The Annals of Applied Statistics
2021, Vol. 15, No. 2, 638–661
https://doi.org/10.1214/20-AOAS1430

A STATISTICAL PIPELINE FOR IDENTIFYING PHYSICAL FEATURES THAT DIFFERENTIATE CLASSES OF 3D SHAPES

BY BRUCE WANG[1], TIMOTHY SUDIJONO[2], HENRY KIRVESLAHTI[3], TINGRAN GAO[4], DOUGLAS M. BOYER[5], SAYAN MUKHERJEE[6] AND LORIN CRAWFORD[7]

[1]*Lewis-Sigler Institute for Integrative Genomics, Princeton University, bruce.wang@princeton.edu*
[2]*Division of Applied Mathematics, Brown University, timothy_sudijono@brown.edu*
[3]*Department of Statistical Science, Duke University, henry.kirveslahti@duke.edu*
[4]*Committee on Computational and Applied Mathematics, Department of Statistics, University of Chicago, gaotingran@gmail.com*
[5]*Department of Evolutionary Anthropology, Duke University, doug.boyer@duke.edu*
[6]*Department of Statistical Science, Department of Computer Science, Department of Mathematics, and Department of Bioinformatics & Biostatistics, Duke University, sayan@stat.duke.edu*
[7]*Microsoft Research New England, lcrawford@microsoft.com*

The recent curation of large-scale databases with 3D surface scans of shapes has motivated the development of tools that better detect global patterns in morphological variation. Studies, which focus on identifying differences between shapes, have been limited to simple pairwise comparisons and rely on prespecified landmarks (that are often known). We present SINATRA, the first statistical pipeline for analyzing collections of shapes without requiring any correspondences. Our novel algorithm takes in two classes of shapes and highlights the physical features that best describe the variation between them. We use a rigorous simulation framework to assess our approach. Lastly, as a case study we use SINATRA to analyze mandibular molars from four different suborders of primates and demonstrate its ability recover known morphometric variation across phylogenies.

1. Introduction. Subimage analysis is an important open problem in both medical imaging studies and geometric morphometric applications. The problem asks which physical features of shapes are most important for differentiating between two classes of 3D images or shapes, such as computed tomography (CT) scans of bones or magnetic resonance images (MRI) of different tissues. More generally, the subimage analysis problem can be framed as a regression-based task: given a collection of shapes, find the properties that explain the greatest variation in some response variable (continuous or binary). One example is identifying the structures of glioblastoma tumors that best indicate signs of potential relapse and other clinical outcomes (Crawford et al. (2020)). From a statistical perspective the subimage selection problem is directly related to the variable selection problem, given high-dimensional covariates and a univariate outcome, we want to infer which variables are most relevant in explaining or predicting variation in the observed response.

Framing subimage analysis as a regression presents several challenges. The first challenge centers around representing a 3D object as a (square integrable) covariate or feature vector. The transformation should lose a minimum amount of geometric information and apply to a wide range of shape and imaging datasets. In this paper we use a tool from integral geometry and differential topology, called the Euler characteristic (EC) transform (Crawford et al. (2020), Curry, Mukherjee and Turner (2019), Ghrist, Levanger and Mai (2018), Turner,

Received October 2019; revised October 2020.
Key words and phrases. 3D image analysis, centrality measures, topological data analysis, Gaussian processes, evolutionary biology, phylogenetics.

Mukherjee and Boyer (2014)), which maps shapes into vectors without requiring prespecified landmark points or pairwise correspondences. This property is central to our innovations.

After finding a vector representation of the shape, our second challenge is quantifying which topological features are most relevant in explaining variation in a continuous outcome or binary class label. We address this classic take on variable selection by using a Bayesian regression model and an information theoretic metric to measure the relevance of each topological feature. Our Bayesian method allows us to perform variable selection for nonlinear functions; we discuss the importance of this requirement in Method Overview and Results.

The last challenge deals with how to interpret the most informative topological features obtained by our variable selection methodology. The EC transform is invertible; thus, we can take the most informative topological features and naturally recover the most imformative physical regions on the shape. In this paper we introduce SINATRA, a unified statistical pipeline for subimage analysis that addresses each of these challenges and is the first subimage analysis method that does not require landmarks or correspondences.

Classically there have been three approaches to modeling random 3D images and shapes: (i) landmark-based representations (Cates, Elhabian and Whitaker (2017), Kendall (1989)), (ii) diffeomorphism-based representations (Dupuis, Grenander and Miller (1998)) and (iii) representations that use integral geometry and excursions of random fields (Worsley (1995)). Landmark-based analysis uses points on shapes that are known to correspond with each other. As a result, any shape can be represented as a collection of 3D coordinates. Landmark-based approaches have two major shortcomings. First, many modern datasets are not defined by landmarks; instead, they consist of 3D CT scans (Boyer et al. (2016), Goswami (2015)). Second, reducing these detailed mesh data to simple landmarks often results in a great deal of information loss.

Diffeomorphism-based approaches have bypassed the need for landmarks. Many tools have been developed that efficiently compare the similarity between shapes in large databases via algorithms that continuously deform one shape into another (Boyer et al. (2011, 2015), Gao, Kovalsky and Daubechies (2019), Gao et al. (2019), Hong, Golland and Zhang (2017), Ovsjanikov et al. (2012)). Unfortunately, these methods require diffeomorphisms between shapes: the map from shape A to shape B must be differentiable, as must the inverse of the map. Such functions are often called "correspondence maps," since they take two shapes and place them in correspondence. There are many applications with no such transformations because of qualitative differences. For example, in a dataset of fruit fly wings some mutants may have extra lobes of veins (Miller (2015)), or, in a dataset of brain arteries many of the arteries cannot be continuously mapped to each other (Bendich et al. (2016)). Indeed, in large databases, such as the MorphoSource (Boyer et al. (2016)), the CT scans of skulls across many clades are not diffeomorphic. Recent algorithms have attempted to overcome this issue by constructing more general "functional" correspondences (Huang et al. (2019), Rustamov et al. (2013)) which can be established even across shapes having different topology. However, there is a real need for 3D image analysis methods that do not require correspondences.

Previous work (Turner, Mukherjee and Boyer (2014)) introduced two topological transformations for shapes, the persistent homology (PH) transform and the EC transform. These tools from integral geometry first allowed for pairwise comparisons between shapes or images without requiring correspondence or landmarks. Since then, mathematical foundations of the two transforms and their relationship to the theory of sheaves and fiber bundles have been established (Curry, Mukherjee and Turner (2019), Ghrist, Levanger and Mai (2018)). Detailed mathematical analyses have also been provided (Curry, Mukherjee and Turner (2019)). A nonlinear regression framework, which uses the EC transform to predict outcomes of disease free survival in glioblastoma (Crawford et al. (2020)), is most relevant to this paper. This works shows that the EC transform reduces the problem of regression with shape covariates into a problem in functional data analysis (FDA) and that nonlinear regression models

are more accurate than linear models when predicting complex phenotypes and traits. The SINATRA pipeline further enhances the relation between FDA and topological transforms by enabling variable selection with shapes as covariates.

Beyond the pipeline, this paper includes software packaging to implement our approach and a detailed design of rigorous simulation studies which can assess the accuracy of subimage selection methods. The freely available software comes with several built-in capabilities that are integral to subimage analyses in both biomedical studies and geometric morphometric applications. First, SINATRA does not require landmarks or correspondences in the data. This means that the algorithm can be implemented on both datasets with raw unaligned shapes, as well as those that have been axis-aligned during preprocessing (see Supplementary Material; Wang et al. (2021)). Second, given any dataset of 3D images, SINATRA outputs evidence measures that highlight the physical regions on shapes that explain the greatest variation between two classes. In many applications, users may suspect a priori that certain landmarks have greater variation across groups of shapes (e.g., via the literature). To this end, SINATRA also provides P-values and Bayes factors that detail how likely any region is identified by chance (Sellke, Bayarri and Berger (2001)).

In this paper we describe each mathematical step of the SINATRA pipeline and demonstrate its power and utility via simulations. We also use a dataset of mandibular molars from four different genera of primates to show that our method has the ability to: (i) further understanding of how landmarks vary across evolutionary scales in morphology and (ii) visually detail how known anatomical aberrations are associated to specific disease classes and/or case-control studies.

2. Method overview. The SINATRA pipeline implements four key steps (Figure 1). First, SINATRA summarizes the geometry of 3D shapes (represented as triangular meshes) by a collection of vectors (or curves) that encode changes in their topology. Second, a nonlinear Gaussian process model, with the topological summaries as input, classifies the shapes. Third, an effect size analog and corresponding association metric is computed for each topological feature used in the classification model. These quantities provide evidence that a given topological feature is associated with a particular class. Fourth, the pipeline iteratively maps the topological features back onto the original shapes (in rank order according to their association measures) via a reconstruction algorithm. This highlights the physical (spatial) locations that best explain the variation between the two groups. Details of our implementation choices are detailed below, with theoretical support given in the Supplementary Material (Wang et al. (2021)).

2.1. *Topological summary statistics for* 3*D shapes.* In the first step of the SINATRA pipeline, we use a tool from integral geometry and differential topology called the Euler characteristic (EC) transform (Crawford et al. (2020), Curry, Mukherjee and Turner (2019), Ghrist, Levanger and Mai (2018), Turner, Mukherjee and Boyer (2014)). For a mesh $\mathcal{M}$, the Euler characteristic is an accessible topological invariants derived from

$$\chi = \#V(\mathcal{M}) - \#E(\mathcal{M}) + \#F(\mathcal{M}), \tag{1}$$

where $\{\#V(\mathcal{M}), \#E(\mathcal{M}), \#F(\mathcal{M})\}$ denote the number of vertices (corners), edges and faces of the mesh, respectively. An EC curve $\chi_v(\mathcal{M})$ tracks the change in the Euler characteristic with respect to a given filtration of length l in direction v (Figure 1(a) and (b)). Theoretically, we first specify a height function $h_v(\boldsymbol{x}) = \boldsymbol{x}^\intercal v$ for vertex $\boldsymbol{x} \in \mathcal{M}$ in direction v. We then use this height function to define sublevel sets (or subparts) of the mesh $\mathcal{M}_v^a$ in direction v, where $h_v(\boldsymbol{x}) \leq a$. In practice, the EC curve is $\chi(\mathcal{M}_v^a)$ computed over a range of l filtration steps in direction v (Figure 1(b)).

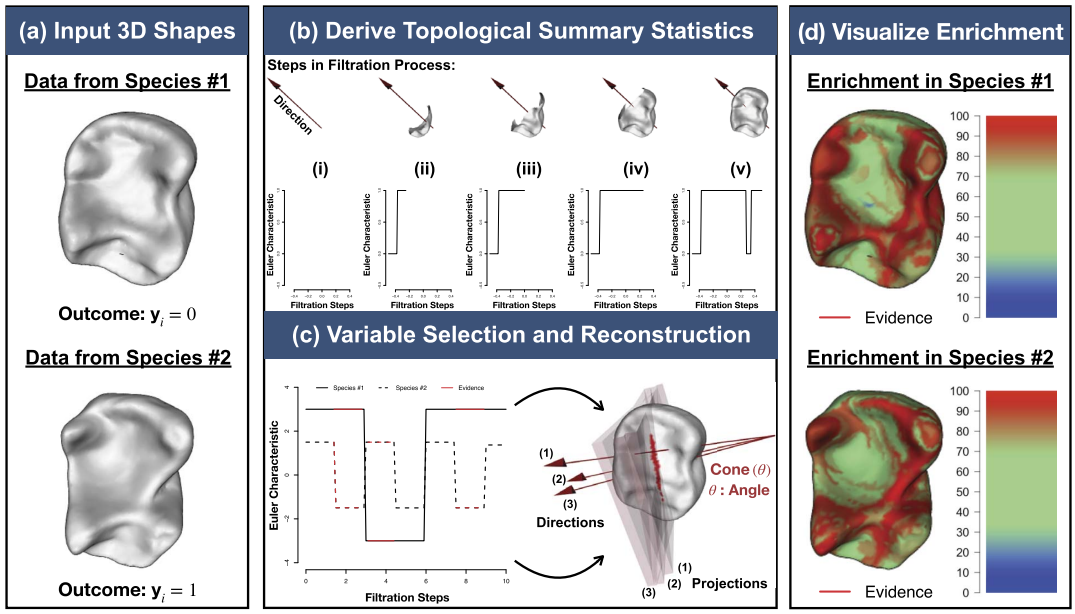


FIG. 1. *Schematic overview of SINATRA, a novel statistical framework for feature selection and association mapping with 3D shapes. (a) The SINATRA algorithm requires the following inputs: (i) Aligned shapes represented as meshes; (ii)* $\mathbf{y}$*, a binary vector denoting shape classes; (iii)* r*, the radius of the bounding sphere for the shapes; (iv)* c*, the number of cones of directions; (v)* d*, the number of directions within each cone; (vi)* θ*, the cap radius used to generate directions in a cone and (vii)* l*, the number of sublevel sets (i.e., filtration steps) to compute the Euler characteristic (EC) along a given direction. Guidelines for how to choose the free parameters are given in Supplementary Table* 1*. (b) We select initial positions uniformly on a unit sphere. Then, for each position we generate a cone of* d *directions within angle* θ *using Rodrigues' rotation formula (Belongie (1999)), resulting in a total of* $m = c \times d$ *directions. For each direction we compute EC curves with* l *sublevel sets. We concatenate the EC curves along all the directions for each shape to form vectors of topological features of length* $p = l \times m$*. Thus, for a study with* n*-shapes, an* $n \times p$ *design matrix is statistically analyzed using a Gaussian process classification model. (c) Evidence of association for each topological feature vector are determined using relative centrality measures. We reconstruct corresponding shape regions by identifying the vertices (or locations) on the shape that correspond to "statistically associated" topological features. (d) This enables us to visualize the enrichment of physical features that best explain the variance between the two classes. The heatmaps display vertex evidence potential on a scale from* $[0 - 100]$*. A maximum of* 100 *represents the threshold at which the first shape vertex is reconstructed, while* 0 *denotes the threshold when the last vertex is reconstructed.*

The EC transform is the collection of EC curves across a set of directions $v = 1, \ldots, m$, and maps a 3D shape into a concatenated $p = (l \times m)$-dimensional feature vector. For a study with n-shapes, an $n \times p$ design matrix $\mathbf{X}$ is statistically analyzed, where the columns denote the Euler characteristic computed at a given filtration step and direction. Each sublevel set value, direction and set of shape vertices used to compute an EC curve are stored for the association mapping and projection phases of the pipeline. Previously, Curry et al. proved sufficiency, stating that the theoretical upper bound on the minimum number of directions m required for the EC transform to preserve all information for a family of 3D shapes (see Theorem 7.14 in Curry, Mukherjee and Turner (2019)) is estimated by

$$m \geq (2b_\delta + 1)\left[1 + \frac{3}{\delta}\right]^2 + O\left(\frac{3b_\delta}{\delta^2}\right)^6, \tag{2}$$

where δ is a lower bound on the "curvature" at every vertex on the mesh and b_δ is a uniform upper bound on the number of critical values for the Euler characteristic curves when viewed in any given δ-ball of directions. While this upper bound may not yield optimal results in practice, we do use this theory to guide the collection of topological statistics with the general

notion that considering larger values of m will lead to more robust summarization of shape variation. In this paper we use a series of simulations and sensitivity analyses to outline empirical trends and develop intuition behind practically choosing the number of directions m and setting the granularity of sublevel filtrations l for real data.

2.2. *Statistical model for shape classification*. In the second step of the SINATRA pipeline, we use (weight-space) Gaussian process probit regression to classify shapes based on their topological summaries generated by the EC transformation. Namely, we specify the following (Bayesian) hierarchical model (Neal (1997, 1999), Nickisch and Rasmussen (2008), Rasmussen and Williams (2006), Williams and Barber (1998)):

$$\mathbf{y} \sim \mathcal{B}(\boldsymbol{\pi}), \qquad \mathrm{g}(\boldsymbol{\pi}) = \Phi^{-1}(\boldsymbol{\pi}) = \boldsymbol{f}, \quad \boldsymbol{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}), \tag{3}$$

where $\mathbf{y}$ is an n-dimensional vector of Bernoulli distributed class labels, $\boldsymbol{\pi}$ is an n-dimensional vector representing the underlying probability that a shape is classified as a "case" (i.e., $y = 1$), $\mathrm{g}(\cdot)$ is a probit link function with $\Phi(\cdot)$ the cumulative distribution function (CDF) of the standard normal distribution and $\boldsymbol{f}$ is an n-dimensional vector estimated from the data.

The key objective of SINATRA is to use the topological features in $\mathbf{X}$ to find the physical 3D properties that best explain the variation across shape classes. To do so, we use kernel regression, where the utility of generalized nonparametric statistical models is well established due to their ability to account for various complex data structures (Cheng et al. (2019), Heckerman et al. (2016), McDonald et al. (2019), Rodriguez-Nieva and Scheurer (2019), Swain et al. (2016), Zhang, Dai and Jordan (2011)). Generally, kernel methods posit that $\boldsymbol{f}$ lives within a reproducing kernel Hilbert space (RKHS) defined by some (nonlinear) covariance function, which implicitly account for higher-order interactions between features, leading to more complete classifications of data (Jiang and Reif (2015), Pillai et al. (2007), Schölkopf, Herbrich and Smola (2001)). To this end, we assume $\boldsymbol{f}$ is normally distributed with mean vector $\mathbf{0}$, and the covariance matrix $\mathbf{K}$ is defined by the radial basis function $\mathbf{K}_{ij} = \exp\{-\theta\|\mathbf{x}_i - \mathbf{x}_j\|^2\}$ with bandwidth θ set using the median heuristic to maintain numerical stability and avoid additional computational costs (Chaudhuri et al. (2017)). The full model specified in equation (3) is commonly referred to as "Gaussian process classification" or GPC.

2.3. *Interpretable feature* (*variable*) *selection*. To estimate the model in equation (3), we use an elliptical slice sampling Markov chain Monte Carlo (MCMC) algorithm (Supplementary Material, Section 1.1, Wang et al. (2021)). Since we take a "weight-space" view on Gaussian processes, the model fitting procedure scales with the number of 3D meshes in the data, rather than with the number of topological features. The MCMC algorithm allows samples from the approximate posterior distribution of $\boldsymbol{f}$ (given the data) and also allows for the computation of an effect size analog for each topological summary statistic (Singleton et al. (2017), Crawford et al. (2018, 2019)),

$$\boldsymbol{\beta} = (\mathbf{X}^\intercal\mathbf{X})^\dagger\mathbf{X}^\intercal\boldsymbol{f}, \tag{4}$$

where $(\mathbf{X}^\intercal\mathbf{X})^\dagger$ is the generalized inverse of $(\mathbf{X}^\intercal\mathbf{X})$.

These effect sizes represent the nonparametric equivalent to coefficients in linear regression using generalized least squares. SINATRA uses these weights and assigns a measure of relative centrality to each summary statistic (first panel Figure 1(c)) (Crawford et al. (2019)). This criterion evaluates how much information in classifying each shape is lost when a particular topological feature is removed from the model. This loss is determined by computing the Kullback–Leibler divergence (KLD) between: (i) the conditional posterior distribution

$p(\boldsymbol{\beta}_{-j}|\beta_j = 0)$ with the effect of the jth topological feature set to zero and (ii) the marginal posterior distribution $p(\boldsymbol{\beta}_{-j})$ with the effects of the jth feature integrated out,

$$\text{KLD}(\beta_j) = \int_{\boldsymbol{\beta}_{-j}} \log\left(\frac{p(\boldsymbol{\beta}_{-j})}{p(\boldsymbol{\beta}_{-j}|\beta_j = 0)}\right) p(\boldsymbol{\beta}_{-j}) \mathrm{d}\boldsymbol{\beta}_{-j}, \quad j = 1, \ldots, p, \tag{5}$$

which has a closed form solution when the posterior distribution of the effect sizes is assumed to be (approximately) Gaussian (Supplementary Material, Section 1.2, Wang et al. (2021)). Finally, we normalize to obtain an association metric for each topological feature, $\gamma_j = \text{KLD}(\beta_j)/\sum \text{KLD}(\beta_l)$.

There are two main takeaways from this formulation. First, the KLD is nonnegative and equals zero if and only if the posterior distribution of $\boldsymbol{\beta}_{-j}$ is independent of the effect β_j. Intuitively, this says that removing an unimportant shape feature has no impact on explaining the variance between shape classes. Second, $\boldsymbol{\gamma}$ is bounded on the unit interval $[0, 1]$ with the natural interpretation of providing relative evidence of association for shape features; higher values suggest greater importance. For this metric the null hypothesis assumes that every feature equally contributes to the total variance between shape classes, while the alternative proposes that some features are more central than others (Crawford et al. (2019)). As we show in the Results and Supplementary Material, when the null assumption is met, SINATRA displays association results that appear uniformly distributed and effectively indistinguishable (Wang et al. (2021)).

2.4. *Shape reconstruction*. After obtaining association measures for each topological feature, we map this information back onto the physical shape (second panel Figure 1(c) and (d)). We refer to this process as *reconstruction*, as this procedure recovers regions that explain the most variation between shape classes (Supplementary Material, Section 1.3, Wang et al. (2021)). Intuitively, we want to identify vertices on the shape that correspond to the topological features with the greatest association measures.

Begin by considering d directions within a cone of cap radius or angle θ, which we denote as $\mathcal{C}(\theta) = \{\nu_1, \ldots, \nu_d|\theta\}$. Next, let $\mathcal{Z}$ be the set of vertices whose projections onto the directions in $\mathcal{C}(\theta)$ are contained within the collection of "significant" topological features; for every $z \in \mathcal{Z}$, the product $z \cdot \nu$ is contained within a sublevel set (taken in the direction $\nu \in \mathcal{C}(\theta)$) that shows high evidence of association in the feature selection step.

A reconstructed region is then defined as the union of all mapped vertices from each cone, or $\mathcal{R} := \bigcup_i \mathcal{Z}_i$. We use cones because vectors of Euler characteristics, taken along directions close together, express comparable information. That similiarity lets us leverage findings between them to increase our power of detecting truly associated shape vertices and regions—as opposed to antipodal directions where the lack of shared information may do harm when determining reconstructed manifolds (Supplementary Material, Section 1.4, Wang et al. (2021)) (Curry, Mukherjee and Turner (2019), Fasy et al. (2018), Oudot and Solomon (2018)).

2.5. *Visualization of enriched shape regions*. Once shapes have been reconstructed, we can visualize the relative importance or "evidence potential" for each vertex on the mesh with a simple procedure. First, we sort the topological features from largest to smallest according to their association measures $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_p$. Next, we iterate through the sorted measures $T_k = \gamma_k$ (starting with $k = 1$) and reconstruct the vertices corresponding to the topological features in the set $\{j : \gamma_j \geq T_k\}$.

The evidence potential for each vertex is defined as the largest threshold T_k at which it is reconstructed for the first time, because vertices with earlier "birth times" in the reconstruction are more important relative to vertices that appear later. We illustrate these values via

heatmaps over the reconstructed meshes (Figure 1(d)). For consistency across different applications and case studies, we set the coloring of these heatmaps on a scale from $[0 - 100]$. A maximum value of 100 represents the threshold value at which the first vertex is born, while 0 denotes the threshold when the last vertex on the shape is reconstructed. Under the null hypothesis, where there are no meaningful regions differentiating between two classes of shapes, (mostly) all vertices appear to be born relatively early and at the same time (Supplementary Figure 4). This is not the case under the alternative.

2.6. *Algorithm and implementation*. Source code for replication is available in the Supplementary Materials (Wang et al. (2021)) and freely available online at https://github.com/lcrawlab/SINATRA. The SINATRA algorithm requires these inputs:

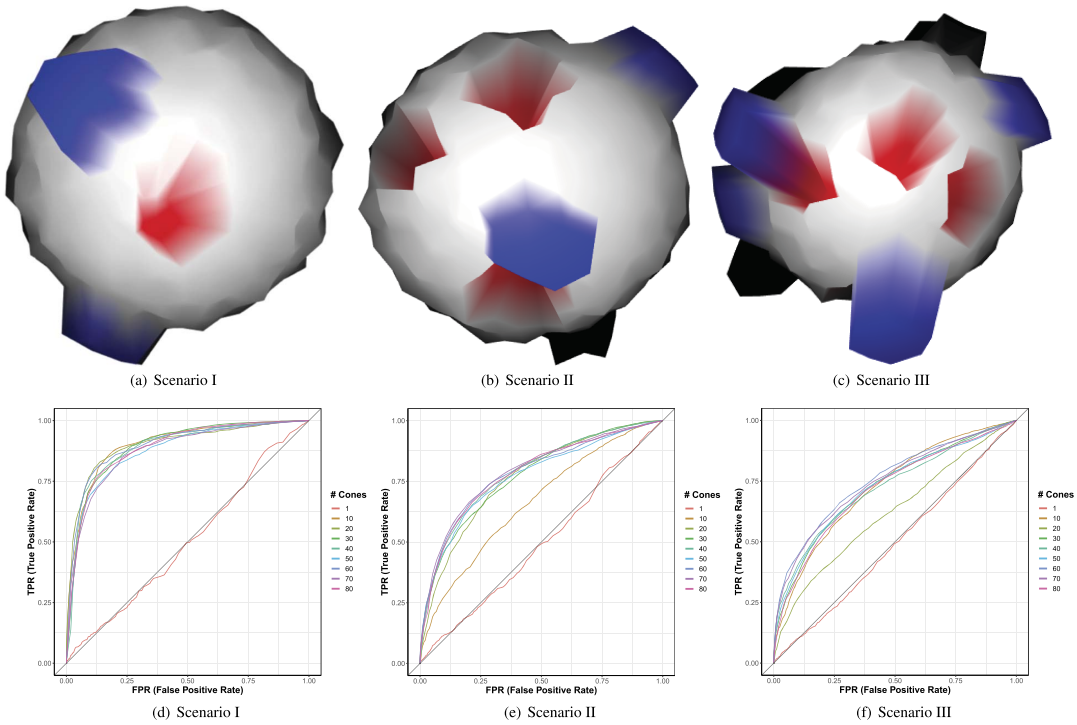
- shapes represented as meshes (unaligned or preprocessed and axis-aligned);
- $\mathbf{y}$, a binary vector denoting shape classes;
- r, the radius of the bounding sphere for the shapes (which we usually set to $1/2$ since we work with meshes normalized to the unit ball);
- c, the number of cones of directions;
- d, the number of directions within each cone;
- θ, the cap radius used to generate directions in a cone;
- l, the number of sublevel sets (i.e., filtration steps) to compute the Euler characteristic (EC) along a given direction.

A table with general guidelines for how set these free parameters in practice is given in Supplementary Table 1. In the next section we discuss strategies for choosing values for the free parameters through simulation studies. A table detailing the scalability for the current algorithmic implementation of SINATRA can also be found in the Supplementary Material (see Supplementary Table 2). For some context, runtime for SINATRA is dependent upon both the number of meshes in a given study (n) as well as the number of cones of directions (c), directions within each cone (d) and sublevel sets (l) used to compute the EC curves. Note that the latter three parameters all contribute to the total number of topological statistics used summarize each mesh (i.e., $p = l \times m$ EC features taken over $m = c \times d$ total directions). Overall, the SINATRA algorithm scales (approximately) linearly in the number of shapes. For example, while holding the values of $\{c = 25, d = 5, l = 25\}$ constant, it takes $\sim$60 and $\sim$196 seconds to analyze datasets with $n = 25$ and $n = 100$ shapes, respectively. The rate limiting step in the SINATRA pipeline is the computation of the association metric for each topological summary statistic (see equation (5)). To mitigate this burden, we use approximations such that the cost of calculating each γ_j is made up of p-independent $O(p^2)$ operations which can be parallelized (see derivations provided in Supplementary Material, Section 1.2, Wang et al. (2021)).

3. Results.

3.1. *Simulation study*: *Perturbed spheres*. We begin with a proof-of-concept simulation study to demonstrate both the power of our proposed pipeline and how different parameter value choices affect its ability to detect associated features on 3D shapes. Again, a table with general guidelines for how set the free parameters within the SINATRA pipeline can be found in Supplementary Table 1. Here, we take 100 spheres and perturb regions, or collections of vertices, on their surfaces to create two classes with a two-step procedure:

1. We generate a fixed number of (approximately) equidistributed regions on each sphere—some number u regions to be shared across classes and the remaining v regions to be unique to class assignment.



(a) Scenario I (b) Scenario II (c) Scenario III

(d) Scenario I (e) Scenario II (f) Scenario III

FIG. 2. *Power and sensitivity analysis for detecting associated vertices across different classes of perturbed spheres. We generate* 100 *shapes by partitioning unit spheres into regions* 10 *vertices wide, centered at* 50 *equidistributed points. Two classes* (50 *shapes per class*) *are defined by shared* (*blue protrusions*) *and class-specific* (*red indentations*) *characteristics. The shared or "nonassociated" features are chosen by randomly selecting u regions and pushing the sphere outward at each of these positions. To generate class-specific or "associated" features, v distinct regions are chosen for a given class and perturbed inward. We vary these parameters and analyze three increasingly more difficult simulation scenarios*: (*a*) $u = 2$ *shared and* $v = 1$ *associated*; (*b*) $u = 6$ *shared and* $v = 3$ *associated, and* (*c*) $u = 10$ *shared and* $v = 5$ *associated. In panels* (*d*)–(*f*), *ROC curves depict the ability of SINATRA to identify vertices located within associated regions as a function of increasing the number of cones of directions used in the algorithm. These results give empirical evidence that seeing more of a shape* (*i.e., using more unique directions*) *generally leads to an improved ability to map back onto associated regions. Other SINATRA parameters were fixed*: $d =$ *five directions per cone,* $\theta = 0.15$ *cap radius used to generate directions in a cone and* $l = 30$ *sublevel sets per filtration. Results are based on* 50 *replicates in each scenario.*

2. To create each region, we perturb the k closest vertices $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k\}$ by a prespecified scale factor α and add some random normally distributed noise $\epsilon_i \sim \mathcal{N}(0, 1)$ by setting $\boldsymbol{x}_i^* := \boldsymbol{x}_i\alpha + \epsilon_i$ for $i = 1, \ldots, k$.

We consider three scenarios based on the number of shared and unique regions between shape classes (Figure 2(a)–(c)). We choose $(u, v) = (2, 1)$ (scenario I), $(6, 3)$ (scenario II) and $(10, 5)$ (scenario III) and set all regions to be $k = 10$ vertices.

Each sequential scenario represents an increase in degree of difficulty, because class-specific regions should be harder to identify in shapes with more complex structures. We analyze 50 unique simulated datasets for each scenario. In each dataset only the v-region vertices used to create class-specific regions are defined as true positives, and we quantify SINATRA's ability to prioritize these true vertices using receiver operating characteristic (ROC) curves plotting true positive rates (TPR) against false positive rates (FPR) (Supplementary Material, Section 2, Wang et al. (2021)). We then evaluate SINATRA's power as a function of its free parameter inputs: c number of cones, d number of directions per cone, direction generating cap radius θ and l number of sublevel sets per filtration. We iteratively vary each parameter while holding the others as constants $\{c = 25, d = 5, \theta = 0.15, l = 30\}$. Figures displayed in

the main text are based on varying the number of cones (Figure 2(d)–(f)), while results for the other sensitivity analyses can be found in the Supplementary Material (Supplementary Figures 1–3).

As expected, SINATRA's performance is consistently better when shapes are defined by a few prominent regions (e.g., scenario I) vs. when shape definitions are more complex (e.g., scenarios II and III), because each associated vertex makes a greater individual contribution to the overall variance between classes (i.e., $\mathbb{V}(\mathbf{y})/10 > \mathbb{V}(\mathbf{y})/30 > \mathbb{V}(\mathbf{y})/50$). Similar trends in performance have been shown during the assessment of high-dimensional variable selection methods in other application areas (Crawford et al. (2017), Li et al. (2015), Zhu and Stephens (2018)).

This simulation study also demonstrates the general behavior and effectiveness of the SINATRA algorithm as a function of different choices for its free input parameters. First, we assess how adjusting the number of cones of directions used to compute Euler characteristic curves changes power. Computing topological summary statistics over just a single cone of directions (i.e., $c = 1$) is ineffective at capturing enough variation to identify class-specific regions (Figure 2(d)–(f)) which supports the intuition that seeing more of a shape leads to an improved ability to understand its complete structure (Crawford et al. (2020), Curry, Mukherjee and Turner (2019), Turner, Mukherjee and Boyer (2014)). Our empirical results show that more power can be achieved by summarizing the shapes with filtrations taken over multiple directions. In practice, we suggest specifying multiple cones $c > 1$ and utilizing multiple directions d per cone (see monotonically increasing power in Supplementary Figure 1).

While the other two parameters (θ and l) do not have monotonic properties, their effects on SINATRA's performance still have natural interpretations. For example, when changing the angle between directions within cones from $\theta \in [0.05, 0.5]$ radians, we observe that power steadily increases until $\theta = 0.25$ radians and then slowly decreases afterward (Supplementary Figure 2). This supports previous theoretical results that cones should be defined by directions in close proximity to each other (Curry, Mukherjee and Turner (2019)) but not so close that they explain the same local information with little variation.

Perhaps most importantly, we must understand how the number of sublevel sets l (i.e., the number of steps in the filtration) used to compute Euler characteristic curves affects the performance of the algorithm. As we show in the next section, this function depends on the types of shapes being analyzed. Intuitively, for very intricate shapes, coarse filtrations with too few sublevel sets cause the algorithm to miss or "step over" very local undulations in a shape. For the spheres simulated in this section, class-defining regions are global-like features, and so finer filtration steps fail to capture broader differences between shapes (Supplementary Figure 3); however, this failure is less important when only a few features decide how shapes are defined (e.g., scenario I). In practice, we recommend choosing the angle between directions within cones θ and the number of sublevel sets l via cross validation or some grid-based search.

As a final demonstration, we show what happens when we meet the null assumptions of the SINATRA pipeline (Supplementary Figure 4). Under the null hypothesis our feature selection measure assumes that all 3D regions of a shape equally contribute to explaining the variance between classes; that is, no one vertex (or corresponding topological characteristics) is more important than the others. We generate synthetic shapes under the two cases when SINATRA fails to produce significant results: (a) two classes of shapes that are effectively the same (up to some small Gaussian noise) and (b) two classes of shapes that are completely dissimilar. In the first simulation case there are no "significantly associated" regions, and, thus, no group of vertices stand out as important (Supplementary Figure 4(a)). In the latter simulation case, shapes between the two classes look nothing alike; therefore, all vertices contribute to class definition, but no one feature is key to explaining the observed variation (Supplementary Figure 4(b)).

3.2. *Simulation study*: *Caricatured shapes*. Our second simulation study modifies computed tomography (CT) scans of real Lemuridae teeth (one of the five families of Strepsirrhini primates commonly known as lemurs) (Boyer et al. (2011)) using a well-known caricaturization procedure (Sela, Aflalo and Kimmel (2015)). We fix the triangular mesh of an individual tooth and specify class-specific regions centered around known biological landmarks (Figure 3) (Boyer et al. (2011)). For each triangular face contained within a class-specific region, we apply a corresponding affine transformation, positively scaled, that smoothly varies on the triangular mesh and attains its maximum at the biological landmark used to define the region (Supplementary Material, Section 3, Wang et al. (2021)). We caricature 50 different teeth with two steps (Figure 3(a)):

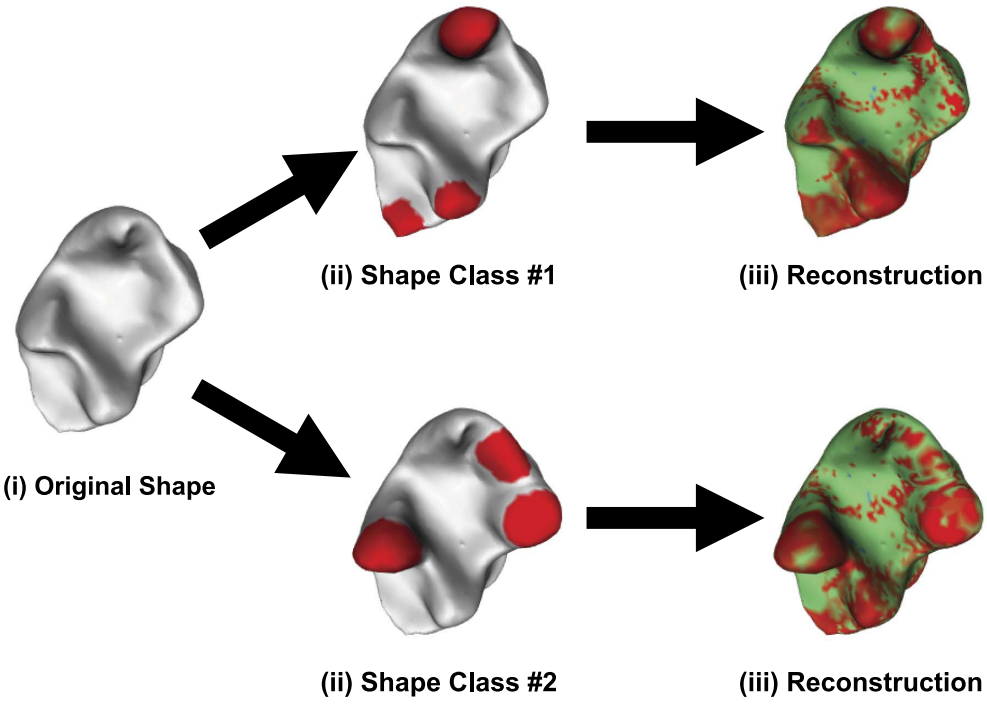
1. Assign v of a given tooth's landmarks to be specific to one class and v' to be specific to the other class.
2. Perform the caricaturization: Multiply each face in the v and v' class-specific regions by a positive scalar (i.e., exaggerated or enhanced). Repeat 25 times (with some small noise per replicate) to create two equally-sized classes of 25 shapes.

We explore two scenarios by varying the number of class-specific landmarks v and v' that determine the caricaturization in each class. First, we set both $v, v' = 3$; next, we fix $v, v' = 5$. Like the simulations with perturbed spheres, the difficulty of the scenarios increases with the number of caricatured regions. We evaluate SINATRA's ability to identify the vertices involved in the caricaturization using ROC curves (Supplementary Material, Section 2, Wang et al. (2021)), and we assess this estimate of power as a function of the algorithm's free parameter inputs. While varying each parameter, we hold the others as constants $\{c = 15, d = 5, \theta = 0.15, l = 50\}$. Figures in the main text are based on varying the number of cones c (Figure 3(b) and (c)); results for the other sensitivity analyses can be found in the Supplementary Material (Supplementary Figures 5–7).

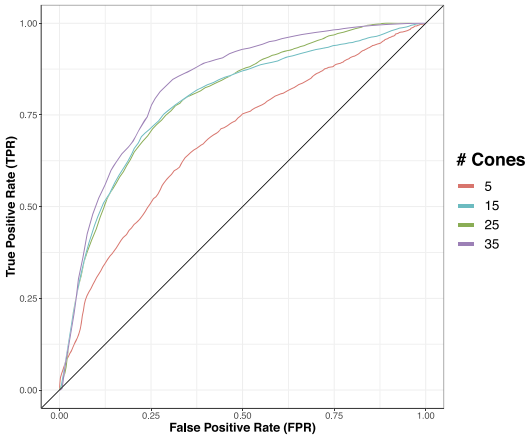
Overall, using fewer caricatured regions results in better (or at least comparable) performance. Like the simulations with perturbed spheres, SINATRA's power increases monotonically with an increasing number of cones and directions used to compute the topological summary statistics (Figures 3(b), 3(c) and Supplementary 5). For example, at a 10% FPR with $c = 5$ cones, we achieve 30% TPR in scenario I experiments and 35% TPR in scenario II. Increasing the number of cones to $c = 35$ improves power to 52% and 40% TPR for scenarios I and II, respectively. Trends from the previous section continue when choosing the angle between directions within cones (Supplementary Figure 6) and the number of sublevel sets (Supplementary Figure 7). Results for the perturbed spheres suggest that there is an optimal cap radius for generating directions in a cone. Since we are analyzing shapes with more intricate features, finer filtrations lead to more power.

3.3. *Simulation study*: *Method comparisons*. In this subsection we compare the power of SINATRA to other state-of-the-art sub-image selection methods. Here, we revisit the same simulation scenarios using the perturbed spheres and caricatured teeth simulation schemes, respectively. Once again, we use ROC curves plotting TPR vs. FPR to assess the ability of each method to identify vertices in class-specific associated regions. For SINATRA we use a grid search to choose the algorithm's free parameters. This led to us setting $\{c = 60, d = 5, \theta = 0.15, l = 30\}$ for the perturbed spheres and $\{c = 35, d = 5, \theta = 0.15, l = 50\}$ for the caricatured teeth. Note that these final values follow the trends we observed from the sensitivity analyses presented in the previous two subsections. We then consider the following three types of competing approaches:

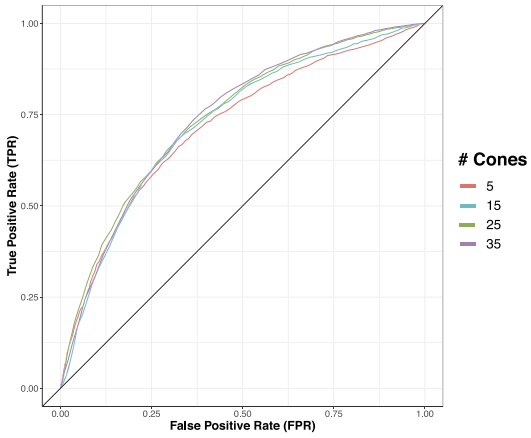
- *Vertex-level regularization (baseline)*: Using the fact that all the vertices within a set of simulated meshes are in one-to-one correspondence, we vectorize the shapes by concatenating the (x, y, z)-coordinates of each vertex into a single vector. In this way we have



(a) Caricature Simulation Flowchart



(b) Scenario I (3 Peaks)



(c) Scenario II (5 Peaks)

FIG. 3. *Power and sensitivity analysis for detecting associated vertices across different classes of caricatured shapes.* (*a*) *We modify real Lemuridae molars using the following caricaturization procedure*: (*i*) *Fix the triangular mesh of an individual tooth*; (*ii*) *Assign* v *of the known landmarks for the tooth* (Boyer *et al.* (2011)) *to be specific to one class and* v' *to be specific to the other. The caricaturization is performed by positively scaling each face within these regions so that class-specific features are exaggerated. We repeat* 25 *times* (*with some small added noise*) *to create two classes of* 25 *shapes.* (*iii*) *The synthetic shapes are analyzed by SINATRA to identify the associated regions. We consider two scenarios by varying the number of class-specific landmarks that determine the caricaturization in each class. In scenario I, we set* $v, v' = 3$; *and in scenario II,* $v, v' = 5$. *In panels* (*b*) *and* (*c*), *ROC curves depict the ability of SINATRA to identify vertices located within associated regions, as a function of the number of cones of directions used in the algorithm. Other SINATRA parameters were fixed*: $d = 5$ *directions per cone,* $\theta = 0.15$ *cap radius used to generate directions in a cone and* $l = 50$ *sublevel sets per filtration. Results are based on* 50 *replicates in each scenario.*

a $(3 \times q)$-dimensional vector representation of each shape in our dataset, where q is the number of vertices that each mesh contains. In other words, each vertex is directly represented by the three entries of the vector (once for each dimension of the shape). This results in a final $n \times (3 \times q)$ design matrix, where n is the total number of shapes in the data. To perform subimage selection, we implement elastic-net regularization (Zou and Hastie (2005)) using the `glmnet` package (Friedman, Hastie and Tibshirani (2010)) in `R` (with the penalization term chosen via cross-validation) to assign sparse individual coefficients to each column of the design matrix. Power is then determined by either ranking the mean or maximum of the three coefficients corresponding to the (x, y, z)-coordinates for each vertex.

- *Landmark-level regularization*: We implement this method by taking advantage of the fact that the simulated objects in our study are generated from perturbations of a base shape. Here, we generate landmarks by selecting $v = \{500, 2000\}$ equidistributed points on a base shape (in the sense of Euclidean distance). Next, we concatenate the (x, y, z)-coordinates of the k closest vertices to each point and obtain landmark-specific vectors in $\mathbb{R}^{3k}$. This results in a final $n \times (3 \times v \times k)$ design matrix, where, again, n is the total number of shapes in the data. To perform subimage selection with this method, we use a logistic regression model with group-lasso regularization (Simon et al. (2013)), where the groups are predefined as the coordinates belonging to a given landmark and the group-based penalization term is chosen via cross-validation with the `gglasso` (Yang and Zou (2015)) in `R`. Note that we consider a group-lasso penalty in order to encourage selection of important landmarks as a whole, rather than individual vertices. True and false positives are then determined by ranking each landmarks based on the magnitude of their coefficients.
- *Limit shapes algorithm (Huang et al. (2019))*: This algorithm builds upon a "functional map network" consisting of (dual representations of) point-by-point pairwise correspondences between all pairs of shapes in a dataset. From here, a consistent latent basis can be extracted and used for constructing a "limit shape" underlying the given collection of 3D objects but represented in a latent space. Deviations of individual shapes from this "limit shape," represented as functions defined on these shapes, provide a visual guide highlighting sources of variability that often carry rich semantic meanings. In our simulations we generate maps between all mesh pairs using a state-of-the-art pipeline that searches for the optimal "bounded-distortion map" and interpolates as many "Gaussian process landmarks" as possible (Gao et al. (2019)). These point-to-point maps of bounded conformal distortion are then converted into functional maps which are essentially the same maps but represented under a different system of bases of eigenfunctions of the discrete Laplace–Beltrami operators. The resulting functional maps are directly fed into the limit shapes workflow (Huang et al. (2019)). To perform subimage selection, we compute a vertex importance vector using the first 10 eigenfunctions of the discrete Laplace–Beltrami operator corresponding to the 10 smallest eigenvalues. Vertex weights were summarized within the eigenfunctions by scaling the absolute value of each entry ψ_i by the maximum vertex weight within that eigenfunction such that $w_j = |\psi_j| / \max\{|\psi_1|, \ldots, |\psi_n|\}$. After norming weights within the eigenfunctions, we obtain a final vector of vertex weights by taking the entrywise maximum of each vertex across the 10 eigenfunctions. Using this procedure enables us to rank features from limit shapes and to generate ROC curves. To imitate realistic conditions where point-by-point maps are not known a priori, we also introduce a (misspecified) variation of limit shapes in which the functional map network is partially scrambled.

Figures 4 and 5 display the performance of SINATRA and each of the competing methods on the perturbed spheres and caricatured teeth, respectively, across 50 replicates in each simulation scenario. There are a few important takeaways from these comparisons. First, the group

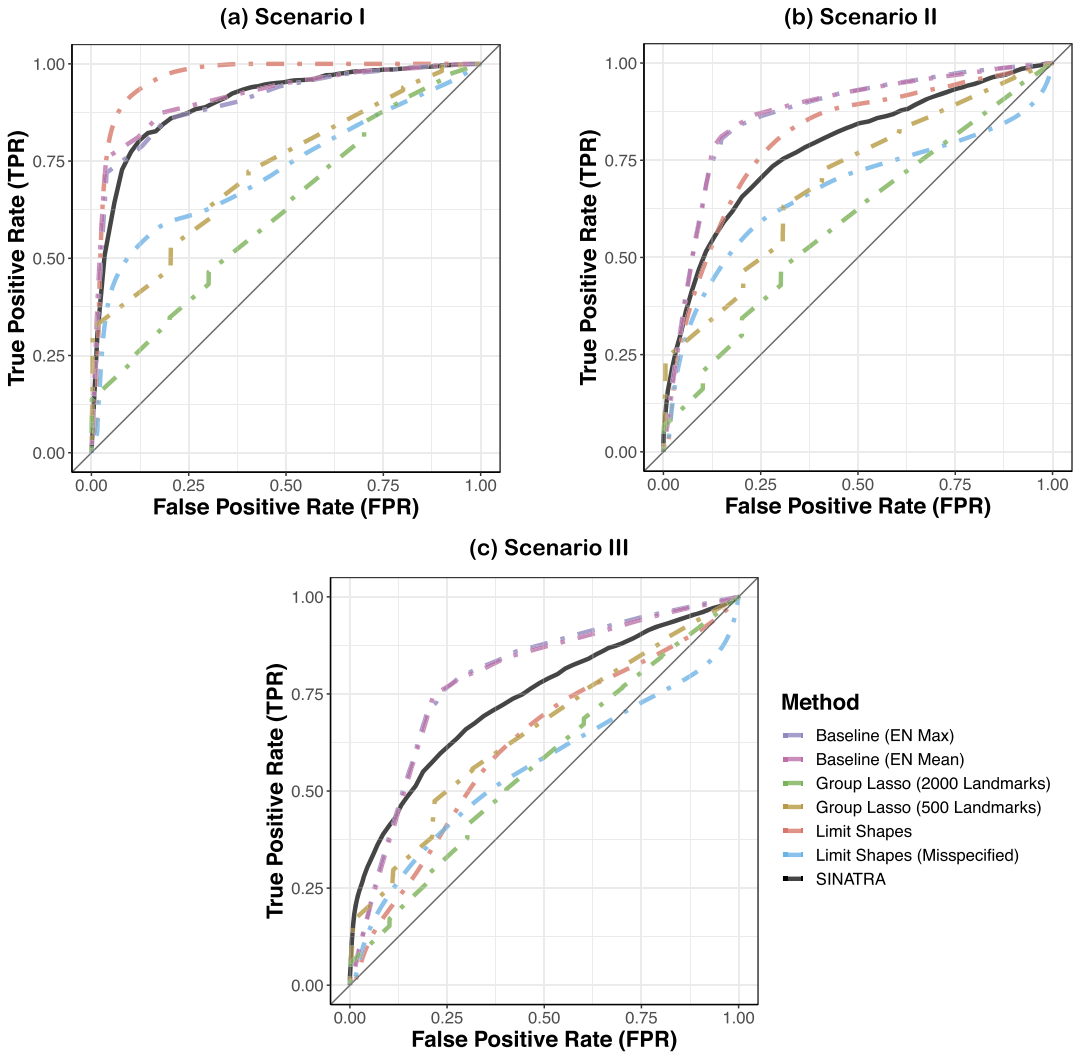


FIG. 4. *Receiver operating characteristic (ROC) curves comparing the performance of SINATRA with competing methods in the perturbed sphere simulations. Here, we generate* 100 *shapes by partitioning unit spheres into regions* 10 *vertex-wide, centered at* 50 *equidistributed points. Two classes* (50 *shapes per class*) *are defined by the number of shared* u *and class-specific* v *characteristics. We vary these parameters and analyze three increasingly more difficult simulation scenarios*: (*a*) $u = 2$ *shared and* $v = 1$ *associated*; (*b*) $u = 6$ *shared and* $v = 3$ *associated, and* (*c*) $u = 10$ *shared and* $v = 5$ *associated. The ROC curves depict the ability of SINATRA to identify vertices located within associated regions using parameters* $\{c = 60, d = 5, \theta = 0.15, l = 30\}$ *chosen via a grid search. We compare SINATRA to three methods. The first baseline concatenates the* (x, y, z)*-coordinates of all vertices on the sphere and treats them as features in a data frame. It then uses elastic-net regularization to assign sparse individual coefficients to each coordinate. For this method we assess power by either taking the mean or maximum of the coefficient values corresponding to each vertex. The second method assumes* 500 *or* 2000 *equally spaced landmarks across each mesh and implements a group-lasso penalty on the collection on vertices within these regions to rank associated features. Lastly, we compare the limit shapes algorithm (Huang et al.* (2019)) *where normalized vertex weights are used to determine true and false positives. Note that the limit-shapes algorithm requires a functional correspondence map between all pairs of meshes in the data; therefore, we display results for this algorithm both when a correspondence map is known and when the map has been misspecified. Results in each ROC curve are based on* 50 *replicates in each scenario.*

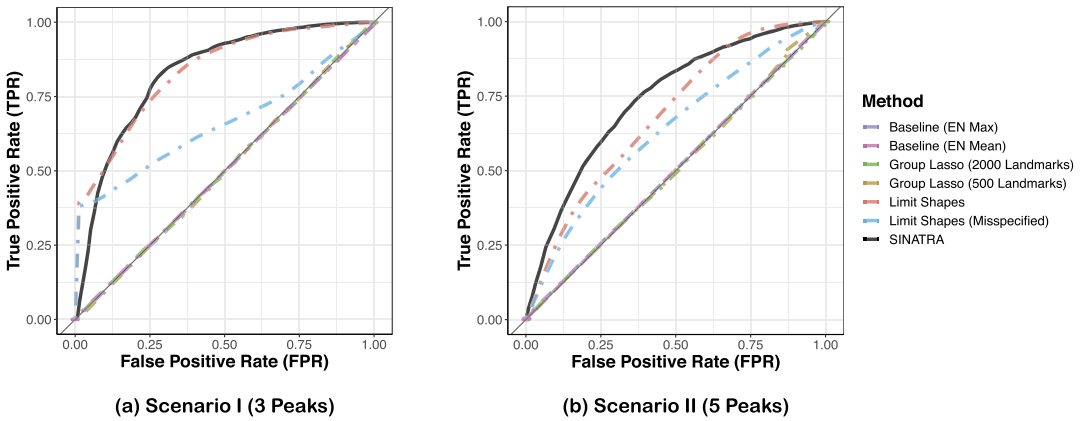


FIG. 5. *Receiver operating characteristic (ROC) curves comparing the performance of SINATRA with competing methods in the caricatured teeth simulations. Here, we modify real Lemuridae molars using a caricaturization procedure where we assign a v number of known landmarks for a tooth (Boyer et al. (2011)) to be specific to one class and v' to be specific to the other. We then consider two scenarios by varying the number of class-specific landmarks. In scenario I we set $v, v' = 3$, and in scenario II, $v, v' = 5$. The ROC curves depict the ability of SINATRA to identify vertices located within associated regions using parameters $\{c = 35, d = 5, \theta = 0.15, l = 50\}$ chosen via a grid search. We compare SINATRA to three methods. The first baseline concatenates the (x, y, z)-coordinates of all vertices on each caricatured tooth and treats them as features in a data frame. It then uses elastic-net regularization to assign sparse individual coefficients to each coordinate. For this method we assess power by either taking the mean or maximum of the coefficient values corresponding to each vertex. The second method assumes* 500 *or* 2000 *equally spaced landmarks across each mesh and implements a group-lasso penalty on the collection on vertices within these regions to rank associated features. Lastly, we compare the limit-shapes algorithm (Huang et al. (2019)) where normalized vertex weights are used to determine true and false positives. Note that the limit-shapes algorithm requires a functional correspondence map between all pairs of meshes in the data; therefore, we display results for this algorithm both when a correspondence map is known and when the map has been misspecified. Results in each ROC curve are based on* 50 *replicates in each scenario.*

lasso landmark-based method consistently performs the worst among all approaches that we consider, regardless of the number of landmarks used. This is likely due to overregularization where landmarks containing only a few vertices with nonzero effects are still be treated as nonassociated with class variation. The vertex-level elastic net baselines perform the best under the perturbed sphere simulations, particularly when the variation between shape classes is sparsely driven by just a few associated regions (see Figure 4). Furthermore, since this approach uses a simple regression, it scales linearly with the number of vertices on each shape. The limit shapes algorithm also had a quicker runtime than SINATRA which took approximately 30 minutes to run per analysis (Supplementary Table 2). While the limit shapes approach performs generally well in each simulation scenario, its performance drops significantly when the functional mapping input into the algorithm is misspecified. This highlights an important and practical advantage of SINATRA which maintains its utility even when such user-specified point-by-point correspondences are unknown. Overall, the performance of SINATRA remains competitive in all settings but clearly becomes the best approach in the caricatured teeth simulations where there is an increase in the overall complexity of the meshes that are analyzed (see Figure 5). We hypothesize that the simple vertex-level regularization approaches struggle on the caricatured teeth because the coordinate-based representation becomes less effective at capturing the varying topology and geometry between shapes.

3.4. *Recovering known morphological variation across genera of primates.* As an application of our pipeline, with "ground truth" or known morphological variation, we consider a

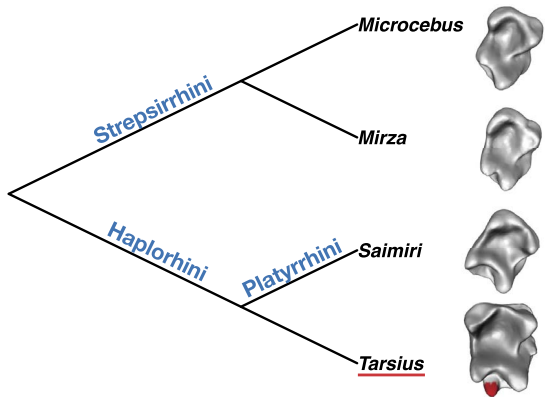


FIG. 6. *Illustration of the relationship between the phylogenetics and unique paraconids in molars belonging to primates in Tarsius genus. We carry out three pairwise comparisons to analyze the physical differences between Tarsius molars and teeth from*: (*i*) *Saimiri*, (*ii*) *Mirza and* (*iii*) *Microcebus genus. Here, we depict the phylogentic relationship between these groups. Morphologically, we know that tarsier teeth have an additional high cusp* (*highlighted in red*) *which allows this genus of primate to eat a wider range of foods* (*Crompton, Savage and Spears* (1998)). *With these data we want to assess SINATRA's ability to find this region of interest* (*ROI*).

dataset of CT scans of $n = 59$ mandibular molars from two suborders of primates: Haplorhini (which include tarsiers and anthropoids) and Strepsirrhini (which include lemurs, galagos and lorises). From the haplorhine suborder, 33 molars came from the genus *Tarsius* (Boyer et al. (2011), Gao (2015, 2021)) and *nine* molars from the genus *Saimiri* (St Clair and Boyer (2016)). From the strepsirrhine suborder, 11 molars came from the genus *Microcebus* and *six* molars from the genus *Mirza* (Boyer et al. (2011), Gao (2015, 2021)); both are lemurs.

We chose this specific collection of molars because morphologists and evolutionary anthropologists understand variations of the paraconid, the cusp of a primitive lower molar. The paraconids are retained only by *Tarsius* and do not appear in the other genera (Figure 6) (Guatelli-Steinberg (2003), St Clair and Boyer (2016)). Using phylogenetic analyses of mitochondrial genomes across primates, Pozzie et al. estimate divergence dates of the subtree composed of *Microcebus* and *Mirza* from *Tarsius* at *five* million years before the branching of *Tarsius* from *Saimiri* (Pozzi et al. (2014)). We want to see if SINATRA recovers the information that the paraconids are specific to the *Tarsius* genus. We also investigate if variation across the molar is associated to the divergence time of the genera.

For these analyses we consider two types of data preprocessing procedures. In the first, the meshes of all teeth were pre-aligned, centered at the origin, and normalized within a unit sphere using the `auto3dgm` software (Puente (2013)) (Supplementary Figure 8). In the second, we first conduct the Euler characteristic transformation on the raw data for each tooth, and then we *implicitly* normalize the 3D meshes by aligning their EC curves (Supplementary Figures 9 and 10). This latter analysis is used to demonstrate the effectiveness of SINATRA in settings where a priori correspondences between shapes are unavailable (details in Supplementary Material, Section 4, Wang et al. (2021)). Since *Tarsius* is the only genus with the paraconid in this sample, we use SINATRA to perform three pairwise classification comparisons (*Tarsius* against *Saimiri*, *Mirza* and *Microcebus*, respectively) and assess SINATRA's ability to prioritize/detect the location of the paraconid as the region of interest (ROI). Based on our simulation studies, we run SINATRA with $c = 35$ cones, $d = 5$ directions per cone, a cap radius of $\theta = 0.25$ to generate each direction and $l = 75$ sublevel sets to compute topological summary statistics. In each comparison we evaluate the evidence for each vertex based on the first time that it appears in the reconstruction: this is the evidence potential for a vertex. A heatmap for each tooth provides visualization of the physical regions that are most

distinctive between the genera (Figure 7). In this figure we also display results from implementing the limit-shapes algorithm (Huang et al. (2019)) on the `auto3dgm` prealigned data as a baseline. Once again, for this method we use the normalized vector weights to determine the importance of each vertex in the data.

To assess the ability of the limit shapes and SINATRA to find *Tarsius*-specific paraconids, we use a null-based scoring method. We place a paraconid landmark on each *Tarsius* tooth and consider the $K = \{10, 50, 100, 150, 200\}$ nearest vertices surrounding the landmark's centermost vertex. This collection of $K + 1$ vertices defines our ROI. Within each ROI we weight the limit shape or SINATRA-computed evidence potentials by the surface area (or area of the Voronoi cell) encompassed by their corresponding vertices and then sum the scaled potentials together across the ROI vertices. This aggregated value, which we denote as τ^*, represents a score of association for the ROI. To construct a "null" distribution and assess the strength of any score τ^*, we randomly select $N = 500$ other "seed" vertices across the mesh of each *Tarsius* tooth and uniformly generate N-"null" regions that are K-vertices wide. We then compute similar (null) scores $\tau_1, \ldots, \tau_N$ for each randomly generated region. A "P-value"-like quantity (for the ith molar) is then generated by

$$P_i = \frac{1}{N+1} \sum_{t=1}^{N} \mathbb{I}\big(\tau_i^* \leq \tau_t\big), \tag{6}$$

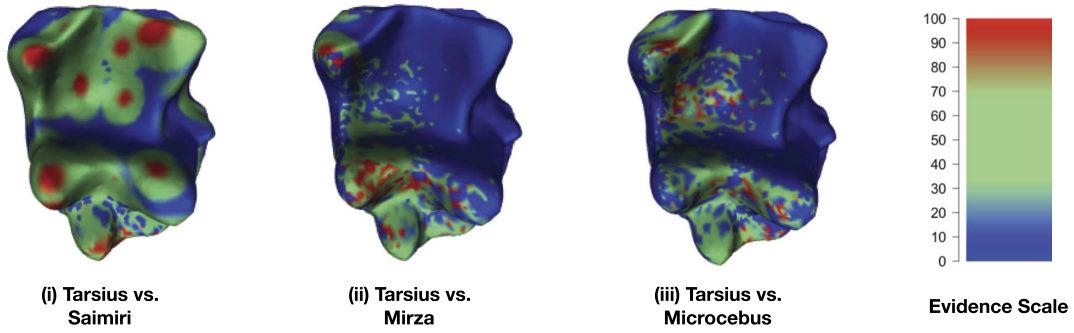
where $\mathbb{I}(\cdot)$ is an indicator function and a smaller P_i means more confidence in either method's ability to find the desired paraconid landmark. To ensure the robustness of this analysis, we generate the N-random null regions in two ways: (i) using a K-nearest neighbors (KNN) algorithm on each of the N-random seed vertices (Cover and Hart (2006)) or (ii) manually constructing K-vertex wide null regions with surface areas equal to that of the paraconid ROI (Supplementary Material, Section 5, Wang et al. (2021)). In both settings we take the median of the P_i values in equation (6) across all teeth and report them for each genus and choice of K combination; see the first half of Tables 1–2 for the SINATRA implementations and Supplementary Table 3 for results with limit shapes, respectively.

Using P-values as a direct metric of evidence can cause problems. For example, moving from $P = 0.03$ to $P = 0.01$ does not increase evidence for the alternative hypothesis (or against the null hypothesis) by a factor of 3. To this end, we use a calibration formula that transforms a P-value to a bound/approximation of a Bayes factor (BF) (Sellke, Bayarri and Berger (2001)), the ratio of the marginal likelihood under the alternative hypothesis H_1 vs. the null hypothesis H_0,

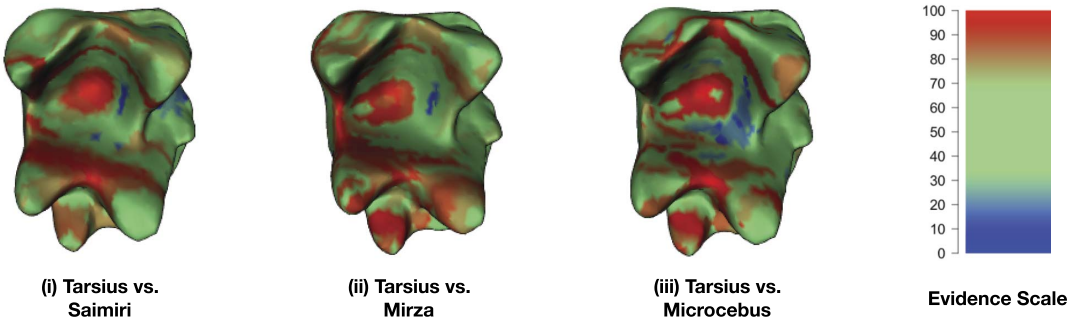
$$BF(P_i)_{10} = \big[-e\, P_i \log(P_i)\big]^{-1}, \tag{7}$$

for $P_i < 1/e$ and $BF(P_i)_{10}$ is an estimate of $\Pr(H_1|\mathcal{M})/\Pr(H_0|\mathcal{M})$, where $\mathcal{M}$ are the molars as meshes and H_0 and H_1 are the null and alternative hypotheses, respectively. The second half of Tables 1–2 and Supplementary Table 3 report these calibrated Bayes factor estimates.

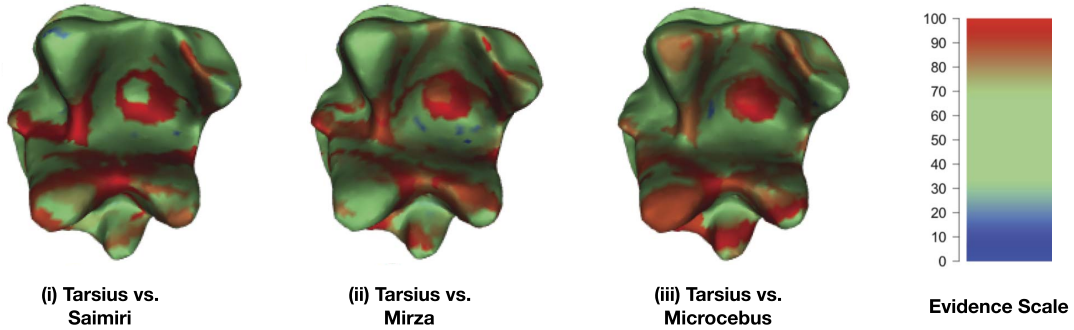
Overall, the limit-shapes algorithm performs generally well on these data; however, the sparsity in the resulting vertex weights hinders clear visual enrichment of the paraconid in the *Tarsius* molar (e.g., Figure 7(a)). As a reminder, there is the additional limitation of needing correctly specified point-by-point maps between each tooth in the data to achieve sufficient power. Based on our simulations, if these user inputs had been misspecified, the limit-shapes approach would have experienced an even more difficult time identifying the region of interest. When using SINATRA on the prealigned data, the paraconid ROI is most strongly enriched in the comparisons between the *Tarsius* and either of the strepsirrhine primates, rather than for the *Tarsius-Saimiri* comparison (e.g., Figure 7(b)). This trend remains consistent even when SINATRA is implemented on the raw data without correspondence maps



(a) Limit Shapes (Baseline) on `Autodgm` Aligned Teeth



(b) SINATRA on `Autodgm` Aligned Teeth



(c) SINATRA on ECT Aligned Teeth

FIG. 7. *Real data analysis aimed at detecting unique paraconids in molars belonging to primates in the Tarsius genus. We carry out three pairwise comparisons to analyze the physical differences between Tarsius molars and teeth from*: (*i*) *Saimiri*, (*ii*) *Mirza and* (*iii*) *Microcebus genus. Here, we consider the region of interest* (*ROI*) *to be the additional high cusp in the Tarsius molars* (*highlighted in red in Figure* 6). *In this figure we give an example of the mesh reconstruction from each class comparison using the Limit Shapes algorithm* (*Huang et al.* (2019)) *and SINATRA. Panels* (*a*) *and* (*b*) *show results on shapes that were prealigned with the* `auto3dgm` *software* (*Puente* (2013)). *Panel* (*c*) *illustrates results using SINATRA without any prior knowledge of pairwise correspondence maps between shapes in the data. In this experiment we first conduct the Euler characteristic transformation on the raw data for each tooth. Then, we implicitly normalize the meshes by aligning the EC curves via the approximate grid search procedure detailed in Section* 4 *of the Supplementary Material. Overall, results are consistent with the phylogeny of the primates as well as with our previous simulation studies. Genetically, Tarsius differ more from the Mirza and Microcebus genera, rather than from Saimiri. As a result, SINATRA finds the unique paraconid in the former two comparisons because of the appropriate genetic distance, rather than in the latter case where molar structures are more similar. This trend occurs whether or not we have prealigned data. The heatmaps in each panel display vertex evidence potential on a scale from* $[0-100]$. *A maximum of* 100 *represents the threshold at which the first shape vertex is reconstructed, while* 0 *denotes the threshold when the last vertex is reconstructed.*

TABLE 1

Null region experiment to evaluate SINATRA's ability to find paraconids in Tarsius molars using meshes with correspondence maps. Here, we assess how likely it is that SINATRA finds the region of interest (ROI) by chance. In this experiment, meshes were aligned using the `auto3dgm` *software (Puente (2013)). To produce the results above, we generate* 500 *"null" regions on each Tarsius tooth using: (i) a KNN algorithm and (ii) an equal-area approach (Supplementary Material, Section* 5*). Next, for each region we sum the evidence potential or "birth times" of all its vertices. We compare how many times the aggregate scores for the ROI is less than those for the null regions. The median of these "P-values" (P), and their corresponding calibrated Bayes factors (BF) when median $P < 1/e$, across all teeth are provided above for the three primate comparisons. Results with values P-values less than* 0.1 *and BFs greater than* 1.598 *are in bold.*

	Test	Region Size	*Tarsius* vs. *Saimiri*	*Tarsius* vs. *Mirza*	*Tarsius* vs. *Microcebus*
P-values	KNN	10	4.75×10^{-1}	3.39×10^{-1}	2.14×10^{-1}
		50	2.89×10^{-1}	2.10×10^{-1}	1.56×10^{-1}
		100	2.14×10^{-1}	$\mathbf{2.20 \times 10^{-2}}$	$\mathbf{6.19 \times 10^{-2}}$
		150	1.99×10^{-1}	$\mathbf{1.80 \times 10^{-2}}$	$\mathbf{6.59 \times 10^{-2}}$
		200	2.22×10^{-1}	$\mathbf{2.99 \times 10^{-2}}$	$\mathbf{9.18 \times 10^{-2}}$
	Equal-Area	10	3.21×10^{-1}	2.10×10^{-1}	1.84×10^{-1}
		50	2.81×10^{-1}	1.72×10^{-1}	1.26×10^{-1}
		100	2.40×10^{-1}	$\mathbf{4.39 \times 10^{-2}}$	$\mathbf{8.78 \times 10^{-2}}$
		150	2.59×10^{-1}	$\mathbf{3.79 \times 10^{-2}}$	$\mathbf{8.18 \times 10^{-2}}$
		200	2.55×10^{-1}	$\mathbf{4.39 \times 10^{-2}}$	$\mathbf{9.98 \times 10^{-2}}$
Bayes Factors	KNN	10	—	1.003	1.115
		50	1.025	1.122	1.269
		100	1.115	**4.381**	**2.136**
		150	1.145	**5.087**	**2.053**
		200	1.101	**3.505**	**1.678**
	Equal-Area	10	1.009	1.122	1.181
		50	1.031	1.215	1.409
		100	1.074	**2.681**	**1.722**
		150	1.051	**3.016**	**1.796**
		200	1.055	**2.681**	**1.599**

(e.g., Figure 7(c))—although we note that the enrichment of the ROI is not as distinct which is most likely due to shape alignments being slightly less precise with topological summary statistics vs. using known landmarks. We suspect that the difference in enrichment across primate comparisons is partly explained by the divergence times between the genera: *Tarsius* is more recently diverged from *Saimiri* than from the strepsirrhines. This conjecture is consistent with the intuition from our simulation studies, where classes of shapes with sufficiently different morphology result in more accurate identification of unique ROI. On the other hand, the *Tarsius-Saimiri* comparison is analogous to the simulations under to the null model: with too-similar molars, no region appears key to explaining the variance between the two classes of primates.

4. Discussion. In this paper we introduce SINATRA, the first statistical pipeline for submage analysis that does not require landmarks or correspondence points between images. We use simulations to demonstrate properties of SINATRA, and we illustrate the practical utility of SINATRA on real data. There are many potential extensions to our proposed framework. First, in its current formulation and software implementation SINATRA is limited to binary classification, but we believe that extensions to multi-class problems and regression with continuous responses are trivial. To analyze continuous traits and phenotypes in many evolutionary applications, one must first disentangle adaptation and heredity (Anderson, Willis and

TABLE 2

Null region experiment to evaluate SINATRA's ability to find paraconids in Tarsius molars using meshes without any prior knowledge of correspondence maps. Here, we assess how likely it is that the SINATRA algorithm finds the region of interest (ROI) by chance. In this experiment we first conduct the Euler characteristic transformation on the raw data for each tooth. Then, we implicitly normalize the collection of 3D meshes by aligning their EC curves; see the description of approximate grid search and alignment procedure detailed in Section 4 of the Supplementary Material. To produce the results in the table above, we generate 500 "null" regions on each Tarsius tooth using: (i) a KNN algorithm and (ii) an equal-area approach (Supplementary Material, Section 5). Next, for each region we sum the evidence potential or "birth times" of all its vertices. We compare how many times the aggregate scores for the ROI is less than those for the null regions. The median of these "P-values" (P) and their corresponding calibrated Bayes factors (BF) when median $P < 1/e$, across all teeth are provided above for the three primate comparisons. Results with values P-values less than 0.1 and BFs greater than 1.598 are in bold.

	Test	Region Size	*Tarsius* vs. *Saimiri*	*Tarsius* vs. *Mirza*	*Tarsius* vs. *Microcebus*
P-values	KNN	10	3.99×10^{-1}	6.09×10^{-1}	2.06×10^{-1}
		50	4.15×10^{-1}	3.23×10^{-1}	$\mathbf{9.58 \times 10^{-2}}$
		100	3.83×10^{-1}	1.72×10^{-1}	$\mathbf{7.59 \times 10^{-2}}$
		150	3.79×10^{-1}	1.16×10^{-1}	$\mathbf{8.38 \times 10^{-2}}$
		200	3.97×10^{-1}	1.52×10^{-1}	$\mathbf{1.00 \times 10^{-1}}$
	Equal-Area	10	2.97×10^{-1}	4.49×10^{-1}	1.54×10^{-1}
		50	3.37×10^{-1}	3.39×10^{-1}	1.20×10^{-1}
		100	3.87×10^{-1}	2.16×10^{-1}	$\mathbf{7.58 \times 10^{-2}}$
		150	4.43×10^{-1}	1.66×10^{-1}	$\mathbf{9.18 \times 10^{-2}}$
		200	4.57×10^{-1}	2.25×10^{-1}	$\mathbf{8.98 \times 10^{-2}}$
Bayes Factors	KNN	10	—	—	1.131
		50	—	1.008	**1.637**
		100	—	1.216	**1.882**
		150	—	1.474	**1.770**
		200	—	1.286	**1.598**
	Equal-Area	10	1.020	—	1.278
		50	1.004	1.003	1.447
		100	—	1.119	**1.881**
		150	—	1.235	**1.678**
		200	—	1.095	**1.700**

Mitchell-Olds (2011), Chen et al. (2012), Gienapp et al. (2008), Lai et al. (2019)). The standard approach for this disentanglement is to explicitly account for the hierarchy of descent by adding genetic covariance or kinship across species to the likelihood, either via phylogenetic regression (Graven (1989)) or linear mixed models (e.g., the animal model) (Henderson (1984)). Modeling covariance structures also arises in statistical and quantitative genetics applications where individuals are related (Kang et al. (2008, 2010), Zhou and Stephens (2012)). The SINATRA framework uses a Bayesian hierarchical model that is straightforward to adapt to analyze complex covariance structures in future work.

A second natural extension would be to apply the SINATRA framework to radiomic studies where the goal is to detail the correlation between clinical imaging features and genomic assays (Crawford et al. (2020)). This would require improving the efficacy of SINATRA's subimage selection capabilities in data settings where intraclass heterogeneity is high. For example, when studying the progression of glioblastoma multiforme (GBM)—a glioma that materializes into aggressive, cancerous tumor growths within the human brain—magnetic resonance images (MRIs) can vary greatly between patients harboring the same oncogenic mutations. As demonstrated in the current work, SINATRA is generally most powered in sce-

narios where just a few global features drive the variation between phenotypic classes (e.g., Figures 2–5 and Supplementary Figures 1–7). In the examples that we consider here, class definitions are derived from base shapes or governed by some rule of phylogeny. This keeps intraclass heterogeneity low and facilitates the ability of SINATRA to detect varying patterns between shape topologies. As a result, SINATRA is well suited for anthropologic-based applications and other similar domains. However, in the radiomics context it is possible that the signature of a given tissue type is made up of small focal lesions that are collectively important in explaining clinical outcomes. Moving forward, more work needs to be done to transfer the utility of SINATRA to cases where inter-class variation is driven by local fluctuations in shape morphology.

Code availability. Source code for replication is available in the Supplementary Materials (Wang et al. (2021)). A package for implementing the SINATRA pipeline is freely available at https://github.com/lcrawlab/SINATRA and is written in `R` (version 3.5.3). As part of this procedure: (i) inference for the Gaussian process classification (GPC) model using elliptical slice sampling was carried out using the `R` package `FastGP` (version 1.2) (Gopalan and Bornn (2015)) and (ii) the computation of effect sizes and association measures for the Euler characteristic curves was done with the "RelATive cEntrality (RATE)" source code in `R` (version 1.0.0; https://github.com/lorinanthony/RATE) (Crawford et al. (2019)). Visualizing the reconstructed regions outputted by SINATRA was done using the package `rgl` (version 0.100.19) (Adler, Nenadic and Zucchini (2003)) and general utility functions for triangular meshes from the package `Rvcg` (version 0.18) (Schlager et al. (2017)). Furthermore, preprocessing steps for the meshes examined in the study were performed using `Morpho` (version 2.60) (Schlager et al. (2017, 2018)) and `auto3dgm` (Version 1.00) (Puente (2013)).

Data availiability. The current study makes use of two real shape datasets. The first consists of Lemuridae teeth, a specific genera of Cercopithecidae (Old World monkeys; http://www.wisdom.weizmann.ac.il/~ylipman/CPsurfcomp/) (Boyer et al. (2011)). The second is comprised of mandibular molars from two different suborders of the primate: Haplorhini ("dry-nosed" primates; https://gaotingran.com/codes/codes.html) and Strepsirrhini ("moist-nosed" primates; http://morphosource.org/Detail/ProjectDetail/Show/project_id/89). From the first suborder we have 33 molars from the *Tarsius* (Boyer et al. (2011), Gao (2015, 2021)) and nine molars from the *Saimiri* (St Clair and Boyer (2016)) genera. In the second suborder, we have 11 molars from the *Microcebus* and six molars from the *Mirza* genera (Boyer et al. (2011), Gao (2015, 2021)). In the initial analysis the meshes of all teeth were aligned using `auto3dgm` (Puente (2013)). This algorithm establishes correspondences between uniformly placed landmarks on each tooth such that each mesh has the same orientation (e.g., Supplementary Figure 8). After alignment the molars were translated to be centered at the origin and normalized to be enclosed within a unit ball. In a secondary analyses we demonstrate how to implement SINATRA in settings where a priori correspondences between shapes are unavailable. Here, we first conduct the Euler characteristic transformation on the raw data for each tooth, and then we "implicitly" normalize the meshes by aligning the EC curves (e.g., Supplementary Figures 9 and 10). Details on the approximate grid search procedure used to do this is given in full detail in Section 4 of the Supplementary Material (Wang et al. (2021)). These `auto3dgm` and ECT quality controlled meshes were then used to demonstrate the utility of the SINATRA pipeline, respectively.

Acknowledgments. We would like to thank the Editor, Associate Editor and two anonymous referees for their constructive comments. We would also like to thank Ani Eloyan, Anthea Monod, Jenny Tung, Katharine Turner and Christine Wall for helpful conversations and suggestions. We are also very appreciative to Ruqi Huang and Maks Ovsjanikov for sharing their `MATLAB` implementation of limit shapes.

Funding. This research was partly supported by grants P20GM109035 (COBRE Center for Computational Biology of Human Disease; PI Rand) and P20GM103645 (COBRE Center for Central Nervous; PI Sanes) from the NIH NIGMS, 2U10CA180794-06 from the NIH NCI and the Dana Farber Cancer Institute (PIs Gray and Gatsonis), an Alfred P. Sloan Research Fellowship and a David & Lucile Packard Fellowship for Science and Engineering awarded to LC. A majority of this research was conducted using computational resources and services at the Center for Computation and Visualization (CCV), Brown University. SM would like to acknowledge partial funding from HFSP RGP005, NSF DMS 17-13012, NSF BCS 1552848, NSF DBI 1661386, NSF IIS 15-46331, NSF DMS 16-13261 as well as high-performance computing partially supported by grant 2016-IDG-1013 from the North Carolina Biotechnology Center. Lastly, TG was supported by NSF Grant No. DMS-1439786 while in residence at the Institute for Computational and Experimental Research in Mathematics (ICERM) in Providence, RI, during the Computer Vision Semester Program in Spring 2019 and partially by NSF DMS-1854831 afterward. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of any of the funders.

Author contributions statement. LC conceived the study. SM and LC developed the methods. BW, TS, and HK developed the algorithms and implemented the software. DB designed sampling strategy for the molar analysis. TG constructed the shape caricaturization and provided the baseline of correspondence-based methods. All authors performed the analyses, interpreted the results and wrote and revised the manuscript. BW, TS, and HK contributed equally to the study. Correspondences for the manuscript should be addressed to SM and LC.

Competing financial interests. The authors have declared that no competing interests exist.

SUPPLEMENTARY MATERIAL

Supplement to "A statistical pipeline for identifying physical features that differentiate classes of 3D shapes" (DOI: 10.1214/20-AOAS1430SUPPA; .pdf). This file contains supplementary derivations, figures, and tables referenced in the main text.

Source code for "A statistical pipeline for identifying physical features that differentiate classes of 3D shapes" (DOI: 10.1214/20-AOAS1430SUPPB; .zip). This file contains source code for the methods and analyses described in this paper.

REFERENCES

ADLER, D., NENADIC, O. and ZUCCHINI, W. (2003). RGL: An R-library for 3D visualization with OpenGL. In *Proceedings of the* 35*th Symposium of the Interface*: *Computing Science and Statistics*, *Salt Lake City* **35** 1–11.

ANDERSON, J. T., WILLIS, J. H. and MITCHELL-OLDS, T. (2011). Evolutionary genetics of plant adaptation. *Trends Genet*. **27** 258–266.

BELONGIE, S. (1999). Rodrigues' rotation formula. From MathWorld—A Wolfram Web Resource, created by Eric W. Weisstein. Available at http://mathworld.wolfram.com/RodriguesRotationFormula.html.

BENDICH, P., MARRON, J. S., MILLER, E., PIELOCH, A. and SKWERER, S. (2016). Persistent homology analysis of brain artery trees. *Ann. Appl. Stat.* **10** 198–218. MR3480493 https://doi.org/10.1214/15-AOAS886

BOYER, D. M., LIPMAN, Y., CLAIR, E. S., PUENTE, J., PATEL, B. A., FUNKHOUSER, T., JERNVALL, J. and DAUBECHIES, I. (2011). Algorithms to automatically quantify the geometric similarity of anatomical surfaces. *Proc. Natl. Acad. Sci. USA* **108** 18221–18226. https://doi.org/10.1073/pnas.1112822108

BOYER, D. M., PUENTE, J., GLADMAN, J. T., GLYNN, C., MUKHERJEE, S., YAPUNCICH, G. S. and DAUBECHIES, I. (2015). A new fully automated approach for aligning and comparing shapes. *Anat. Rec.* (*Hoboken*) **298** 249–276.

BOYER, D. M., GUNNELL, G. F., KAUFMAN, S. and MCGEARY, T. M. (2016). Morphosource: Archiving and sharing 3-D digital specimen data. *The Paleontological Society Papers* **22** 157–181.

CATES, J., ELHABIAN, S. and WHITAKER, R. (2017). Shapeworks: Particle-based shape correspondence and visualization software. In *Statistical Shape and Deformation Analysis* 257–298. Elsevier, Amsterdam.

CHAUDHURI, A., KAKDE, D., SADEK, C., GONZALEZ, L. and KONG, S. (2017). The mean and median criteria for kernel bandwidth selection for support vector data description. In *IEEE International Conference on Data Mining Workshops* (*ICDMW*), 2017 842–849.

CHEN, J., KÄLLMAN, T., MA, X., GYLLENSTRAND, N., ZAINA, G., MORGANTE, M., BOUSQUET, J., ECKERT, A., WEGRZYN, J. et al. (2012). Disentangling the roles of history and local selection in shaping clinal variation of allele frequencies and gene expression in Norway spruce (Picea abies). *Genetics* **191** 865–881.

CHENG, L., RAMCHANDRAN, S., VATANEN, T., LIETZÉN, N., LAHESMAA, R., VEHTARI, A. and LÄHDESMÄKI, H. (2019). An additive Gaussian process regression model for interpretable non-parametric analysis of longitudinal data. *Nat. Commun.* **10** 1798.

COVER, T. and HART, P. (2006). Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **13** 21–27.

CRAWFORD, L., ZENG, P., MUKHERJEE, S. and ZHOU, X. (2017). Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. *PLoS Genet.* **13** e1006869.

CRAWFORD, L., WOOD, K. C., ZHOU, X. and MUKHERJEE, S. (2018). Bayesian approximate Kernel regression with variable selection. *J. Amer. Statist. Assoc.* **113** 1710–1721. MR3902240 https://doi.org/10.1080/01621459.2017.1361830

CRAWFORD, L., FLAXMAN, S. R., RUNCIE, D. E. and WEST, M. (2019). Variable prioritization in nonlinear black box methods: A genetic association case study. *Ann. Appl. Stat.* **13** 958–989. MR3963559 https://doi.org/10.1214/18-AOAS1222

CRAWFORD, L., MONOD, A., CHEN, A. X., MUKHERJEE, S. and RABADÁN, R. (2020). Predicting clinical outcomes in glioblastoma: An application of topological and functional data analysis. *J. Amer. Statist. Assoc.* **115** 1139–1150. MR4143455 https://doi.org/10.1080/01621459.2019.1671198

CROMPTON, R. H., SAVAGE, R. and SPEARS, I. R. (1998). The mechanics of food reduction in Tarsius bancanus. Hard-object feeder, soft-object feeder or both? *Folia Primatol.* (*Basel*) **69 Suppl 1** 41–59. https://doi.org/10.1159/000052698

CURRY, J., MUKHERJEE, S. and TURNER, K. (2019). How many directions determine a shape and other sufficiency results for two topological transforms. Available at arXiv:1805.09782.

DUPUIS, P., GRENANDER, U. and MILLER, M. I. (1998). Variational problems on flows of diffeomorphisms for image matching. *Quart. Appl. Math.* **56** 587–600. MR1632326 https://doi.org/10.1090/qam/1632326

FASY, B. T., MICKA, S., MILLMAN, D. L., SCHENFISCH, A. and WILLIAMS, L. (2018). Challenges in reconstructing shapes from Euler characteristic curves. Available at arXiv:1811.11337.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.

GAO, T. (2015). Hypoelliptic diffusion maps and their applications in automated geometric morphometrics, Ph.D. thesis, Duke Univ.

GAO, T. (2021). The diffusion geometry of fibre bundles: Horizontal diffusion maps. *Appl. Comput. Harmon. Anal.* **50** 147–215. MR4162316 https://doi.org/10.1016/j.acha.2019.08.001

GAO, T., KOVALSKY, S. Z. and DAUBECHIES, I. (2019). Gaussian process landmarking on manifolds. *SIAM J. Math. Data Sci.* **1** 208–236. MR3949707 https://doi.org/10.1137/18M1184035

GAO, T., KOVALSKY, S. Z., BOYER, D. M. and DAUBECHIES, I. (2019). Gaussian process landmarking for three-dimensional geometric morphometrics. *SIAM J. Math. Data Sci.* **1** 237–267. MR3949708 https://doi.org/10.1137/18M1203481

GHRIST, R., LEVANGER, R. and MAI, H. (2018). Persistent homology and Euler integral transforms. *J. Appl. Comput. Topol.* **2** 55–60. MR3873179 https://doi.org/10.1007/s41468-018-0017-1

GIENAPP, P., TEPLITSKY, C., ALHO, J. S., MILLS, J. A. and MERILA, J. (2008). Climate change and evolution: Disentangling environmental and genetic responses. *Mol. Ecol.* **17** 167–178.

GOPALAN, G. and BORNN, L. (2015). FastGP: An R package for Gaussian processes. Available at arXiv:1507.06055.

GOSWAMI, A. (2015). Phenome10K: A free online repository for 3-D scans of biological and palaeontological specimens. Website.

GRAVEN, A. (1989). The phylogenetic regression. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **326** 87–99.

GUATELLI-STEINBERG, D. (2003). Primate dentition: An introduction to the teeth of non-human primates. *Am. J. Phys. Anthropol.* **121** 189–189.

HECKERMAN, D., GURDASANI, D., KADIE, C., POMILLA, C., CARSTENSEN, T., MARTIN, H., EKORU, K., NSUBUGA, R. N., SSENYOMO, G. et al. (2016). Linear mixed model for heritability estimation that explicitly addresses environmental variation. *Proc. Natl. Acad. Sci. USA* **113** 7377–7382.

HENDERSON, C. R. (1984). *Applications of Linear Models in Animal Breeding*. Guelph, Ont.: University of Guelph. Includes index.

HONG, Y., GOLLAND, P. and ZHANG, M. (2017). Fast geodesic regression for population-based image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 317–325. Springer, Berlin.

HUANG, R., ACHLIOPTAS, P., GUIBAS, L. and OVSJANIKOV, M. (2019). Limit shapes–A tool for understanding shape differences and variability in 3D model collections. *Comput. Graph. Forum* **38** 187–202.

JIANG, Y. and REIF, J. C. (2015). Modeling epistasis in genomic selection. *Genetics* **201** 759–768.

KANG, H. M., ZAITLEN, N. A., WADE, C. M., KIRBY, A., HECKERMAN, D., DALY, M. J. and ESKIN, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics* **178** 1709–1723.

KANG, H. M., SUL, J. H., SERVICE, S. K., ZAITLEN, N. A., KONG, S.-Y., FREIMER, N. B., SABATTI, C. and ESKIN, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42** 348–354.

KENDALL, D. G. (1989). A survey of the statistical theory of shape. *Statist. Sci.* **4** 87–120. MR1007558

LAI, Y.-T., YEUNG, C. K. L., OMLAND, K. E., PANG, E.-L., HAO, Y., LIAO, B.-Y., CAO, H.-F., ZHANG, B.-W., YEH, C.-F. et al. (2019). Standing genetic variation as the predominant source for adaptation of a songbird. *Proc. Natl. Acad. Sci. USA* **116** 2152–2157.

LI, F., ZHANG, T., WANG, Q., GONZALEZ, M. Z., MARESH, E. L. and COAN, J. A. (2015). Spatial Bayesian variable selection and grouping for high-dimensional scalar-on-image regression. *Ann. Appl. Stat.* **9** 687–713. MR3371331 https://doi.org/10.1214/15-AOAS818

MCDONALD, K. R., BRODERICK, W. F., HUETTEL, S. A. and PEARSON, J. M. (2019). Bayesian nonparametric models characterize instantaneous strategies in a competitive dynamic game. *Nat. Commun.* **10** 1808.

MILLER, E. (2015). Fruit flies and moduli: Interactions between biology and mathematics. *Notices Amer. Math. Soc.* **62** 1178–1184. MR3408062 https://doi.org/10.1090/noti1290

NEAL, R. M. (1997). Monte Carlo implementation of Gaussian process models for Bayesian regression andMonte Carlo implementation of Gaussian process models for Bayesian regression and classification. Technical Report No. 9702, Dept. of Statistics, Univ. Toronto.

NEAL, R. M. (1999). Regression and classification using Gaussian process priors. In *Bayesian Statistics*, 6 (*Alcoceber*, 1998) 475–501. Oxford Univ. Press, New York. MR1723510

NICKISCH, H. and RASMUSSEN, C. E. (2008). Approximations for binary Gaussian process classification. *J. Mach. Learn. Res.* **9** 2035–2078. MR2452620

OUDOT, S. and SOLOMON, E. (2018). Inverse problems in topological persistence. Available at arXiv:1810.10813.

OVSJANIKOV, M., BEN-CHEN, M., SOLOMON, J., BUTSCHER, A. and GUIBAS, L. (2012). Functional maps: A flexible representation of maps between shapes. *ACM Trans. Graph.* **31** 30:1–30:11.

PILLAI, N. S., WU, Q., LIANG, F., MUKHERJEE, S. and WOLPERT, R. L. (2007). Characterizing the function space for Bayesian kernel models. *J. Mach. Learn. Res.* **8** 1769–1797. MR2332448

POZZI, L., HODGSON, J. A., BURELLA, A. S., RAAUMB, R. L. and DISOTELL, T. R. (2014). Primate phylogenetic relationships and divergence dates inferred from complete mitochondrial genomes. *Mol. Phylogenet. Evol.* **75** 165–183.

PUENTE, J. (2013). Distances and algorithms to compare sets of shapes for automated biological morphometrics, Ph.D. thesis, Princeton Univ., Princeton, NJ.

RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR2514435

RODRIGUEZ-NIEVA, J. F. and SCHEURER, M. S. (2019). Identifying topological order through unsupervised machine learning. *Nat. Phys.*

RUSTAMOV, R. M., OVSJANIKOV, M., AZENCOT, O., BEN-CHEN, M., CHAZAL, F. and GUIBAS, L. (2013). Map-based exploration of intrinsic shape differences and variability. *ACM Trans. Graph.* **32** 1–12.

SCHLAGER, S., ZHENG, G., LI, S. and SZÉKELY, G. (2017). Morpho and Rvcg—Shape analysis in R: R-packages for geometric morphometrics, shape analysis and surface manipulations. In *Statistical Shape and Deformation Analysis*: *Methods*, *Implementation and Applications* 217–256. Academic Press, San Diego.

SCHLAGER, S., PROFICO, A., DI VINCENZO, F. and MANZI, G. (2018). Retrodeformation of fossil specimens based on 3D bilateral semi-landmarks: Implementation in the R package "Morpho". *PLoS ONE* **13** e0194073.

SCHÖLKOPF, B., HERBRICH, R. and SMOLA, A. J. (2001). A generalized representer theorem. In *Computational Learning Theory* (*Amsterdam*, 2001). *Lecture Notes in Computer Science* **2111** 416–426. Springer, Berlin. MR2042050 https://doi.org/10.1007/3-540-44581-1_27

SELA, M., AFLALO, Y. and KIMMEL, R. (2015). Computational caricaturization of surfaces. *Comput. Vis. Image Underst.* **141** 1–17.

SELLKE, T., BAYARRI, M. J. and BERGER, J. O. (2001). Calibration of p values for testing precise null hypotheses. *Amer. Statist.* **55** 62–71. MR1818723 https://doi.org/10.1198/000313001300339950

SIMON, N., FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2013). A sparse-group lasso. *J. Comput. Graph. Statist.* **22** 231–245. MR3173712 https://doi.org/10.1080/10618600.2012.681250

SINGLETON, K. R., CRAWFORD, L., TSUI, E., MANCHESTER, H. E., MAERTENS, O., LIU, X., LIBERTI, M. V., MAGPUSAO, A. N., STEIN, E. M. et al. (2017). Melanoma therapeutic strategies that select against resistance by exploiting MYC-driven evolutionary convergence. *Cell Rep.* **21** 2796–2812.

ST CLAIR, E. M. and BOYER, D. M. (2016). Lower molar shape and size in prosimian and platyrrhine primates. *Am. J. Phys. Anthropol.* **161** 237–258.

SWAIN, P. S., STEVENSON, K., LEARY, A., MONTANO-GUTIERREZ, L. F., CLARK, I. B. N., VOGEL, J. and PILIZOTA, T. (2016). Inferring time derivatives including cell growth rates using Gaussian processes. *Nat. Commun.* **7** 13766.

TURNER, K., MUKHERJEE, S. and BOYER, D. M. (2014). Persistent homology transform for modeling shapes and surfaces. *Inf. Inference* **3** 310–344. MR3311455 https://doi.org/10.1093/imaiai/iau011

WANG, B., SUDIJONO, T., KIRVESLAHTI, H., GAO, T., BOYER, D. M., MUKHERJEE, S. and CRAWFORD, L. (2021). Supplement to "A statistical pipeline for identifying physical features that differentiate classes of 3D shapes." https://doi.org/10.1214/20-AOAS1430SUPPA.

WANG, B., SUDIJONO, T., KIRVESLAHTI, H., GAO, T., BOYER, D. M., MUKHERJEE, S. and CRAWFORD, L. (2021). Source Code for "A statistical pipeline for identifying physical features that differentiate classes of 3D shapes." https://doi.org/10.1214/20-AOAS1430SUPPB.

WILLIAMS, C. K. I. and BARBER, D. (1998). Bayesian classification with Gaussian processes. *IEEE Trans. Pattern Anal. Mach. Intell.* **20** 1342–1351.

WORSLEY, K. J. (1995). Estimating the number of peaks in a random field using the Hadwiger characteristic of excursion sets, with applications to medical images. *Ann. Statist.* **23** 640–669. MR1332586 https://doi.org/10.1214/aos/1176324540

YANG, Y. and ZOU, H. (2015). A fast unified algorithm for solving group-lasso penalize learning problems. *Stat. Comput.* **25** 1129–1141. MR3401877 https://doi.org/10.1007/s11222-014-9498-5

ZHANG, Z., DAI, G. and JORDAN, M. I. (2011). Bayesian generalized kernel mixed models. *J. Mach. Learn. Res.* **12** 111–139. MR2773550

ZHOU, X. and STEPHENS, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44** 821–825.

ZHU, X. and STEPHENS, M. (2018). Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. *Nat. Commun.* **9** 4361.

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327 https://doi.org/10.1111/j.1467-9868.2005.00503.x