Editors' Suggestion

Applications of persistent homology in nuclear collisions

Greg Hamilton,[1] Travis Dore,[2,3] and Christopher Plumberg[2,4,*]
[1]*Institute for Condensed Matter Theory, Department of Physics, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA*
[2]*Illinois Center for Advanced Studies of the Universe, Department of Physics, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA*
[3]*Fakultät für Physik, Universität Bielefeld, D-33615 Bielefeld, Germany*
[4]*Natural Science Division, Pepperdine University, Malibu, California 90263, USA*

(Received 11 October 2022; accepted 29 November 2022; published 23 December 2022)

We introduce a novel set of observables associated to the rapidly developing field of *persistent homology* for the quantitative characterization of nuclear collisions and their evolution. Persistent homology allows for the identification of topological and homological characteristics of distributions in multidimensional spaces. We demonstrate here how to apply the tool kit of persistent homology to the extraction of novel clustering signatures and the identification of long-range flow correlations in the particle production process of nuclear collisions.

DOI: 10.1103/PhysRevC.106.064912

I. INTRODUCTION

The field of relativistic nuclear collisions exists to explore the properties of quantum chromodynamic (QCD) matter at asymptotically large temperatures and densities [1,2]. In addition to creating a novel phase of deconfined matter known as the *quark-gluon plasma* (QGP) [3], nuclear collisions provide insights into the equation of state of nuclear matter [4–6], conjectured topological characteristics of the QCD vacuum [7–9], input into models of neutron star structure and mergers [10–12], and much more [13–15]. To date, a vast number of observables have been used to probe nuclear collisions, including particle multiplicities [16,17], p_T and rapidity distributions [18,19], anisotropic flow [20–24], jet quenching [25–27], fluctuations and correlations of conserved charges [28,29], interferometry and femtoscopy [30,31], and a litany of others [32–35].

What all of these observables have in common is that they are constructed from *point clouds*. A point cloud, as defined in this paper, is simply a distribution of points in some d-dimensional space (cf. Fig. 1). Point clouds generically arise as finite samples from an underlying continuous distribution and may reflect nontrivial topological structure present in the latter. In the case of nuclear collisions, each collision (or "event") emits a number of particles which are detected, and whose three-momenta can be measured experimentally. The fundamental insight of this paper is to treat these emitted particles as a point cloud in momentum space, where each particle exists as a point with three-dimensional coordinates given by its three-momentum $\vec{p}$ as measured by the detector. One therefore has access experimentally to an *ensemble* of point clouds which can be mined for insights into the underlying dynamics and properties of the nuclear collisions which produced them.

The tool kit of persistent homology (PH) has been designed for exactly this purpose. PH has developed rapidly in recent years as one of the foremost techniques for nonparametrically identifying important topological characteristics of large datasets, including point cloud distributions. PH is best suited to identifying topological or homological features of a given point cloud, including clustering, bubbles, filaments, holes, walls, and so on. It has been applied in a vast number of other disciplines, including the description and evolution of cosmic structure [36–38], Bose-Einstein condensates [39], phase diagrams [40], confinement in non-Abelian lattice gauge theory [41], and even the assembly and disassembly of multispecies ecological systems [42]. While PH yields access to topological features at varying degrees of scales, it also implicitly probes multiorder correlational structure. Indeed, PH has strong connections to robust results in Morse theory and quantifies large-scale structure much like the Minkowski functionals for convex bodies, which can be interpreted in terms of integrated connected correlation functions at all orders [43–45]. Further, deep connections to the Gauss-Bonnet theorem and the Euler characteristic [43] render PH a promising extension of traditional statistical methods for analyzing discrete point clouds, and make it a natural candidate for developing new ways of probing nuclear collisions.

Our topological approach complements and extends a recent explosion of machine learning techniques in high-energy physics, most prominently tools utilizing neural networks like graph neural networks [46]. What is more, persistent homology is easily folded into a machine learning pipeline and can help identify the topological structure of information as it passes through the layers of a neural network [47]. Developing a formalism for applying PH to nuclear collision phenomenology therefore broadens the possible avenues for connecting it to the field of machine learning.

*christopher.plumberg@pepperdine.edu

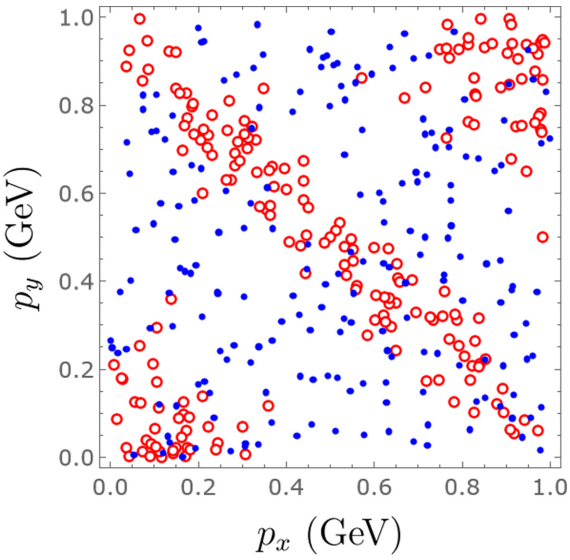

FIG. 1. A toy example of a point cloud in two-dimensional momentum space. Each point represents a particle emitted from an event, where the point cloud distributions of the solid blue points clearly differ from that of the open red circles. These are the sort of qualitative features which persistent homology is designed to access.

In this paper, we show how the PH tool kit can be applied to the description of nuclear collisions and the momentum-space point clouds they produce. For illustrative purposes, and as a proof of concept, we use it to introduce novel ways of quantifying both clustering effects and anisotropic collective flow in realistic nuclear collision simulations. Our goal is to exhibit some of the ways in which PH could be leveraged to provide new insights into the evolution and phenomenology of nuclear collisions.

The outline of this paper is as follows. In Sec. II, we provide a brief review of the main results, tools, concepts, and techniques employed in PH, and summarize the pipeline by which we analyze a given point cloud in this work. Then, in Sec. III, we discuss specifically how we apply these results in the context of nuclear collisions, and introduce several PH observables which are designed to probe familiar nuclear collision phenomenology. Sections IV and V present an illustrative proof of concept for the application of our novel methods to realistic nuclear collisions, based on established simulation packages for modeling the real-time evolution of these systems. Finally, we conclude with an assessment of the prospects for applying PH in nuclear collisions and suggest further questions which our novel approach could help to clarify.

II. CONCEPTS OF PERSISTENT HOMOLOGY

In this section we present an overview of PH, focusing especially on its use as a tool to investigate higher-order correlational structures. We begin with a brief description of PH in general and introduce the pipeline we use to apply PH to simulated nuclear collision data. For the sake of clarity, we illustrate the most important concepts from our pipeline using a toy dataset which contains artificial topological structures that are naturally probed by PH.

A. PH overview

At a high level, PH quantifies how topology persists with respect to a variational parameter. This variational parameter introduces a nesting (or *filtration*) of topological spaces. Each of these topological spaces is typically triangulated, yielding what is formally known as a *simplicial complex* [an example is shown in Fig. 2(b)]. In turn, by making use of the techniques of simplicial homology [48], each simplicial complex in a filtration can be connected with a sequence of homology groups. The key insight of persistent homology is to track how these homology groups change as a function of the filtration parameter, thus providing insight into the topological structure reflected in the point cloud. The output of this process can be usually represented by a plot known as a *persistence diagram*, and may be analyzed further by means of various observables designed to isolate relevant features of a given point cloud ensemble.

There are several excellent reviews [49,50] which we encourage the reader to consult for further discussion of PH in general. In the remainder of this section we provide a description of the specific PH pipeline which we apply to the output of simulated nuclear collisions. The pipeline consists of three main steps, described below, with technical details of each step deferred to Appendix A.

B. Summary of PH pipeline

We illustrate the results of our pipeline when applied to a toy dataset in Fig. 2. The dataset itself is a two-dimensional (2D) Euclidean point cloud, denoted X, and depicted in Fig. 2(a). While X is a point cloud (and therefore has trivial topological structure), it was sampled from a continuous distribution consisting of two concentric annuli, and clearly reflects the nontrivial topology of the latter. Some additional points have also been added as noise. The topologically nontrivial loop structures exhibited by this toy dataset can be characterized by following three main steps: (1) performing a Delaunay triangulation and associated field estimation, (2) conducting a superlevel set filtration, and (3) identifying homological features of interest and evaluating relevant observables which quantify these features. We now discuss each of these three steps in greater detail.

1. Delaunay triangulation and field estimation

First, we generate a Delaunay triangulation of X, as shown in Fig. 2(b). The Delaunay triangulation is a nonparametric triangulation such that, in Euclidean space, no points appear in the circumcircle interior of any triangle [51]. The boundary of the Delaunay triangulation is called the *convex hull* of the point cloud.

After constructing the triangulation, we use a technique known as Delaunay triangulation field estimation (DTFE) to define a density field $f(x)$ on the points $x \in X$ [52]. The DTFE assigns to each point a density which depends upon the area of adjacent triangles and thus correlates closely with the density

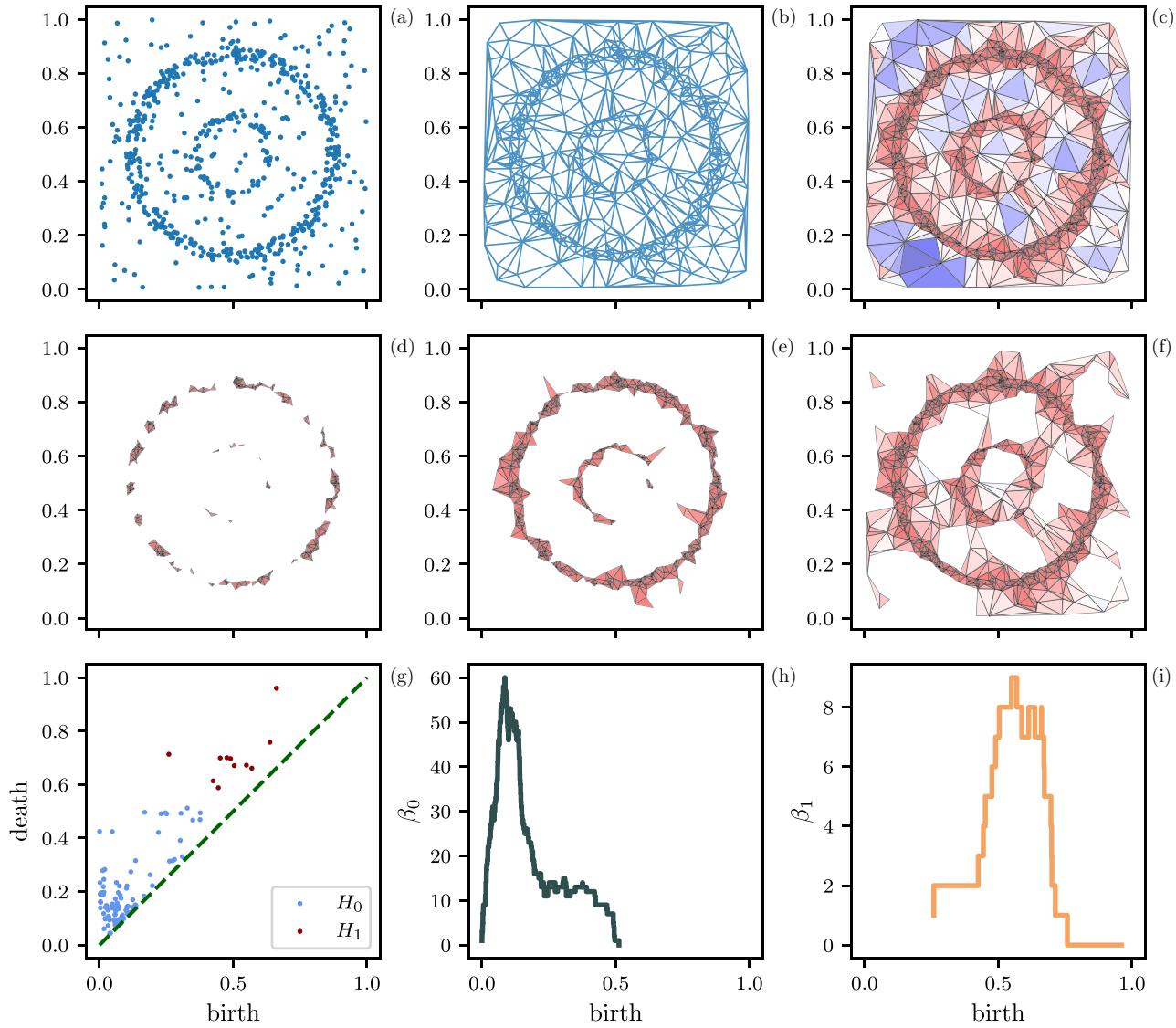

FIG. 2. PH pipeline used in this work. (a) Random point cloud with multiplicity $n \approx 2000$. (b) The Delaunay triangulation; note the prevalence of long "sliver" triangles at the boundary. (c) The DTFE with the color of any triangle roughly scaling with average density of its vertices: warm (red) colors denote high-density regions, while cool (blue) colors denote low-density regions. The precise determination of the densities is described in the main text. Superlevel set of the density field whereby (d) 95%, (e) 75%, and (f) 50% of the points have density lower than the threshold ε. (g) The resultant persistence diagram for H_0, H_1. (h) Betti curve for zero dimensions. (i) Betti curve for one dimensions.

of neighboring points in the vicinity (cf. Appendix A for more details). The results of the DTFE are depicted in Fig. 2(c). The density field has been normalized to the range [0,1], although we use the unnormalized density when analyzing nuclear collisions below. The colors of the triangles correspond roughly to the average $f(x)$ of the vertices of the triangle. The two annuli reflect regions of higher density, and are thus denoted by redder colors; lower-density regions are colored blue.

2. *Superlevel set filtration*

After constructing the Delaunay triangulation and performing the field estimation, we introduce a variational parameter ε and consider the set of points $L_\varepsilon^+ := \{x | f(x) \geqslant \varepsilon\}$, which is the collection of points in the cloud where the density is greater than or equal to ε. This kind of set is referred to as a *superlevel set* of the density field $f(x)$. By definition, we have $L_\varepsilon^+ \subseteq L_{\varepsilon'}^+$ whenever $\varepsilon' \leqslant \varepsilon$. L_ε^+ is thus a filtration, where ε plays the role of the filtration parameter.

For a given value of ε, a simplicial complex K_ε can be constructed in the following way. First, the vertices (or *0-simplices*) in K_ε are identified with the points $x \in L_\varepsilon^+$. Once the 0-simplices have been specified, the 1-simplices are identified as the edges in the triangulation that connect two vertices. Similarly, any triangles are identified as 2-simplices, tetrahedra as 3-simplices, and so on. Thus, for example, any triangle which is formed by three edges (or 1-simplices) in K_ε is "filled in" and included in K_ε as a 2-simplex. The same applies to higher-order simplices.

Figures 2(d)–2(f) show K_ε for three values $\varepsilon_d \geqslant \varepsilon_e \geqslant \varepsilon_f$. The thresholds $\varepsilon_{d/e/f}$ in the figure correspond to points in, respectively, the top 5%, the top 25%, and the top 50% of the densities provided by the DTFE. Note that each simplicial complex is a subcomplex of the subsequent complex. Note also that the outer annulus is a fully connected component in Fig. 2(d), and that the "loop" structure is clearly formed. A particular filtration thus generates a corresponding sequence of nested simplicial complexes which can reflect topological structures underlying the original point cloud.

3. Identify homological features

Once a filtration and sequence of simplicial complexes have been defined, the associated homology groups follow immediately. For any ε, we compute the simplicial homology of K_ε to obtain a direct sum of the homology groups (vector spaces) reflected in K_ε. The rank of each homology group is known as its *Betti number*, β_i, where i denotes the dimension. Intuitively, β_i counts the number of homological features of dimension i: β_0 is the number of connected components, β_1 is the number of nonhomologous loops, and so on. In Figs. 2(h) and 2(i) we show the Betti numbers β_0 and β_1, respectively, as a function of ε; these plots are known as *Betti curves*. The nesting of simplicial complexes ensures a common basis to track which homology groups persist through the filtration $\{\varepsilon\}_{\varepsilon\in[0,1]}$.

Finally, by tracking the homology groups through the filtration one obtains a *persistence diagram* (PD), shown in Fig. 2(g). The abscissa ("birth" or b axis) is the filtration value at which a homological feature first appears, while the ordinate ("death" or d axis) indicates at which filtration value the same homological feature vanishes. The blue markers, denoted H_0 to indicate the zero-dimensional (0D) homology group, correspond to the point cloud's connected components. H_1 denotes the one-dimensional (1D) homology group and represents the loops (formally, *cycles*) present in the point cloud. The distribution of points in the PD thus serves as a "fingerprint" for the topology of the point cloud: the farther points in the PD are from the diagonal, the "longer-lived" the corresponding homological feature. We define the difference $d - b$ as the "lifetime" of the homological feature. Thus, long-lived features correspond to large-scale structure, while short-lived features correspond to noise and local curvature [47,53].

The example given here is identical to the PH pipeline we apply to our simulated collision data, save a few caveats. In this work we consider point clouds in (ϕ, y) coordinates, where

$$p_x = p_T \cos\phi, \tag{1}$$

$$p_y = p_T \sin\phi, \tag{2}$$

$$p_z = m_T \sinh y, \tag{3}$$

and y is the rapidity $m_T \equiv \sqrt{m^2 + p_T^2}$. For consistency, we must have periodic boundary conditions in the ϕ direction, which complicates the Delaunay triangulation and tends to generate spurious edge effects induced by the subsequent DTFE. To avoid these complications, we impose a rapidity cut $|y| \leqslant 2$, which implies that our observables outlined below are defined within this rapidity interval. We discuss this further in Appendix A.

Furthermore, the pipeline we present here is not the only way PH could be applied to nuclear collisions. For instance, although we have employed (ϕ, y) coordinates in this work, one could also consider analyzing particle spectra in three dimensions using coordinates $\vec{p} = (p_x, p_y, p_z)$. Similarly, we use a density-based filtration previously applied in the context of cosmological models for galactic morphology [54], but there are several alternative ways to perform PH on a point cloud as well, as we discuss in Appendix B. We leave a more in-depth analysis of these various possibilities and extensions to future work.

III. OBSERVABLES

While in principle the PD contains a great deal of information about a single point cloud's persistent topology, in practice it can be difficult to analyze its *statistical* properties for an ensemble of point clouds. For this reason, it is often desirable to introduce a scalar quantity or functional (known as a *topological summary*), derived from the PD, for which one can readily formulate and quantify relevant statistical properties. Many ways to do this have been discussed in the literature, including persistence landscapes [55], persistence images [56], and statistics on the birth, death, and lifetime distributions [57]. Each topological summary yields unique insights into fluctuations of the intrinsic topology of an ensemble of point clouds. In this work we focus on four such summaries, two of which actually incorporate more information than the PD alone. These summaries thus provide observables which can be applied to an ensemble of nuclear collisions.

A. Fractal dimension

The first observable we consider is known as the *fractal dimension*. While the birth and death distributions of homological features are interesting in their own right, the *lifetime* distribution in particular quantifies how persistent topological features are in the underlying point cloud. Moreover, we can study how the lifetime distribution changes as the size (or *multiplicity*) of a point cloud representing some dynamical process increases. The scaling of persistent topology with respect to multiplicity thus gives rise to a notion of fractality which we can quantify, and which was formally described in Ref. [58].

Given a point cloud with multiplicity n and its corresponding PD, let PD_i denote the restriction to homological features of dimension i. Then let $E^i_\alpha := \sum_{(b,d)\in PD_i}(d-b)^\alpha$ denote a sum of powers of the lifetimes in PD_i. The fractal dimension is then defined as

$$\dim PH_i := \frac{\alpha}{1-\beta}, \quad \beta = \lim_{n\to\infty}\frac{\log\langle E^i_\alpha\rangle}{\log n}. \tag{4}$$

Here $\langle\cdot\rangle$ denotes averaging over persistence diagrams with the same multiplicity n [58]. Intuitively, Eq. (4) measures how the sum of powers of the lifetimes scales with multiplicity. Small values of α emphasize the small lifetime features (e.g.,

local clustering), while large α probes more global features. The homological fractal dimension defined here has close connections to the box-counting and correlation dimensions, and has been explored in the context of identifying critical exponents and fractal dimensions for dynamical processes [58]. Moreover, a notion of fractal dimension was recently explored in the context of jet classification [59].

B. Betti curves

As noted above, the Betti number β_n is simply the rank of the nth homology group and reflects the number of important topological features of dimension n at a given stage in the filtration. While the Betti curve is insensitive to the relative lifetimes of homological features, it does give an indication of the point in the filtration at which homological features are likely to exist. The structure of the Betti curve, in particular, the maximum, was recently used in the context of identifying phase transitions in quantum many-body systems [60]. As we explore in Sec. V, the Betti curve also serves as a cluster distribution function with respect to scale, i.e., yielding the (unnormalized) probability of m clusters at scale ε. These cluster distribution functions form an important set of observables in cosmological studies of galactic morphology [61], leading us here to consider their relevance for nuclear collisions as well.

C. Cluster entropy

While the 0D Betti curve is a useful indicator for the distribution of clusters with respect to filtration value, the Betti number is insensitive to the multiplicity of individual clusters. For instance, given a point cloud of n points with three clusters, the Betti number β_0 does not distinguish between three clusters with equal numbers of $n/3$ points each, versus one cluster with $n-2$ points and the other two clusters with one point each.

To access this cluster multiplicity information, we exploit the fact that our density-based filtration amounts to a parametrized hierarchical clustering scheme due to the Delaunay triangulation. This hierarchical clustering is quite similar to the clustering scheme used to identify jets [62]; indeed, some collinear, infrared-safe jet clustering algorithms also make use of the Delaunay triangulation in (ϕ, y) space [63]. The output of hierarchical clustering is an object known as a *dendrogram* (or *merge tree*), wherein the lengths of branches between merges indicate the filtration interval in which a given cluster exists. The number of leaves of a branch yields the multiplicity of the cluster at that filtration level. Figure 3 shows a small illustrative example; note that the heights of the leaves are nonuniform because, in our filtration, the leaves appear at values related to the density. The dendrogram thus provides a way of quantifying the distribution of cluster multiplicities as a function of filtration.

Given the number of points in each cluster as a function of the filtration parameter, we introduce a novel observable which we refer to as the *cluster entropy*, which acts as a topological summary of agglomerative clustering. For a point cloud of multiplicity n, at each filtration level ε we define

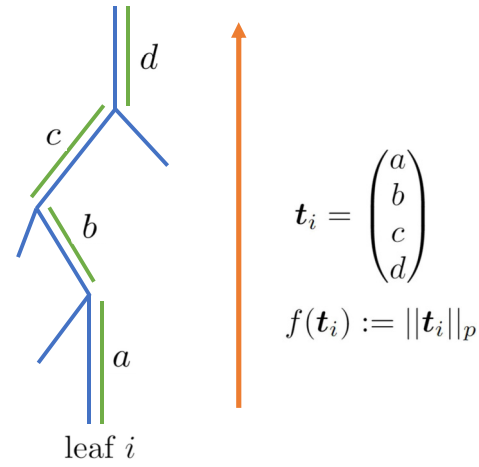

FIG. 3. A pictorial representation of a merge tree, or dendrogram. The filtration from leaves to root runs up the page, with the height of a leaf corresponding to the value at which a hadron appears in the filtration. The length of the green bars from the lowest leaf to the root indicate the lifetimes of the clusters that include the lowest leaf. The set of these lifetimes for each leaf i forms the vector $\boldsymbol{t}_i$.

$$H(\varepsilon) = -\sum_{C_i \in \mathcal{C}(\varepsilon)} p_i \log p_i, \tag{5}$$

where $i \in \mathcal{C}(\varepsilon)$ is the set of clusters at ε and $p_i = |C_i|/n(\varepsilon)$. Here $n(\varepsilon)$ is the number of points that exist at filtration value ε, so that $n(\varepsilon) \leqslant n$. Note that p_i is a proper probability distribution, and so the cluster entropy defined here is the Shannon entropy [64] of the cluster probability distribution. The cluster entropy indicates the degree of "mixedness" in the distribution of points among clusters. This statistic naturally generalizes to the Rényi and Tsallis entropies, which explore the "rarity" or heavy tails of a distribution [64].

D. Local clustering statistics

A particularly important phenomenon frequently studied in nuclear collisions is that of *local clustering*, which can arise in a variety of contexts, including Bose-Einstein correlations [65], the QCD critical point [66–68], jet identification [69], and n-body correlations arising from collective flow [70]. By "local clustering," we mean any significant deviation from a uniformly distributed point cloud which may be correlated with position within the point cloud. In this sense, local clustering provides a generalized notion of ordinary clustering, which implies deviations from a uniform distribution but need not specify where the clustering takes place. Since we wish to characterize nonuniform point cloud distributions using PH in a way which *can* depend on the specific region of momentum space (e.g., for anisotropic flow), it is therefore essential to have a way of quantifying local clustering as well.

However, while persistence diagrams provide structural statistics on point clouds, PH alone does not retain information regarding other, nontopological degrees of freedom, such as whether or not clustering behavior is more likely in one part of the point cloud than another. While PH has been used to

identify anisotropy in point cloud distributions and to quantify local curvature [53], PH does not explicitly retain positional degrees of freedom.

However, the form of PH we pursue in this work comes with an object that does retain positional degrees of freedom: the dendrogram. Each leaf in a dendrogram corresponds to a point in the point cloud, and the lengths of branches between merges indicate how long a cluster persists before being merged into another cluster. To each leaf (point) we identify a vector $\boldsymbol{t}_i$, the components of which are the lengths of the branches along the path from the leaf i to the root of the dendrogram. We depict a simple dendrogram in Fig. 3, wherein the green bars denote the relevant branches from the leaf i (bottom of the plot) up to the root. Every leaf in the dendrogram therefore has a corresponding vector $\boldsymbol{t}_i$. Taking the p-norm of each $\boldsymbol{t}_i$ then yields a novel statistic $f(\boldsymbol{t}_i) := ||\boldsymbol{t}_i||_p$ (which we refer to as the leaf's *p-norm*) on the point cloud that reflects the local clustering statistics: for large p the p-norm emphasizes large-scale clustering, while small $p < 1$ (technically a seminorm) targets local clustering and local curvature. Here the p-norm for $\boldsymbol{t}_i = (t_0, \ldots, t_m)$ is defined as

$$||\boldsymbol{t}_i||_p = \left(\sum_j t_j^p \right)^{1/p}. \tag{6}$$

This novel clustering statistic has to our knowledge not appeared in the topological data analysis literature, though recent approaches to merge trees that incorporate higher-dimensional homological information (known as decorated merge trees) have touched on similar ideas [71–73]. Given the importance of anisotropy in higher-order correlation functions in the context of flow, we see the local clustering statistics observable as a reasonable step towards bridging the gap between persistent homology and traditional correlational metrics in nuclear collisions.

We thus have introduced four observables—the fractal dimension, the Betti curves, the cluster entropy, and the p-norm associated to dendrogram leaves—which can be readily applied to the analysis of nuclear collisions. These observables are designed to characterize different aspects of point clouds typically produced in nuclear collisions. In the next section we discuss the simulation framework to which these observables will be applied.

IV. NUCLEAR COLLISION SIMULATIONS

We now present our approach to generating realistic simulations of Pb+Pb collisions at LHC energies. In addition to applying PH to the simulated Pb+Pb events themselves, it is also crucial to establish a suitable background as a reference against which to compare any proposed signals. Clearly for PH a suitable background should also account for the intrinsic topology of the ambient space in which a point cloud is situated, such as the cylindrical topology of the (ϕ, y) coordinate system. As noted previously, the periodicity in the ϕ coordinate introduces some technical complications which we discuss more fully in Appendix A. Below we discuss the details of our simulation framework and describe how we construct backgrounds for the PH observables we consider here.

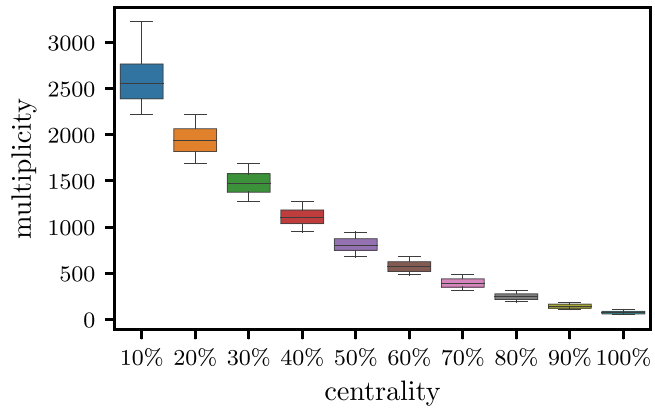

FIG. 4. Box plot showing the distribution of multiplicities as a function of centrality class.

A. Hydrodynamic simulations

In this study, we have considered an ensemble of 10 000 Pb+Pb collision events at $\sqrt{s_{NN}} = 2.76$ TeV, simulated using the Duke Bayesian tune of the iEBE-VISHNU framework to LHC p+Pb and Pb+Pb data [74–76]. This approach couples together T$_\mathrm{R}$ENTo initial conditions [77] with a conformal, prehydrodynamic free-streaming phase [78,79], a boost-invariant hydrodynamic phase [74,80] using the Denicol-Niemi-Molnar-Rischke (DNMR) formalism [81], and a hadronic afterburner UrQMD [82,83]. We use the maximum-likelihood parameters [76] for the transport coefficients. The hydrodynamic evolution is terminated at a hypersurface of constant $T_{FO} = 150$ MeV, corresponding to an energy density of $e_{FO} \approx 0.265\ \mathrm{MeV/fm}^3$. After the hydrodynamic phase is completed, particles are sampled from the freeze-out hypersurface and fed into URQMD. For each collision event, URQMD yields a list of particles which were emitted by that collision, together with their momentum space coordinates, after all rescattering has finished. The particle lists generated by URQMD are then used as input to the PH pipeline described in Sec. II. Since our focus in this work is on PH observables, we do not explore more standard nuclear collision observables here, such as those which have already been extensively studied and compared with data elsewhere [75,76].

B. Centrality classes and background construction

To study the centrality dependence of PH observables, we divide the ensemble of events into ten deciles. The multiplicity distribution is shown versus centrality class in Fig. 4. We exclude collisions with output multiplicity less than 50 particles, as the PH statistics are rendered highly unstable for very small point clouds. Smaller event multiplicities can be probed by considering a sufficiently large number of events, a task we defer to future work.

Once the simulated nuclear collision data have been generated, we construct a background against which to compare

the PH observables extracted from individual events. The way we do this is similar in spirit to the usual "mixed event" approach employed in experimental analyses (e.g., Ref. [84]): we combine all simulated events in the same centrality class into a single large, uncorrelated event, where each event has been rotated by a different random angle $\delta\phi \in [0, 2\pi]$. We then sample uniformly from this combined event to obtain an event in the same centrality class as the original events. The original events are referred to as "signal" events, while the mixed events are referred to as "background" events. Within each centrality class we generate background events matching the empirical multiplicity distribution of signal events. This procedure thus provides a reference against which to test the significance of our PH observables.

C. PH analysis

Each signal or background event includes a list of discrete particles with momentum-space coordinates. Each event is individually supplied as input to our PH pipeline, so that our analysis is carried out on an event-by-event basis. This yields an ensemble of signal PH observables and another ensemble of background PH observables, where the latter are used to establish a baseline for the former.

For the PH calculations themselves we use the open-source computational package `giotto-ph` [85]. The `treelib` package is used for generating the dendrograms. With these tools we construct and analyze the four different PH observables discussed above: (i) fractal dimension, (ii) Betti curves, (iii) cluster entropy, and (iv) local clustering statistics. Once the observables are constructed for both sets of events, either the ratio or the difference between the signal and background is taken, depending on the specific observable under consideration (as discussed below).

V. RESULTS

In this section we describe our results obtained by applying PH to point clouds generated from Pb+Pb collisions and compare our PH observables in signal events to those generated from background events. We employ the novel statistical summaries outlined in Sec. II and present our results within each centrality class unless otherwise specified.

For several observables we compute the difference between the signal and background events rather than the ratio. Our reasoning is that several topological summaries describe the number or magnitude of homological features as a function of filtration, and therefore the difference in topology is functionally more appropriate than the ratio. This is in contrast to n-point correlation functions, wherein "dividing out" the background is a more natural procedure [84].

For our PH pipeline we employ a sublevel set filtration of the functional $\ell(v) = (\sum_{t\in\Delta(v)} \text{Area}(t))^{1/2}$; here v is a vertex (point) in the Delaunay triangulation, $\Delta(v)$ is the set of triangles adjacent to v, and $\text{Area}(t)$ denotes the area of the triangle t in the (ϕ, y) plane. As discussed in Appendix A, this sublevel set filtration is equivalent to a superlevel set filtration of a density functional; the square root ensures units of (angular) distance. Put more plainly, filtering from small to large values of ℓ is equivalent to (up to a monotonic map) filtering from large to small density.

To assess the effect of nontopological density fluctuations, we also consider some of the observables under a modified filtration $\tilde{\ell} := \ell/\langle\ell\rangle$, where $\langle\ell\rangle$ denotes the mean value of ℓ *within an event*. This filtration modification is performed for all events in a centrality class prior to computing an observable, and we explicitly note both in the text and in figures which filtration we use.

Finally, for the fractal dimension calculations we omit any infinitely long-lived homological features. In zero dimensions this omission corresponds to the largest connected component, while in one dimension we omit the topological loop indicative of the topology of the cylinder.

A. Persistence diagrams

We begin our analysis by examining how the persistence diagrams from the signal collisions differ from those of the background events. For each event type (signal or background) we aggregate the persistence diagrams within each centrality class and compute a count-normalized 2D histogram. We then compute the difference in histograms between the signal events and the background events, the results of which for homological dimension zero and one are depicted in Figs. 5 and 6, respectively. Each subplot for centrality classes 0–10% through 70–80% depicts by color where the persistent homology of the two event types substantially differ: red regions indicate where the signal events have a stronger concentration of persistent topological features, while the blue regions indicate where the signal events have fewer persistent features with respect to the background.

We first note that there is a strong tendency for the signal collisions to have a concentration of persistence features at earlier filtration values, as most easily seen in Figs. 5(c)–5(e). As we travel up the diagonal, we notice that the concentration of signal persistence features gives way to a suppression of persistent features relative to the background (the blue "stripe" running perpendicular to the diagonal). This stripe is less prominent for 50–60% and higher centrality classes. This is consistent with the presence of stronger elliptic flow v_2 in mid-central collisions than in central collisions [86], which produces more points in plane than out of plane and thus leads to a more rapid formation of structure as a function of filtration than an event with vanishing v_2.

Similar to the 0D case, we also compare the 1D persistence diagrams in Fig. 6. We see again the striping behavior (predominance of the signal lifetimes early in the filtration, followed by a suppression relative to the background). The 1D PDs display a different overall shape from the 0D PDs: the 0D PDs show a broader distribution of death values early in the filtration, while the 1D PDs exhibit a distribution of death values that broadens later in the filtration. This implies that loops born later in the filtration persist over a larger range of filtration values. Note as well that, due to our density-based filtration, the 1D PD is effectively measuring the propensity and relative scale of high-density regions surrounding low-density regions. These fluctuations can be interpreted in terms

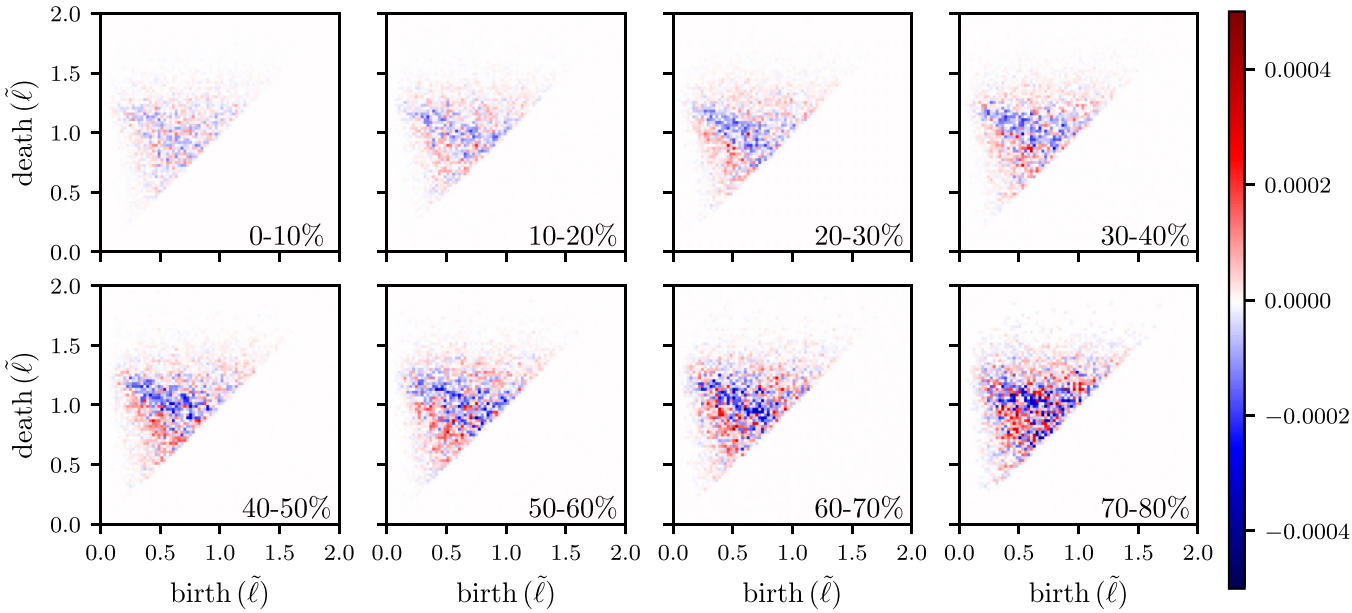

FIG. 5. Differences in 0D PDs for each of the centrality classes. Note that the abscissa and ordinate are in terms of $\tilde{\ell}$.

of local curvature and quantified through fractal dimensions, as we explore next.

B. Fractal dimensions

As noted above in Sec. III, the fractal dimension can serve as a measure of fractality and local curvature. Recall that the fractal dimension effectively measures how the p-norm (here we use α instead of p) of the PD lifetimes scales with the multiplicity of the underlying point cloud. A small α probes local curvature, while large α probes more global structure.

In Fig. 7(a) we depict the fractal dimension of the 0D homology as a function of α, both for the signal and background events. The standard error is estimated from a linear regression of the slope β and then propagated through to the fractal dimension. Figure 7(b) shows the difference in fractal dimensions between signal and background events. We first note that, for small α, the difference in fractal dimension between signal and background is quite substantial, indicating that the persistent homology identifies a higher degree of clustering and fractality in the signal collisions for zero dimensions. Given that the ambient space is effectively a closed cylinder, it is perhaps not surprising that the fractal dimensions are $\sim$2. For larger values of α both the signal and background collisions steadily converge and become statistically hard to distinguish. Given that our analysis is for midrapidity observ-

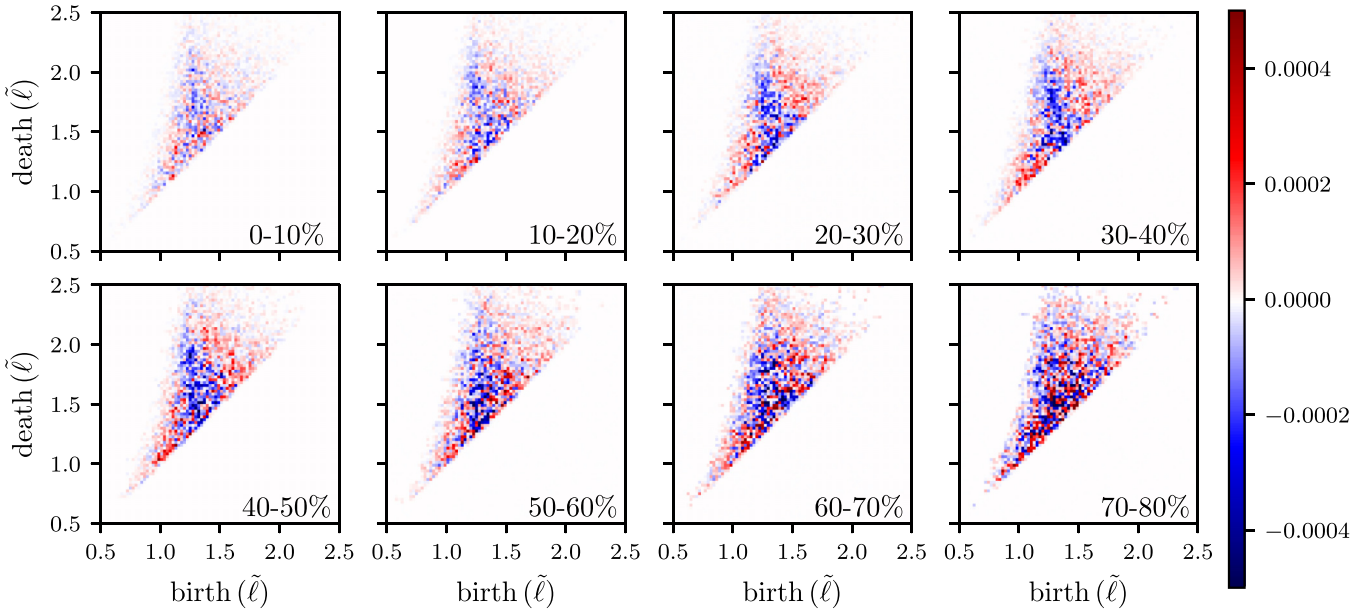

FIG. 6. Differences in 1D PDs for each of the centrality classes. Note that the abscissa and ordinate are in terms of $\tilde{\ell}$.

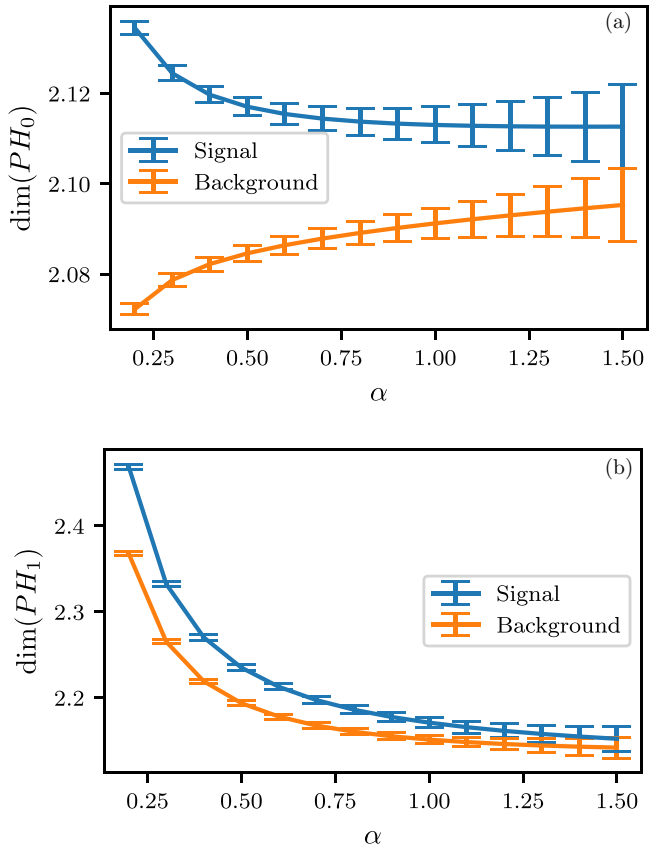

FIG. 7. (a) Fractal dimension dim(PH_0) as defined in Eq. (4) as a function of α. (b) Corresponding fractal dimension for 1D homology.

ables $|y| \leqslant 2$, this lack of distinction follows from the limited volume of our bounded cylinder.

In Figs. 7(c) and 7(d) we show the fractal dimension for the 1D homology and the difference in fractal dimension between the signal and background events. Curiously, for one-dimensional PH both the signal and background events have the same monotonic decrease in dimension, though the fractal dimension differences widen as a function of α. Moreover, in both zero and one dimension the fractal dimension for the signal events is higher than the background. A simple explanation is that the higher degree of clustering in the signal events tends to form shorted-lived loops. As the α-norms for $\alpha < 1$ emphasize small features, this implies the fractal dimension for one dimension is larger for the signal events, though the gap appears to close for sufficiently large α.

C. Betti curves

While the fractal dimension yields important distinctions between the signal and background events, the fractal dimension is insensitive to when in the filtration homological features are most prominent. The Betti curve, described in Sec. III, gives more precise insight into the distribution of homological features as a function of filtration.

1. Mean of $\beta_i(\ell)$

In Figs. 8(a)–8(h) we show the mean and standard error of the Betti curves $\beta_i(\ell)$ and $\beta_i(\tilde{\ell})$ for signal collisions in each centrality class.

To construct the mean and standard error of the signal Betti curves, we compute the average (or standard error) over all events while holding the filtration value fixed. Repeating this process for a large number of filtration values yields Fig. 8. The mean β_i curves both start at zero and end at one. For the β_0 curve, we begin the filtration with no clusters and end with the cluster representing the entire point cloud. For the

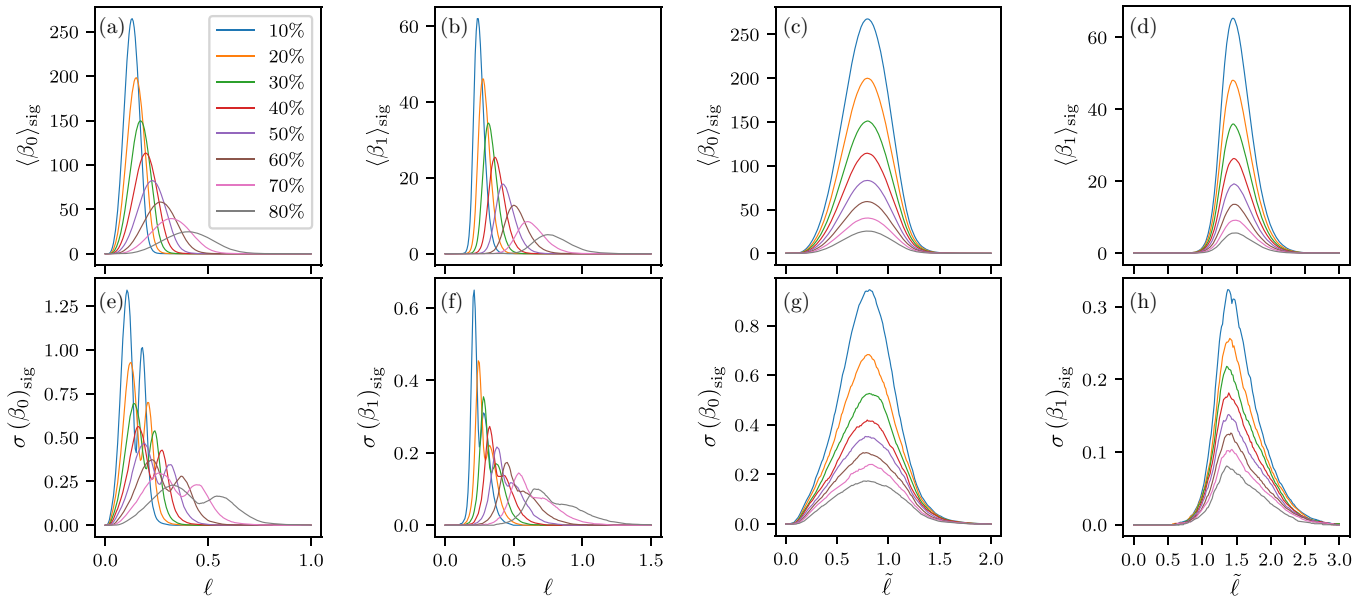

FIG. 8. (a) Mean β_0 calculated by averaging within a centrality class while holding a set of sampling filtration values fixed. The same procedure was performed for β_1 in (b). [(c) and (d)] The same statistics for the $\tilde{\ell}$ filtration. [(e) and (f)] The standard error of β_0 and β_1, respectively, and [(g) and (h)] identical statistics for the $\tilde{\ell}$ filtration.

β_1 curve, we begin with no loops and end with no loops apart from the topological loop represented by the closed cylinder. However, this loop has infinite lifetime and is therefore ignored in the persistence diagram and fractal dimension calculations. Due to the DTFE construction and our filtration definition, clusters can appear at any point during the filtration.

Figure 8(a) shows that each centrality class possesses a single peak, located at a value of ℓ which increases with decreasing multiplicity. This has a straightforward interpretation: events with large multiplicity have a smaller average interparticle spacing than events with small multiplicity, which distribute a smaller number of particles over the same region in momentum space. Consequently, large multiplicity events will establish connections at smaller values of the filtration parameter, and conversely for small multiplicities. Similarly, the merging of separate clusters (and eventual reduction of β_0 to zero) also proceeds more rapidly with ℓ at high multiplicity than low multiplicity. As a result, the peak is reached first by the most central collisions and last by the most peripheral collisions. We note in passing that the Betti curve of each centrality class in Fig. 8(a) is very well described by a Weibull distribution, which has been shown elsewhere to reproduce well the distribution of cluster sizes and nearest-neighbor distributions in random point clouds [87]. Although we do not further explore this issue here, we speculate that measuring *deviations* of β_0 curves from a Weibull distribution may provide a useful way of quantifying nontrivial correlations in point clouds in general. We defer a careful discussion of this possibility to future work.

In Fig. 8(c) we depict $\beta_0(\tilde{\ell})$; note that under this rescaling the peaks of the Betti curve all roughly coincide. This supports the conclusion that the rate of cluster formation and interparticle spacing distributions should depend strongly on multiplicity and therefore the density; thus, rescaling the filtration removes some, but not all, of this effect.

Figure 8(b) shows the mean $\beta_1(\ell)$ curve, wherein we observe similar behavior as the $\beta_0(\ell)$ curve: the maximal number of nonhomologous loops existing at any point in the filtration steadily decreased with multiplicity, while the location of the peak increases in filtration value as a function of centrality class. Rescaling to $\tilde{\ell}$ yields Fig. 8(d); again we see the peaks $\beta_1(\tilde{\ell})$ all roughly coincide across centrality classes.

2. *Standard error of $\beta_i(\ell)$*

In Fig. 8(e) we show the standard deviation of the Betti curve $\beta_0(\ell)$ (i.e., the fluctuations about the mean Betti curve) for each centrality class. The $\sigma(\beta_0)$ in each centrality class exhibits two distinctive peaks which are most prominent in the most central collisions. We note also that the second peak is typically slightly smaller than the first. This can again be straightforwardly understood in terms of event-by-event fluctuations in the scale and location parameters of the underlying Weibull distribution extracted from a single nuclear collision. The slightly smaller second peak is then a consequence of the resulting fluctuations in the shallower slope as $\langle\beta_0\rangle$ descends from its peak value. Both properties of the complete β_0 distribution shown in Fig. 8 are thus consistent with fluctuations of the average density within a single centrality class. As a confirmation of this analysis, the bimodality disappears under the rescaled $\tilde{\ell}$ filtration, as shown in Fig. 8(g).

We show the standard error for the $\beta_1(\ell)$ and $\beta_1(\tilde{\ell})$ curves in Figs. 8(f) and 8(h). Quite similar to the $\beta_0(\ell)$ standard error, we observe a bimodality that disappears under the rescaling $\ell \to \tilde{\ell}$.

3. *$\beta_i(\ell)$ difference*

In Fig. 9 we show the difference between the signal and background events in mean $\beta_i(\tilde{\ell})$: the left panel shows $\beta_0(\tilde{\ell})$ while the right shows $\beta_1(\tilde{\ell})$.

Beginning with β_0, we note for all centrality classes common behavior: a small peak in the signal β_0, followed by a large dip wherein the signal β_0 is lower than the background β_0, followed by a final larger peak in the signal β_0 relative to the background.

The initial uptick in β_0 and large dip at $\tilde{\ell} \approx 1$ is of course consistent with our expectation, already reflected in Figs. 5 and 6, to have more clustering earlier in the filtration, and thus a lower number of clusters and a smaller Betti number than the corresponding background. However, the tendency to have more clustering at the beginning of the filtration appears to drop off for larger centrality classes (see the 50–60% and 60–70% centrality classes). This characteristic rise-and-dip pattern observed in the mid-central collisions implies an initial enhancement of clustering in the signal events as compared with the background which is a direct consequence of enhanced local clustering resulting from collective, anisotropic flow. The final peak in $\beta_0(\tilde{\ell})$ implies that, late in the filtration, there is a resurgence of clusters in the signal events relative to the background. Since late filtration corresponds to low-density regions of the point cloud, the signal events have a preference for stronger fluctuations in density within low-density regions, particularly in low centrality classes.

For the $\beta_1(\tilde{\ell})$ difference we observe a similar rise-and-dip behavior, though for the 0–10% and 10–20% centrality classes the initial peak is stronger than the final peak. Furthermore, the locations of the peaks and dips occur considerably later in the $\tilde{\ell}$ filtration than the corresponding features in the β_0 curves. This is consistent with the onset of loop formation occurring somewhat later than the beginning of cluster formation, relative to the background events.

D. Cluster entropy

As noted above, the Betti curve is indifferent to the relative distribution of points between clusters at each level of the filtration. Since our filtration amounts to a hierarchical clustering scheme, we can access the number of points per cluster and calculate a "cluster entropy," defined in Sec. III. To reiterate, the cluster entropy $H(\tilde{\ell})$ is given by

$$H(\tilde{\ell}) = -\sum_{C_i \in \mathcal{C}(\tilde{\ell})} p_i(\tilde{\ell}) \log p_i(\tilde{\ell}), \tag{7}$$

where $i \in \mathcal{C}(\tilde{\ell})$ is the set of clusters at value $\tilde{\ell}$ and $p_i(\tilde{\ell}) = |C_i|/n(\tilde{\ell})$, $n(\tilde{\ell})$ being the number of points that exist at value

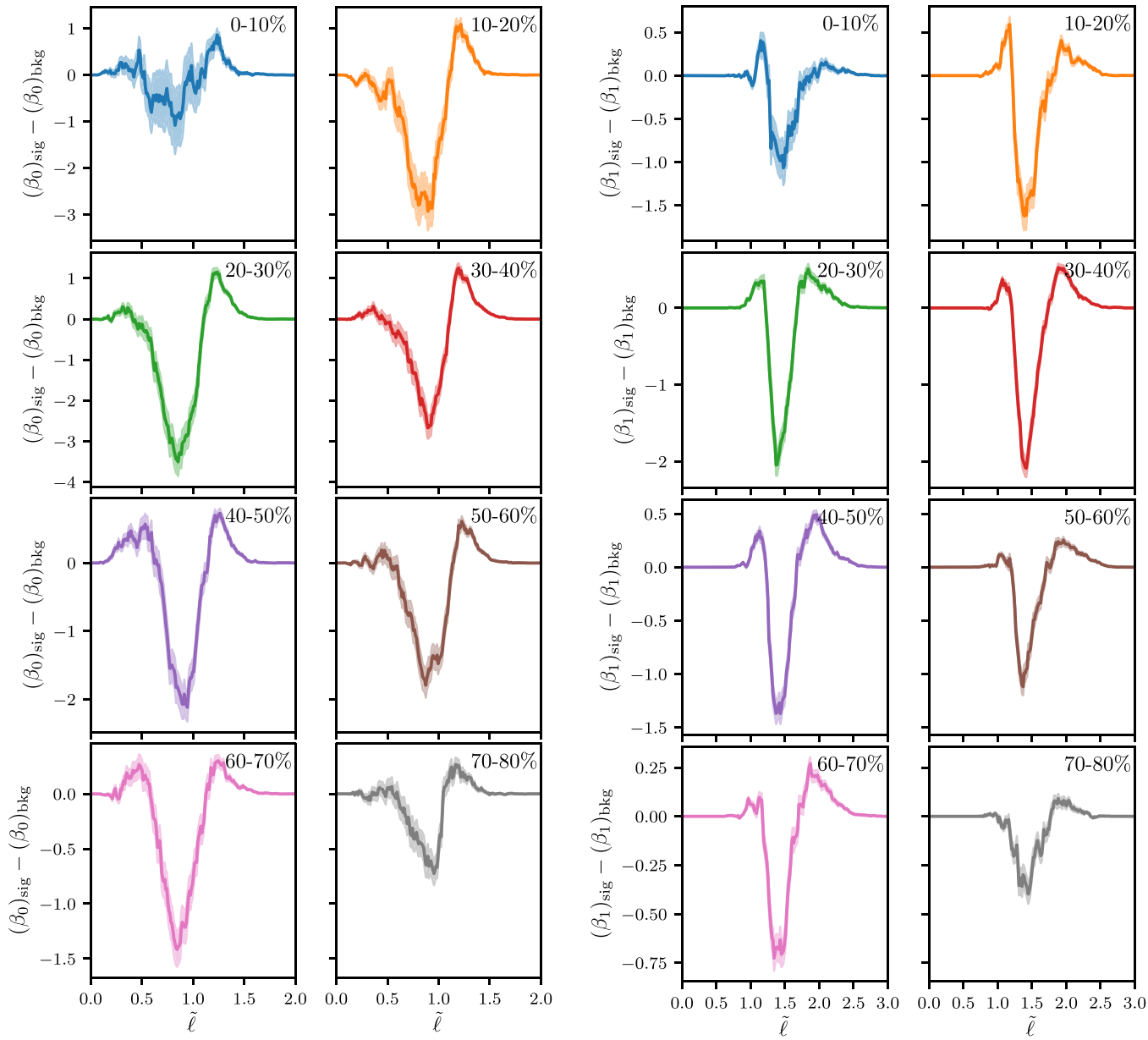

FIG. 9. Left: Difference in $\beta_0(\tilde{\ell})$ between the signal and background for each centrality class. Right: Difference in $\beta_1(\tilde{\ell})$ for each centrality class.

$\tilde{\ell}$ (recall that, by our DTFE, points enter during the filtration and thus do not exist at the beginning of the filtration).

In what follows we also compute the mean number of points per cluster, $p(\tilde{\ell}) = \frac{1}{|\mathcal{C}|}\sum_{C_i\in\mathcal{C}} p_i(\tilde{\ell})$. To ensure fair comparison of events within the same centrality class but different multiplicity, we divide by the multiplicity N and compute $p(\tilde{\ell})/N$.

In Fig. 10(a) we show the mean cluster entropy $H(\tilde{\ell})$ for the signal events as a function of filtration and centrality classes. Note for each centrality class the curve begins at the formation of the first cluster, as the entropy is undefined prior to this point. We observe across all centrality classes a rise in the cluster entropy to a peak that scales in magnitude with the average multiplicity of the centrality class, followed by a steep descent to a vanishing cluster entropy which coincides with the merging of all points into one connected component.

In Fig. 10(b) we show the difference in cluster entropy between the signal and background events. Strong fluctuations in the cluster entropy for small $\tilde{\ell}$ evolve into a consistent suppression in signal cluster entropy around $\tilde{\ell} \approx 1.0$, followed by an increase at $\tilde{\ell} \approx 1.25$ before the signal and background cluster entropies converge to a common value of zero.

In Fig. 10(c) we depict the $p(\tilde{\ell})/N$ for the signal events and plot the difference in $p(\tilde{\ell})/N$ between the signal and background in Fig. 10(d). Note that for all centrality classes save the 70–80% class the difference in $p(\tilde{\ell})/N$ is positive early in the filtration, indicating signal events have more points per cluster. This reflects our expectation that, early

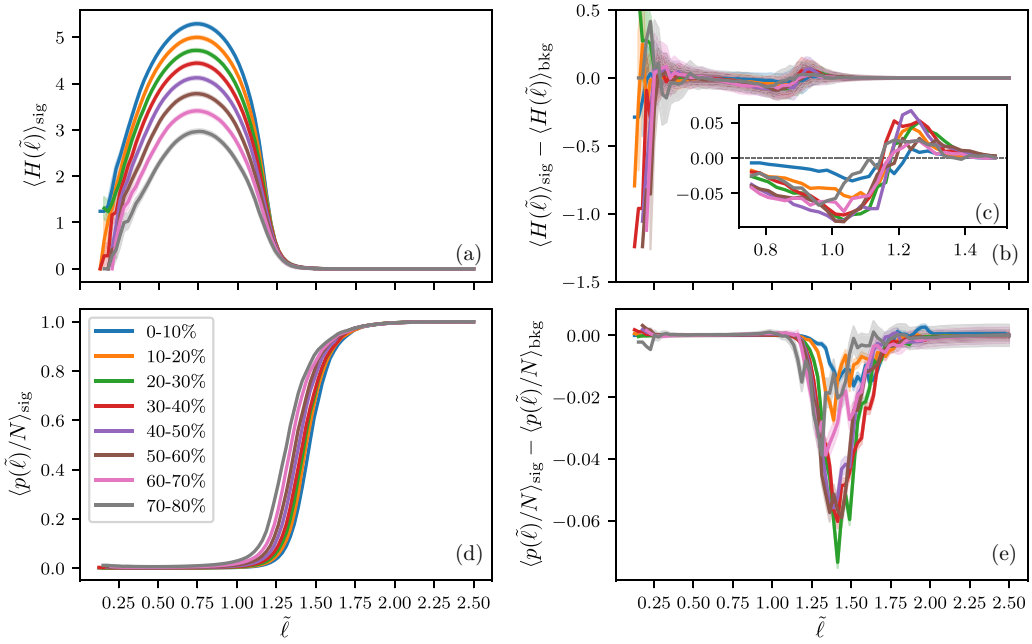

FIG. 10. Cluster statistics as a function of centrality class and filtration. Each plot is computed by averaging over events in a centrality class with a set of sampling filtration values fixed. (a) Average cluster entropy $H(\tilde{\ell})$ for the signal events. (b) Difference in cluster entropy between the signal and background events with enhanced resolution in the inset (c) (the uncertainty bands have been removed from the inset to enhance visibility). (d) Mean number of points per cluster, dividing by the multiplicity. (e) Difference in mean number of points per cluster (dividing by the multiplicity) between the signal and background events.

in the filtration, more local clustering implies more clusters to start relative to the background. At $\tilde{\ell} \approx 1.25$ the mean points per cluster are strongly suppressed for the signal events relative to the background. Note that by this time in the filtration the cluster entropies of the signal and background have largely converged. This sharp reduction in mean points per cluster coincides with the uptick in $\beta_0(\tilde{\ell})$ in Fig. 9 around $1.0 \leqslant \tilde{\ell} \leqslant 1.5$, which indicates the signal point clouds have a late-filtration increase of clusters with small multiplicities. These small clusters drive down $p(\tilde{\ell})$. Due to the small size of these new clusters, the cluster entropy is relatively unchanged, as the largest size cluster dominates.

E. Local clustering statistics

Each of the observables explored so far does not *explicitly* depend upon the relative spatial degrees of freedom in the point cloud, i.e., the interparticle angular correlations. This is unfortunate, as one might expect some dynamical processes to introduce anisotropies, the spatial statistics of which would be of interest. This is particularly true for hydrodynamical flow, the effects of which we have seen above are implicit in a number of the observables discussed previously.

As noted in Sec. III, however, one benefit of the PH pipeline is a dendrogram which reflects local degrees of freedom in the system and captures which clusters merge when in the filtration. Using the local clustering statistic introduced in Sec. III, we calculated, for each 0D PD, the p-norm of the $\boldsymbol{t}_i$ for each hadron. To recall, $\boldsymbol{t}_i$ for hadron indexed i has as components the lengths of branches between merges of clusters containing i (cf. Fig. 3). Our process for computing the local clustering statistics is as follows. First, we computed for each dendrogram the p-norm $||t_i||^p$ as a function of leaf i, which we define as $f_p(\phi_i)$. Second, we determined the mean p-norm within a centrality class and within equal-sized bins (width 0.01 rad) of the azimuthal angle ϕ. Note that we integrate over the rapidity range $|y| \leqslant 2$, by our DTFE construction. We denote the average p-norm as $\langle f_p(\phi)\rangle_X$, where X denotes averaging over either the signal (sig) or background (bkg) events and ϕ now denotes a bin. Third, we took the ratio $g(\phi) := \langle f_p(\phi)\rangle_{sig}/\langle f_p(\phi)\rangle_{bkg}$. Finally, we extracted the v_2 Fourier coefficient of $g(\phi)$, normalized by the average of $g(\phi)$; in particular, we follow standard practice [88] and define the complex quantity

$$V_2 \equiv v_2 e^{2i\Psi_2} = \frac{\int_0^{2\pi} g(\phi) e^{2i\phi} d\phi}{\int_0^{2\pi} g(\phi) d\phi}. \tag{8}$$

We emphasize that, despite using notation similar to that which is normally used in nuclear collision phenomenology, the quantity in Eq. (8) should *not* be confused with the usual measure of elliptic flow, which is computed in a completely different way [70].

We evaluated Eq. (8) for $p = \{0.25, 0.5, 0.75, 1.0, 1.5\}$ and within each centrality class. Figure 11 shows the extracted v_2 coefficient as a function of centrality class and p-norm. The distinctive peak in the flow magnitude v_2 in mid-central collisions is characteristic of the geometry-driven flow anisotropy observed in nucleus-nucleus collisions [89,90]. This demonstrates that anisotropic flow can indeed be accessed and quantified using PH. Note that for illustrative purposes we have ignored complications of our procedure relating to the

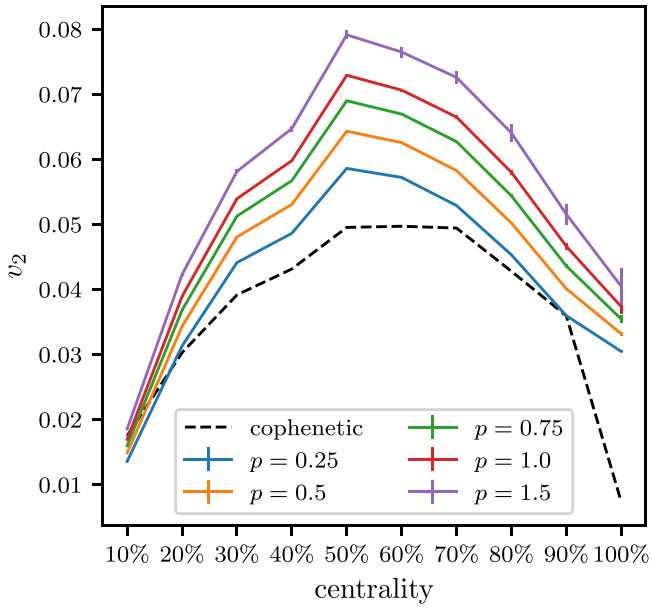

FIG. 11. Extracted v_2 coefficient of the average p-norm local clustering statistic, shown as a function of centrality class and p. The dashed line represents the v_2 coefficient extracted from the cophenetic distance function shown in Fig. 12.

estimation of the elliptic flow plane [91], which we take in this work to coincide with the positive x axis in the transverse plane. In principle, sensitivity of flow observables to the nth flow plane angle Ψ_n can be significant and is typically avoided by working instead with multiparticle correlation functions, which by construction do not depend on Ψ_n [70].

A thorough generalization of the analysis we present here to one which is similarly insensitive to Ψ_n will be deferred to future work. However, we briefly consider one possible way of doing this which involves introducing a PH-based notion of separation between pairs of hadrons. This notion is provided by the cophenetic distance, which we discuss next.

F. Cophenetic distance

As noted above, the dendrogram encodes both the clustering information (through the heights at which branches merge) as well as positional information (by looking at the positional coordinates of the leaves). One natural statistic that couples the physical degrees of freedom to the clustering statistics is the cophenetic distance. The cophenetic distance $d_c(i, j)$ between two points i, j is the height in the dendrogram (value of the filtration) at which the corresponding leaves merge into a single cluster [92].

This distance function can be naturally extended to a correlation functional $d_c(\Delta\phi, \Delta y) := \langle\langle d_c(i, j)\rangle\rangle$ such that points i, j have separation $(\Delta\phi, \Delta y)$, and $\langle\langle\,\rangle\rangle$ denotes averaging over dendrograms within a centrality class. This *cophenetic correlation function* is readily generalizable to nth-order correlation functions (e.g., the cophenetic distance between three points, etc.).

In Fig. 12 we depict the ratio of cophenetic distance functions $d_c^{sig}(\Delta\phi, \Delta y)/d_c^{bkg}(\Delta\phi, \Delta y)$, where the superscripts indicate the type of event. As in the case of the local p-norm clustering statistics, we extracted a corresponding v_2 Fourier coefficient for the correlation distance function. Defining

$$\tilde{g}(\Delta\phi) \equiv \int_{-\Delta y_{\max}}^{\Delta y_{\max}} d\Delta y \, \frac{d_c^{sig}(\Delta\phi, \Delta y)}{d_c^{bkg}(\Delta\phi, \Delta y)}, \tag{9}$$

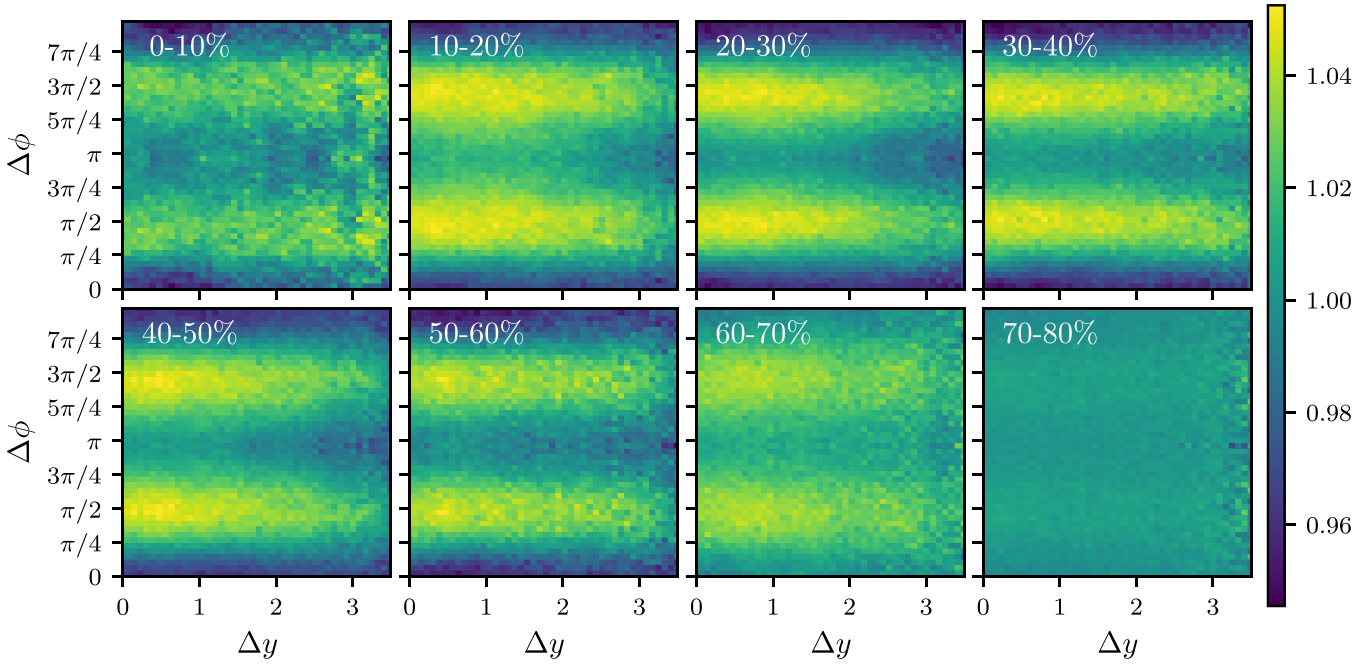

FIG. 12. The ratio of cophenetic correlation functions $d_c^{sig}(\Delta\phi, \Delta y)/d_c^{bkg}(\Delta\phi, \Delta y)$ for different centrality classes. We note the long-range (in rapidity) enhancement which exhibits the oscillation characteristic of elliptic flow.

and setting $\Delta y_{\max} = 3$, we compute the cophenetic v_2 coefficient according to

$$v_2^{\text{cophenetic}} = \left(\frac{\int_0^{2\pi} \tilde{g}(\Delta\phi) e^{2i\Delta\phi} d\Delta\phi}{\int_0^{2\pi} \tilde{g}(\Delta\phi) d\Delta\phi} \right)^{1/2}. \tag{10}$$

The result is plotted in Fig. 11. The cophenetic v_2 shows good agreement with the v_2 coefficient from the local clustering.

A large cophenetic distance between two particles implies a longer time for them to merge into a single cluster, so that sparsely populated regions lead to an enhanced cophenetic distance (relative to the background); conversely, densely populated regions produce suppression of the cophenetic distance in those regions. Thus the elliptic flow signal emerges here (with peaks at $\Delta\phi \approx \pi/2$ and $3\pi/2$ and valleys at $\Delta\phi \approx 0$ and π), qualitatively consistent with that obtained by other two-particle methods [89–91].

VI. CONCLUSIONS

In this work we demonstrated how to use persistent homology to probe correlational and topological structures in the particle distributions produced by nuclear collisions. Our primary aim is to advance the notion of a point cloud as a useful and flexible perspective for characterizing the properties of these particle distributions. We utilized a density-based filtration of ensembles of point clouds to identify large- and small-scale topological structure, and introduced several new tools and observables to probe aspects of nuclear collision phenomenology. Most importantly, we augmented existing tools from topological data analysis to incorporate spatial degrees of freedom, in an effort specifically to quantify anisotropies indicative of hydrodynamical flow. While these topologically minded observables have close, intuitive connections to traditional correlational measures, it remains the subject of future work to formalize the relationships between PH observables and standard statistical measures. In this vein, we briefly note several directions in which our analysis could be improved.

Applications of PH to nuclear collisions have been largely unexplored save for a small collection of works [93]. One fundamental challenge of PH is cleanly relating PH observables to traditional correlational measures like n-point connected correlation functionals. The density-based filtration used here has strong connections to Minkowski functionals, Euler characteristics, and integrals of connected correlation functions [43,44], but a "dictionary" mapping from PH to standard statistical measures has (to our knowledge) yet to be constructed.

Our PH observables are also closely tied to recent work on energy flow polynomials, where correlational structures are tied to functionals of the angular degrees of freedom in the resultant point cloud [94]. Another further direction of pursuit is defining robust observables that better incorporate the positional degrees of freedom, like the local clustering statistic given here. One could imagine statistics like cluster correlation functions being generalized to PH observables, and we intend in future work to explore this further.

Finally, PH is not limited solely to point clouds and could profitably be applied to other aspects of nuclear collisions. For instance, PH has been successfully applied to discretized scalar fields (e.g., on a hyperlattice) through the use of *cubical* complexes [95,96]. In the hydrodynamical phase of the evolution of the QGP one could leverage PH at each phase of the flow to form a sequence of topological fingerprint "snapshots" of the dynamical system. The evolution of these snapshots as a time series can be further explored by leveraging distance measures between PDs [95]. A recent exploration of using PH to probe Rayleigh convection leveraged these tools to identify Lyapunov exponents and critical behavior [95]. A similar pipeline could be used to probe the emergence of turbulent or critical behavior of QGP.

ACKNOWLEDGMENTS

The authors would like to thank Jorge Noronha and Jacquelyn Noronha-Hostler for insightful discussions on potential applications of persistent homology to nuclear collisions. The authors also thank Ante Bilandzic and Mauricio Martinez for providing useful comments on the manuscript. C.P. was supported by the US DOE Grant No. DE-SC0020633. G.H. acknowledges support from the Department of Energy (DOE) Grant No. DESC0020165. T.D. acknowledges support from the ICASU Graduate Fellowship. This work made use of the Illinois Campus Cluster, a computing resource that is operated by the Illinois Campus Cluster Program (ICCP) in conjunction with the National Center for Supercomputing Applications (NCSA) and which is supported by funds from the University of Illinois at Urbana-Champaign.

APPENDIX A: DETAILS OF DTFE

In this Appendix we more fully describe the DTFE, outline the correspondence between superlevel and sublevel set filtrations, and document our procedure for implementing the DTFE in the context of nuclear collisions.

Given the Delaunay triangulation of a point cloud in an ambient space, let v denote a vertex in the triangulation, and define Δ_v as the set of n-simplices adjacent to v. We define a functional

$$f_p(v) = \left(\sum_{t \in \Delta(v)} V(t) \right)^p, \tag{A1}$$

where $V(t)$ denotes the volume of an n-simplex t. The choice $p = -1$ defines the density field $f(v) := f_{-1}(v)$, and we continue the density field onto the rest of the triangulation via a piecewise linear interpolation. The intuition for the density function $f(v)$ is that, in regions with a high concentration of points (with respect to the standard Lebesgue measure), the Delaunay triangulation tends to build many n-simplices with small volume. Vertices in high-concentration areas therefore coface several small-volume n-simplices, which implies $f(v)$ is larger in high-concentration regions than in lower-concentration regions.

While performing a superlevel set filtration on $f(v)$ is perfectly valid, in this work we elect to instead perform a

sublevel set filtration on the "inverse density field"

$$\ell(v) := f_{1/2}(v) = \left(\sum_{t \in \Delta(v)} V(t) \right)^{1/2} . \tag{A2}$$

Apart from being computationally more efficient, performing the sublevel set filtration on $\ell(v)$ is equivalent to a superlevel set filtration on $f(v)$, up to some monotonic map between the filtration values. In other words, both level set filtrations start at large "densities" and end at low "densities."

While the sublevel sets of the linearly interpolated inverse density ℓ are submanifolds, the PH pipeline requires simplicial complexes. Thankfully, the homology of the submanifold is unaltered by instead considering the flag complex of the subgraph of the Delaunay triangulation formed by vertices in the sublevel set [54]. Put differently, the topology can only change when vertices are added to the sublevel set filtration, and so the PH pipeline need only consider when vertices enter and exit the filtration, as the edges only appear when both vertices appear. The flag complex is computationally efficient, since it only requires knowledge of the value $\ell(v)$ and the Delaunay triangulation, which is generally sparse.

The Delaunay triangulation in (ϕ, y) coordinates is essentially a triangulation in two dimensions with one periodic boundary condition, which one can visualize as "unrolling" the infinite cylinder by cutting along the rapidity axis. To properly account for the periodicity of the ϕ coordinate, we computationally leverage a trick of duplicating the point cloud twice over (i.e., if X is the point cloud, generate two more point clouds $X(\phi + 2\pi)$, $X(\phi - 2\pi)$), compute the standard Delaunay triangulation in 2D Euclidean space, and then remove all vertices from $X(\phi + 2\pi)$, $X(\phi - 2\pi)$ that are not part of the original point cloud X (along with all corresponding simplices). The area of each triangle in the triangulation is then calculated as a function of $\Delta\phi$, Δy (where $\Delta\phi$ is the proper angular separation) and is therefore invariant under a Lorentz boost along the beam axis.

While this process is well established in the literature and properly accounts for our periodic degree of freedom, one unfortunate consequence is the presence of edge effects along the boundary of the point cloud in the rapidity direction. One can easily see this edge effect artifact in Fig. 2(b): the upper boundary exhibits "sliver" triangles due to the low density along that direction (these slivers are permitted because the circumcircle extends into a region with no points). The consequence of this edge artifact is that the "sliver" triangles have a low area, and therefore contribute to a larger $f(v)$ for positive p. Thus, the boundary points appear to have a larger $f(v)$ than would be implied by the Lebesgue measure. This edge effect substantially affects the PH pipeline by "turning on" points and simplices at low values of $\ell(v)$ prematurely early.

There are several different ways to combat this issue, each with its own advantages and disadvantages. One approach is to artificially introduce periodic boundary conditions in the y direction as well, such that the high-magnitude rapidity points are less likely to produce "sliver" triangles. However, given the rapidity direction is noncompact, a choice has to be made regarding where to "cut" along the rapidity direction so as to introduce the periodicity. This choice is somewhat arbitrary and might in and of itself introduce spurious edge effects. Another approach is simply to perform the DTFE with one periodic boundary condition, and then introduce a rapidity cut. The advantage of this approach is computational efficiency, at the cost of a choice of where to introduce the rapidity cut. Based upon our numerical experiments of which high-magnitude rapidity points contribute to the edge artifacts, we elected to choose a rapidity cut $|y| < 2$ for our PH pipeline. This rapidity cut functionally implies we compute the DTFE for the full point cloud, but exclude the points that lie beyond the rapidity cut for the PH calculation. While we believe this choice to be appropriate and cogent for our work, we designate future work to assessing other methods of mitigating these edge effects.

APPENDIX B: DIFFERENT PH PROTOCOLS

In this Appendix, we discuss alternative PH protocols, in the interest of inspiring and informing future work.

The PH pipeline performed in this work utilized a sublevel set density filtration which explicitly depended upon the Delaunay triangulation. The triangulation of the manifold was necessary to specify when two points should be joined together. However, given access to the distances between any two points (either the three-momenta Euclidean distance or the cylindrical distance in (ϕ, y)), the Vietoris-Rips (VR) filtration could have been performed. In the VR pipeline two points i, j are connected with an edge whenever $d(i, j) < \varepsilon$, where ε is the filtration parameter and $d(\cdot, \cdot)$ is the distance function. Constructing the flag complex on the resultant graph yields the Vietoris-Rips complex. Note that in this construction the points are assumed to have existed at the beginning of the construction. This has large implications for the Betti number and the cluster entropy, as it implies the zeroth Betti number begins the filtration at its largest value.

One disadvantage of the VR pipeline is the propensity for large chains of points to merge. This effect is well documented in the context of single-linkage clustering and can be combatted through other agglomerative schemes like Wald or centroid clustering. However, more complicated clustering schemes that avoid these chaining effect are less amenable to higher-dimensional simplicial homology (e.g., clusters, not points, are merged in Wald clustering, and so a notion of 1D intracluster homology is difficult to define).

The VR pipeline can be modified to have nonuniform open sets around each point. For example, one can build ellipsoids with principal axes that are point dependent, or radii of balls that scale both with the filtration value and some local functional [97].

Two frequent limitations of PH are computationally large point clouds and the presence of outliers. While large point clouds have less impact on computing 0D homology, higher-dimensional homology becomes computationally intensive. One workaround is to leverage the robustness of PH through the use of a witness complex, wherein a subset of points are used as landmarks used to "witness" simplicial complexes that reflect the underlying topology [49,98]. The witness complex can also be extended through subsampling methods to be robust with respect to outliers [99].

[1] P. Jacobs and X.-N. Wang, Prog. Part. Nucl. Phys. **54**, 443 (2005).
[2] G. Baym, Nucl. Phys. A **956**, 1 (2016).
[3] E. Shuryak, Rev. Mod. Phys. **89**, 035001 (2017).
[4] P. Braun-Munzinger and J. Wambach, Rev. Mod. Phys. **81**, 1031 (2009).
[5] C. Ratti, Rep. Prog. Phys. **81**, 084301 (2018).
[6] J. Noronha-Hostler, P. Parotto, C. Ratti, and J. M. Stafford, Phys. Rev. C **100**, 064910 (2019).
[7] D. Kharzeev, Phys. Lett. B **633**, 260 (2006).
[8] D. E. Kharzeev, L. D. McLerran, and H. J. Warringa, Nucl. Phys. A **803**, 227 (2008).
[9] D. E. Kharzeev and J. Liao, Nat. Rev. Phys. **3**, 55 (2021).
[10] H. Tan, J. Noronha-Hostler, and N. Yunes, Phys. Rev. Lett. **125**, 261104 (2020).
[11] V. Dexheimer, J. Noronha, J. Noronha-Hostler, C. Ratti, and N. Yunes, J. Phys. G **48**, 073001 (2021).
[12] E. R. Most, S. P. Harris, C. Plumberg, M. G. Alford, J. Noronha, J. Noronha-Hostler, F. Pretorius, H. Witek, and N. Yunes, Mon. Not. R. Astron. Soc. **509**, 1096 (2021).
[13] C. Aidala *et al.* (PHENIX Collaboration), Nat. Phys. **15**, 214 (2019).
[14] J. Adamczewski-Musch *et al.* (HADES Collaboration), Nat. Phys. **15**, 1040 (2019).
[15] M. Bluhm *et al.*, Nucl. Phys. A **1003**, 122016 (2020).
[16] J. Adam *et al.* (ALICE Collaboration), Phys. Rev. Lett. **116**, 222302 (2016).
[17] S. Acharya *et al.* (ALICE Collaboration), Phys. Rev. C **101**, 044907 (2020).
[18] S. Chatrchyan *et al.* (CMS Collaboration), J. High Energy Phys. 08 (2011) 141.
[19] L. Adamczyk *et al.* (STAR Collaboration), Phys. Rev. Lett. **112**, 032302 (2014).
[20] G. Aad *et al.* (ATLAS Collaboration), J. High Energy Phys. 11 (2013) 183.
[21] B. B. Abelev *et al.* (ALICE Collaboration), Phys. Rev. C **90**, 054901 (2014).
[22] L. Adamczyk *et al.* (STAR Collaboration), Phys. Rev. Lett. **115**, 222301 (2015).
[23] S. Acharya *et al.* (ALICE Collaboration), Phys. Rev. C **97**, 024906 (2018).
[24] S. Acharya *et al.* (ALICE Collaboration), J. High Energy Phys. 09 (2018) 006.
[25] B. I. Abelev *et al.* (STAR Collaboration), Phys. Rev. C **80**, 064912 (2009).
[26] B. Abelev *et al.* (ALICE Collaboration), J. High Energy Phys. 03 (2014) 013.
[27] S. Chatrchyan *et al.* (CMS Collaboration), Phys. Rev. Lett. **113**, 132301 (2014); **115**, 029903(E) (2015).
[28] B. Abelev *et al.* (ALICE Collaboration), Phys. Rev. Lett. **110**, 152301 (2013).
[29] L. Adamczyk *et al.* (STAR Collaboration), Phys. Lett. B **785**, 551 (2018).
[30] K. Aamodt *et al.* (ALICE Collaboration), Phys. Lett. B **696**, 328 (2011).
[31] A. Adare *et al.* (PHENIX Collaboration), Phys. Rev. C **92**, 034914 (2015).
[32] S. V. Afanasiev *et al.* (NA49 Collaboration), Phys. Rev. C **66**, 054902 (2002).
[33] S. Chatrchyan *et al.* (CMS Collaboration), J. High Energy Phys. 05 (2012) 063.
[34] S. Acharya *et al.* (ALICE Collaboration), J. High Energy Phys. 02 (2019) 150.
[35] S. Acharya *et al.* (ALICE Collaboration), Phys. Lett. B **833**, 137338 (2022).
[36] X. Xu, J. Cisewski-Kehe, S. B. Green, and D. Nagai, Astron. Comput. **27**, 34 (2019).
[37] M. Biagetti, A. Cole, and G. Shiu, J. Cosmol. Astropart. Phys. 04 (2021) 061.
[38] G. Wilding, K. Nevenzeel, R. van de Weygaert, G. Vegter, P. Pranav, B. J. T. Jones, K. Efstathiou, and J. Feldbrugge, Mon. Not. R. Astron. Soc. **507**, 2968 (2021).
[39] D. Spitz, J. Berges, M. Oberthaler, and A. Wienhard, SciPost Phys. **11**, 060 (2021).
[40] I. Donato, M. Gori, M. Pettini, G. Petri, S. De Nigris, R. Franzosi, and F. Vaccarino, Phys. Rev. E **93**, 052138 (2016).
[41] D. Spitz, J. M. Urban, and J. M. Pawlowski, arXiv:2208.03955.
[42] M. T. Angulo, A. Kelley, L. Montejano, C. Song, and S. Saavedra, Nat. Ecol. Evol. **5**, 1091 (2021).
[43] J. Schmalzing and T. Buchert, Astrophys. J. **482**, L1 (1997).
[44] J. Schmalzing, arXiv:astro-ph/9710302.
[45] A. Wiegand, T. Buchert, and M. Ostermann, Mon. Not. R. Astron. Soc. **443**, 241 (2014).
[46] S. Thais, P. Calafiura, G. Chachamis, G. DeZoort, J. Duarte, S. Ganguly, M. Kagan, D. Murnane, M. S. Neubauer, and K. Terao, arXiv:2203.12852.
[47] H. Adams and M. Moy, Front. Artif. Intell. **4**, 54 (2021).
[48] R. Ghrist, Bull. Am. Math. Soc. **45**, 61 (2008).
[49] N. Otter, M. A. Porter, U. Tillmann, P. Grindrod, and H. A. Harrington, EPJ Data Sci. **6**, 17 (2017).
[50] G. Carlsson, in *Handbook of Homotopy Theory* (Chapman and Hall/CRC, London, 2020), pp. 297–329.
[51] S. Fortune, in *Computing in Euclidean Geometry* (World Scientific, Singapore, 1995), pp. 225–265.
[52] R. van de Weygaert, B. J. Jones, E. Platen, and M. A. Aragón-Calvo, in *2009 Sixth International Symposium on Voronoi Diagrams* (IEEE, Piscataway, NJ, 2009), pp. 3–30.
[53] P. Bubenik, M. Hull, D. Patel, and B. Whittle, Inverse Probl. **36**, 025008 (2020).
[54] P. Pranav, H. Edelsbrunner, R. van de Weygaert, G. Vegter, M. Kerber, B. J. T. Jones, and M. Wintraecken, Mon. Not. R. Astron. Soc. **465**, 4281 (2017).
[55] P. Bubenik, in *Topological Data Analysis* (Springer, Berlin, 2020), pp. 97–117.
[56] H. Adams, T. Emerson, M. Kirby, R. Neville, C. Peterson, P. Shipman, S. Chepushtanova, E. Hanson, F. Motta, and L. Ziegelmeier, J. Mach. Learn. Res. **18**, 218 (2017).
[57] N. Atienza, R. Gonzalez-Díaz, and M. Soriano-Trigueros, Pattern Recognit. **107**, 107509 (2020).
[58] J. Jaquette and B. Schweinhart, Commun. Nonlinear Sci. Numer. Simul. **84**, 105163 (2020).
[59] J. Davighi and P. Harris, Eur. Phys. J. C **78**, 334 (2018).
[60] B. Olsthoorn and A. V. Balatsky, arXiv:2110.10214.
[61] N. Hamaus, P. Sutter, and B. D. Wandelt, Proc. Int. Astron. Union **11**, 538 (2014).
[62] C. S. Greenberg, S. Macaluso, N. Monath, A. Dubey, P. Flaherty, M. Zaheer, A. Ahmed, K. Cranmer, and A. McCallum, in *Uncertainty in Artificial Intelligence* (Machine Learning Research Press, 2021), pp. 2061–2071.
[63] M. Cacciari and G. P. Salam, Phys. Lett. B **641**, 57 (2006).
[64] J. M. Amigó, S. G. Balogh, and S. Hernández, Entropy **20**, 813 (2018).

[65] W. A. Zajc, Phys. Rev. D **35**, 3396 (1987).
[66] E. Shuryak and J. M. Torres-Rincon, Phys. Rev. C **100**, 024903 (2019).
[67] E. Shuryak and J. M. Torres-Rincon, Phys. Rev. C **101**, 034914 (2020).
[68] D. DeMartini and E. Shuryak, Phys. Rev. C **104**, 024908 (2021).
[69] M. Cacciari, G. P. Salam, and G. Soyez, J. High Energy Phys. 04 (2008) 063.
[70] U. Heinz and R. Snellings, Annu. Rev. Nucl. Part. Sci. **63**, 123 (2013).
[71] M. Pegoraro, arXiv:2108.13108.
[72] M. Pont, J. Vidal, J. Delon, and J. Tierny, IEEE Trans. Visualization Comput. Graphics **28**, 291 (2021).
[73] J. Curry, H. Hang, W. Mio, T. Needham, and O. B. Okutan, J. Appl. Comput. Topol. **6**, 371 (2022).
[74] C. Shen, Z. Qiu, H. Song, J. Bernhard, S. Bass, and U. Heinz, Comput. Phys. Commun. **199**, 61 (2016).
[75] J. S. Moreland, J. E. Bernhard, and S. A. Bass, Phys. Rev. C **101**, 024911 (2020).
[76] J. E. Bernhard, J. S. Moreland, and S. A. Bass, Nat. Phys. **15**, 1113 (2019).
[77] J. S. Moreland, J. E. Bernhard, and S. A. Bass, Phys. Rev. C **92**, 011901(R) (2015).
[78] W. Broniowski, W. Florkowski, M. Chojnacki, and A. Kisiel, Phys. Rev. C **80**, 034902 (2009).
[79] J. Liu, C. Shen, and U. Heinz, Phys. Rev. C **91**, 064906 (2015); **92**, 049904(E) (2015).
[80] H. Song and U. W. Heinz, Phys. Rev. C **77**, 064901 (2008).
[81] G. S. Denicol, H. Niemi, E. Molnar, and D. H. Rischke, Phys. Rev. D **85**, 114047 (2012); **91**, 039902(E) (2015).
[82] S. A. Bass *et al.*, Prog. Part. Nucl. Phys. **41**, 255 (1998).
[83] M. Bleicher *et al.*, J. Phys. G **25**, 1859 (1999).
[84] S. Chatrchyan *et al.* (CMS Collaboration), Phys. Lett. B **724**, 213 (2013).
[85] J. B. Pérez, S. Hauke, U. Lupo, M. Caorsi, and A. Dassatti, arXiv:2107.05412.
[86] K. Aamodt *et al.* (ALICE Collaboration), Phys. Rev. Lett. **105**, 252302 (2010).
[87] S. Chandrasekhar, Rev. Mod. Phys. **15**, 1 (1943).
[88] M. Luzum and H. Petersen, J. Phys. G **41**, 063102 (2014).
[89] J. Adams *et al.* (STAR Collaboration), Phys. Rev. C **72**, 014904 (2005).
[90] J. Adam *et al.* (ALICE Collaboration), Phys. Rev. Lett. **116**, 132302 (2016).
[91] S. Chatrchyan *et al.* (CMS Collaboration), Phys. Rev. C **87**, 014902 (2013).
[92] R. R. Sokal and F. J. Rohlf, Taxon **11**, 33 (1962).
[93] L. Li, T. Liu, and S.-J. Xu, arXiv:2006.12446.
[94] P. T. Komiske, E. M. Metodiev, and J. Thaler, J. High Energy Phys. 04 (2018) 013.
[95] M. Kramar, R. Levanger, J. Tithof, B. Suri, M. Xu, M. Paul, M. F. Schatz, and K. Mischaikow, Physica D **334**, 82 (2016).
[96] S. Kaji, T. Sudo, and K. Ahara, arXiv:2005.12692.
[97] S. Kalisnik and D. Lesnik, arXiv:2006.09194.
[98] V. De Silva and G. E. Carlsson, in *Symposium on Point Based Graphics* (Eurographics Association, 2004), pp. 157–166.
[99] B. J. Stolz, arXiv:2103.14743.