A Comparative Study of the Perceptual Sensitivity of Topological Visualizations to Feature Variations

Tushar M. Athawale, Bryan Triana, Tanmay Kotha, Dave Pugmire, and Paul Rosen

Abstract—Color maps are a commonly used visualization technique in which data are mapped to optical properties, e.g., color or opacity. Color maps, however, do not explicitly convey structures (e.g., positions and scale of features) within data. Topology-based visualizations reveal and explicitly communicate structures underlying data. Although we have a good understanding of what types of features are captured by topological visualizations, our understanding of people’s perception of those features is not. This paper evaluates the sensitivity of topology-based isocontour, Reeb graph, and persistence diagram visualizations compared to a reference color map visualization for synthetically generated scalar fields on 2-manifold triangular meshes embedded in 3D. In particular, we built and ran a human-subject study that evaluated the perception of data features characterized by Gaussian signals and measured how effectively each visualization technique portrays variations of data features arising from the position and amplitude variation of a mixture of Gaussians. For positional feature variations, the results showed that only the Reeb graph visualization had high sensitivity. For amplitude feature variations, persistence diagrams and color maps demonstrated the highest sensitivity, whereas isocontours showed only weak sensitivity. These results take an important step toward understanding which topology-based tools are best for various data and task scenarios and their effectiveness in conveying topological variations as compared to conventional color mapping.

Index Terms—Perception & cognition, computational topology-based techniques, comparison and similarity.

1 INTRODUCTION

The scale and complexity of scientific datasets have reached a level that makes directly communicating the details of the data through visualization exceedingly difficult [37]. Even if such data can be rendered, the complexity of the output far exceeds what humans can directly interpret [82, 87]. One approach to resolving this issue is to use topological data analysis (TDA) to summarize the data [79, 90]. TDA shows promise in this area because it provides a suite of tools that can summarize n-dimensional scalar field data as features or hierarchies in ways that are intuitive to humans [68, 89]. TDA-based visualizations have been fundamental to understanding feature variations in multiple real-world applications, including fluid dynamics [31, 60] and combustion [8]. Although what data features TDA can summarize is quite well understood (e.g., the relationship of critical points of a scalar function), what TDA-based visualizations communicate to a user is not. In this paper, we describe our study of the sensitivity of TDA-based visualization methods, namely, isocontours, Reeb graphs, and persistence diagrams, to a reference color mapping visualization in communicating feature variations within a scalar field.

In practice, color mapping (see Fig. 2a) is often used to directly render the scalar field, but such an approach leaves the function topology to be inferred by the viewer. On the other hand, TDA-based visualizations show abstractions of the scalar field topology. The isocontour visualization (see Fig. 2b) acts as a topographic map of the function by using a series of level sets that directly show changes in the topology of the function. Reeb graphs (see Fig. 2c) act as a skeletal summary of the isocontour visualization by tracking the evolution, i.e., creation, destruction, merging, and splitting, of the contour level sets and specifically

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a non-exclusive, paid up, irrevocable, world-wide license to publish or reproduce the published form of the manuscript, or allow others to do so, for U.S. Government purposes. The DOE will provide public access to these results in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).
We briefly discuss related research, including the evaluation of visualizations, topology-based data visualization, and sensitivity analysis.

2.1 Evaluation of Visualization

Although there are many approaches to visualizing complex scientific data, measuring how good a particular visualization is has been a nontrivial challenge for researchers [42]. Visualization quality is determined by the ability of users to effectively and efficiently extract knowledge from complex, large-scale data. Various visualization components (e.g., hardware requirements, software implementation costs, interactivity, accuracy of visualizations, and perception and cognition) have been evaluated to understand the overall impact of visualizations on decision-making [33, 58, 69, 86]. One promising approach to evaluating the effectiveness of visualizations in decision-making is conducting human-subject studies. Human-subject studies enable researchers to perform quantitative assessment of visualization parameters (e.g., errors and time associated with decisions, comparison with alternative visualizations, and the sensitivity of decisions to a person’s domain and visualization expertise), which can potentially help reinforce the scientific foundation of visualization [43, 62].

Multiple studies have assessed the effectiveness of various scientific data visualizations. Color mapping is one of the fundamental methods used in data visualization because scalar data can be encoded into intuitive visual attributes, such as hue and opacity, to convey data features. The choice of a color map, however, strongly influences human perception [55, 92]. Liu and Heer [50] quantified the perceptual performance of users for different color maps by presenting tasks that required color comparisons. A similar study was performed to understand the effectiveness of 2D [45] and 3D [34] vector field visualizations by evaluating the quality of decisions in identifying flow features (e.g., critical points and direction of particle advection). An interesting approach to measuring cognitive load experienced when perceiving information from visualizations is through the analysis of EEG patterns of subjects [1].

2.2 Topology-Based Data Visualization

As a key tool in scientific visualization, topological data analysis (TDA) [38, 79, 90] enables the understanding of abstract structures present in data. Visualization applications in numerous scientific domains (molecular dynamics, combustion science [9], fluid dynamics [47], and radio astronomy [72]) have demonstrated the strengths of TDA in unraveling complexities of the underlying data. A few of the fundamental topological descriptors for univariate data include level sets [51], critical points [57], Reeb graphs [26], contour trees [14], and persistence diagrams [20] (see Sect. 3 for technical details on these descriptors). Other topological descriptors include Morse-Smale complexes [28], which are widely used topological abstractions that segment the domain into cells with uniform gradient behavior. Reeb spaces [27], Jacobi sets [80], and fiber surfaces/feature level sets [13, 41] extend the concepts of Reeb graphs, critical points, and level sets, respectively, to multivariate data. Furthermore, there have been significant research developments in leveraging topological features for effective feature tracking of scientific data [7, 10, 19, 30, 66, 75, 83].

Previous works have also investigated novel metrics to quantitatively compare scalar fields via topological descriptors. For example, the distance between persistence diagrams can be measured as the cost of matching features between diagrams, e.g., using bottleneck [21] or Wasserstein [26] distance. Further, kernel-based methods [16, 70] have been developed to quantify the distance between persistence diagrams that are more suitable for machine learning tasks. An array of metrics have been devised, including interleaving distance [56], edit distance [76], stable distance [4], and Wasserstein distance [65], that quantitatively measure the distance between merge trees, the building blocks of contour trees and Reeb graphs. Bauer et al. [2] proposed functional distortion metric, and Lan et al. [46] leveraged the idea of intrinsic interleaving distances between merge trees [35] to compare Reeb graphs. The recent survey paper by Yan et al. [90] provides an overview of the state of the art in quantitative measures for comparing topological descriptors of scalar fields.

2.3 Sensitivity Analysis

Since we evaluate the sensitivity of visualizations to amplitude and positional variations of scalar fields, we discuss how sensitivity analysis has been shown to help improve data representations and scientific analyses. Saltelli et al. [73] discussed in detail a broad spectrum of statistical techniques for understanding the sensitivity of functions to
The Reeb graph is a structural abstraction that provides insight into the topology of a function. The Reeb graph is determined by the critical points of the scalar field. The number of connected components of a level set for each function value depends on the critical points of the function. Both $f_a$ and $f_e$ have a level set with one connected component, whereas $f_b$ has two. The center point for each connected component, $C_i$, is labeled as $G_i$. By tracking the points, $G_i$, from $-\infty \rightarrow +\infty$, the Reeb graph structure emerges, which summarizes the topology succinctly. Center Panel: The four types of critical points, including (a) a local minimum, which has one outgoing edge; (e) a local maximum, which has one incoming edge; (f) the merge saddle, which has two incoming edges that merge into a single outgoing edge; and (g) the split saddle, which has one edge split into two. Right Panel: For two paired critical points, $A$ and $B$, the persistence between them is $d = |f(B) - f(A)|$, where $f$ is the function applied to the manifold. (i) To reduce visual clutter, the Reeb graph can be pruned by removing any birth-death pair with $d < \varepsilon$.

### 3.3 Persistence Diagrams

The structure of the Reeb graph can be summarized with a multiset of points, known as a persistence diagram. Persistence diagrams highlight the more prominent features in a Reeb graph and are a stable representation of the function.

Points of the persistence diagram are formed by pairing critical points in one of three configurations, saddle-minimum, saddle-maximum, and saddle-saddle pairs. Without a loss of generality, we describe saddle-minimum pairing, which depends on the sublevel sets of the function and corresponds to the Elder’s rule [26]. As $f_e$ increases continuously from $-\infty \rightarrow +\infty$, let $B$ be a saddle point of the function with function value $f(B)$.

Persistence is used to classify the importance of a feature and can be used to differentiate topological signal from noise. One method of denoising a Reeb graph is to prune noise via persistence simplification [29], as illustrated in Fig. 3i. For a given feature to be considered noise, its persistence, $d$, must be less than a certain $\varepsilon$ threshold. The features are removed from the output graph by deleting associated nodes and reconnecting the graph [24]. Selecting an optimal threshold is difficult to automate because being too aggressive can lead to the removal of important features, and being too relaxed may result in a noisy dataset. Therefore, users of persistence often rely on manual selection and tuning of $\varepsilon$ to gain insights from Reeb graphs.

### 4 Visualizations Evaluated

We compare a reference visualization type—color maps—to three TDA-based visualizations—isocountours, Reeb graphs, and persistence diagrams—that reveal the topology of a function applied to a manifold.

#### 4.1 Reference Visualization: Color Maps

Color maps are a commonly used visualization and are seen as an intuitive way to interpret data (see Fig. 4a) [55, 92]. Given a function, $f$, applied to the surface of a manifold, $\mathcal{M}$, with a global minimum of $f_{\text{min}}$ and a global maximum of $f_{\text{max}}$, a color scale, $K$, is a set of colors mapped between $K_{\text{min}}$ and $K_{\text{max}}$ respectively. Practically, color maps are implemented with a discrete set of colors, and in general, a given input, $f_i$, is linearly interpolated to convey the continuity of the function. However, multi-hue color maps, e.g., viridis, are formed by tracing curves through perceptually-uniform color models, e.g., CIEXYZ [48].
We model prominent data features as a mixture of Gaussian distributions; while perception is biased (e.g., pure green may appear brighter than pure blue) [50, 77]. In a recent study, Liu et al. [50] compared various color maps and concluded that the *viridis* map was the most effective at presenting data in a way that enabled users to ascertain features correctly. Cooper et al. [22] presented a list of color maps that maintain better perceptual ordering than the rainbow color map, among which the viridis color map is a reasonable choice for a reference implementation.

**Design:** One challenge for color map visualizations is selecting the best color map to use. For example, the size of physical marks used varies what the user perceives [77]. Further, perceived color differences are not necessarily uniform—colors may be mathematically equidistant, while perception is biased (e.g., pure green may appear brighter than pure blue) [50, 77]. In a recent study, Liu et al. [50] compared various color maps and concluded that the *viridis* map was the most effective at presenting data in a way that enabled users to ascertain features correctly. Cooper et al. [22] presented a list of color maps that maintain better perceptual ordering than the rainbow color map, among which the viridis color map is a reasonable choice for a reference implementation.

**Interpretation:** When observing a color map visualization, the user should look for cold spots (purple hue) and hot spots (yellow hue) to understand the positions of local minima and maxima (critical points) of the scalar field. Fig. 4a shows cold and hot spots on the rabbit model.

**Interaction:** Interactive rotation and zooming help overcome occlusion and enable investigating the data in finer detail.

### 4.2 Topology-Based Visualization: Isocontours

The visualization of level sets for different isovalues (see Sect. 3.1), called isocontouring, provides insight into the structural evolution of a function as well as the distribution of isolvels across the manifold on which the data are sampled (see Fig. 4b).

**Design:** Let $f_{\min}$ and $f_{\max}$ denote the global minimum and global maximum values of a function sampled on a manifold. The interval $[f_{\min}, f_{\max}]$ is partitioned into $L$ equally spaced isolvels. Then, a series of level sets for the isolvels is extracted using the marching triangles algorithm [39], and they are overlaid on top of the model.

**Interpretation:** When observing function topology with isocontours, the most important features are the visual patterns that arise as the contours wrap around the model. Formation or merging of contours will occur around the critical points, so identifying those features may create a better picture of the function's topology. The topological evolution of isocontours is illustrated with red-dotted boxes in Fig. 4b.

**Interaction:** We define the level of detail for isocontours as the number of isolvels, i.e., $L$, where $L$ is a user-selectable parameter. Larger values of $L$ result in higher level of detail, and smaller values of $L$ result in lower level of detail (see the skull dataset example in Fig. 4b). The parameter $L$ can be modified to change the level of detail, thereby enhancing the perception of the structure of the function. The view can also be rotated and zoomed to overcome occlusion and enhance details.

### 4.3 Topology-Based Visualization: Reeb Graphs

Reeb graphs (see Fig. 4c) are used to visualize the topological structure as a skeleton (see Sect. 3.2).

**Design:** Our visualization embeds the Reeb graph onto the model, similar for how it is handled in the Topology Toolkit [54, 81]. The nodes of the Reeb graph, which represent the critical points of the function, are overlaid on the model. The edges represent the flow of the function (contours) between critical points. Therefore, a naive edge connection scheme would not suffice because it would not represent the flow of data correctly. We utilized the centroids of isocontours to draw the trajectory of the arcs throughout the model.

**Interpretation:** Since the Reeb graph is embedded in the model, its features can be directly correlated with the function on the model. The primary features to look for in a Reeb graph are the positions of critical points, which are represented as blue dots on the model, and the arc trajectories between blue dots, which are represented as red curves.

**Interaction:** We define the level of detail in the context of the Reeb graph as the number of arcs of the graph. For high-density noise, the Reeb graph visualization might look cluttered owing to the high level of detail arising from noise, as shown in the bottom right image of the skull in Fig. 4c. Visual clutter can be interactively mitigated by adjusting the level of detail using persistence-guided pruning [29], which removes low persistence critical point pairs from the visualization, as shown in the bottom left image of the skull in Fig. 4c. Notably, persistence-based noise reduction may remove longer edges before short ones, which may be counter-intuitive. Occlusion and clutter can be further mitigated by rotating or zooming the view.

### 4.4 Topology-Based Visualization: Persistence Diagrams

Persistence diagrams (see Fig. 4d) are scatterplots that can be used to illustrate the birth-death pairs of topological data for contour trees, Reeb graphs, and other TDA-based techniques. A birth-death pair corresponds to the pair of minimum/maximum and saddle critical points that result in the creation and disappearance of a connected component, respectively, as described in Sect. 3.3. Persistence diagrams provide a stable view of the function from a topological perspective, as they show the relationship of the critical points in the Reeb graph.

**Design:** To generate the persistence diagrams, Reeb graphs are generated for the dataset, and the birth-death pairs are then identified, as described in Sect. 3.3. The pairs are then visualized in the persistence diagram. The horizontal axis of a persistence diagram is labeled as *birth* and denotes the appearance of a feature. The vertical axis is labeled as *death* and denotes the disappearance of a feature. The blue dot represents local saddle-minimum or saddle-maximum pairs. The red dots represent saddle-saddle pairs, which are formed by holes/tunnels in the model. The diagonal line indicates where *birth = death*, and the
vertical distance from the diagonal is a quantified measure of persistence. Points near the diagonal have a low persistence (i.e., regarded as noise), whereas those farther away are more important. High- and low-persistence features for the rabbit are highlighted in Fig. 4d.

Interpretation: When evaluating data with the persistent diagram, one must note the points far away from the diagonal and their color/type, as these points denote the more persistent features. Secondarily, one may consider the distribution and patterns of points along the diagonal (e.g., feature clustering for the rabbit in Fig. 4d).

Interaction: We define the level of detail in the context of persistence diagrams as the number of points representing critical point pairs. We provide the functionality to interactively set a persistence threshold, which removes low-persistence points for observing different levels of detail, which is depicted in the bottom of Fig. 4d.

5 Sensitivity Analysis of Topological Visualizations

We describe the method used for measuring the sensitivity of visualizations to variations in data using a crowdsourced experiment.

5.1 Data Model

For our study, we model prominent data features as a mixture of 3D Gaussian distributions mapped onto a 2-manifold. Specifically, a given dataset has two components. The first is a 2-manifold triangle mesh, $\mathcal{M}$, embedded in 3D space. The second component is a scalar field, $f$, which is a mixture of Gaussian distributions and noise defined on the vertices of the mesh. Gaussian mixtures are commonly utilized in scientific literature to model data (e.g., in Vidal et al. [88] and Yan et al. [91]), and scalar fields on 2-manifolds are observed in many real-world applications (see Fig. 1).

5.1.1 Scalar Field Generation

Scalar values were generated per vertex using a combination of several data features and noise. The value at each vertex was calculated as

$$f(x, y, z) = \mathcal{N}(x, y, z) + \sum_{i=1}^{NOF} S_i (y, z),$$

where $\mathcal{N}$ was the noise function, NOF (number of features) was the number of salient data features, and $S_i$ was the data field feature.

Data Features: To simplify the definition of data features, isotropic 3D Gaussian functions were used:

$$S_i (y, z) = a_i \cdot \exp\left(\frac{-(x-x_i)^2 + (y-y_i)^2 + (z-z_i)^2}{2\sigma^2}\right),$$

where $a_i$ was the amplitude of the feature, which was 1 by default; $(x_i, y_i, z_i)$ was the feature position, which was set using a random vertex on the model; and $\sigma$ was the standard deviation, which was fixed to 1.

Noise: Noise was added using the following function:

$$\mathcal{N}(x, y, z) = SN \cdot \text{Perlin}(x, y, z),$$

where $SN$ is the amplitude of the noise, and $\text{Perlin}$ is the noise function. The level of noise was specified using an input signal-to-noise ratio (SNR). Because the default feature amplitude was 1, $SN = 1/\text{SNR}$. $\text{Perlin}$ was a standard Perlin noise function [64], the frequency of which was set during our data calibration (see Sect. 6.5).

5.2 Sensitivity for 1D Gaussian Distributions

Without a loss of generality, we describe visualization sensitivity in the context of 1D Gaussian distributions. At the end of the section, we discuss how our observations extend to the 3D Gaussian distributions described in the previous section.

5.2.1 Position and Amplitude Variation

We are interested in evaluating the sensitivity of the topological visualizations described in Sect. 4.2-Sect. 4.4 as compared to the reference color mapping (Sect. 4.1) to changes in the input function. In particular, we evaluate the sensitivity of these visualization techniques to two variation types. First, we are interested in changes in the position of topologically important features of the data. For example, in Fig. 5a, the feature of interest is moved from A to B. Second, we are interested in sensitivity to changes in the scale or amplitude of topologically important features. As shown in Fig. 5b, the maximum value of the peak decreases from A to B.

5.2.2 Measuring Sensitivity

To test the sensitivity of the visualizations, we consider a scenario in which a participant must pick from two experimental visualizations which is the most similar to a baseline visualization. Let DB denote the baseline dataset corresponding to the original position and amplitude parameters B. Let $DA_0$ and $DA_1$ denote the two datasets corresponding to parameters $A_0$ and $A_1$, respectively, representing either variations of amplitude or position for a single feature. Let $VA_0$, $VA_1$, and VB be visualizations generated using the same visualization technique, either color maps, isocontours, Reeb graphs, or persistence diagrams. We present a participant with these visualizations, and the participant has to decide which of $VA_0$ and $VA_1$ is closer to VB.

Consider the 1D examples in Fig. 6. For positional variation (see Fig. 6a), given a baseline visualization for the position parameter B, the participant must select between the visualizations for parameters $A_0$ or $A_1$, which they believe are more similar to the baseline visualization. Similarly, for the amplitude variation (see Fig. 6b), the participant must select between the visualizations for parameters $A_0$ or $A_1$, which is more similar to the baseline visualization. For amplitude variation, either $A_0$ or $A_1$ is always larger than B, and the other is smaller than B to ensure the participants are not directly comparing $A_0$ and $A_1$.

To evaluate the sensitivity of a visualization to variations in features, we measure how often participants can correctly select between the visualizations generated using the test being performed. For positional variations, $d(A, B)$ is the geodesic distance between the location of the feature being moved in A and B. For amplitude variations, the absolute difference between the amplitude of the feature being modified is used (i.e., $d(A, B) = |A - B|$).

Intuitively, $d$ measures how different in distance or amplitude the variations are from the baseline. A larger $d$ value implies that one stimulus is much more similar to the baseline and would therefore be more likely to be selected by a participant. Consider the example shown in Fig. 6. For both positional and amplitude variations, it is not easy to decide which visualization is perceptually more similar to the baseline visualization when $d(A_0, B) \approx d(A_1, B)$ (see Fig. 6(top)). In contrast, when $d(A_1, B)$ is significantly smaller than $d(A_0, B)$ (see Fig. 6(bottom)), it is relatively easy to visually decide that the 1D plot for parameter $A_1$ is closer to the baseline visualization.

Generally speaking, the sensitivity of a quantity of interest $Y$ with respect to a parameter $X$ is defined as the first derivative $\partial Y/\partial X$ and can be estimated through linear regression for the observed data [73]. In perceptual psychology, however, the Weber-Fechner Law [32] notes that the minimal detectable increase in the stimulus is proportional to the stimulus magnitude (Weber’s law), and the perceived intensity is proportional to logarithm of the actual intensity (Fechner’s inference). Thus, the sensitivity of visualizations is captured by the change in the accuracy of user decisions with respect to the change in parameter $A'$ using logistic regression (see the analysis description in Sect. 6.7).

5.3 Generating Positional and Amplitude Variations in 3D

To test the sensitivity, three scalar functions are generated for each parameter $B$, which removes low-persistence points for observing different levels of detail, which is depicted in the bottom of Fig. 4d.
6.1 Hypotheses

To understand the sensitivity of the four visualizations, we consider the topological aspects of the function that each visualization shows. Fig. 2 illustrates the sensitivity of each visualization to positional and amplitude variations, assisting in the formulation of our hypotheses.

**Color Maps:** For color mapping, we have chosen the viridis color map (see Sect. 4.1 for more details). Thus, the brightest spot in the color map should also indicate the location of the peak of the feature. Therefore, we hypothesize that color maps will be sensitive to variations in both the amplitude and the position of features.

**Isocontours:** Similar to color maps, the concentric rings of isocontours show where function extrema occur. However, the rings themselves provide no direct indication of the function value. Therefore, we hypothesize that isocontours will be sensitive to variations in the position of features but not sensitive to the amplitude of features.

**Reeb Graphs:** Reeb graphs show the interconnectedness of critical points in the function. However, they do not provide any indication of the value of those critical points. Therefore, we hypothesize that Reeb graphs will be sensitive to variations in the position of features but not sensitive to the amplitude of features.

**Persistence Diagrams:** Persistence diagrams show only the birth and death of critical features, which correspond to the amplitude of a birth-death pair. Positional variations are visible indirectly only when pairs switch. Therefore, we hypothesize that persistence diagrams will be sensitive to variations in the amplitude of features but not sensitive to their position.

6.2 Experimental Interface

The experiment consisted of a web page with four demographic questions, a tutorial describing all of the visualization types and interactive capabilities (e.g., rotation and zooming), four practice questions, 24 experimental trials, and a post-experiment questionnaire. The tutorial and practice questions serve to familiarize participants, especially those with no expertise or limited prior experience in visualization, with the four visualization types. Specifically, we present guidelines for participants summarizing what specific patterns to look for when decoding the visualizations (similar to the interpretation descriptions in Sect. 4). Please refer to pages 4-10 of the supplementary material illustrating the tutorial and practice questions for a single user. The responses to our post-experiment questionnaire (see pp. 39-47 of the supplementary material and Sect. 7.4) capture the thought process of participants and the difficulties they encountered when interpreting the four visualization types and making decisions.

The main experimental interface (see Fig. 7) consists of the three visualizations, always of the same type, including a baseline in the center and experimental dataset choices to the left and right (see Fig. 7(a)). An orbiting camera with zoom was used so participants could view the models from every angle, which enabled the participants to observe features that would otherwise be hidden and to investigate details. The navigation was coordinated to ensure that changes to orientation or zoom were applied to all three visualizations. Visualizations that allow level-of-detail adjustments had a slider for this purpose (see Fig. 7(b)). We asked the participants which of the two experimental visualization choices was more similar to the baseline (see Fig. 7(c)). The participants had to make one of three choices: (1) left, (2) right, or (3) unsure if they were unable to answer with any confidence. The correct answer (i.e., the more similar visualization) was randomly shuffled between the left and right sides. For each trial, the participant was allowed up to 60 seconds, after which the experiment automatically advanced.

6.3 Model Selection

Models were selected to balance the desire for realistically complex geometry while also limiting cognitive strain caused by interpreting the model. We, therefore, selected familiar 3D shapes, namely, biological models from humans and animals (see examples in the supplementary material). To avoid categorical bias, we selected four models from each of the categories: human busts, human anatomical models, human extremities, land animals, sea animals, and birds. The goldenRetriever, lion, rabbit, horse, skull, turtle, shark, fish, owl, parrot, and bird come from the Princeton Shape Benchmark repository [74]. The bimba, bust, windfish, handFist, and handPointPrep are from the AIM@SHAPE repository [23]. The heart and kidney are from the University of Michigan’s BlueLink Anatomy TOOL [85]. The duck, frederick, lincoln, foot1, and foot2 were acquired from Free3D [67]. The models were prepared and cleaned using Blender 3D’s [3] mesh retopology and malformed mesh detection tools. All models were then normalized to fit in a unit box at the origin.

6.4 Variables

The independent variables of our experiment include the 3D model; visualization types (color maps, isocontours, Reeb graphs, or persistence diagrams); variation type (position or amplitude); values of A₀, A₁, A′; SNR; and the NOF. However, we specifically focus on evaluating visualization types, variation types, and A′ values. The dependent variables are the accuracy and sensitivity of selection, time taken, mouse interactions (rotate and zoom), and level-of-detail interactions.
6.5 Parameter Calibration
Determining parameter ranges was an important problem because the data complexity had to be tuned to ensure that the task was neither too difficult nor too easy for all visualization types. In particular, we were focused on identifying reasonable values of A’, SNR, and NOF. Our parameters went through a four-stage calibration process. (1) Initially, parameters were set using the research team’s observations of values. (2) Next, we ran an internal study using the research team and lab members to check the difficulty of tasks and adjust accordingly. (3) Then, we ran a small-scale (40-person) preliminary study on Amazon Mechanical Turk. (4) A final internal study was repeated using the research team and lab members. At each stage, parameters were adjusted to calibrate difficulty and experiment length and to improve the experimental interface.

6.6 Data Generation
A Python script was used to generate the model and function configuration for all participants in the experiment. The within-subject design consisted of 24 trials. Each participant was presented with trials that had the following characteristics:

- 1 trial per 3D model
- 6 trial per visualizations type: {color map, isocontour, Reeb graph, persistence diagram}
- 12 trial per variation type: {position, scale}
- 6 trial per A’: {0.15, 0.30, 0.45, 0.60}
- 8 trial per SNR value: {80, 90, 100}
- 3 trial per NOF in the range: [2, 9]

Thus, we ensured that each user was presented with the datasets generated using a balanced distribution of parameters, including visualization type, function type, A’, SNR, and NOF. Additionally, A0 was randomly selected such that A0 ∈ [0, 1], and AI was selected such that AI ∈ [0.1, 0.9] while also satisfying the A’ requirement. For amplitude variation, AI is added to B, and AI is subtracted from B, or vice versa, in a random manner to guarantee that one stimulus would have a larger amplitude, and one stimulus would have a smaller amplitude relative to the baseline.

6.7 Data Collection and Analysis Methodology
Datasets were pre-generated using a Python script and the parameters described Sect. 6.6. The model, function, Reeb graph, and persistence data were stored in JSON format and ZIP compressed to improve the file transfer rate. The experiment was run on a custom-built Node.js web server. Three.js [78] was used for color map, isocontour, and Reeb graph rendering. D3.js [5] was used for persistent diagram rendering. The visualizations were custom implementations. The answers from each participant were recorded in a JSON document with entries containing participant ID, visualization type, parameters for the stimuli, participant’s selection, interaction information (i.e., click, scroll, slider move counts), and time taken.

We utilized three statistical tools in the analysis. First, a binomial test, which determines if an observed distribution deviated from an expected distribution, was used to determine whether the accuracy of the visualizations was statistically significant from a null hypothesis of 50% (i.e., guessing would achieve a score of approximately 50%). Next, to determine if the differences in accuracy for each method were significant, we utilized a χ² contingency test, which determines if the observed distributions in one or more categories deviated from an expected distribution with a null hypothesis that all methods had identical accuracy. Finally, binary logistical regression is commonly used as a statistical model for hypothesis testing of a continuous independent variable and a binary dependent variable [40]. We used a logit function (i.e., binary logistical regression) to evaluate whether the accuracy of visualization methods was sensitive to increases in A’ with a null hypothesis that it was not. For all tests, we consider significance to be p < .05, but we report exact p values to 3 digits for completeness.

7 RESULTS
We conducted the institutional review board-approved study using participants from Amazon Mechanical Turk. The experiment generally took less than 15 minutes, and each participant was compensated $2 USD. There were 120 participants filtered by region (US and Canada) and HIT Approval Rate (> 95%), of which 18 provided problematic data (i.e., the participant failed to engage with the study because they did not answer the complete 24 questions, frequently timed out during the experiment, or their median response time was less than 5 seconds). Hence, we analyzed 2,448 trials from the remaining 102 participants.

Participant Demographics: 57% of the participants were male, 42% female, and 1% nonbinary. All participants were 18 years old or older (see Fig. 8a). 92% of the participants reported having casual, minimal, or no visualization experience (see Fig. 8b).

In Sect. 7.1, we present the overall accuracy of participants for position and amplitude variations in data features. Next, Sect. 7.2 reports the sensitivity of participants’ decisions to feature variations. Finally, we report observations of the timing and interactions and collected feedback from participants in Sect. 7.3 and Sect. 7.4, respectively.

7.1 Overall Task Accuracy
We begin by evaluating the accuracy of the different visualizations.

Positional Variation: The results for positional variation can be seen in Table 1. To test whether the overall accuracy was statistically significant, a binomial test was conducted on each visualization, with a null hypothesis of 50% accuracy (i.e., guessing). All visualizations except Reeb graphs showed significance. To test for the significance of the differences in the overall accuracy between methods, we performed a χ² contingency test with the null hypothesis that all methods perform identically. The result, χ²(3, N = 1191) = 8.09, p = .044, shows the differences are significant. Overall, color maps performed the best of all visualizations, persistence diagrams and isocontours are almost tied in terms of accuracy, and Reeb graphs were the least accurate. However, it is important to note that our concern is with the sensitivity as A’ increases, not the overall accuracy.

Table 1: Summary of the results for the positional variation trials. Accuracy was measured as N_correct/(N_correct + N_incorrect). Bold indicates statistical significance (p < .05) calculated using a binomial test.

| Method                  | Trials | N_correct | N_incorrect | Accuracy | | Accuracy | Overall | p < .05 |
|-------------------------|--------|-----------|-------------|----------|----------|---------|---------|
| Color maps              | 300    | 187       | 103         | 62.3%    | 61.4%    | 56.6%   | 67.4%   |
| Isocontours             | 287    | 163       | 114         | 54.7%    | 59.3%    | 56.7%   | 56.0%   |
| Reeb graphs             | 307    | 156       | 141         | 42.7%    | 47.4%    | 55.1%   | 58.9%   |
| Persistence diagrams    | 331    | 188       | 133         | 56.5%    | 66.2%    | 55.2%   | 50.6%   |

Amplitude Variation: The results for amplitude variation can be seen in Table 2. The same binomial test was conducted to verify that accuracy was statistically significant. Again, all methods except Reeb graphs showed significance. A χ² contingency test was also used to compare the accuracy of the methods. The result, χ²(3, N = 1187) = 7.58, p = .056, shows the differences are just outside the range of significance. Nevertheless, persistence diagrams offered the best accuracy, followed by color maps, isocontours, and finally, Reeb graphs. However, we again note that our primary focus is sensitivity to changes in A’, not overall accuracy.

Statistics of Unsure Selections: Tables 1 and 2 show a small number of unsure selections, and participants answered the questions more than 95% of the time for all visualizations. Therefore, the results did not show any significant differences between visualization types.

Fig. 8: Bar charts of the demographic survey of participants.
We next evaluate whether using any of the visualizations resulted in a upward trend; however, it is slightly outside of statistical significance.

7.2 Sensitivity to Variation

Here, we evaluate the sensitivity of different visualizations to changes in features (i.e., $A'$). Please refer to Sect. 5.2.2 describing visualization sensitivity with respect to parameter $A'$. Ideally, as $A'$ increases, the accuracy of selections should also increase for a visualization.

Positional Variation: Fig. 9(top) shows the bar charts and the logistic regression of positional variation accuracy for different $A'$ values for each visualization type (see Table 1 for exact accuracies). For color maps and isocontours, the nearly flat line and large $p$ values indicate that they are not sensitive to increases in $A'$. Persistence diagrams show a downward trend that, although not statistically significant, suggests that higher $A'$ values were detrimental to the accuracy. Finally, the Reeb graph is the only method that shows a statistically significant upward trajectory. In other words, for positional variations, Reeb graphs are the only method sensitive to changes in $A'$.

Amplitude Variation: Fig. 9(bottom) shows the bar charts and logistic regression for different $A'$ values of each visualization for changes in amplitude (see Table 2 for exact accuracies). Color maps show a statistically significant upward trend. Isocontours also have an upward trend; however, it is slightly outside of statistical significance. Reeb graphs show a flat line and high $p$-value, which indicates that they are not sensitive to changes in $A'$. Lastly, persistence diagrams react the most to increases in $A'$, as they show the steepest statistically significant upward trend.

7.3 Time and Interaction

We next evaluate whether using any of the visualizations resulted in longer time or more interactions from participants.

7.4 Qualitative Feedback

At the end of the experiment, participants had the opportunity to provide several forms of qualitative feedback. The responses provided by individual participants are documented in pp. 39-47 of the supplement.

7.5 Perceived Difficulty of Visualization Methods: Participants had the option to indicate which visualizations they thought were easiest and most challenging to use for the comparison task (see Fig. 10c). The overwhelming majority considered color maps to be the easiest. The results for the hardest to use were mixed, but they did indicate that Reeb graphs and persistence diagrams tended to be harder to understand, even though they had the best performance on positional and amplitude variations, respectively.

Participant Feedback: Participants were also asked to provide additional information on what approaches they used to answer the questions. The typical response for color maps was that they tried to find the variation that shared the most significant number of cold and hard events.

Table 2: Summary of the results for the amplitude variation trials. Accuracy was measured as $N_{correct}/(N_{correct} + N_{incorrect})$. Bold indicates statistical significance ($p < .05$) calculated using a binomial test.

<table>
<thead>
<tr>
<th>Method</th>
<th>Trials</th>
<th>$N_{correct}$</th>
<th>$N_{incorrect}$</th>
<th>$A'$ = 15</th>
<th>$A'$ = 30</th>
<th>$A'$ = 45</th>
<th>$A'$ = 60</th>
<th>Accuracy</th>
<th>Overall</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Color maps</td>
<td>311</td>
<td>178</td>
<td>122</td>
<td>48.8%</td>
<td>52.9%</td>
<td>60.3%</td>
<td>67.1%</td>
<td>48.8%</td>
<td>55.8%</td>
<td>52.8% (p = .022)</td>
</tr>
<tr>
<td>Isocontours</td>
<td>329</td>
<td>177</td>
<td>140</td>
<td>48.7%</td>
<td>49.4%</td>
<td>54.9%</td>
<td>62.5%</td>
<td>52.8%</td>
<td>55.8%</td>
<td>52.8% (p = .024)</td>
</tr>
<tr>
<td>Persistence diagrams</td>
<td>304</td>
<td>155</td>
<td>142</td>
<td>52.8%</td>
<td>48.1%</td>
<td>48.5%</td>
<td>53.9%</td>
<td>52.8%</td>
<td>55.8%</td>
<td>52.8% (p = .001)</td>
</tr>
<tr>
<td>Reeb graph</td>
<td>279</td>
<td>172</td>
<td>101</td>
<td>47.4%</td>
<td>57.5%</td>
<td>65.7%</td>
<td>81.3%</td>
<td>47.4%</td>
<td>57.5%</td>
<td>57.5% (p = .000)</td>
</tr>
</tbody>
</table>

Time Taken: The time it took for participants to answer each question was similar for each visualization, as shown in Fig. 10a. We also found no clear trend between the time it took a participant to answer a question and the accuracy of their answer.

Interactions: We also evaluated accuracy when participants used some form of interaction (i.e., mouse click, mouse scroll, slider movement). We found no significant relationship between the use of interactions and the accuracy of the answers, as shown in Fig. 10b.

Fig. 10: (a) A box plot summarizing the time it took participants to answer for each visualization type. (b) A bar chart showing the accuracy of answers with and without interactions for each visualization type. (c) Results from the post-experiment survey that asked participants what visualization type was the easiest and hardest to use.
were overall in the least agreement with our expectations. In terms of overall, we can confirm color map sensitivity to amplitude variations Through our analysis of accuracy and sensitivity, color maps excelled technically, this confirms our hypothesis (that isocontours have no sen-

variations of their supported type, whereas other variations may be lost. In addition, Reeb graphs showed no statistically significant sensitivity in the visualization. In addition, Reeb graphs showed no statistically significant sensitivity to variations in amplitude, thereby confirming this hypothesis as well. Overall, as predicted, Reeb graphs showed sensitivity to position variations but not amplitude variations.

**Persistence Diagrams:** The results for persistence diagrams showed that, as hypothesized, they were not sensitive to positional variations. Interestingly, they showed a non-statistically significant negative trend as $A'$ increased. We speculate that this is due to artifacts caused by birth-death pairing switches that occur when critical points move apart. For amplitude variation, persistence diagrams showed both the highest accuracy and sensitivity to variations in $A'$, thereby confirming this hypothesis. Overall, as predicted, persistence diagrams were sensitive to amplitude variations but were not sensitive to positional variations.

**8.2 Implications**

There are several important implications for our findings.

**Rejected Hypotheses:** The several hypotheses that were rejected might be as important as those that were confirmed. Our hypotheses come from multiple decades worth of combined experience in scientific visualization. These rejected hypotheses potentially signify important misconceptions about the effectiveness of certain visualization types that further studies may illuminate.

**Color Maps and Isocontours:** One surprising result to us was the relative strength of color maps and the weakness of isocontours. Through our analysis of accuracy and sensitivity, color maps excelled in several aspects, whereas isocontours stood out in none.

**Reeb Graphs and Persistence Diagrams:** Reeb graphs and persistence diagrams have demonstrated very precise utility within the context of our study of positional variation for Reeb graphs and amplitude variation for persistence diagrams. This quality can be seen as both an asset and a liability because it means each will highlight variations of their supported type, whereas other variations may be lost.