SINATRA: A Sub-Image Analysis Pipeline for Selecting Features that Differentiate Classes of 3D Shapes

Bruce Wang[1,*], Timothy Sudijono[2,*], Henry Kirveslahti[3,*], Tingran Gao[4], Douglas M. Boyer[5], Sayan Mukherjee[3,6-8], and Lorin Crawford[9-11,†]

1 Data Science Initiative, Brown University, Providence, RI, USA
2 Division of Applied Mathematics, Brown University, Providence, RI, USA
3 Department of Statistical Science, Duke University, Durham, NC, USA
4 Committee on Computational and Applied Mathematics, Department of Statistics, University of Chicago, Chicago, IL, USA
5 Department of Evolutionary Anthropology, Duke University, Durham, NC, USA
6 Department of Computer Science, Duke University, Durham, NC, USA
7 Department of Mathematics, Duke University, Durham, NC, USA
8 Department of Bioinformatics & Biostatistics, Duke University, Durham, NC, USA
9 Department of Biostatistics, Brown University, Providence, RI, USA
10 Center for Statistical Sciences, Brown University, Providence, RI, USA
11 Center for Computational Molecular Biology, Brown University, Providence, RI, USA

*** Authors Contributed Equally**
† **Corresponding E-mail: lorin_crawford@brown.edu**

Abstract

It has been a longstanding challenge in geometric morphometrics and medical imaging to infer the physical locations (or regions) of 3D shapes that are most associated with a given response variable (e.g. class labels) without needing common predefined landmarks across the shapes, computing correspondence maps between the shapes, or requiring the shapes to be diffeomorphic to each other. In this paper, we introduce SINATRA: the first statistical pipeline for sub-image analysis which identifies physical shape features that explain most of the variation between two classes without the aforementioned requirements. We also illustrate how the problem of 3D sub-image analysis can be mapped onto the well-studied problem of variable selection in nonlinear regression models. Here, the key insight is that tools from integral geometry and differential topology, specifically the Euler characteristic, can be used to transform a 3D mesh representation of an image or shape into a collection of vectors with minimal loss of geometric information. Crucially, this transform is invertible. The two central statistical, computational, and mathematical innovations of our method are: (1) how to perform robust variable selection in the transformed space of vectors, and (2) how to pullback the most informative features in the transformed space to physical locations or regions on the original shapes. We highlight the utility, power, and properties of our method through detailed simulation studies, which themselves are a novel contribution to 3D image analysis. Finally, we apply SINATRA to a dataset of mandibular molars from four different genera of primates and demonstrate the ability to identify unique morphological properties that summarize phylogeny.

Significance

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 pre-specified landmarks (that are often expert-derived). We present 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 molars from suborders of primates and demonstrate its ability recover known morphometric variation across phylogenies.

Introduction

Sub-image analysis is an important, yet open, problem in both medical imaging studies and geometric morphometric applications. One statement of this problem is, given two classes of 3D images or shapes (e.g. computed tomography (CT) scans of bones or magnetic resonance images (MRI) of different tissues), which physical features on the shapes are most important to defining a particular class label. More generally, the sub-image analysis problem can be framed as a regression-based task, where one is given a collection of shapes and the goal is to find the properties that explain the greatest variation in some response variable (continuous or binary). For example, one may be interested in identifying the structures of glioblastoma tumors that best indicate signs of potential relapse and other clinical outcomes [1]. From a statistical perspective, the sub-image selection problem is therefore directly related to the variable selection problem — given high-dimensional covariates and a univariate outcome, we want to infer which of the variables are most relevant in explaining or predicting variation in the observed response.

There are several challenges in framing sub-image analysis as a regression. The first challenge centers around representing a 3D object as a (square integrable) covariate or feature vector. The desired transformation should have minimal loss in geometric information and should also be applicable to a wide range of shape and imaging datasets. In this paper, we will use a tool from integral geometry and differential topology called the Euler characteristic (EC) transform [1–4], which sufficiently maps shapes into vectors without requiring pre-specified landmark points or pairwise correspondences. This property will be central to our innovations. Once we are given a vector representation of the shape, the second challenge in framing sub-image analysis as a regression-based problem is quantifying which topological features are most relevant in explaining variation in a continuous outcome or binary class label. This is the classic take on variable selection which we address using a Bayesian regression model and an information theoretic metric to measure the relevance of each topological feature. Importantly, our Bayesian method allows us to perform variable selection for nonlinear functions — again, we will discuss the importance of this requirement later. The last challenge deals with how to interpret the most informative topological features obtained by our variable selection methodology. An important property of the EC transform is that it is invertible; thus, we can take the most informative topological features and naturally recover the physical regions on the shape that are most informative. In this paper, we introduce SINATRA: a unified statistical pipeline for sub-image analysis that addresses each of these challenges and is the first sub-image 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 [5], (ii) diffeomorphism-based representations [6], and (iii) representations that use integral geometry and excursions of random fields [7]. The main idea behind landmark-based analysis is that there are points on shapes that are known to be in correspondence with each other. As a result, any shape can be represented as a collection of 3D coordinates. The shortcoming with landmark-based approaches is twofold. First, many modern datasets are not defined by landmarks; instead, they are consist of 3D CT scans [8, 9]. Second, reducing these detailed mesh data to simple landmarks often results in a great deal of information loss. Alternatively, diffeomorphism-based approaches have bypassed the need for landmarks. There has also been a great deal of progress in developing tools that efficiently compare the similarity between shapes in large databases via algorithms that continuously deform one shape into another [10–14]. Unfortunately, these methods require that shapes be diffeomorphic: a continuous transformation between two shapes that places them in correspondence. There are many applications where shapes and images cannot be placed in correspondence because of qualitative differences. For example, in

a dataset of fruit fly wings, some mutants may have extra lobes of veins [15]; or, in a dataset of brain arteries, many of the arteries cannot be continuously mapped to each other [16]. Indeed, in large databases such as the MorphoSource [9], the CT scans of skulls across many clades will not be diffeomorphic. Thus, there is a real need for 3D image analysis methods that do not require correspondences.

In previous work [2], two topological transformations for shapes were introduced: the persistent homology (PH) transform and the EC transform were introduced. 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 [3, 4]. Detailed mathematical analyses have also been provided [3]. Most relevant to our approach, in this paper, is a nonlinear regression framework which uses the EC transform to predict outcomes of disease free survival in glioblastoma [1]. The two major takeaways from this work is 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, other notable contributions of this paper include software packaging to implement our approach, and a detailed design of rigorous simulation studies that may be used to assess the accuracy of sub-image selection methods. The freely available software comes with several built-in capabilities that are integral to sub-image analyses in both biomedical studies and geometric morphometric applications. First, and foremost, SINATRA does not require landmarks or correspondences in the data. Second, given a dataset of normalized and axis aligned 3D images, SINATRA will output evidence measures that highlight the physical regions of shapes that are most variable between two predefined classes. There are many applications where users may suspect *a priori* that certain landmarks may vary across groups of shapes (e.g. via the literature). To this end, SINATRA also provides notions of statistical "significance" for any region of interest (ROI) by computing p-values and Bayes factor estimates that effectively detail how likely it is to be informative by chance [17].

Throughout the rest of the paper, we will describe each mathematical step of the SINATRA pipeline, and demonstrate its power and utility via simulations. We will 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.

Results

SINATRA Pipeline Overview

The SINATRA pipeline generally implements four key steps (Fig. 1). First, the geometry of 3D shapes (represented as triangular meshes) is summarized 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 variables, is used to classify 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 a notion of evidence that a given topological feature is associated with a particular class. Fourth, the topological features are iteratively mapped back onto the original shapes (in rank order according to their association measures) via a reconstruction algorithm. This allows us to highlight 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 SI Appendix.

Step One: Topological Summary Statistics for 3D 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 [1–4]. Briefly, for a mesh $\mathcal{M}$, the Euler characteristic is one of the accessible topological invariants derived using the following summation

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

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 $\chi_\nu(\mathcal{M})$ tracks the change in the Euler characteristic, with respect to a given filtration of length l in direction ν (Figs. 1(a) and (b)). Mathematically, this is done by first specifying a height function $h_\nu(\boldsymbol{x}) = \boldsymbol{x}^\intercal \nu$ for vertex $\boldsymbol{x} \in M$ in direction ν. We then use this height function to define sublevel sets (or subparts) of the mesh $\mathcal{M}_\nu^a$ in direction ν, where $h_\nu(\boldsymbol{x}) \leq a$. The EC curve is simply $\chi(\mathcal{M}_\nu^a)$ over a range of l filtration steps over a (Fig. 1(b)). The EC transform is the collection of EC curves across a set of directions $\nu = 1, \ldots, m$, and effectively 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 to be 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. Note that notions of sufficiency, stating the m number of directions and the l range of sublevel set values required for the EC transform to preserve all information for a family of shapes, have been previously provided [3]. In this paper, we will use simulations to outline empirical procedures and develop intuition behind these quantities.

Step Two: 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 [18–22]

$$\mathbf{y} \sim \mathcal{B}(\boldsymbol{\pi}), \qquad \mathrm{g}(\boldsymbol{\pi}) = \Phi^{-1}(\boldsymbol{\pi}) = \boldsymbol{f}, \qquad \boldsymbol{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}), \tag{2}$$

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 $\Phi(\cdot)$ being 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 accomplish this objective, we use kernel regression where the utility of generalized nonparametric statistical models is well-established due their ability to account for various complex data structures [23–28]. Generally, kernel methods posit that $\boldsymbol{f}$ lives within a reproducing kernel Hilbert space (RKHS) defined by some (nonlinear) covariance function that implicitly account for higher-order interactions between features, leading to more complete classifications of data [29–31]. To this end, we assume $\boldsymbol{f}$ to be normally distributed with mean vector $\mathbf{0}$, and covariance matrix $\mathbf{K}$ 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 [32]. The full model specified in Equation (2) is commonly referred to as "Gaussian process classification" or GPC.

Step Three: Feature (Variable) Selection. To estimate the model in Equation (2), we use an elliptical slice sampling Markov chain Monte Carlo (MCMC) algorithm (SI Appendix Section 1.1). This enables samples to be taken 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 [33–35]

$$\boldsymbol{\beta} = (\mathbf{X}^\intercal \mathbf{X})^\dagger \mathbf{X}^\intercal \boldsymbol{f}, \tag{3}$$

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 ordinary least squares. SINATRA uses these weights and assigns a measure of relative centrality to each summary statistic (first panel Fig. 1(c)) [35]. Specifically, this criterion evaluates how much information in classifying each shape is lost when a particular topological feature is removed from the model. This 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 j-th topological feature being set to zero, and (ii) the marginal posterior distribution $p(\boldsymbol{\beta}_{-j})$ with the effects of the j-th feature being integrated out. Namely,

$$\mathrm{KLD}(\beta_j) = \int_{\boldsymbol{\beta}_{-j}} \log\left(\frac{p(\boldsymbol{\beta}_{-j})}{p(\boldsymbol{\beta}_{-j} \,|\, \beta_j = 0)}\right) p(\boldsymbol{\beta}_{-j})\, \mathrm{d}\boldsymbol{\beta}_{-j} \qquad j = 1, \ldots, p. \tag{4}$$

which has a closed form solution when the posterior distribution of the effect sizes is assumed to be (approximately) Gaussian (SI Appendix 1.2). Finally, we normalize to obtain an association metric for each topological feature, $\gamma_j = \mathrm{KLD}(\beta_j)/\sum \mathrm{KLD}(\beta_l)$. There are two main takeaways from this formulation. First, the KLD is a non-negative quantity, and equals zero if and only if the posterior distribution of $\boldsymbol{\beta}_{-j}$ is independent of the effect β_j. Intuitively, this is equivalent to saying that removing an unimportant shape feature will have no impact on explaining the variance between shape classes. The second key takeaway is that $\boldsymbol{\gamma}$ is bounded on the unit interval $[0, 1]$, with the natural interpretation of providing relative evidence of association for shape features; higher values suggesting 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 indeed more central to this explanation than others [35]. As we will show in the coming sections, when the null assumption is met, SINATRA will display association results that are appear uniformly distributed and effectively indistinguishable.

Step Four: Reconstruction. After obtaining association measures for each topological feature, we map this information back onto the physical shape (second panel Fig. 1(c) and 1(d)). We refer to this process as *reconstruction*, as this procedure recovers regions that explain the most variation between shape classes (SI Appendix Section 1.3). Intuitively, the goal is to identify vertices on the shape that correspond to topological features with the greatest association measures. Begin by considering d directions all 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 — meaning, 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 $\mathcal{R} := \bigcup_i \mathcal{Z}_i$. The choice to use cones is motivated by the idea that vectors of Euler characteristics taken along directions close together will express comparable information, allowing us to leverage findings between them and increase our power of detecting truly associated shape vertices and regions — this as opposed to antipodal directions where the lack of shared information may do harm when determining reconstructed manifolds (SI Appendix Section 1.4) [3, 36, 37].

Visualization of Enrichment. Once shapes have been reconstructed, we can visualize the relative importance or "evidence potential" for each vertex on the mesh. This is computed using the following simple procedure. First, we sort the topological features from largest to smallest, in descending order, according to their association measures $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_p$. Next, we iteratively move through the sorted measures $T_k = \gamma_k$ (starting with $k = 1$), and we reconstruct the vertices corresponding to the topological features in the set $\{j : \gamma_j \geq T_k\}$. The evidence potential for each vertex is then defined as the largest threshold T_k at which it is reconstructed for the first time. Here, the key intuition is that 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 (Fig. 1(d)). For consistency across different applications and case studies, we set the coloring of these heatmaps to be on a scale from $[0 - 100]$. Here, 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 will appear to be born relatively early and at the same time. This will not be the case under the alternative.

Algorithmic Overview and Implementation. Software for implementing the steps in the SINATRA pipeline is carried out in `R` code, which is freely available at `https://github.com/lcrawlab/SINATRA`. This algorithm requires the following inputs: (i) axis 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 (which we usually set to $1/2$ since we work with meshes normalized to the unit ball); (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. In the next two sections, we discuss strategies for how to choose values for the free parameters through simulation studies. A table detailing scalability for the current algorithmic implementation can be found in SI Appendix (see Table S1).

Simulation Study: Perturbed Spheres

We begin with a simple proof-of-concept simulation study to demonstrate the power of our proposed pipeline and illustrate how different parameter value choices will affect its ability to detect truly associated features on 3D shapes. To do so, we take 100 spheres and perturb regions on their surfaces to create two equally sized classes. This is done by using the following two-step procedure:

- First, we generate a fixed number of (approximately) equidistributed points on each sphere: some number u regions to be shared across classes, and the remaining v regions to be unique to class assignment.
- Second, within each region, we perturb the k closest vertices $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k\}$ by a pre-specified scale factor α and add some random normally distributed noise $\epsilon_i \sim \mathcal{N}(0, 1)$. Formally, this specified as $\boldsymbol{x}_i^* := \boldsymbol{x}_i\alpha + \epsilon_i$ for $i = 1, \ldots, k$.

We consider three scenarios based on the number of shared and unique regions between shape classes (Figs. 2(a)-2(c)). Specifically, 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-wide. Intuitively, each sequential scenario represents an increase in degree of difficulty. Class-specific regions should be harder to identify in shapes with more complex structures. We analyze fifty different simulated datasets for each of the three scenarios. In each simulated dataset, only 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) (SI Appendix Section 2). 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. Here, we iteratively vary each parameter across a wide range of appropriate values, while holding the others at fixed constants $\{c = 25, d = 5, \theta = 0.15, l = 30\}$. Figures displayed in the main text are based on varying the number of cones (Figs. 2(d)-2(f)), while results for the other sensitivity analyses can be found in SI Appendix (Figs. S1-S3).

As expected, SINATRA's ability to detect associated regions depends on the proportion of shape class variance $\mathbb{V}(\mathbf{y})$ that is explained by each of the corresponding associated vertices. More specifically, the algorithm's performance is consistently better when shapes are defined by just a few prominent

regions (e.g. scenario I) versus when shape definitions are more complex (e.g. scenarios II and III). This is because, in the former setting, associated vertices make greater individual-level contributions to the overall variance between classes (i.e. $\mathbb{V}(\mathbf{y})/10 > \mathbb{V}(\mathbf{y})/30 > \mathbb{V}(\mathbf{y})/50$). Note that similar trends in performance have been shown during the assessment of high-dimensional variable selection methods in other application areas [38–40].

This simulation study also allows us to demonstrate the general behavior and effectiveness of the SINATRA algorithm as a function of different choices for its free input parameters. First, we assess what happens to our power when we adjust the number of cones of directions used to compute Euler characteristic curves. The key takeaway for this parameter is that 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 (Figs. 2(d)-2(f)). This supports the intuition that seeing more of a shape leads to an improved ability to understand its complete structure [1–3]. Our empirical results show that this can be achieved by summarizing the shapes with filtrations taken over multiple directions. As a result, in practice, we suggest specifying multiple cones $c > 1$ and utilizing multiple directions d per cone (see monotonically increasing power in Fig. S1). While the other two parameters 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 afterwards (Fig. S2). This supports previous theoretical results that state cones should be defined by directions that are in close proximity to each other [3]; but not too close such that they effectively explain the same local information with little variation. Lastly, and perhaps most importantly, is understanding the performance of the algorithm as a function of the number of sublevel sets l (i.e. the number of steps in the filtration) used to compute Euler characteristic curves. As we will show in the next section, this depends on the types of shapes being analyzed. Intuitively, for very intricate shapes, coarse filtrations with too few sublevel sets will 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 this information (Fig. S3); however, this is less important when only a few features decide how shapes are defined (e.g. scenario I). To this end, 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 the null assumptions of the SINATRA pipeline are met (Fig. S4). Recall that, 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 or more central than the others. Here, we generate synthetic shapes under the two cases when SINATRA will fail 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 distinctively stand out as being important (Fig. S4(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 central or key to explaining the observed variation (Fig. S4(b)).

Simulation Study: Caricatured Shapes

We further assess the SINATRA pipeline using a second simulation study where we modify computed tomography (CT) scans of real Lemuridae teeth (one of the five families of Strepsirrhini primates commonly known as lemurs) [41] using a well-known caricaturization procedure [42]. Briefly, we fix the triangular mesh of an individual tooth and specify class-specific regions centered around expert-derived biological landmarks (Fig. 3) [10]. For each triangular face contained within a class-specific region, we multiply a corresponding affine transformation by a positive scalar that smoothly varies on the triangular mesh and attains maximum value at the biological landmark used to define the region (SI Appendix Section 3).

We caricature 50 different teeth according to the following procedure (Fig. 3(a)):

- First, we take the expert-derived landmarks for a given tooth, and assign v of them to be specific to one class and v' to be specific to the other class.
- Second, we perform the caricaturization where each face in the v and v' class-specific regions is multiplied by a positive scalar (i.e. exaggerated or enhanced). This is repeated twenty-five times (with some small noise per replicate) to create two equally-sized classes of 25 shapes.

Here, we explore two scenarios by varying the number of class-specific landmarks v and v' that determine the caricaturization in each class. In the first, we set both $v, v' = 3$; while, in the second, we fix $v, v' = 5$. As in the previous 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 (SI Appendix Section 2), and we again assess this estimate of power as a function of the algorithm's free parameter inputs. While varying each parameter, we hold the others at fixed constants $\{c = 15, d = 5, \theta = 0.15, l = 50\}$. Figures described in the main text are based on varying the number of cones (Figs. 3(b) and 3(c)), and results for the other sensitivity analyses can be found in SI Appendix (Figs. S5-S7).

Overall, as noted above, scenarios where classes are determined using fewer caricatured regions result in better (or at least comparable) performance than scenarios which used more regions. Similar to the simulations with perturbed spheres, we observe that SINATRA's power increases monotonically with an increasing number of cones and directions used to compute the topological summary statistics (Figs. 3(b), 3(c), and S5). For example, at a 10% FPR with $c = 5$ cones, we achieve 30% TPR in scenario I experiments and 35% 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 also remain consistent when choosing the angle between directions within cones (Fig. S6) and the number of sublevel sets (Fig. S7). Results for the former again suggest that there is an optimal cap radius to be used when generating directions in a cone. For the latter, since we are analyzing shapes with more intricate features, finer filtrations lead to more power.

Recovering Known Morphological Variation Across Genera of Primates

As a real application of our pipeline, with "ground truth" or known morphological variation, we consider a 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, there were 33 molars from the genus *Tarsius* [10, 43, 44] and 9 molars from the genus *Saimiri* [45]. From the strepsirrhine suborder, we have two examples of lemurs with 11 molars coming from the genus *Microcebus* and 5 molars being derived from the genus *Mirza* [10, 43, 44]. The meshes of all teeth were aligned, translated to be centered at the origin, and normalized to be enclosed within a unit sphere (SI Appendix Section 4 and Fig. S8).

This specific collection of molars was selected because morphologists and evolutionary anthropologists have come to understand variation 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 (Fig. 4(a)) [45, 46]. Phylogenetic analyses of mitochondrial genomes across primates place estimates of divergence dates of the subtree composed of *Microcebus* and *Mirza* from *Tarsius* at 5 million years before the branching of *Tarsius* from *Saimiri* [47]. Our main objective is to see if SINATRA recovers the information that the paraconids are specific to the *Tarsius* genus and whether variation across the molar is associated to the divergence time of the genera.

Since *Tarsius* is the only genus with the paraconid in this sample, we used SINATRA to perform three pairwise classification comparisons (*Tarsius* against *Saimiri*, *Mirza*, and *Microcebus*, respectively), and assessed SINATRA's ability to prioritize/detect the location of the paraconid as the region of interest

(ROI). Based on our findings in the 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. Again, we refer to this as the evidence potential for a vertex. We then display this information via a heatmap for each tooth (Fig. 4(b)), which allows us to visualize the physical regions that are most differential between the genera.

To assess the strength of SINATRA's ability to find *Tarsius*-specific paraconids, we make use of a null-based scoring method. Here, we place an expert-derived 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, the SINATRA computed evidence potentials are weighted by the surface area (or area of the Voronoi cell) encompassed by their corresponding vertices, and then summed together. This aggregated value, which we will 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. Similar (null) scores $\tau_1, \ldots, \tau_N$ are then computed for each randomly generated region. A "p-value"-like quantity (for the i-th molar) is then generated by following

$$P_i = \frac{1}{N+1} \sum_{t=1}^{N} \mathbb{I}(\tau_i^* \leq \tau_t), \tag{5}$$

where $\mathbb{I}(\cdot)$ denotes an indicator function, and a smaller P_i can be interpreted as having more confidence in SINATRA's ability to find the desired paraconid landmark. To ensure the robustness of this analysis, we generate the N-random null regions via one of two ways: (i) using a K-nearest neighbors (KNN) algorithm on each of the N-random seed vertices [48], or (ii) manually constructing K-vertex wide null regions such that they have surface areas equal to that of the paraconid ROI (SI Appendix Section 5). In both settings, we take the median of the P_i values in Equation (5) across all teeth, and report them for each genus and choice of K combination (see the first half of Table 1). Notedly, using p-values as a direct metric of evidence can be problematic. 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, a calibration formula has been provided that transforms a p-value to a bound/approximation of a Bayes factor (BF) [17], the ratio of the marginal likelihood under the alternative hypothesis H_1 versus the null hypothesis H_0, via the formula

$$BF(P_i)_{10} = [-e\, P_i \log(P_i)]^{-1}, \tag{6}$$

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. Table 1 reports the calibrated Bayes factor estimates as well.

Overall, we observe that the paraconid ROI is more strongly enriched in the comparisons between the *Tarsius* and either of the strepsirrhine primates, rather than for the *Tarsius*-*Saimiri* comparison. We suspect this difference is partly explained by the divergence times between these genera: *Tarsius* is more recently diverged from *Saimiri* than from the strepsirrhines. This conjecture is consistent with the intuition we developed in 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 the molars being too similar, no region appears key to explaining the variance between the two classes of primates.

Discussion

In this paper, we introduce SINATRA: the first statistical pipeline for sub-image analysis that does not require landmarks or correspondence points between images. We state properties of SINATRA using simulations and illustrate the practical utility of SINATRA on real data. The current formulation and software for SINATRA is limited to the classification setting. Extending the model and algorithm to the regression setting with continuous responses is trivial. There are many evolutionary applications where adaptation and heredity must first be disentangled in the analyses of continuous traits and phenotypes. The standard approach for this is to explicitly account for the hierarchy of descent by adding genetic covariance or kinship across species to the likelihood either via phylogenetic regression [49] or linear mixed models (e.g. the animal model) [50]. Modeling covariance structures also arises in statistical and quantitative genetics applications where individuals are related [51–53]. The SINATRA framework uses a Bayesian hierarchical model that is straightforward to adapt to analyze complex covariance structures in future work.

Acknowledgements

The authors would like to thank Ani Eloyan, Anthea Monod, Jenny Tung, and Christine Wall for helpful conversations and suggestions. 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), as well as by an Alfred P. Sloan Research Fellowship (No. FG-2019-11622) 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. 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. All authors performed the analyses, interpreted the results, and wrote and revised the manuscript.

Competing Financial Interests

The authors have declared that no competing interests exist.

Figures and Tables

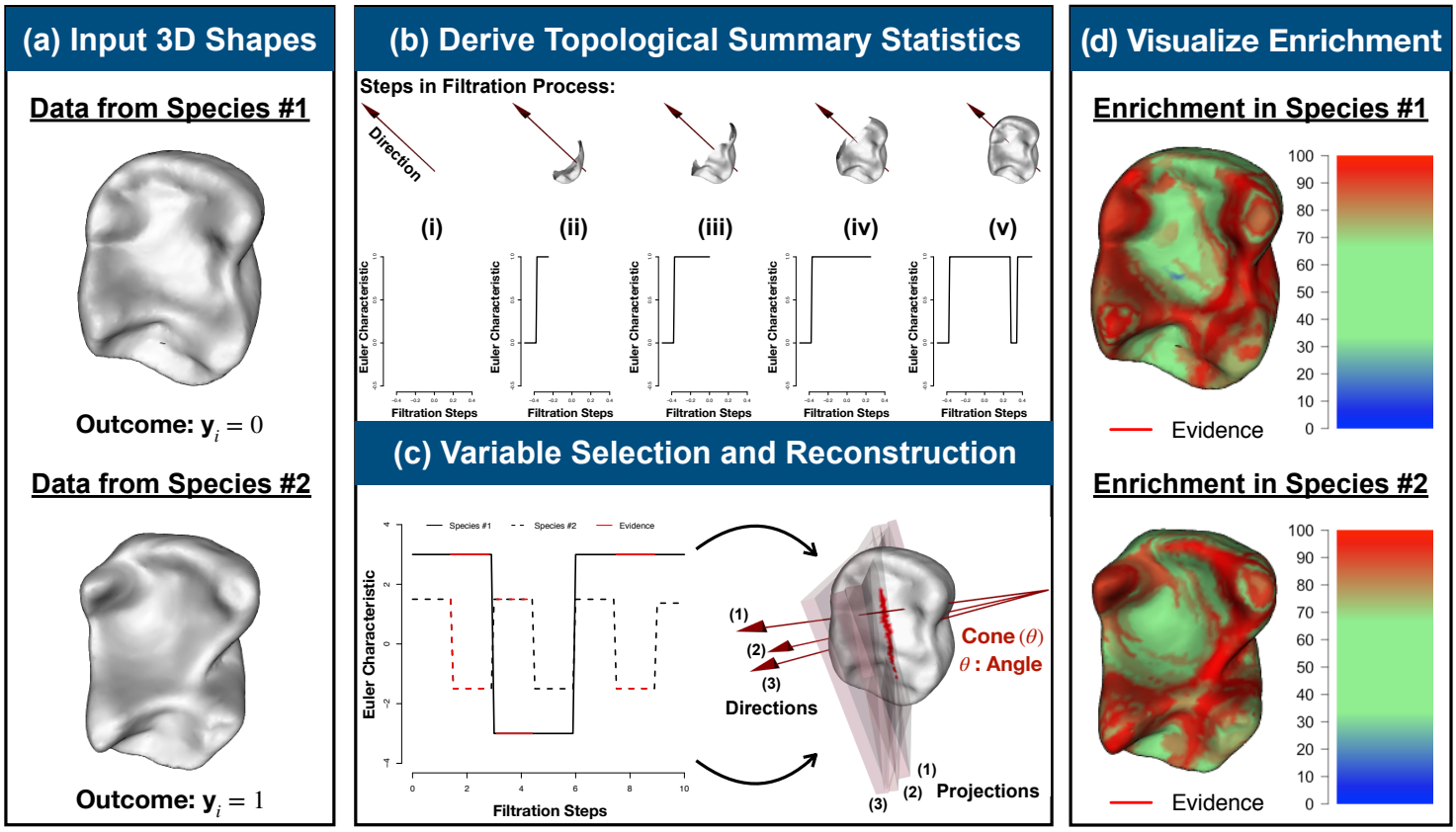
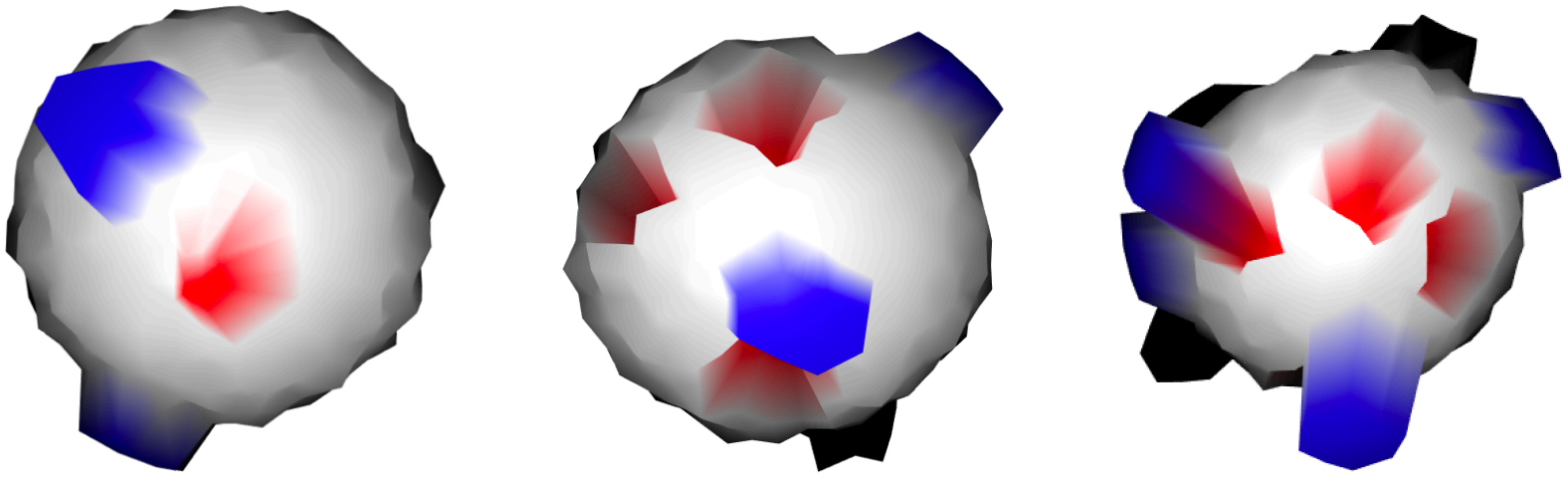
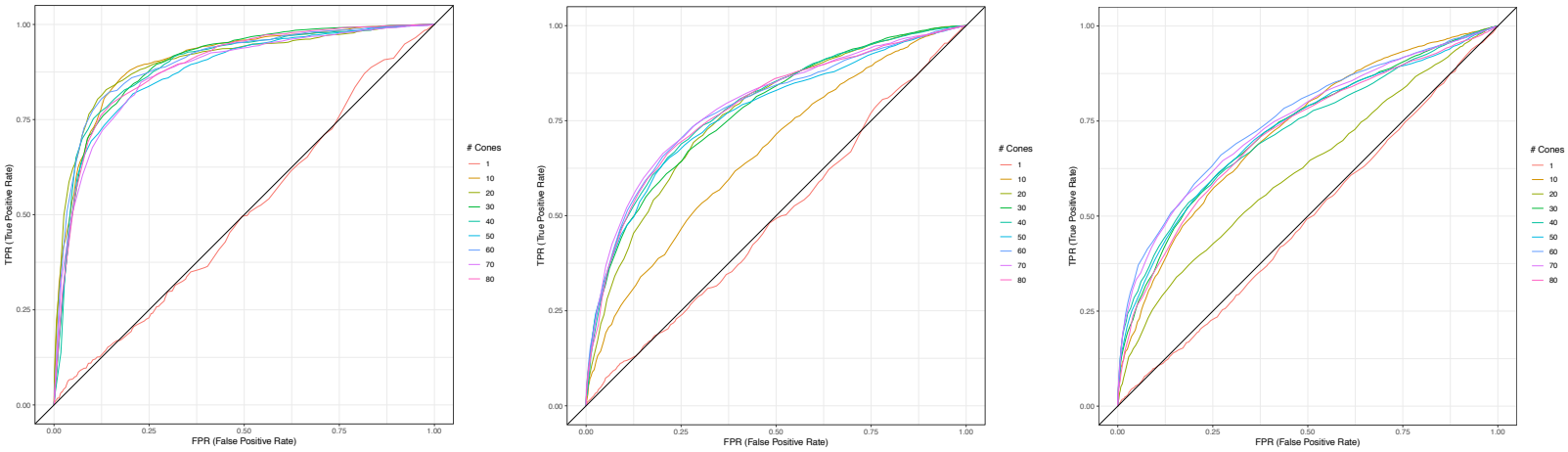


Figure 1. Schematic overview of SINATRA: a novel statistical framework for feature selection and association mapping with 3D shapes. (a) The SINATRA algorithm requires the following inputs: (i) aligned shapes represented as meshes; (ii) $\mathbf{y}$, a binary vector denoting shape classes; (iii) r, the 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. **(b)** We first 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 [54], 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 curve 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. Using these 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.

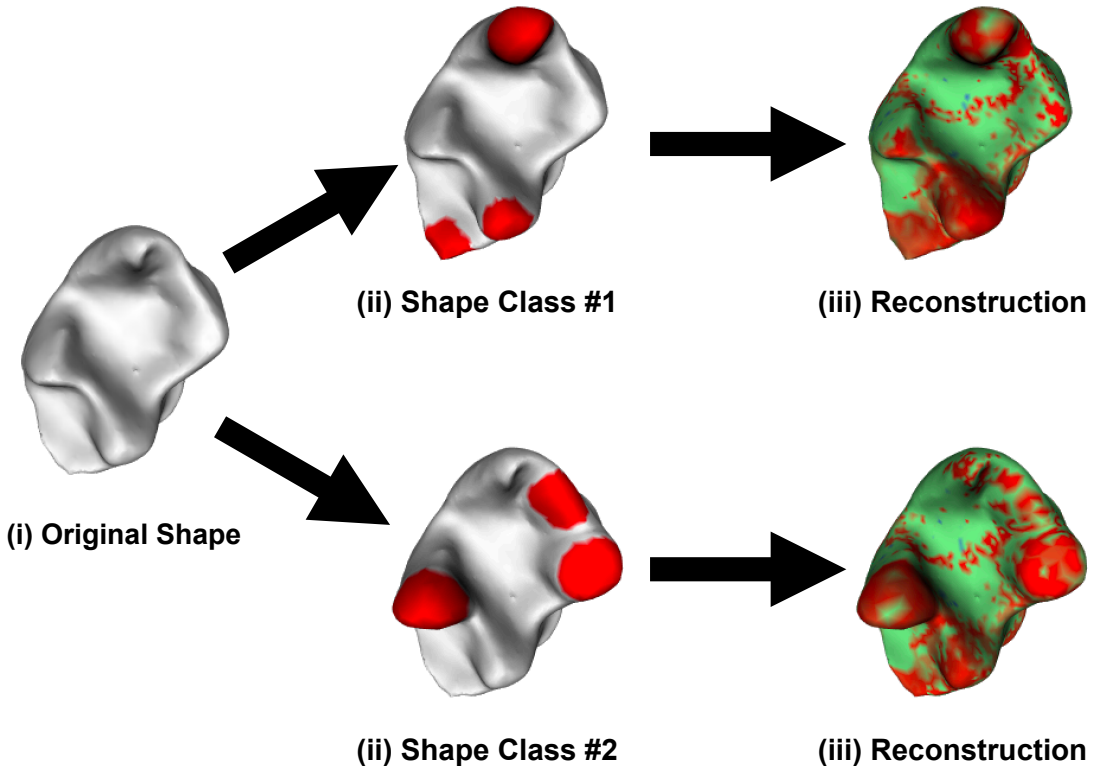


(a) Scenario I (b) Scenario II (c) Scenario III

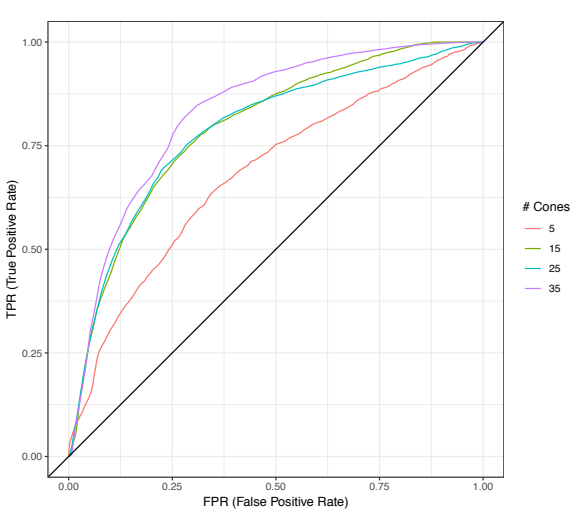


(d) Scenario I (e) Scenario II (f) Scenario III

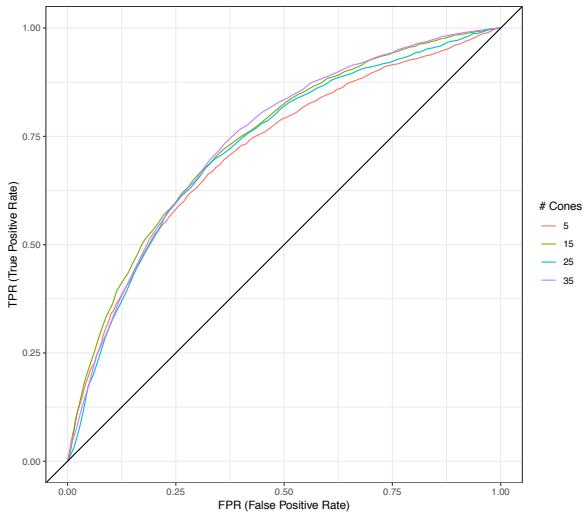
Figure 2. Power analysis for detecting associated vertices across different classes of perturbed spheres. Here, we generate 100 shapes by partitioning unit spheres into 10 vertex-wide regions, centered at 50 equidistributed points. Two classes (50 shapes per class) are defined by shared (blue protrusions) and class-specific (red indentations) characteristics. The shared or "non-associated" features are chosen by randomly selecting u regions and pushing the sphere outward at each of these positions. This is done for all shapes, regardless of class. 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 at the following: $d = 5$ 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 fifty replicates in each scenario.



(a) Caricature Simulation Flowchart

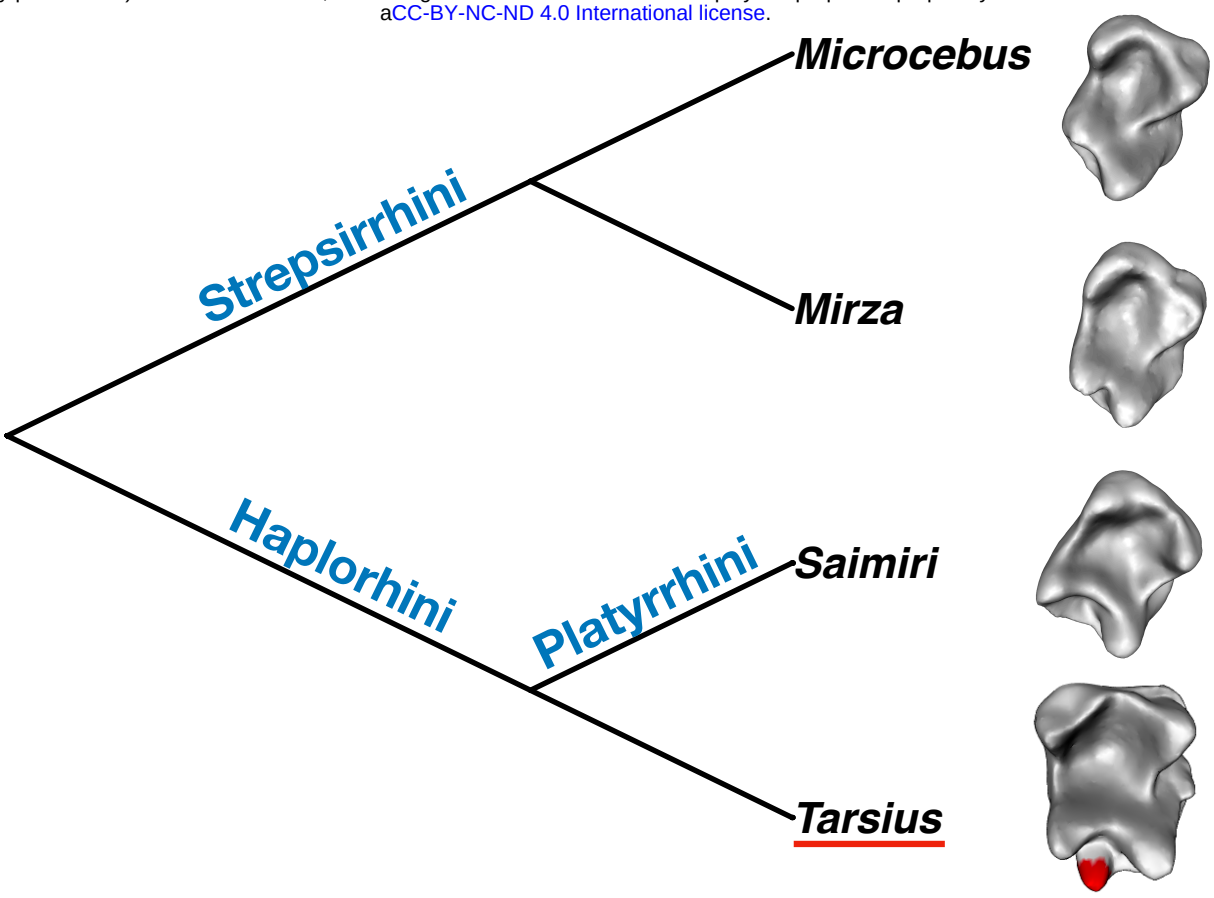


(b) Scenario I (3 Peaks)

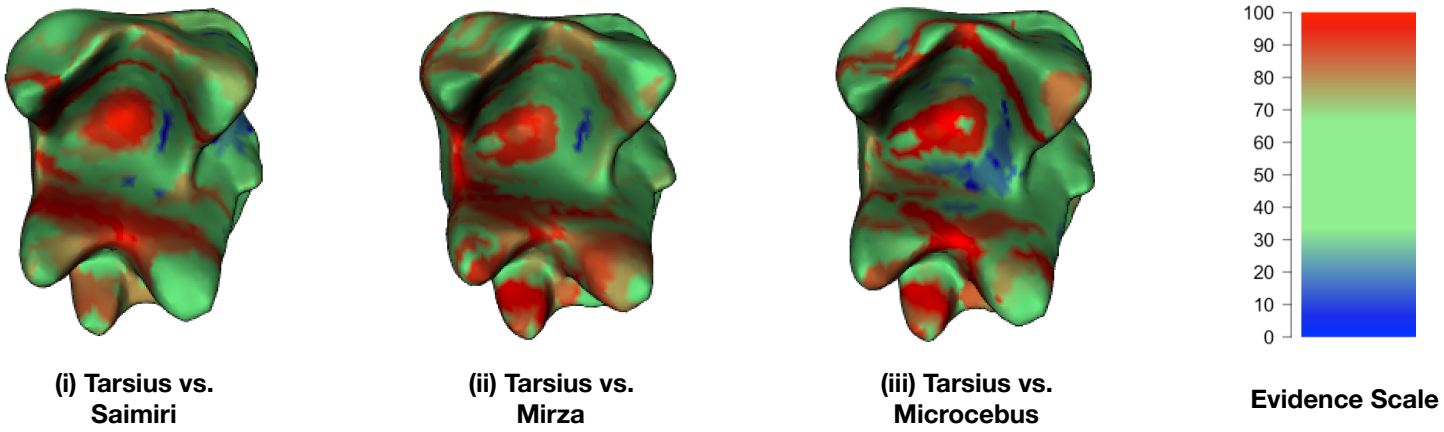


(c) Scenario II (5 Peaks)

Figure 3. Power analysis for detecting associated vertices across different classes of caricatured shapes. (a) Here, we modify real Lemuridae molars using the following caricaturization procedure. (i) First, we fix the triangular mesh of an individual tooth. (ii) Next, we take expert-derived landmarks for the tooth [10], and assign v of them to be specific to one class and v' to be specific to the other. The caricaturization is performed by multiplying each face within these regions by positive scalars so that class-specific features are exaggerated. This is repeated twenty-five times (with some small added noise) to create two equally-sized 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 increasing the number of cones of directions used in the algorithm. Other SINATRA parameters were fixed at the following: $d = 5$ 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 fifty replicates in each scenario.



(a) Associated Feature and Phylogenetic Relationship



(b) Example of Reconstructions per Comparison

Figure 4. Real data analysis aimed at detecting unique paraconids in molars belonging to primates in *Tarsius* genus. Here, we carry out three different pairwise comparisons where we analyze the physical difference between *Tarsius* molars and teeth from (i) *Saimiri*, (ii) *Mirza*, and (iii) *Microcebus* genus, respectively. In panel **(a)**, 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 reduce a wider range of foods [55]. The goal of this analysis is to assess SINATRA's ability to find this region of interest (ROI). In panel **(b)**, we show an example of the reconstruction resulting from each comparison. Intuition behind these results is consistent both with the phylogeny of the primates, as well as with our previous simulation studies. Genetically, *Tarsius* differ more from the *Mirza* and *Microcebus* genus, rather than from *Saimiri*. As a result, SINATRA is powered to find the unique paraconid in the former two comparisons because of the appropriate genetic distance, rather than in the latter case where molar structure is much more similar. 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.

	Test	Region Size	*Tarsius* vs. *Saimiri*	*Tarsius* vs. *Mirza*	*Tarsius* vs. *Microcebus*
P-Values (P)	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 (BF)	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**

Table 1. Null region experiment to evaluate SINATRA's ability to find paraconids in *Tarsius* molars. Here, the goal is to assess how likely it is that SINATRA finds the region of interest (ROI) by chance. To do so, we first generate 500 "null" regions on each *Tarsius* tooth using (i) a KNN algorithm and (ii) an equal-area approach (SI Appendix Section 5). Next, for each region, we sum the evidence potential or "birth times" of all the vertices it contains. Then, 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", 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 given in bold.

References

1. Crawford L, Monod A, Chen AX, Mukherjee S, Rabadán A. Functional data analysis using a topological summary statistic: The smooth Euler characteristic transform. arXiv. 2017;p. 1611.06818. Available from: https://arxiv.org/abs/1611.06818.

2. Turner K, Mukherjee S, Boyer DM. Persistent homology transform for modeling shapes and surfaces. Inf Inference. 2014;3(4):310–344. Available from: https://academic.oup.com/imaiai/article-abstract/3/4/310/724811?redirectedFrom=fulltext.

3. Curry J, Mukherjee S, Turner K. How many directions determine a shape and other sufficiency results for two topological transforms. arXiv. 2018;p. 1805.09782. Available from: https://arxiv.org/abs/1805.09782.

4. Ghrist R, Levanger R, Mai H. Persistent homology and Euler integral transforms. J Appl and Comput Topology. 2018;2(1-2):55–60. Available from: https://link.springer.com/article/10.1007/s41468-018-0017-1.

5. Kendall DG. A Survey of the Statistical Theory of Shape. Statist Sci. 1989;4(2):87–99. Available from: https://doi.org/10.1214/ss/1177012582.

6. Dupuis P, Grenander U. Variational problems on flows of diffeomorphisms for image matching. Q Appl Math. 1998;LVI(3):587–600. Available from: http://dl.acm.org/citation.cfm?id=298828.298844.

7. Worsley KJ. Estimating the number of peaks in a random field ssing the Hadwiger characteristic of excursion Sets, with applications to medical images. Ann Stat. 1995;23(2):640–669. Available from: https://doi.org/10.1214/aos/1176324540.

8. Goswami A. Phenome10K: a free online repository for 3-D scans of biological and palaeontological specimens; 2015. Available from: www.phenome10k.org.

9. Boyer DM, Gunnell GF, Kaufman S, McGeary TM. Morphosource: archiving and sharing 3-D digital specimen data. The Paleontological Society Papers. 2016;22:157–181.

10. Boyer DM, Lipman Y, St Clair E, Puente J, Patel BA, Funkhouser T, et al. Algorithms to automatically quantify the geometric similarity of anatomical surfaces. Proc Natl Acad Sci U S A. 2011;108(45):18221. Available from: http://www.pnas.org/content/108/45/18221.abstract.

11. Ovsjanikov M, Ben-Chen M, Solomon J, Butscher A, Guibas L. Functional maps: a flexible representation of maps between shapes. ACM Trans Graph. 2012;31(4):30:1–30:11. Available from: http://doi.acm.org/10.1145/2185520.2185526.

12. Boyer DM, Puente J, Gladman JT, Glynn C, Mukherjee S, Yapuncich GS, et al. A new fully automated approach for aligning and comparing shapes. Anat Rec (Hoboken). 2015;298(1):249–276. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/ar.23084.

13. Gao T, Kovalsky SZ, Daubechies I. Gaussian process landmarking on manifolds. SIAM J Math Data Sci. 2019;1(1):208–236. Available from: https://epubs.siam.org/doi/abs/10.1137/18M1184035.

14. Gao T, Kovalsky S, Boyer D, Daubechies I. Gaussian process landmarking for three-dimensional geometric morphometrics. SIAM J Math Data Sci. 2019;1(1):237–267. Available from: https://epubs.siam.org/doi/abs/10.1137/18M1203481.

15. Miller E. Fruit flies and moduli: interactions between biology and mathematics. Notices of the AMS. 2015;62(10):1178–1184.

16. Bendich P, Marron JS, Miller E, Pieloch A, Skwerer S. Persistent homology analysis of brain artery trees. Ann Appl Stat. 2016 03;10(1):198–218. Available from: https://doi.org/10.1214/15-AOAS886.

17. Sellke T, Bayarri MJ, Berger JO. Calibration of p-values for testing precise null hypotheses. Am Stat. 2001;55(1):62–71.

18. Neal RM. Monte Carlo implementation of Gaussian process models for Bayesian regression and-Monte Carlo implementation of Gaussian process models for Bayesian regression and classification. Dept. of Statistics, University of Toronto; 1997. 9702.

19. Neal RM. Regression and classification using Gaussian process priors. Bayesian Anal. 1998;6:475.

20. Williams CKI, Barber D. Bayesian classification with Gaussian processes. IEEE Trans Pattern Anal Mach Intell. 1998;20(12):1342–1351. Available from: https://ieeexplore.ieee.org/document/735807/.

21. Rasmussen CE, Williams CKI. Gaussian processes for machine learning. Cambridge, MA: MIT Press; 2006.

22. Nickisch H, Rasmussen CE. Approximations for binary Gaussian process classification. J Mach Learn Res. 2008;9(10):2035–2078.

23. Zhang Z, Dai G, Jordan MI. Bayesian generalized kernel mixed models. J Mach Learn Res. 2011;12:111–139.

24. Heckerman D, Gurdasani D, Kadie C, Pomilla C, Carstensen T, Martin H, et al. Linear mixed model for heritability estimation that explicitly addresses environmental variation. Proc Natl Acad Sci U S A. 2016;113(27):7377–7382. Available from: http://www.pnas.org/content/113/27/7377.abstract.

25. Swain PS, Stevenson K, Leary A, Montano-Gutierrez LF, Clark IBN, Vogel J, et al. Inferring time derivatives including cell growth rates using Gaussian processes. Nat Comm. 2016;7(12):13766. Available from: https://doi.org/10.1038/ncomms13766.

26. Cheng L, Ramchandran S, Vatanen T, Lietzén N, Lahesmaa R, Vehtari A, et al. An additive Gaussian process regression model for interpretable non-parametric analysis of longitudinal data. Nat Comm. 2019;10(1):1798. Available from: https://doi.org/10.1038/s41467-019-09785-8.

27. McDonald KR, Broderick WF, Huettel SA, Pearson JM. Bayesian nonparametric models characterize instantaneous strategies in a competitive dynamic game. Nat Comm. 2019;10(1):1808. Available from: https://doi.org/10.1038/s41467-019-09789-4.

28. Rodriguez-Nieva JF, Scheurer MS. Identifying topological order through unsupervised machine learning. Nat Phys. 2019;Available from: https://doi.org/10.1038/s41567-019-0512-x.

29. Schölkopf B, Herbrich R, Smola AJ. A generalized representer theorem. In: Proceedings of the 14th Annual Conference on Computational Learning Theory and and 5th European Conference on Computational Learning Theory. London, UK, UK: Springer-Verlag; 2001. p. 416–426. Available from: http://dl.acm.org/citation.cfm?id=648300.755324.

30. Pillai NS, Wu Q, Liang F, Mukherjee S, Wolpert R. Characterizing the function space for Bayesian kernel models. J Mach Learn Res. 2007;8:1769–1797.

31. Jiang Y, Reif JC. Modeling epistasis in genomic selection. Genetics. 2015;201:759–768.

32. Chaudhuri A, Kakde D, Sadek C, Gonzalez L, Kong S. The mean and median criteria for kernel bandwidth selection for support vector data description. Data Mining Workshops (ICDMW), 2017 IEEE International Conference on. 2017;p. 842–849. Available from: `https://ieeexplore.ieee.org/abstract/document/8215749/`.

33. Singleton KR, Crawford L, Tsui E, Manchester HE, Maertens O, Liu X, et al. Melanoma therapeutic strategies that select against resistance by exploiting MYC-driven evolutionary convergence. Cell Rep. 2017;21(10):2796–2812.

34. Crawford L, Wood KC, Zhou X, Mukherjee S. Bayesian approximate kernel regression with variable selection. J Am Stat Assoc. 2018;113(524):1710–1721. Available from: `https://doi.org/10.1080/01621459.2017.1361830`.

35. Crawford L, Flaxman SR, Runcie DE, West M. Predictor variable prioritization in nonlinear models: a genetic association case study. Ann Appl Stat. 2019;13(2):958–989. Available from: `https://projecteuclid.org/euclid.aoas/1560758434`.

36. Fasy BT, Micka S, Millman DL, Schenfisch A, Williams L. Challenges in reconstructing shapes from Euler characteristic curves. arXiv. 2018;p. 1811.11337.

37. Oudot S, Solomon E. Inverse problems in topological persistence. arXiv. 2018;p. 1810.10813. Available from: `https://arxiv.org/abs/1810.10813`.

38. Li F, Zhang T, Wang Q, Gonzalez MZ, Maresh EL, Coan JA. Spatial Bayesian variable selection and grouping for high-dimensional scalar-on-image regression. Ann Appl Stat. 2015;9(2):687–713. Available from: `https://projecteuclid.org:443/euclid.aoas/1437397107`.

39. Crawford L, Zeng P, Mukherjee S, Zhou X. Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. PLoS Genet. 2017;13(7):e1006869–. Available from: `https://doi.org/10.1371/journal.pgen.1006869`.

40. Zhu X, Stephens M. Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. Nat Comm. 2018;9(1):4361.

41. Lipman Y. Collection of Lemuridae teeth; 2011. Available from: `http://www.wisdom.weizmann.ac.il/~ylipman/CPsurfcomp/`.

42. Sela M, Aflalo Y, Kimmel R. Computational caricaturization of surfaces. Comput Vis Image Underst. 2015;141:1–17. Available from: `https://doi.org/10.1016/j.cviu.2015.05.013`.

43. Gao T. Hypoelliptic diffusion maps and their applications in automated geometric morphometrics. Duke University; 2015.

44. Gao T. The diffusion geometry of fibre bundles: horizontal diffusion maps. arXiv. 2016;p. 1602.02330.

45. St Clair EM, Boyer DM. Lower molar shape and size in prosimian and platyrrhine primates. Am J Phys Anthropol. 2016;161(2):237–258.

46. Guatelli-Steinberg D. Primate dentition: an introduction to the teeth of non-human primates. Am J Phys Anthropol. 2003;121(2):189–189. Available from: `https://doi.org/10.1002/ajpa.10194`.

555 47. Pozzi L, Hodgson JA, Burella AS, Raaumb RL, Disotell TR. Primate phylogenetic relationships and
556 divergence dates inferred from complete mitochondrial genomes. Mol Phylogenet Evol. 2014;75:165–
557 183.

558 48. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theor. 2006;13(1):21–27.
559 Available from: https://doi.org/10.1109/TIT.1967.1053964.

560 49. Graven A. The phylogenetic regression. Philos Trans R Soc Lond B Biol Sci. 1989;326(1233):87–99.

561 50. Henderson CRCR, of Guelph U. Applications of linear models in animal breeding. Guelph, Ont. :
562 University of Guelph; 1984. Includes index.

563 51. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, et al. Efficient control of
564 population structure in model organism association mapping. Genetics. 2008;178(3):1709–1723.
565 Available from: http://www.genetics.org/content/178/3/1709.abstract.

566 52. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong Sy, Freimer NB, et al. Variance component model
567 to account for sample structure in genome-wide association studies. Nat Genet. 2010;42(4):348–354.
568 Available from: http://dx.doi.org/10.1038/ng.548.

569 53. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat
570 Genet. 2012;44(7):821–825.

571 54. Belongie S. Rodrigues' rotation formula. From MathWorld–A Wolfram Web Resource, created by
572 Eric W Weisstein http://mathworld wolfram com/RodriguesRotationFormula html. 1999;.

573 55. Crompton RH, Savage R, Spears IR. The mechanics of food reduction in Tarsius bancanus. Hard-
574 object feeder, soft-object feeder or both? Folia Primatol (Basel). 1998;69(Suppl 1):41–59.