Measuring hidden phenotype: Quantifying the shape of barley seeds using the Euler Characteristic Transform

Erik J. Amézquita[1], Michelle Y. Quigley[2], Tim Ophelders[4], Jacob B. Landis[5,6,7], Daniel Koenig[7], Elizabeth Munch[1,3,†], Daniel H. Chitwood[1,2,†]

[1]Department of Computational Mathematics, Science & Engineering, [2]Department of Horticulture, [3]Department of Mathematics, Michigan State University, East Lansing, MI, USA

[4]Department of Mathematics and Computer Science, TU Eindhoven, Eindhoven, the Netherlands

[5]School of Integrative Plant Science, Section of Plant Biology and the L.H. Bailey Hortorium, Cornell University, Ithaca, NY, USA

[6]BTI Computational Biology Center, Boyce Thompson Institute, Ithaca, NY, USA

[7]Department of Botany & Plant Sciences, University of California, Riverside, CA, USA

Short title: Measuring the shape of barley using topology

Keywords: Topological Data Analysis; Euler characteristic transform; mathematical biology; data science; shape

[†]To whom correspondence should be addressed:

Dr. Daniel H. Chitwood
1066 Bogue St, East Lansing, MI 48824
(517) 353-0462
chitwoo9@msu.edu

Dr. Elizabeth Munch
428 S Shaw Ln # 3115, East Lansing, MI 48824
(517) 432-0619
muncheli@msu.edu

Abstract

Shape plays a fundamental role in biology. Traditional phenotypic analysis methods measure some features but fail to measure the information embedded in shape comprehensively. To extract, compare, and analyze this information embedded in a robust and concise way, we turn to Topological Data Analysis (TDA), specifically the Euler Characteristic Transform (ECT). TDA measures shape comprehensively using mathematical terms based on algebraic topology features. To study its use, we compute both traditional and topological shape descriptors to quantify the morphology of 3121 barley seeds scanned with X-ray Computed Tomography (CT) technology at 127 micron resolution. The ECT measures shape by analyzing topological features of an object at thresholds across a number of directional axes. We optimize the number of directions and thresholds for classification to 158 and 8 respectively, creating vectors of length 1264 that are topological signatures for each barley seed. Using these vectors, we successfully train a support vector machine to classify 28 different accessions of barley based on the 3D shape of their grains. We observe that combining both traditional and topological descriptors classifies barley seeds to their correct accession better than using just traditional descriptors alone. This improvement suggests that TDA is thus a powerful complement to traditional morphometrics to describe comprehensively a multitude of shape nuances which are otherwise not picked up. Using TDA we can quantify aspects of phenotype that have remained "hidden" without its use, and the ECT opens the possibility of accurately reconstructing objects from their topological signatures.

1 Introduction

There is a discrepancy between the information embedded in biological forms that we can discern with our senses versus that which we can quantify. Methods to comprehensively quantify phenotype are not commensurate with the thoroughness and speed with which genomes can be sequenced. High-throughput phenotyping has enabled us to collect large amounds of phenotyping data (Andrade-Sanchez et al., 2013; Araus and Cairns, 2014; Tanabata et al., 2012); nonetheless, we are not maximizing the information extracted from the data we collect.

One framework for extracting information embedded within data is to consider its shape. From a morphological perspective, the form of biological organisms is both data and literal shape simultaneously. Landmark-based approaches based on Procrustean superimposition (Bookstein, 1997) and Fourier-based decomposition of closed outlines (Kuhl and Giardina, 1982; Lestrel, 1997) comprise traditional morphometric methods. These approaches measure shape comprehensively, but are limited to either a geometric perspective that only considers the distances and relative positions of data points to each other or to a frequency domain transform of a closed contour. We thus turn to topology, the mathematical discipline that studies shape in a more abstract sense.

Topological Data Analysis (TDA) is a set of tools that arise from the perspective that all data has shape and that shape is data (Lum et al., 2013; Munch, 2017). TDA treats the data as if made of elementary building blocks as in Figure 1A: points, edges, squares, and cubes, referred to as 0-, 1-, 2-, and 3-dimensional *cells* respectively. A collection of cells is referred to as a *cubical complex*, or complex, for short.

Cubical complexes are both a natural and consistent way to represent image data (Kovalevsky, 1989). Given a grayscale image as shown in Figure 1A, we follow a strategy similar to Wagner et al. (2012) to construct a cubical complex: A nonzero pixel will correspond to a vertex in our complex. If two pixels are adjacent—in the 4-neighborhood sense—we say that there is an edge between the corresponding vertices in the complex. If 4 pixels in the image form a 2×2 square, we will consider a square in our complex between the corresponding 4 vertices. Additionally, for the 3D image case, if 8 voxels—the 3D equivalent of pixels—make a $2 \times 2 \times 2$ cube, we will draw a cube in our complex between the corresponding 8 vertices.

TDA seeks to describe the shape of our data based on the number of relevant topological features found in the corresponding complex. For instance, the complex in Figure 1A has two distinct, separate pieces colored in blue and red respectively, formally referred to as *connected components*. This complex also has 8 edges forming the outline of a square without an actual red block filling it—edges thickened for emphasis—this is referred to as a *loop*. In higher dimensions, we could also consider hollow blocks containing *voids*. We can even go a step further and summarize these topological features with a single value known as the *Euler characteristic*, represented by the Greek letter χ, defined for voxel-based images as

$$\chi = \#(\text{connected components}) - \#(\text{loops}) + \#(\text{voids}).$$

The Euler characteristic is a topological invariant, that is, it will remain unchanged under any smooth transformation applied to our shape. The well-known but surprising Euler-Poincaré formula states that χ can be computed easily as

$$\chi = \#(\text{Vertices}) - \#(\text{Edges}) + \#(\text{Faces}) - \#(\text{Cubes}).$$

This equivalence can be seen in the cubical complex from Figure 1A, where

$$\begin{aligned}\chi &= 18 \text{ vertices} - 19 \text{ edges} + 2 \text{ faces} \\ &= 2 \text{ connected components} - 1 \text{ loop} + 0 \text{ voids} = 1.\end{aligned}$$

The Euler characteristic by itself might be too simple. Nonetheless, we can extract more information out of our data-based complex if we think of it as a dynamic object that grows in number of vertices, edges, and faces across time. As our complex grows, we may observe significant changes in χ. The changes in χ can be thought as a topological signature of the shape, referred to as an *Euler characteristic curve (ECC)*. The growth of the complex is defined by a *filter function* which assigns a real number value to each voxel. For reasons discussed later, we will focus on directional filters which assign to each voxel its height as if measured from a fixed direction.

As an example, consider the cubical complex of a barley seed and the direction corresponding to the adaxial-abaxial axis in Figure 1B. Voxels at the top of the seed will be assigned the lowest values, while voxels at the bottom will obtain the highest values. We then consider 32 equispaced, increasing thresholds $t_1 < t_2 < \ldots < t_{32}$ which define 32 different slices of equal thickness along the adaxial-abaxial axis. We start by computing the Euler characteristic of the first slice, that is, all the voxels with filter value less than t_1. Next we aggregate the second slice, which are all the voxels with filter value less than t_2, and recompute the Euler characteristic. We repeat the procedure for the 32 slices. For instance in Figure 1C, we observe that we started with scattered voxels which are thought of as many connected components which may explain the high Euler characteristic values. As we keep adding slices, we connect most of the stray voxels into fewer but larger connected components, and simultaneously, we might have created loops as seen in t_4 and t_6. This merging of connected

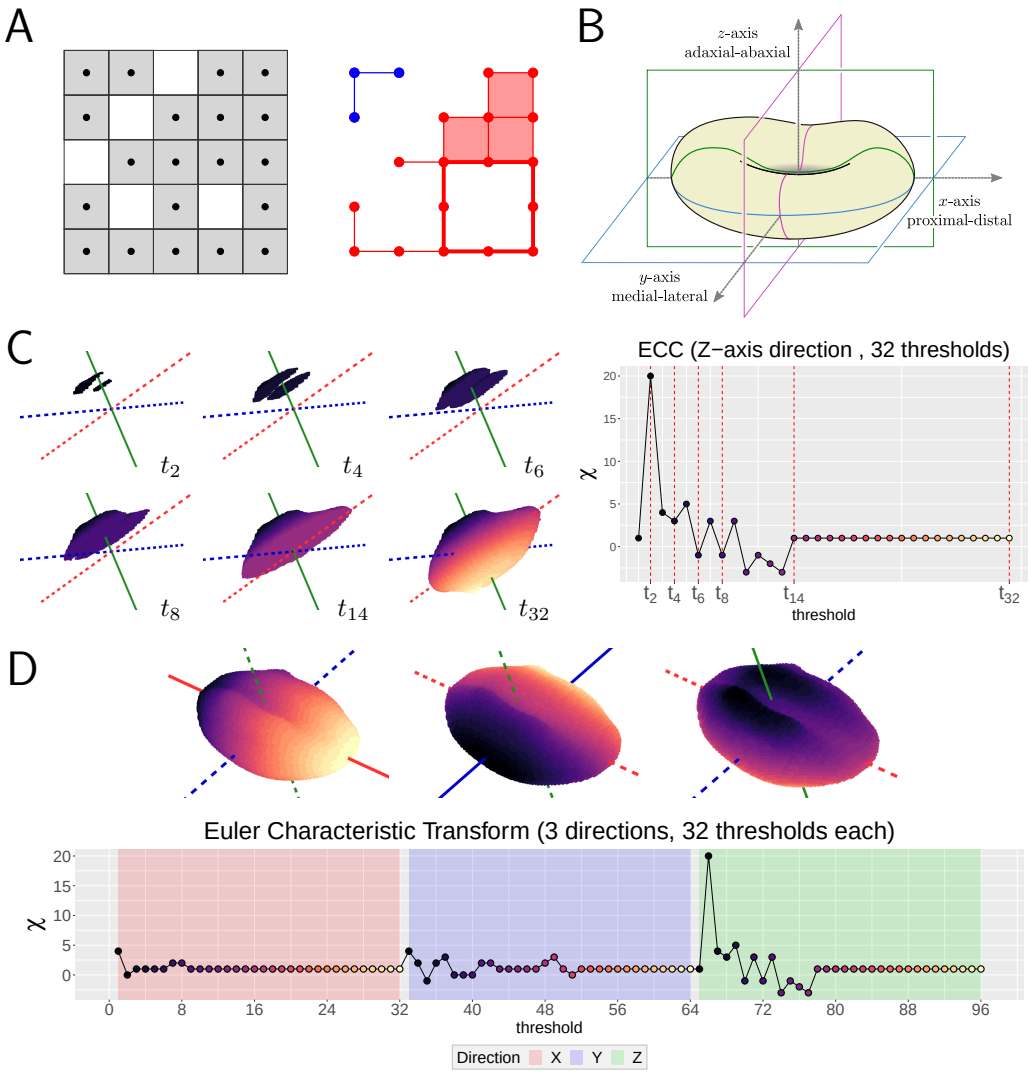

Figure 1: Extracting topological shape signatures from barley seeds. **A.** A binary image (left) is treated as a cubical complex (right). This cubical complex has 2 connected components, 1 loop, 0 voids. The distinct connected components are colored in blue and red respectively. The loop is emphasized with thicker edges. **B.** The barley seeds were aligned so that their proximal-distal, medial-lateral, and adaxial-abaxial axes corresponds to the X, Y, Z-axes in space. **C.** Example of an Euler Characteristic Curve (ECC) as we filter the barley seed across the adaxial-abaxial axis (depicted as a solid, green line) through 32 equispaced thresholds. **D.** The Euler Characteristic Transform (ECT) consists of concatenating all the ECCs corresponding to all possible directions. In this example, we concatenate 3 ECCs corresponding to the X, Y, Z directions.

components, and formation and closing of loops might explain the fluctuation of the Euler characteristic between positive and negative values. Finally, after more than half of the slices have been considered, at t_{14}, we observe that no new loops are formed, and every new voxel will simply be part of the single connected component. Thus, the Euler characteristic remains constant at 1. The ECC is precisely the sequence of different Euler characteristic values as we add systematically individual slices along the chosen direction.

To get a better sense of how the Euler characteristic changes overall, we can compute several ECCs corresponding to different directional filters. For example, in Figure 1D, we choose three directions in total corresponding to the proximal-distal, medial-lateral, and adaxial-abaxial axes respectively. Each filter produces an individual ECC, which we later concatenate into a unique large signal known as the *Euler Characteristic Transform (ECT).*

There are two important reasons to use ECT over other TDA techniques. First, the ECT is computationally inexpensive, since it is based on successive computations of the Euler characteristic, which is simply an alternating sum of counts of cells. This inexpensiveness is especially relevant as we are dealing with thousands of extremely high-resolution 3D images. Assuming that we have already treated the image as a cubical complex, we can compute a single ECC in linear time with respect to the number of voxels in the image (Richardson and Werman, 2014). We can thus compute the ECT of a 50,000-voxel seed scan with 150 directions in less than two seconds on a traditional PC. The second reason to use the ECT is its provable invertibility and statistical sufficiency. As proved by Turner et al. (2014), and later extended by Curry et al. (2018) and Ghrist et al. (2018), if we compute all possible directional filters we would have sufficient information to reconstruct the original shape. Moreover, this ECT is a sufficient statistic that effectively summarizes all information regarding

158 shape. Although there are infinite possible directional filters, there is ongoing
159 research into defining a sufficient finite number of directions such that we can
160 effectively reconstruct shapes based solely on their finite ECT (Belton et al.,
161 2018; Betthauser, 2018; Curry et al., 2018; Fasy et al., 2019). Nonetheless, a
162 computationally efficient reconstruction procedure for large 3D images remains
163 elusive.

164 Here we show the use of ECTs to correctly describe the shape of barley seeds
165 as a proof of concept. We scanned a collection of barley panicles comprising
166 28 different accessions with X-ray CT technology at 127 micron resolution.
167 These scans were later digitally processed to isolate 3121 individual grains.
168 With individual seeds, we quantified their morphology using both traditional and
169 topological shape descriptors. To verify the descriptor correctness, we trained a
170 support vector machine (SVM) to determine the accession of individual grains
171 based on their shape alone. Our experiment shows that SVMs perform better
172 whenever topological information is taken into account, which suggests that
173 the ECT measures shape that is "hidden" from traditional shape descriptors.

174 2 Materials and Methods

175 We selected 28 barley accessions with diverse spike morphologies and geographi-
176 cal origins for our analysis (Harlan and Martini, 1929, 1936, 1940). In November
177 of 2016, seeds from each accession were stratified at 4C on wet paper towels
178 for a week, and germinated on the bench at room temperature. Four day old
179 seedlings were transferred into pots in triplicate and arranged in a completely
180 randomized design in a greenhouse. Day length was extended throughout the
181 experiment using artificial lighting (minimum 16h light / 8h dark). After the

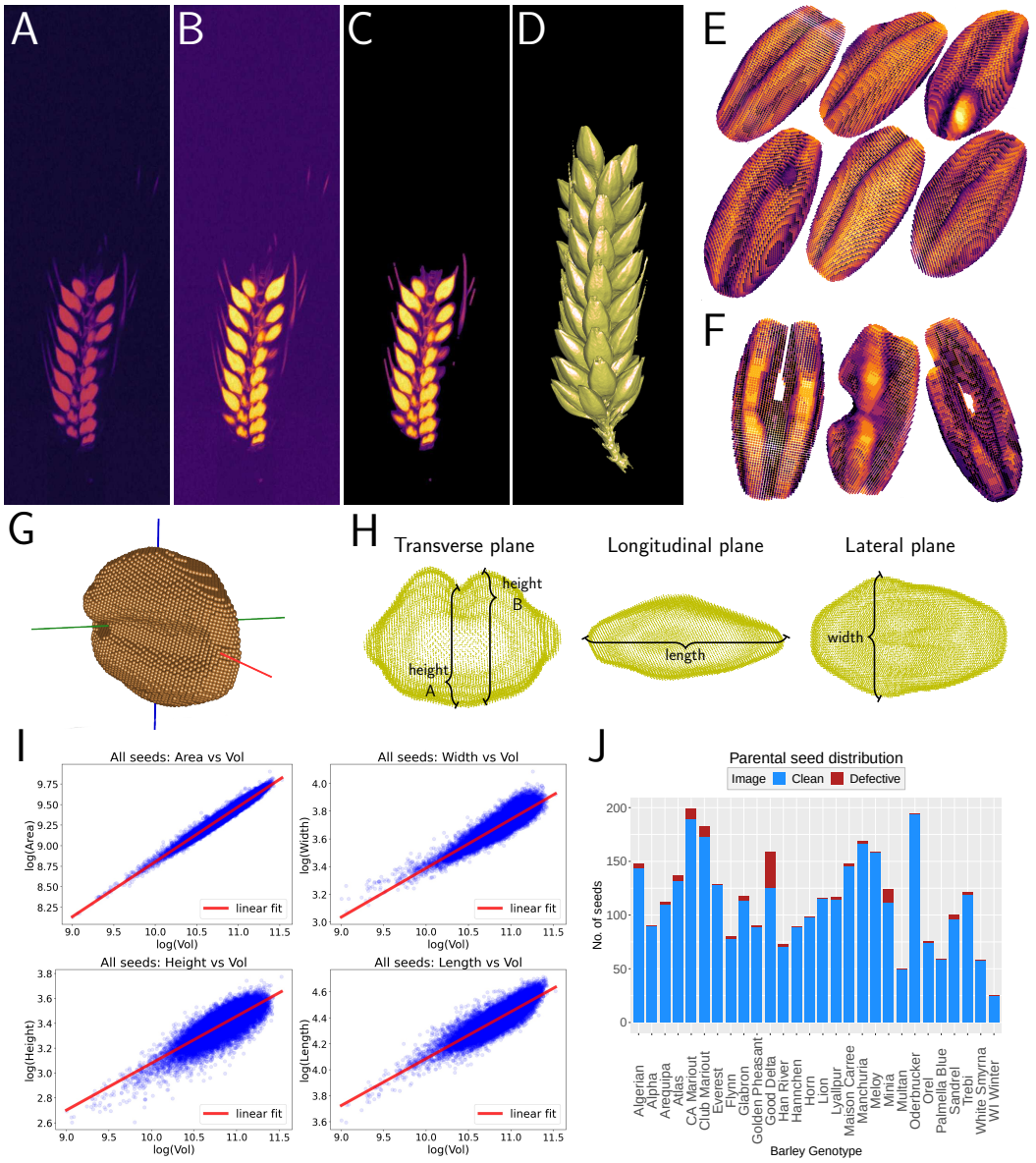

Figure 2: Barley image processing. The morphology measurements were extracted from 3D voxel-based images of the barley panicles. Before any analysis was done, the **A.** raw X-ray CT scans of the panicles had their **B.** densities normalized, **C.** air and other debris removed, and awns pruned. **D.** After automating these image processing steps, we could finally work with a large collection of clean, 3D panicles. **E.** An extra digital step segmented the individual seeds for each barley spike. **F.** Example of incomplete or broken seeds which were removed from the data set. **G.** The seeds were aligned according to their principal components, which allowed us to **H.** measure a number of traditional shape descriptors. **I.** The damaged seeds were initially identified as outliers of the allometry plots. **J.** The total number of clean and defective seeds measured from each accession. Defective seeds were not concentrated in a particular accession.

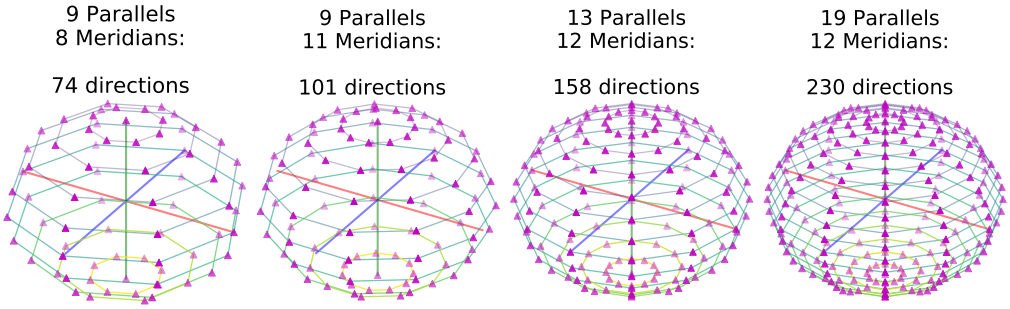

Figure 3: Directions chosen to compute the ECT. The sphere was split into a equispaced fixed number of parallels and meridians in each case. The directions were the taken from the intersections.

plants reached maturity and dried, a single spike was collected from each replicate for scanning at Michigan State University. The scans were produced using the North Star Imaging X3000 system and the included efX software, with 720 radiographs per scan. The X-ray source was set to a voltage of 75 kV, current of 100 μA, and focal spot length of 0 microns. The 3D reconstruction of the spikes was computed with the efX-CT software, obtaining a final voxel size of 127 microns. The intensity values for all raw reconstructions was standardized, the air and debris thresholded out, and awns digitally pruned—Figures 2A-2D. We digitally isolated all the seeds as in Figure 2E, and thus obtained a collection of 3121 seeds in total. The details of varieties and their number of seeds can be found in the supplement Table 1. Due to the large volume of data, we used python to automate the image processing pipeline for all panicles and grains.

To make the comparison of different directional filters comparable across seeds, all the seeds were aligned with respect to their first three principal components. This alignment corresponds to the proximal-distal, medial-lateral, and adaxial-abaxial axes respectively as depicted in Figures 1B or 2G. With this alignment we were able to measure the length, width, heights, surface area and volume of each seed as depicted in Figure 2H. We also computed the convex hull for each seed and measure its surface area and volume. Finally, we computed the ratios

Table 1: Sample size of seed scans used for each individual accession. The seeds come from a three panicles per accession setup. 3121 seeds were used in total.

Accession	num	Accession	num	Accession	num
Algerian	144	Golden Pheasant	89	Minia	112
Alpha	90	Good Delta	126	Multan	50
Arequipa	110	Han River	71	Oderbrucker	194
Atlas	132	Hannchen	89	Orel	74
California Mariout	189	Horn	98	Palmella Blue	59
Club Mariout	173	Lion	116	Sandrel	96
Everest	128	Lyallpur	115	Trebi	119
Flynn	78	Maison Carree	146	White Smyrna	58
Glabron	114	Manchuria	167	Wisconsin Winter	25
		Meloy	159		

of seed surface area and volume to its convex hull surface area and volume respectively. In total we measured 11 different traditional shape descriptors. Outliers in the allometry plots helped us identify and remove damaged seeds, as in Figures 2I and 2J.

As a proof of concept, we explored how topological descriptors varied as we varied both the number of different directions and the number of uniformly spaced thresholds. In total, for every seed we computed the ECT considering 74, 101, 158, and 230 different directions. We emphasized directions toward the seed's cleft, which correspond to directions close to both north and south poles. Refer to Figure 3. For each direction, we produced ECCs with 4, 8, 16, 32, and 64 thresholds.

For every seed we computed a very high dimensional vector of topological information, usually above 1000 dimensions. In general, high-dimensional vectors

tend to produce distorted prediction and regression results (Köppen, 2000), so we sought to reduce the topological information to just a few dimensions. As proposed originally by Schölkopf et al. (1998), we employed a non-linear kernel principal component analysis (KPCA) with a Laplacian kernel to aggressively reduce the ECT vectors to just 12 dimensions, usually less than 1% of the original ECT dimension. Hereafter, by topological descriptors we will refer to the ECT vectors after being reduced in dimension with KPCA.

We then sought to test the descriptiveness of both traditional and topological measures. To this end, we trained three non-linear support vector machines (SVM) (Burges, 1998) to characterize and predict the seeds from 28 different accessions based on three different collection of descriptors: traditional, topological, and combining both traditional and topological descriptors. In every case, the descriptors were centered and scaled to variance 1 prior to classification. Given that SVM is a supervised learning method, we partitioned our data into training and testing sets. In our case, we randomly sampled 80% of the seeds from every accession as our training data set. The remaining 20% was used to test the accuracy of our prediction model. We repeated this SVM setup 100 times and considered the average accuracy and confusion matrices as final results.

3 Results

Using either exclusively traditional or topological shape descriptors produces a comparable classification results. With either collection of descriptors as seen in Table 2, the machine is able to correctly determine the grain variety roughly 55% of the time. For comparison, by simply guessing randomly the variety, we

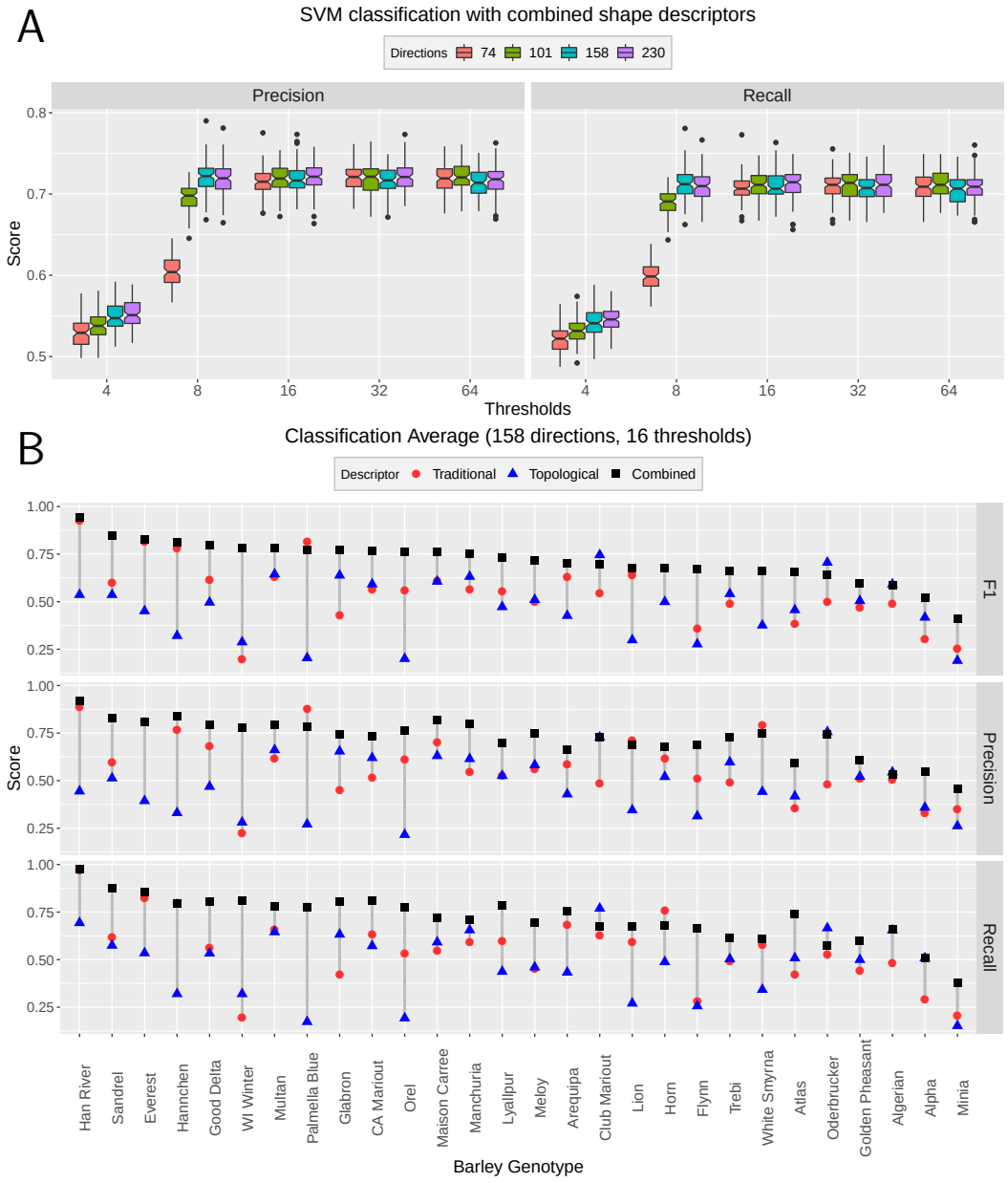

Figure 4: Classification results for different parameters and shape descriptors. A. The ECT was computed by concatenating 158 different directional filters as in Figure 3 with 8 thresholds each. This choice is due to the fact that increasing either the number of directions or thresholds did not improve classification scores when using combined shape descriptors. **B.** Combined shape descriptors in general outperform the separate use of traditional or topological shape descriptors. Combined shape descriptors produce the best precision, recall, and F_1 classification scores for most of the barley accessions.

Shape descriptors	No. of descriptors	Scores (weighted average ± standard deviation) Precision	Recall	F_1
Traditional	11	0.57 ± 0.058	0.56 ± 0.019	0.55 ± 0.019
Topological (ECT + KPCA)	12	0.51 ± 0.063	0.51 ± 0.020	0.50 ± 0.020
Combined (Trad. + Topo.)	23	0.72 ± 0.055	0.71 ± 0.018	0.71 ± 0.018

Table 2: SVM classification accuracy of barley seeds from 28 different founding lines after 100 randomized training and testing sets. The ECT was computed with 158 directions (as in Figure 3) and 8 thresholds. Since we are in a multi-class classification setting we first computed the precision, recall, and F_1 scores for each founding line. Later, we computed the weighted average for each score, where the weight depended on the number of test seeds for each of the barley lines. Observe that the use of combined descriptors outperforms the use of traditional descriptors.

Assuming t distribution			Assuming normal distribution		
	Traditional	Topological		Traditonal	Topological
Topological	8.6×10^{-3}	$*$	Topological	6.7×10^{-5}	$*$
Combined	$< 2 \times 10^{-16}$	$< 2 \times 10^{-16}$	Combined	$< 2 \times 10^{-16}$	$< 2 \times 10^{-16}$

Table 3: Small Quade post-hoc p-values (with Bonferroni correction) suggest that different descriptors produce statistically different SVM results.

would expect to be correct just $1/28 \times 100 \approx 4\%$ of the time. Thus, both sets of descriptors do capture important morphological patterns that can be picked up by a computer. Moreover, our overall prediction accuracy increases beyond 70% if we use both traditional and topological measures to characterize seed shape. This is even more striking if we consider that we aggressively reduced the dimension of the ECTs. A Friedman test (Friedman, 1937) among the three accuracy results produces a p-value of 8.1×10^{-8}, which suggests that the three SVM classifiers are statistically different. Since we are comparing only three classifiers, we can rely better on a Quade post-hoc pairwise test (Quade, 1979) as suggested in (Conover, 1998). The p-values are reported in Table 3.

The results presented on Tables 2 and 3 are based on an ECT computed with 158 directions (refer to Figure 3) and 8 thresholds. As shown in Figure 4A, We chose this parameters on the observation that increasing either the number of thresholds or directions did not improve classification results, and potentially contributed to diminishing returns.

4 Discussion

Traditional morphometrics has been used on ancient cereal grains to reveal fundamental trends in morphological changes across space and time (Bouby, 2001; Coster and Field, 2015). Historical evidence shows that barley seeds became smaller as the crop moved from Mediterranean climates to Northwest Europe to account for colder temperatures and higher sunlight variance, shedding some insight on the timeline of barley domestication in Central Asia (Motuzaite Matuzeviciute et al., 2018). Similarly, grains became rounder and the spikes became more compact as they moved to higher altitude sites in Nepal (Tanno and Willcox, 2012). Differences become more subtle if we compare accessions that originated from similar regions and time periods. Geometric Morphometrics (GMM) has provided a more detailed characterization of the grains. For example, GMM can successfully tell apart barley grains from einkorn (*Triticum monococcum*) and emmer (*Triticum dicoccum*) grains (Bonhomme et al., 2017); it can be used to distinguish two-row vs six-row barley seeds (Ros et al., 2014); and it can establish unique morphological characteristics of land races to deduce their possible historical origins (Wallace et al., 2019).

Morphometrics has a number of drawbacks in our proposed X-ray scan setting. GMM may have trouble if there are no clear homologous points and currently

most of the discipline has focused on 2D images rather than large 3D X-ray CT scans (Dryden and Mardia, 2016). We thus turn to topology. In recent years, TDA has produced promising results in diverse biological problems, like histological image analysis (Qaiser et al., 2019), viral phylogenetic trees description (Chan et al., 2013), and active-binding sites identification in proteins (Kovacev-Nikolic et al., 2016). In plant biology, the Euler characteristic has been used successfully used to define the morphospace of more than 180,000 leaves from seed plants (Li et al., 2018), and to characterize the shape of apple leaves (Migicovsky et al., 2018) and the 3D structure of grapevine inflorescences (Li et al., 2019).

The Euler characteristic provides important shape information for barley seeds as well. We observe that the topological shape descriptors provide an overall similar characterization performance than the traditional shape descriptors. As seen from Table 2, both kinds of shape descriptors provide similar precision and recall scores. Notice however that some specific barley varieties are more easily distinguishable with the topological lens but not with traditional measures, and vice-versa. For instance if we focus on the F_1 scores in Figure 4B, Glabron and Alpha report considerably higher classification accuracies whenever using topological information compared to using only traditional measures. Moreover, some lines such as Club Mariout and Oderbrucker are better characterized using exclusively topological features, since combining traditional measures just muddles classification results. On the other hand, our topological descriptors perform poorly whenever we try to distinguish lines such as Palmella Blue and Hannchen, as these lines seem much better characterized by traditional measures alone. Finally, some lines like Wisconsin Winter or Flynn reported poor classification results whenever we limited ourselves to just topological or traditional measures; however, our classification accuracy improved dramatically

as we combined both descriptors.

A more careful exploration on the directions used to compute the ECT could reveal more shape information and improve the classification results described above. Of particular note, we could do a more exhaustive ECT analysis and observe if there is a particular directional filter that contributes the most morphological information. A related question is to explore how the ECT and subsequent results vary if we pick randomly distributed directions—or according to any other probability distribution—instead of regularly distributed ones as in Figure 3. We can also do a more systematic experimentation with different dimension reduction algorithms, and classification techniques afterward, in order to improve the results presented above.

The Euler characteristic is a simple yet powerful way to reveal features not readily visible to the naked eye. There is "hidden" morphological information that traditional and geometric morphometric methods are missing. The Euler characteristic, and Topological Data Analysis in general, can be readily computed from any given image data, which makes it an extremely versatile tool to use in a vast number of biology-related applications. TDA provides a comprehensive framework to detect and compare these important morphological nuances for different barley accessions, nuances that can be distinguished by just analyzing the external shape structure of individual grains rather than working with the barley spike as a whole. These "hidden" shape nuances at the seed level, if properly detected, can provide surprisingly enough information to characterize specific accessions. Our results suggest a new exciting path, driven mainly by morphological information, to explore further the phenotype-genotype relationship in barley and many more plant species.

5 Software and data availability

All of our code is available at the https://github.com/amezqui3/demeter/ repository. This includes the image processing pipeline to clean the raw scans and segment the seeds (python), the computation of the ECTs (python), and the SVM classification and analysis (R). A collection of jupyter notebook tutorials is also provided in order to ease the usage and understanding of the different components of the data processing and data analyzing pipelines.

6 Acknowledgements

Dan Chitwood is supported by the USDA National Institute of Food and Agriculture, and by Michigan State University AgBioResearch. The work of Elizabeth Munch is supported in part by the National Science Foundation through grant CCF-1907591. Daniel Koenig is supported by an award from the National Science Foundation Plant Genome Research Program (IOS-2046256) and funding from the USDA NIFA (CA-R-BPS-5154-H).

References

Andrade-Sanchez P, Gore MA, Heun JT, Thorp KR, Carmo-Silva AE, French AN, Salvucci ME, White JW (2013). Development and evaluation of a field-based high-throughput phenotyping platform. *Functional Plant Biology* **41**(1), 68–79. doi:10.1071/FP13126.

Araus JL, Cairns JE (2014). Field high-throughput phenotyping: the new crop breeding frontier. *Trends in Plant Science* **19**(1), 52–61. doi:10.1016/j.tplants.2013.09.008.

Belton RL, Fasy BT, Mertz R, Micka S, Millman DL, Salinas D, Schenfisch A, Schupbach J, Williams L (2018). Learning simplicial complexes from persistence diagrams. arXiv:1805.10716v2.

Betthauser LM (2018). *Topological reconstruction of grayscale images*. Ph. D. thesis, University of Florida, Gainesville, Florida.

Bonhomme V, Forster E, Wallace M, Stillman E, Charles M, Jones G (2017). Identification of inter- and intra-species variation in cereal grains through geometric morphometric analysis, and its resilience under experimental charring. *Journal of Archaeological Science* **86**, 60 – 67. doi:10.1016/j.jas.2017.09.010.

Bookstein FL (1997). *Morphometric Tools for Landmark Data: Geometry and Biology*. Geometry and Biology. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511573064.

Bouby L (2001). L'orge à deux rangs (*Hordeum distichum*) dans l'agriculture gallo-romaine : données archéobotaniques. *ArchéoSciences, revue d'Archéométrie*, 35–44. doi:10.3406/arsci.2001.999.

Burges CJ (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* **2**(2), 121–167. doi:10.1023/A:1009715923555.

Chan JM, Carlsson G, Rabadán R (2013). Topology of viral evolution. *Proceedings of the National Academy of Sciences* **110**(46), 18566–18571. doi:10.1073/pnas.1313480110.

Conover WJ (1998). *Practical Nonparametric Statistics* (3rd ed.). Wiley Series in Probability and Statistics. Wiley.

Coster AC, Field JH (2015). What starch grain is that? — a geometric morphometric approach to determining plant species origin. *Journal of Archaeological Science* **58**, 9 – 25. doi:10.1016/j.jas.2015.03.014.

Curry J, Mukherjee S, Turner K (2018). How many directions determine a shape and other sufficiency results for two topological transforms. arXiv:1805.09782.

Dryden IL, Mardia KV (2016). *Statistical Shape Analysiswith Applications in R* (2 ed.). John Wiley & Sons Ltd. doi:10.1002/9781119072492.

Fasy BT, Micka S, Millman DL, Schenfisch A, Williams L (2019). The first algorithm for reconstructing simplicial complexes of arbitrary dimension from persistence diagrams. arXiv:1912.12759.

Friedman M (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association* **32**(200), 675–701. doi:10.1080/01621459.1937.10503522.

Ghrist R, Levanger R, Mai H (2018). Persistent homology and Euler integral transforms. *Journal of Applied and Computational Topology* **2**(1), 55–60. doi:10.1007/s41468-018-0017-1.

Harlan HV, Martini ML (1929). A composite hybrid mixture. *Agronomy Journal* **21**(4), 487–490. doi:10.2134/agronj1929.00021962002100040014x.

Harlan HV, Martini ML (1936). *Problems and results in barley breeding*. Washington, DC: US Department of Agriculture.

Harlan HV, Martini ML (1940). A study of methods in barley breeding. Technical Report 720, US Department of Agriculture, Washington, DC.

Köppen M (2000). The curse of dimensionality. In *5th Online World Conference on Soft Computing in Industrial Applications (WSC5)*, Volume 1, pp. 4–8.

Kovacev-Nikolic V, Bubenik P, Nikolić D, Heo G (2016). Using persistent homology and dynamical distances to analyze protein binding. *Statistical Applications in Genetics and Molecular Biology* **15**(1), 19–38. doi:10.1515/sagmb-2015-0057.

Kovalevsky V (1989). Finite topology as applied to image analysis. *Computer Vision, Graphics, and Image Processing* **46**(2), 141 – 161. doi:10.1016/0734-189X(89)90165-5.

Kuhl FP, Giardina CR (1982). Elliptic Fourier features of a closed contour. *Computer Graphics and Image Processing* **18**(3), 236 – 258. doi:10.1016/0146-664X(82)90034-X.

Lestrel PE (Ed.) (1997). *Fourier Descriptors and their Applications in Biology*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511529870.

Li M, An H, Angelovici R, Bagaza C, Batushansky A, Clark L, Coneva V, Donoghue MJ, Edwards E, Fajardo D et al. (2018). Topological data analysis as a morphometric method: Using persistent homology to demarcate a leaf morphospace. *Frontiers in Plant Science* **9**, 553. doi:10.3389/fpls.2018.00553.

Li M, Klein LL, Duncan KE, Jiang N, Chitwood DH, Londo JP, Miller AJ, Topp CN (2019). Characterizing 3D inflorescence architecture in grapevine using X-ray imaging and advanced morphometrics: implications for understanding cluster density. *Journal of Experimental Botany* **70**(21), 6261–6276. doi:10.1093/jxb/erz394.

Lum PY, Singh G, Lehman A, Ishkanov T, Vejdemo-Johansson M, Alagappan M, Carlsson J, Carlsson G (2013). Extracting insights from the shape of complex data using topology. *Scientific Reports* **3**(1236). doi:10.1038/srep01236.

Migicovsky Z, Li M, Chitwood DH, Myles S (2018). Morphometrics reveals complex and heritable apple leaf shapes. *Frontiers in Plant Science* **8**, 2185. doi:10.3389/fpls.2017.02185.

Motuzaite Matuzeviciute G, Abdykanova A, Kume S, Nishiaki Y, Tabaldiev K (2018). The effect of geographical margins on cereal grain size variation: Case study for highlands of Kyrgyzstan. *Journal of Archaeological Science: Reports* **20**, 400 – 410. doi:10.1016/j.jasrep.2018.04.037.

Munch E (2017). A user's guide to topological data analysis. *Journal of Learning Analytics* **4**, 47–61. doi:10.18608/jla.2017.42.6.

Qaiser T, Tsang YW, Taniyama D, Sakamoto N, Nakane K, Epstein D, Rajpoot N (2019). Fast and accurate tumor segmentation of histology images using persistent

425 homology and deep convolutional features. *Medical Image Analysis* **55**, 1 – 14.
426 doi:10.1016/j.media.2019.03.014.

427 **Quade D** (1979). Using weighted rankings in the analysis of complete blocks with
428 additive block effects. *Journal of the American Statistical Association* **74**(367),
429 680–683. doi:10.1080/01621459.1979.10481670.

430 **Richardson E, Werman M** (2014). Efficient classification using the Euler characteristic.
431 *Pattern Recognition Letters* **49**, 99 – 106. doi:10.1016/j.patrec.2014.07.001.

432 **Ros J, Evin A, Bouby L, Ruas MP** (2014). Geometric morphometric analysis of
433 grain shape and the identification of two-rowed barley (*Hordeum vulgare subsp.*
434 *distichum L.*) in southern France. *Journal of Archaeological Science* **41**, 568 – 575.
435 doi:10.1016/j.jas.2013.09.015.

436 **Schölkopf B, Smola A, Müller KR** (1998). Nonlinear component analysis as a
437 kernel eigenvalue problem. *Neural Computation* **10**(5), 1299–1319. doi:10.1162/
438 089976698300017467.

439 **Tanabata T, Shibaya T, Hori K, Ebana K, Yano M** (2012). SmartGrain: High-
440 throughput phenotyping software for measuring seed shape through image analysis.
441 *Plant Physiology* **160**(4), 1871–1880. doi:10.1104/pp.112.205120.

442 **Tanno Ki, Willcox G** (2012). Distinguishing wild and domestic wheat and barley
443 spikelets from early Holocene sites in the Near East. *Vegetation History and*
444 *Archaeobotany* **21**(2), 107–115. doi:10.1007/s00334-011-0316-0.

445 **Turner K, Mukherjee S, Boyer DM** (2014). Persistent homology transform for
446 modeling shapes and surfaces. *Information and Inference* **3**(4), 310–344. doi:
447 10.1093/imaiai/iau011.

448 **Wagner H, Chen C, Vuçini E** (2012). Efficient computation of persistent homology for
449 cubical data. In Peikert R, Hauser H, Carr H, Fuchs R (Eds.), *Topological Methods in*
450 *Data Analysis and Visualization II: Theory, Algorithms, and Applications*, pp. 91–106.

Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-23175-9_7.

Wallace M, Bonhomme V, Russell J, Stillman E, George TS, Ramsay L, Wishart J, Timpany S, Bull H, Booth A et al. (2019). Searching for the origins of *Bere* barley: a geometric morphometric approach to cereal landrace recognition in archaeology. *Journal of Archaeological Method and Theory* **26**(3), 1125–1142. doi:10.1007/s10816-018-9402-2.