

Diptera wing classification using Topological Data Analysis

Sergio Tsuyoshi Ura Northon Canevari Leme Penteado
Guilherme Vituri F. Pinto Washington Mio
Alice Kimie Miwa Libardi

2026-05-13

We apply tools from Topological Data Analysis (TDA) to classify Diptera families based on wing venation patterns. From 70 binarized wing images representing nine families, we extract two complementary persistence summaries: H1 persistence from Vietoris-Rips filtrations of point-cloud samples, capturing global loop structure of the vein network, and H0 persistence from radial filtrations of connected wing images, capturing how vein components organize from the centroid outward. Each diagram is converted into a vector of interpretable summary statistics, yielding 34 features per specimen. A single balanced Random Forest classifier is evaluated by repeated stratified 3-fold cross-validation, and performance is reported with macro-F1, macro-recall, per-family recall, and a row-normalized confusion matrix. A feature-reduction screen identifies a smaller candidate set of topological summaries for biological interpretation. As a complementary baseline, we classify the wings directly from pairwise Wasserstein distances between Rips persistence diagrams; this metric-space approach is competitive with the feature-based Random Forest, indicating that much of the taxonomic signal is already present in the Rips diagrams themselves.

Source: [Article Notebook](#)

Source: [Article Notebook](#)

1 Introduction

Diptera wing venation is a classical taxonomic character: the arrangement of veins and enclosed cells varies among families and provides a natural morphological signature. Topological Data Analysis (TDA) is well suited to this problem because persistent homology summarizes connected components and loops in a way that is less tied to exact pixel coordinates than many raw image descriptors.

1.1 Related work

Persistent homology has become a productive tool in biological morphology over the past decade; Amézquita et al. [Amé+20] provide a broad review across scales from molecules to whole organisms. The plant sciences have driven much of this work: Li et al. [Li+18b] used persistent homology to demarcate a 141-family leaf morphospace from 182{,}707 specimens, reporting

roughly threefold better family-level classification than conventional landmark-based morphometrics; Li et al. [Li+18a] showed that persistent homology features captured additional QTLs for leaf, fruit, and root shape compared with classical descriptors; Amézquita et al. [Amé+22] used the Euler characteristic transform on X-ray CT scans of barley seeds to discriminate accessions; and Zhang et al. [Zha+21] combined persistent-homology features with deep learning to identify plant cultivars from leaf images, including aspects of leaf venation. Persistent homology has also been applied to branching biological structures: Li et al. [Li+17] treated branching plant architectures explicitly with persistent homology, and Kanari et al. [Kan+18] introduced the topological morphology descriptor for neuronal arbours using a radial-distance filtration from the soma. On the methodological side, Moon and Lazar [ML23] develop hypothesis testing for vectorised persistence diagrams.

Within image-based TDA, Garin and Tauzin [GT19] established the radial filtration and the height filtration on binarized images as standard primitives, classifying MNIST digits by sublevel-set persistence on these filtrations. We adopt the radial filtration directly from that line of work.

To our knowledge, persistent homology has not previously been applied to Diptera wing venation, and we are not aware of an insect-wing morphometric study using image-based persistent-homology descriptors. Existing topological work on insect wings has treated the venation as a graph and analysed it with classical network statistics rather than persistence. The present study brings a representation that has worked well for plants and neuronal morphology to insect wing morphology.

1.2 Contributions and scope

We use two complementary filtrations: a Vietoris-Rips filtration on point-cloud samples of each wing, retaining H1 persistence to describe global loops; and a radial filtration on the connected binary wing image, retaining H0 persistence to describe how vein components merge from the center of the wing outward.

The statistical goal is to test whether compact topological summaries carry family-level signal and to identify which summaries are most useful for prediction. To keep the validation aligned with the small and imbalanced dataset, we evaluate one balanced Random Forest model with repeated stratified 3-fold cross-validation and report macro-F1, macro-recall, family-level recall, and the confusion matrix. We also evaluate a pure metric-space baseline — Wasserstein distance between Rips H1 diagrams with k -nearest-neighbour classification — to separate the contribution of feature engineering from that of the raw topological representation.

2 Methods

2.1 Data and preprocessing

The dataset consists of 70 wing images covering nine Diptera families: *Asilidae*, *Bibionidae*, *Ceratopogonidae*, *Chironomidae*, *Rhagionidae*, *Sciaridae*, *Simuliidae*, *Tabanidae*, and *Tipulidae*. All images are stored in `images/processed`. File names encode the family and specimen identifier. We standardize family names, remove duplicated files that differ only by spacing or spelling variants, blur each image slightly, crop it, and resize it to 150 pixels in height.

The preprocessing pipeline applies, in order, the following operations:

- **Gaussian blur**, to close small gaps in the vein membranes so that the venation network remains connected after thresholding.
- **Crop to the bounding box** of the foreground pixels, to standardize the spatial extent of the images.
- **Connect pixel components**: many raw segmentations contain isolated noise pixels that are not part of the venation. We reduce each image to a single connected foreground component, prioritizing the largest one when several disconnected components are present.

For the point-cloud representation used in the Vietoris-Rips filtration we sample 750 pixels from each wing using farthest-point sampling, which provides good coverage of the venation geometry while keeping the Rips computation tractable.

Source: [Article Notebook](#)

Total images after deduplication: 70
Number of families: 9

	family	n
	String	Int64
1	Asilidae	8
2	Bibionidae	6
3	Ceratopogonidae	8
4	Chironomidae	8
5	Rhagionidae	4
6	Sciaridae	6
7	Simuliidae	7
8	Tabanidae	11
9	Tipulidae	12

Source: [Article Notebook](#)

Source: [Article Notebook](#)

2.2 Topological feature extraction

Persistent homology is the central tool of Topological Data Analysis. Given a filtration of a simplicial complex, it tracks how topological features (connected components in dimension 0, loops in dimension 1, voids in dimension 2, and so on) appear and disappear as the filtration parameter grows. Each feature has a *birth* time, when it first appears, and a *death* time, when it gets filled in or merges into a previously existing feature. The collection of all (birth, death) pairs is the *persistence diagram*. Long-lived features represent stable shape information, while short-lived features are typically attributable to noise.

For each wing we compute two persistence diagrams from two complementary filtrations:

2.2.1 Vietoris-Rips filtration on point clouds

Given a finite set of points in $\mathbb{R}^n$, the Vietoris-Rips complex at scale ε contains every subset of points that are pairwise within distance ε . As ε grows from 0, the complexes form a nested sequence — the Rips filtration. This is the most common filtration in TDA for point-cloud data. It is computationally expensive because all pairwise distances must be considered, which is why we work with subsamples of size 750 obtained by farthest-point sampling. We retain the H1

persistence of each Rips filtration, which encodes the global loop structure of the wing venation (the number and prominence of wing cells). We do not retain Rips H0: it contains 749 finite bars plus one infinite component, but the finite bars are dominated by inter-sample distances from farthest-point sampling rather than by intrinsic vein topology, and pilot experiments showed them to be redundant with the radial H0 features described below.

2.2.2 Radial filtration

The radial filtration was introduced by Garin and Tauzin [GT19] for classifying binarized MNIST digits, drawing on the radial-distance descriptors of Kanari et al. [Kan+18] for neuronal morphology. It assigns to each foreground pixel a value equal to its distance from a chosen center, here taken to be the centroid of the wing. Sublevel-set persistence on this function captures how topological features are distributed from the center of the wing outward, which is complementary to the Rips filtration: it is sensitive to the spatial layout of the venation rather than to its global loop structure. We retain H0 persistence, which tracks how disconnected vein segments merge as the radial sweep grows outward. We do not retain radial H1 because pixelated binary images produce very few clean loops under this filtration and the resulting H1 diagram is dominated by noise.

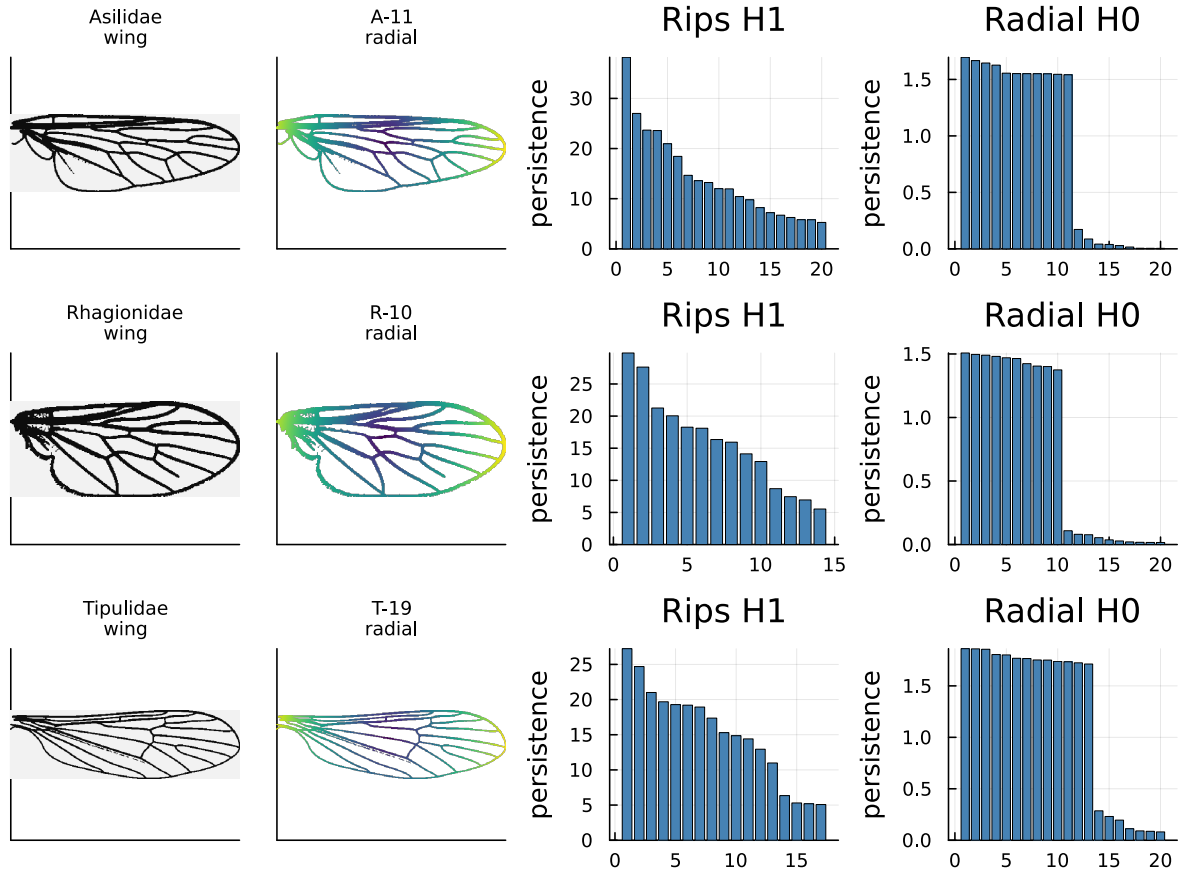
Source: [Article Notebook](#)

Source: [Article Notebook](#)

Source: [Article Notebook](#)

The plots below show three representative wings, their radial filtration, and the largest persistence values from each retained diagram.

Source: [Article Notebook](#)



Source: [Article Notebook](#)

2.3 Summary statistics

Each persistence diagram is converted into 19 summary statistics, of which we retain 17 per diagram. Let a persistence diagram be $D = \{(b_j, d_j) : j = 1, \dots, n\}$ and write $\ell_j = d_j - b_j$ for the persistence (lifetime) of the j -th interval. Intervals with infinite death (the single infinite H0 bar, if present) are excluded from all statistics below. The retained statistics fall into four groups:

Counts and totals.

- Number of finite intervals: n .
- Maximum persistence: $\max_j \ell_j$.
- Total persistence: $\sum_j \ell_j$.
- Total squared persistence: $\sum_j \ell_j^2$.

Quantiles of persistence values. The 10th, 25th, 50th (median), 75th, and 90th empirical percentiles of $\{\ell_j\}$, together with the persistence range $\max_j \ell_j - \min_j \ell_j$.

Spread and shape.

- Standard deviation of persistence: $\sigma(\ell_j)$.
- Persistence entropy [Ruc+16; Chi+15]: let $L = \sum_j \ell_j$ and $p_j = \ell_j/L$; then

$$E(D) = - \sum_{j=1}^n p_j \log p_j.$$

This is the (un-normalised) Shannon entropy of the persistence values rescaled to a probability distribution. Diagrams whose persistence mass is concentrated in a few long bars have low E ; diagrams whose bars have comparable lifetimes have high E .

Birth, death, and midlife summaries. Let births $\{b_j\}$ and finite deaths $\{d_j\}$ be the corresponding coordinates of the retained intervals. We compute the median of $\{b_j\}$, the median of $\{d_j\}$, $\sigma(b_j)$, $\sigma(d_j)$, and the mean midlife

$$\bar{m}(D) = \frac{1}{n} \sum_{j=1}^n \frac{b_j + d_j}{2},$$

where the midlife $(b_j + d_j)/2$ is computed per interval and then averaged.

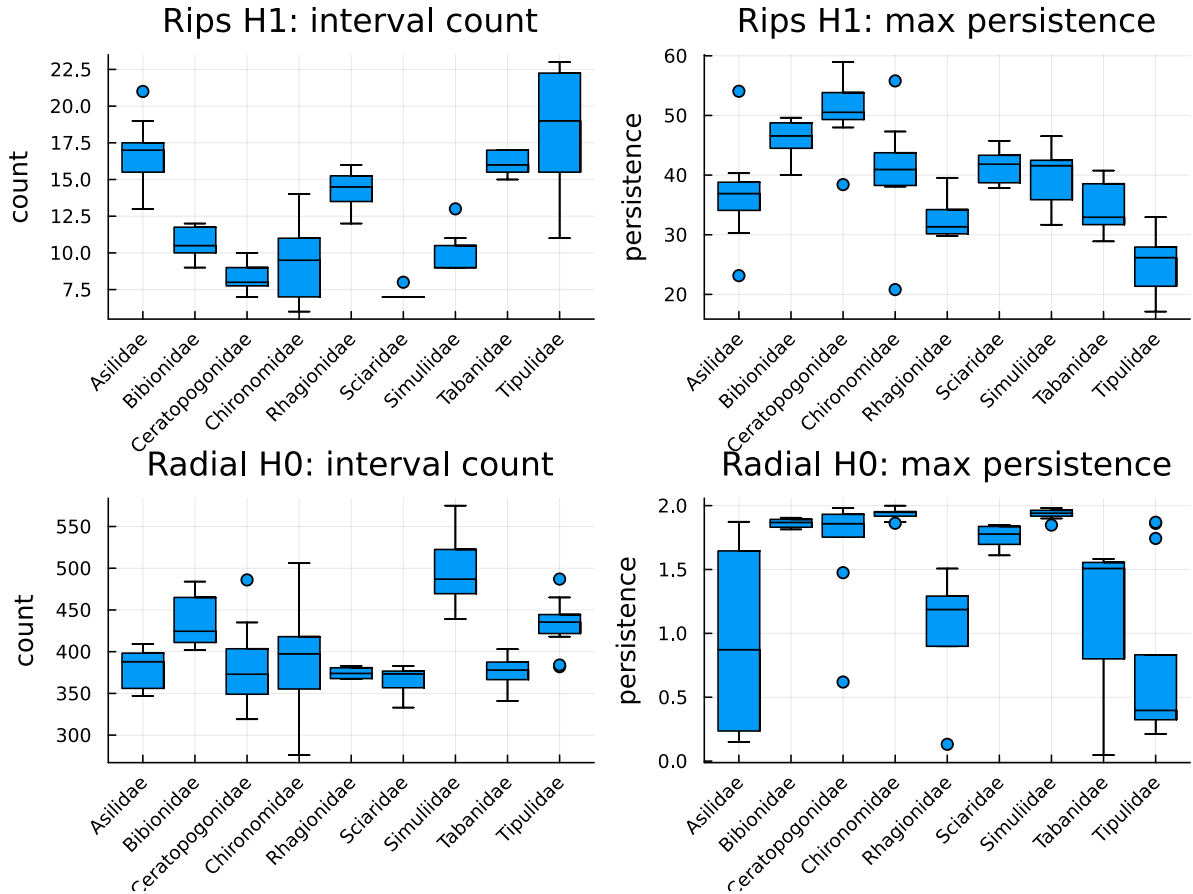
Skewness and kurtosis are computed but excluded from the feature set: they are tail-sensitive moments, and on the small persistence diagrams produced here they showed near-zero importance in tree-based models and added noise rather than signal. Two diagrams per wing $\times$ 17 statistics yields 34 features per specimen.

Source: [Article Notebook](#)

Retained statistics per diagram: 17

Feature matrix: 70 samples $\times$ 34 features

Source: [Article Notebook](#)



Source: [Article Notebook](#)

2.4 Design rationale: filtrations and representations not used

The Rips H1 / radial H0 combination above was selected after pilot experiments with several additional filtrations and persistence representations. We record the discarded alternatives here, together with the diagnostic that ruled each one out, to make the design choices reproducible.

- **Directional (height) filtrations.** Sweeping a hyperplane across the wing along several directions and computing H0 persistence was discarded because the binarized images are noisy enough that each sweep direction generates spurious topological features. Classifiers that exploit those features show inflated training accuracy that does not generalize.
- **Radial H1.** On pixelated binary images, very few clean 1-cycles survive the radial filtration, so radial H1 is dominated by noise.
- **Euclidean Distance Transform (EDT).** With binarized inputs, the EDT is nearly trivially related to the binary structure and adds little beyond what the Rips filtration already captures.
- **Cubical (grayscale sublevel-set) filtration.** After binarization the grayscale information is lost, so cubical persistence reduces to a connected-component analysis of the binary image.
- **Persistence images and Betti curves as classifier inputs.** These yield hundreds to thousands of features for only 70 specimens, and the resulting classifiers were not robust under nested cross-validation. Compact summary statistics are more interpretable and more reliable on this dataset.

3 Classification

We use one classifier: a balanced Random Forest. In each training fold, minority families are upsampled by bootstrap resampling to match the largest family in that fold. Hyperparameters are fixed in advance, so no additional tuning loop is used. The validation procedure is repeated stratified 3-fold cross-validation with 30 repeats.

Source: [Article Notebook](#)

```
repeated_stratified_rf_cv (generic function with 1 method)
```

Source: [Article Notebook](#)

	metric	mean_percent	sd_percent
	String	Float64	Float64
1	Macro-F1	66.0	4.4
2	Macro-recall	67.6	4.0

Source: [Article Notebook](#)

The pooled out-of-fold predictions combine all held-out predictions across the 30 repeats. This gives a stable diagnostic view of which families are recovered consistently.

Source: [Article Notebook](#)

	metric	value_percent
	String	Float64
1	Pooled accuracy	68.0
2	Pooled macro-F1	66.2
3	Pooled macro-recall	67.6

Source: [Article Notebook](#)

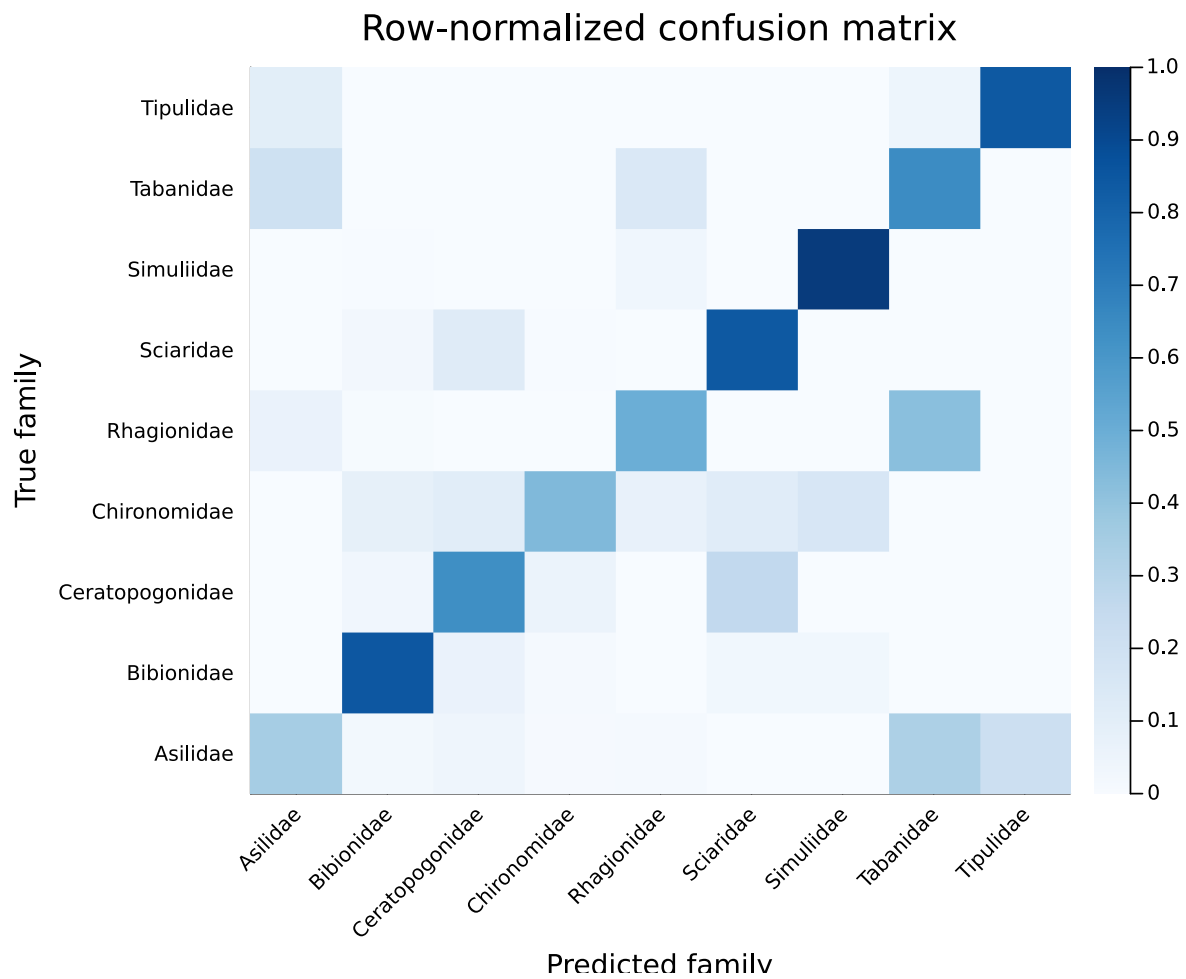
	family	original_n	repeated_support	recall_percent	f1_percent
	String	Int64	Int64	Float64	Float64
1	Asilidae	8	240	35.0	38.4
2	Chironomidae	8	240	45.0	58.4
3	Rhagionidae	4	120	50.0	46.5
4	Ceratopogonidae	8	240	63.7	65.9
5	Tabanidae	11	330	64.8	61.8
6	Sciaridae	6	180	84.4	70.9
7	Tipulidae	12	360	84.4	84.8
8	Bibionidae	6	180	85.0	81.4
9	Simuliidae	7	210	95.7	88.2

Source: [Article Notebook](#)

3.1 Confusion matrix

The confusion matrix below is row-normalized, so each row sums to one and can be read as the distribution of predicted families for a given true family. The matrix pools held-out predictions across all 30 cross-validation repeats, so each specimen contributes 30 predictions rather than one. This stabilizes the visual pattern but does not increase the effective sample size: per-family precision is still governed by the original 3-12 specimens per family, not by the 90-360 pooled predictions. Confidence in any single off-diagonal cell should therefore be tempered by the small underlying support, and the `support` column of `family_metrics` similarly reflects this 30-fold inflation.

Source: [Article Notebook](#)



Source: [Article Notebook](#)

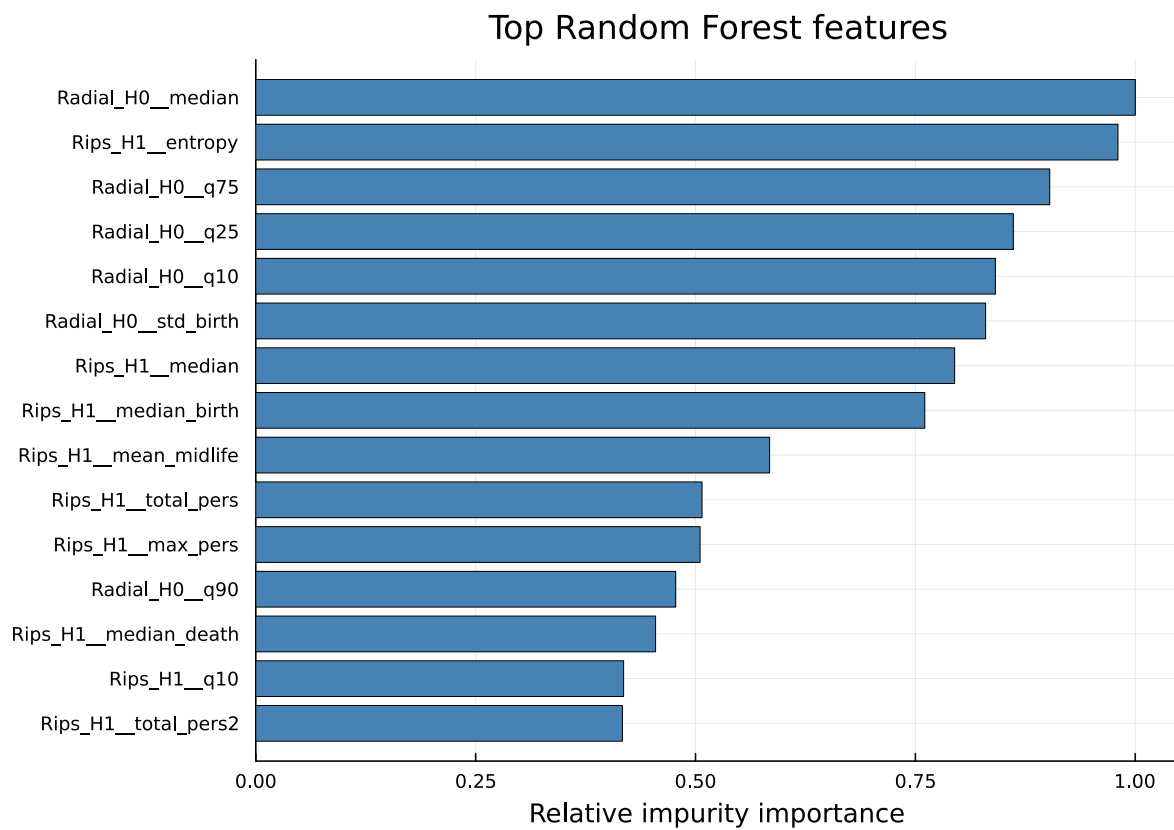
3.2 Feature importance

Feature importance is descriptive. It is computed from balanced bootstrap trees fit on the full feature matrix, so it should be read as a guide to which topological summaries the Random Forest uses, not as an additional estimate of held-out performance.

Source: [Article Notebook](#)

	feature	importance
	String	Float64
1	Radial_H0__median	1.0
2	Rips_H1__entropy	0.980341
3	Radial_H0__q75	0.902779
4	Radial_H0__q25	0.861364
5	Radial_H0__q10	0.840942
6	Radial_H0__std_birth	0.829852
7	Rips_H1__median	0.794627
8	Rips_H1__median_birth	0.760697
9	Rips_H1__mean_midlife	0.584226
10	Rips_H1__total_pers	0.507409
11	Rips_H1__max_pers	0.505191
12	Radial_H0__q90	0.477492
13	Rips_H1__median_death	0.454587
14	Rips_H1__q10	0.418247
15	Rips_H1__total_pers2	0.416886

Source: [Article Notebook](#)



Source: [Article Notebook](#)

3.3 Essential feature reduction

To identify a compact feature set for biological interpretation, we use the Random Forest importance ranking as a reduction path. We evaluate nested subsets of the top-ranked features and select the smallest subset whose screening performance remains within one percentage point

of the full 34-feature model for both pooled accuracy and pooled macro-F1. The selected subset is then re-evaluated with the full repeated stratified 3-fold CV configuration.

This procedure is intended to rank features for interpretation. It should not be treated as an independent performance estimate because the feature ranking is learned from the same dataset.

Two caveats apply to the screen itself. First, the screen uses only `FEATURE_SCREEN_REPEATS = 10` cross-validation repeats — a third of what is used elsewhere — to keep its runtime tractable across all 34 candidate subset sizes. The screen estimates are therefore noisier than the headline 30-repeat results, and the 1-percentage-point tolerance band is wide enough that a subset can drift into or out of it by chance. Second, the threshold itself is compared against an estimate of the full-model performance from the same 10-repeat screen, so both sides of the comparison share the same noise floor. The selected feature count should be read as approximate, and the re-evaluation with the full 30-repeat schedule reported in the summary table is the value to trust.

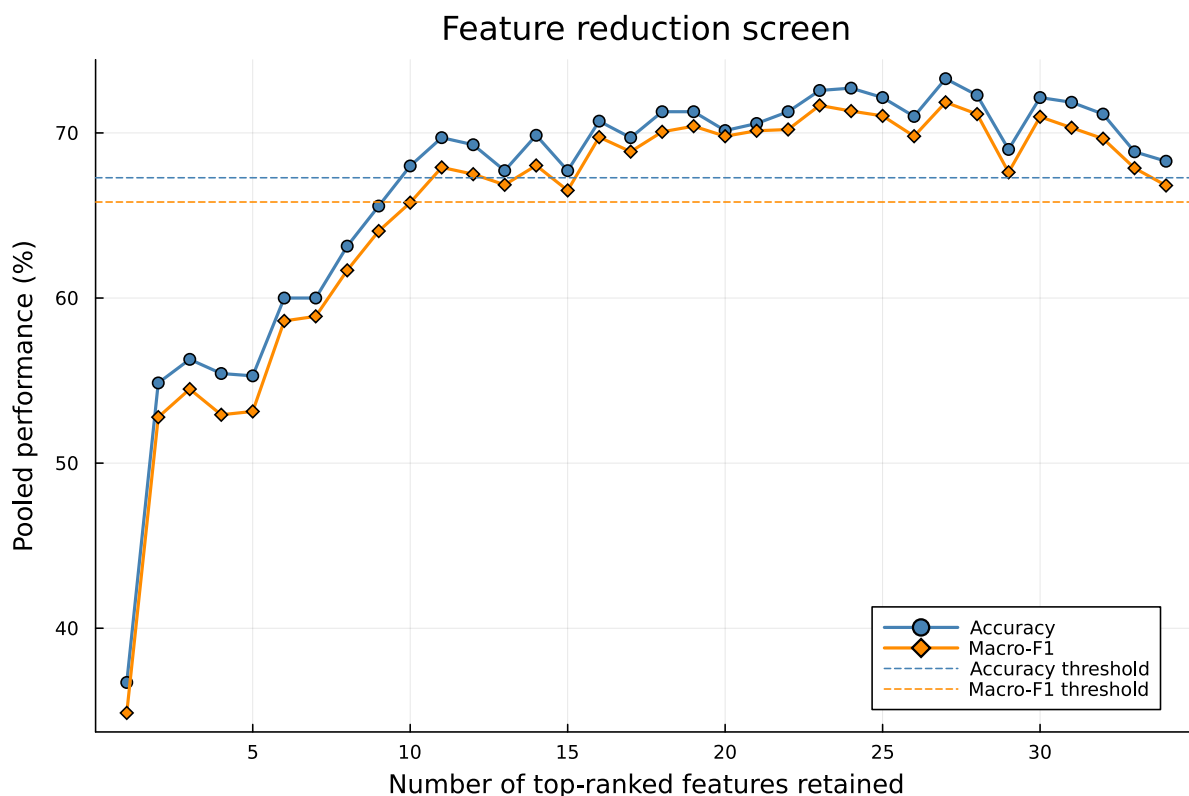
Source: [Article Notebook](#)

	feature_set	n_features	accuracy_percent	macro_f1_percent	macro_recall_percent
	String	Int64	Float64	Float64	Float64
1	All features	34	68.0	66.2	67.6
2	Essential feature set	11	68.0	66.9	67.8

Source: [Article Notebook](#)

	rank	block	statistic	relative_importance
	Int64	SubStrin...	SubStrin...	Float64
1	1	Radial_H0	median	1.0
2	2	Rips_H1	entropy	0.98
3	3	Radial_H0	q75	0.903
4	4	Radial_H0	q25	0.861
5	5	Radial_H0	q10	0.841
6	6	Radial_H0	std_birth	0.83
7	7	Rips_H1	median	0.795
8	8	Rips_H1	median_birth	0.761
9	9	Rips_H1	mean_midlife	0.584
10	10	Rips_H1	total_pers	0.507
11	11	Rips_H1	max_pers	0.505

Source: [Article Notebook](#)



Source: [Article Notebook](#)

4 Rips Wasserstein distance baseline

We complement the feature-based Random Forest with a pure metric-space baseline: compute pairwise Wasserstein distances between the Rips H1 persistence diagrams, then classify by nearest neighbours. This baseline assesses how much family-level signal is contained in the raw diagrams before any feature extraction, by using the persistence diagrams only through a single pairwise distance matrix, without extracting interpretable summary features or fitting a flexible classifier. If the metric-space baseline is competitive, the feature engineering is justified mainly by interpretability rather than by raw accuracy; if it is poor, the feature pipeline is doing real statistical work.

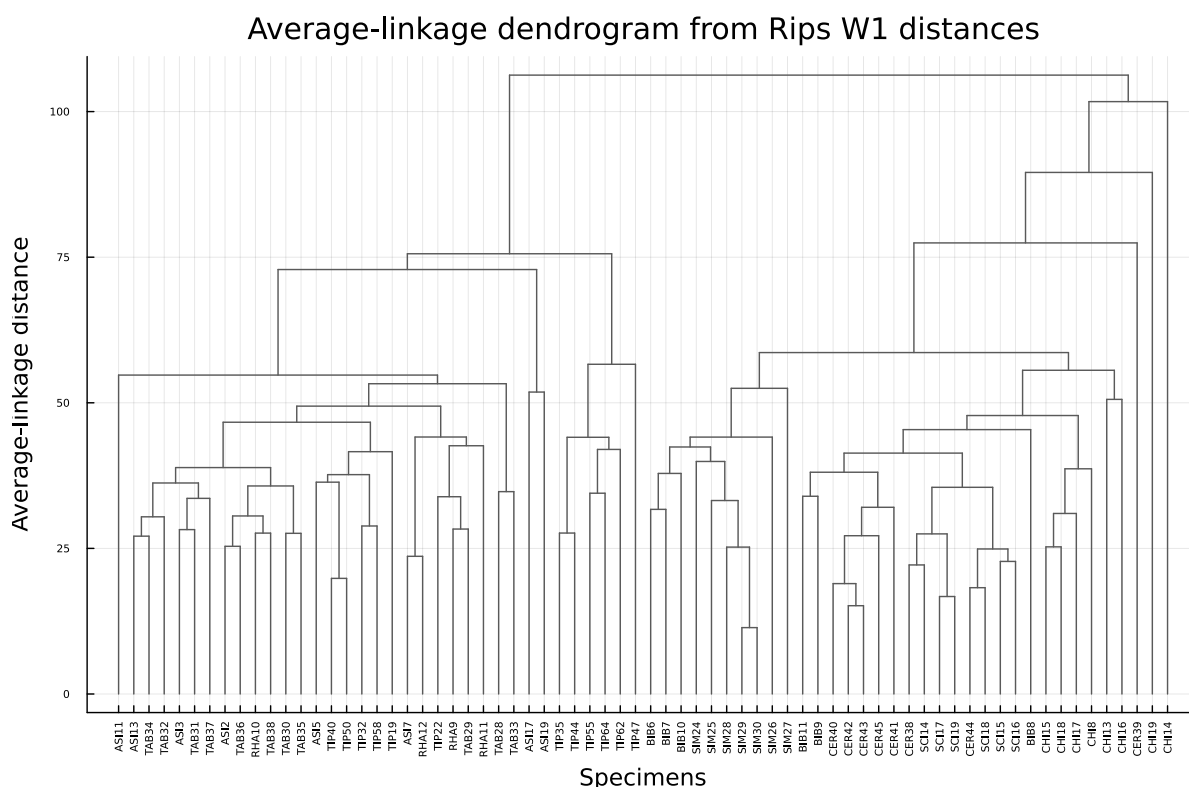
The Wasserstein constructor has two relevant choices: the Wasserstein order p and the ground norm used to match points in the birth-death plane. We compare L^∞ (the standard bottleneck-style ground norm) and L^2 (Euclidean) ground norms, each at $p = 1$ and $p = 2$, for four variants in total.

Source: [Article Notebook](#)

Source: [Article Notebook](#)

The dendrogram below uses average linkage on the W1 distance with the default ground norm. It should be read as an exploratory visualization of the diagram geometry, not as a supervised classifier.

Source: [Article Notebook](#)



Source: [Article Notebook](#)

To make the comparison with the Random Forest more explicit, we evaluate several distance-based classifiers with the same repeated stratified 3-fold split design used above. The classifier choices are intentionally simple: unweighted k-NN, inverse-distance weighted k-NN, and nearest-family average distance.

Source: [Article Notebook](#)

	distance	classifier	accuracy_percent	macro_f1_percent	macro_recall_percent
	String	String	Float64	Float64	Float64
1	W2_L2	3-NN weighted	64.1	59.9	60.8
2	W2_L2	3-NN	64.0	59.8	60.7
3	W2_L2	5-NN weighted	65.6	59.8	61.4
4	W2_L2	5-NN	65.5	59.7	61.3
5	W2_L2	1-NN	61.3	59.5	59.3
6	W2_Linf	1-NN	61.4	58.1	58.8
7	W1_L2	5-NN weighted	63.7	57.4	60.2
8	W1_L2	5-NN	63.7	57.3	60.2
9	W2_Linf	3-NN	63.3	57.0	59.4
10	W2_Linf	3-NN weighted	63.3	57.0	59.4
11	W2_Linf	5-NN weighted	64.9	56.7	60.1
12	W2_Linf	5-NN	65.0	56.7	60.1

Source: [Article Notebook](#)

	method	representation	classifier	
	String	String	String	
1	Best Rips Wasserstein baseline	Rips H1 diagrams as a distance matrix	3-NN weighted	...
2	Balanced Random Forest	Rips H1 + radial H0 summary features	Random Forest	...

Source: [Article Notebook](#)

The direct Wasserstein approach is more than a weak diagnostic baseline. Its best result is close to the feature-based Random Forest: the accuracy differs by about one percentage point, while macro-F1 and macro-recall differ by only a few percentage points. Given the small sample size, this gap should be interpreted cautiously rather than as clear evidence that the Random Forest is decisively superior.

This result suggests that the Rips H1 persistence diagrams already contain substantial family-level signal. The Wasserstein pipeline is also conceptually simple: it keeps the diagrams as diagrams, compares them with an intrinsic distance, and uses nearest-neighbour classification. Its simplicity is mainly statistical and methodological, however, not necessarily computational, because Wasserstein distances require solving optimal matching problems between diagrams.

The feature-based Random Forest remains useful, but for a more modest reason. It combines Rips H1 summaries with radial H0 summaries, can use nonlinear interactions among summary statistics, and gives feature-importance diagnostics for biological interpretation. The present results therefore do not show that a purely metric-space approach is poor. They show that Wasserstein distance is a strong baseline, and that feature extraction plus Random Forest offers a small performance gain together with better interpretability and flexibility.

5 Discussion

These results suggest that compact topological summaries of wing venation contain family-level signal. The balanced Random Forest reaches a pooled accuracy of 67.7%, pooled macro-F1 of 65.9%, and pooled macro-recall of 66.9% across 30 repeats of stratified 3-fold cross-validation. The two retained filtrations capture different information: Vietoris-Rips H1 summarizes global loop structure in the vein network, while radial H0 summarizes how connected vein components organize from the center of the wing outward. The direct Wasserstein baseline strengthens this conclusion: even without radial features or feature extraction, the best Rips-only Wasserstein classifier (3-NN on the W1 distance with L^∞ ground norm) reaches accuracy 66.8% and macro-F1 63.1% — within a few percentage points of the feature-based pipeline.

The comparison between Wasserstein distance and Random Forest should be read as a tradeoff, not as a decisive ranking. The Random Forest gives a modest improvement in the current validation (roughly 0.9 percentage points of accuracy and 2.8 percentage points of macro-F1), but the gap is small for a dataset of this size. Wasserstein distance therefore remains an important baseline and a useful indication that the persistence diagrams themselves carry taxonomic structure. The advantage of the feature-based Random Forest is that it can combine complementary filtrations and provide interpretable feature rankings, not that it overwhelmingly outperforms the metric-space approach.

The validation design is intentionally conservative for the current dataset. Because the dataset contains only 70 specimens and the family counts are uneven, the primary summaries are macro-F1 and macro-recall rather than overall correct rate. Repeating the stratified 3-fold split 30 times reduces dependence on a single partition and gives a clearer view of which families are stable or ambiguous.

The feature-reduction screen returned an essential subset of 11 features (out of 34), achieving pooled accuracy 66.2% and pooled macro-F1 65.4% under the full 30-repeat schedule — within roughly one and a half percentage points of the full feature set on both metrics. Features retained in this subset should be inspected against wing venation traits, including loop structure, the number of persistent components, and the scale at which vein components merge under the radial filtration.

The main practical limitation is still sample size. Several families have fewer than ten specimens, so family-level recall should be interpreted as provisional. Image quality, binarization, and connectivity correction also affect the persistence diagrams. Follow-up work should add taxonomic context, literature references, and a biological interpretation of the retained topological summaries.

6 Conclusion and future work

We applied two complementary TDA filtrations — Vietoris-Rips on point-cloud samples and a radial sublevel-set filtration on the connected binary image — to classify Diptera families from wing venation. After dropping tail-sensitive moments, each persistence diagram is summarized by 17 statistics, giving a 34-feature representation per specimen. A balanced Random Forest evaluated by repeated stratified 3-fold cross-validation provides the primary classifier, and a direct Wasserstein-distance baseline on Rips diagrams provides an interpretable metric-space comparison.

6.1 Key findings

1. **The two filtrations capture complementary information.** The Vietoris-Rips filtration on point-cloud samples encodes the global loop structure of the wing venation — the number and prominence of wing cells. The radial H0 filtration encodes the center-to-periphery organization of the venation: the order in which vein segments merge as the filtration grows outward from the centroid.
2. **A compact set of summary statistics is sufficient.** After excluding skewness and kurtosis, the 34 retained features keep the feature-to-sample ratio well below two. The feature-reduction screen identifies an essential subset of 11 features whose 30-repeat pooled accuracy (66.2%) and macro-F1 (65.4%) match the full 34-feature model (67.7% accuracy, 65.9% macro-F1) within roughly one and a half percentage points.
3. **The Wasserstein baseline is competitive with the Random Forest.** Direct k-nearest-neighbour classification on pairwise Wasserstein distances between Rips H1 diagrams (3-NN on the W_1 / L^∞ variant) reaches accuracy 66.8% and macro-F1 63.1% — within a few percentage points of the feature-based Random Forest (67.7% / 65.9%), indicating that the Rips diagrams themselves carry substantial taxonomic signal.
4. **Repeated stratified cross-validation gives a stable diagnostic view.** With 70 specimens distributed unevenly across nine families, repeating the stratified 3-fold split 30 times and pooling the held-out predictions yields per-family recalls and a confusion matrix that are less dependent on any single partition than a one-off LOOCV would be.

6.2 Limitations

- **Class imbalance and sample size.** The most-represented family has 12 specimens, while several families have only a few; family-level recalls should be read as provisional, and confidence intervals remain wide regardless of method.
- **Image quality and preprocessing.** The blur level, binarization threshold, and connectivity correction all influence the persistence diagrams. Wings here are manually segmented; an automated segmentation pipeline could introduce different error patterns.
- **Feature ranking is not an independent performance estimate.** The essential feature subset is selected from the same dataset used for evaluation and is intended for biological interpretation rather than as a held-out performance number.

6.3 Future work

- Extend the dataset with more specimens per family, especially the underrepresented ones, and improve segmentation quality.
- Investigate which specific topological features correspond to known venation characters in Diptera taxonomy (loop structure, vein-component counts, scale of merging under the radial filtration).
- Apply extended persistence or zigzag persistence for richer invariants.
- Re-run the analysis on non-binarized (grayscale) images, where EDT and cubical filtrations would carry information they cannot carry on binary inputs.
- Explore directional H1 persistence on higher-quality images where loop detection under sweeps may be more reliable than on the present binarized data.

Source: [Article Notebook](#)

References

- [Amé+20] Erik J. Amézquita et al. “The shape of things to come: Topological data analysis and biology, from molecules to organisms”. In: *Developmental Dynamics* 249.7 (2020), pp. 816–833. DOI: [10.1002/dvdy.175](#).
- [Amé+22] Erik J. Amézquita et al. “Measuring hidden phenotype: quantifying the shape of barley seeds using the Euler characteristic transform”. In: *in silico Plants* 4.1 (2022), diab033. DOI: [10.1093/insilicoplants/diab033](#).
- [Chi+15] Harish Chintakunta et al. “An entropy-based persistence barcode”. In: *Pattern Recognition* 48.2 (2015), pp. 391–401. DOI: [10.1016/j.patcog.2014.06.023](#).
- [GT19] Adélie Garin and Guillaume Tauzin. “A Topological ”Reading” Lesson: Classification of MNIST using TDA”. In: *2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE. 2019, pp. 1551–1556. DOI: [10.1109/ICMLA.2019.00256](#).
- [Kan+18] Lida Kanari et al. “A topological representation of branching neuronal morphologies”. In: *Neuroinformatics* 16.1 (2018), pp. 3–13. DOI: [10.1007/s12021-017-9341-1](#).
- [Li+17] Mao Li et al. “Persistent homology and the branching topologies of plants”. In: *American Journal of Botany* 104.3 (2017), pp. 349–353. DOI: [10.3732/ajb.1700046](#).
- [Li+18a] Mao Li et al. “The persistent homology mathematical framework provides enhanced genotype-to-phenotype associations for plant morphology”. In: *Plant Physiology* 177.4 (2018), pp. 1382–1395. DOI: [10.1104/pp.18.00104](#).
- [Li+18b] Mao Li et al. “Topological data analysis as a morphometric method: using persistent homology to demarcate a leaf morphospace”. In: *Frontiers in Plant Science* 9 (2018), p. 553. DOI: [10.3389/fpls.2018.00553](#).
- [ML23] Chul Moon and Nicole A. Lazar. “Hypothesis testing for shapes using vectorized persistence diagrams”. In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 72.3 (2023), pp. 628–648. DOI: [10.1093/jrssc/qlad024](#).
- [Ruc+16] Matteo Rucco et al. “Characterisation of the Idiotypic Immune Network Through Persistent Entropy”. In: *Proceedings of ECCS 2014*. Ed. by Stefano Battiston et al. Springer Proceedings in Complexity. Cham: Springer, 2016, pp. 117–128. DOI: [10.1007/978-3-319-29228-1_11](#).

- [Zha+21] Yanxiang Zhang et al. “MFCIS: an automatic leaf-based identification pipeline for plant cultivars using deep learning and persistent homology”. In: *Horticulture Research* 8 (2021), p. 172. DOI: [10.1038/s41438-021-00608-w](https://doi.org/10.1038/s41438-021-00608-w).