Measuring the Effects of Scalar and Spherical Colormaps on Ensembles of DMRI Tubes

Jian Chen, *Member, IEEE*, Guohao Zhang, *Student Member, IEEE*,
Wesley Chiou, David H. Laidlaw, *Fellow, IEEE*, and Alexander P. Auchus

Abstract—We report empirical study results on the color encoding of ensemble scalar and orientation to visualize diffusion magnetic resonance imaging (DMRI) tubes. The experiment tested six scalar colormaps for average fractional anisotropy (FA) tasks (grayscale, blackbody, diverging, isoluminant-rainbow, extended-blackbody, and coolwarm) and four three-dimensional (3D) spherical colormaps for tract tracing tasks (uniform gray, absolute, eigenmaps, and Boy's surface embedding). We found that extended-blackbody, coolwarm, and blackbody remain the best three approaches for identifying ensemble average in 3D. Isoluminant-rainbow colormap led to the same ensemble mean accuracy as other colormaps. However, more than 50% of the answers consistently had higher estimates of the ensemble average, independent of the mean values. The number of hues, not luminance, influences ensemble estimates of mean values. For ensemble orientation-tracing tasks, we found that both Boy's surface embedding (greatest spatial resolution and contrast) and absolute colormaps (lowest spatial resolution and contrast) led to more accurate answers than the eigenmaps scheme (medium resolution and contrast), acting as the uncanny-valley phenomenon of visualization design in terms of accuracy. Absolute colormap broadly used in brain science is a good default spherical colormap. We could conclude from our study that human visual processing of a chunk of colors differs from that of single colors.

Index Terms—Ensemble visualization, diffusion magnetic resonance imaging, quantitative validation, colormap.

1 Introduction

Exploratory vector and tensor field visualizations studying regions of interest or a group of objects at a time [1] count on the human visual system to extract statistical information from features. Perceiving average or other statistical features from a group of similar items, called *ensemble perception* [2] [3], is a robust visual phenomenon studied largely in vision science that operates across a host of visual dimensions: size [4], orientation [5], position [6], motion [7], speed [8], number [9], identities [10], structures [11], and luminance [12].

The applicability of these vision science results to visualizations is anecdotal because of at least two methodological barriers between these two domains. Vision science studies are intended to capture static views, separate perception and cognition from interaction, and also separate domain-specific uses from visual stimuli. In contrast, in visual exploration these factors must be integrated. Additionally, spatial visualization features, such as continuity, symmetry, and clusters, may not be present in images.

Working in collaboration with brain scientists, we have recognized two challenges for showing three-dimensional (3D) diffusion magnetic resonance imaging (DMRI) tractography. The first is to support univariate representations of an ensemble of scalar values. Scalars are commonly encoded in one-dimensional (1D) colormaps, e.g., showing fractional anisotropy (FA) measured at every voxel to quantify disease states [13]. Though univariate coloring has been extensively studied in two-dimensional (2D) data visualizations (see the excellent reviews by Zhou and Hansen [14] and Silva et al. [15]), 3D color ensembles may introduce constraints in three respects. First, the univariate schemes of luminance and hue combination that work well in 2D may not apply in 3D, because luminance contrast can belie color constancy and distort 3D shape perception due to lighting [16]. Second, shading prevents the use of dark colors [17] [18], thus reducing the number of differentiable color steps. Third, interpreting ensembles may not require visually deriving individual values [3]. Since scientific data are often continuous, the human visual system may well optimize strategies for efficient visual detection [19].

The second challenge is showing structural connectivities from tracts (often rendered as tubes). This task requires the viewer to visually segment collections of tracts of various orientations. Szafir et al. call this type of task *ensemble subset extraction* [20]; Phadke et al. call it *attribute value exploration* [1]. DMRI tracts, unlike data in these studies are continuous in space and carry domain-specific attributes such as *symmetry* and *proximate regions* [21] [22]. Tracts are often colored with spherical colormaps, i.e., every point on a sphere is assigned a color to the orientation. Novel tract colormaps have been explored concerning *tract locality* (trajectories closer in the high-dimensional space remain close in the low-dimensional space) [23], *angular uniformity* (angular difference is perceptually uniform) [21], and *spatial resolution* (lines with different directions get different colors) [22]. No design knowledge exists, however, to quantify the practicality of spherical colormaps in visualization.

- *J. Chen is with the Computer Science and Engineering Department, The Ohio State University, OH 43210; E-mail: chen.8028@osu.edu.*
- *G. Zhang and W. Chiou are with the Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore County, MD 21025. E-mail: {guohaozhang, wchiou1}@umbc.edu.*
- *D.H. Laidlaw is with Computer Science Department, Brown University, RI 02912. E-mail: dhl@cs.brown.edu.*
- *A.P. Auchus is with the Neurology Department at the University of Mississippi Medical Center, Jackson, MS 39216. E-mail: aauchus@umc.edu.*

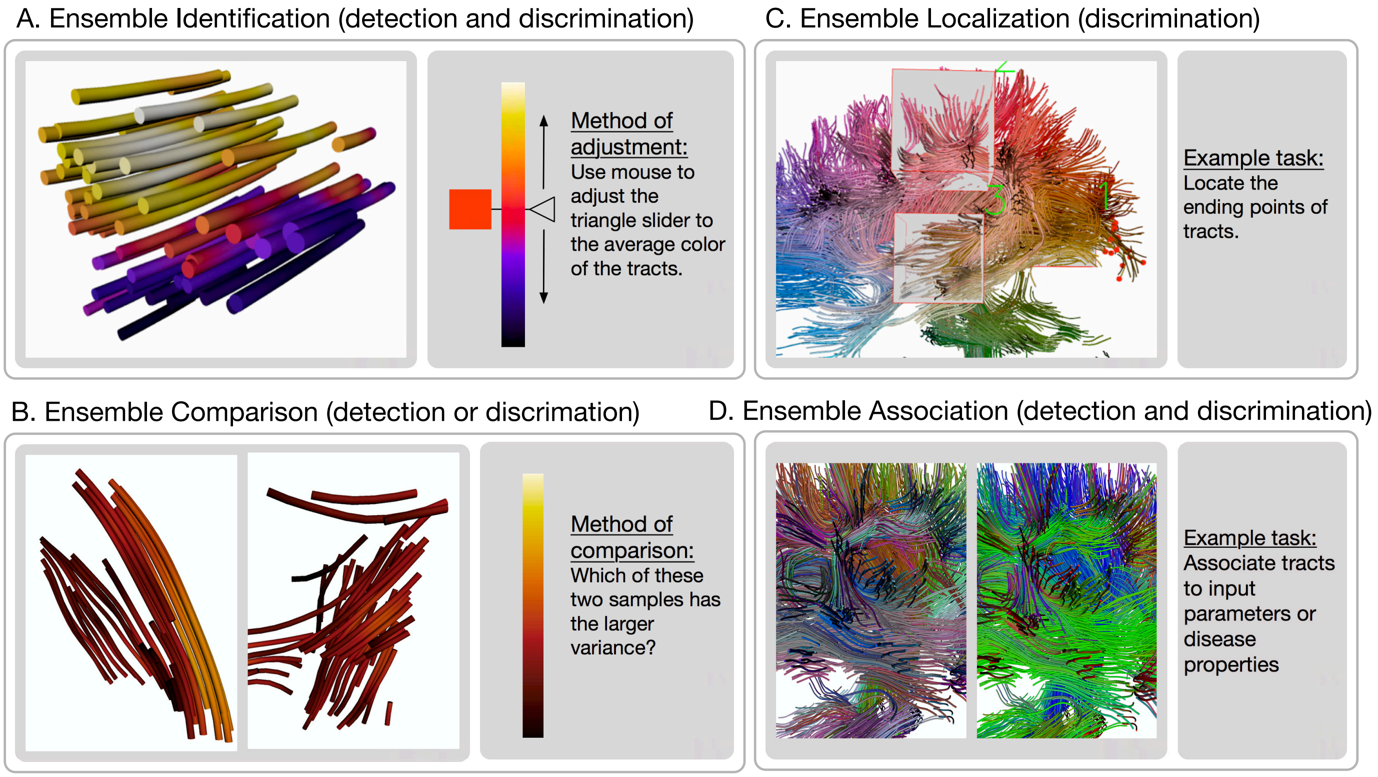

A. Ensemble Identification (detection and discrimination)
Method of adjustment:
Use mouse to adjust the triangle slider to the average color of the tracts.
C. Ensemble Localization (discrimination)
Example task:
Locate the ending points of tracts.
B. Ensemble Comparison (detection or discrimation)
Method of comparison:
Which of these two samples has the larger variance?
D. Ensemble Association (detection and discrimination)
Example task:
Associate tracts to input parameters or disease properties

Fig. 1: Four Types of Common Ensemble Tasks (*Identification*, *Comparison*, *Localization*, and *Association*) and Some Selected Methods for Testing Ensemble Representations.

The present work addresses these two important challenges by first summarizing a set of ensemble tasks of *identification*, *localization*, *comparison*, and *association* (Fig. 1). We then examine two identification tasks, ensemble average (Fig. 2) and orientation (Table 1 and Fig. 3), by evaluating state-of-the-art coloring methods. Specifically, we answer the following questions: *How reliable are colormaps for deriving ensemble averages from 3D spatially distributed tracts? Which colormaps are applicable to ensemble average? What is the most effective spherical colormap for presenting orientation in ensemble representations?*

Our work makes the following contributions.

- Formally proposes and expands ensemble visualization concepts inspired by vision science.
- Establishes new measurement metrics for bias analysis and associates this bias to data distribution in evaluation when recommending colormaps.
- Derives some design recommendations for spatially continuous datasets for ensemble average and orientation discrimination.

2 Terms and Related Work

Our work draws upon work related to (1) ensemble color studies in vision science and (2) univariate and spherical representations in visualization. In this section, we broaden the definition of ensemble representation in visualization and connect color theory to ensembles and relevant study results.

2.1 Ensemble Representation: Definition

The *ensemble* concept in visualization often refers to a collection of datasets and is perhaps best known as ensemble simulation and uncertainty quantifications [24] [25]. Ensemble has been broadly studied in vision science (e.g., [26]), where *ensemble representation* is used to explore how humans use statistical regularities in a group of similar objects to process information [27].

Our current work supports this recent broad perspective on the role of visual statistical processing and embraces the idea that these visualization tasks, whether from ensemble simulations (e.g., statistical properties such as uncertainty [28] [29]) or not (e.g., overviews and detecting global features in flow fields [30] and areas or sets [31]), share the property that multiple measurements are combined to give rise to a higher-level statistical description.

Following this new human information processing perspective, we formalize **ensemble representation** as an umbrella term encompassing existing 3D visualization methods that demand the human visual system to derive statistical attributes from visualization. For example, correlated textures along vector fields help humans derive ensemble patterns to see flow movement; glyphs enable efficient visual assessment of "a chunk of flow" [32].

2.2 Color Ensembles

Representing **color ensembles** concerns how our visual system derives statistical information through visual processing of color features. Color ensembles can facilitate scalable visual inspection. Mauly and Franklin [33] study a series

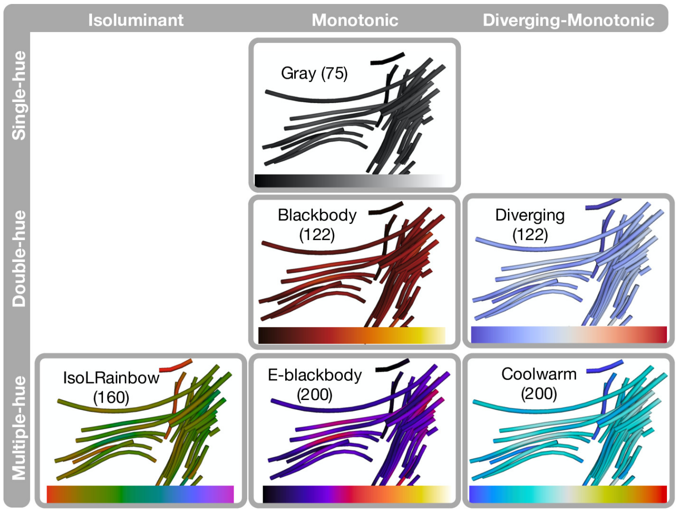


Fig. 2: Six Univariate Colormaps for Ensemble Average Task (Task 1) and Their Attributes. The numbers after colormap names are arc-lengths in the L*A*B* color space. Arc-length is computed with CIEDE 2000 by summing the ΔE^*_{00} values along the curve following the colormap in the L*A*B* color space, where ΔE^*_{00} specifies the perceptual uniformity between adjacent points along the curve [41].

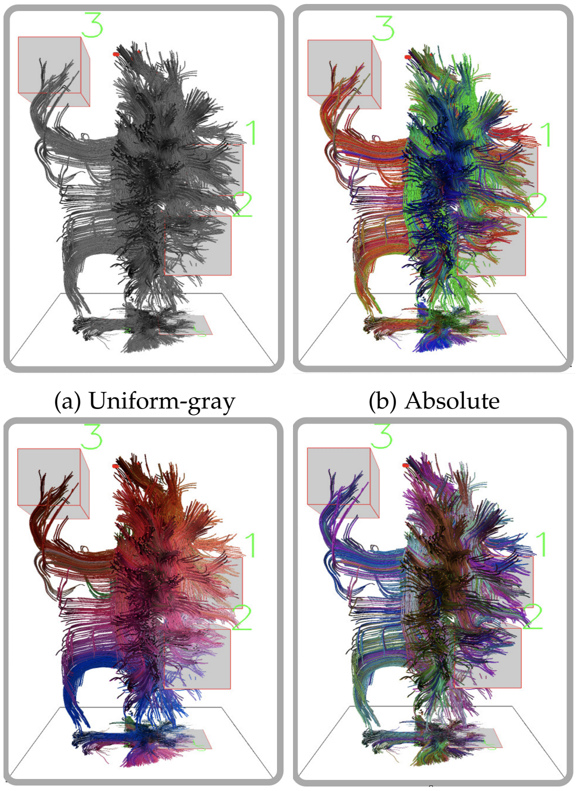
of uniformly colored circular elements ranging from 4 to 48 items subtended at 12, 20, and 28 just-noticeable differences (JNDs) and suggest that the accuracy is insensitive to changes in the number of elements in an ensemble. Only reaction time is longer for ensembles with more hues.

Dedicated human visual processing of color ensembles may also exist [34]. Human vision can effectively discount spectral variations and assign stable colors to objects to achieve consistent scene [35] and color constancy [16]. In a 2D time-varying chart visualization, Correll et al. find that color is more effective than position for showing averages and distributions [36]; this result is contradictory to classical design recommendations in which position is more accurate than color for quantitative comparisons, when ensemble is not required [37]. Also reliable average estimates can be made from two hues of red-blue, blue-green, and yellow-green [38] and categorical boundaries can be accurately labeled for greenish-blue, bluish-green [39], and gray-scale alike textures [40].

These intriguing results on hue ensembles, mostly presented in vision science, seem to refute the idea that hues cannot be described in terms of magnitude but only as qualitative experiences. They may be effective for ensemble averages and boundary detection when the data or hue variance is localized and small. In this work, we choose several multihue colormaps, such as extended-blackbody and coolwarm (Fig. 2). We also use a well-designed rainbow colormap, *aka* Kindlmann et al.'s isoluminant-rainbow [42]. We compare these approaches against other double- and single-hue methods.

TABLE 1: Four Spherical Colormaps for Ensemble Orientation Task (Task 2): Their Attributes.

Colormaps	Contrast	Angular uniformity	Spatial resolution
Uniform-gray	Very low	No	Very low
Absolute [21]	Low	No	Low
Eigenmaps [23]	Medium	Yes	Medium
Boy's Surface [22]	High	Yes	High



(a) Uniform-gray (b) Absolute

(c) Eigenmaps (d) Boy's Surface

Fig. 3: Four Spherical Colormaps for Ensemble Orientation Task (Task 2).

2.3 Univariate Coloring

The most influential color studies lie in univariate colormap design and characterizations (e.g., color harmony and categories [43] [44], metrics [45], and modeling [46]). Silva et al. [15] and Zhou and Hansen [14] summarize important color characteristics in univariate colormap design, such as *ordering* (colormaps must preserve the order in data), *separation* (different data must be perceived differently) [47], and *uniformity* (perceived differences in color must accurately reflect numerical data differences). Among these characteristics, uniformity is believed most important for showing quantitative data [48]. Rainbow colormap is believed to be poor at showing quantitative data because it lacks nearly all these attributes.

This design knowledge leads us to adopt several univariate maps suggested by Moreland [18], including extended-blackbody (monotonic luminance and multihue), blackbody (perceptually uniform, monotonic luminance, and multihue), coolwarm (perceptually uniform, two-hues and monotonic on each side), and diverging (two-hues and perceptually uniform and monotonic on each side) (Fig. 2). Some of them have also been incorporated in the popular 3D visualization tools VTK and Paraview.

2.4 Vector and Tensor Field Evaluation

Pioneering 3D vector and tensor field studies have largely focused on univariate comparisons, such as vector speed between two locations [32], tracing a single tract [49], reading quantities at sampling sites [50], and showing depth and distances between adjacent occluded tracts [51] [52]. An exception is the study by Acevedo and Laidlaw [53] in which participants discriminate boundaries through a set of size-varying circles and must visually derive groups from visualization.

Borkin et al.'s work [54] closely resembles ours in terms of colormap comparisons to support seeing in 3D. That study compares rainbow and diverging colormaps for detecting regions of heart diseases after projecting 3D artery flow patterns to 2D and finds that a rainbow colormap decreases detection rates [54]. The present work builds on these studies but expands the scope in two important ways: we measure more tasks to understand ensemble averages discrimination and orientation detection, and our tasks are in 3D. We further formalize the task space in Chen et al. [55] for ensemble univariate and orientation discrimination.

2.5 Continuous Ensemble Spherical Colormaps

Knowledge about effective spherical colormap design is limited, despite their importance for showing tensor and vector fields. To show brain connectivity through tracts, Pajevic and Pierpaoli [21] use elegant solutions through extensive studies on *rotation* and *mirror symmetry*. The absolute values of the xyz-components of the principal diffusion tensor eigenvectors are mapped directly to RGB color-triples. The advantages of this *absolute* approach include: (1) user familiarity with RGB colors associated with a vertebrate direction, (2) high contrast between vertebrate directions, and (3) four-way symmetry (left-right, dorsal-ventral, anterior-posterior, and antipodal.) Even though this absolute encoding approach provides a seemingly low-resolution view of tract orientation, our brain scientist collaborators suggest that this colormap dominates brain science because it conveys most important transverse, sagittal, and coronal directions.

Other solutions reveal patterns and increase spatial resolutions. Kindlmann et al. introduced a *hue-ball* approach and a barycentric map for direct volume rendering of tensor fields by assigning color and opacity based on the direction of the principal eigenvector and anisotropy type of the diffusion tensor [56]. An attractive characteristic of this approach is its high contrast between adjacent tracts: they are colored with bright, saturated colors spanning from red, yellow, green, cyan, blue to purple. Demiralp et al. [22] use *Boy's real projective plane immersion* to visualize the direction of brain tracts. This *Boy's surface* coloring possesses good *locality* and *contrast* by showing the finest details, and has the greatest *spatial resolution* of all spherical colormaps (Table 1).

Vision science has studied multihue mainly as a pattern-segmentation mechanism for identifying structural variations. Maule et al. [57] suggest that there may be a functional limit to the amount of variance that can be rapidly encoded by summary statistics of set discriminations. Such set discriminations, though close to our orientation discrimination, can prescribe methods only for discrete clusters. No study exists to our knowledge to explore to what extent continuous spherical coloring of ensemble line field would be most beneficial. Our study compares four techniques to understand the effectiveness of ensemble orientation discriminations. Our expectation is largely driven by the vision science literature positioning that colormaps with higher resolution could improve the orientation detection.

3 Brain DMRI Data Characterization and Ensemble Tasks

This section first describes the data and task characterization by following Munzner's data and task abstraction method [58], and then presents our measurement method. The objective is to establish the importance of ensembles in the different tasks inspired by real-world relevant uses in DMRI analysis alluding to different questions that must be answered. The data and task abstraction also helps choose study tasks and encourages reuse of our study results.

3.1 Brain DMRI Data Characterization

DMRI measures water diffusion as a second-order positive-definite tensor [59]. Water diffusion patterns have been analyzed comprehensively by brain scientists to study anatomical fibrous structures. Modern advances have extended to meta-analysis of brain cohorts [60]. Visualization design guidelines for understanding complex spatial structures have also been a recent focus [61], albeit disproportionately small in the amount of empirical work directly focused on evaluation. Preim et al. [62] have surveyed perceptually-motivated 3D visualization for medical imaging visualization, but focused on depth and shading. This challenge in coloring MRI datasets is often cited as a top visualization challenge [19].

The first and most reliable benchmark measurement is fractional anisotropy (FA) [63]. FA, a normalized scalar, measures the water diffusion patterns: a value of zero means that apparent diffusivities do not depend on direction, for instance when diffusion is restricted equally in all directions (e.g., in gray matter.) A value of one means that diffusion occurs only along one axis and is fully restricted (e.g., in white matter.) Brain scientists are concerned with average FAs in regions containing a set of voxels or tracts. In this study, FAs are in the range of [0.2, 1] and average FAs are [0.25, 0.85].

Another important measurement is brain structural connectivities [55]. A continuous diffusion tensor field is first constructed from the measured DMRI data. Tracts are then computed at voxel sampling locations via tractography, a 3D technique for representing brain structural connectivity [64]. We terminate tract tracing when the FA value is less than 0.2. The tracts are depicted to show connectivity information. A group of tracts sharing similar orientations is called a *bundle*. Some studies use template-based approaches to derive and color tracts to show anatomical connectivity; others attempt to visualize the structures independent of templates. Our current work studies four major bundles labeled by our collaborators.

Several brain analysis tools and methods have supported colormaps. For example, DTI Studio lets users manually assign selected tracts a color as well as use the default randomly assigned colors for individual tracts [65]. 3D Slicer lets users select among a large variety of colormaps or customize their own for visualizing variables [66]. While these tools offer great flexibility, our results can give users more informed design choices among techniques and tools.

3.2 Ensemble Task Characterization: Four Types

We obtain the following measurable low-level tasks (Fig. 1). In each category, we separate detection (e.g., which is higher?) and discrimination (e.g., how much higher?) tasks inspired by Borgo et al. [67] and Zhao et al. [50], so as to address the goal of design for perceptually accurate visualizations.

1) *Ensemble identification* is performed when the goal is to read mean values or estimate the probability distributions of values from *similar* objects. Some typical identification tasks are: what are the average FA values (Fig. 1(A))? Where is the boundary between regions of different anatomical structures? Do the two bundles belong to different groups? What is the average brain?
2) *Ensemble comparison* is useful to compare multiple ensembles or items or identify the most common outputs. Some example tasks are: are the left and the right hemispheres of CC different? If so, by how much? Do the brain cortical surface shapes differ? The task in Fig. 1(B) compares between diseases outcomes in cohorts.
3) *Ensemble localization* asks the viewer to find where a certain ensemble value or attribute (e.g., inliers and outliers) is located within the data. Fig. 1(C) stresses visual lookup and asks where the lesion is in the brain. Where are regions of maximum and minimum mean FA values?
4) *Ensemble association* involves determining the associative relationships between or among *related* objects. Fig. 1(D) shows the average tracts computed from ensembles. Some example tasks are: which of these two average brains is associated with dementia? And at what state of the dementia? Using a simulator and after varying parameter A, what are the associated brain regions sensitive to these inputs, and what is the distribution of the changes among these output ensembles?

Based on this task characterization, we choose to study the first *identification* task in the current work because they are most common DMRI visualization challenges.

3.3 Metric

There are several considerations in measuring the ensemble representations. We divide the data or the colormap into bins to represent sub-regions. This is because a region of interest (ROI) in a spatial volume is likely to be localized to a group of data points. Also, we can associate the data distributions in each bin to color distributions in a colormap to understand colormap usefulness. For example, the spread or variance of the resulting distribution in each bin in a colormap reflects the ensemble average performance. The shape of the results also reflects the sensitivity of features or dimensions to the ensembles. Robust sensitivity to summary statistics will yield a narrow distribution. A function can also be fitted to the data to reveal sensitivity to the discriminative threshold to measure accuracy. In this work, for ensemble average we divide the input data into 12 bins and randomly sample the data such that the sample dataset is a high-fidelity representation of the DMRI tract attributes. For orientation detection tasks, we follow past practice and measure the responses to spherical colormaps by randomly sampling the input.

4 Ensemble Experiment for Brain DMRI Visualizations

The objective here is to determine which ensemble colormaps are more accurate for showing DMRI datasets. We are particularly interested in the first task type, ensemble identification (of mean and orientation) (Section 3.2).

4.1 Exploratory Study Method

Given our own experience, our collaborators' subjective choices, several pilot study results (see Appendix B) and the vision science literature, we carry out an exploratory study by comparing several scalar and spherical colormaps in two tasks. One could use a hypothesis-driven confirmatory analysis. However, due to the lack of visualization literature in this ensemble visualization area, we resisted in the beginning to pose hypotheses. Rather, the goal in this exploratory study is to try to explore some of our initial expectations and report many outcomes (efficacy, effectiveness, and associations of data and result distributions) to facilitate deeper insight into color representation and future experiments. Some of our expectations include: (1) multiple-hue colormaps could support ensemble average tasks; (2) Higher spatial resolution could improve orientation accuracy for ensemble orientation detection; And (3) having color may be better than no-color uniform representation for identifying orientations.

4.2 Three-Dimensional Ensemble Tasks

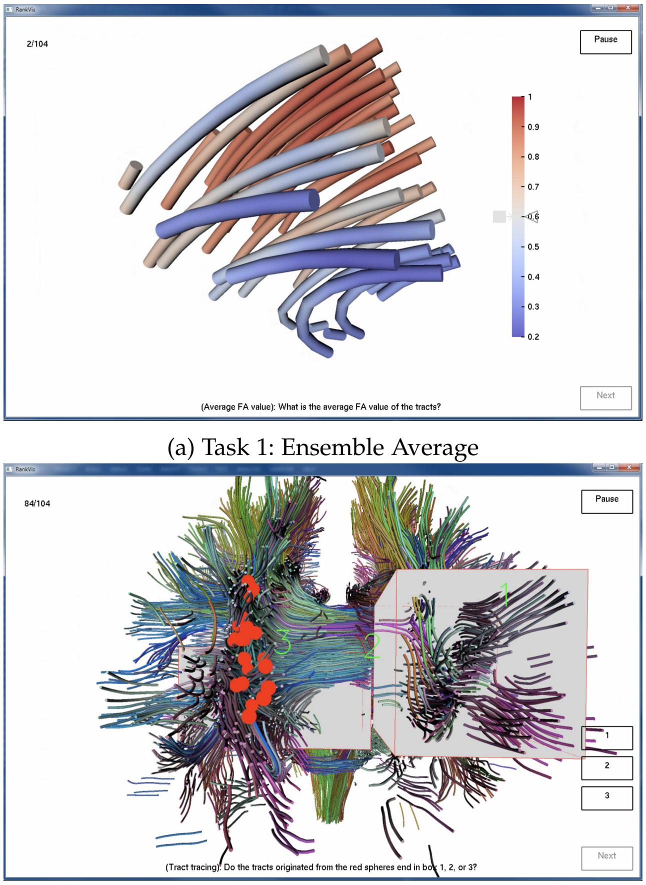
4.2.1 Task 1: Ensemble Average (Discrimination task)

Figure 4a shows an example task in which participants were asked to label the average FA values of the brain areas sampled in a ROI. The participants indicate their answer for each task by dragging the slider on the screen to show the average color. The answers are evenly distributed along the 12 bins (see Section 5.3) so that participants are not biased.

4.2.2 Task 2: Ensemble Orientation (Detection task)

Figure 4b shows an example task in which participants were asked to find the one box among three in which the endpoints of the tracts marked by red spheres lay at one end of the tracts. Participants were told that the marked tracts in the same bundle followed the same orientation (anterior-posterior, dorsal-ventral, or left-right).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2898438, IEEE Transactions on Visualization and Computer Graphics



(Average FA value): What is the average FA value of the tracts?

(a) Task 1: Ensemble Average

(Tract tracing): Do the tracts originated from the red spheres end in box 1, 2, or 3?

(b) Task 2: Ensemble Orientation

Fig. 4: Two Ensemble Identification Tasks in the Empirical Study. (a) What is the average value of the tracts? This example uses the diverging colormap. (b) Do the tubes originating from the red spheres end in box 1, 2 or 3? This example uses a Boy's surface colormap.

4.3 Choosing Ensemble Colormaps

Figures 2 and 3 and Table 1 summarize the colormap choices and their characteristics. The following two subsections give a detailed explanation of these choices. Readers not interested in these details can go directly to Section 4.4.

4.3.1 Six Univariate Colormaps for Ensemble Average

Six univariate colormaps shown in Figure 2 are measured in task 1 (Ensemble Average). These colormaps are chosen due to their popularity, relevance to our collaborators' recommendations. All color interpolation is performed using linear interpolation in this L*A*B* color space. The dark part is cut out to keep the arc-lengths among colormaps as close as possible and to consider low sensitivity of human vision to low luminance values. Appendix A shows the colormap profile in the L*A*B* color space.

The *grayscale colormap* uses a single-hue and monotonic luminance with arc-length 75.

The *blackbody colormap* is a double-hue and monotonic luminance map inspired by the wavelengths of light from blackbody radiation. We use arc-length 122 instead of 145 to match that of the diverging map. We removed the dark end due to the low sensitivity to low luminance values.

The *diverging* colormap contains two hues and increases/decreases luminance monotonically with arc-length 122. The closer the color is to the center of the colormap, the higher the luminance.

The *isoluminant-rainbow colormap* displays multihue rainbow with arc-length 160. It is isoluminant for the standard viewer, with the luminance level of 50.

The *extended blackbody colormap* is a monotonic luminance colormap and adds blue and purple hues to the blackbody map described above with arc-length 200.

The *coolwarm colormap* has monotonically increased and decreased luminance. This colormap has the same luminance range and variations along the luminance direction as the diverging map; it adds yellow and cyan hues to the diverging map because these two hues are common transitions in coolwarm colormaps that use red and blue.

4.3.2 Four Spherical Colormaps for Ensemble Orientation

The four spherical colormaps shown in Figure 3 are used in task 2 (ensemble tract tracing).

Uniform gray is used as a control condition.

Absolute RGB color-triples uses Pajevic's approach [21] in which the three different orientations (left-right, dorsal-ventral, anterior-posterior) are represented as red (R), green (G), and blue (B). Each tract uses a constant color indicating its global orientation. This is a most popular spherical colormap applied in brain science.

Eigenmap embedding implements the method of Brun et al. [23]. It assigns colors to tracts based on the similarities among tracts. The tracts become points in the embedded low-dimensional space [68] and the similarity of tracts is measured using the closeness of these points and a similarity matrix. The 3D coordinates of the points are normalized to fit into the displayable range of the L*A*B* color space and the corresponding colors are used for the tracts.

The *Boy's surface embedding* implements the method of Demiralp et al. [22], a one-to-one mapping between an orientation and a location in a color space based on a Boy's surface immersion in the color space. The embedding is also *angular uniform*, i.e., the larger the difference in tract orientations, the larger the perceptual difference in their colors.

4.4 Diffusion MRI Datasets

For ensemble average tasks, the average FA values were in the range [0.25, 0.85]. We evenly divided this range into 12 bins and the step size was 0.05. We randomly sampled within the four brain regions (here corpus callosum (CC), cortical spinal tracts (CST), inferior frontal occipital fasciculus (IFO), and inferior longitudinal occipitotemporal fasciculus (ILF)) by randomly placing boxes in these regions. We then took an equal number of samples in each bin from these samples. Fig. 5 shows the variance of the data in these 12 bins. We see that the lower and higher mean FA would have narrower spread (smaller variance) than those in the middle; this is the unique domain-specific data attribute.

Because ensemble mean is affected by variance [33], one way to conduct a study is to control the variance in each bin and measure the color effectiveness in each bin. We did not do this in order to retain a high-fidelity representation of

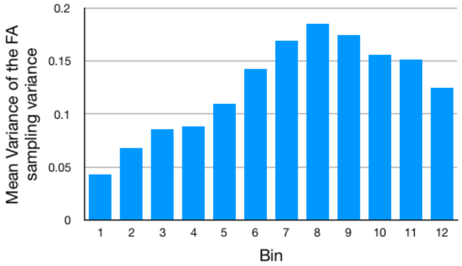


Fig. 5: Domain-Specific Data Attribute: The spreads (variances) of all AverageFA data in the 12 bins in our random samples are smaller in the lower bins ($\leq$ 5) and become most spread (with larger variances) when the bins ids in bins $[7, 9]$. From bin 1 to bin 12, the average FAs are 1: [0.25, 0.3), 2: [0.3, 0.35), 3: [0.35, 0.4), ..., 12: [0.8, 0.85] respectively.

tractography features; otherwise, we would have to produce artificial data to control the spread in each bin.

For orientation tasks, tractography data were computed from source DMRI images captured from a normal human brain at resolution $0.9375mm \times 0.9375mm \times 4.52mm$. Data are also sampled from four major bundles, CC, CST, IFO, and ILF. All tracts are rendered using tubes.

4.5 Experimental Design

Within-participant design was used for both tasks: i.e., each participant examined all colormaps. The independent variable is colormap. The dependent variables are completion time, accuracy, and subjective ratings. For task type 1 of ensemble average with 6 colormaps, each participant performed 12 instances (evenly distributed in the 12 bins) using the 6 colormaps (72 trials). Six instances of data (two CST, two CC, one ILF, and one IFO sample) and the six colormaps form a Latin square. No data was repetitively used by the same participant.

For task type 2 of the ensemble orientation using four colormaps, each participant performed eight instances of every colormap (32 trials). Again, datasets were not reused by the same participant. We ordered the four bundles and the four colormaps by a 4×4 Latin square. The order of the trials for each colormap was randomized.

Each participant performed $72 + 32 = 104$ sub-tasks.

4.6 Participants, Apparatus, and Environment

A total of 24 participants (17 male and 7 female) took part in the study: two medical professionals, seven computer science students, and 15 students from other disciplines (mechanical engineering, math, and global studies). Their average age was 27.8 years with standard deviation 4.0.

The program runs on a Linux desktop with a 27" monitor (BenQ GTG XL 2720Z, pixel resolution 1920 $\times$ 1080). Gamma was adjusted daily to ensure uniform perceived brightness: the gamma value used for the display was 2.2.

The lighting used fixed-pipeline OpenGL rendering with per-vertex lighting and Gouraud shading. We used a traditional three-point lighting scheme. Key and fill lights were placed in relation to a preset camera with 35mm focal length and the key light is at the top left of the scene, the location assumed by most human observers. Lighting placement and intensity were chosen to generate images with contrast and lighting properties appropriate for the data and human assumptions. For example, the key and fill lights were elevated and slightly to the left and right of the observer. All lights were white. The screen background color was white.

TABLE 2: Main Effects of Colormap on Accuracy and Task Completion Time. Here Avg. stands for ensemble average task, Ori. for ensemble orientation task, C for color, and P for participant. The *large* effect sizes are in **bold** and the medium ones in *italic*.

Avg.	C on error	$F_{(5,1728)} = 0.98$, p=0.43	d=0.16
	C on time	$F_{(5,1728)} = 6.23$**, p<0.0001**	*d=0.31*
	P on error	$F_{(23,1728)} = 2.77$**, p<0.0001**	**d=0.71**
	P on time	$F_{(23,1728)} = 50.24$**, p<0.0001**	**d=3.72**
Ori.	C on accuracy	$\chi^2_{(3,768)}$**=13.94, p=0.0030**	V=0.13
	C on time	$\chi^2_{(3,768)}$=0.67, p=0.57	d=0.13
	P on accuracy	$\chi^2_{(23,768)}$=2.35, p=0.43	V=0.17
	P on time	$\chi^2_{(23,768)}$**=4.35, p<0.0001**	**d=1.56**

4.7 Procedure

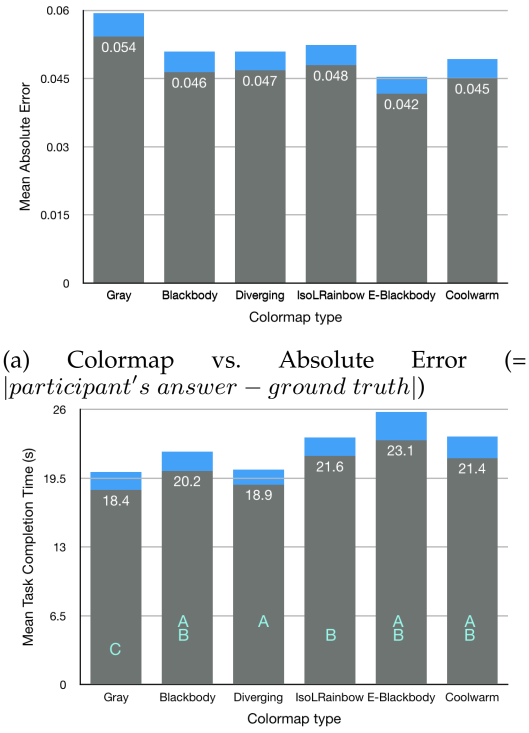
Participants were tested for normal vision and passed the Ishihara Color Vision test. They received general information about brain structure and about DMRI techniques and their medical uses. The training session, which lasted about 15 minutes, ensured that the participants understood the coloring and tasks.

Task completion time was recorded from the time when the visualization was shown on the screen to the time when the final answer button was clicked. Participants were told to be as accurate and as fast as possible, and that accuracy was more important than time. They were also told to rotate the data to better interpret the structures. They had to finish a task in order to go to the next one. No time limit was set on each task. They could take a break at any time. After finishing all sub-tasks using each colormap, they selected from a 7-point scale (1 (worst) to 7 (best)) on the computer screen to rate the map they just used. Finally, participants were interviewed for their comments. Participants took about an hour on average to finish this study and received monetary compensation. No fatigue was reported.

We conducted three pilot studies comparing performance with a total of 50 participants (including three brain scientists) to refine our experimental procedure. These pilot study participants were not used in the formal study. We recruited brain scientists to collect some domain-specific comments related to brain sciences on the color encoding methods. The main difference between expert and novice groups, as observed in our previous studies, was that experts took longer to complete task because they were more interested in examining the data. Our pilot studies revealed no significant difference in task completion time and accuracy between medical school students and other college students without medical backgrounds.

5 Results

We collected 1728 and 768 data points for the *ensemble average* and *orientation* tasks accordingly. To summarize, we



(a) Colormap vs. Absolute Error (= $|participant's\ answer - ground\ truth|$)

(b) Colormap vs. Absolute Task Completion Time

Fig. 6: Ensemble Average Tasks: Mean Absolute Error and Task Completion Time. The gray bars are the means and blue the 95% confidence intervals. (A). *Absolute error* = $|participant's\ answer - ground\ truth|$. (B). Colormaps labeled with the same letter belong to the same group in the post-hoc analysis.

found that the rainbow colormap was as accurate as other colormaps for ensemble average. The extended-blackbody, blackbody, coolwarm are the most accurate colormaps. Having some colors in the spherical colormap choices is always better than not.

5.1 Overview of Analysis Approaches and Summary Statistics

Results were analyzed by tasks. Table 2 shows the statistical analysis of accuracy and task completion time measured using the following statistical approaches. For both tasks, we examine the main effect of colormap on error and task completion time using the SAS GLM procedure. A post-hoc analysis using the Tukey Studentized Range test (HSD) is performed when we observe a significant main effect.

Task 1 (average tasks) performance is analyzed using several methods. Task completion time is converted to log-based to obtain a close-to-normal distribution. We compute *error* by the distance from the participants' answers to the ground truth and use the formula $error$ = $\log_2|participant's\ answer - ground\ truth| + 8$, following Cleveland and McGill [37]. We explore the accuracy of these ensemble colormaps using two additional measurements.

- **Accuracy.** Accuracy is percentage of correct answers. We threshold the error to measure whether an answer is correct, where the threshold $\delta \in$ [0.01-0.1], with step size 0.01 and error is $|participant's\ answer - ground\ truth|$. An answer is considered correct when error falls in δ.
- **Directional bias.** We compute whether or not the colormaps bias observers towards values larger or smaller than ground truth.

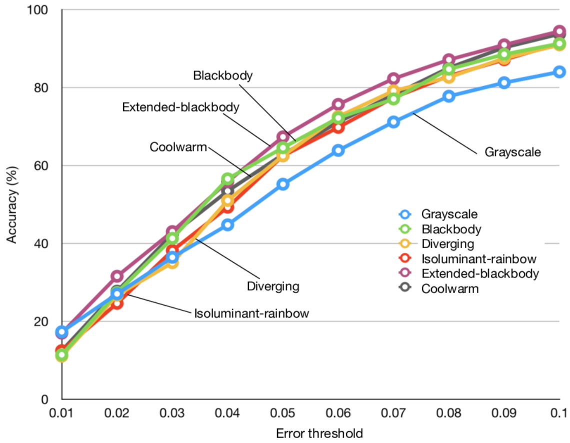


Fig. 7: Ensemble Average Tasks: Colormap Accuracy. An answer is considered correct if the *absolute error* (= $|participant's\ answer - ground\ truth|$) is less than the *error threshold* (the X axis).

The accuracy data in orientation tasks are binary and are analyzed using logistic regression and reported using the p value from the Wald χ^2 test. When the p value is less than 0.05, variable levels with 95% confidence interval of pairwise difference of odds ratios not overlapping are considered significantly different. The χ^2 test with the "$freq$" procedure is used to examine whether or not there is a significant correlation between the main effect (the colormap or participant) and accuracy.

We measure effect sizes using Cohen's d for time and task type I error and Cramer's V for correctness to understand the practical significance [69]. We used Cohen's benchmarks for "small"(0.07-0.21), "medium" (0.21-0.35), and "large" (> 0.35) effects.

5.2 Average Tasks: Summary Statistics

For ensemble average tasks, colormap was not a significant main effect on error (Table 2 and Fig. 6a). A general trend was that extended blackbody had the least error and gray had the most. Two-way interaction between participant and colormap on error was not significant.

Colormap and participant were significant main effects on time (Table 2 and Figure 6b). The post-hoc analysis suggests three Tukey groups: (gray), (blackbody, isoluminant-rainbow, extended-blackbody, and coolwarm), and (blackbody, diverging, extended-blackbody, and coolwarm). The extended-blackbody and coolwarm maps led to the longest task completion time and the gray, though efficient, had the highest error. Two-way interaction between participants and colormaps on time was not significant.

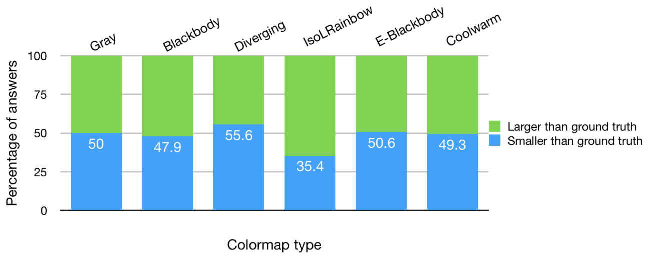


Fig. 8: Directional Biases by Colormap. More participants tend to overshoot (report larger than the ground truth in green) when using isoluminant-rainbow. Using the diverging colormap, more participants underestimated the ensemble average (in blue). Gray, extended-blackbody, and coolwarm had the minimum directional biases.

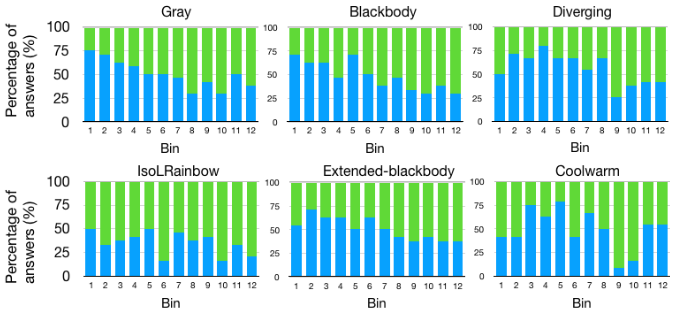


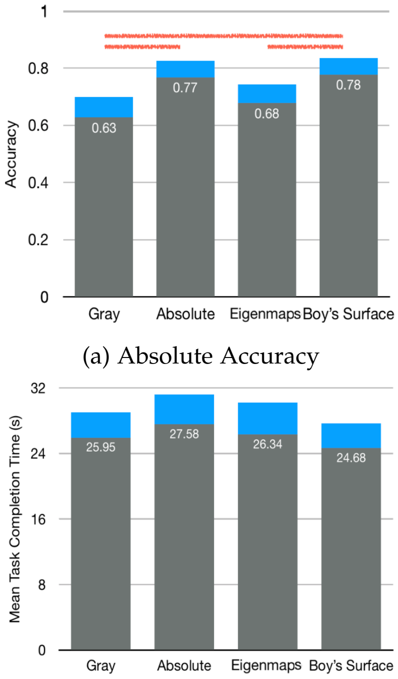
Fig. 9: Directional Biases by Colormap and Bin. More than 50% larger-than-ground-truth answers appeared in *all* 12 bins for isoluminant-rainbow.

5.3 Average Tasks: Color Sensitivity and Directional Bias

We computed the colormap sensitivity by measuring the percentage of correct answers or accuracy (Fig. 7). We first computed the mean absolute error. Fig. 7 showed that gray had on average the lowest accuracy among all colormaps (Fig. 6a). Gray-scale had similar accuracy to other colormaps when the error threshold was as low as 0.01. A general trend was that the slope of improvement was least for gray-scale to improve colormap accuracy (Fig. 7).

Directional bias measures if observers consistently chose larger or smaller values than the ground truth using a colormap. We found that more answers using isoluminant-rainbow were biased towards higher values, while the diverging color slightly towards lower answers (Fig. 8). All other colormaps of blackbody, extended-blackbody, and coolwarm showed about even distributions between higher and lower participants' answers.

We further analyzed the bias distribution in the 12 bins (Fig. 9). We found that more than 50% of the answers overshoot (selected larger than ground-truth) when using isoluminant-rainbow in *all* bins. Correlations between the data variance and colormap absolute error showed that these two variables were statistically significantly correlated for all but isoluminant-rainbow. This result may indicate that the ensemble behaviors of isoluminant-rainbow might not be as predicable, despite its accuracy for ensemble average was comparable to other colormaps.



(a) Absolute Accuracy

(b) Absolute Task Completion Time

Fig. 10: Ensemble Orientation Tasks: Mean Time and Accuracy. The color schemes connected by the orange line are significantly different.

5.4 Orientation Tasks: Summary Statistics

The second row in Table 2 shows the statistical results. Fig. 10a shows mean accuracy (percentage correct answers) and time and 95% confidence intervals from the mean. Colormap had a significant main effect on accuracy but not on task completion time. The Boy's surface embedding and the absolute embedding led to most accurate answers for following tracts and were among the most efficient (Figs. 10a and 10b). It was also noticeable that Boy's surface and absolute improved accuracy by 15% and 14% respectively compared to the gray.

5.5 Subjective Ratings and Comments

Participants' ratings and comments provide useful insights into how the usefulness of the colormaps was perceived. Participants' subjective rating of the usefulness of these colormaps, from high to low are: average tasks: coolwarm (5), extended-blackbody (4.96), blackbody (4.96), diverging (4.75), isoluminant-rainbow (4.4), and grayscale (3.7); Orientation tasks: absolute (5.3), eigenmaps (5.3), Boy's surface (4.8), and uniform-gray (2). Grayscale in task 1 and uniform-gray in task 2 were rated least useful.

The interviews revealed that those who liked the *absolute* method found it the simplest to understand and easiest for following the tracts because of its symmetry. In addition, the less chaotic color changes helped them recognize the orientations better. Those who disliked *absolute* thought that tracts looked too similar to differentiate, showing the tradeoffs between colormap *similarity* and *resolution*. Most participants were relatively neutral on the *Boy's surface*, considering it similar to the *eigenmaps* method in terms of

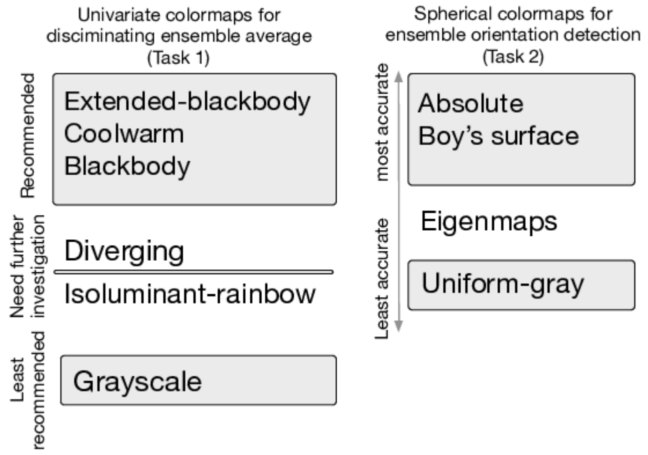


Fig. 11: Ranking of Colormaps for Ensemble Average and Orientation Tasks. Colormaps in the same gray boxes are in no particular order in our recommendations.

hue uses (spatial resolution). Participants commented that *"it (Boy's surface) was useful to have some different hues, but too many hues made the visualization less intuitive"*, while others stated that the *"right amount of hues of eigenmaps provided enough discriminations between values without overloading one's perception capability."*

6 Discussion

This section discusses our results. Fig. 11 shows our recommendations for choosing colormaps for the two ensemble tasks studied here.

6.1 isoluminant-rainbow Does Not Decrease the Mean Accuracy, but Introduces Biases

The most interesting result may be that the isoluminant-rainbow does *not* introduce greater error on average tasks (Fig. 7). This efficiency result may agree with those in vision science because humans can average hues [57]. However, none of the vision science studies to our knowledge drills down to the empirical study results to examine whether or not participants would be biased towards higher or lower than ground truth. The fact that isoluminant-rainbow introduces higher overshooting needs to be further studied, perhaps by explicitly controlling the variance in data for us to learn the colormap behaviors. Rainbow colors are known to be poor for univariate encoding due to the lack of uniformity and ordering and because they produce artificial boundaries in data. We could conclude from our study that human visual processing of a chunk of colors for quantitative discrimination tasks differs from that of single colors.

We do not recommend this isoluminant-rainbow map for ensemble average tasks. Instead, we propose to further explore *how* and *why* multihue works for *limited* capacity ensemble processing. This is mainly because the biases in isoluminant-rainbow are consistently independent of the variances in data (Figs. 5 and 9). The rainbow map certainly uses a set of semantically meaningful colors that would ease human understanding and our brain scientist collaborators particularly love rainbows; however, rainbow maps may still violate Trumbo's color design heuristics that *"the basic information should be displayed in a clear and logical fashion so that it may be decoded with precision and without continual references to the key (labeled scheme)"* and *"if small neighboring regions produce illusion of color over larger map areas, these illusions should not give misleading information"* [47].

6.2 Multihue Maps Improve Ensemble Accuracy

We ran a statistical analysis to examine whether or not hue or luminance affected error or task completion time. We found that hue had a significant main effect on time ($F(2, 1728) = 4.99$, $p = 0.0069$). The post-hoc analysis showed that multihue colormaps led to statistically significantly longer task completion time than single-hue (gray) colormaps.

The multihue extended-blackbody and the coolwarm colormaps had the lowest absolute error. This accuracy result of extended-blackbody agreed with 2D study results, though we did not observe significant differences. There may be at least two reasons for the benefits. First, one might think these two colormaps had the largest arc-length and thus yielded slightly better results than other maps. The other, perhaps primary reason for the benefits is that the multihue lets participants quickly determine the target-region first before formulating their answers. This conjecture may be supported by visual inspection of three cases of different FA distributions (Fig. 12): mid-average FA with large variance, high-average FA with mid-variance, and low-average with small variance. We may observe in all three cases that the many-hue colormaps in the last three columns help quickly locate the target regions on the colormap into which the answers fall.

6.3 That Many Colormaps Work Well Also Shows the Power of Human Visual Systems in Judging Ensemble Averages

We did not observe differences in accuracy among colormaps when measuring the distance of participants' answers from the ground truth. This result suggests the power of visual ensembles for quantitative estimates.

Since the application domain is in neuroscience, it would be reasonable to assume accuracy and error reduction are more important than time. Balancing all considerations of error, correctness, and bias in these colormaps, we rank them in the order shown in Fig. 11, where extended-blackbody, coolwarm, and blackbody seem to work well. Isoluminant-rainbow and diverging are worth further investigations. Grayscale is not recommended because of their higher absolute error. Though we cannot say whether the poor performance of grayscale was caused by its simultaneous contrast or its sole luminance channel, the result indeed is in agreement with the literature on 2D colorization.

6.4 Local Contrast and Resolution Together Might Be a Decisive Property for Ensemble Orientation

Our results present an uncanny valley effect where the highest and lowest resolution maps (Boy's surface and absolute) are better than the mid-resolution (eigenmaps).

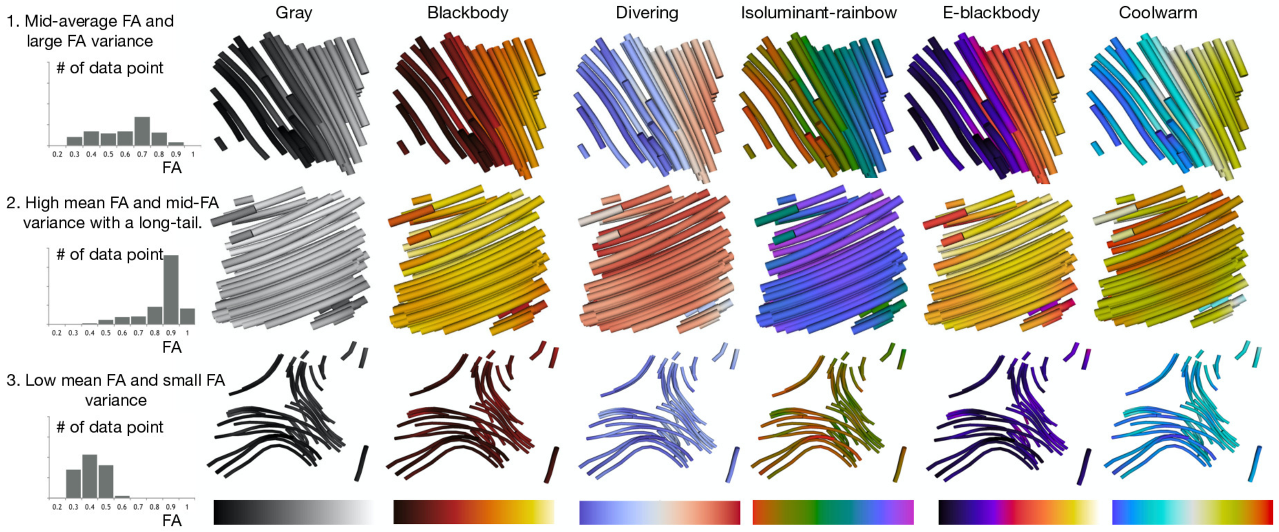


Fig. 12: Example Dataset Distribution and Their Colormaps: top: high-variance; middle: higher mean FA and narrow long-tail; bottom: low mean FA and narrow variance.

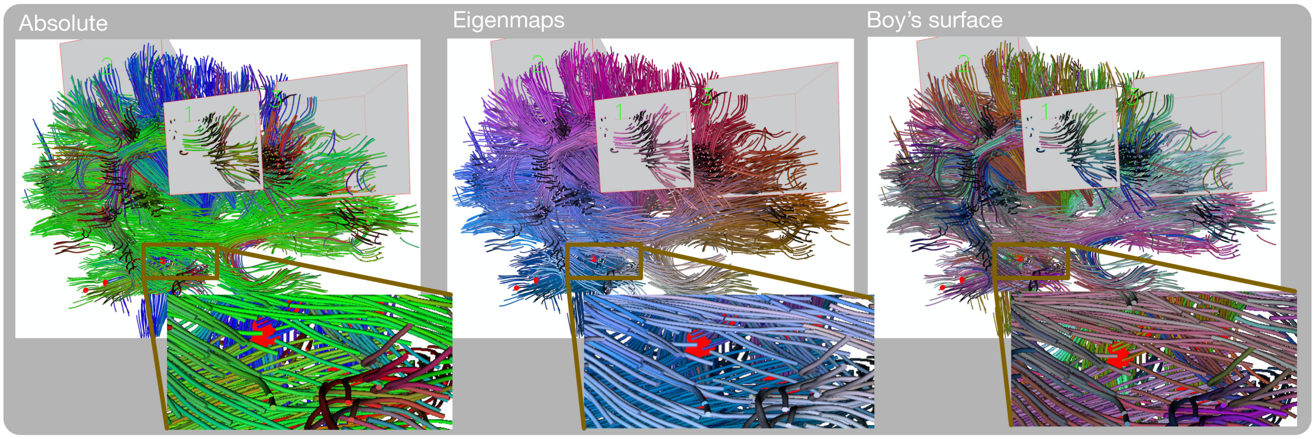


Fig. 13: An example from the empirical study in which all participants got correct answers using absolute and Boy's surface but only half of the participants got correct answers with eigenmaps. Red dots are sources. The correct answer is Box 2. Eigenmaps tend to show similar colors while the other two methods produce visually distinguishable ones. All three boxes looked pinkish using eigenmaps and are more visually distinguishable when using absolute and Boy's surface.

All colormaps with colors improve accuracy compared to the uniform-gray.

Overall, our results did not suggest that *resolution* contributes to higher accuracy in 3D space, since both *Boy's surface* and *absolute* reduced errors. The *eigenmaps* had reasonable resolution, as does *Boy's surface*, but lowered accuracy. We inspected specific examples to understand *when Boy's surface* and *absolute* succeeded and *eigenmaps* failed. Fig. 13 shows one of these examples for which participants achieved 100% accuracy using *Boy's surface* and *absolute* but 50% when using *eigenmaps*. We see that, while *eigenmaps* provided regional coloring, the adjacent regions had relatively lower contrast than other two approaches. This observation may suggest that the combination of local contrast and spatial resolution might be a decisive property.

Absolute colormap was commonly used in brain science and worked well in our study. Boy's surface generates colors that seem to strike the right balance in the spatial resolution and contrast for this spatial structure determination. Finally, the data sample varies so no dataset is seen twice by the same participants. For the eigenmaps, this setting means that the colors for the same tracts in different datasets would change, while the same tube would always be given the same color with the other maps.

We therefore recommend *absolute* as a good default; *Boy's surface* shows similar performance to *absolute* for coloring DMRI ensemble, as shown in Fig 11.

6.5 Reuse of Our Results to Ensemble Representations

We sought to further our understanding of the color ensembles to generate concrete implications for visual analysis of brain DMRI tractography datasets. In general, both tasks suggest that high-contrast localized colormaps may have

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2898438, IEEE Transactions on Visualization and Computer Graphics

helped both ensemble average and orientation tasks. Reuse of our results in other domains would have to take into account domain specificities of data, task (e.g., [70]), and user. Several areas could benefit from our work, such as weather forecasting [29], hurricane track prediction [71], and motion or movement trajectories [72] [73], because direct trajectory depiction has been informative. The most suitable reuse would be when the datasets have relatively low variance, so that colormaps can be localized to a smaller region scalar data visualizations. Similarly, the spherical colormap for line field visualizations might also be domain-dependent. In our case, the tracts are following three major orientations. We also did not consider other tract shapes, when appropriate distance measures were needed for maximal performance.

6.6 Participants' Experiences

Participants in this study have different backgrounds, and an ideal condition might be to use only brain scientists, clinicians, or medical school students. One major reason for the background differences was that we had access to only a few brain scientists. We used as many as possible in the study because we wanted to collect their comments related to the brain science domain.

Also, we followed Munzner's approach [58] of abstracting tasks into a level suitable for empirical study. In other words, these tasks could be performed by a trained participant. This may explain why we did not observe differences in task completion time and accuracy between students with and without medical backgrounds. Several user studies in flow visualization have used non-domain experts, suggesting that non-domain-expert is a viable option in empirical studies [74].

6.7 Using Ensemble for Visualization Design

It is intuitive to think that hue, due to its categorical effect (e.g. yellow or red), would interfere with the ensemble coloring, thus making representing a multihue average difficult. However, this turns out not to be the case. In vision science, ensemble is believed to be used by the human visual system to address our severely limited visual working memory. We can quickly derive patterns that guide our attention towards the most useful information. Scientific data is often highly structured and may carry redundant structures. When there is redundancy, it is possible to sample and filter to produce optimal views. For example, a handful of past visualization work has shown that implicit or explicit representation of sets of objects as groups or ensembles can guide observers' attention to process only the most relevant incoming information (e.g., explicit depiction of a group of objects in clusters [75], grouping interfaces to augment exploration workflows [76] [77] or using spatial patterns to form texture pattern to guide observers' behavior [40]). We believe there will be an opportunity to create a compressed and efficient ensemble representation of information, such as ensemble overviews, to guide visual attention to the areas more relevant to the targets. Our design guidelines can perhaps be used in devising algorithmic colormaps, similar to that of [78].

6.8 Limitations and Future Work

Our study is only a first step towards understanding ensemble tasks in visualizations. Although this study can suggest *what* colormap to choose for ensemble representation, we may need to build computational models or isolate factors (e.g., hue and luminance for ensemble average and resolution and uniqueness for ensemble orientation) to explain *how* these colormaps are used by our visual system. Effectiveness of these coloring approaches needs to be studied further for other discrimination and detection tasks.

Our study would suggest further work. Since multihue colormaps in general improved ensemble average accuracy, one could run studies to systematically control the mean and variance of the ensemble datasets to model the ensemble performance. Viewers make a two-alternative forced-choice judgment about which visualization method contains the larger average value. Sensitivities are measured based on the differences between the values. A psychometric function fitted to the data reveals sensitivity to the discriminative threshold to measure accuracy. Using this method, we could answer questions about *why* and *when* multihue average will be effective and how variance influences the effectiveness and efficiency.

7 Conclusion

This study is the first (to our knowledge) to define and compare color ensembles for 3D DMRI tractography visualizations. Results from the study provide the following insights for choosing 3D tube ensemble coloring.

- The most interesting result was that the isoluminant-rainbow performed reasonably well, though it did lead to more reporting bias towards higher-than-ground-truth values than other colormaps.
- Extended-blackbody, coolwarm, and blackbody are reasonably accurate for ensemble average in 3D. Our analysis showed that hue had much larger influence on error than luminance.
- Our study on the ensemble orientation discrimination supports the proposition that having some colors is better than no color at all.
- Colormaps with better local contrast and resolution together (e.g., *Boy's surface* and *absolute*) are most desirable for orientation discrimination tasks such as ensemble tract tracing.

Appendix A
The Univariate Colormaps in the L*A*B* Colorspace

Fig. 14 shows the scalar colormaps in the L*A*B* color space. The curve in each figure shows the trajectory of colormaps and their three projections in the color space. All color interpolation is performed using linear interpolation in this space.

We used the Rogowitz-Kalvin [79] and Kindlmann-Reinhard-Creem approaches [42] to help visually inspect colormaps to test their luminance profile. This method utilizes our sensitivity to luminance variations in human faces

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2898438, IEEE Transactions on Visualization and Computer Graphics

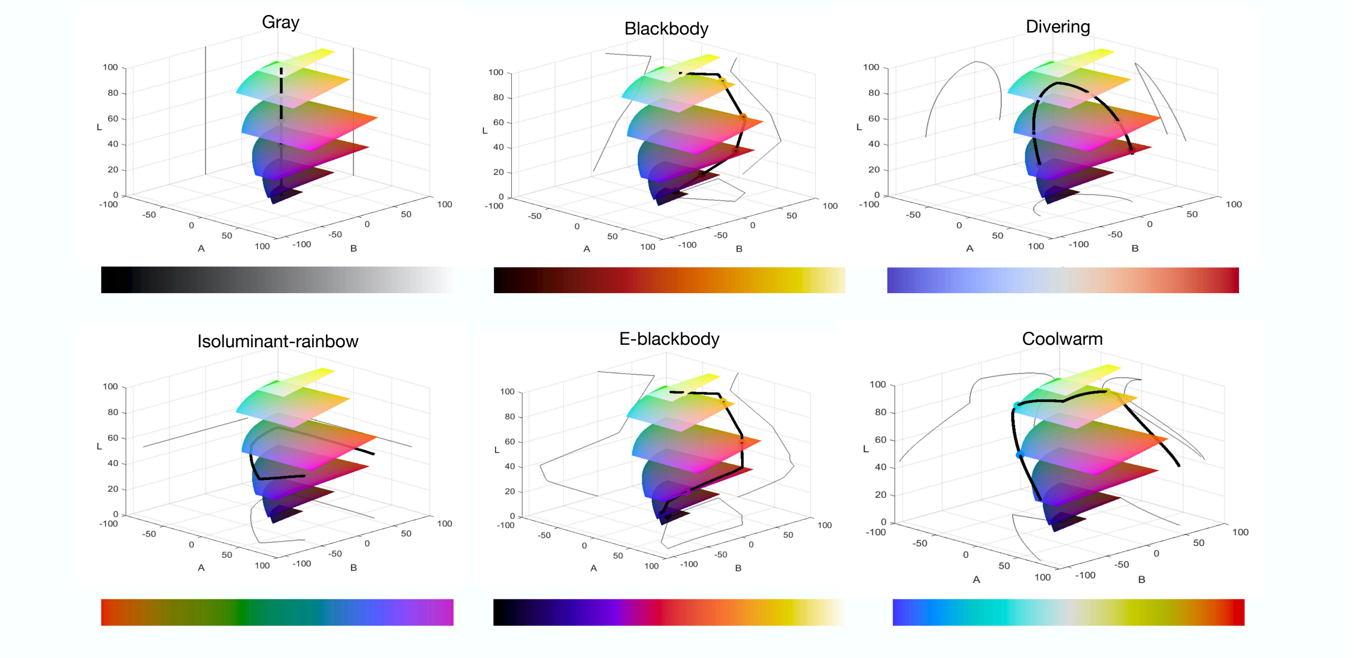


Fig. 14: Colormap Profile for Showing Scalars in Task 1 (AverageFA tasks) in the L*A*B* Color Space. L-planes from bottom to top are L=5, 20, 40, 60, 80, and 95.

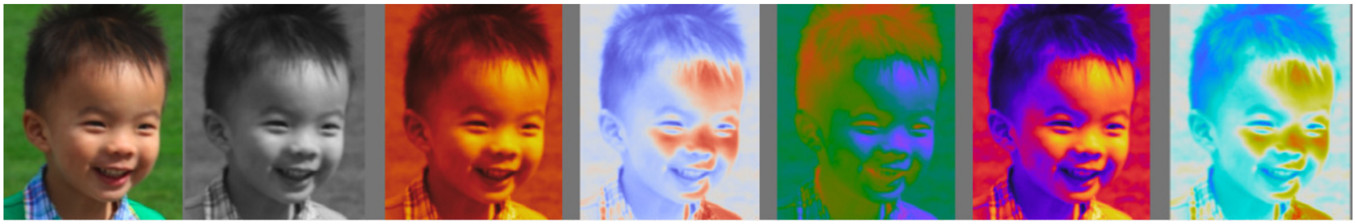


Fig. 15: Using Faces to Examine the Luminance Profile of Colormaps (from left to right): original image, gray, blackbody, diverging, isoluminant-rainbow, extended-blackbody, and coolwarm colormaps.

to select colormaps. Fig. 15 shows samples of faces generated by these six colormaps with our online tool. The faces with isoluminant-rainbow and diverging colormaps are less recognizable than all others. The rainbow and coolwarm colormaps help distinguish different values: one can clearly see red (high) values around the nose and under the eyes.

Appendix B
Pilot Studies

A set of preparatory pilot studies was designed to test the performance characteristics and capabilities of study designs, measures, and procedures under consideration for use in this and subsequent studies. These pilot studies help identify relevant factors that could create barriers to subsequent study completion. Each study involves 12-24 participants depending on the experimental setting.

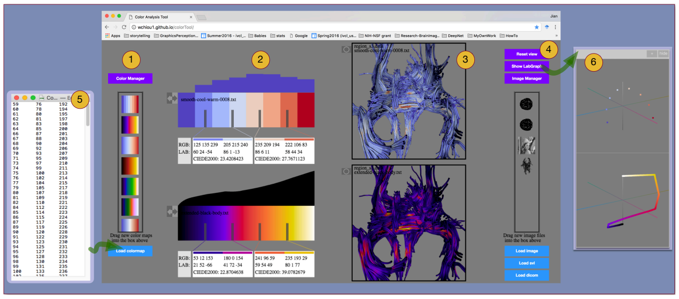
The first study compared multiple visual marks of color, size, and texture to measure what visual variables would be suitable for ensemble representations, as estimated by their potential effectiveness relative to 2D visualization. We observed the benefits of coloring over size and texture in ensemble accuracy in that study. The second study compared several colormaps generated using the algorithmic approach in Wijffelaars et al. [78]. A distinct observation in both studies was that multihue colormaps did not downgrade ensemble average accuracy. These results differ from our current knowledge of multihue rainbow colormap for continuous quantitative data visualization. The third pilot study used the same conditions as those reported here so that we could refine the procedure.

Our pilot study on ensemble orientation studied colormaps of absolute, hue-ball, similarity, and Boy's surfaces. In general, we observed that the similarity method did not improve task performance, and we subsequently removed this method from our study. Hue-ball, absolute, and the Boy's surface achieved similar accuracy and were among the best. After considering these design choices in conjunction with vision science literature, we decided to use eigenmaps, gray, absolute, and Boy's surface in our study to have a meaningful range of variation of color attributes in the current study.

Appendix C
Coloring Tool Website

Our own tool (Fig. 16) is hosted at http://wchiou1.github.io/colorTool/(Fig. 16). During

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2898438, IEEE Transactions on Visualization and Computer Graphics



```
59    76   192
60    78   194
61    80   195
62    81   197
63    83   198
64    85   200
66    87   201
67    88   203
68    90   204
69    92   206
70    93   207
71    95   209
73    97   210
74    99   211
75   100   213
76   102   214
77   104   215
79   105   217
80   107   218
81   109   219
82   110   221
84   112   222
85   114   223
86   115   224
87   117   225
89   119   226
90   120   228
91   122   229
93   123   230
94   125   231
95   127   232
96   128   233
98   130   234
99   131   235
100  133   236
```

RGB: 125 135 239 205 215 240
LAB: 60 24 -54 86 1 -13
CIEDE2000: 23.4208423

RGB: 235 209 194 222 106 83
LAB: 86 6 11 58 44 34
CIEDE2000: 27.7671123

RGB: 53 12 153 180 0 154
LAB: 21 52 -66 41 72 -34
CIEDE2000: 22.8704638

RGB: 241 96 59 235 193 29
LAB: 59 54 49 80 1 77
CIEDE2000: 39.0782679

Fig. 16: Exploratory Color Comparison Tool.

the evaluation process, we found that using a coloring tool to quickly provide side-by-side comparison made our discussion with the brain scientists very effective and efficient. The direct manipulation interface lets users directly drag and drop plain-text colormaps. It can display both 2D image and 3D geometry examples.

Acknowledgments

The authors would like to thank Drs. Peter Kochunov, L. Elliot Hong, and Neda Jahanshad for their discussions on color uses in brain science, Dr. Bernice Rogowitz for a discussion on colormap uses in real-world applications, Dr. Jeremy Wolfe for his thorough review and comments on this manuscript, and the anonymous reviewers for their constructive comments. The authors also thank the participants at University of Maryland, Baltimore Country, University of Mississippi Medical School, University of Maryland Medical School, and Veterans Affairs Medical Center of Providence, RI for their time and effort. We thank Katrina Avery for her editorial support. This work was supported in part by NSF IIS-1302755, CNS-1531491, DBI-1260795, IIS-1018769, and DUE-0817106, and NIST MSE-70NANB13H181. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of National Institute of Standards and Technology (NIST) or the National Science Foundation (NSF.)

Jian Chen is the corresponding author.

References

[1] M. N. Phadke, L. Pinto, F. Alabi, J. Harter, R. M. Taylor II, X. Wu, H. Petersen, S. A. Bass, and C. G. Healey, "Exploring ensemble visualization," in *Proceedings of SPIE*, vol. 8294, no. 82940B (12 pages), 2012.

[2] A. Y. Leib, A. Kosovicheva, and D. Whitney, "Fast ensemble representations for abstract visual impressions," *Nature Communications*, vol. 7 (article number 13186), 2016.

[3] A. Chetverikov, G. Campana, and Á. Kristjánsson, "Representing color ensembles," *Psychological Science*, vol. 28, no. 10, pp. 1510–1517, 2017.

[4] D. Ariely, "Seeing sets: Representation by statistical properties," *Psychological Science*, vol. 12, no. 2, pp. 157–162, 2001.

[5] N. Robitaille and I. M. Harris, "When more is less: Extraction of summary statistics benefits from larger sets," *Journal of Vision*, vol. 11, no. 12, pp. 1–8, 2011.

[6] G. A. Alvarez and A. Oliva, "The representation of simple ensemble visual features outside the focus of attention," *Psychological Science*, vol. 19, no. 4, pp. 392–398, 2008.

[7] D. W. Williams and R. Sekuler, "Coherent global motion percepts from stochastic local motions," *Vision Research*, vol. 24, no. 1, pp. 55–62, 1984.

[8] S. N. Watamaniuk and A. Duchon, "The human visual system averages speed information," *Vision Research*, vol. 32, no. 5, pp. 931–941, 1992.

[9] D. Burr and J. Ross, "A visual sense of number," *Current Biology*, vol. 18, no. 6, pp. 425–428, 2008.

[10] M. F. Neumann, S. R. Schweinberger, and A. M. Burton, "Viewers extract mean and individual identity from sets of famous faces," *Cognition*, vol. 128, no. 1, pp. 56–63, 2013.

[11] A. Oliva and A. Torralba, "Building the GIST of a scene: The role of global image features in recognition," *Progress in Brain Research*, vol. 155, pp. 23–36, 2006.

[12] B. Bauer, "Does Stevens's power law for brightness extend to perceptual brightness averaging?" *The Psychological Record*, vol. 59, no. 2, p. 171, 2009.

[13] T. C. Chua, W. Wen, M. J. Slavin, and P. S. Sachdev, "Diffusion tensor imaging in mild cognitive impairment and Alzheimer's disease: a review," *Current Opinion in Neurology*, vol. 21, no. 1, pp. 83–92, 2008.

[14] L. Zhou and C. D. Hansen, "A survey of colormaps in visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 8, pp. 2051–2069, 2016.

[15] S. Silva, B. S. Santos, and J. Madeira, "Using color in visualization: A survey," *Computers & Graphics*, vol. 35, no. 2, pp. 320–333, 2011.

[16] G. D. Finlayson, S. D. Hordley, and P. M. Hubel, "Color by correlation: A simple, unifying framework for color constancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1209–1221, 2001.

[17] M. H. Kim, T. Weyrich, and J. Kautz, "Modeling human color perception under extended luminance levels," in *ACM SIGGRAPH*, vol. 28, no. 3 (10 pages), 2009.

[18] K. Moreland, "Why we use bad color maps and what you can do about it," *Electronic Imaging*, vol. 2016, no. 16, pp. 1–6, 2016.

[19] M. Christen, D. A. Vitacco, L. Huber, J. Harboe, S. I. Fabrikant, and P. Brugger, "Colorful brains: 14 years of display practice in functional neuroimaging," *NeuroImage*, vol. 73, pp. 30–39, 2013.

[20] D. A. Szafir, S. Haroz, M. Gleicher, and S. Franconeri, "Four types of ensemble coding in data visualizations," *Journal of Vision*, vol. 16, no. 5, pp. 1–19, 2016.

[21] S. Pajevic, C. Pierpaoli *et al.*, "Color schemes to represent the orientation of anisotropic tissues from diffusion tensor data: application to white matter fiber tract mapping in the human brain," *Magnetic Resonance in Medicine*, vol. 42, no. 3, pp. 526–540, 1999.

[22] C. Demiralp, J. F. Hughes, and D. H. Laidlaw, "Coloring 3D line fields using Boy's real projective plane immersion," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 6, pp. 1457–1464, 2009.

[23] A. Brun, H.-J. Park, H. Knutsson, and C.-F. Westin, "Coloring of DT-MRI fiber traces using Laplacian eigenmaps," in *International Conference on Computer Aided Systems Theory*, 2003, pp. 518–529.

[24] K. Potter, A. Wilson, P.-T. Bremer, D. Williams, C. Doutriaux, V. Pascucci, and C. R. Johnson, "Ensemble-vis: A framework for the statistical visualization of ensemble data," in *IEEE International Conference on Data Mining Workshops*, 2009, pp. 233–240.

[25] J. Wang, S. Hazarika, C. Li, and H.-W. Shen, "Visualization and visual analysis of ensemble data: A survey," *IEEE Transactions on Visualization and Computer Graphics*, 2018.

[26] D. Whitney and A. Y. Leib, "Ensemble perception," *Annual Review of Psychology*, vol. 69, no. 12, pp. 1–25, 2017.

[27] G. A. Alvarez, "Representing multiple objects as an ensemble enhances visual cognition," *Trends in Cognitive Sciences*, vol. 15, no. 3, pp. 122–131, 2011.

[28] K. Potter, P. Rosen, and C. Johnson, "From quantification to visualization: A taxonomy of uncertainty visualization approaches," *Uncertainty Quantification in Scientific Computing*, pp. 226–249, 2012.

[29] J. Sanyal, S. Zhang, J. Dyer, A. Mercer, P. Amburn, and R. Moorhead, "Noodles: A tool for visualization of numerical weather model ensemble uncertainty," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 1421–1430, 2010.

[30] R. S. Laramee, H. Hauser, H. Doleisch, B. Vrolijk, F. H. Post, and D. Weiskopf, "The state of the art in flow visualization: Dense and texture-based techniques," in *Computer Graphics Forum*, vol. 23, no. 2, 2004, pp. 203–221.

[31] N. Max, P. Hanrahan, and R. Crawfis, "Area and volume coherence for efficient visualization of 3d scalar functions," *Proceedings of the Workshop on Volume Visualization*, pp. 27–33, 1990.

[32] A. Forsberg, J. Chen, and D. H. Laidlaw, "Comparing 3D vector field visualization methods: A user study," *IEEE Transactions on*

Visualization and Computer Graphics, vol. 15, no. 6, pp. 1219–1226, 2009.

[33] J. Maule and A. Franklin, "Effects of ensemble complexity and perceptual similarity on rapid averaging of hue," *Journal of Vision*, vol. 15, no. 4, pp. 1–18, 2015.

[34] J. Haberman, T. F. Brady, and G. A. Alvarez, "Individual differences in ensemble perception reveal multiple, independent levels of ensemble representation," *Journal of Experimental Psychology: General*, vol. 144, no. 2, pp. 432–446, 2015.

[35] J. Haberman and D. Whitney, "Ensemble perception: Summarizing the scene and broadening the limits of visual processing," *From Perception to Consciousness: Searching with Anne Treisman*, pp. 339–349, 2012.

[36] M. Correll, D. Albers, S. Franconeri, and M. Gleicher, "Comparing averages in time series data," in *Proceedings of ACM SIGCHI*, 2012, pp. 1095–1104.

[37] W. S. Cleveland and R. McGill, "Graphical perception: Theory, experimentation, and application to the development of graphical methods," *Journal of the American Statistical Association*, vol. 79, no. 387, pp. 531–554, 1984.

[38] J. Webster, P. Kay, and M. A. Webster, "Perceiving the average hue of color arrays," *Journal of the Optical Society of America. A, Optics, Image Science, and Vision*, vol. 31, no. 4, pp. A283–A292, 2014.

[39] O. Wright, C. Biggam, C. Hough, C. Kay, and D. Simmons, "Effects of stimulus range on color categorization," *New Directions in Colour Studies*, pp. 265–276, 2011.

[40] H. Zhao and J. Chen, "Bivariate separable-dimension glyphs can improve visual analysis of holistic features," *arXiv: https://arxiv.org/abs/1712.02333v1*, 2017.

[41] G. Sharma, W. Wu, and E. N. Dalal, "The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations," *Color Research & Application*, vol. 30, no. 1, pp. 21–30, 2005.

[42] G. Kindlmann, E. Reinhard, and S. Creem, "Face-based luminance matching for perceptual colormap generation," in *Proceedings of the Conference on Visualization*, 2002, pp. 299–306.

[43] G. Hu, Z. Pan, M. Zhang, D. Chen, W. Yang, and J. Chen, "An interactive method for generating harmonious color schemes," *Color Research & Application*, vol. 39, no. 1, pp. 70–78, 2014.

[44] C. C. Gramazio, D. H. Laidlaw, and K. B. Schloss, "Colorgorical: Creating discriminable and preferable color palettes for information visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 521–530, 2017.

[45] R. Bujack, T. L. Turton, F. Samsel, C. Ware, D. H. Rogers, and J. Ahrens, "The good, the bad, and the ugly: A theoretical framework for the assessment of continuous colormaps," *IEEE Transactions on Visualization and Computer Graphics*, vol. 1, no. 1, pp. 923 – 933, 2018.

[46] D. A. Szafir, "Modeling color difference for visualization design," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 392–401, 2018.

[47] B. E. Trumbo, "A theory for coloring bivariate statistical maps," *The American Statistician*, vol. 35, no. 4, pp. 220–226, 1981.

[48] C. Ware, T. L. Turton, F. Samsel, R. Bujack, D. H. Rogers, K. Lawonn, N. Smit, and D. Cunningham, "Evaluating the perceptual uniformity of color sequences for feature discrimination," in *EuroVis Workshop on Reproducibility, Verification, and Validation in Visualization. The Eurographics Association*, 2017.

[49] D. Penney, J. Chen, and D. H. Laidlaw, "Effects of illumination, texture, and motion on task performance in 3d tensor-field streamtube visualizations," in *IEEE Pacific Visualization Symposium*, 2012, pp. 97–104.

[50] H. Zhao, G. W. Bryant, W. Griffin, J. E. Terrill, and J. Chen, "Validation of splitvectors encoding for quantitative visualization of large-magnitude-range vector fields," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 6, pp. 1691–1705, 2017.

[51] F. Ritter, C. Hansen, V. Dicken, O. Konrad, B. Preim, and H.-O. Peitgen, "Real-time illustration of vascular structures," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, pp. 877–884, 2006.

[52] P. Svetachov, M. H. Everts, and T. Isenberg, "DTI in context: illustrating brain fiber tracts in situ," in *Computer Graphics Forum*, vol. 29, no. 3, 2010, pp. 1023–1032.

[53] D. Acevedo and D. Laidlaw, "Subjective quantification of perceptual interactions among some 2D scientific visualization methods," *IEEE Transactions on Visualization and Computer Graphics*, vol. 12, no. 5, pp. 1133–1140, 2006.

[54] M. Borkin, K. Gajos, A. Peters, D. Mitsouras, S. Melchionna, F. Rybicki, C. Feldman, and H. Pfister, "Evaluation of artery visualizations for heart disease diagnosis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2479–2488, 2011.

[55] J. Chen, H. Cai, A. P. Auchus, and D. H. Laidlaw, "Effects of stereo and screen size on the legibility of three-dimensional streamtube visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2130–2139, 2012.

[56] G. Kindlmann and D. Weinstein, "Hue-balls and lit-tensors for direct volume rendering of diffusion tensor fields," in *Proceedings of the Conference on Visualization*, 1999, pp. 183–189.

[57] J. Maule, C. Witzel, and A. Franklin, "Getting the gist of multiple hues: metric and categorical effects on ensemble perception of hue," *Journal of the Optical Society of America. A, Optics, Image Science, and Vision*, vol. 31, no. 4, pp. A93–102, 2014.

[58] T. Munzner, "A nested model for visualization design and validation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 6, pp. 921–928, 2009.

[59] S. Mori and J. Zhang, "Principles of diffusion tensor imaging and its applications to basic neuroscience research," *Neuron*, vol. 51, no. 5, pp. 527–539, 2006.

[60] C. Zhang, M. Caan, T. Höllt, E. Eisemann, and A. Vilanova, "Overview+detail visualization for ensembles of diffusion tensors," in *Computer Graphics Forum*, vol. 36, no. 3, 2017, pp. 121–132.

[61] T. Isenberg, P. Isenberg, J. Chen, M. Sedlmair, and T. Möller, "A systematic review on the practice of evaluating visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2818–2827, 2013.

[62] B. Preim, A. Baer, D. Cunningham, T. Isenberg, and T. Ropinski, "A survey of perceptually motivated 3D visualization of medical image data," in *Computer Graphics Forum*, vol. 35, no. 3, 2016, pp. 501–525.

[63] P. Kochunov, H. Ganjgahi, A. Winkler, S. Kelly, D. K. Shukla, X. Du, N. Jahanshad, L. Rowland, H. Sampath, B. Patel, P. O'Donnell, Z. Xie, S. A. Paciga, C. R. Schubert, J. Chen, G. Zhang, P. M. Thompson, T. E. Nichols, and H. L. Elliot, "Heterochronicity of white matter development and aging explains regional patient control differences in schizophrenia," *Human Brain Mapping*, vol. 37, no. 12, pp. 4673–4688, 2016.

[64] S. Zhang, C. Demiralp, and D. H. Laidlaw, "Visualizing diffusion tensor MR images using streamtubes and streamsurfaces," *IEEE Transactions on Visualization and Computer Graphics*, vol. 9, no. 4, pp. 454–462, 2003.

[65] H. Jiang, P. C. van Zijl, J. Kim, G. D. Pearlson, and S. Mori, "DtiStudio: resource program for diffusion tensor computation and fiber bundle tracking," *Computer Methods and Programs in Biomedicine*, vol. 81, no. 2, pp. 106–116, 2006.

[66] S. Pieper, M. Halle, and R. Kikinis, "3D slicer," in *IEEE International Symposium on Biomedical Imaging: Nano to Macro*, 2004, pp. 632–635.

[67] R. Borgo, J. Dearden, and M. W. Jones, "Order of magnitude markers: An empirical study on large magnitude number detection," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 2261–2270, 2014.

[68] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[69] J. Cohen, *Statistical power analysis for the behavioral sciences*. New York: Academic Press, 1988.

[70] C. Tominski, G. Fuchs, and H. Schumann, "Task-driven color coding," in *Proceedings of International Conference on Information Visualisation*, 2008, pp. 373–380.

[71] J. Cox and M. Lindell, "Visualizing uncertainty in predicted hurricane tracks," *International Journal for Uncertainty Quantification*, vol. 3, no. 2, pp. 143–156, 2013.

[72] J. Chen, M. Kostandov, I. Pivkin, D. Riskin, D. Willis, S. Swartz, and D. Laidlaw, "Visual analysis of dimensionality reduction for exploring bat flight kinematics in a virtual environment," in *Proceedings of the 15th Joint Eurographics Conference on Virtual Environments*, 2009, pp. 77–84.

[73] N. Andrienko, G. Andrienko, and S. Rinzivillo, "Leveraging spatial abstraction in traffic analysis and forecasting with visual analytics," *Information Systems*, vol. 57, pp. 172–194, 2016.

[74] Z. Liu, S. Cai, J. E. Swan, R. J. Moorhead, J. P. Martin, and T. Jankun-Kelly, "A 2D flow visualization user study using explicit flow synthesis and implicit task design," *IEEE Transactions on*

Visualization and Computer Graphics, vol. 18, no. 5, pp. 783–796, 2012.

[75] Z. Peng, E. Grundy, R. S. Laramee, G. Chen, and N. Croft, "Mesh-driven vector field clustering and visualization: An image-based approach," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 2, pp. 283–298, 2012.

[76] G. Li, A. C. Bragdon, Z. Pan, M. Zhang, S. M. Swartz, D. H. Laidlaw, C. Zhang, H. Liu, and J. Chen, "VisBubbles: a workflow-driven framework for scientific data analysis of time-varying biological datasets," in *ACM SIGGRAPH Asia Posters*, 2011, p. 27.

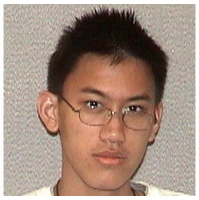
[77] E. D. Ragan, A. Endert, J. Sanyal, and J. Chen, "Characterizing provenance in visualization and data analysis: an organizational framework of provenance types and purposes," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 31–40, 2016.

[78] M. Wijffelaars, R. Vliegen, J. J. Van Wijk, and E.-J. Van Der Linden, "Generating color palettes using intuitive parameters," in *Computer Graphics Forum*, vol. 27, no. 3, 2008, pp. 743–750.

[79] B. E. Rogowitz and A. D. Kalvin, "The "which Blair project": A quick visual method for evaluating perceptual color maps," in *Proceedings of the Conference on Visualization*, 2001, pp. 183–190.

Jian Chen received the PhD degree in Computer Science from Virginia Polytechnic Institute and State University (Virginia Tech). She did her post-doctoral work in the Department of Computer Science at Brown University. She is an Associate Professor in Computer Science and Electrical Engineering at The Ohio State University where she directs the Interactive Visual Computing Laboratory (IVCL). Her research interests include design and evaluation of visualization techniques and virtual reality. She is a member of the IEEE and the IEEE Computer Society.

Guohao Zhang is a PhD student in the Department of Computer Science and Electrical Engineering at University of Maryland, Baltimore County. He received his B.E. degree in Engineering Physics from Tsinghua University in 2012. His research interests include design and evaluation of visualization techniques and 3D visualizations. He is a student member of IEEE.



Wesley Chiou is an undergraduate student in the Department of Computer Science and Electrical Engineering at University of Maryland, Baltimore County. His research interest is human-computer interaction and visualization.

David H. Laidlaw received the PhD degree in computer science from the California Institute of Technology, where he also did post-doctoral work in the Division of Biology. He is a professor in the Computer Science Department at Brown University. His research centers on applications of visualization, modeling, computer graphics, and computer science to other scientific disciplines. He is a fellow of IEEE.

Alexander P. Auchus Dr. Alexander P. Auchus holds degrees from Johns Hopkins University and from Washington University in St. Louis. He is an elected fellow of the American Neurological Association, the American Academy of Neurology, and the American Geriatrics Society. He has served on the faculty of Emory University, Case Western Reserve University, and University of Tennessee. His present position is Professor and McCarty Chair of Neurology at the University of Mississippi Medical Center. Dr. Auchus's research interests are in neuroimaging biomarkers for Alzheimer's disease and other dementias.