robust estimators and show that they attain the fast $\sqrt{n}$ convergence rate, enabling valid and tractable inference even in fully nonparametric settings. We further derive new stability bounds for weighted silhouettes and develop a formal hypothesis test for the null of no topological causal effect. Our approach substantially broadens the expressive capacity of causal inference, as demonstrated in numerical studies, enabling rigorous analysis of structural effects in complex systems and opening new directions for causal investigation in diverse, modern data structures.

2 Preliminaries: Topological Tools for Machine Learning

This section provides a brief overview of the foundational concepts in TDA and introduces key notations used throughout the paper. For a more comprehensive treatment, we refer the reader to Edelsbrunner & Harer (2010); Chazal & Michel (2021).

Simplex and Simplicial Complex. Let $u_0, \ldots, u_k$ be affinely independent points in $\mathbb{R}^d$. A *k-simplex* is the convex hull of the $k+1$ points, $\sigma_k = \text{conv}\{u_0, \ldots, u_k\}$ (e.g., 0-simplex is a vertex, 1-simplex is an edge, 2-simplex is a triangle, etc.). τ is a *face* of σ_k if it is the convex hull of any non-empty subset of the $k+1$ vertices of σ_k, denoted $\tau \leq \sigma_k$. A *simplicial complex* K is a finite collection of simplices such that (i) the face of any simplex in K is also in K, and (ii) the intersection of two simplices in K is either empty or a face of both simplices.

Persistent Homology and Diagrams. A *filtration* $\mathcal{F} = \{K(a) \subset K : a \in \mathbb{R}\}$ is a collection of nested simplicial complexes such that $a \leq b$ implies $K(a) \subset K(b)$. Filtrations are often constructed using a monotonic *filtration function* $f : K \to \mathbb{R}$, where f is monotonic in the sense that $f(\tau) \leq f(\sigma)$ given $\tau \leq \sigma$, i.e., τ is a face of σ. By defining $K(a) := f^{-1}(-\infty, a]$, we have $K(a) \subset K(b)$ whenever $a \leq b$. Given a filtration $\mathcal{F}$, *persistent homology* provides a multi-scale topological representation of the data by tracking the birth and death of d-dimensional homological features (e.g., $d = 0$: connected components, $d = 1$: loops, $d = 2$: cavity, etc.). We denote the d-th homology group as H_d. A homological feature is said to be born at a and to die at b if it appears in $K(a)$ and disappears in $K(b)$. The set of all such birth-death pairs (a, b) can be represented as points in a plane, which gives us the *persistence diagram*. A *persistence diagram* $\mathcal{D}(\mathcal{F})$ is defined as a multiset of points in $\mathbb{R}^{2+} := \{(a, b) \in (\mathbb{R} \cup \infty)^2 : a < b\}$. We let $\mathcal{D}_d(\mathcal{F})$ denote the persistence diagram corresponding to d-dimensional homological features. For notational simplicity, the filtration notation $\mathcal{F}$ will be dropped when understood from context.

Weighted Silhouettes. The multiset nature of persistence diagrams complicates the application of conventional operations, motivating the need for alternative representations that facilitate analysis. One such representation is *weighted silhouette* (Chazal et al., 2014), which is a mapping of persistence diagrams into a functional Hilbert space. Given a persistence diagram $\mathcal{D}$, we first define a piecewise linear tent function $\Lambda_p : \mathbb{R} \to \mathbb{R}$ for each $p = (a, b) \in \mathcal{D}$ such that

$$\Lambda_p(t) = \max\{0, \min\{t - a, b - t\}\}. \tag{1}$$

A collection of functions $\text{kmax}_{p\in\mathcal{D}}\{\Lambda_p(t)\},\ k \in \mathbb{N},\ t \in \mathbb{T}$ is referred to as the *persistence landscape* of $\mathcal{D}$ (Bubenik et al., 2015), where kmax is the k-th largest value in the set and $\mathbb{T}$ is some compact interval in $\mathbb{R}$. When $\mathcal{D}$ has N off-diagonal points, the *weighted silhouette* is a weighted average of the same tent functions (1) used in persistence landscapes:

$$\phi(t; \mathcal{D}) = \frac{\sum_{j=1}^{N} w_j \Lambda_{p_j}(t)}{\sum_{j=1}^{N} w_j}, \quad p_j \in \mathcal{D},\ t \in \mathbb{T}.$$

To reflect topological importance, we generally wish to assign more weights to birth–death pairs with longer lifespans. Thus, the *power-weighted silhouette* assigns weights $w_j = |b_j - a_j|^r$ such that

$$\phi(t; \mathcal{D}, r) = \frac{\sum_{j=1}^{N} |b_j - a_j|^r \Lambda_{p_j}(t)}{\sum_{j=1}^{N} |b_j - a_j|^r}, \quad p_j = (a_j, b_j) \in \mathcal{D},\ t \in \mathbb{T}, \tag{2}$$

for $0 < r \leq \infty$, where large r implies the dominance of most persistence pairs, as opposed to dominance of low persistence pairs for small r. For simplicity, we omit the notation r and assume that all subsequent results hold for an arbitrary choice of r. Both landscapes and silhouettes convert persistence diagrams into functional summaries. Landscapes preserve detailed rank information, while silhouettes provide a more efficient and noise-robust summary via weighted averaging.

The following lemma establishes that power-weighted silhouette functions are 1-Lipschitz, a property that will be instrumental in the inferential procedures developed in Section 5.

Lemma 2.1 (Lipschitz Stability of the Weighted Silhouette)**.** *For any $\delta > 0$, it follows that*

$$\mathbb{E}\Big[\sup_{|s-t|\leq\delta} |\phi(s; \mathcal{D}) - \phi(t; \mathcal{D})|\Big] \ \leq\ \delta.$$

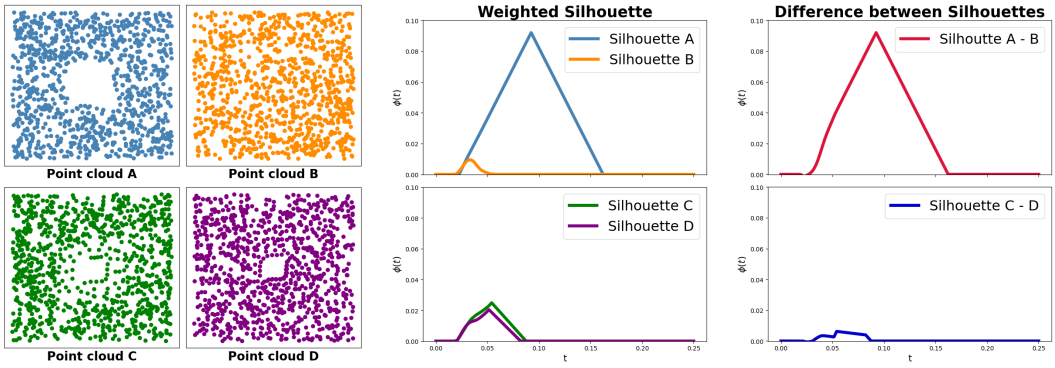
Choice of Filtration. Computing PH descriptors requires specifying a filtration, i.e., a simplicial complex K and a function $f : K \to \mathbb{R}$. The choice depends on the problem and the data structure, and standard constructions are tailored to each modality to reflect intrinsic structure. For example, Vietoris–Rips and Alpha filtrations are standard choices for point-cloud data, while sub- or superlevel filtrations on cubical complexes are natural for grid-structured data such as digital images. For graph data, common options include clique filtrations and the persistent homology transform (PHT) (Turner et al., 2014), which applies height filtrations from multiple directions. By incorporating these different filtration types, our framework readily extends to a broad range of complex data modalities.

3 Framework

Let $\{Z_i = (X_i, A_i, Y_i)\}_{i=1}^n$ denote an i.i.d. observed sample. $A_i \in \mathcal{A} = \{0, 1\}$ is a binary treatment variable such that $A_i = 1$ if subject i is treated and $A_i = 0$ otherwise. $X_i \in \mathcal{X} \subseteq \mathbb{R}^p$ is a p-dimensional vector of covariates. Let $\mathcal{F}_i$ denote the *filtration* of simplicial complexes embedded in $\mathbb{R}^d$ constructed from Y_i, and let $\mathcal{F}_i^a$ denote the corresponding *potential* filtration that would be observed for the potential outcome Y_i^a under treatment $A = a$. Let $\phi_i(t; \mathcal{D}_d) = \phi(t; \mathcal{D}_d(\mathcal{F}_i))$ and $\phi_i^a(t; \mathcal{D}_d^a) = \phi(t; \mathcal{D}_d(\mathcal{F}_i^a))$ denote the power-weighted silhouettes (2) of Y_i and Y_i^a, respectively, for d-dimensional homological features. We denote by $\mathbb{T}$ the domain of definition of ϕ. Our target parameter of interest is the *topological average treatment effect* (TATE), defined as a collection of power-weighted silhouettes; $\psi = (\psi_0(t), \ldots, \psi_{d-1}(t))$ where

$$\psi_d(t) = \mathbb{E}\left[\phi^1(t; \mathcal{D}_d^1) - \phi^0(t; \mathcal{D}_d^0)\right]. \tag{3}$$

Each function $\psi_d(t)$ represents the difference in average silhouettes, capturing the treatment-induced topological variation in d-dimensional homology across treatment groups. Specifically, we will have



(a) (b) (c)

Figure 2: Silhouette functions revealing differences in 1-dimensional features in the ORBIT dataset (Adams et al., 2017). (a) Point clouds A, B, C, D; (b) their power-weighted silhouettes; (c) silhouette contrasts $\phi_A - \phi_B$ (top) and $\phi_C - \phi_D$ (bottom). Strong signals in $\phi_A - \phi_B$ indicate new 1-dimensional features in A relative to B, whereas near-zero $\phi_C - \phi_D$ shows little change between C and D.

$\psi_d(t) > 0$ if, on average, the treated group exhibits *more* or *stronger* d-dimensional features at scale t; $\psi_d(t) < 0$ if those features diminish under treatment. Thus, the mapping $t \mapsto \psi_d(t)$ is a *functional* causal effect that traces how treatment alters topology across all filtration scales.

There are several compelling advantages to interpreting (3) as treatment-induced topological effects. First, $\psi_d(t)$ is inherently scale-aware, as the silhouette function preserves the filtration parameter t, thereby allowing $\psi_d(t)$ to localize changes in topology across different geometric scales. Second, it exhibits robustness to noise: the use of power-weighted silhouettes effectively downweights ephemeral features, enabling $\psi_d(t)$ to emphasize persistent structural differences that are more likely to be meaningful. Third, unlike raw persistence diagrams, $\psi_d(t)$ is vectorizable, residing in a separable Hilbert space, which facilitates theoretical analysis and integration into gradient-based machine learning frameworks. Fourth, $\psi_d(t)$ can be studied within the framework of recent advances in functional causal inference. Collectively, $\psi_d(t)$ can be viewed as a topology-aware analogue of the average treatment effect, capturing how an intervention reshapes the d-dimensional homological structure of the underlying data manifold across treatment groups. One limitation is that silhouette functions may obscure the exact number of changing homological features, as they aggregate multiple tent functions into a single weighted summary. In principle, the TATE can be defined using individual landscape functions across all homology orders up to a specified level; this may yield a finer topological resolution, at the cost of reduced information efficiency (see Section 2).

The motivating example in Figure 2 illustrates how the target parameter (3) captures structural variation. The point cloud pairs (A, B) and (C, D) in Figure 2(a) represent potential outcomes for treated and control groups. The treatment induces a loop in A relative to the control, resulting in a pronounced silhouette difference in Figure 2(c), whereas the difference for (C, D) remains near zero, indicating negligible structural change. The sign of the silhouette difference curve encodes the direction of topological change: positive values correspond to newly emerged features under treatment, and negative values to features lost relative to control.

For identification, we invoke the standard causal assumptions: (C1) *Consistency*, $\mathcal{F} = \mathcal{F}^a$ whenever $A = a$; (C2) *No unmeasured confounding*, $A \perp\!\!\!\perp \phi^a(t; \mathcal{D}_d^a) \mid X$; and (C3) *Positivity*, $\mathbb{P}(A = a \mid X) > 0$ almost surely for all $a \in \mathcal{A}$ and homology dimensions $d \in \mathbb{N}$. Despite the increased complexity of our target parameter, the identification conditions closely parallel those in standard causal inference settings (e.g., Imbens & Rubin, 2015). We assume that Assumptions (C1)–(C3) hold throughout the paper. Note that component-wise identification is sufficient for identifying the functional effect $\psi_d(t)$ at each filtration scale t. Then, for each d, the target $\psi_d(t)$ is identified as

$$\psi_d(t) = \mathbb{E}\left[\mathbb{E}\{\phi(t; \mathcal{D}_d) \mid X, A = 1\} - \mathbb{E}[\phi(t; \mathcal{D}_d) \mid X, A = 0\}\right] \tag{4}$$

$$= \mathbb{E}\left\{\frac{A\phi(t; \mathcal{D}_d)}{\pi(X)} - \frac{(1 - A)\phi(t; \mathcal{D}_d)}{1 - \pi(X)}\right\}. \tag{5}$$

The identifying expressions above motivate the use of plug-in and inverse probability weighting estimators, which will be discussed in greater detail in the following section.

4 ESTIMATION

In this section, we develop an estimation strategy for ψ_d defined in (3). For convenience, we let

$$\pi(x) = \mathbb{P}(A = 1 \mid X = x), \quad \mu_a(t, x; d) = \mathbb{E}[\phi(t; \mathcal{D}_d) \mid X = x, A = a],$$

denote the *propensity score* and the *silhouette regression function*, respectively. Their respective estimators are denoted by $\hat{\pi}$ and $\hat{\mu}_a$. We let $\widehat{\mathbb{P}}$ denote the empirical distribution on which our nuisance estimators $\hat{\mu}_a$ and $\hat{\pi}$ are estimated. We assume access to a separate sample distribution $\mathbb{P}_n$ independent of $\widehat{\mathbb{P}}$, used for constructing our estimator. Throughout, we fix the homology dimension d and, when the context is clear, omit its dependence from the target parameter and estimators, since the subsequent theoretical results do not depend on d. Motivated by the identification results in (4) and (5), we propose the *plug-in* (PI) regression and *inverse probability weighting* (IPW) estimators as

$$\widehat{\psi}_{PI}(t) = \mathbb{P}_n \left\{ \hat{\mu}_1(t, X; d) - \hat{\mu}_0(t, X; d) \right\}, \tag{6}$$

$$\widehat{\psi}_{IPW}(t) = \mathbb{P}_n \left\{ \frac{A\phi(t; \mathcal{D}_d)}{\hat{\pi}(X)} - \frac{(1 - A)\phi(t; \mathcal{D}_d)}{1 - \hat{\pi}(X)} \right\}, \tag{7}$$

respectively. $\widehat{\psi}_{PI}$ and $\widehat{\psi}_{IPW}$ inherit their convergence rates directly from those of the estimators for μ_a and π, respectively. IPW estimators are generally preferred, as domain knowledge about the treatment assignment mechanism is often more readily available and can be effectively incorporated (Imbens & Rubin, 2015). In our setting, $\widehat{\psi}_{\text{IPW}}$ is particularly advantageous for estimating the TATE, as estimating the functional regression μ_a is substantially more challenging than estimating the standard propensity score π.

A more efficient estimator can be constructed using tools from semiparametric efficiency theory. Let

$$\varphi(t, Z; \eta) = \mu_1(t, X; d) - \mu_0(t, X; d) + \left\{ \frac{A}{\pi(X)} - \frac{1 - A}{1 - \pi(X)} \right\} \left\{ \phi(t; \mathcal{D}_d) - \mu_A(t, X; d) \right\}, \tag{8}$$

where $\eta = \{\pi, \mu_a\}_a$, a set of the nuisance functions. φ is the uncentered efficient influence function (EIF) for the target functional (3). The EIF enables construction of the efficient semiparametric estimator by de-biasing the PI or IPW estimators, where we may achieve local minimax lower bounds (Bickel et al., 1993; van der Vaart, 2002; Tsiatis, 2006; Kennedy, 2016). This also yields desirable properties for our estimator, such as double robustness or general second-order bias, which allows us to relax nonparametric conditions on nuisance function estimation. It is immediate to see that $\mathbb{E}\{\varphi(t, Z; \eta)\} = \psi(t)$. Based on the EIF (8), we construct an efficient augmented inverse-probability-weighted (AIPW) estimator as

$$\widehat{\psi}_{AIPW}(t) = \mathbb{P}_n \left\{ \varphi(t, Z; \widehat{\eta}) \right\} \equiv \mathbb{P}_n \left\{ \widehat{\varphi}(t) \right\}, \tag{9}$$

where $\widehat{\eta} = \eta(\widehat{\mathbb{P}})$, i.e., $\{\hat{\pi}, \hat{\mu}_a\}$.

To our knowledge, there are two main approaches for constructing AIPW estimators; one is based on empirical process conditions, and the other is to use sample splitting. One may assume that the function class for φ_d, η, and the corresponding estimators are not too complex (e.g., Donsker or low-entropy type conditions), but this would limit the flexibility of the nuisance estimators. To avoid this, alternatively, we can use sample splitting (or cross-fitting) to allow for arbitrarily complex nuisance estimators. Both approaches can be viewed as ways to avoid using the same data twice, one for constructing relevant nuisance components, and the other for de-biasing, which can introduce a threat of overfitting (Chernozhukov et al., 2018). We refer the interested readers to Kennedy (2016; 2024) and references therein. We remain methodologically agnostic regarding the choice of estimation approach and assume independence between $\mathbb{P}_n$ and $\widehat{\mathbb{P}}$. Nonetheless, for the sake of simplicity and clarity, we adopt the sample splitting method as the default throughout this paper.

Remark 1 (Sample splitting)**.** *For nuisance estimation, independent samples $\mathbb{P}_n$ and $\widehat{\mathbb{P}}$ can be obtained via random data splitting. Full-sample efficiency is recoverable through cross-fitting (e.g., Chernozhukov et al., 2018; Newey & Robins, 2018). For simplicity and clarity, we focus on a single split, though extending to multiple splits is straightforward (e.g., Kennedy, 2019; 2023).*

5 INFERENCE

Conditions for achieving desirable statistical properties, such as $\sqrt{n}$-rate convergence and asymptotic normality, are well established in the case of scalar responses for ATE estimation (e.g., Kennedy, 2016; 2024). In contrast to the conventional setting, the problem addressed here requires additional structural assumptions to ensure valid statistical inference.

We begin with a simplified setting in which the treatment assignment mechanism is fully known, as in randomized experiments. Specifically, we assume that the true propensity score π is known almost surely. In this case, the IPW estimator in (7) is unbiased. Furthermore, under relatively mild conditions, the following weak convergence result can be established.

Theorem 5.1. *Let* $\xi(Z,t) = \left\{\frac{A}{\pi(X)} - \frac{1-A}{1-\pi(X)}\right\}\phi(t;\mathcal{D}_d)$. *Assume that for any* $s,t \in \mathbb{T}$, *cov*$(\xi(Z,s),\xi(Z,t)) < \infty$. *Then we have*

$$\sqrt{n}(\widehat{\psi}_{IPW}(t) - \psi) \to \mathbb{G} \text{ weakly in } \ell^\infty(\mathbb{T}),$$

where $\mathbb{G}$ *is a mean-zero Gaussian process with covariance function cov* $[\mathbb{G}(s),\mathbb{G}(t)] =$ *cov*$(\xi(Z,s),\xi(Z,t))$.

Hence, when the propensity score is known, $\widehat{\psi}_{\text{IPW}}$ converges in distribution to a Gaussian process.

In observational settings, it is often preferable to use the AIPW estimator defined in (9). Hereafter, we use the notation $\|\cdot\|_{\mathbb{P},q}$ to denote the $L_q(\mathbb{P})$-norm. To analyze the asymptotic properties, we introduce the following set of assumptions.

(A1) $\|1/\widehat{\pi}\|_\infty < \infty$ and $\|1/(1-\widehat{\pi})\|_\infty < \infty$.

(A2) *Sample splitting*: The nuisance estimators are computed in a separate independent sample.

(A3) *Cross-sectional consistency*: For every $t \in \mathbb{T}$, $\|\widehat{\varphi}(t) - \varphi(t)\|_{\mathbb{P},2} = o_{\mathbb{P}}(1)$.

(A4) *Rate condition on nuisance estimation*: For every $t \in \mathbb{T}$, $\|\hat{\pi}(X) - \pi(X)\|_{\mathbb{P},2}\left\{\sum_{a\in\mathcal{A}}\|\hat{\mu}_a(t,X;d) - \mu_a(t,X;d)\|_{\mathbb{P},2}\right\} = o_{\mathbb{P}}(n^{-1/2})$, $\forall t \in \mathbb{T}$..

(A5) *Uniform consistency*: For all $x \in \mathcal{X}$, $\mu_a(t,x;d)$ is uniformly Lipschitz in t on $\mathcal{T}$, and following conditions hold:

$$\sup_{t\in\mathcal{T},x\in\mathcal{X}}\left|\widehat{\mu}_a(t,x;d) - \mu_a(t,x;d)\right| = o_{\mathbb{P}}(1), \quad \sup_{x\in\mathcal{X}}\left|\widehat{\pi}(x) - \pi(x)\right| = o_{\mathbb{P}}(1).$$

Assumptions (A1) - (A5) are standard and often employed in semiparametric causal inference. (A1) is a mild boundedness condition. (A2) and (A3) are required to control the associated empirical process term (see Remark 1). Note that one may replace (A2) with a Donsker-type condition on ψ and $\widehat{\psi}_{\text{AIPW}}$. (A4) requires the von Mises remainder to be asymptotically negligible, ensuring double robustness (Kennedy, 2024), and is typically combined with the uniform consistency conditions in Assumption (A5) (e.g., Kennedy, 2019; Kennedy et al., 2023).

It is worth noting that our regularity conditions on the functional regression estimator from Assumption (A5) is weaker than that used in (Testa et al., 2025):

(A5′) *Regularity condition on* $\hat{\mu}_a$: For $\delta > 0$ and $a \in \mathcal{A}$, the silhouette regression function estimator $\hat{\mu}_a(t,X;d)$ satisfies $\mathbb{E}\left[\sup_{|s-t|\le\delta}|\hat{\mu}_a(s,X;d) - \hat{\mu}_a(t,X;d)|\right] \le L_\phi\delta$, for some positive constant L_ϕ.

(A5′) directly imposes a Lipschitz modulus on the estimator $\widehat{\mu}_a$'s random sample paths. In contrast, (A5) leaves $\widehat{\mu}_a$ unspecified apart from uniform consistency. Thus in our case, the estimator itself can even be non-smooth, provided it converges uniformly to a Lipschitz target, which is often guaranteed by mild smoothness of the data-generating mechanism.

Depending on the estimation strategy, this regularity condition on the functional regression estimator can, in fact, be entirely relaxed, as illustrated in the following example.

Example 5.1 (Functional Linear Smoother)**.** *One can use a functional linear smoother for the scalar-on-function regression μ_a (e.g., Reiss et al., 2017), where the estimator takes the form $\hat{\mu}_a(t, X; d) = L_d \mathbf{\Phi}$, with $\mathbf{\Phi} = (\phi_1, \ldots, \phi_n)$ and $L_d \in \mathbb{R}^{n \times n}$ denoting the smoothing matrix. In this case, we have*

$$\mathbb{E}\Big[\sup_{|s-t|\leq\delta} |\hat{\mu}_a(s, X; d) - \hat{\mu}_a(t, X; d)|\Big] = \mathbb{E}\Big[L_d \mathbb{E}\Big\{\sup_{|s-t|\leq\delta} |\phi(s; \mathcal{D}) - \phi(t; \mathcal{D})|\Big\}\Big] \leq \mathbb{E}(L_d)\delta,$$

where the inequality follows from Lemma 2.1. Consequently, Assumption (A5) can be completely omitted. A notable special case is the ordinary least squares estimator.

Pointwise $\sqrt{n}$-consistency and asymptotic normality follow naturally as consequences of semiparametric efficiency theory.

Proposition 5.2. *Assume that Assumptions (A1) - (A4) hold. Then for fixed $t \in \mathbb{T}$,*

$$\sqrt{n}\left(\widehat{\psi}_{AIPW}(t) - \psi\right) \xrightarrow{d} N\left(0, \mathit{var}(\varphi(t))\right),$$

and $\widehat{\psi}_{AIPW}(t)$ is efficient, meaning that there exist no other regular estimators that are asymptotically unbiased and have smaller variance.

The following result forms the foundation of our inferential framework.

Theorem 5.3. *Under Assumptions (A1)–(A5), the following weak convergence result holds:*

$$\sqrt{n}(\widehat{\psi}_{AIPW}(t) - \psi) \to \mathbb{G} \text{ weakly in } \ell^\infty(\mathbb{T}),$$

where $\mathbb{G}$ is a mean-zero Gaussian process with covariance function $cov\left[\mathbb{G}(s), \mathbb{G}(t)\right] = cov(\varphi(t, Z; \eta), \varphi(s, Z; \eta))$, and φ is defined in (8).

Building on Theorem 5.3, we may construct confidence bands for inference with simultaneous coverage. There are several approaches to constructing an asymptotically valid $1 - \alpha$ confidence band, depending on the strength of additional assumptions one is willing to impose or the level of computational complexity one is prepared to accommodate. For instance, one may apply the pivotal method proposed by Liebl & Reimherr (2023), or adopt the parametric bootstrap approach developed by Pini & Vantini (2017). For a more detailed discussion and comparison, see Testa et al. (2025).

For many, if not most, practitioners, the primary goal is to test for the presence of topological effects as captured by persistent homology. Hypothesis tests based on estimated, vectorized topological summaries (e.g., silhouettes, persistence images, or landscapes) generally do not yield valid inference in the metric space of persistence diagrams. In contrast, our framework furnishes a formal test of the null of no topological effect. To this end, we first establish stability bounds for weighted silhouettes, which, to our knowledge, have not previously appeared in the literature.

Theorem 5.4. *Let $W_q(\mathcal{D}, \mathcal{D}')$ denote the q-Wasserstein distance between two persistence diagrams $\mathcal{D}$ and $\mathcal{D}'$, and let $m^\star$ denote the corresponding optimal W_1 matching. Assume that*

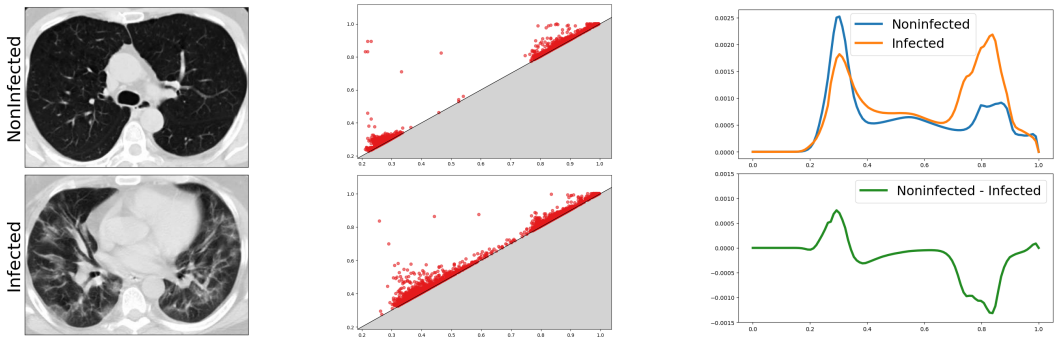
(A6) For the power-weighted silhouette defined in (2), its corresponding persistence diagram $\mathcal{D}$ is bounded such that for all $p_j = (a_j, b_j) \in \mathcal{D}$, $-\infty < a_j \leq b_j < \infty$. Moreover, there exists a global constant $L > 0$ such that, for all $p_j \in \mathcal{D}$ and a given weighting exponent r, $\frac{b_j - a_j}{(b_j - a_j)^r} \leq L$.

Consider weighted silhouette functions ϕ and ϕ' corresponding to $\mathcal{D}$ and $\mathcal{D}'$. Under Assumption (A6),

$$\|\phi - \phi'\|_\infty \leq \left(1 + 2Lr\, c^{\,r-1}\right) W_1(\mathcal{D}, \mathcal{D}'),$$

for some constant $c > 0$ that depends only on r and an upper bound on the persistences in the diagrams $\mathcal{D}, \mathcal{D}'$.

The formal definitions of the Wasserstein distance and the corresponding optimal matching are provided in Appendix B.4. The constant c can be chosen as any global upper bound on the lifetimes (persistences) of the pairs matched under the optimal W_1-matching $m^\star$, whose precise definition is also given in Appendix B.4. Building on the stability result above and the Gaussian weak convergence of our estimator, our framework provides, to our knowledge, the first formal test of the null hypothesis of no topological effect, with asymptotically correct size and consistency against fixed alternatives, as formally stated below.



(a) (b) (c)

Figure 3: (a) CT-scan images of non-infected (top) and infected (bottom) patients; (b) corresponding 0-dimensional persistence diagrams; (c) average silhouettes of non-infected and infected patients given $r = 0.1$ (top) and difference in average silhouettes between non-infected and infected patients.

Corollary 5.5. *Consider the null hypothesis*

$$H_0 : \; W_1(\mathcal{D}^1, \mathcal{D}^0) = 0 \quad a.s.,$$

and the test statistic $T_n := \sqrt{n}\,\|\widehat{\psi}_{\mathrm{AIPW}}\|_\infty$. *Under Assumptions (A1)–(A6), we have* $T_n \;\Rightarrow\; \|\mathbb{G}\|_\infty$, *where* $\mathbb{G}$ *is the Gaussian process from Theorem 5.3. Let* $\widehat{\mathbb{G}}_n$ *be a Gaussian- or Rademacher-multiplier bootstrap process based on the estimated influence function (8), and let* $c_{1-\alpha}$ *be the conditional* $(1-\alpha)$*-quantile of* $\|\widehat{\mathbb{G}}_n\|_\infty$. *Then the test that rejects* H_0 *when* $T_n > c_{1-\alpha}$ *has asymptotic size* α *and is consistent against any fixed alternative with* $\|\psi\|_\infty > 0$.

6 Experiments

To demonstrate the effectiveness of our method, we carry out experiments on two semi-synthetic datasets and one synthetic dataset; results for the synthetic (ORBIT) data appear in Appendix C.3 due to page limitations. For all experiments, we construct a hypothetical dataset (X, A, Y^0, Y^1), where the potential outcome pairs are designed to exhibit distinct topological contrasts across treatment groups. We generate the covariates $X \in \mathbb{R}^5$ from a multivariate Gaussian distribution, imposing a subgroup structure by specifying different mean vectors for each subgroup. Given X, treatment A is assigned with probability $\pi(X) = expit(-0.5X_1 - 0.1X_2 + 0.6X_3 + 0.1X_4 + 0.1X_5 + 0.5X_2X_3 - 0.7X_1X_3)$. This treatment mechanism is designed such that one subgroup has a higher probability of receiving treatment than the other. We model the silhouette regression function μ_a using function-on-scalar regression with a Fourier basis expansion, while the propensity score π is estimated via a random forest classifier. Our goal is to estimate the true topological causal effect based on the observable data. All experiments are repeated over 20 simulations. For complete details of the experimental setup, see Appendix C.

SARS-CoV-2 dataset. The first semi-synthetic experiment uses image, possibly in different sizes, with synthesized covariates and treatment assignments while retaining real outcomes, allowing controlled yet realistic evaluation of causal estimators. We employ the SARS-CoV-2 dataset (Soares et al., 2020), which contains CT-scans collected from real patients who are infected or non-infected by COVID-19. Infected patients exhibit high rates of ground-glass opacities and consolidations that appear as isolated regions in CT-scans, which can be captured by 0-dimensional persistence diagrams of Lower-star filtration (Iqbal et al., 2025). In Figure 3-(a) and (b), the difference between an infected and a non-infected CT scan image is reflected in the associated persistence diagrams. Figure 3-(c)-(top) illustrates the average silhouettes of infected and non-infected patients, where the non-infected group exhibits higher values in [0.2, 0.4] and the infected group exhibits higher values in [0.7, 0.9]. Thus, a treatment can be assumed to be more effective when TATE exhibits larger magnitudes over the interval $[0.2, 0.4]$ (positive direction) and $[0.7, 0.9]$ (negative direction), as illustrated in Figure 3-(c)-(bottom). In this experiment, the true TATE is known, as we manually construct (Y^0, Y^1) by assigning 500 infected samples to Y^0 and subsequently pairing it with Y^1,

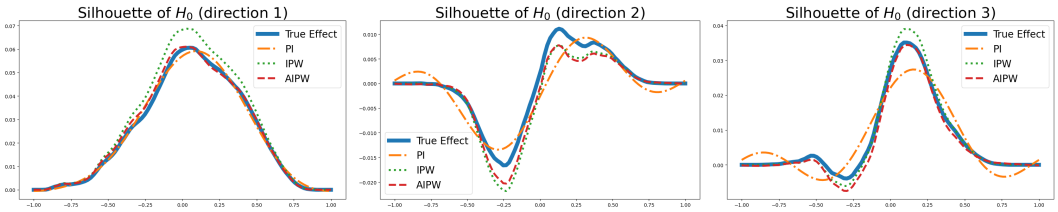


Figure 5: 0-dimensional true silhouette functions and its PI, IPW, AIPW estimates along three directions using persistent homology on the GEOM-Drugs dataset.

which consist of 75% non-infected and 25% infected samples. We compute the PI, IPW, and AIPW estimators from the observed data and compare their estimates with the true effect.

Results. Figure 4 presents the true silhouette function, which by design clearly reflects a causal effect. Although all three estimators reasonably capture the overall shape of the true target, the IPW estimator systematically overestimates the true treatment effect, whereas the PI estimator underestimates it. In contrast, the AIPW estimator provides an accurate reconstruction of the true silhouette function, achieving minimal bias and closely matching the exact shape of the ground truth.

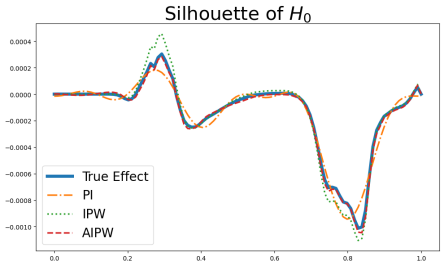


Figure 4: True silhouette and its IPW, PI, AIPW estimates for first-order homological features in the SARS-CoV-2 dataset.

GEOM-Drugs Dataset. Next, we evaluate our framework on graph data through a semi-synthetic experiment with the GEOM-Drugs dataset (Axelrod & Gomez-Bombarelli, 2022), which provides graph-structured representations of molecular compounds. Adopting procedures analogous to the previous experiment, we randomly select 2000 samples to construct 1000 pairs of (Y^0, Y^1). To analyze graph-structure data, we utilize the persistent homology transform defined in (10), which produces a silhouette for each direction ν. For simplicity, and in line with standard practice, we use three representative directions equally spaced on the unit circle.

Results. Consistent with the preceding experiment, Figure 5 shows that the IPW estimator overestimates the true treatment effect, whereas the PI estimator tends to underestimate them. Across all three directional settings, the AIPW estimator delivers the most accurate and reliable approximation of the true silhouettes, consistent with the theoretical results in Section 5.

7 Discussion

We introduce a novel connection between causal inference and TDA, enabling estimation of causal effects that capture not only shifts in mean or variance but also changes in the meaningful intrinsic topological structure of the outcome space. This substantially broadens the expressive capacity of causal inference methodologies and opens new avenues for investigating causal mechanisms in complex, high-dimensional systems.

Several caveats and prospective solutions deserve attention, highlighting fruitful avenues for future investigation. First, the proposed TATE framework is designed to capture macroscopic topological shifts and may be less informative when the primary interest lies in detecting fine-grained, local changes. In such cases, standard causal estimands could be estimated in parallel if possible, potentially after appropriate preprocessing. Relatedly, as discussed in Section 3, silhouette functions do not exactly quantify the number of changing homological features. Nevertheless, the proposed estimators and the analyses in Sections 4 and 5 extend naturally to individual persistence landscape functions, offering finer topological resolution when required. Lastly, the construction of our estimators can be computationally intensive due to the use of persistent homology. This cost may be mitigated by adopting more efficient topological summaries, such as Euler characteristic curves. Additional extensions include adapting the framework to more complex causal settings, such as continuous treatments, instrumental variable designs, or longitudinal exposures.

REFERENCES

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.

Simon Axelrod and Rafael Gomez-Bombarelli. Geom, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1):185, 2022.

Peter J Bickel, Chris AJ Klaassen, Peter J Bickel, Ya'acov Ritov, J Klaassen, Jon A Wellner, and YA'Acov Ritov. *Efficient and adaptive estimation for semiparametric models*, volume 4. Springer, 1993.

Patrick Billingsley. Convergence of probability measures john wiley & sons. *INC, New York*, 2(2.4), 1999.

Peter Bubenik et al. Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.*, 16(1):77–102, 2015.

Zixuan Cang and Guo-Wei Wei. Integration of element-specific persistent homology and machine learning for protein-ligand binding affinity prediction. *International Journal for Numerical Methods in Biomedical Engineering*, 34(2):e2914, 2018.

Gunnar Carlsson. Topological methods for data modelling. *Nature Reviews Physics*, 2(12):697–708, 2020.

Mathieu Carrière, Marco Cuturi, and Steve Oudot. Perslay: A neural network layer for persistence diagrams and new graph topological signatures. In *International Conference on Artificial Intelligence and Statistics*, pp. 2786–2796, 2020.

Mathieu Carrière, Ulrich Bauer, and Steve Oudot. Optimizing persistence diagrams using differentiable topological layers. *International Conference on Machine Learning*, 2021.

Frédéric Chazal and Bertrand Michel. An introduction to topological data analysis: fundamental and practical aspects for data scientists. *Frontiers in artificial intelligence*, 4:108, 2021.

Frédéric Chazal, Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, and Larry Wasserman. Stochastic convergence of persistence landscapes and silhouettes. In *Proceedings of the thirtieth annual symposium on Computational geometry*, pp. 474–483, 2014.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018.

Justin Curry, Sayan Mukherjee, and Katharine Turner. How many directions determine a shape and other sufficiency results for two topological transforms. *Transactions of the American Mathematical Society, Series B*, 9(32):1006–1043, 2022.

Kreske Ecker, Xavier de Luna, and Lina Schelin. Causal inference with a functional outcome. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 73(1):221–240, 2024.

H. Edelsbrunner and J. Harer. *Computational Topology: An Introduction*. Applied Mathematics. American Mathematical Society, 2010. ISBN 9780821849255.

Seyedehsamaneh Gholizadeh and Bianca Zadrozny. A short survey of topological data analysis in time series and systems analysis. *arXiv preprint arXiv:1809.10745*, 2018.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Sohail Iqbal, Hafiz Fareed Ahmed, Talha Qaiser, Muhammad Imran Qureshi, and Nasir Rajpoot. Classification of covid-19 via homology of ct-scan. *Computers in Biology and Medicine*, 193: 110226, 2025.

Edward H Kennedy. Semiparametric theory and empirical processes in causal inference. *Statistical causal inferences and their applications in public health research*, pp. 141–167, 2016.

Edward H Kennedy. Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 114(526):645–656, 2019.

Edward H Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023.

Edward H Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *Handbook of Statistical Methods for Precision Medicine*, pp. 207–236, 2024.

Edward H Kennedy, Sivaraman Balakrishnan, and Max G'sell. Sharp instruments for classifying compliers and generalizing causal effects. 2020.

Edward H Kennedy, Sivaraman Balakrishnan, and LA Wasserman. Semiparametric counterfactual density estimation. *Biometrika*, 110(4):875–896, 2023.

Kwangho Kim, Jisu Kim, and Alessandro Rinaldo. Time series featurization via topological data analysis. *arXiv preprint arXiv:1812.02987*, 2018.

Kwangho Kim, Jisu Kim, Manzil Zaheer, Joon Kim, Frédéric Chazal, and Larry Wasserman. Pllay: Efficient topological layer based on persistent landscapes. *Advances in Neural Information Processing Systems*, 33:15965–15977, 2020.

Violeta Kovacev-Nikolic, Peter Bubenik, Dragan Nikolić, and Giseon Heo. Using persistent homology and dynamical distances to analyze protein binding. *Statistical applications in genetics and molecular biology*, 15(1):19–38, 2016.

Dominik Liebl and Matthew Reimherr. Fast and fair simultaneous confidence bands for functional parameters. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3): 842–868, 2023.

Whitney K Newey and James R Robins. Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*, 2018.

Alessia Pini and Simone Vantini. Interval-wise testing for functional data. *Journal of Nonparametric Statistics*, 29(2):407–424, 2017.

Philip T Reiss, Jeff Goldsmith, Han Lin Shang, and R Todd Ogden. Methods for scalar-on-function regression. *International Statistical Review*, 85(2):228–249, 2017.

Ann E Sizemore, Chad Giusti, and Danielle S Bassett. Importance of the whole: Topological data analysis for the network neuroscientist. *Network Neuroscience*, 3(3):656–673, 2019.

Eduardo Soares, Plamen Angelov, Sarah Biaso, Michele Higa Froes, and Daniel Kanda Abe. Sars-cov-2 ct-scan dataset: A large dataset of real patients ct scans for sars-cov-2 identification. *MedRxiv*, pp. 2020–04, 2020.

Lorenzo Testa, Tobia Boschi, Francesca Chiaromonte, Edward H Kennedy, and Matthew Reimherr. Doubly-robust functional average treatment effect estimation. *arXiv preprint arXiv:2501.06024*, 2025.

Anastasios A Tsiatis. *Semiparametric theory and missing data*, volume 4. Springer, 2006.

Katharine Turner, Sayan Mukherjee, and Doug M Boyer. Persistent homology transform for modeling shapes and surfaces. *Information and Inference: A Journal of the IMA*, 3(4):310–344, 2014.

Aad W van der Vaart. Semiparametric statistics. In *Lectures on probability theory and statistics (Saint-Flour, 1999)*, pp. 331–457. Springer, 2002.

APPENDIX

A MORE BACKGROUND ON TOPOLOGICAL DATA ANALYSIS

Vietoris-Rips Complex. Let X be a finite set of points in $\mathbb{R}^d$. For $r > 0$, the *Vietoris-Rips complex* is a collection of simplices where the distance between any two vertices is smaller than $2r$:

$$\text{Rips}(r) = \{\sigma \subset X | d(u_i, u_j) < 2r, \forall u_i, u_j \in \sigma\}.$$

Notice that $\text{Rips}(r_1) \subset \text{Rips}(r_2)$ when $r_1 \leq r_2$. Thus, we can build a filtration on the Vietoris-Rips complex by monotonically increasing r.

Alpha Complex. Let X be a finite set of points in $\mathbb{R}^d$. For each $u_i \in X$, the *Voronoi cell* of u_i is the set of points that are closest to u_i; $V_{u_i} = \{x \in \mathbb{R}^d | d(u_i, x) \leq d(u_j, x), \forall u_j \in X, u_j \neq u_i\}$. For $r > 0$ and each $u_i \in X$, let us denote the closed r-*ball* with center u_i and radius r as $B_{u_i}(r)$. Then, we define $R_{u_i}(r) = B_{u_i}(r) \cap V_{u_i}$, which is the intersection of each r-ball with its corresponding Voronoi cell. The *Alpha complex* is a collection of simplices such that all $R_{u_i}(r)$ of the vertices in the simplex have an intersection:

$$\text{Alpha}(r) = \{\sigma \subset X | \cap_{u_i \in \sigma} R_{u_i}(r) \neq \emptyset\}.$$

Similar to the Vietoris-Rips complex, we can build a filtration on the Alpha complex by monotonically increasing r.

Cubical Complex. Cubical complex is an analogy of simplicial complex that consists of k-cubes (e.g., vertices, edges, squares, cubes, etc.). It provides a suitable framework for analyzing data that is naturally aligned with a grid structure (e.g., digital images). An *elementary interval* is an interval of form $I = [l, l + 1]$ or $I = [l, l]$ for some $l \in \mathbb{Z}$, where the former interval is called *nondegenerate* and the latter *degenerate*. An *elementary cube* is the finite product of elementary intervals, i.e., $Q = I_1 \times I_2 \times \cdots \times I_n$. The *dimension* of Q is the number of nondegenerate elementary intervals in the product. P is a *face* of Q if $P \subset Q$ where P and Q are both elementary cubes. A *cubical complex* K is a finite collection of elementary cubes such that the face of any cube in K is also in K. A *filtered* cubical complex can be constructed by assigning a filtration value to each of the cubes.

Persistence Landscapes. Persistence landscapes (Bubenik et al., 2015) are a mapping of persistence diagrams into a functional Hilbert space. The *persistence landscape* of a diagram $\mathcal{D}$ is a collection of functions $\{\lambda(k, t; \mathcal{D})\}_{k \in \mathbb{N}}$ defined as

$$\lambda(k, t; \mathcal{D}) = \underset{p \in \mathcal{D}}{\text{kmax}}\{\Lambda_p(t)\}, \quad k \in \mathbb{N},\ t \in \mathbb{T},$$

where $\Lambda_p(t)$ is the tent function in (1), kmax is the k-th largest value in the set, and $\mathbb{T}$ is some compact interval in $\mathbb{R}$. When k is given, we write $\lambda(t; \mathcal{D}, k)$ to denote the k-th persistence landscape. Accordingly, $\lambda(t; \mathcal{D}_d, k)$ is the k-th persistence landscape of d-dimensional homological features. If the corresponding persistence diagram contains N off-diagonal points, all k-th persistence landscapes for $k > N$ are zero.

Persistent Homology Transform. The persistent homology transform (PHT) (Turner et al., 2014) is an injective shape descriptor that consists of persistence diagrams computed from multiple directions. To construct PHT, we first set the filtration function f as follows. Given a simplicial complex K embedded in $\mathbb{R}^d$, let K_0 be the set of 0-simplices in K. For any unit vector $\nu \in S^{d-1}$, we define the filtration function f to be

$$\begin{aligned} f : K \times S^{d-1} &\to \mathbb{R} \\ (\sigma, \nu) &\mapsto \max_{\tau \leq \sigma; \tau \in K_0} \langle \tau, \nu \rangle, \end{aligned}$$

where ν is a direction and $\langle \tau, \nu \rangle$ is the projection of a vertex τ onto ν. When the direction ν is fixed, we write $f_\nu := f(\cdot, \nu)$. Defining $K_\nu(h) := f_\nu^{-1}(-\infty, h]$ for $h \in \mathbb{R}$ naturally induces a sublevel set filtration of nested subcomplexes $\mathcal{F}_\nu := \{K_\nu(h) \subset K : h \in \mathbb{R}\}$, where each subcomplex $K_\nu(h)$ consists of simplices $\sigma \in K$ that satisfy $\langle \tau, \nu \rangle \leq h$ for all vertices $\tau \in \sigma$. Given $\mathcal{F}_\nu$, the *persistent homology transformation* is defined as

$$\text{PHT} : (K, \nu) \mapsto (\mathcal{D}_0(\mathcal{F}_\nu), \ldots, \mathcal{D}_{d-1}(\mathcal{F}_\nu)).$$

A key property of PHT is its injectivity, characterizing it as a complete and sufficient descriptor for shape identification. Notably, PHT remains injective even under a finite set of directions, as long as the number of directions is sufficiently large (Curry et al., 2022). Throughout this work, we adopt power-weighted silhouettes in place of persistence diagrams. Accordingly, PHT is reformulated as

$$\text{PHT} : (K, \nu) \mapsto (\phi(t; \mathcal{D}_0(\mathcal{F}_\nu)), \ldots, \phi(t; \mathcal{D}_{d-1}(\mathcal{F}_\nu)))) \,. \tag{10}$$

It is worth noting that replacing persistence diagrams with power-weighted silhouettes results in a loss of injectivity, in exchange for a more tractable and interpretable representation.

B PROOF

Notation. Hereafter, we let $\|x\|_q$ denote L_q norm for any fixed vector x. For a given function f, we use the notation $\|f\|_{\mathbb{P},q} = [\mathbb{P}(|f|^q)]^{1/q} = \left[\int |f(z)|^q d\mathbb{P}(z)\right]^{1/q}$ as the $L_q(\mathbb{P})$-norm of f. Also, we let $\mathbb{P}$ denote the conditional expectation given the sample operator $\hat{f}$, as in $\mathbb{P}(\hat{f}) = \int \hat{f}(z) d\mathbb{P}(z)$. Notice that $\mathbb{P}(\hat{f})$ is random only if $\hat{f}$ depends on samples, in which case $\mathbb{P}(\hat{f}) \neq \mathbb{E}(\hat{f})$. Otherwise $\mathbb{P}$ and $\mathbb{E}$ can be used exchangeably. For example, if $\hat{f}$ is constructed on a separate (training) sample $\mathsf{D}^n = (Z_1, ..., Z_n)$, then $\mathbb{P}\left\{\hat{f}(Z)\right\} = \mathbb{E}\left\{\hat{f}(Z) \mid \mathsf{D}^n\right\}$ for a new observation $Z \sim \mathbb{P}$. We let $\mathbb{P}_n$ denote the empirical measure as in $\mathbb{P}_n(f) = \mathbb{P}_n\{(f(Z)\} = \frac{1}{n}\sum_{i=1}^n f(Z_i)$. Lastly, we use the shorthand $a_n \lesssim b_n$ to denote $a_n \leq \mathsf{c} b_n$ for some universal constant $\mathsf{c} > 0$.

B.1 PROOF OF LEMMA 2.1

Proof. For notational convenience, we denote $\phi(t) := \phi(t; \mathcal{D}, r)$ for any fixed persistence diagram $\mathcal{D}$ and the power parameter r. Fix $p = (a, b) \in \mathcal{D}$. The tent function Λ_p is linear on $[a, (a+b)/2]$ with slope $+1$ and on $[(a+b)/2, b]$ with slope -1. Hence, for all $s, t \in \mathbb{R}$,

$$|\Lambda_p(s) - \Lambda_p(t)| \leq |s - t|.$$

Namely, each Λ_p is 1–Lipschitz.

Rewrite the weighted silhouette as $\phi(t) = \sum_{j=1}^N \alpha_j \Lambda_{p_j}(t)$ with weights $\alpha_j = w_j / \sum_k w_k \in (0, 1)$ satisfying $\sum_j \alpha_j = 1$. Then it follows that

$$|\phi(s) - \phi(t)| = \left| \sum_j \alpha_j \big(\Lambda_{p_j}(s) - \Lambda_{p_j}(t)\big) \right| \leq \sum_j \alpha_j \, |s - t| = |s - t|,$$

which shows the silhouette functions is also Lipschitz with constant $L = 1$.

Even when the persistence diagram $\mathcal{D}$ is random, the inequality holds pathwise; thus, taking expectations yields

$$\mathbb{E}\Big[\sup_{|s-t| \leq \delta} |\phi(s) - \phi(t)| \Big] \;\leq\; \delta.$$

□

Theorem 5.1 follows as a direct consequence of Theorem 5.3. We therefore focus on proving Theorem 5.3 directly.

B.2 PROOF OF THEOREM 5.3 ASSUMPTION (A5′)

Recall that φ is the uncentered EIF for the target functional (3), satisfying a Von Mises expansion:

$$\psi(t; \widehat{\mathbb{P}}) - \psi(t; \mathbb{P}) = \int \varphi(t, z; \eta) \, d(\widehat{\mathbb{P}} - \mathbb{P})(z) + R_2(\widehat{\mathbb{P}}, \mathbb{P}), \tag{11}$$

where the second-order remainder term is specified as

$$R_2(\widehat{\mathbb{P}}, \mathbb{P}) = \int \left\{ \frac{1}{\pi(X)} - \frac{1}{\hat{\pi}(X)} \right\} \{\mu_1(t, X; d) - \hat{\mu}_1(t, X; d)\} \, \pi(X) \, d\mathbb{P} \\ - \int \left\{ \frac{1}{1-\pi(X)} - \frac{1}{1-\hat{\pi}(X)} \right\} \{\mu_0(t, X; d) - \hat{\mu}_0(t, X; d)\} \, \{1 - \pi(X)\} \, d\mathbb{P}, \, \forall t \in \mathbb{T}. \tag{12}$$

First, we prove the asymptotic normality of finite-dimensional causal effect projections as stated in the following lemma. The proof follows the argument of Theorem 5.31 in van der Vaart (2002); similar techniques to those have also been employed in Kennedy (2019).

Lemma B.1 (Asymptotic Normality for Discrete-Time Estimators)**.** *Let $k \in \mathbb{N}$ and $t_1, \ldots, t_k \in \mathcal{T}$ be fixed. Define*

$$\widehat{\boldsymbol{\psi}}_{t_1,\ldots,t_k} = \left[\widehat{\psi}(t_1), \ldots, \widehat{\psi}(t_k) \right]^\top, \quad \boldsymbol{\psi}_{t_1,\ldots,t_k} = \left[\psi_d(t_1), \ldots, \psi_d(t_k) \right]^\top.$$

Under Assumptions (A1)-(A5), we have

$$\sqrt{n} \left(\widehat{\boldsymbol{\psi}}_{t_1,\ldots,t_k} - \boldsymbol{\psi}_{t_1,\ldots,t_k} \right) \xrightarrow{d} N(0, \Sigma_{t_1,\ldots,t_k}),$$

where the covariance matrix is given by

$$\Sigma_{t_1,\ldots,t_k} = \mathbb{E}\left[\boldsymbol{\varphi}_{t_1,\ldots,t_k}(\mathcal{D}) \boldsymbol{\varphi}_{t_1,\ldots,t_k}(\mathcal{D})^\top \right],$$

for $\boldsymbol{\varphi}_{t_1,\ldots,t_k}(\mathcal{D}) = \left(\varphi(t_1; \mathcal{D}), \ldots, \varphi(t_k; \mathcal{D}) \right)^\top$.

Proof. For notational simplicity, let $\boldsymbol{\varphi} = \boldsymbol{\varphi}_{t_1,\ldots,t_k}(\mathcal{D})$ and $\widehat{\boldsymbol{\varphi}} = \widehat{\boldsymbol{\varphi}}_{t_1,\ldots,t_k}(\mathcal{D}) = \left(\widehat{\varphi}(t_1; \mathcal{D}), \ldots, \widehat{\varphi}(t_k; \mathcal{D}) \right)^\top$. Recall that $\widehat{\varphi}(t; \mathcal{D})$ is defined by

$$\widehat{\varphi}(t; \mathcal{D}) = \hat{\mu}_1(t, X; d) - \hat{\mu}_0(t, X; d) + \left\{ \frac{A}{\hat{\pi}(X)} - \frac{1-A}{1-\hat{\pi}(X)} \right\} \{\phi(t; \mathcal{D}_d) - \hat{\mu}_A(t, X; d)\},$$

where all the nuisance estimators are constructed on a separate, independent sample (Assumption (A2)). Then, consider the following decomposition:

$$\sqrt{n} \left(\widehat{\boldsymbol{\psi}}_{t_1,\ldots,t_k} - \boldsymbol{\psi}_{t_1,\ldots,t_k} \right) = \sqrt{n}(\mathbb{P}_n - \mathbb{P})\boldsymbol{\varphi} + \sqrt{n}(\mathbb{P}_n - \mathbb{P})(\widehat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}) + \mathbb{P}(\widehat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}). \tag{13}$$

By the multivariate central limit theorem, the first term in equation 13 converges to a multivariate Normal distribution with mean 0 and covariance matrix equal to $\Sigma_{t_1,\ldots,t_k}$. The third term in equation 13 is simply $\sqrt{n}R_2$ where R_2 is the remainder term specified in equation 12. Under the rate condition in Assumption (A4), it follows that $\sqrt{n}R_2 = o_{\mathbb{P}}(1)$. Thus, it suffices to show that the second empirical process term is negligible, i.e., of order $o_{\mathbb{P}}(1)$. By the consistency condition (A3) and the triangle inequality,

$$\|\widehat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}\|_{\mathbb{P},2} = O\left(\sum_{j=1}^{k} \|\widehat{\varphi}(t_j) - \varphi(t_j)\|_{\mathbb{P},2} \right) = o_{\mathbb{P}}(1).$$

Applying the multidimensional Chebyshev's inequality to the proof of Lemma 2 in Kennedy et al. (2020), it is immediate to get

$$(\mathbb{P}_n - \mathbb{P})(\widehat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}) = O_{\mathbb{P}}\left(\frac{\|\widehat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}\|_{\mathbb{P},2}}{\sqrt{n}} \right) = o_{\mathbb{P}}(n^{-1/2}).$$

Putting all the pieces together into equation 13, the result follows by Slutsky's theorem. □

Note that Proposition 5.2 follows by Lemma B.1. We are now in a position to prove Theorem 5.3. Here, we impose Assumption (A5′) and follow the proof strategy of Testa et al. (2025). We then show that the same result holds under the weaker Assumption (A5), in Section B.3.

Proof. First, we show the stochastic equicontinuity of $\widehat{\psi}$:

$$\lim_{\delta\to 0}\limsup_{n\to\infty}\mathbb{P}\left[\sup_{|s-t|\le\delta}\left|\widehat{\psi}(s)-\widehat{\psi}(t)\right|\ge\varepsilon\right]=0,$$

for every $\varepsilon>0$. Fix $\delta>0$ and write

$$w(\widehat{\psi};\delta)=\sup_{|s-t|\le\delta}\left|\widehat{\psi}(s)-\widehat{\psi}(t)\right|.$$

For $a\in\{0,1\}$ set $\Delta_a(s,t)=\widehat{\mu}_a(s,X;d)-\widehat{\mu}_a(t,X;d)$ and $R(s,t)=\phi(s;\mathcal{D}_d)-\phi(t;\mathcal{D}_d)-\left\{\widehat{\mu}_A(s,X;d)-\widehat{\mu}_A(t,X;d)\right\}$. Then

$$\widehat{\psi}(s)-\widehat{\psi}(t)=\Delta_1(s,t)-\Delta_0(s,t)+\left\{\tfrac{A}{\widehat{\pi}(X)}-\tfrac{1-A}{1-\widehat{\pi}(X)}\right\}R(s,t).$$

By Assumption (A5), $\mathbb{P}\left[\sup_{|s-t|\le\delta}|\Delta_a(s,t)|\right]\le L_\phi\delta$, $\forall a\in\mathcal{A}$. Since persistent pairs with infinite lifespan are excluded from consideration, we have $\|\phi\|_\infty<\infty$. Further, by Lemma 2.1, $\phi(t)$ is 1-Lipschitz continuous in t, yielding $|\phi(s;\mathcal{D}_d)-\phi(t;\mathcal{D}_d)|\le|s-t|$. Thus, it follows that $\mathbb{P}\left[\sup_{|s-t|\le\delta}|\mathbb{R}(s,t)|\right]\le(L_\phi+1)\delta$. Therefore, by the Markov inequality, for any fixed $\varepsilon>0$, we have

$$\mathbb{P}\left[w(\widehat{\psi};\delta)\ge\varepsilon\right]\le\frac{\mathbb{P}\left[w(\widehat{\psi};\delta)\right]}{\varepsilon}\le\frac{(L_\phi+3)\delta}{\varepsilon},$$

where the second inequality follows by the triangle inequality. Finally, taking $\delta\to 0$ and then $\limsup_{n\to\infty}$ yields $\lim_{\delta\to0}\limsup_{n\to\infty}\mathbb{P}\big[w(\widehat{\psi};\delta)\ge\varepsilon\big]=0$ for every $\varepsilon>0$, as desired. Having established the stochastic equicontinuity of $\widehat{\psi}$, the result follows directly from Lemma B.1 and Theorem 7.5 of Billingsley (1999). □

B.3 Proof of Theorem 5.3 under Assumption (A5)

Assumption (A5′) furnishes the Lipschitz and uniform consistency conditions that guarantee the stochastic equicontinuity, playing a key role in the proof of Theorem 5.3 in Section B.2. Recall that we must verify

$$\lim_{\delta\to0}\limsup_{n\to\infty}\mathbb{P}\Big[\sup_{|s-t|\le\delta}|\widehat{\varphi}(s)-\widehat{\varphi}(t)|\ge\varepsilon\Big]=0,\quad\forall\varepsilon>0.$$

Write $\widehat{\varphi}(t)=\varphi(t)+\big\{\widehat{\varphi}(t)-\varphi(t)\big\}$. Hence

$$\widehat{\varphi}(s)-\widehat{\varphi}(t)=\underbrace{\big\{\varphi(s)-\varphi(t)\big\}}_{\text{(i)}}+\underbrace{\big\{\widehat{\varphi}(s)-\varphi(s)\big\}-\big\{\widehat{\varphi}(t)-\varphi(t)\big\}}_{\text{(ii)}}.$$

We control terms (i) and (ii) separately as follows.

(i). Because each $\mu_a(t,x;d)$ is Lipschitz in t (first part of Assumption (A5′)) and $\phi(t;\mathcal{D}_d)$ is Lipschitz with constant L_ϕ,

$$|\varphi(s)-\varphi(t)|\le L_\varphi|s-t|,\qquad L_\varphi:=2L_\mu+L_\phi.\tag{A}$$

(ii). The uniform consistency conditions in Assumption (A5′) imply that for each $a\in\mathcal{A}$,

$$\sup_{t,x}|\widehat{\mu}_a(t,x;d)-\mu_a(t,x;d)|=o_{\mathbb{P}}(1),\qquad\sup_x|\widehat{\pi}(x)-\pi(x)|=o_{\mathbb{P}}(1).$$

Thus, denoting $\Delta_n:=\sup_t|\widehat{\varphi}(t)-\varphi(t)|$, we have $\Delta_n=o_{\mathbb{P}}(1)$. Consequently, for all s,t,

$$|\widehat{\varphi}(s)-\varphi(s)|+|\widehat{\varphi}(t)-\varphi(t)|\le2\Delta_n=o_{\mathbb{P}}(1).\tag{B}$$

Hence, combining (A) and (B), we obtain the modulus of continuity:

$$\sup_{|s-t|\le\delta}|\widehat{\varphi}(s)-\widehat{\varphi}(t)|\le L_\varphi\delta+2\Delta_n.$$

Choose a deterministic sequence $\delta\downarrow0$ such that $L_\varphi\delta_n\le\varepsilon/2$. Then

$$\mathbb{P}\Big[\sup_{|s-t|\le\delta}|\widehat{\varphi}(s)-\widehat{\varphi}(t)|\ge\varepsilon\Big]\le\mathbb{P}(2\Delta_n\ge\varepsilon/2)\to0,$$

because $\Delta_n=o_{\mathbb{P}}(1)$. Having established stochastic equicontinuity, the remaining steps of the proof proceed unchanged, completing the proof of Theorem 5.3.

B.4 Proof of Theorem 5.4

We first introduce several definitions that will be used in the development of our stability results.

Definition B.1 (Matching)**.** *Let $\triangle$ denote the diagonal of a persistence diagram. A* matching *between two persistence diagrams $\mathcal{D}$ and $\mathcal{D}'$ is defined as a subset $m \subset \mathcal{D} \times \mathcal{D}'$ such that every point in $\mathcal{D} \setminus \triangle$ and $\mathcal{D}' \setminus \triangle$ appears exactly once in m.*

Definition B.2 (Wasserstein distance)**.** *The d-*th Wasserstein distance *between two persistence diagrams $\mathcal{D}$ and $\mathcal{D}'$ is*

$$W_d(D, D') = \inf_{\text{matching } m} \left(\sum_{(p,q)\in m} \|p - q\|_\infty^d \right)^{1/d}.$$

We denote the matching m that satisfies Definition B.2 as the *optimal W_d matching*.

Let $\phi = \phi(t; \mathcal{D})$ and $\phi' = \phi(t; \mathcal{D}')$ be the corresponding power-weighted silhouettes for persistence diagrams $\mathcal{D}$ and $\mathcal{D}'$, respectively. Throughout the remainder of this section, any quantity derived from $\mathcal{D}'$ will be denoted with a superscript $'$. With a slight abuse of notation, we let $p = (b_p, d_p) \in \mathcal{D}$, $q = (b'_q, d'_q) \in \mathcal{D}'$, and denote the power weights corresponding to p and q as $w_p = |d_p - b_p|^r$ and $w'_q = |d'_q - b'_q|^r$, respectively, where $0 < r \leq \infty$. Note that under the diagram boundedness condition of Assumption (A6), the absolute value in the power weights may be omitted. Hence, we assume $w_p = (d_p - b_p)^r$ and $w'_q = (d'_q - b'_q)^r$ throughout this section.

We now establish the following lemma, which serves as a preliminary result essential for the development of the main theorem.

Lemma B.2. *Given a matching $m \subset \mathcal{D} \times \mathcal{D}'$, for any $(p, q) \in m$,*

$$|w_p - w'_q| \leq 2r|c_{pq}^{r-1}| \cdot \|p - q\|_\infty,$$

where c_{pq} is some constant that satisfies the Mean Value Theorem for the function $g(x) = x^r, x \in [0, \infty)$, given points p and q. Consequently,

$$|w_p - w'_q| \leq 2rc^{r-1}\|p - q\|_\infty,$$

for

$$c = \max_{(p,q)\in m} \max\{\, d_p - b_p,\ d'_q - b'_q \,\},$$

so c depends only on the weighting exponent r (through the bound) and on the largest matched lifetime across $\mathcal{D}$ and $\mathcal{D}'$.

Proof.

$$\begin{aligned}
|w_p - w'_q| &= |(d_p - b_p)^r - (d'_q - b'_q)^r| \\
&= |rc_{pq}^{r-1}\{(d_p - b_p) - (d'_q - b'_q)\}| \\
&= r|c_{pq}^{r-1}| \cdot |(d_p - b_p) - (d'_q - b'_q)| \\
&\leq r|c_{pq}^{r-1}| \cdot \{|b_p - b'_q| + |d_p - d'_q|\} \\
&\leq 2r|c_{pq}^{r-1}| \cdot \max\{|b_p - b'_q|, |d_p - d'_q|\} \\
&= 2r|c_{pq}^{r-1}| \cdot \|p - q\|_\infty,
\end{aligned}$$

where the second equality uses the Mean Value Theorem. □

Note that a matching–independent choice for c, which also suffices is

$$c = \max\{\, d_x - b_x : x \in \mathcal{D} \cup \mathcal{D}' \,\}.$$

Notice that when $r = 1$ the bound does not involve c (cf. the special case in the theorem).

Now we give the proof of Theorem 5.4.

Proof. Let $w_p = (d_p - b_p)^r$, $w'_q = (d'_q - b'_q)^r$, and define

$$S = \sum_{p\in\mathcal{D}} w_p, \qquad S' = \sum_{q\in\mathcal{D}'} w'_q, \qquad \phi = \frac{1}{S}\sum_{p\in\mathcal{D}} w_p\Lambda_p, \qquad \phi' = \frac{1}{S'}\sum_{q\in\mathcal{D}'} w'_q\Lambda'_q,$$

where Λ_p denotes the tent function centered at p, i.e., $\|\Lambda'_q\|_\infty = (d'_q - b'_q)/2$. Augmenting the diagrams with the diagonal if needed, the optimal W_1 matching m^* is a bijection between $\mathcal{D}$ and $\mathcal{D}'$, hence

$$\|\phi - \phi'\|_\infty = \left\| \sum_{(p,q)\in m^*} \Big(\frac{w_p\Lambda_p}{S} - \frac{w'_q\Lambda'_q}{S'} \Big) \right\|_\infty \le \sum_{(p,q)\in m^*} \left\| \frac{w_p\Lambda_p}{S} - \frac{w'_q\Lambda'_q}{S'} \right\|_\infty .$$

Split each summand as

$$\left\| \frac{w_p\Lambda_p}{S} - \frac{w'_q\Lambda'_q}{S'} \right\|_\infty \le \left\| \frac{w_p}{S}(\Lambda_p - \Lambda'_q) \right\|_\infty + \|\Lambda'_q\|_\infty \left| \frac{w_p}{S} - \frac{w'_q}{S'} \right|.$$

Summing over $(p,q) \in m^*$ yields

$$\|\phi - \phi'\|_\infty \;\le\; T_1 + T_2, \qquad T_1 := \frac{1}{S}\sum_{(p,q)\in m^*} w_p\|\Lambda_p - \Lambda'_q\|_\infty, \quad T_2 := \sum_{(p,q)\in m^*} \|\Lambda'_q\|_\infty \left| \frac{w_p}{S} - \frac{w'_q}{S'} \right|.$$

Bound for T_1. By the 1-Lipschitz property of the tent functions, we get $\|\Lambda_p - \Lambda'_q\|_\infty \le \|p - q\|_\infty$. Therefore,

$$T_1 \;\le\; \frac{1}{S}\sum_{(p,q)\in m^*} w_p\|p - q\|_\infty \;\le\; \frac{1}{S}\Big(\sum_{p\in\mathcal{D}} w_p\Big)\Big(\sum_{(p,q)\in m^*} \|p - q\|_\infty\Big) \;=\; W_1(\mathcal{D}, \mathcal{D}'),$$

where the second inequality follows by a simple corollary of Hölder's inequality:

$$\sum_i a_i x_i \;\le\; \|a\|_1 \, \|x\|_\infty \;\le\; \|a\|_1 \, \|x\|_1 \;=\; \Big(\sum_i a_i\Big)\Big(\sum_i x_i\Big),$$

for $a_i, x_i \ge 0$, and the last equality is the definition of the W_1 cost of m^*.

Bound for T_2. Observe the algebraic identity

$$\left| \frac{w_p}{S} - \frac{w'_q}{S'} \right| = \frac{|(S' - S)w_p + S(w_p - w'_q)|}{SS'} \;\le\; \frac{|S' - S|\, w_p}{SS'} + \frac{|w_p - w'_q|}{S'}.$$

Hence $T_2 \le T_{2a} + T_{2b}$, where

$$T_{2a} := \sum_{(p,q)\in m^*} \|\Lambda'_q\|_\infty \frac{|S' - S|\, w_p}{SS'}, \qquad T_{2b} := \sum_{(p,q)\in m^*} \|\Lambda'_q\|_\infty \frac{|w_p - w'_q|}{S'}.$$

We first tackle the term T_{2a}. Using Hölder's inequality to separate the sums and cancel S, we have

$$T_{2a} \;\le\; \Big(\sum_{q\in\mathcal{D}'} \|\Lambda'_q\|_\infty\Big) \frac{|S' - S|}{S'}.$$

Since $S' - S = \sum_{(p,q)\in m^*}(w'_q - w_p)$, the triangle inequality and Lemma B.2 yield

$$|S' - S| \;\le\; \sum_{(p,q)\in m^*} |w_p - w'_q| \;\le\; \sum_{(p,q)\in m^*} 2r\, c_{pq}^{r-1}\, \|p - q\|_\infty \;\le\; 2r\, c^{r-1} \sum_{(p,q)\in m^*} \|p - q\|_\infty.$$

Moreover, $\sum_{q\in\mathcal{D}'} \|\Lambda'_q\|_\infty = \frac{1}{2}\sum_{q\in\mathcal{D}'}(d'_q - b'_q)$ and $S' = \sum_{q\in\mathcal{D}'}(d'_q - b'_q)^r$, so by the boundedness condition from Assumption (A6),

$$\frac{\sum_{q\in\mathcal{D}'} \|\Lambda'_q\|_\infty}{S'} = \frac{1}{2} \cdot \frac{\sum_q (d'_q - b'_q)}{\sum_q (d'_q - b'_q)^r} \;\le\; \frac{L}{2}.$$

Therefore

$$T_{2a} \;\le\; \frac{L}{2} \cdot 2r\, c^{\,r-1} \sum_{(p,q)\in m^*} \|p-q\|_\infty \;=\; Lr\, c^{\,r-1}\, W_1(\mathcal{D},\mathcal{D}').$$

The analysis of T_{2b} follows the same steps as above. Again by Lemma B.2 and the ratio boundedness condition from Assumption (A6),

$$T_{2b} \;\le\; \frac{\sum_{q\in\mathcal{D}'} \|\Lambda'_q\|_\infty}{S'} \sum_{(p,q)\in m^*} |w_p - w'_q| \;\le\; \frac{L}{2} \sum_{(p,q)\in m^*} 2r\, c_{pq}^{\,r-1}\, \|p-q\|_\infty \;\le\; Lr\, c^{\,r-1}\, W_1(\mathcal{D},\mathcal{D}').$$

Combining the bounds for T_{2a} and T_{2b} gives

$$T_2 \;\le\; 2Lr\, c^{\,r-1}\, W_1(\mathcal{D},\mathcal{D}').$$

Putting together the bounds for T_1 and T_2, we obtain

$$\|\phi - \phi'\|_\infty \;\le\; \big(1 + 2Lr\, c^{\,r-1}\big)\, W_1(\mathcal{D},\mathcal{D}').$$

Special case $r = 1$. When $r = 1$, Lemma B.2 gives $|w_p - w'_q| \le 2\|p-q\|_\infty$ and

$$\frac{\sum_{q\in\mathcal{D}'} \|\Lambda'_q\|_\infty}{S'} = \frac{\sum_q (d'_q - b'_q)/2}{\sum_q (d'_q - b'_q)} = \frac{1}{2},$$

so each of T_{2a} and T_{2b} is bounded by $W_1(\mathcal{D},\mathcal{D}')$, while $T_1 \le W_1(\mathcal{D},\mathcal{D}')$, which yields $\|\phi - \phi'\|_\infty \le 3\, W_1(\mathcal{D},\mathcal{D}')$. □

C Experiment Details

We perform experiments on three benchmark datasets: the SARS-CoV-2 CT-scan image dataset, the GEOM-Drugs molecular graph dataset, and the ORBIT point cloud dataset. In all experiments, we construct a synthetic counterfactual dataset $\{(X_i, A_i, Y_i^0, Y_i^1)\}_{i=1}^n$ to facilitate the evaluation of estimators against a known true effect. We begin by randomly selecting and pairing two data samples to form each potential outcome pair, assigning one to Y^0 and the other to Y^1 in a manner that induces a clear topological contrast. Next, we generate the covariates X and treatment A according to a stochastic data-generating process and treatment assignment mechanism. The same procedures are identically applied to all experiments, with details outlined below.

Data-generating process. We assume a setting in which a subgroup structure is imposed on the covariates $X \in \mathbb{R}^5$. The covariates are generated from two subgroups, each governed by a multivariate Gaussian distribution with distinct mean vectors μ_1, μ_2, and a common covariance matrix Σ. The covariance matrix Σ is set as a diagonal matrix with standard deviation 0.5, and the mean vector of each subgroup is specified as $\mu_1 = [1, 0.6, -0.7, 2.2, -1]^T$ and $\mu_2 = [0.4, -0.4, -0.6, 3.3, 3.]^T$. Covariates for half of the samples are generated from $N(\mu_1, \Sigma)$, while the remaining half are generated from $N(\mu_2, \Sigma)$.

Treatment mechanism. Given the covariates X, treatment $A \in \{0, 1\}$ is assigned with probability $\pi(X) = expit(-0.5X_1 - 0.1X_2 + 0.6X_3 + 0.1X_4 + 0.1X_5 + 0.5X_2X_3 - 0.7X_1X_3)$. This treatment mechanism is designed such that one subgroup has a higher probability of receiving treatment than the other (see Figure 7-(a)). Upon treatment assignment, the observed outcome is given by $Y = AY^1 + (1 - A)Y^0$.

The aforementioned procedure is repeated 20 times to generate multiple datasets with different realizations of X and A, for each of which the PI, IPW, and AIPW estimators are computed. The estimators are constructed by modeling the silhouette regression function μ_a with function-on-scalar regression employing a Fourier basis expansion, and estimating the propensity score π using a random forest classifier. To assess the performance of each estimator, we examine the pointwise mean and the pointwise 1-standard deviation error bands computed across the 20 estimates.

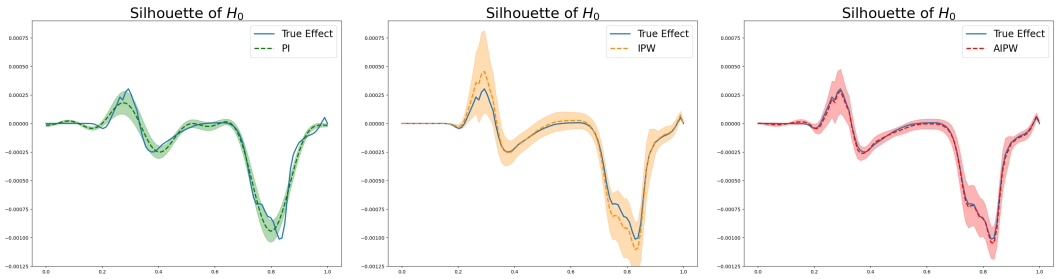
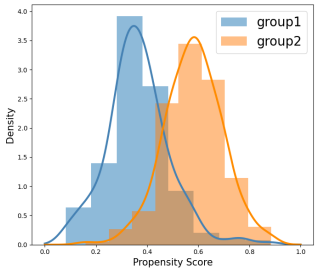
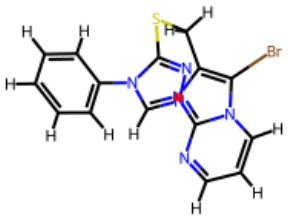


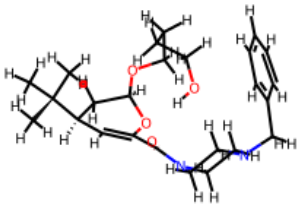
Figure 6: Visualization of the point-wise mean (dotted line) and the point-wise 1-standard deviation error bands (shaded area) of PI, IPW, and AIPW estimators on the SARS-CoV-2 dataset. The true topological causal effect is shown as a blue line.



(a) Density of propensity score by subgroup (Group 2 more likely to be treated).



(b) Sample molecular graph A (Axelrod & Gomez-Bombarelli, 2022).



(c) Sample molecular graph B (Axelrod & Gomez-Bombarelli, 2022).

Figure 7: (a) Propensity score distribution by subgroup; (b–c) representative molecular graphs.

C.1 SARS-CoV-2

The SARS-CoV-2 dataset contains CT-scans collected from real patients in Brazil, who are infected or non-infected by COVID-19. In our experiment, 500 potential outcomes pairs (Y^0, Y^1) are constructed according to the following procedure: (i) 500 infected images are sampled and assigned to Y^0, (ii) 375 non-infected and 125 infected images are sampled and assigned to Y^1, (iii) Y^0 and Y^1 are randomly paired. By experimental design, TATE exhibits treatment effect since 75% of Y^1 is non-infected where as every individual in Y^0 is infected. Silhouettes are computed using sublevel set filtration on a filtered cubical complex (see Appendix A) with $r = 0.1$. For the nuisance estimators, function-on-scalar regression is implemented with 10 basis functions, while random forest is constructed with 100 trees. The point-wise 1-standard deviation error bands of the respective estimators are demonstrated in Figure 6.

C.2 GEOM-Drugs

The GEOM-Drugs dataset consists of graph-structured representations of molecular compounds, as shown in Figure 7-(b) and (c). To analyze graph data, we adopt persistent homology transform (10) with three directions, yielding three separate estimates per homology dimension for each estimator. In our experiment, 1000 potential outcomes pairs (Y^0, Y^1) are constructed according to the following procedure: (i) sample 2000 graph data from original dataset, (ii) assign samples with large silhouette magnitudes to Y^1, and the rest to Y^0, (iii) randomly pair Y^0 and Y^1. This allocation scheme gives rise to significant topological differences between the counterfactual groups. We compute silhouettes with $r = 1$, while the nuisance estimators for μ_a and π are specified with 5 basis functions and 100 trees, respectively. Figure 8 illustrates the PI, IPW, AIPW estimators for 1-dimensional silhouettes, where similar to the previous result, the AIPW estimator consistently provides the most accurate and stable approximation of the true effect across all three directional settings. Figures 9 and 10 provide a visualization of the point-wise mean and the point-wise 1-standard deviation error bands of the estimators for each of the three directions per homology dimension.

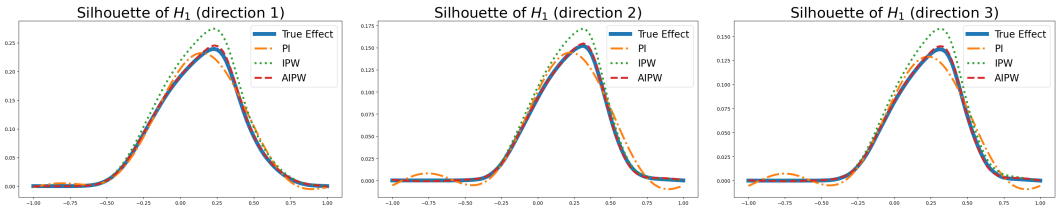


Figure 8: 1-dimensional true silhouette functions and its PI, IPW, AIPW estimates along three directions using persistent homology on the GEOM-Drugs dataset.

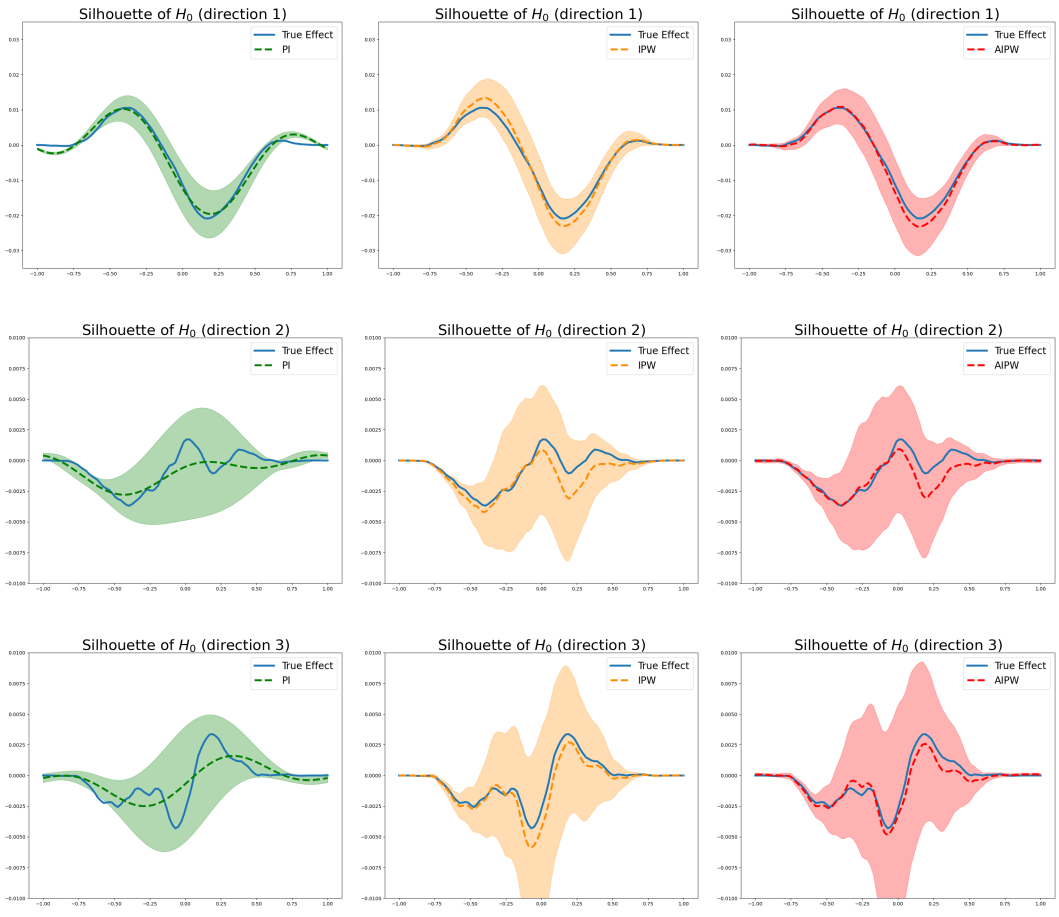


Figure 9: Visualization of the point-wise mean (dotted line) and the point-wise 1-standard deviation error bands (shaded area) of 0-dimensional PI, IPW, and AIPW estimators on the GEOM-Drugs dataset for three directions. The true topological causal effect is shown as a blue line.

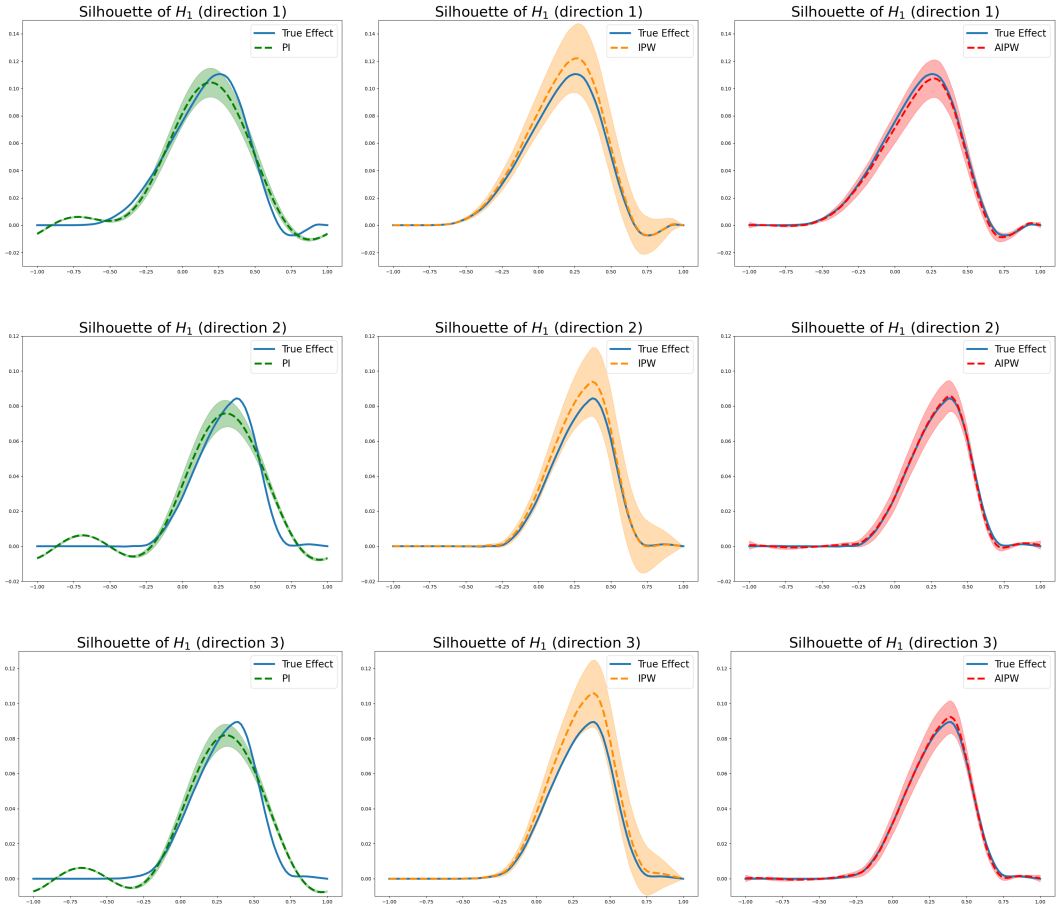


Figure 10: Visualization of the point-wise mean (dotted line) and the point-wise 1-standard deviation error bands (shaded area) of 1-dimensional PI, IPW, and AIPW estimators on the GEOM-Drugs dataset for three directions. The true topological causal effect is shown as a blue line.

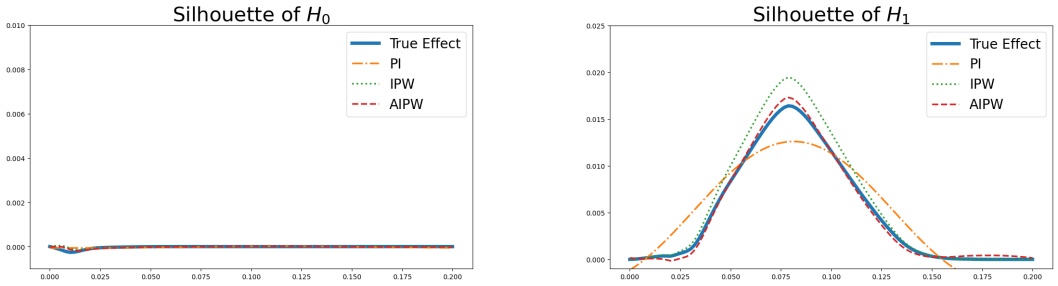


Figure 11: 0-dimensional (left) and 1-dimensional (right) true silhouette functions and its PI, IPW, AIPW estimates for the ORBIT dataset.

C.3 ORBIT

The ORBIT dataset (Adams et al., 2017) is a synthetic point cloud dataset that is generated by simulating different dynamical systems characterized by a parameter r. Given a random initial point $(x_0, y_0) \in [0, 1]^2$ and $r > 0$, we generate point clouds consisting of 1000 points as follows:

$$x_{n+1} = x_n + r y_n(1 - y_n) \mod 1$$
$$y_{n+1} = y_n + r x_n(1 - x_n) \mod 1,$$

In this experiment, we use $r = 3.5, 4, 4.1$ to generate 1000 samples for each value of r, with the resulting point clouds illustrated in Figure 2-(a): $r = 3.5$ (top right), $r = 4$ (bottom), $r = 4.1$ (top left). From each triplet of point clouds generated by $r = 3.5, 4, 4.1$, we randomly select two point clouds and assign the one corresponding to the higher r value as Y^1 with probability 0.7. This procedure yields pairs of matched potential outcomes (Y^0, Y^1) for all 1000 samples, where the treated potential outcome is intentionally designed to possess more pronounced topological features. Here, we use 3 bases to model the silhouette regression function and 100 trees to estimate the propensity score, and the silhouettes are computed using Alpha filtration (see Appendix A) with power weight $r = 3$.

Results. Figure 11 illustrates the true target silhouettes in homology dimensions 0 and 1, which clearly demonstrate a causal treatment effect on first-order homological features of point clouds. In particular, the positive values of the 1-dimensional silhouette indicate the emergence of holes in the treated point cloud. Our aim is to recover the magnitude of this silhouette function, as larger values correspond to more substantial structural changes. Consistent with previous results, the IPW estimator tends to overestimate the treatment effect, whereas the plug-in estimator underestimates it. The AIPW estimator, by contrast, yields a substantially more accurate estimate of the true silhouette function. Moreover, the near-zero silhouette for 0-dimensional homology features suggests that the data's 0-dimensional homology, including connected components, remained largely unchanged after treatment. Figure 12 illustrates the pointwise mean and the pointwise 1-standard deviation error bands for each of the PI, IPW, and AIPW estimators on the ORBIT dataset. The AIPW estimator exhibits near-perfect alignment with the true causal effect, whereas the IPW and plug-in estimators fail to encompass the true effect within their respective error intervals.

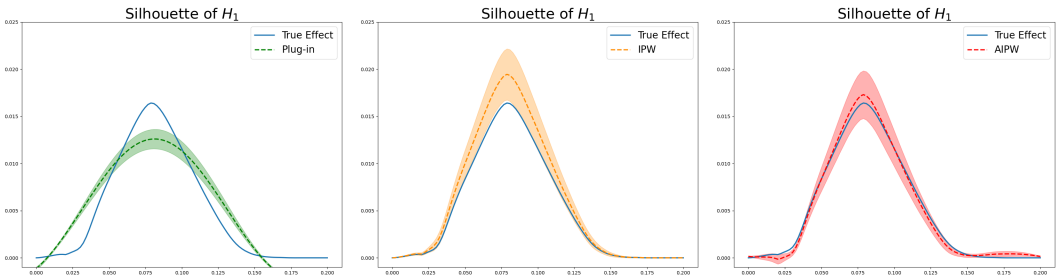


Figure 12: Visualization of the pointwise mean (dotted line) and the pointwise 1-standard deviation error bands (shaded area) of 1-dimensional PI, IPW, and AIPW estimators on the ORBIT dataset. The true topological causal effect is shown as a blue line.