Environmental Science & Technology

pubs.acs.org/est

Article

Building Interpretable Machine Learning Models to Identify Chemometric Trends in Seabirds of the North Pacific Ocean

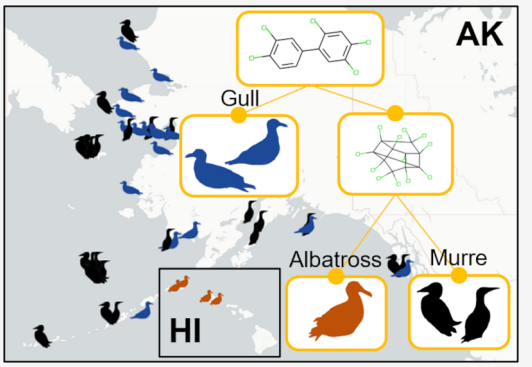
Nathan A. Mahynski,* Jared M. Ragland, Stacy S. Schuur, and Vincent K. Shen

Cite This: *Environ. Sci. Technol.* 2022, 56, 14361−14374

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Marine environmental monitoring efforts often rely on the bioaccumulation of persistent anthropogenic contaminants in organisms to create a spatiotemporal record of the ecosystem. Intercorrelation results from the origin, uptake, and transport of these contaminants throughout the ecosystem and may be affected by organism-specific processes such as biotransformation. Here, we explore trends that machine learning tools reveal about a large, recently released environmental chemistry data set of common anthropogenic pollutants measured in the eggs of five seabird species from the North Pacific Ocean. We modeled these data with a variety of machine learning approaches and found models that could accurately determine a range of taxonomic and spatiotemporal trends. We illustrate a general workflow and set of analysis tools that can be used to identify interpretable models which perform nearly as well as state-of-the-art "black boxes." For example, we found shallow decision trees that could resolve genus with greater than 96% accuracy using as few as two analytes and a *k*-nearest neighbor classifier that could resolve species differences with more than 94% accuracy using only five analytes. The benefits of interpretability outweighed the marginally improved accuracy of more complex models. This demonstrates how machine learning may be used to discover rational, quantitative trends in these systems.



KEYWORDS: *environmental monitoring, reproducibility, exposure, biorepository, STAMP*

INTRODUCTION

Ubiquitous contamination by anthropogenic pollutants in the North Pacific Ocean has been widely acknowledged as an issue of concern. Finding proper indicators of global trends is complicated owing to the complex nature and size of the ecosystem.[1] This is of particular importance in geographic areas (e.g., Alaska in this work) where subsistence hunting provides a significant nutrition source to indigenous peoples, exposing human populations to environmental pollutants.[2] Seabirds are an important group of upper trophic-level marine organisms in which lipophilic contaminants accumulate due to biomagnification. Analyses of seabird tissues, particularly eggs, have historically played important roles in temporal and spatial environmental monitoring of persistent organic pollutants and mercury in North America and Europe.[3−7] Egg contents serve as repositories of chemical information reflecting the environment in which these seabirds lived at the time eggs were laid.[8] Seabirds are widely dispersed over the North Pacific Ocean, so they can act as biomonitors of this body of water.

In 1999, the U.S. Fish and Wildlife Service Alaska Maritime National Wildlife Refuge, the U.S. Geological Survey Biological Resources Division, and the U.S. National Institute of Standards and Technology (NIST) implemented the Seabird Tissue Archival and Monitoring Project (STAMP) to monitor contaminants in Alaska's marine environments, which was expanded to include the U.S. Pacific Islands in 2010. This multiagency project aimed to collect, process, and bank seabird eggs under standardized protocols to avoid external contamination[9−11] with the ultimate goal of using these seabird eggs as biomonitors for environmental contaminants. To date, over 2588 seabird egg clutches from 11 species and 55 different colonies throughout the North Pacific Ocean[12] have been processed, and the homogenized egg contents archived in liquid nitrogen vapor freezers at the NIST Biorepository.[13] Contaminant information at this scale proves invaluable to wildlife managers and citizen stakeholders alike in understanding potential exposure-driven environmental and human health risk factors.[14] Understanding chemometric patterns in a given region enhance the understanding of contaminants in that ecosystem, which may reflect large-scale oceanographic changes.[15]

Machine learning has recently revolutionized modeling capabilities in many scientific disciplines, but state-of-the-art

Received: March 21, 2022
Revised: September 14, 2022
Accepted: September 15, 2022
Published: October 5, 2022

https://doi.org/10.1021/acs.est.2c01894
Environ. Sci. Technol. 2022, 56, 14361−14374

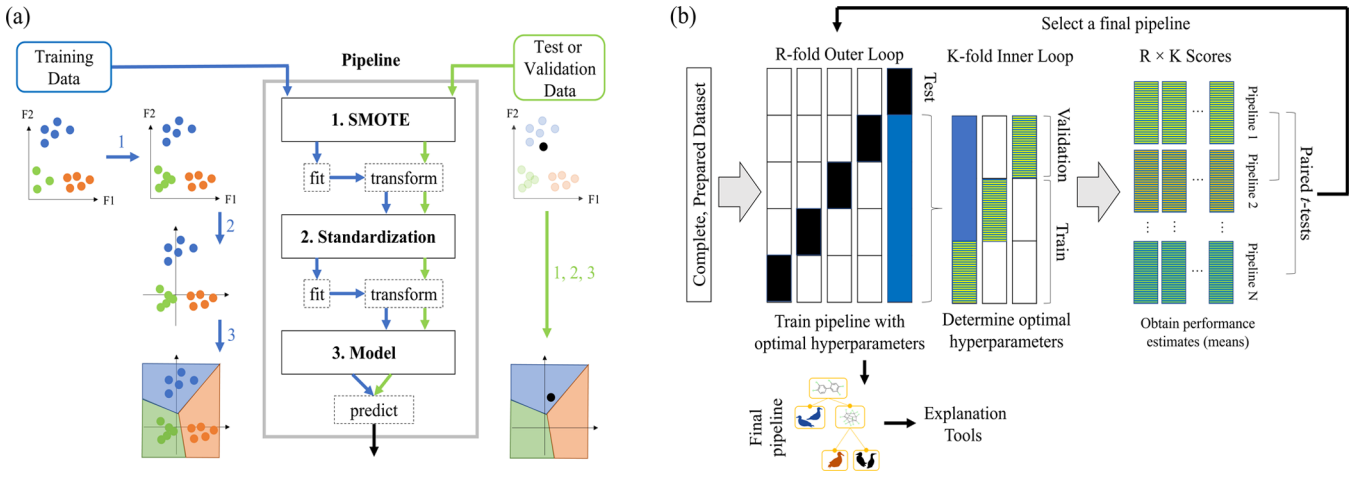


Figure 1. Machine learning methodology used in this study. (a) Example pipeline composed of 3 steps is shown. Training follows the blue arrows, while testing on new data follows the green arrows. (b) Nested strategy employed to estimate the performance of different pipelines and their uncertainty. Paired *t*-tests are performed on the $R \times K$ total validation values to assess the statistical significance of any differences. A pipeline is chosen, usually as a balance between performance and simplicity, and retrained using a single CV loop before being sent for further analysis by explanation tools, if the pipeline's model is not naturally interpretable already.

models for dense, tabular data, such as random forests[16] and gradient-boosted trees,[17,18] tend to be "black boxes".[19−21] While they may perform well, the reasons behind their performance are generally opaque. A variety of model-agnostic explanation methods such as permutation feature importances,[16,22] local interpretable model-agnostic explanations (LIME),[23] and Shapley additive explanations (SHAP)[24,25] have been developed to explain the inner workings of such models. Other "glass box" models, such as (shallow) decision trees or explainable boosting machines,[26] do not require explanation because their internal structure is naturally comprehensible. Explanations or interpretations of models are generally required to engender trust between the user and the model and are essential to understanding the scientific reasoning behind why a model behaves the way it does. In high-risk situations, such as when health factors are involved, there is a strong preference for simple, easily understandable models, even if they are not as powerful as the best black-box models. Model selection is further complicated by the so-called "Rashomon Effect"[21] which occurs when multiple, high-performing models make similar predictions but for different reasons. This occurs when different models identify different trends in data, which often results when features (in this work, individual contaminant levels) are highly correlated or there are more features than necessary to perform a prediction task.[21,27] Models of highly-monitored complex chemical systems, such as the environment, are often susceptible to this.

To this end, in order to explore the utility of machine learning methods on data commonly generated as a part of environmental monitoring efforts, we develop interpretable models to elucidate chemometric trends in oceanographic biomonitors relevant to wildlife conservation and management efforts. Here, we illustrate a standard, systematic workflow that can be used to train, compare, select, and explain different approaches enabling the development of rational models, which are simple to deploy and scientifically defensible. This is the first quantitative modeling effort to be performed on the STAMP database, and demonstrates the utility of machine learning on it; however, this work is not intended to be an exhaustive analysis. Typically, publicly available data are focused on specific project needs and are too heterogeneous to provide the scale needed for this study. We leveraged the combination of sample metadata housed by the NIST Biorepository and analytical chemistry data produced by NIST and its partners to classify seabird eggs based on their chemometric signatures. Our results suggest that chemometric data used for the identification and classification of tissue samples may be captured with simultaneously high-performing and interpretable machine learning models; these models provide scientific insight into the observed trends, and this workflow may serve as an invaluable tool in future analyses of such data.

MATERIALS AND METHODS

Data. All data preprocessing, analysis, and visualizations described here were performed in Python 3.7. Results were plotted with seaborn v0.11.1,[28] matplotlib v3.3.4,[29] dtreeviz v1.3.7 (https://github.com/parrt/dtreeviz), and cartopy v0.18.0[30] packages. We analyzed the STAMP data set consisting of persistent bioaccumulative contaminants found in the eggs of seabirds from the North Pacific Ocean collected between 1999 and 2010. The instrumentation and data set preparation are described in detail in ref 31, so we will only summarize the key points here. The original data set contained a total of 487 samples which were collected and heterogeneously analyzed for 174 ubiquitous contaminants including brominated diphenyl ethers, polychlorinated biphenyls (PCBs), per- and polyfluorinated alkyl substances, organochlorine pesticides (OCPs), mercury, butyltins, and stable isotope ratios.

Measurements in this data set are a result of various different historical investigations and thus are not homogeneous across all samples. Consequently, a subset of this data was selected that contains no missing values, and thus represents a uniform set of information. This subset contained only OCPs, PCBs, and mercury measurements. The species available in this subset currently include Common murre (*Uria aalge*, Pontoppidan 1763), Thick-billed murre (*Uria lomvia*, Linnaeus 1758), Glaucous gull (*Larus hyperboreus*, Gunnerus 1767), Glaucous-winged gull (*Larus glaucescens*, Naumann 1840), and

https://doi.org/10.1021/acs.est.2c01894
Environ. Sci. Technol. 2022, 56, 14361−14374

Laysan albatross (*Phoebastria immutabilis*, Rothschild 1893). The number of species present, and analytes tested for, may be expanded by additional analysis in the future but is beyond the scope of the current work which aims to interrogate existing data from this recently-released data set.

The subset of 5 seabird species represents 3 different genera nesting at numerous colonies and contains a total of 461 samples each tested for 49 different analytes or sets of analytes after the preprocessing described in ref 32. A complete list of these analytes, which are sometimes combined into a single feature due to, for example, coelution, is available in the Supporting Information. While there are instances where analytes were not detected, in this subset a test was always performed. When a nondetect occurred, the value was randomly imputed to a value below the limit of detection. Imputing to fixed values, such as zero, can artificially bias classification and regression models. Random values tend to destroy artificial statistical correlation with target variables, such as the origin of a sample, which may lead to more conservative models.

Machine Learning Pipelines. In this work, we focus on training various machine learning models for classification tasks. Henceforth, a "class" refers to a category and should not be confused with the same term used to describe the formal biological taxonomy of seabirds. To systematically investigate different machine learning algorithms and preprocessing steps, we employed a construct known as a "pipeline" in the imbalanced-learn v0.7.0 package.[33] A pipeline is essentially a series of steps that may be chained together in a sequence to produce a composite model. Figure 1a illustrates an example pipeline composed of three steps: class-balancing (SMOTE), standardization, and a final predictive model. When training data is provided to the pipeline the initial steps both fit and transform the data. Fitting stores internal parameters that enable the transformation; for example, standardization measures the mean, $\overline{x}$ and standard deviation, s, of the sampled data so that a feature may be autoscaled, $x_{\text{new}} = \frac{x - \overline{x}}{s}$. When testing or validation data is provided, no fitting is performed; only the transformation is performed to pass data from one step to the next. This final step is always the predictive model which is trained on the data that was transformed by the preceding steps. All models used in this work are available in the scikit-learn v0.22.2[34] and XGBoost packages.[18]

The pipeline construct also allows preprocessing steps to be added or removed. One common step is to balance the classes when the number of samples from each category is very different. The synthetic minority oversampling technique (SMOTE)[35] was sometimes used to balance the classes by interpolating between pairs of observations to create synthetic data; we combined it with the edited nearest neighbors approach (SMOTEENN)[36−38] to reduce noisy synthetic samples that result from interpolation toward extreme points, which could be outliers. However, in tree-based models bias from imbalanced classes can also generally be prevented by weighting the data inversely proportional to the observed frequency of a given class. The use or disuse of this is an example of a hyperparameter that is tuned during cross-validation (CV). All hyperparameters associated with each pipeline were optimized by using a grid search with stratified CV, where the relative proportion of classes is kept constant across all folds.

Training and Evaluating Pipelines. We considered 9 different models in pipelines: a simple decision tree,[39] random forest,[16] a gradient-boosted tree as implemented in XGBoost,[18,40] linear discriminant analysis (LDA) and quadratic discriminant analysis,[41] LDA for dimensionality reduction followed by logistic regression,[41,42] support vector classifiers with linear and radial basis function kernels,[41] naïve Bayes,[43,44] and k—nearest neighbors (KNN).[41] Hyperparameters for, and descriptions of, these models can be found in the scikit-learn[34] and XGBoost documentation.[18] Autoscaling was included in all pipelines.

Pipelines were trained and evaluated using a nested approach graphically illustrated in Figure 1b. An initial outer loop splits the data into R repeats with disjoint held-out test sets; the remaining data is then split into K-folds (inner loop) to perform CV. The inner loop uses $K - 1$ folds of data (Train) to fit the pipeline with one set of hyperparameters and then tests the performance on the remaining fold (Validation); this is repeated K times to determine an average. The K scores from the model with the best average score are retained. This entire process is repeated on each of the R different data splits to obtain a total of $R \times K$ validation set performances, which reflect how well a pipeline and its hyperparameters can be optimized given different possible training sets. Each outer loop effectively ignores the held-out $1/R$ fraction of the data (Test) which serves only to help decorrelate the different repeats by ensuring they do not use the exact same set of data. Conventional nested CV[45] uses this test data to obtain an unbiased estimate of the generalization error, whereas this scheme is essentially repeated "flat CV" whose K performances may be biased; however, extensive studies on real-world data sets using the models we employ in this work suggest this bias does not significantly affect the ranking of these classifiers.[46]

Each of the R pipelines may have slightly different hyperparameters, so the final production model is obtained by performing a single CV (outer) loop on the entire data set at the end; with no held-out data to compare against, its accuracy, a, and uncertainty are reported as those obtained for that pipeline during the nested loop process. The data shuffling and splitting are the same for all pipelines; thus, each pipeline trains on the same $R \times K$ training sets. This allows us to perform a paired t-test to ascertain if one pipeline outperforms another in a statistically significant manner. For each task we attempt such as training models to predict species, we performed a 1-sided test (H_0: $\overline{d} \leq 0$, H_1: $\overline{d} > 0$) for each pipeline against the one with the best mean performance. Here, $\overline{d}$, is the mean difference in accuracy, $d_{i,j} = a_{i,j}^{\text{best}} - a_{i,j}^{\text{other}}$, between the two pipelines

$$\overline{d} = \frac{1}{R \times K} \sum_{i=1}^{R} \sum_{j=1}^{K} d_{i,j} \tag{1}$$

$$s^2 = \frac{1}{R \times K - 1} \sum_{i=0}^{R} \sum_{j=0}^{K} (d_{i,j} - \overline{d})^2 \tag{2}$$

Since data folds overlap between the runs, the $R \times K$ validation scores are not completely independent, which leads to high Type I error in a standard t-test.[47] To account for this, we used Bouckaert and Frank's corrected repeated K-fold CV test,[47,48] which employs Nadeau and Bengio's correction[49] to combat this effect. In the original formulation, the K-fold CV is assumed to be repeated on the same (entire) data set R times;

https://doi.org/10.1021/acs.est.2c01894
Environ. Sci. Technol. 2022, 56, 14361−14374

however, the nested formulation helps further decorrelate the R different repeats which also helps counteract this effect. The final t-statistic is

$$t = \frac{\bar{d}}{\sqrt{\left(\frac{1}{R\times K} + \frac{\rho}{1-\rho}\right)s^2}} \tag{3}$$

where $\rho = n_{\text{test}}/(n_{\text{test}} + n_{\text{train}}) = 1/K$.

We make the practical conclusion that any pipeline for which we fail to reject H_0 is essentially as good as the best performer, whereas the rejection of H_0 occurs when there is statistical evidence that the first pipeline is better. We generally search for the "simplest" model that does not statistically underperform the best model or is at least very close. Simplicity is subjective and is left to the user to decide the relative merits of simplicity versus accuracy. Unless otherwise stated we used $R = 5$ and $K = 2$ with a significance level of $\alpha = 0.05$.

Other Preprocessing. All preprocessing steps which require information about the target variable, such as the species when performing SMOTE, are considered supervised since they incorporate information about the final goal of the model. All supervised methods are included in the pipelines so that any bias from the observed training data is accounted for in the CV. We performed 2 unsupervised preprocessing steps before beginning the analysis. Imputation of nondetects to random values below their limit of detection is the first example of this. The second is to remove correlated analytes. This is described in detail in ref 31 and in the Supporting Information, but it essentially involves performing agglomerative hierarchical clustering with Ward's linkage[50] on the Spearman rank-order correlation, r_S, between analytes. This results in an unsupervised clustering of the analytes; the number of clusters changes based on the threshold distance, t, allowed between different clusters. When $t = 0$, all analytes are unique and are alone in their own "cluster;" as t increases, clusters of multiple analytes form in which members are considered to contain similar information. As a result, we select a single analyte from each cluster.

This strategy enables one to choose, for example, a set of analytes that fall into categories of similar types (e.g., all PCBs), or those that can all be measured on the same instrument. It can also be used to prioritize analyte selection when some are missing from parts of agglomerative data sets (e.g., from different analytical methods), or prioritize selection by *ad-hoc* choices (e.g., preferentially selecting analytes with known health impacts) backed by pattern similarity. This enables models to be built using a set of variables that are practically simpler and potentially more cost-effective to obtain. To perform selection, analytes are categorized and the entropy of a selection is

$$S = -\sum_{i}^{N_{\text{cat}}} f_i \ln(f_i) \tag{4}$$

where N_{cat} is the total number of different categories, and f_i is the fraction of selected analytes that belong to category i.

Given a set of clusters and categorized analytes, we searched for an optimal selection by minimizing S with a Monte Carlo method. We initially selected a random set of analytes; subsequently, perturbations were proposed by selecting a random cluster and a new random analyte within it. This new set of choices was accepted with a probability given by a Metropolis criteria[51]

$$p = \min[1,\ \exp(-(S_{\text{new}} - S_{\text{current}})/T)] \tag{5}$$

where T is a parameter that controls how often a more disordered set of analytes is chosen; for a more ordered set $S_{\text{new}} < S_{\text{current}}$ so that $p = 1$ which is always accepted. After 1000 rounds with $T = 0.1$ a final selection was made. This process was repeated 5 times to ensure the selection with the lowest S_{cat} was found.

In this work, we labeled analytes by the categories described in the Data section, such as OCP, PCB, mercury, stable isotope ratio, and so forth. The 49 analytes used in this work fall exclusively into the first 3 categories with PCBs being the most numerous; consequently, as the number of clusters decreased (increasing t), PCBs tend to dominate the features chosen. The SI contains a complete list of analytes chosen as the number of clusters was systematically decreased from 49 to 5.

EXPLANATION TOOLS

SHAP. After a pipeline has been developed and decided upon, it should be explained. SHAP[24,25] is a framework, based on cooperative game theory, to explain the output of any model by fairly allocating credit to each of the features in the model. A SHAP value for a feature (e.g., level of PCB 110) represents the average change in the model's output (e.g., probability of class membership) due to the value of that feature. For example, a random forest can be used to predict the probability, $p(\text{C}|\vec{X})$, that a sample originates at a given colony, C. A SHAP value can be computed for each analyte in the observation, $x \in \vec{X}$, which indicates how much of the probability is due to x. The average probability of the class plus the sum of the SHAP values for an instance yields the model's output, $p(\text{C}|\vec{X})$. The mean absolute SHAP value across all observations in a set serves as an indicator of which features play a predominant role in determining the likelihood a sample originates from a colony. While SHAP values do not necessarily indicate causality, they are a powerful tool that reveals how the model is using its features (in this case, analytes) to make its predictions. Importantly, it is model-agnostic, so it may be used as a unified approach to explain all models, and SHAP itself has been shown to be a generalization of a number of other feature-based explanation techniques, such as LIME[23] and quantitative input influence.[52]

Jensen−Shannon Divergence. When performing classification, groups of individual analytes, which uniquely "fingerprint" a class are both interesting and useful in highly explainable models. A fingerprint emerges when the distributions of an analyte for two categories (such as different species) are disjoint, or nearly so. Decision trees naturally make use of such analytes to split nodes into different sets in decreasing impurity. Such a tree will try to choose the numerically optimal analyte to split a node; however, there may be other analytes that could yield splits that are nearly as good but are much more naturally understandable or preferable for another reason. We computed the Jensen−Shannon divergence (JSD)[53,54] to identify analytes that naturally distinguish a class from others to identify such alternatives. The JSD is a convenient metric for quantifying the difference between two distributions, P and Q; it is a symmetrized form of the Kullback−Leibler divergence (KLD) and is bounded between

https://doi.org/10.1021/acs.est.2c01894
Environ. Sci. Technol. 2022, 56, 14361−14374

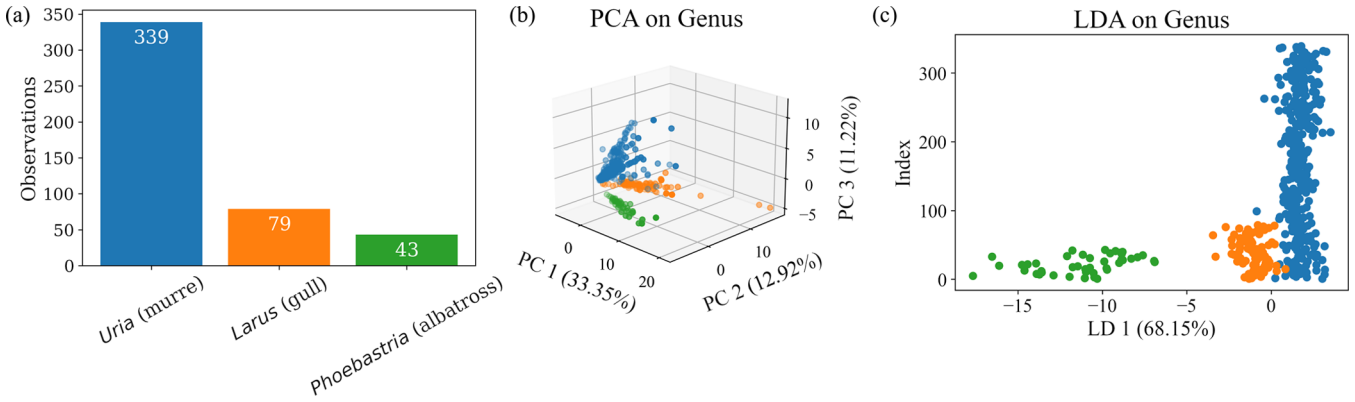


Figure 2. Summary of data set labeled by genus. (a) Class imbalance found in the data, with over 70% coming from murres. (b) PCA performed on the entire data set illustrates that a three-dimensional separation may be possible. (c) LDA on the entire data set illustrates that a single dimension, or discriminant, is sufficient to separate the genera.

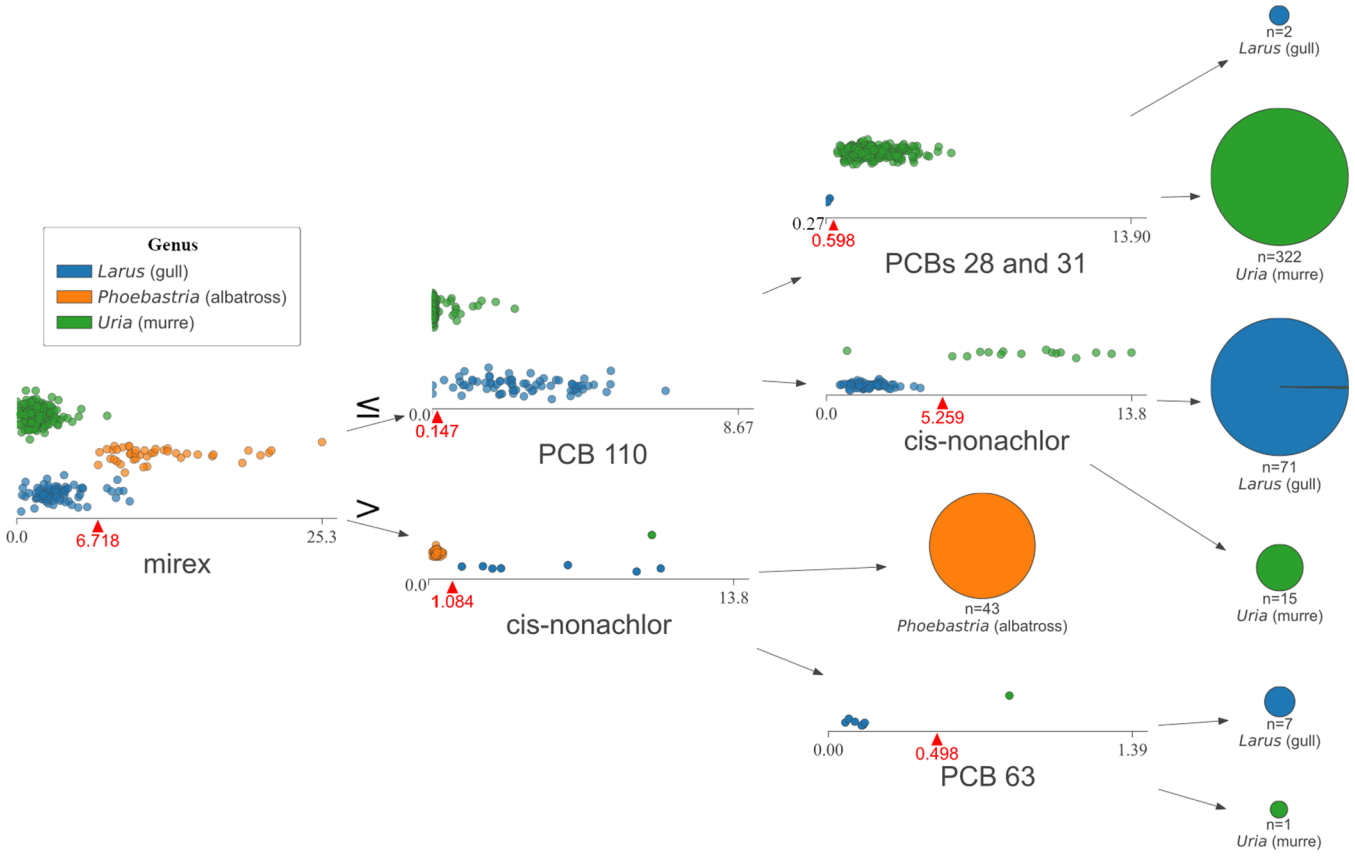


Figure 3. Decision tree model trained on a SMOTEENN-balanced data set to predict genus. The optimal tree depicted was determined by stratified, 10-fold CV and has an estimated accuracy of 97.94 ± 0.67%. Predictions for the entire (unbalanced) data set are illustrated, and a jittered strip plot of data at each node is shown; the value (ng/g wet mass) each node is split at is given in red.

0 (identical distributions) and 1 (no overlap) when a logarithm with base 2 is used.

$$\mathrm{JSD}(P\|Q) = \frac{\mathrm{KLD}_2(P\|M) + \mathrm{KLD}_2(Q\|M)}{2} \tag{6}$$

where $M = (P + Q)/2$ is the mean of the two distributions, and the KLD with base n is defined as

$$\mathrm{KLD}_n(P\|Q) = \sum_x P(x)\log_n\left(\frac{P(x)}{Q(x)}\right) \tag{7}$$

This is a useful tool for intuitively suggesting when an analyte used by a decision tree, or similar model, might be exchanged for a different analyte to construct a slightly different, more interpretable model with nearly identical performance.

When computing a one-vs-all (OvA) JSD for an analyte, Q was computed as a normalized histogram using data from all the classes except the one in question, for example, a combination of PCB 100 for murre and gulls if we are focusing on albatross; the distribution of that class, P, was also normalized, and the JSD between the two was computed. This

https://doi.org/10.1021/acs.est.2c01894
Environ. Sci. Technol. 2022, 56, 14361−14374

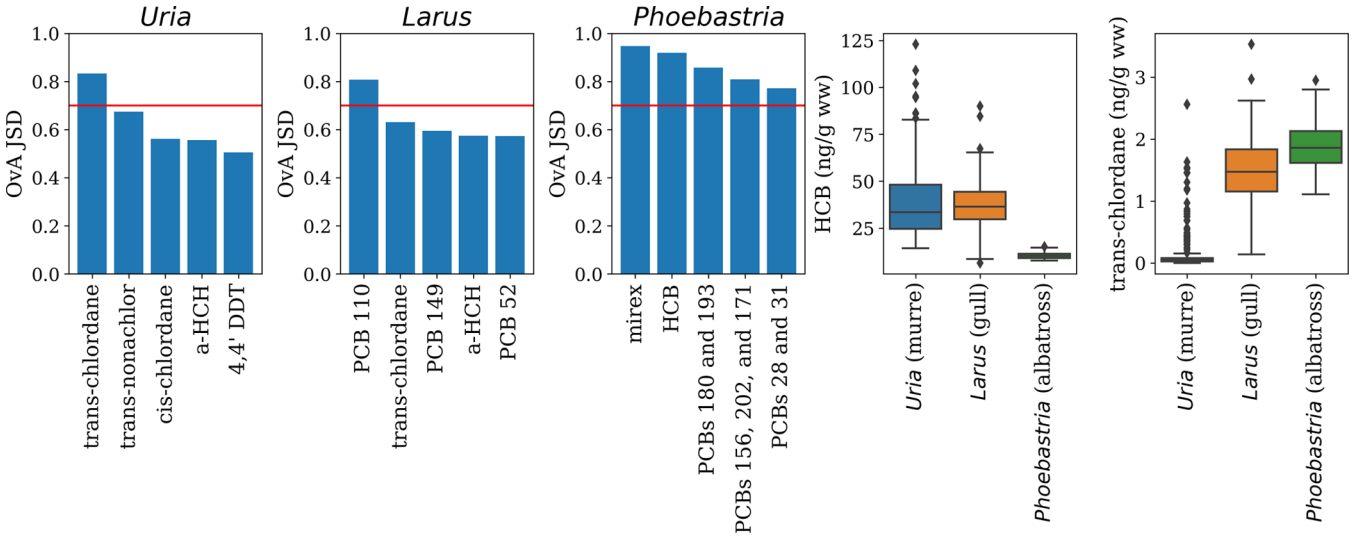


Figure 4. OvA JSD for analytes that best distinguish one genus from the others in the data set. An arbitrary threshold of 0.7 is shown in red as a guide to the eye. The leading analyte for *Phoebastria* is mirex, already shown in Figure 3; the second analyte, HCB is shown here for comparison (OvA JSD = 0.92). The best analyte for distinguishing *Uria* is *trans*-chlordane, with a level typically much lower than *Larus* and *Phoebastria* (OvA JSD = 0.83).

OvA JSD indicates how uniquely the analyte is distributed for the class in question relative to all others. We also computed a pairwise JSD between two classes (e.g., murre versus albatross); in this case, two normalized distributions are compared, and no combination is required.

RESULTS AND DISCUSSION

Classifying Genus. First, we examined what, if any, chemometric differences exist between seabirds based on their taxonomy. Three genera are present in the data: *Uria* (murre), *Larus* (gulls), and *Phoebastria* (albatross). There is a substantial amount of class imbalance, as illustrated in Figure 2a; *Uria* comprises 73.5% of the data, with 17.1% coming from *Larus* and the remainder from *Phoebastria*. Principal component analysis (PCA), after autoscaling the data, revealed that 3 dimensions were sufficient to naturally separate the genera in a visually intuitive fashion; however, this unsupervised approach proved to be less efficient than LDA, which was able to separate the data in only a single dimension (cf. Figure 2b).

We performed 10-fold, stratified CV with LDA and computed the leading coefficients, or scalings, on the analytes in the top linear discriminant, LD 1, to understand how substantially they affect the classification. The full results are shown in Supporting Information Figure S1, which illustrates that the levels of PCBs 153, 132, and 170 play the most important role in determining LD 1; however, given the proximity of *Larus* to *Uria* along this dimension, we cannot discount the combined importance of the other analytes in accurately determining a sample's position along this axis. While another classifier, such as logistic regression, may perform well when trained on the dimensionally compressed data by creating "vertical" partitions between the classes in Figure 2c, the LDA classifier itself, which does not rely on any such reduction, yielded a 98.70 ± 1.44% accuracy (one standard deviation) across the 10-fold tests.

To classify new data, however, would still require the measurement of 49 different analytes. Decision trees are a nonlinear alternative, which we found required far less information to achieve comparable performance. Moreover, the significant class imbalance can be addressed in situ in tree-based models by adjusting class weights to be inversely proportional to their observation frequency. We considered simple tree classifiers in 3 cases: (1) when no class balancing was performed, (2) when balanced by weighting points inversely proportional to their class frequency, and (3) when SMOTEENN was used to balance the data by creating additional synthetic observations. 5 × 2 nested, stratified CV was used to optimize these pipelines' hyperparameters and estimate the generalization error. All three pipelines performed similarly, with validation set accuracies of 96.69 ± 1.15%, 97.12 ± 1.36%, and 97.94 ± 0.67%.

Taking the last approach, we performed 10-fold, stratified CV on the data set to identify the optimal hyperparameters for such a tree; remarkably, the optimal depth was found to be only 3. This tree is depicted in Figure 3. The imbalanced tree from case (1) was found to be nearly identical, differing primarily at the stump, where it used *trans*-chlordane instead of mirex (cf. Supporting Information Figure S2). This is quite rational as mirex distinguishes the minority *Phoebastria* class very well from the other two as shown in Figure 3, whereas *trans*-chlordane is effective at the same task for *Uria*. Chlordane components have been previously found to correlate with arctic seabird phylogeny, distinguishing thick-billed murre from other nonmurre species.[55]

These simple decision trees tend to result when there are analytes that uniquely "fingerprint" a single class of interest in an OvA fashion. For genus classification, mirex and PCB 110 serve as excellent fingerprints within this specific set of species; however, deeper trees can make use of multiple serial comparisons and could be designed around analytes that show only good pairwise rather than OvA separability. We used the JSD to compare the distribution of each analyte for a given genus to the distribution of the other genera; the results are shown in Figure 4. Good fingerprints have an OvA JSD close to 1. Clearly, the distribution of PCB 110 for *Larus* deviates strongly from the other genera indicating that it is a

https://doi.org/10.1021/acs.est.2c01894
Environ. Sci. Technol. 2022, 56, 14361−14374

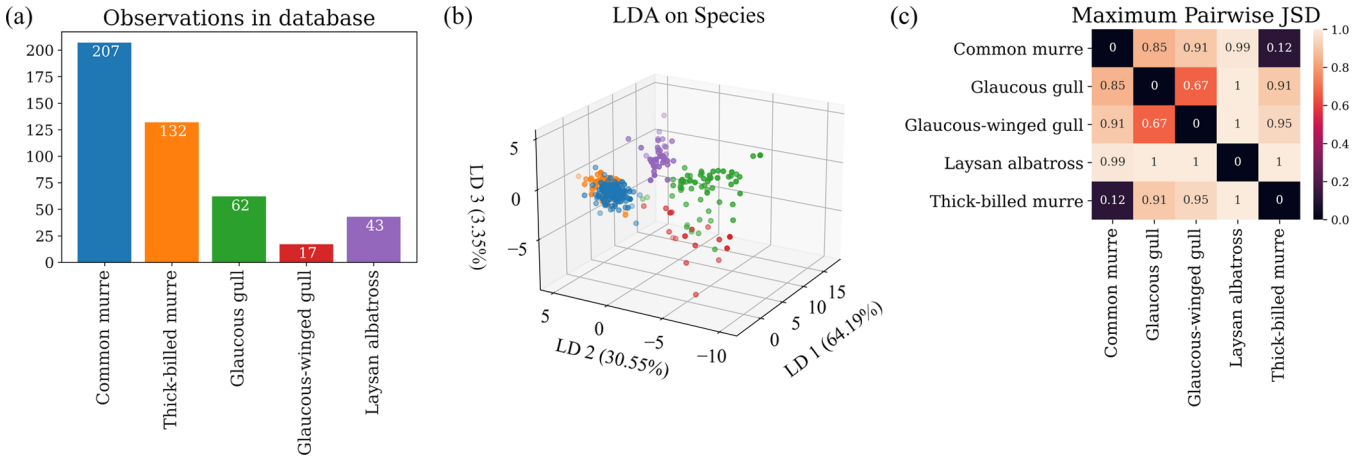


Figure 5. Predicting a sample's species. (a) Class imbalance observed for the 5 species present. (b) LDA on the standardized dataset; colors correspond to (a). (c) Maximum pairwise JSD between pairs of species. The analyte each JSD corresponds to is given in Supporting Information Table S1.

good candidate. Similarly, the *Uria* are well characterized by their *trans*-chlordane measurements (cf. Figure 4).

For the *Phoebastria*, however, there are many analytes that can be used to distinguish them from the other genera; this is unsurprising given that these samples were collected from 4 colonies in Hawaii (Barking Sands PMRF and North Shore, Kauai, and Kaena Point and Kaneohe MCBH, Oahu) whereas *Uria* and *Larus* samples originate from Alaska. While these albatross feed in Alaskan waters, they breed in Hawaii. The leading fingerprint for *Phoebastria* is an elevated level of mirex, which was historically used as an insecticide on pineapple plantations in Hawaii. Although registration was canceled in 1977, stockpiles were still found as late as 2000.[56,57]

Still, there are numerous other analytes that could serve as fingerprints for *Phoebastria*. For example, 5 × 2 nested, stratified CV was used to train a pipeline using SMOTEENN with a decision tree model using only mirex, PCB 110, and *trans*-chlordane, which resulted in a performance of 96.75 ± 0.90%. Comparatively, another tree trained similarly, but with HCB instead of mirex, yielded 96.75 ± 0.64%. Both were limited to a depth of 4, but are nearly identical to each other in terms of performance and to the tree given in Figure 3. HCB is characteristically lower in *Phoebastria* than *Uria* or *Larus*; HCB has high mobility and is expected to migrate from the area of release toward the poles based on global fractionation theory.[58] The *Uria* and *Larus* species are distributed more closely to the poles than the *Phoebastria*, in line with this theory's prediction.

In principle, only $N - 1$ analytes that are excellent at OvA comparisons are required to separate N classes; for example, if we start with a group of 3 classes (A, B, and C) and the first comparison splits away class A, we need only one more comparison to split B and C apart. Indeed, trees trained to use only 2 of these analytes (such as PCB 110 and *trans*-chlordane or mirex and PCB 110) yielded nearly identical performances. The latter resulted in 96.64 ± 0.83% when constrained to have only 3 leaf nodes (one for each class, cf. Supporting Information Figure S3); note that this tree is very similar to the first 2 levels of the tree in Figure 3. We conclude that this tree is not unique in its simplicity and performance; further, the JSD provides clear intuitive guidance as to which analytes could be used in simple models as chemometric fingerprints for these genera. While PCB 110 and mirex may be analyzed in a single run using electron impact gas chromatography−mass spectrometry (GC−MS), *trans*-chlordane is more accurately measured using negative chemical ionization GC−MS. Thus, machine learning may be capable of providing insights that reduce analytical efforts and costs; for example, if the goal is only to separate genera or investigate general contaminant trends rather than monitor for specific analytes of interest, such approaches may provide an adequate understanding with less data collection.

Classifying Species. In this data set, we have at most 2 different species present from the same genus (cf. Figure 5a). Therefore, the leaves of a very good genus classification tree are expected to have at most 2 species present, which suggests that as little as one additional comparison would be required to split them apart and create a species classifier. However, Figure 5b shows the result of LDA when compressing the data into three dimensions, which illustrates a pathological problem regardless of dimensionality or the complexity of any model we used. While the Laysan albatross separates well from the other species, and the two gull species show signs of separating from each other along the LD 3 axis, the two murre species are largely indistinguishable.

While an OvA JSD can identify analytes that uniquely differentiate a species from all others, a pairwise comparison illuminates the maximum degree to which we might expect a single pair of classes to be differentiable. Figure 5c illustrates this; most species are highly distinguishable from each other, with the exception of the two murres. In the case of murres, the most discriminating analyte was found to be PCB 206, which is quite poor. Supporting Information Table S1 contains more details; the analytes generally align with the genus-level classification results. Polynomial feature engineering was not found to improve the situation.

Of course, more advanced models may be able to leverage patterns that emerge from complex combinations of many analytes. To this end, we considered a range of models, summarized in Figure 6. Tracks in Figure 6 which are grayed out indicate where the one-sided null hypothesis that the model's mean performance is indistinguishable from the best performing one was rejected. For the models considered, ensemble methods (gradient-boosted trees and random forests) performed the best; yet, these models are bounded above by an accuracy of roughly 80%. Using SMOTEENN to synthetically balance the data set during training did not

https://doi.org/10.1021/acs.est.2c01894
Environ. Sci. Technol. 2022, 56, 14361−14374

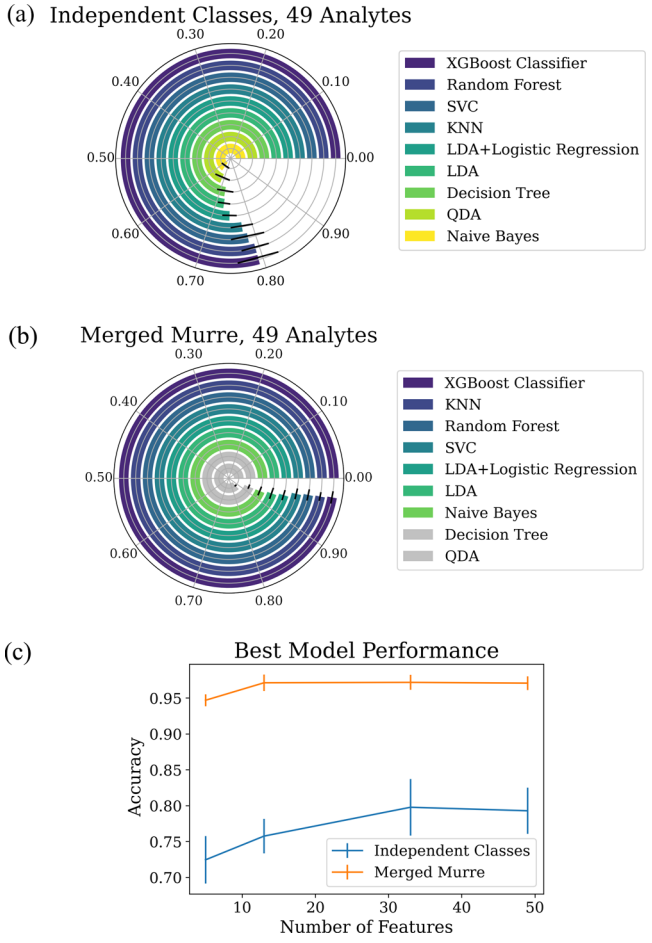


Figure 6. Performance summary of different machine learning models trained to classify species. Error bars correspond to one standard deviation from a 5 × 2 nested, stratified CV. (a) Results when all species are considered separate classes. (b) Results when the two murre species were merged into a single class. Gray bars indicate the rejection of the null hypothesis (α = 0.05) that a model's mean performance is no worse than the best. (c) Performance and standard deviation of the best model when trained on a different number of features.

improve these results. This upper bound is rational given that 3 of the 5 species are relatively easy to distinguish (gulls and albatross); assuming a model can identify those with nearly perfect accuracy, while the murre species are essentially guessed randomly with a 50% accuracy, a weighted average is roughly 80%. When the two murre species were combined into a single class, classification models achieved accuracies in excess of 97%, as expected. Once again, ensemble methods performed the best, though a KNN classifier performed nearly as well as the leading gradient-boosted tree.

Classes of contaminants tend to have similar physical and chemical properties, and thus, tend to be correlated when measured in the environment.[58,59] We performed hierarchical clustering based on Spearman rank-order correlations[60] and, from these clusters, selected features that maximized their categorical similarity. We performed the same tests for a variety of subsets of analytes, summarized in Figure 6c. When the murre species were kept separate, the mean accuracy generally decreased as features were removed. When they were combined, however, it made almost no difference. This is not surprising since the OvA JSD calculations suggest that decision trees can be very successful in separating $N = 4$ categories with as little as $N - 1 = 3$ features; indeed, tree-based models were found to be among the best models.

The best 4 models (gradient-boosted tree, random forest, support vector classifier, and KNN) tended to remain the best models regardless of the number of analytes used or whether or not the murre species were combined, though their relative rank fluctuated (cf. Supporting Information Figure S7); since we failed to reject the null hypothesis for these pipelines, we treat the performance of these as models as essentially equivalent. Of these, the KNN model is the most naturally interpretable. When 10-fold, stratified CV was performed on the data set to optimize the KNN hyperparameters, we found the optimum corresponds to $k = 1$ (first nearest neighbor) with a Manhattan distance, rather than a Euclidean one. The generalization performance estimated from the 5 × 2 nested, stratified CV in Figure 6b is 96.86 ± 0.76% when all 49 analytes are used. However, when only 5 analytes are used (PCBs 18, 105, 107, 110, and 201), the accuracy only drops to 94.69 ± 0.83%; in fact, the KNN model is the best performing one when only these 5 analytes are allowed (cf. Supporting Information Figure S7).

Furthermore, we note that when no more than 13 analytes were allowed, the decision tree model performed only slightly worse than the leading 4 models (95.28 ± 1.13%) and turned out to have a particularly simple, balanced structure. A class-balanced decision tree was retrained on 67% of the data set with 10-fold, stratified CV; it is shown in Figure 7, and had a performance of 95.42% on the remaining held-out data. At a depth of 2, the tree has already split the data into the 4 classes quite well (92.81% accuracy on the test set); one additional level improves the tree, but it is quite interpretable and uses many of the analytes suggested in Figure 4.

Overall, these results suggest that because we could not detect a significant chemometric difference between the two murre species, it is reasonable to consider these congeneric species as a single class for oceanographic and environmental monitoring; these species are known to exhibit differential foraging behavior,[61] though apparently, this had no discernible impact on chemometric signatures found in this work. This suggests local environmental factors at the collection site have a more pronounced effect than differential foraging on the egg contaminant levels of the 49 analytes focused on in this study (OCPs, PCBs, and mercury); other analytes, such as stable isotope ratios, are expected to reflect this information. We will examine this in more detail in the next section. The success of the KNN classifier further suggests that this data set represents a good reference set that may be of use to wildlife managers, as techniques such as those described here become more commonplace.

Classifying Colony. Next, we examined trends in the geographic origin of samples. There are 35 total colonies in the data set; however, many contain only a small number of observations, which make them difficult to characterize with any statistical certainty. We chose a threshold of 10 samples and considered only the data from colonies with at least this many entries, limiting the data from 461 samples to 350, distributed across the leading 13 colonies shown in Supporting Information Figure S8a. Furthermore, given the chemometric trends that enable the prediction of taxonomy, we wish to avoid drawing any conclusions about geographic trends that are simply due to sampling from species-specific colonies. For example, all albatross samples originate from Hawaii. As shown

https://doi.org/10.1021/acs.est.2c01894
Environ. Sci. Technol. 2022, 56, 14361−14374

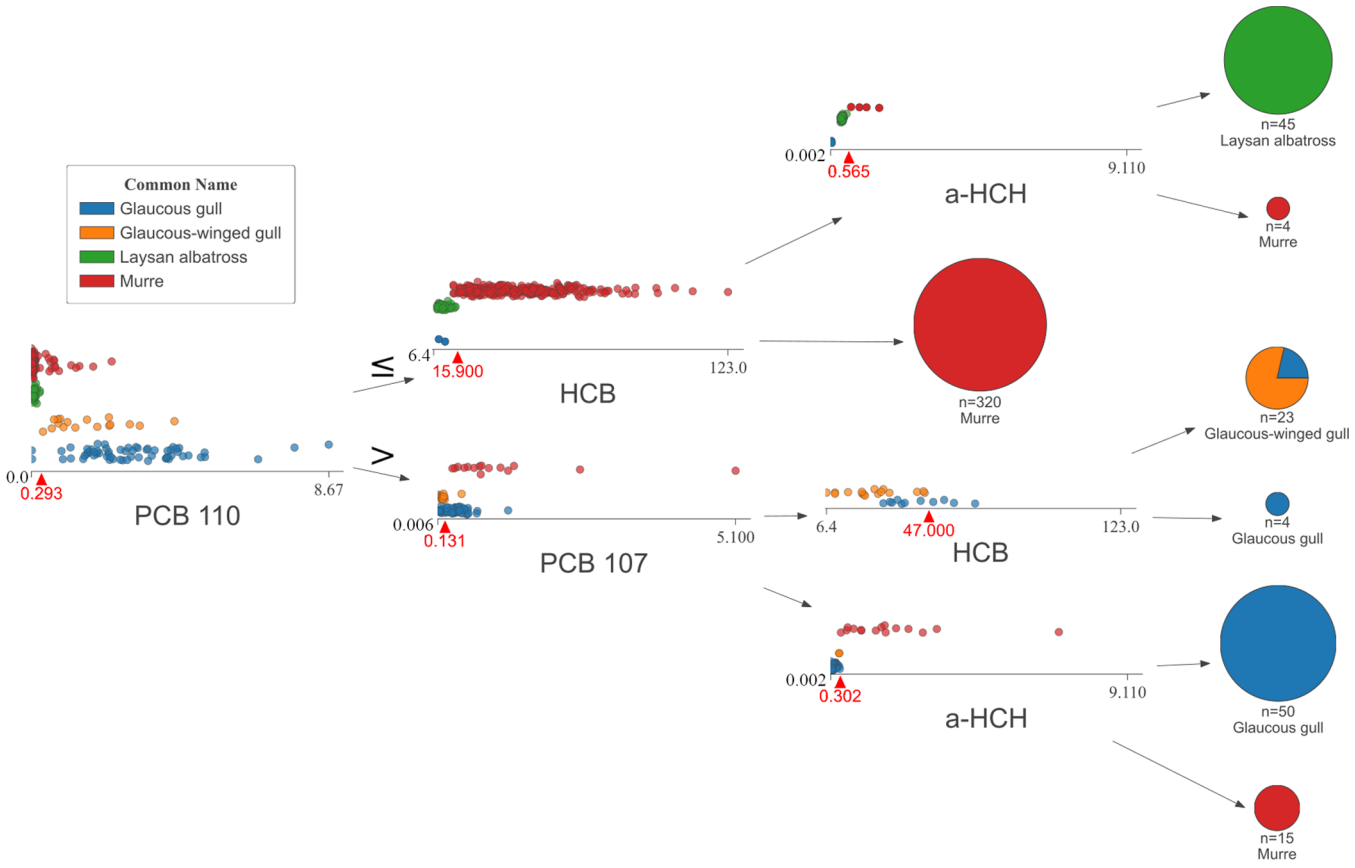


Figure 7. Decision tree trained to predict species if the murres are considered a single class. 10-fold, stratified CV was performed on a 67:33 train/test split of the data. Note that the 4 classes are already well separated after 2 levels. Predictions for the entire (unbalanced) data set are illustrated, and a jittered strip plot of data at each node is shown; the value (ng/g wet mass) each node is split at is given in red.

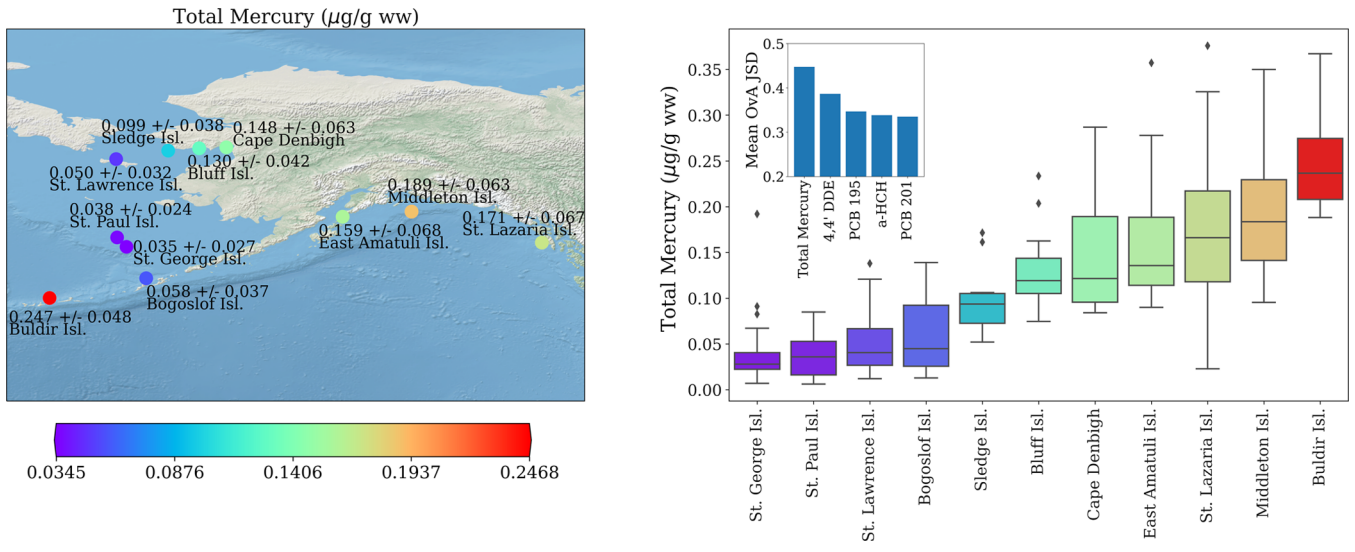


Figure 8. The geographic distribution (plate carrée projection using cartopy[30]) of mercury across murre colonies with at least 10 samples. The mean plus or minus one standard deviation is reported. A box-and-whisker plot shows the distribution at each location. Colonies are ordered by their mean, not their median values. The OvA JSD was computed (using 25 bins) for each analyte at each colony; the top 5 averaged across all colonies is reported in the inset.

in Supporting Information Figure S8b, gull and albatross samples in this reduced data set originate from a limited set of colonies. Murre, however, are both the most numerous and widely distributed, so we restricted our analysis to only the murre data (11 colonies). As shown previously, there is essentially no detectable chemometric difference in this data set between the congeneric murre species, so we may reasonably consider them as a single class.

We computed the OvA JSD for all analytes when datapoints were grouped by colony; total mercury, in particular, stood out

https://doi.org/10.1021/acs.est.2c01894
Environ. Sci. Technol. 2022, 56, 14361−14374

as having a relatively unique level at different locations. The inset in Figure 8 shows the largest OvA JSD values for analytes when averaged across all colonies (cf. Supporting Information Figure S9). Buldir Isl. is volcanically active, so its high mercury content is unsurprising. The next three colonies, Middleton Isl., St. Lazaria Isl., and East Amatuli Isl. are all along the Gulf of Alaska; the following three, Cape Denbigh, Bluff Isl., and Sledge Isl., are found in Norton Sound. The remaining colonies are located throughout the Bering Sea. Broadly speaking, with the exception of Buldir Isl., colonies located near or along the mainland of Alaska have elevated levels of mercury relative to those further offshore. In Alaska, this trend holds even if we examine all data, regardless of species or number of samples per colony (cf. Supporting Information Figure S10). Notably, samples taken from Nome have the second highest mean level, second only to Buldir Isl. It has been speculated that outflow from the Yukon River and associated cinnabar deposits, combined with elevated carbon output, may increase the methylation and therefore bioavailability of mercury in this area.[62]

The analyte with the second highest mean OvA JSD is 4,4′ DDE. In this case, St. Lazaria Isl. exhibits particularly high levels, but in general, colonies along the Gulf of Alaska tend to have similar values, which are elevated relative to those in the Bering Sea. If we consider results from all species and all colonies, those near Norton Sound tend to be more varied. 4,4′ DDE is the major metabolite of the pesticide DDT. While banned in the Stockholm Convention, exceptions have been granted for the control of malaria in high-risk areas. Certain physical and chemical properties of 4,4′ DDE, combined with enhanced wet and dry deposition in the eastern Gulf of Alaska, suggest elevated levels are expected at St. Lazaria. In fact, many analytes have elevated levels at this colony, consistent with this expectation.[63,64]

Stable isotope ratios of carbon ($\delta^{13}C$) and nitrogen ($\delta^{15}N$) were also measured, though their low frequency and asymmetric distribution disqualified them from consideration as features in previous taxonomy models. Carbon did not reveal discernible geographic trends (cf. Supporting Information Figure S14); however, nitrogen revealed a breakdown of the region into 3 broad categories: Norton Sound, offshore in the Bering Sea, and other coastal areas (cf. Supporting Information Figure S15). While $\delta^{15}N$ measurements have been previously used to describe the trophic level of seabirds, we posit that this is a result of baseline differences in food webs, as previously found for Thick-billed murre.[65−67]

Figure 8 illustrates that total mercury, and perhaps 4,4′ DDE (cf. Supporting Information), can be used to make a simple decision tree model, but clearly does not provide enough information to uniquely fingerprint each colony on its own. As in the case of murre, we can search for more advanced models when features suggested by the JSD are insufficient; however, a random forest was found to consistently outperform all other models, so we selected this for further investigation.

Figure 9a shows the results of 5 × 4 nested, stratified CV to assess performances of various models; the random forest achieved 78.77 ± 3.16% when only 33 of the 49 analytes were included in the model, following decorrelation. The accuracy was nearly identical when all 49 analytes were included, but removing more analytes (down to 13) caused the mean accuracy to drop more than 6%. An 80:20 train/test split was performed where 5-fold, stratified CV was used to optimize a class-balanced random forest's size on the training set. The

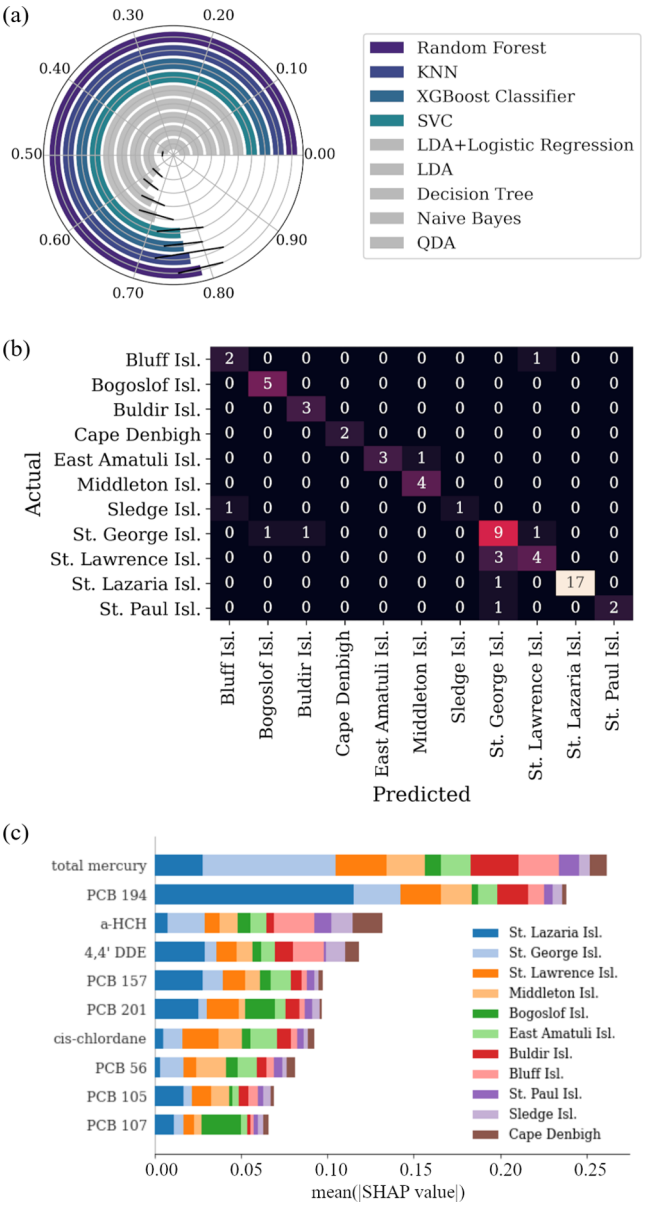


Figure 9. Classifying murre colonies. (a) Performance of different models estimated by nested, stratified 5 × 4 CV when 33 decorrelated analytes were used; one standard deviation is shown as the error. The color scheme follows the convention in Figure 6. (b) Confusion matrix for a random forest model trained on 80% of the data and tested (shown) on the remaining 20%. (c) Overall SHAP summary of the importance of the top 10 analytes in the model is shown in (b).

optimal forest contained 270 trees and had an 82.54% accuracy on the test set; the confusion matrix is shown in Figure 9b. While roughly 80% accuracy was considered insufficient for species classification, due to the small number of different classes; here, it is clear that the errors tend to be geographically correlated. For example, St. Lawrence, St. George, and St. Paul Islands are neighboring colonies in the Bering Sea; the model makes most of its mistakes when it confuses these three with each other. Furthermore, Sledge Isl. and Bluff Isl. are neighbors in Norton Sound, while St. George and Bogoslof Isl. are neighbors further south.

Thus, the model reflects geographic trends reasonably well, though specific regions may be somewhat diffuse and difficult to precisely define. Proximal colonies were similar enough to

https://doi.org/10.1021/acs.est.2c01894
Environ. Sci. Technol. 2022, 56, 14361−14374

suggest that they may be combined to improve sampling power if the goal is to monitor regional water bodies, allowing for other colonies to serve the same purpose when colonies targeted for monitoring efforts cannot be sampled (e.g., due to weather disruptions, sampling impacts on colony health, and unproductive years). We used the SHAP methodology to explain the random forest's probability of assigning a sample to a colony; this provides insight into the way the model determines characteristic regional differences.[24,25]

Figure 9c shows the overall summary of the top 10 features, determined by the mean absolute SHAP value summed over all the colonies. The levels of total mercury and PCB 194 are exceptionally important and have rational origins. We found that St. Lazaria Isl. has a uniquely elevated level of PCB 194 (cf. Supporting Information Figure S13) compared to other colonies and has an OvA JSD of 0.90, the highest of any analyte for any colony (cf. Supporting Information Figure S9). Simply put, the model tends to rely on the level of PCB 194 to decide if a sample originates from St. Lazaria Isl., then uses the total mercury content to narrow down the remaining possibilities. As shown in Figure 8, most colonies have a fairly unique mean mercury level, though its distribution may overlap with several neighbors. The remainder of the analytes are used to differentiate between these neighbors, creating complex but unique fingerprints. The identification of a different analyte (PCB 194) that can provide a good signature for St. Lazaria Isl. is critical to the model, since mercury levels at this colony display the broadest distribution of any colony (cf. Figure 8); this would likely confuse any model relying on mercury levels alone, necessitating additional discriminating information.

Analyzing Temporal Trends. The models found in this study are premised on the bioaccumulation of anthropogenic contaminants. Due to ongoing pollution or, conversely, regulatory restrictions and environmental remediation efforts, it is not clear whether the trends in the features (analytes) of our models are biased based on the year in which they originate since it is plausible that they may increase or decrease over time. To establish whether or not the models we have developed are expected to be applicable to timeframes outside the sampling window (1999−2010), we sought to establish if there exist any significant temporal trends.

We have already shown chemometric trends that correlate with taxonomy and geography, so we chose to consider only samples from individual colonies for individual species to avoid the confounding factor that data was not analyzed evenly over time or sample location (cf. Supporting Information Figure S17). Based on the available data, only St. Lazaria Isl. had sufficient data for reasonable analysis. This is a murre colony from which 89 samples were available; 53 were Common murre and 36 were Thick-billed murre (cf. Supporting Information Figure S8b). As suggested by our previous results, we ignored this difference. We computed the Spearman rank-order correlation between the mean analyte level and the collection year. This is a nonparametric measure of the monotonicity of the relationship between the two, which allows us to assess if an analyte's level is generally increasing ($r_S > 0$) or decreasing ($r_S < 0$) over time. We also computed the p value for a two-sided hypothesis test whose null hypothesis is that the year and mean mass fraction are uncorrelated ($r_S = 0$). We caution this is generally only reliable for large data sets, but we found it to give reasonable results.

We chose a significance level of $\alpha = 0.05$ and used this as a threshold; only analytes with $p < \alpha$ were considered to have significant trends. Figure 10 shows the results. No analytes

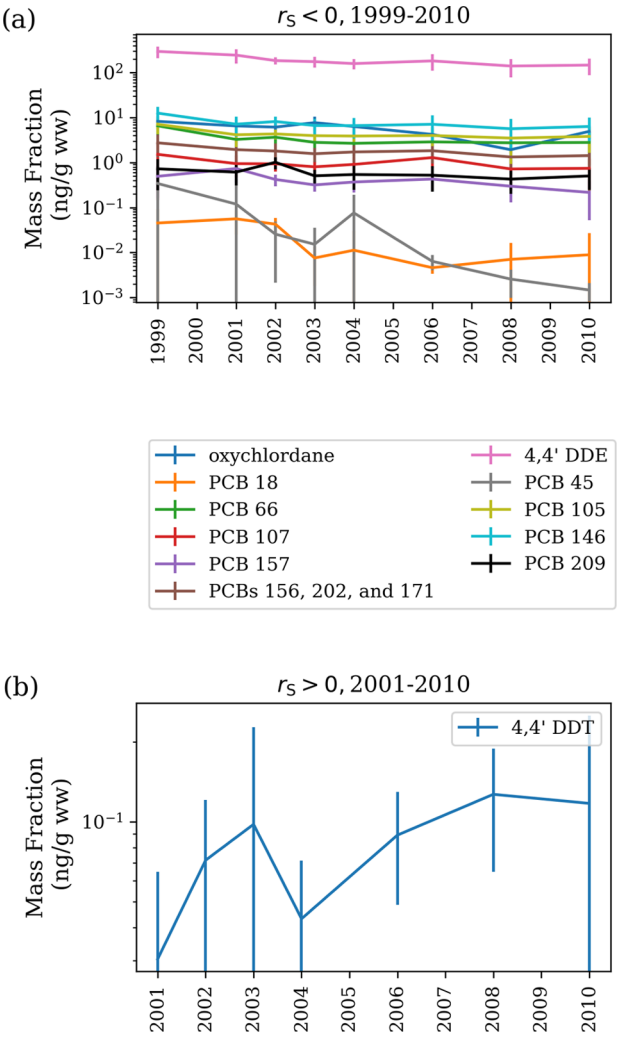


Figure 10. Mean mass fractions of analytes measured in murres from St. Lazaria Isl. exhibiting a statistically significant trend over time. Error bars correspond to one standard deviation. (a) Analytes trending downward as determined by their Spearman rank-order correlation, r_S, using data from all available years. (b) Only analyte with a statistically significant $r_S > 0$ when data from the years 1999 and 2000 are removed.

were found to have $r_S > 0$ with significance, which is encouraging. Of the original 49 analytes, only 11 were found to be decreasing over time at St. Lazaria Isl.; however, samples collected from 1999 to 2000 were analyzed using GC−electron capture detection (ECD), while in subsequent years GC−MS was used. The GC−ECD method may have resulted in measurements that appear elevated relative to GC−MS, which is expected to skew the results. This trend is more apparent when not plotted in the log scale (cf. Supporting Information Figure S18). Repeating the calculation on data only from the year 2001 and later (79 out of 89 total measurements), we found that the downward trend of only the final 5 analytes in Figure 10a was significant (second column in legend, 4,4′ DDE—PCB 209). We also note that 4,4′ DDT exhibited a significant upward trend if the earlier data was removed, but not if that data was included.

https://doi.org/10.1021/acs.est.2c01894
Environ. Sci. Technol. 2022, 56, 14361−14374

In all cases, however, the trends are very weak; error bars in Figure 10 show one standard deviation of the measurements made for a given year. This is consistent with the expectation that these p values are not expected to be as reliable for such a small (roughly 10 year-averaged values) set of measurements. Moreover, this agrees with the earlier preliminary investigation of seabirds in this area, which was unable to discern any temporal trends.[62] St. George Isl. had the second most observations; repeating this analysis for those murres yielded similar results (cf. Supporting Information Figure S19), though the specific analytes exhibiting a trend varied relative to St. Lazaria Isl. Regardless, analytes playing significant roles in models described earlier in this work were generally not among those exhibiting a significant temporal trend. While local variations are certainly possible, this suggests that overall environmental levels for analytes important to these discriminatory models are reasonably stable, justifying the use of these pollutants as features to distinguish taxonomy and geographical origin.

We further note that the differences which arise in this analysis due to differences in measurement technique highlight the importance of biobanking efforts. As a difference was observed that may be due to analytical variations, despite proper use of reference materials (SRM 1946 Lake Superior Fish Tissue),[68] additional aliquots of all samples are banked at the NIST Biorepository and are available to be reanalyzed in the future. This enables the determination of whether trends are accurate or are an artifact of changing instrumentation and analytical methodology.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.est.2c01894.

> Absolute value of the coefficients on the leading discriminant axis for genus; decision tree models for genus; decorrelating analytes using hierarchical clustering; dendrogram from Ward clustering of the Spearman rank-order correlations; heatmap of the Spearman rank-order correlations; machine learning models to predict species with statistical comparison; colonies from which samples in the database originated; OvA JSD for each murre colony for all analytes considered; geographic distribution of various analytes; detailed SHAP analysis of the random forest model to predict colony; and additional temporal trends (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Nathan A. Mahynski − *Chemical Sciences Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899-8320, United States;* orcid.org/0000-0002-0008-8749; Email: nathan.mahynski@nist.gov

Authors

Jared M. Ragland − *Chemical Sciences Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899-8320, United States;* orcid.org/0000-0002-8055-2432

Stacy S. Schuur − *Chemical Sciences Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899-8320, United States;* orcid.org/0000-0001-5926-4631

Vincent K. Shen − *Chemical Sciences Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899-8320, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.est.2c01894

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Contribution of the National Institute of Standards and Technology, not subject to US Copyright.

■ REFERENCES

(1) Brown, T. M.; Takada, H. Indicators of marine pollution in the North Pacific Ocean. *Arch. Environ. Contam. Toxicol.* **2017**, *73*, 171−175.

(2) Fernández-Llamazares, Á.; Garteizgogeascoa, M.; Basu, N.; Brondizio, E. S.; Cabeza, M.; Martínez-Alier, J.; McElwee, P.; Reyes-García, V. A state-of-the-art review of indigenous peoples and environmental pollution. *Integrated Environ. Assess. Manag.* **2020**, *16*, 324−341.

(3) AMAP. *AMAP Assessment Report: Arctic Pollution Issues*, 1998.

(4) Becker, P. H.Biomonitoring with birds. *Trace Metals and Other Contaminants in the Environment*; Elsevier, 2003; Vol. *6*, Chapter 19, pp 677−736.

(5) Pol, S. S. V.; Becker, P. R. Monitoring contaminants in seabirds: The importance of specimen banking. *Mar. Ornithol.* **2007**, *35*, 113−118.

(6) Rigét, F.; et al. Temporal trends of persistent organic pollutants in Arctic marine and freshwater biota. *Sci. Total Environ.* **2019**, *649*, 99−110.

(7) Gagné, T. O.; Johnson, E. M.; Hyrenbach, K. D.; Hagemann, M. E.; Bass, O. L.; MacDonald, M.; Peck, B.; Houtan, K. S. V. Coupled trophic and contaminant analysis in seabirds through space and time. *Environ. Res. Commun.* **2019**, *1*, 111006.

(8) Hollamby, S.; Afema-Azikuru, J.; Waigo, S.; Cameron, K.; Rae Gandolf, A. R.; Norris, A.; Sikarskie, J. G. Suggested guidelines for use of avian species as biomonitors. *Environ. Monit. Assess.* **2006**, *118*, 13−20.

(9) National Institute of Standards and Technology. Seabird Tissue Archival and Monitoring Project (STAMP). https://go.usa.gov/xHU58, (accessed on Sept 06, 2021).

(10) York, G. W.; Porter, B. J.; Pugh, R. S.; Roseneau, D. G.; Simac, K.; Becker, P. R.; Thorsteinson, L. K.; Wise, S. A., *Seabird Tissue Archival and Monitoring Project: Protocol for Collecting and Banking Seabird Eggs; Internal Report NISTIR 6735*, 2001.

(11) Rust, L. B.; Pugh, R. S.; Amanda, J.; Stacy, S.; Becker, P. R.; Roseneau, D. G., *Seabird Tissue Archival and Monitoring Project: Project overview, and updated protocols for collecting and banking seabird eggs; Internal Report NISTIR 7678*, 2010.

(12) Schuur, S. S.STAMP samples banked and analyzed. https://www.easymapmaker.com/map/STAMP, (accessed on Dec 07, 2021).

(13) National Institute of Standards and Technology. The NIST Biorepository. https://go.usa.gov/xHUNp, (accessed on Dec 07, 2021).

(14) Dietz, R.; et al. Current state of knowledge on biological effects from contaminants on arctic wildlife and fish. *Sci. Total Environ.* **2019**, *696*, 133792.

(15) Kalia, V.; Schuur, S. S.; Hobson, K. A.; Chang, H. H.; Waller, L. A.; Hare, S. R.; Gribble, M. O. Relationship between the pacific decadal oscillation (PDO) and persistent organic pollutants in sympatric Alaskan seabird (Uria aalge and U. lomvia) eggs between 1999 and 2010. *Chemosphere* **2021**, *262*, 127520.

(16) Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5−32.

(17) Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y.LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Adv. Neural Inf. Process. Syst.*, **2017**; *30*.
(18) Chen, T.; Guestrin, C.Xgboost: A scalable tree boosting system. *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery, 2016; pp 785−794.
(19) Mjolsness, E.; DeCoste, D. Machine learning for science: state of the art and future prospects. *Science* **2001**, *293*, 2051−2055.
(20) Karpatne, A.; Atluri, G.; Faghmous, J. H.; Steinbach, M.; Banerjee, A.; Ganguly, A.; Shekhar, S.; Samatova, N.; Kumar, V. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Trans. Knowl. Data Eng.* **2017**, *29*, 2318−2331.
(21) Rudin, C.; Chen, C.; Chen, Z.; Huang, H.; Semenova, L.; Zhong, C. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Stat. Surv.* **2022**, *16*, 1−85.
(22) Covert, I.; Lundberg, S. M.; Lee, S.-I. Explaining by Removing: A Unified Framework for Model Explanation. *J. Mach. Learn. Res.* **2021**, *22*, 1−90.
(23) Ribeiro, M. T.; Singh, S.; Guestrin, C."Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery, 2016; pp 1135−1144.
(24) Lundberg, S. M.; Lee, S.-I.*Advances in Neural Information Processing Systems 30*; Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc., 2017; pp 4765−4774.
(25) Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2020**, *2*, 2522−5839.
(26) Nori, H.; Jenkins, S.; Koch, P.; Caruana, R.Interpretml: A unified framework for machine learning interpretability. Submitted Sept 19, 2019. arXiv:1909.09223, (accessed on Sept 06, 2022).
(27) Dong, J.; Rudin, C. Exploring the cloud of variable importance for the set of all good models. *Nat. Mach. Intell.* **2020**, *2*, 810−824.
(28) Waskom, M. L. Seaborn: Statistical data visualization. *J. Open Source Software* **2021**, *6*, 3021.
(29) Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **2007**, *9*, 90−95.
(30) Met Office. *Cartopy: A Cartographic Python Library with a Matplotlib Interface*, version 0.16.; BibSonomy, 2010.
(31) Mahynski, N. A.; Ragland, J. M.; Schuur, S. S.; Pugh, R.; Shen, V. K. Seabird Tissue Archival and Monitoring Project (STAMP) data from 1999-2010. *J. Res. Natl. Inst. Stand. Technol.* **2021**, *126*, 1−7.
(32) National Institute of Standards and Technology. *STAMP Dataset 1999-2010*. https://github.com/mahynski/stamp-dataset-1999-2010, (accessed on Aug 18, 2022).
(33) Lemaître, G.; Nogueira, F.; Aridas, C. K. Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.* **2017**, *18*, 1−5.
(34) Pedregosa, F.; et al. scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(35) Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321−357.
(36) Batista, G. E.; Prati, R. C.; Monard, M. C.A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*; ACM, 2004; Vol. *6*, pp 20−29.
(37) Wilson, D. L. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Trans. Syst. Man Cybern.* **1972**, *2*, 408−421.
(38) Tomek, I. An experiment with the edited nearest-neighbor rule. *IEEE Trans. Syst. Man Cybern.* **1976**, *6*, 448−452.
(39) Steinberg, D.*The Top Ten Algorithms in Data Mining*; Chapman and Hall/CRC, 2009; pp 179−201.
(40) Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **2001**, 1189−1232.
(41) Hastie, T.; Tibshirani, R.; Friedman, J. H.*The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer, 2009; Vol. *2*.
(42) Kleinbaum, D. G.; Dietz, K.; Gail, M.; Klein, M.; Klein, M.*Logistic Regression*; Springer, 2002.
(43) Kononenko, I.Comparison of inductive and naive Bayesian learning approaches to automatic knowledge acquisition. *Current Trends in Knowledge Acquisition*; IOS Press, 1990; Vol. *331*.
(44) Hand, D. J.; Yu, K. Idiot's Bayes—not so stupid after all? *Int. Stat. Rev.* **2001**, *69*, 385−398.
(45) Cawley, G. C.; Talbot, N. L. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.* **2010**, *11*, 2079−2107.
(46) Wainer, J.; Cawley, G. Nested cross-validation when selecting classifiers is overzealous for most practical applications. *Expert Syst. Appl.* **2021**, *182*, 115222.
(47) Bouckaert, R. R.; Frank, E.Evaluating the replicability of significance tests for comparing learning algorithms. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*; Springer, 2004; pp 3−12.
(48) Bouckaert, R. R.Choosing between two learning algorithms based on calibrated tests. *Proceedings of the Twentieth International Conference on Machine Learning*; ICML, 2003; pp 51−58.
(49) Nadeau, C.; Bengio, Y. Inference for the generalization error. *Mach. Learn.* **2003**, *52*, 239−281.
(50) Ward, J. H., Jr Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* **1963**, *58*, 236−244.
(51) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087−1092.
(52) Datta, A.; Sen, S.; Zick, Y. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. *IEEE Symp. Secur. Priv.* **2016**, 598−617.
(53) Manning, C.; Schutze, H.*Foundations of Statistical Natural Language Processing*; MIT Press, 1999.
(54) Fuglede, B.; Topsoe, F.Jensen-Shannon divergence and Hilbert space embedding. *International Symposium on Information Theory, 2004 ISIT 2004 Proceedings*; IEEE, 2004; p 31.
(55) Fisk, A.; Moisey, J.; Hobson, K.; Karnovsky, N.; Norstrom, R. Chlordane components and metabolites in seven species of Arctic seabirds from the Northwater Polynya: Relationships with stable isotopes of nitrogen and enantiomeric fractions of chiral components. *Environ. Pollut.* **2001**, *113*, 225−238.
(56) McEwen, F.; Beardsley, J.; Hapai, J.; Su, T. Laboratory tests with candidate insecticides for control of the big-headed ant, Pheidole megacephala (Fabricius). *Proc. Hawaii. Entomol. Soc.* **1979**, *13*, 119−124.
(57) Department of Health State of Hawaii. *Hazard Evaluation and Emergency Response Office Activities for FY 2001*; HEER office, 2002.
(58) Wania, F.; Mackay, D.*The Global Fractionation of Persistent Organic Pollutants*; U.S. Department of Energy, 1996; pp 1−24.
(59) Scheringer, M.; Strempel, S.; Hukari, S.; Ng, C. A.; Blepp, M.; Hungerbuhler, K. How many persistent organic pollutants should we expect? *Atmos. Pollut. Res.* **2012**, *3*, 383−391.
(60) Kokoska, S.; Zwillinger, D.*CRC Standard Probability and Statistics Tables and Formulae*; CRC Press, 2000.
(61) Kokubun, N.; Yamamoto, T.; Sato, N.; Watanuki, Y.; Will, A.; Kitaysky, A.; Takahashi, A. Foraging segregation of two congeneric diving seabird species (common and thick-billed murres) breeding on St. George Island, Bering Sea. *Biogeosci. Discuss.* **2015**, *12*, 18151−18183.
(62) Day, R. D.; Roseneau, D. G.; Pol, S. S. V.; Hobson, K. A.; Donard, O. F.; Pugh, R. S.; Moors, A. J.; Becker, P. R. Regional, temporal, and species patterns of mercury in Alaskan seabird eggs: Mercury sources and cycling or food web effects? *Environ. Pollut.* **2012**, *166*, 226−232.
(63) Ritter, L.; Solomon, K.; Forget, J.; Stemeroff, M.; O'leary, C.A review of selected persistent organic pollutants. *International*

Programme on Chemical Safety (IPCS). PCS/95.39. Geneva: World Health Organization; Citeseer, 1995; Vol. 65.
(64) Pearson, C.; Howard, D.; Moore, C.; Obrist, D. Mercury and trace metal wet deposition across five stations in Alaska: Controlling factors, spatial patterns, and source regions. *Atmos. Chem. Phys.* **2019**, *19*, 6913−6929.
(65) Hobson, K. A.; Piatt, J. F.; Pitocchelli, J. Using stable isotopes to determine seabird trophic relationships. *J. Anim. Ecol.* **1994**, *63*, 786−798.
(66) Pol, S. S. V.; Hobson, K. A.; Becker, P. R.; Day, R. D.; Ellisor, M. B.; Pugh, R. S.; Roseneau, D. G. Geographic differences in organic contaminants and stable isotopes (δ 13 C, δ 15 N) in thick-billed murre (Uria lomvia) eggs from Alaska. *Journal of Environmental Monitoring* **2011**, *13*, 699−705.
(67) Elliott, K. H.; Braune, B. M.; Elliott, J. E. Beyond bulk δ15N: Combining a suite of stable isotopic measures improves the resolution of the food webs mediating contaminant signals across space, time and communities. *Environ. Int.* **2021**, *148*, 106370.
(68) Pol, S. S. V.; Ellisor, M. B.; Pugh, R. S.; Becker, P. R.; Poster, D. L.; Schantz, M. M.; Leigh, S. D.; Wakeford, B. J.; Roseneau, D. G.; Simac, K. S. Development of a murre (Uria spp.) egg control material. *Anal. Bioanal. Chem.* **2007**, *387*, 2357−2363.

https://doi.org/10.1021/acs.est.2c01894
Environ. Sci. Technol. 2022, 56, 14361−14374