Identification of the Species Constituents of Maggot Populations Feeding on Decomposing Remains—Facilitation of the Determination of Post Mortem Interval and Time Since Tissue Infestation through Application of Machine Learning and Direct Analysis in Real Time-Mass Spectrometry.

Samira Beyramysoltan,[1] Mónica I. Ventura,[1] Jennifer Y. Rosati,[2] Justine E. Giffen-Lemieux [1] and Rabi A. Musah[1*]

[1]Department of Chemistry, University at Albany, State University of New York, 1400 Washington Avenue, Albany, NY 12222, USA

[2] John Jay College of Criminal Justice, 524 West 59th St, New York, NY 10019, USA

*Corresponding author: rmusah@albany.edu

SUPPORTING INFORMATION

Contained within this supporting information document are 4 additional figures, one Scheme and 5 tables referenced in the text: Aggregation of blow fly larvae on a pig carcass; representative DART-HRMS spectra representing various species of larvae; the merits plots of the aggregated hierarchical conformal predictor in analysis of the actual and *y*-permuted data for class assignment in the hypothesized significance level threshold range of 0-0.2; the algorithm of the aggregated hierarchical predictor; makeup of the 70% aqueous ethanol suspensions representing mixture types; information on the input data and parameters for the learning of the neural networks; classification tree performance for all calibration and test samples using the com-mon strategies for accuracy computation; and percentages of errors, multiple predictions, not assigned predictions, and false positive and false negative rates at the threshold significance levels.

Supporting Figures.

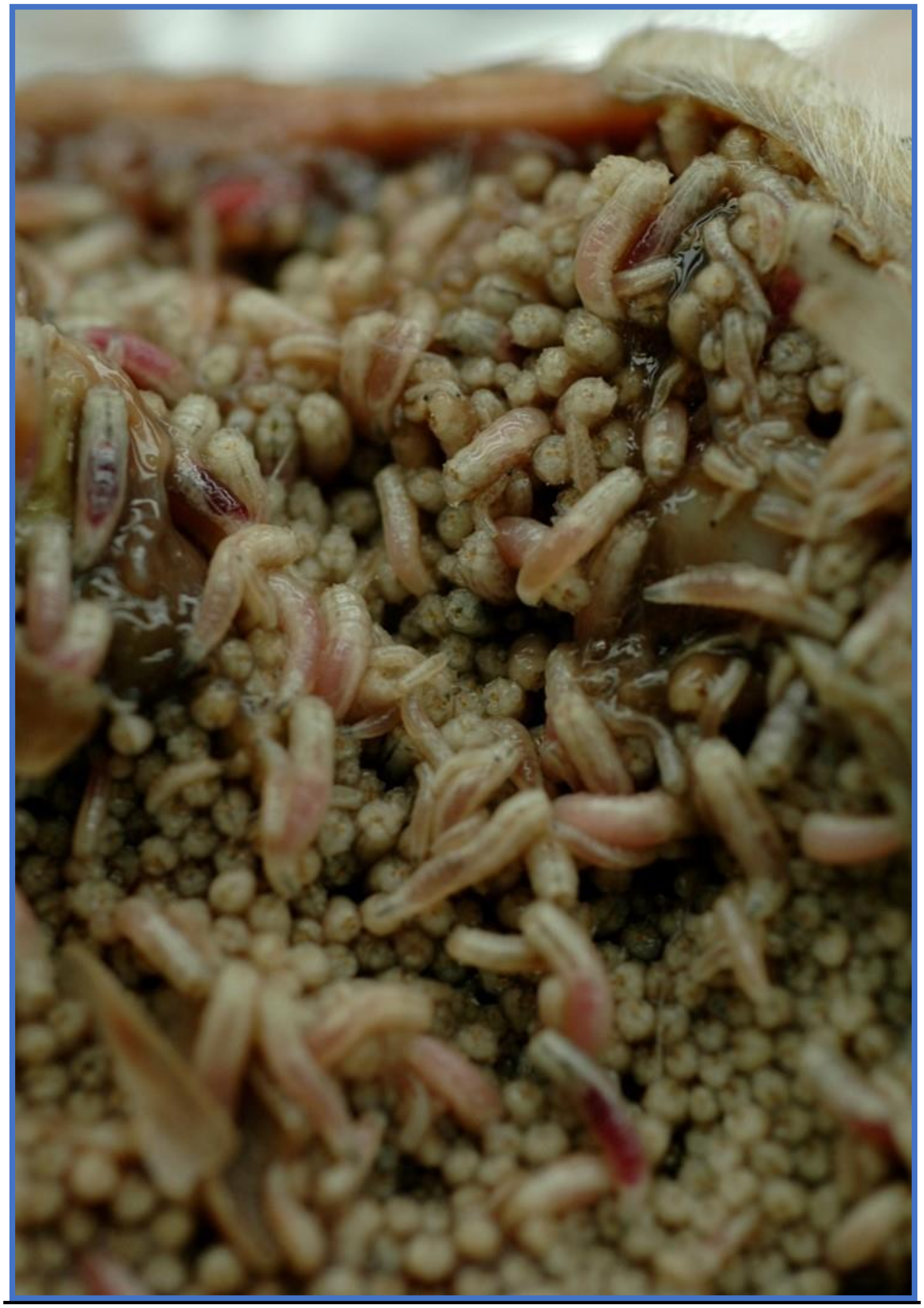

Figure S1. Close up image of blow fly larvae colonization of a piglet carcass.

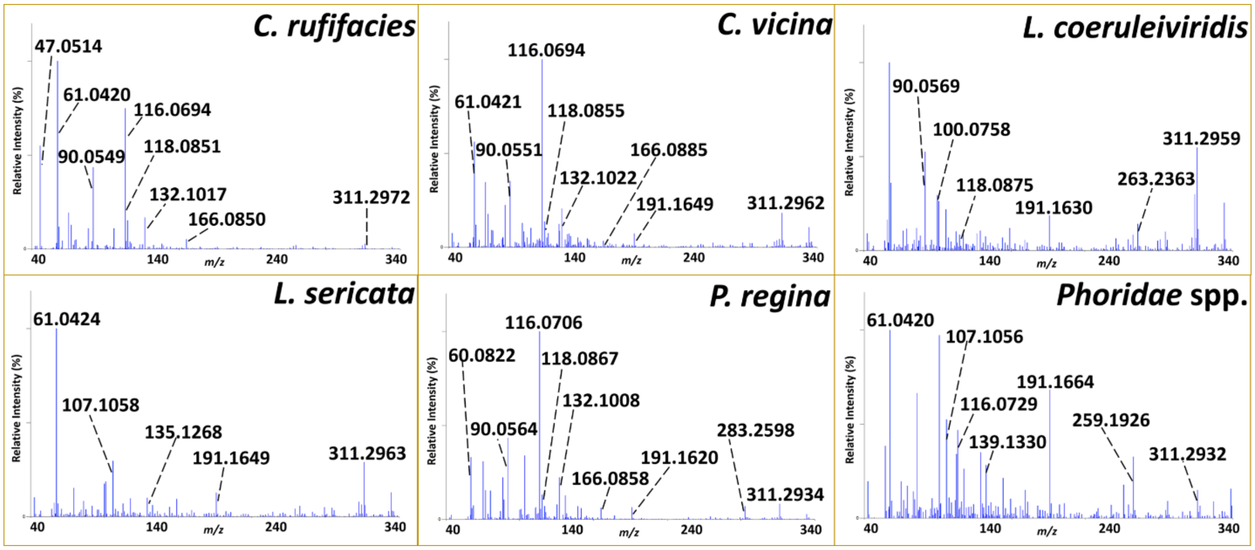

Figure S2. Results of DART-HRMS analyses of aqueous ethanol suspensions of six species of necrophagous fly larvae. All analyses were performed in positive-ion mode at 350 °C. Each panel represents one individual species and is an average of 5 analyses for each sample. The spectra generated for each species are unique and can be used as a fingerprint for the identification of an individual species. The figure is reprinted with permission from: Beyramysoltan, S., Giffen, J. E., Rosati, J. Y. & Musah, R. A. Direct analysis in real time-mass spectrometry and kohonen artificial neural networks for species identification of larva, pupa and adult life stages of carrion insects. Anal. Chem. 90, 9206-9217, doi:https://doi.org/10.1021/acs.analchem.8b01704 (2018). Copyright © 2018 American Chemical Society.

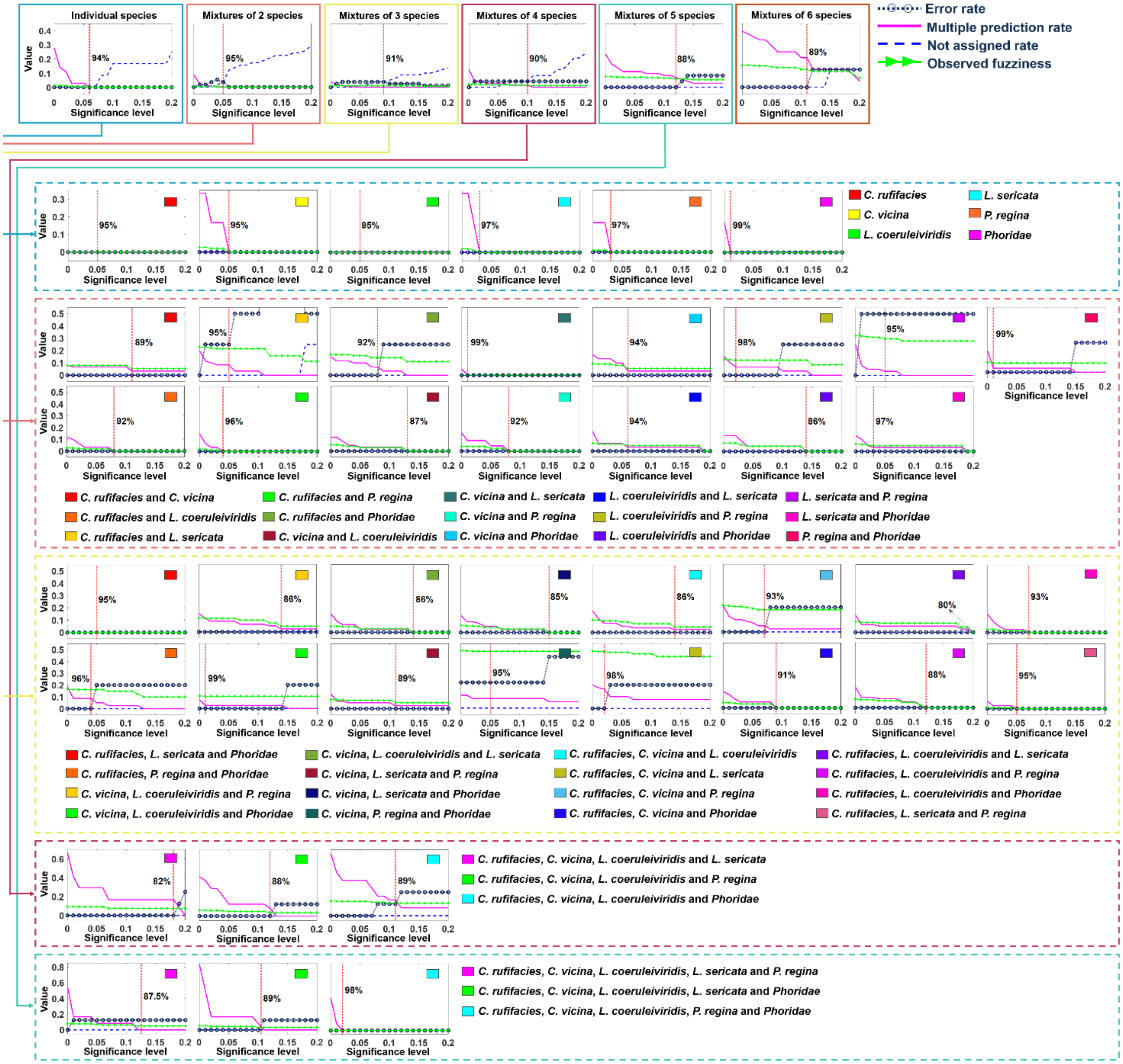

Figure S3. Merits of the conformal predictor (i.e. not assigned, multiple species prediction, observed fuzziness and error rates) for class assignment for the hypothesized significance level threshold range of 0-0.2 for local validity of class labels in the classification nodes of the classification tree. As shown by the dashed vertical lines in each plot, a specific significance level was defined for assignment of samples to a specific class in the hierarchical classification tree.

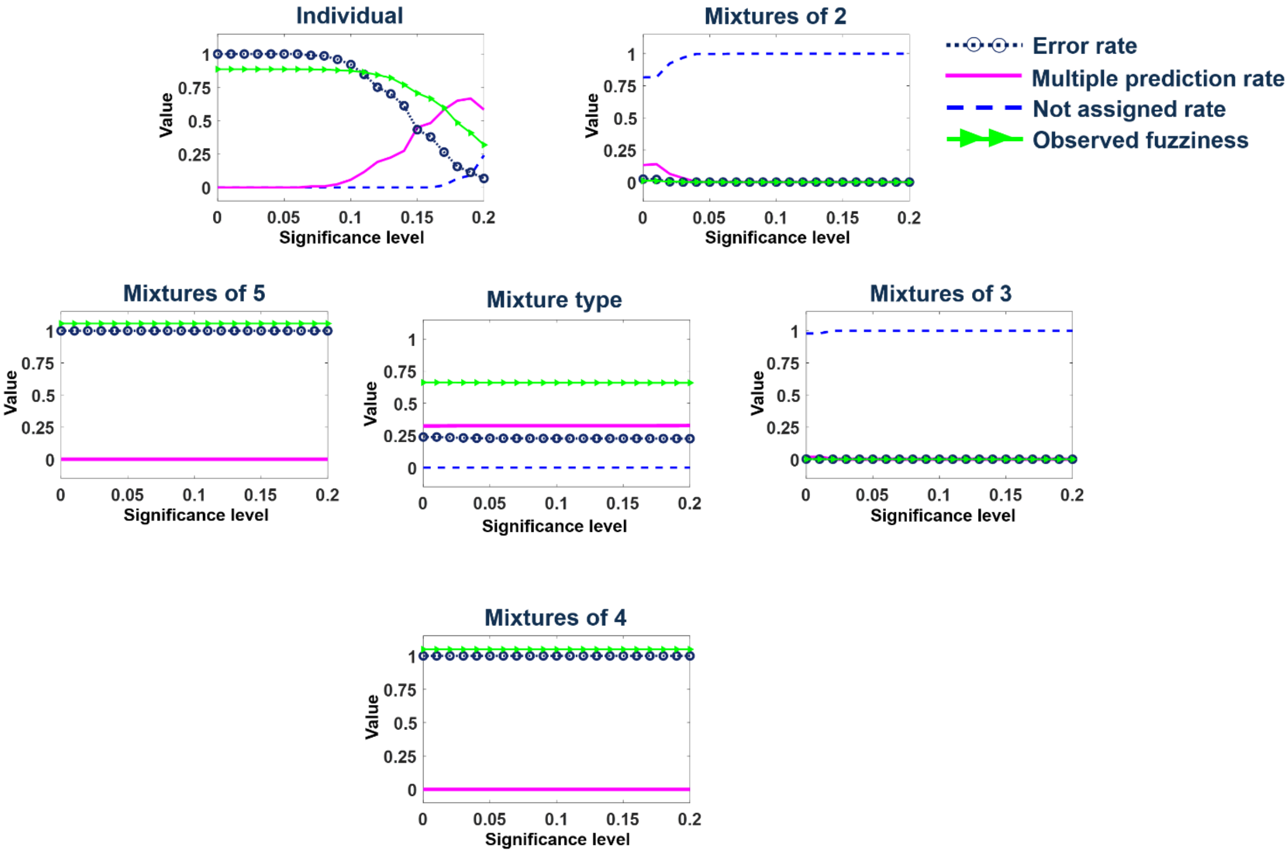

Figure S4. Efficiency and validity merits of the conformal predictor (i.e. the not assigned, multiple prediction, observed fuzziness and error rates) in the analysis of 10 *y*-permuted data sets, for class assignment at the hypothesized significance level threshold range of 0-0.2. The illustrated results show the overall validity in the classification nodes of the tree for the two discrimination levels (i.e. mixture type at the first discrimination level, and at the second level, the components of each mixture type, such as single component, and mixtures of 2, 3, 4, and 5 species).

Scheme S1. The algorithm of the aggregated hierarchical predictor.

```
Input: A data set: $\{(x_1, y_1), \dots, (x_{n-1}, y_{n-1})\}$, where n represents the number of objects and the new sample $(x_n, y_n)$

Perform bootstrap resampling without replacement and provide $q$ number of training and calibration sets, designated as t and c respectively.
The bootstrap output is: $q$ training sets $\{(x_1, y_1), \dots, (x_{n1}, y_{n1})\}_t$, and $q$ calibration sets $\{(x_{n1+1}, y_{n1+1}), \dots, (x_{n-1}, y_{n-1})\}_c$,

Dt=$\{(x_1, l_1), \dots, (x_{n1}, l_{n1})\}_t$
Dc=$\{(x_{n1+1}, l_{n1+1}), \dots, (x_{n-1}, l_{n-1})\}_c$

for $i\_b = 1$ through $q$:
    Train feed-forward network using: Dt
    Calculate the non-conformity measure, $\alpha_{global} = \mathrm{A}(\mathrm{Dc})$
                        Non-conformity score, α (details in Scheme S3-Appendix 1)
    Calculate $p_{\mathrm{global}}$ for $(x_n, y_n)$ for class $k\_g$, where $k\_g \in \{1, \dots, \mathrm{n}c\}$

    $p(\alpha^{k\_g}_{global,i\_b}) = \left[\frac{\mathrm{count}\{i \in \{\mathrm{n1}+1, \dots, n-1\} \mid y_i = k\_g \& \alpha_i^{k\_g} \geq \alpha_n^{k\_g}\}}{\mathrm{count}\{i \in \{\mathrm{n1}+1, \dots, n-1\} \mid y_i = k\_g\}}\right]_{i\_b}$

    for $i\_t$=1 through $l$, where $l$ refers to mixture types
        Dt=filter(Dt, mixture type)
        Dc=filter(Dc, mixture type)
        Train feed-forward network using: Dt
        Calculate the non-conformity measure, $\alpha_{\ type} = A(\mathrm{Dc})$
        Calculate $p_{\mathrm{type}}$ for $(x_n, y_n)$ for class $k\_t$, where $k\_t \in \{1, \dots, \mathrm{n}c\}$

        $p(\alpha^{k\_t}_{type,i\_b}) = \left[\frac{\mathrm{count}\{i \in \{\mathrm{n1}+1, \dots, n-1\} \mid y_i = k\_t \& \alpha_i^{k\_t} \geq \alpha_n^{k\_t}\}}{\mathrm{count}\{i \in \{\mathrm{n1}+1, \dots, n-1\} \mid y_i = k\_t\}}\right]_{i\_b}$
    end
end

$p\left(\alpha^{\mathrm{k\_g}}_{global}\right) = \frac{\sum_{i\_b}(p(\alpha^{k\_g}_{global,i\_b}))}{l}$
$p\left(\alpha^{\mathrm{k\_t}}_{\mathrm{type}}\right) = \frac{\sum_{i\_b}(p(\alpha^{k\_t}_{type,i\_b}))}{l}$
If $p\left(\alpha^{\mathrm{k\_g}}_{global}\right) > \varepsilon^{k\_g}$ and $p\left(\alpha^{\mathrm{k\_t}}_{\mathrm{type}}\right) > \varepsilon^{k\_t}$
    $\tau_{global} = [k\_g], \tau_{type} = [k\_t]$, where $\tau$ defines the prediction region
else
    $\tau = \varphi$
end
```

Scheme S1-Appendix 1. Non-conformity measures

Three non-conformity measures (shown in Eq. S1-3) were used in this study. In these equations, *o* defines the prediction output of the neural network, and *nc*, *k* and *i* are the number of classes, class indices and sample numbers respectively.

$$\alpha_i^k = 1 - \frac{o_i^k}{\sum_{j\in\{1,\dots,nc\}} o_i^j} = 1 - o_i^k \tag{S1}$$

$$\alpha_i^k = \frac{max_{j\in\{1,\dots,nc\},j\neq k} o_i^j}{o_i^k} \tag{S2}$$

$$\alpha_i^{1:nc} = \text{Euclidean distance}(o_i, \mathrm{I}(nc \times nc)), where\ \mathrm{I}\ is\ identity\ matrix \tag{S3}$$

Scheme S1-Appendix 2. Performance merits of conformal predictor

A predictor is considered to have made an error when the predicted region does not contain the true label, and the error rate refers to the number of observations predicted incorrectly. The NA rate reveals the fraction of samples that are not assigned to classes within the defined prediction region. The multiple prediction rate is the size of the multiple predictions region (around the tested samples which involve the true label), and the OF is defined as the sum of all *p*-values for the incorrect class labels. The NA rate, E-criterion, and OF equations are presented in Eq. S4-6 respectively, where *nt* indicates the number of samples in the test set (10% of the samples in the data matrix); and $|\tau_i^\varepsilon|$ denotes the size of the prediction region, where τ_i^ε is the prediction region of sample *i* at significance level *ε*.

$$\mathrm{NA} = \frac{1}{nt}\sum_{i=1}^{nt} 1_{\{|\tau_i^\varepsilon|=\emptyset\}}, \tag{S4}$$

$1_{\{|\tau_i^\varepsilon|=\emptyset\}}$ designates the indicator function of $\{|\tau_i^\varepsilon| = \emptyset\}$ (it has value 1 if $\{|\tau_i^\varepsilon| = \emptyset\}$ happens and 0 if not)

$$\mathrm{E-criteria} = \frac{1}{nt}\sum_{i=1}^{nt} 1_{\{|\tau_i^\varepsilon|>1\}} \frac{|\tau_i^\varepsilon|}{nc}, i \in \{1, \dots, nt \,\big|\, |\{y_i\} \in \tau_i^\varepsilon|\},$$

where y_i is true label of sample I (S5)

$1_{\{|\tau_i^\varepsilon|>1\}}$ denotes the indicator function of $\{|\tau_i^\varepsilon| > 1\}$ (it has value 1 if $\{|\tau_i^\varepsilon| > 1\}$ happens and 0 if not)

$$\mathrm{OF} = \frac{1}{nt}\sum_{i=1}^{nt} \sum_{y\neq y_i} p_i^y\,, i \in \{1, \dots, nt\} \tag{S6}$$

Scheme S1-Appendix 3. The general sequence of steps in Scheme 3

The general sequence of steps in Scheme 3 involved: (1) Inputting of the data and its corresponding labels; (2) Applying 100 × bootstrap resampling without replacement to provide q number of training and calibration sets; (3) Within each bootstrap iteration (i_b): (a) training of a hierarchical classification tree which was then used for prediction of test and calibration samples; (b) computing the non-conformity measure for the test and calibration samples; and (c) estimating the *p*-values based on the comparison of the non-conformity measures for test and calibration samples. For hypothetical classes defined as: k_g (for mixture-type determination) and k_t (for determination of the identities of the species within a mixture type), two label conditional *p*-values, termed p_{global} ($p(\alpha^{k\_g}_{global,i\_b})$) and p_{type} ($p(\alpha^{k\_t}_{type,i\_b})$) were computed for making assignments for each test sample; and (4) using the average of the resulting calculated *p*-values acquired from the 100 × bootstrap resampling, for class assignment of test samples.

Supporting Tables.

Table S1. Makeup of the 70% aqueous ethanol suspensions representing mixture types of the 6 species *C. rufifacies*, *C. vicina*, *L. coeruleiviridis*, *L. sericata*, *P. regina*, and *Phoridae*. Columns A-F indicate the components (the species types) within each mixture type. The columns in the table display the volume percent of mixture sample components in a total volume of 25 µL. For instance, for the mixture type *C. rufifacies* and *C. vicina* (i.e. a two species mixture), seven solutions were made in the following proportions of A:B:70% aqueous ethanol respectively: 20:80:0; 45:45:10; 80:20:0; 45:10:45; 10:45:45; 5:10:85; and 10:5:85.

Sample Number	%A	%B	%C	%D	%E	%F	70% aqueous ethanol
Mixtures of 2 species							
1	20	80					0
2	45	45					10
3	80	20					0
4	45	10					45
5	10	45					45
6	5	10					85
7	10	5					85
Mixtures of 3 species							
1	10	80	10				0
2	10	10	80				0
3	80	10	10				0
4	45	45	10				0
5	10	45	45				0
6	45	10	45				0
7	33	33	33				1
8	90	5	5				0
9	80	10	5				5
Mixtures of 4 species							
1	10	10	10	10			60
2	25	25	25	25			0
3	45	10	10	10			25
4	10	45	10	10			25
5	10	10	45	10			25
6	10	10	10	45			25
7	70	10	10	10			0
8	10	70	10	10			0
9	10	10	70	10			0
10	10	10	10	70			0
11	10	10	25	55			0
12	55	25	10	10			0
13	10	55	25	10			0
14	25	10	55	10			0
15	55	10	10	25			0
Mixtures of 5 species							
1	5	5	5	5	5		75
2	80	5	5	5	5		0
3	5	80	5	5	5		0
4	5	5	80	5	5		0
5	5	5	5	80	5		0
6	5	5	5	5	80		0
7	40	40	5	5	5		5
8	40	5	40	5	5		5
9	40	5	5	40	5		5
10	40	5	5	5	40		5
11	20	20	20	20	20		0
12	30	5	30	5	30		0
13	30	30	30	5	5		0
14	5	30	5	30	30		0
15	5	30	30	30	5		0
16	5	10	30	10	5		40
Note: Total volume of the prepared samples is 25 µL.							

Table S1 (continued).

Sample Number	%A	%B	%C	%D	%E	%F	70% aqueous ethanol
Mixtures of 6 species							
1	5	5	5	5	5	5	70
2	70	5	5	5	5	10	0
3	5	70	5	10	5	5	0
4	5	5	70	5	5	10	0
5	10	5	5	70	5	5	0
6	5	5	10	5	70	5	0
7	5	10	5	5	5	70	0
8	30	10	30	10	10	10	0
9	30	10	10	30	10	10	0
10	30	10	10	10	30	10	0
11	20	20	20	20	10	10	0
12	30	20	30	5	10	5	0
13	10	10	30	10	10	30	0
14	10	30	5	10	10	30	5
15	40	20	20	10	5	5	0
16	15	15	15	15	15	15	10
Note: Total volume of the prepared samples is 25 µL.							

Table S2. Dimensions of the training, calibration and test datasets for the neural networks in each iteration of the bootstrap resampling in the nodes representing the mixture types, and the components of the mixtures of 2, 3, 4, 5 and 6 species. The optimized hyperparameters for each node, acquired by the Bayesian method (i.e. the hidden layer size, learning rate, and regularization parameter), are displayed.

Model	**Data sets**			**Optimized hyperparameters**		
	Training dataset	Calibration data matrix	Test data matrix	Hidden layer size	Learning rate	Regularization parameter
Mixture type	1260×350	423×350	206×350	15	0.012	0.5
Individual species	60×350	18×350	12×350	14	0.003	0.5
Mixtures of 2 species	347×350	120×350	58×350	44	0.001	0.5
Mixtures of 3 species	482×350	160×350	80×350	37	0.025	0
Mixtures of 4 species	156×350	53×350	24×350	30	0.986	0.55
Mixtures of 5 species	162×350	54×350	24×350	36	0.978	0

Table S3. Classification tree performance for all calibration samples in the 100 × bootstrap resampling. Significance metrics for the learning of the model, featuring actual data versus 10 × *y*-permuted data are presented. Each row shows the results associated with the indicated classification node of the classification tree. The three metrics used to compute the prediction accuracy of the neural networks are: max values of the neural network response; and response values at the significance thresholds of >0.5 and > 0.4.

Models	**Actual data**			***y*-Permuted data**		
	Max value	**> 0.5**	**> 0.4**	**Max value**	**> 0.5**	**> 0.4**
Mixture type	94.4	92.6	94.3	38.2	0	0.06
Individual species	99 .2	98.4	99	17	0	0
Mixtures of 2 species	85.7	76.9	82.6	0.07	0	0
Mixtures of 3 species	79.9	66.9	74.1	0.07	0	0
Mixtures of 4 species	82.4	77.7	81.4	33.4	0.02	17.2
Mixtures of 5 species	88	85	88	33	0.02	16.5

Table S4. Classification tree performance for all test samples in the 100 × bootstrap resampling. Significance metrics for the learning of the model, featuring actual data versus 10 × *y*-permuted data are presented. Each row shows the results associated with the indicated classification node of the classification tree. The three metrics used to compute the prediction accuracy of the neural networks are: max values of the neural network response; and response values at the significance thresholds of >0.5 and > 0.4.

Models	**Actual data**			***y*-permuted data**		
	Max value	**> 0.5**	**> 0.4**	**Max value**	**> 0.5**	**> 0.4**
Mixture type	93.2	91	92.7	38	0	0
Individual species	100	100	100	17	0	0
Mixtures of 2 species	82.8	79.3	82.8	0.04	0	0
Mixtures of 3 species	88.8	72.5	82.5	0.07	0	0
Mixtures of 4 species	75	75	75	35	0	0
Mixtures of 5 species	91.7	91.7	91.7	36.7	0	0

Table S5. Error percentages, multiple predictions, and not assigned predictions at the indicated significance thresholds for each label assignment in the classification tree. The corresponding false positive and negative rates are also reported. Gray shading reports the values associated with the first discrimination level, while the colored cells show values associated with the second level of discrimination.

Discrimination level		**Specification**					
First	**Second**	Significance level	Error rate	Multiple prediction	Not assigned	fp*	fn**
Single species		**0.060**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
	C. rufifacies	**0.050**	0.000	0.000	0.000	0.000	0.000
	C. vicina	**0.050**	0.000	0.000	0.000	0.000	0.000
	L. coeruleiviridis	**0.050**	0.000	0.000	0.000	0.000	0.000
	L. sericata	**0.030**	0.000	0.000	0.000	0.000	0.000
	P. regina	**0.030**	0.000	0.000	0.000	0.000	0.000
	Phoridae	**0.010**	0.000	0.000	0.000	0.000	0.000
Mixtures of 2		**0.050**	**0.000**	**0.000**	**0.034**	**0.000**	**0.000**
	C. rufifacies and *C. vicina*	**0.110**	0.000	0.030	0.000	0.000	0.000
	C. rufifacies and *L. coeruleiviridis*	**0.080**	0.000	0.000	0.000	0.018	0.000
	C. rufifacies and *L. sericata*	**0.050**	0.250	0.080	0.000	0.000	0.250
	C. rufifacies and *P. regina*	**0.040**	0.000	0.000	0.000	0.000	0.000
	C. rufifacies and *Phoridae*	**0.080**	0.000	0.070	0.000	0.000	0.000
	C. vicina and *L. coeruleiviridis*	**0.130**	0.000	0.000	0.000	0.000	0.000
	C. vicina and *L. sericata*	**0.010**	0.000	0.000	0.000	0.000	0.000
	C. vicina and *P. regina*	**0.080**	0.000	0.000	0.000	0.000	0.000
	C. vicina and *Phoridae*	**0.060**	0.000	0.030	0.000	0.000	0.000
	L. coeruleiviridis and *L. sericata*	**0.060**	0.000	0.030	0.000	0.000	0.000
	L. coeruleiviridis and *P. regina*	**0.020**	0.000	0.070	0.000	0.000	0.000
	L. coeruleiviridis and *Phoridae*	**0.140**	0.000	0.000	0.000	0.000	0.000
	L. sericata and *P. regina*	**0.050**	0.500	0.030	0.000	0.000	0.000
	L. sericata and *Phoridae*	**0.030**	0.000	0.030	0.000	0.000	0.000
	P. regina and *Phoridae*	**0.010**	0.000	0.030	0.000	0.000	0.000

*Refers to the false positive rate

**Refers to the false negative rate

Table S5 (continued).

Discrimination level		**Specification**					
First	**Second**	Significance level	Error rate	Multiple prediction	Not assigned	fp*	fn**
Mixtures of 3		**0.090**	**0.037**	**0.000**	**0.038**	**0.000**	**0.025**
	C. rufifacies, *L. sericata* and *Phoridae*	**0.050**	0.000	0.000	0.000	0.000	0.000
	C. rufifacies, *P. regina* and *Phoridae*	**0.040**	0.000	0.087	0.000	0.000	0.000
	C. vicina, *L. coeruleiviridis* and *P. regina*	**0.140**	0.000	0.025	0.000	0.000	0.000
	C. vicina, *L. coeruleiviridis* and *Phoridae*	**0.010**	0.000	0.025	0.000	0.000	0.000
	C. vicina, *L. coeruleiviridis* and *L. sericata*	**0.140**	0.000	0.000	0.000	0.000	0.000
	C. vicina, *L. sericata* and *P. regina*	**0.110**	0.000	0.025	0.000	0.013	0.000
	C. vicina, *L. sericata* and *Phoridae*	**0.150**	0.000	0.000	0.000	0.000	0.000
	C. vicina, *P. regina* and *Phoridae*	**0.050**	0.200	0.075	0.000	0.000	0.200
	C. rufifacies, *C. vicina* and *L. coeruleiviridis*	**0.140**	0.000	0.025	0.000	0.000	0.000
	C. rufifacies, *C. vicina* and *L. sericata*	**0.020**	0.000	0.162	0.000	0.000	0.000
	C. rufifacies, *C. vicina* and *P. regina*	**0.070**	0.000	0.050	0.000	0.000	0.000
	C. rufifacies, *C. vicina* and *Phoridae*	**0.090**	0.000	0.000	0.000	0.000	0.000
	C. rufifacies, *L. coeruleiviridis* and *L. sericata*	**0.200**	0.000	0.000	0.000	0.000	0.000
	C. rufifacies, *L. coeruleiviridis* and *P. regina*	**0.120**	0.000	0.000	0.000	0.000	0.000
	C. rufifacies, *L. coeruleiviridis* and *Phoridae*	**0.070**	0.000	0.000	0.000	0.000	0.000
	C. rufifacies, *L. sericata* and *P. regina*	**0.050**	0.000	0.000	0.000	0.000	0.000

*Refers to the false positive rate

**Refers to the false negative rate

Table S5 (continued).

Discrimination level		**Specification**					
First	**Second**	Significance level	Error rate	Multiple prediction	Not assigned	fp*	fn**
Mixtures of 4		**0.100**	**0.042**	**0.014**	**0.042**	**0.005**	**0.042**
	C. rufifacies, *C. vicina*, *L. coeruleiviridis* and *L. sericata*	**0.180**	0.000	0.170	0.000	0.188	0.000
	C. rufifacies, *C. vicina*, *L. coeruleiviridis* and *P. regina*	**0.120**	0.000	0.080	0.000	0.000	0.000
	C. rufifacies, *C. vicina*, *L. coeruleiviridis* and *Phoridae*	**0.110**	0.125	0.170	0.000	0.188	0.125
Mixtures of 5		**0.120**	**0.000**	**0.069**	**0.000**	**0.000**	**0.000**
	C. rufifacies, *C. vicina*, *L. coeruleiviridis*, *L. sericata* and *P. regina*	**0.125**	0.125	0.000	0.000	0.000	0.000
	C. rufifacies, *C. vicina*, *L. coeruleiviridis*, *L. sericata* and *Phoridae*	**0.110**	0.125	0.000	0.000	0.000	0.125
	C. rufifacies, *C. vicina*, *L. coeruleiviridis*, *P. regina* and *Phoridae*	**0.020**	0.000	0.000	0.000	0.062	0.000
Mixtures of 6		**0.110**	**0.125**	**0.210**	**0.000**	**0.010**	**0.000**

*Refers to the false positive rate

**Refers to the false negative rate