Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: http://ees.elsevier.com

ELSEVIER

Mixture analyses of air-sampled pollen extracts can accurately differentiate pollen taxa

Leszek J. Klimczak [a], Cordula Ebner von Eschenbach [b], Peter M. Thompson [c,d], Jeroen T.M. Buters [b], Geoffrey A. Mueller [a,*]

[a] National Institute of Environmental Health Sciences, USA
[b] Center of Allergy & Environment (ZAUM), Member of the German Center for Lung Research (DZL), Technische Universität München/Helmholtz Center, Munich, Germany
[c] Molecular Education, Technology and Research Innovation Center, North Carolina State University, Raleigh, NC, USA
[d] Department of Molecular and Structural Biochemistry, North Carolina State University, Raleigh, NC, USA

ARTICLE INFO

Keywords
Pollen
NMR
Metabolomics
Mixtures
Exposure
Aerobiology

ABSTRACT

The daily pollen forecast provides crucial information for allergic patients to avoid exposure to specific pollen. Pollen counts are typically measured with air samplers and analyzed with microscopy by trained experts. In contrast, this study evaluated the effectiveness of identifying the component pollens using the metabolites extracted from an air-sampled pollen mixture. Ambient air-sampled pollen from Munich in 2016 and 2017 was visually identified from reference pollens and extracts were prepared. The extracts were lyophilized, rehydrated in optimal NMR buffers, and filtered to remove large proteins. NMR spectra were analyzed for pollen associated metabolites. Regression and decision-tree based algorithms using the concentration of metabolites, calculated from the NMR spectra outperformed algorithms using the NMR spectra themselves as input data for pollen identification. Categorical prediction algorithms trained for low, medium, high, and very high pollen count groups had accuracies of 74% for the tree, 82% for the grass, and 93% for the weed pollen count. Deep learning models using convolutional neural networks performed better than regression models using NMR spectral input, and were the overall best method in terms of relative error and classification accuracy (86% for tree, 89% for grass, and 93% for weed pollen count). This study demonstrates that NMR spectra of air-sampled pollen extracts can be used in an automated fashion to provide taxa and type-specific measures of the daily pollen count.

1. Introduction

Pollen allergies in the United States affect 30% of the population (Salo et al., 2014). Similarly, in Europe 40% of the people are estimated to suffer from pollen allergy (D'amato et al., 2007; Burbach et al., 2009). After a diagnosis of pollen allergy, patients can treat the symptoms with medication, and depending on the severity, immunotherapy may be suggested. However, avoidance of the allergen source is a primary part of the recommendations for a patient (Platts-Mills, 2004). For pollen allergies, the daily pollen count is the fundamental source of this information.

The pollen count is determined from air samplers that trap particulate matter by impaction (Verein-Deutsche-Ingenieure, 2017). The particulate matter is typically analyzed using microscopy by trained experts to identify up to 30 different pollen species. To improve either the accuracy, speed, or cost various other methods have been explored including image classification (Oteros et al., 2015; Marcos et al., 2015; Holt et al., 2011), DNA next generation sequencing (Kraaijeveld et al., 2015; Brennan et al., 2019), and chemical analysis (Buters et al., 2015). In Bavaria, multiple air samplers with robotic automated image classification have been installed representing a major investment (Buters et al., 2018). In terms of artificial intelligence, pollen grain images have been identified with deep learning artificial neural networks (Sevillano and Aznarte, 2018), and time series analyses of pollen forecasts have been attempted (Valencia et al., 2019). Other spectroscopic methods that have been examined include SERS Raman scattering (Seifert et al., 2016) and Fourier transform Raman scattering. (Zimmermann, 2010; Bagcioglu et al., 2015). While many of these systems are highly accurate, they are all still improving and other methodologies may be more efficient in other locations.

In our recent analysis of metabolites in extracts of major pollen species, we noticed that the metabolite concentrations could be used to differentiate the major pollen types: tree, grass, and weed (Mueller et al., 2016). We therefore hypothesized that measuring the metabolites in an extract of air sampled pollen may be useful in differentiating the pollen taxa that were present. A number of technical challenges needed to be resolved that were not present in the

[*] Corresponding author. 111 T.W. Alexander Dr., Research Triangle Park, NC, 27709, USA.
E-mail address: Geoffrey.Mueller@nih.gov (G.A. Mueller)

https://doi.org/10.1016/j.atmosenv.2020.117746
Received 18 March 2020; Received in revised form 24 June 2020; Accepted 26 June 2020
Available online xxx
1352-2310/© 2020.

previous study. First, purified pollens were previously used at very high concentrations in order to sample as many metabolites as possible but air sampled pollen rarely approaches the amounts previously used. Sensitivity of the NMR methodology was a concern. Second, numerous species-specific metabolites like the isoflavones and flavanones were apparent in the NMR data but there was not enough material to chemically characterize them further. However, because the NMR spectral signals are reproducible, these chemical fingerprints could be an asset in pollen type identification using algorithms trained to recognize the patterns. Third, the same species of pollen are known to contain different amount of allergen depending on day or location of harvest (Buters et al., 2012, 2015; Galan et al., 2013), and it is not known if this variability exists in pollen metabolites too.

Herein, we explore the use of NMR data of air sampled pollen extracts to analyze the pollen types present. First, we examined the taxa that were identified and how the data characteristics influenced the direction of predictive algorithms. Then we identified metabolites from the extract data and looked for patterns in the metabolites that might be predictive of type. In addition, we explored using the NMR spectra without human analysis as input to the predictive modeling. Interestingly, deep learning methods that utilized the NMR spectra only were the best performing of the predictive algorithms developed.

2. Materials and methods

2.1. Sampling

The pollen for the Hirst-type pollen trap were collected on a Melinex-tape on a 7-days rotating drum. The Melinex tape was then cut in 24 hrs pieces and analyzed according international standards by visual identification at 400 × magnification using Safranine red stain and pure pollen samples as a reference (Galan et al., 2014). Pollen samples for extracts were obtained using a high-volume Chemvol® cascade impactor (Butraco Inc., Son, Netherlands) collected daily at the same location on polyurethane PM 10 foam filters as described by Buters et al. (2012). There were 119 samples from the peak pollen season 2016 in Munich, Germany (April 1 to July 30, 2016) and 10 samples from 2017. The 10 samples from 2017 were all from high pollen count days in order to improve the number of samples with various taxa present. In brief: 800 l min $^{-1}$ ambient air was sampled daily on polyurethane foam filters with the impactor equipped with different size class stages. The polyurethane PM 10 filters were pre-washed twice using the same buffer as used in the allergen extraction protocol, followed by 3 washes of distilled water and dried at 37 Celsius. The filter was cut into 3 equal parts and stored at −80 °C. For extraction, the filter slices of impacted pollen were thawed to room temperature.

2.2. Extraction

Samples were extracted from two of the PM 10 foam filter slices in 15 ml/slice of an 0.1 M ammonium bicarbonate buffer pH 8.1 containing 0.1% Bovine Serum Albumin (BSA). Extraction was done for 4 h at room temperature in a head-over-head rotator. Extracts were aliquoted for various analyses, frozen at −80 °C and lyophilized (Buters et al., 2015). For the NMR analysis 4 ml of dried extract was utilized; a separate 4 ml portion was used for the extract duplicates annotated in Supplemental Table 2. The dried extract was re-hydrated in 700 μl with a buffer of 50% ^2H_2O, and 50% ^{2}H-phosphate buffered saline (^{2}H-PBS) and 0.2 mM 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS), which was termed 50%-DPBS. DSS is used for a concentration and chemical shift reference, and the low salt ^{2}H-buffer is designed for maximum NMR sensitivity. The samples were filtered with 3 kDa Amicon concentrator to remove large molecules. This removed broad peaks and resulted in a flat NMR baseline for accurate quantitation and identification of metabolite peaks. The concentrators were rinsed and centrifuged twice with ^2H_2O to remove the glycerol coating on the filters; however some exogenous glycerol remained, see below.

2.3. NMR

NMR data were acquired on a Bruker 700 MHz Avance NEO spectrometer with a TCI cryoprobe and 60-sample changer. The pulse sequence was a 1D presaturation NOESY with 100 ms mixing time, 4s data acquisition, 1s recycle delay, with 12 steady state scans and 1024 scans for about 1.5 h of total sampling. The data were processed and analyzed with Chenomx version 8.4 (Alberta, Canada) to measure metabolite concentrations. Metabolites were identified with the help of the standard chemical library in Chenomx, by assessing the splitting patterns and corresponding peak intensities. Concentrations were measured with reference to DSS. The total number of spectra analyzed includes several duplicate data sets. The extraction was repeated and analyzed for 12 days in 2016, and the NMR data acquisition was repeated for 6 days. These are annotated in Supplementary Table 1. The final data set includes 129 days plus 18 replicates for 147 spectra analyzed.

2.4. NMR metabolite analysis and machine learning

The metabolite, NMR, and pollen data were imported into MATLAB (Mathworks, Massachusetts, USA) for analysis. The following regression and classification algorithms were tested to predict the pollen count: linear, binary tree, support vector machines, ensemble learners, and gaussian process. For classification of low, medium, high, and very-high pollen counts the following were tested: decision trees, discriminant analysis, naïve Bayes classifiers, support vector machines, K-nearest neighbor methods, and ensemble classifiers. For each method 2–7 variants of the method parameters were tested to optimize its performance. As described below, the input data was tested using various scaling techniques to improve the predictions. Since the size of the data set is considered small (n = 147), the predictive value of the regression and classification models were assessed by randomly dividing the data into training (90%) and testing (10%) sets. This was repeated 10 times and reported as the ten-fold cross validated performance. This was not feasible for deep learning where the data was considered more limited and leave-one-out cross validation was performed. In this case, all combinations of the training set (n = 146) and test data (n = 1) were assessed in the model building, and repeated 5 times with different initial conditions.

Principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) (Van Der Maaten and Hinton, 2008) analyses were also performed with MATLAB.

2.5. Additional NMR spectral processing

For the NMR spectral data as input to machine learning and deep learning, several additional pre-processing measures were taken. The signals from water and DSS were excised from the input because they were very large compared to the metabolite data. The signal intensities in the NMR data varied over 10 orders of magnitude, making it difficult to discern rare compounds with low intensities. The intensities were scaled by taking the square root to reduce the range of signal intensities to 5 orders of magnitude. In data not shown this appeared to improve performance. However, this scaling can also have the detrimental effect in that it can enhance the noise. The NMR data was also smoothed with the MATLAB 'loess' function to reduce noise and avoid overfitting artifacts in the very large input vectors. Regarding the size of the input vector, the NMR data was acquired with over 70,000 points but was downsampled by binning adjacent data points to ~2000 data points, which preserved reasonable spectral resolution, smoothed some noise artifacts, and reduced the computational time for training.

2.6. Deep learning

After an initial exploratory analysis of various deep learning methods, convolution neural networks were selected for predictive analysis. Details of the spectral processing, network architectures, and initial tests are provided as supplementary material. In order to understand which parts of the spectrum contributed most to the prediction, a series of mathematical disruptions were applied to a series of non-overlapping windows of constant size across the whole spectrum. The results with and without disruptions were compared.

3. Results

3.1. Data analysis

The various pollen types, and taxa that were identified in Munich during the 2016 pollen season are graphed in 3 different ways in the panels of Fig. 1. Panel A shows the groups of tree, grass, weed, and indeterminate pollen versus day of the year, while Panel B more finely separates the top 14 taxa by total pollen count versus day. The trees dominate the pollen count early in the year, while the grass pollen emerges around day 150. Weed pollen emerges slightly later and continues longer than the grass pollen until the end of the sampling. Panel C shows the total pollen counts for the entirety of 2016 sorted by total. Panels B shows that the tree taxa dominate the total pollen count on a daily basis, but the total annual pollen from weeds (Urticaceae) and grasses (Poacea) are the in the top four in panel C.

Fig. 1 illustrates several features of the pollen count data that will need to be considered in a predictive algorithm. First, the pollen counts vary over 3 orders of magnitude for a given type or taxa. Second, there are days between which multiple types of pollen can also vary orders of magnitude. Third, from panel C some of the taxa which are rare will be difficult to predict from NMR data, which is traditionally signal limited. Based on this consideration, further analysis considered only the 10 most abundant pollens, which included 6 deciduous trees, *Pinus*, Poacea (grasses), and Urticacea (nettles or weeds). The category indeterminant was excluded from further analysis. *Pinus* was the lowest of the 10 abundant pollens and as will be seen below it was problematic to create predictive models, suggesting this was a good cut off for further modeling.

Fig. 2 shows example NMR data of the pollen extract from day 164, plotting intensity versus frequency. In terms of quality, there is good water suppression (around 4.7 ppm) and flat baselines at the edges so that neither can interfere with accurate quantitation. The black line is the raw NMR data, and the red line indicates peaks that were identified and fit with the Chenomx software. Not all peaks could be identified, but most of the intense peaks were consistently identified in all spectra.

Nineteen compounds were readily identified in the NMR data and key peaks are highlighted in Fig. 2: acetate, alanine, betaine, citrate, DSS, formate, glycerol, lactate, proline, and sucrose. Less abundant metabolites included choline, ethanol, glucose, methylguanidine, N-acetylglycine, O-acetylcholine, pyruvate, succinate, and trigonelline. Importantly, methylguanidine, N-acetylglycine, and O-acetylcholine are pseudonyms for plant compounds that resemble these metabolites but could not be confirmed as such. In other words, N-acetylglycine is an alias for an N-acetyl peak typically in the chemical

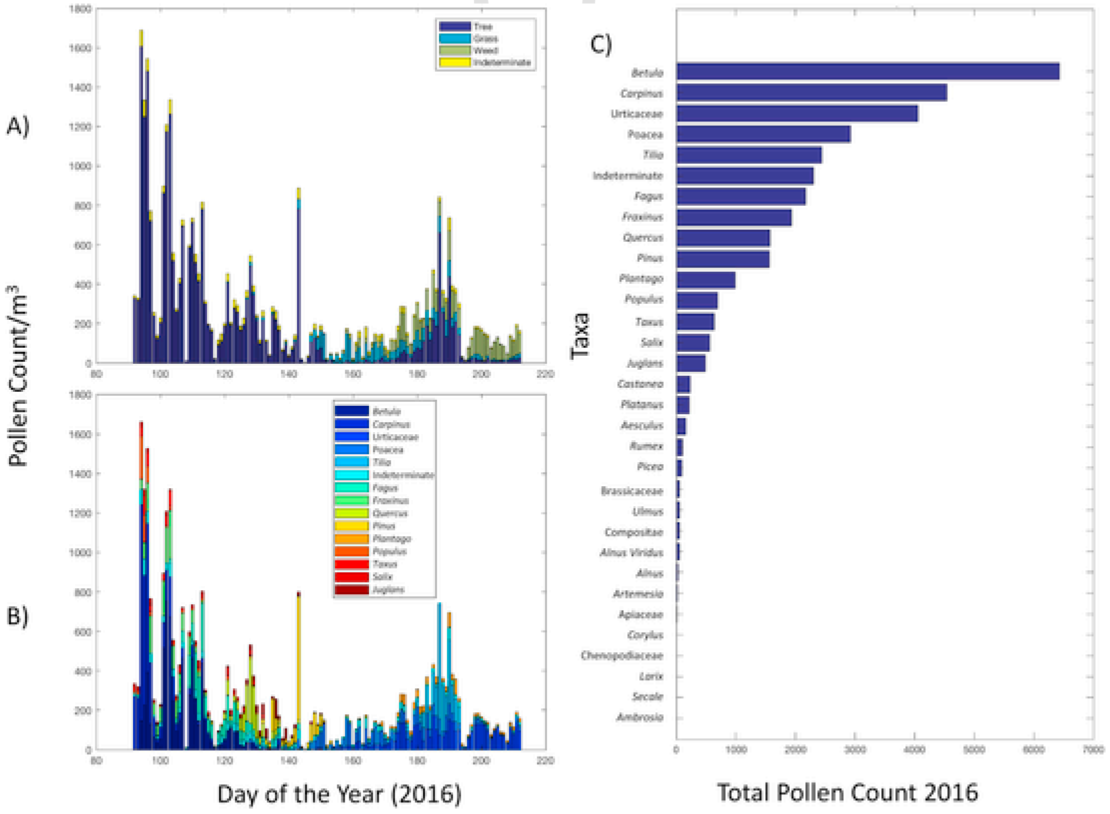

Fig. 1. Pollen Sampled in Munich 2016. Panels A and B show the pollen count/m^3 sampled by day of the year (2016) subdivided into A) Tree/Grass/Weed and B) by Taxa. Panel C shows the total pollen for the year sorted by amount.

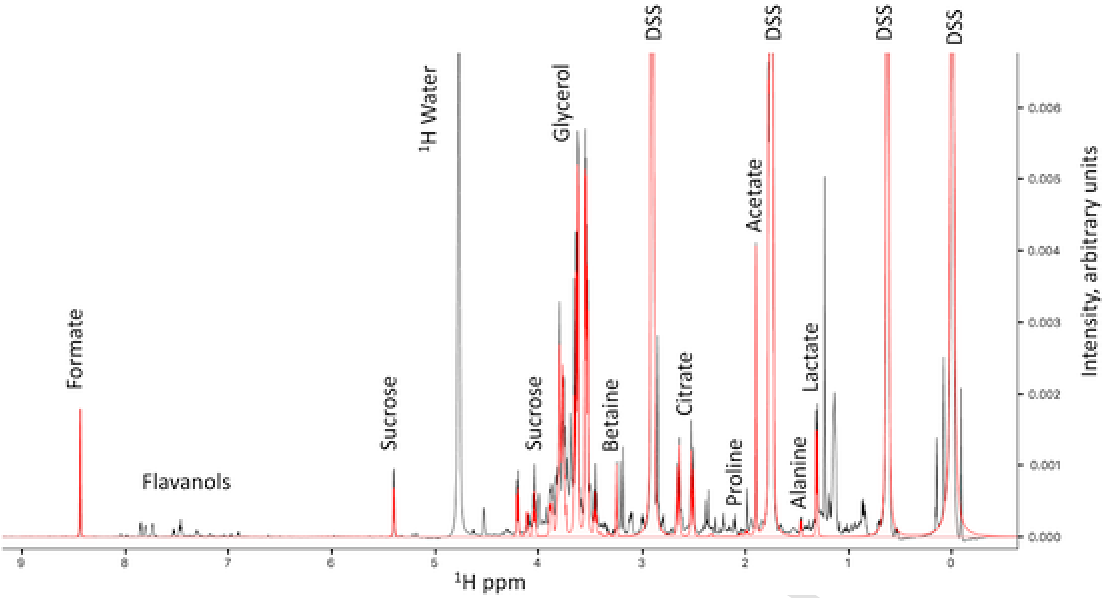

Fig. 2. Example NMR data. Fig. 2 shows an example NMR spectrum of Day 164 in black overlayed with identified metabolites and the metabolite-simulated spectrum in red. DSS is the concentration and chemical shift standard, see Methods. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

range of N-acetylglycine that was found consistently in many of the spectra. The DSS was added as a reference standard and the glycerol is most likely acquired from the concentrator filters. A control washing experiment suggested that the approximately 20 μM glycerol in each sample is not endogenous to the pollen (data not shown). Hence this was excluded from further analysis. Supplemental Fig. 1 plots the concentration of each metabolite measured, at each day of calendar year sampling.

The concentration data was probed for correlations between the amount of the 3 main pollen types, and the concentration of the metabolites. Fig. 3 shows the pairwise correlations of the different metabolites with the tree, grass, weed pollen count sorted for positive and negative correlations. Looking at the y-axis for each type, no very strong associations were found a single metabolite, likely reflecting the convoluted nature of the pollen data. The strongest associations both positive and negative were with the calendar day of the year. The day negatively correlated with tree pollen, which makes sense as most trees pollinate early in the spring. The ordering of metabolite correlations (x-axis) with either tree, grass, or weed pollen was unique to each type. Taken together, this does suggest that underlying patterns might be found in the metabolite concentration data that would be predictive of the pollen count.

To further probe if there were underlying patterns in the NMR metabolite or spectral data that reflected the major pollen categories various algorithms for dimensionality reduction were examined for clustering. Fig. 4 shows some of the more promising figures using either t-SNE or principal component analysis that suggested the input data may be predictive of the various categories. Fig. 4 A-C shows a t-SNE analysis of the metabolite data where the days are colored for whether the total tree (A), grass (B), or weed (C) pollen count was in the High, Medium, or Low categories. See Supplementary Table 2 for the categorical ranges. In Fig. 4A the high tree days primarily cluster to the left and in Fig. 4C the high weed days cluster tightly on the right. In 4B the high grass days cluster to the right and overlap significantly with the high weed days, as one might expect from examining the underlying pollen count data in Fig. 1A. Fig. 4 D-F shows a principal component analysis of the NMR spectral data with similar conclusions. The high tree days are separated from the high grass days, and the high weed days appear to be a subset of the high grass days. This figure is encouraging in that there are underlying patterns in the data which might not be immediately apparent in the correlations of Fig. 3 that could be exploited with a predictive algorithm.

3.2. Developing predictive models using metabolite input

In data not shown, we experimented with the various regression and classification algorithms listed above in the methods. Consistently, the ensemble boosted regression, and binary decision tree algorithms were superior to the others for these data. The RMSD and correlation of the predicted versus true pollen count were very similar for both methods and the best results from either are reported in Table 1 with 10-fold cross-validated performance. Table 1 shows strong correlations of the actual versus predicted values (R^2), except for *Pinus* (see below). The root mean squared error (RMSE) shows reasonable deviations proportional to the total pollen. For a proportional comparison, the root relative squared error (RRSE) column can be used. For example, *Betula* has an RMSE of 114 while Poaceae has an RMSE of 18, but both have an RRSE of 0.52 because the total birch tree pollen is simply much greater than the total for grasses on a given day.

To show the predictive accuracy from Table 1 visually, Fig. 5 plots the actual and predicted pollen counts versus day of the year for the groups and taxa studied. Fig. 5 highlights again the strong correlations between the actual (black circle) and predicted (red x) values for the models. *Pinus* showed the worst correlation among the taxa studied. As there was one very high pollen count day, the value was rarely predicted well with 10-fold cross validation. The plot shown in Fig. 5 is one of the better of the 10 predictions. The genus *Pinus* was the lowest total pollen count chosen for analysis (Fig. 1C), and occurred over a very limited time range. Overall, the results are encouraging that the NMR data, combined with the calendar day, can make a successful predictive model.

A key to the performance of the regression was to sort the input data for the 9 strongest correlation magnitudes, either positive or negative, for each taxa and discard the other input data. Then all 2^9 unique combinations of the input data were tested for their predictive capacity with 10-fold cross validation. Typically, the top ten fits

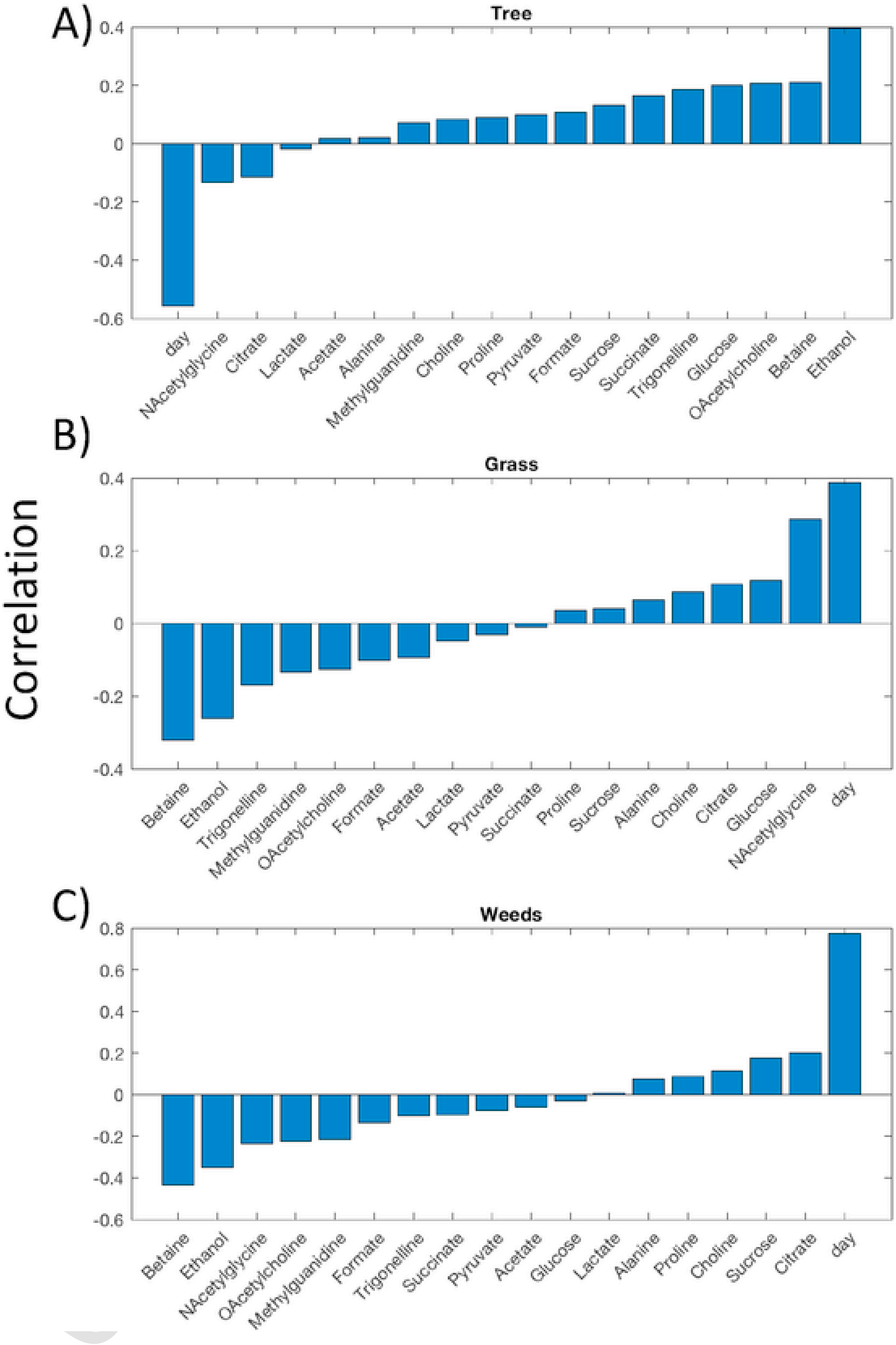

Fig. 3. Pairwise Correlations. The correlation of the input (calendar day or metabolite) versus the pollen count for A) tree pollen, B) grass pollen, and C) weed pollen.

were very close in their performance. Thus, to assess which input data was the most important, the top ten performing models were analyzed in Supplementary Fig. 2. The bar graphs show which inputs were used most often in the ten best models for each taxa. The calendar day was almost always utilized in the best fit, but a variety of different metabolites were used in different priorities by the vari-

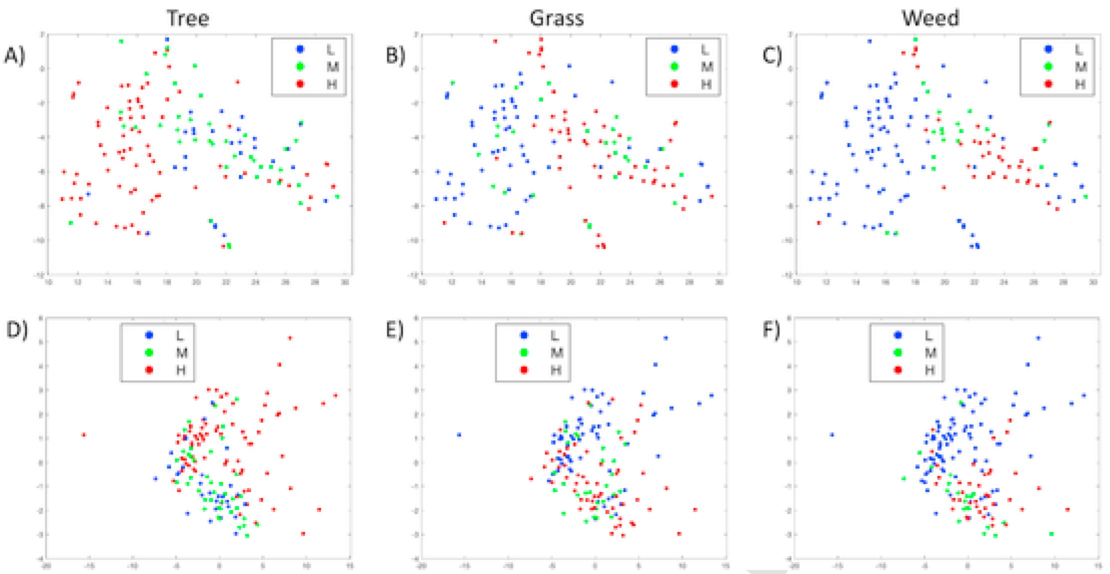

Fig. 4. Dimensionality reduction of the NMR data. Panels A–C show a t-SNE analysis of the metabolite data colored for high (red), medium (green), and low (blue) pollen count days for tree (A), grass (B), and weed (C). Panels D–E show a PCA plot of the first two principal components of the NMR spectral input colored as above for tree (D), grass (E), and weed (F) pollen counts. The PCA and t-SNE x and y axes are unitless and scaled for maximum dispersion of points. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 1
Performance of the Regression and Convolution Neural Networks. The predictive value of the different inputs and models were assessed.

Modeled Pollen Count	Input	Model	R2	RMSE	RRSE	Classification Accuracy	Classifier Accuracy
Betula	Metabolites	Regression	0.90	114	0.52	N.A.	N.A.
Carpinus	Metabolites	Regression	0.81	81	1.10	N.A.	N.A.
Fagus	Metabolites	Regression	0.77	26	0.80	N.A.	N.A.
Fraxinus	Metabolites	Regression	0.89	65	0.73	N.A.	N.A.
Quercus	Metabolites	Regression	0.84	21	0.90	N.A.	N.A.
Tillia	Metabolites	Regression	0.90	42	1.30	N.A.	N.A.
Pinus	Metabolites	Regression	0.27	55	0.62	N.A.	N.A.
Tree (Deciduous)	Metabolites	Regression	0.87	212	0.55	82%	90%
	NMR Spectra	Regression	0.73	325	0.76	56%	73%
Tree (All)	Metabolites	Regression	0.78	270	0.78	62%	74%
	NMR Spectra	Regression	0.67	365	0.87	54%	75%
	NMR Spectra	CNN	0.92	220	0.48	82%	86%
Poacea (Grass)	Metabolites	Regression	0.89	18	0.52	84%	82%
	NMR Spectra	Regression	0.60	34	0.94	62%	73%
	NMR Spectra	CNN	0.85	21	0.57	84%	89%
Urticacea (Weed)	Metabolites	Regression	0.88	23	0.48	91%	93%
	NMR Spectra	Regression	0.83	29	0.60	80%	88%
	NMR Spectra	CNN	0.94	17	0.35	88%	93%

Abbreviations: CNN- convolutional neural network.
R ^{2}- Peason's correlation coefficient.
RMSE-root mean squared error.
RRSE-root relative squared error.

ous models. Without the calendar day input the average correlation between actual and predicted values was only 0.3.

The National Allergy Board of the United States (NAB) typically does not identify pollen by genera in its forecast but instead categorizes it by the three types: tree, grass, and weed. To model this, the tree pollen counts for the individual genera were summed. The results were poorer than the individual trees and we hypothesized that this might be due to the *Pinus* pollen. By removing this genus the results improved as can be seen in Table 1 and Fig. 5 under the heading DTree for deciduous trees. The pollen counts can also be modeled in categorical manner using the metabolites and calendar day as input, according the values prescribed by the NAB in Supplemen-

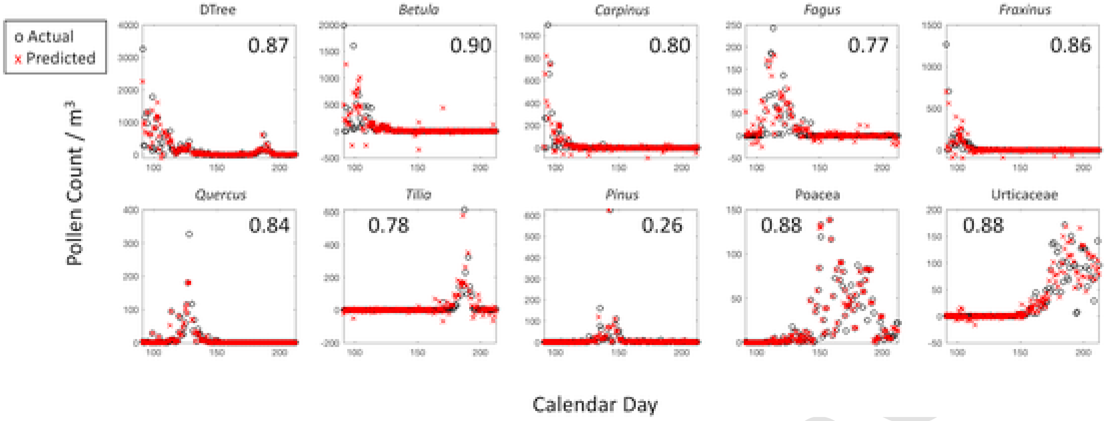

Fig. 5. Predicted pollen taxa compared to known pollen count. Each graph shows the actual pollen count per m^3 for the calendar day with a black circle, and the predicted value from the best model with a red x. The correlation coefficient of the predicted and actual value is shown inset for cross-validated performance. DTree is the sum of all deciduous trees. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

tary Table 2. The results of these predictive models can also give some perspective on the RMSD and RMSE values reported in Table 1. The accuracy of predicting the correct category, i.e., low, medium, high, and very high (for a few days in tree pollen season), is reported in Table 1. Weed pollen counts achieved the highest categorical predictive accuracy using 10-fold cross validation at 93%, tree pollen counts at 90%, and grass pollen at 82%. Grass is the least abundant of the pollens, hence it is most likely the one to be most difficult for NMR methods to measure accurately. The influence of the date as an input parameter was assessed for the categorical models as well. Without the date input, the accuracy of predicting the correct categorical pollen count was 59%, 63%, and 76% for the tree, grass, and weed categories respectively. The individual tree genera were not considered for categorical analysis because of the limited time range for which the pollen count was not 'low'. In other words, guessing 'low' for the entire date range was potentially 85% accurate.

3.3. Machine learning predictions using NMR spectra

The NMR spectra has numerous peaks from plant compounds that could not be uniquely identified, as shown in Fig. 2 by the difference between the red and black lines. This includes many flavanols and other unidentified compounds. This suggested that regression models directly using the spectral data directly may improve the classification. However, Table 1 shows the results were typically worse than using the metabolites alone. We suggest that the input vectors are too large for successful analysis, even using partial least squares discriminant analysis (PLSDA) regression techniques that seek to simplify large input vectors. However, the classification accuracy for grass and weed pollens was on par with the metabolite data.

3.4. Deep learning using NMR spectra

In order to more fully explore the potential of the spectral data beyond standard regression techniques, deep learning methods were examined. There were a number of challenges with the pollen data set compared to a typical deep learning data set, which includes the small number of training samples. Leave-one-out cross validation with 5 replicates was used to address the small number of training samples to insure that the predictions are not biased by any one sample. Other technical issues are addressed in the supplemental material.

Another issue was the large number of samples with low pollen counts. Sample weights were added to compensate for the underrepresented (high) pollen counts to improve the achieved predictions. As shown in Fig. 6, using sample weights caused the predictions for higher pollen counts to align closer with the straight line of perfect predictions, but – not surprisingly – the corresponding predictions for lower pollen counts were now scattered away from the actual values. The correlation coefficient of actual and predicted pollen counts was 0.919 without sample weights and 0.908 with sample weights thus, the overall accuracy of predictions was not improved as shown in Supplementary Fig. 4. Neither was the accuracy of predictions improved when the numeric predicted pollen counts were converted to categorical prediction (Supplementary Fig. 4). The predictions of pollen count for grass and weed pollen showed similarly high correlation coefficients of 0.854 and 0.945, respectively (Fig. 6), as well as accuracies when converted to categorized predictions (Supplementary Fig. 4). The overall accuracies are reported in Table 1. The results were slightly better when training was performed directly on categorical models, instead of just categorizing numerical predictions from regression models (listed as "Classification Accuracy" in Table 1; see also Supplementary Fig. 5). Similar to the regression model, including sample weights for underrepresented pollen counts did not result in increase of overall accuracy, although the accuracy for the Very High count group was dramatically improved.

In the cases of tree and grass pollen count predictions, the classification predictions improved over traditional regression models, and in fact the grass classification using deep learning and the NMR spectral input was the best method found (Table 1). The accuracy of the weed pollen count predictions using deep learning versus regression were equivalent. Also, note that for deep learning all the deciduous and pine trees were pooled together, and calendar day of the year was not included as input.

We identified the areas of the spectra that were responsible for predictions in the deep learning regression models by introducing various localized disruptions of the input spectra ("in silico smudging") and analyzing their effect on the predictions. While many areas of the spectra showed no effect (the same prediction before and after the disruption), disruption of multiple areas resulted both in increases (yellow) and decreases (purple) of the predicted outcome values in the regression models (Supplementary Fig. 6). After testing a series of scanning windows of various sizes, we identified two major, although still similar to each other, patterns: a larger-scale pattern with the window size of 200 pixels and more and a finer small scale pattern with the window size of 50 pixels and less. Interestingly, the average width of an NMR peak in our dataset is around 100 pixels, so these two patterns appear to correspond to disruptions

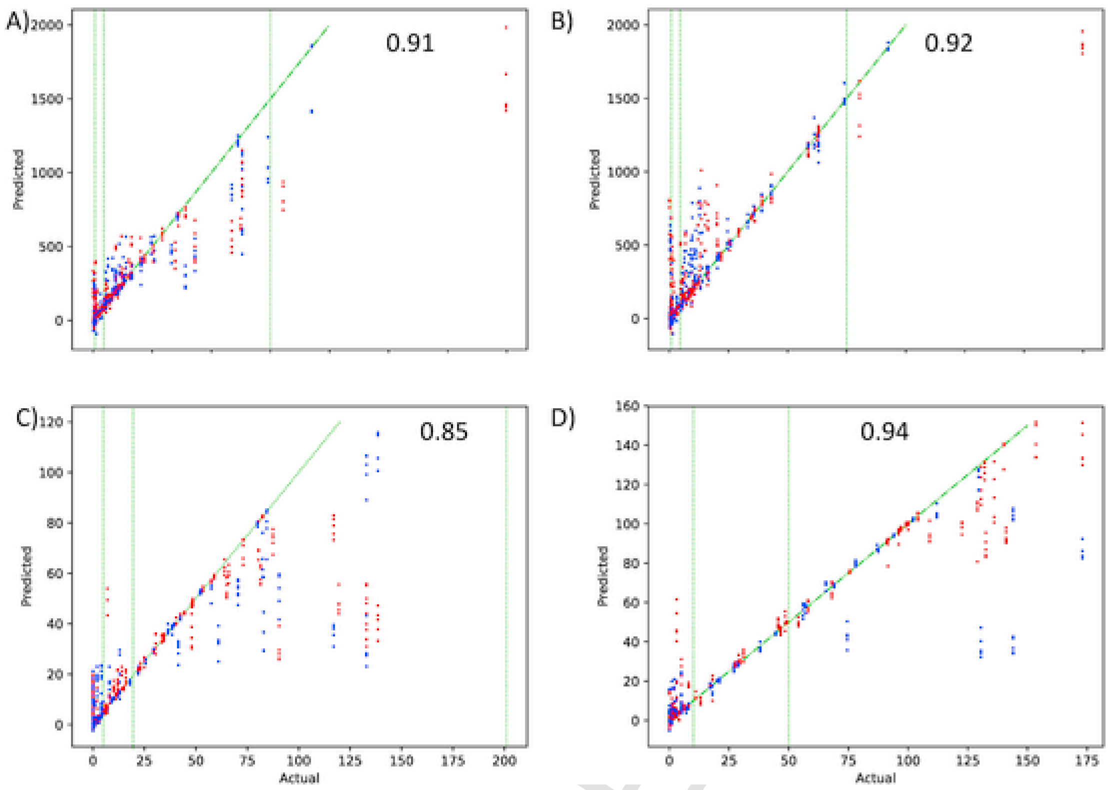

Fig. 6. Correlation of Actual and Predicted Pollen Counts. A, without sample weights for Trees B, with sample weights for Trees. C, Grass; D, Weed. The correlation coefficient of actual versus predicted is inset for leave-one-out cross validation. Dotted vertical green lines indicate boundaries between pollen count groups: L, M, H, VH. Dotted & Dashed diagonal green line indicates perfect agreement of the actual and predicted values. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

of either groups of multiple peaks or a single peak, respectively corresponding to various metabolites. Some spectra were more sensitive to disruption than others (for grass and weed, the resulting changes had to be plotted on a log scale to accommodate the wide range of values) – reflecting some instability of the trained models due to the low number of training samples. In spite of such instabilities, the spectra show consistent patterns of sensitivity to disruption, indicating the underlying peak components that are responsible for predictions. Given the low number of training samples and the observed instabilities of the trained models, it was difficult to derive detailed identification of the underlying metabolites, such as that illustrated in Fig. 2. However, with more example data and more robustly trained models this may be possible.

4. Discussion

This paper sought to demonstrate that the metabolite concentrations or spectral data from a mixture of pollens could be used to identify the component pollens. As a real-world example of pollen mixtures we utilized ambient air-sampled pollen from Munich primarily in 2016 (Fig. 1) and supplemented with data from 2017. NMR spectra of pollen extracts (Fig. 2) were examined for metabolites (Supplemental Fig. 1). Despite the general insensitivity of NMR, we were able to successfully measure the concentrations of 18 metabolites. When we looked for correlations between metabolites and the types of pollen in Fig. 3 there were no obvious correlations for an individual metabolite with the tree, grass, or weed pollen types. However, the ordering of importance for the metabolites was unique for each suggesting there might be trends in the multidimensional data that could be exploited in a predictive algorithm. This suggestion was reinforced by looking for patterns using t-SNE multidimensional decomposition methods (Fig. 4A–C). This more clearly showed that the metabolite data likely did contain tree, grass, or weed pollen specific data. The PCA in Fig. 4 D-F also showed a different type of input data, the NMR spectra with no analysis, also contained patterns specific for tree, grass, or weed pollen counts. This gave us confidence that predictive methods could be developed.

Therefore, we tested predictive algorithms using traditional machine learning with either metabolite concentrations or spectral input, and deep learning was tested using spectral input (Figs. 5 and 6, Table 1). Table 1 shows that regression and decision-tree based algorithms using the concentration of metabolites, calculated from the NMR spectra outperformed algorithms using the NMR spectra themselves as input data for pollen identification. This was somewhat surprising as we anticipated the additional peaks in the NMR spectra might contain valuable pollen type data. We suspected that the characteristics of the NMR data (redundancy, large size, large dynamic range) might be better suited for deep learning methods. Therefore, we attempted to train deep learning models using convolutional neural networks (CNN) similar to those that perform image recognition. Table 1 shows that the CNN performed better than regression models using NMR spectral input, and were the overall best method in terms of relative error and classification accuracy (86% for tree, 89% for grass, and 93% for weed pollen count). This is especially impressive considering that the CNNs did not require the calendar day as part of their input vector; they only used the NMR spectra.

This study is unique in that it utilizes a metabolomic or mixture analysis to identify pollens from an ambient air-sampled extract. NMR and Mass Spectrometry (MS) are commonly used for metabolomic and mixture analyses. The sensitivity of NMR is much less than MS, but in NMR there is a direct relationship between the signal intensities and the concentration of the metabolites. The direct proportionality makes this advantageous for using the peak intensities to measure the amount of sample (e.g. pollen) when this may be unknown. NMR analyses of other plant metabolites has been utilized to differentiate the geographic origin of hazelnut, coffee, and olive crops just to name a few examples (Dais and Hatzakis, 2013; Bachmann et al., 2018; Consonni and Cagliani, 2018). There have also been analyses of pollen metabolites by both NMR and MS (Mueller et al., 2016; Gilles et al., 2011). From our previous NMR analysis, we were motivated to see if we could use pollen metabolite information to differentiate pollen.

Looking into then future, development of an automated prediction procedure should benefit in particular from the successful application of deep learning methods, which require very little data processing and can be applied directly to NMR spectra, and potentially mass spectrometry data, without the need for manual feature extraction and curation. We achieved quite satisfactory prediction accuracies using a very small set of 147 samples and it should be possible to improve upon these accuracies using a more comprehensive training set – resulting in a very robust predictor. If we imagine a potential future workflow, air-sampled pollen could be extracted directly into NMR amenable buffers instead of being lyophilized and rehydrated. Previously we found excellent extraction of metabolites in only 30 min. Then filtration of the sample through a 3 kDa filter to remove proteins can take up to an hour in a spin-concentrator. The NMR spectra acquired here took 90 min and the computer processing of the data could be handled automatically. In terms of cost, an hour of NMR time can be $30-$60. We imagine this could be run in parallel with other existing analyses of pollen, or pollen metabolites. In summary, it should be possible to develop an efficient and automated procedure that capitalizes on these findings. Alternatively, other methodologies that measure metabolites like mass spectrometry may be similarly successful.

As a cautionary note, there is the strong possibility that, while an automated pipeline would likely have widespread geographic applicability, the models trained on pollen sampled in Munich may not be as applicable across the globe. For example, in regions of the U.S. ragweed pollen dominates as opposed to *Urticaceae* in Munich. And given the numerous papers on how the metabolites in other plants are sensitive to geography, the Munich pollen results may not be applicable to pollen sampled as close as Vienna. There will likely be a need for local optimization using studies like those conducted here to calibrate the metabolites to the sampling region.

5. Conclusion

This paper demonstrates the proof of principle that chemical analyses of the pollen metabolome may be useful in differentiating pollens, and that it can be done in an automated fashion using various machine learning and deep learning algorithms.

CRediT authorship contribution statement

Leszek J. Klimczak: Formal analysis, Investigation, Methodology, Data curation, Writing - original draft. **Cordula Ebner von Eschenbach:** Formal analysis, Investigation, Methodology, Writing - original draft. **Peter M. Thompson:** Investigation, Methodology, Data curation, Writing - original draft. **Jeroen T.M. Buters:** Investigation, Data curation, Writing - original draft. **Geoffrey A. Mueller:** Conceptualization, Formal analysis, Investigation, Methodology, Data curation, Writing - original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported in part by the Intramural Research Program of the National Institute of Environmental Health Sciences, (Z01-ES102906, GAM). This work was performed in part by the Molecular Education, Technology and Research Innovation Center (METRIC) at NC State University, which is supported by the State of North Carolina.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.atmosenv.2020.117746.

References

Bachmann, R, Klockmann, S, Haerdter, J, Fischer, M, Hackl, T, 2018. (1)H NMR spectroscopy for determination of the geographical origin of hazelnuts. J. Agric. Food Chem. 66, 11873–11879.

Bagcioglu, M, Zimmermann, B, Kohler, A, 2015. A multiscale vibrational spectroscopic approach for identification and biochemical characterization of pollen. PloS One 10, e0137899.

Brennan, G L, Potter, C, De Vere, N, Griffith, G W, Skjoth, C A, Osborne, N J, Wheeler, B W, Mcinnes, R N, Clewlow, Y, Barber, A, Hanlon, H M, Hegarty, M, Jones, L, Kurganskiy, A, Rowney, F M, Armitage, C, Adams-Groom, B, Ford, C R, Petch, G M, Poller, G E N C, Creer, S, 2019. Temperate airborne grass pollen defined by spatio-temporal shifts in community composition. Nat. Ecol. Evol. 3, 750–754.

Burbach, G J, Heinzerling, L M, Edenharter, G, Bachert, C, Bindslev-Jensen, C, Bonini, S, Bousquet, J, Bousquet-Rouanet, L, Bousquet, P J, Bresciani, M, Bruno, A, Canonica, G W, Darsow, U, Demoly, P, Durham, S, Fokkens, W J, Giavi, S, Gjomarkaj, M, Gramiccioni, C, Haahtela, T, Kowalski, M L, Magyar, P, Murakozi, G, Orosz, M, Papadopoulos, N G, Rohnelt, C, Stingl, G, Todo-Bom, A, Von Mutius, E, Wiesner, A, Wohrl, S, Zuberbier, T, 2009. GA(2)LEN skin test study II: clinical relevance of inhalant allergen sensitizations in Europe. Allergy 64, 1507–1515.

Buters, J, Prank, M, Sofiev, M, Pusch, G, Albertini, R, Annesi-Maesano, I, Antunes, C, Behrendt, H, Berger, U, Brandao, R, Celenk, S, Galan, C, Grewling, L, Jackowiak, B, Kennedy, R, Rantio-Lehtimaki, A, Reese, G, Sauliene, I, Smith, M, Thibaudon, M, Weber, B, Cecchi, L, 2015. Variation of the group 5 grass pollen allergen content of airborne pollen in relation to geographic location and time in season. J. Allergy Clin. Immunol. 136, 87–95 e6.

Buters, J, Schmidt-Weber, C, Oteros, J, 2018. Next-generation pollen monitoring and dissemination. Allergy 73, 1944–1945.

Buters, J T M, Thibaudon, M, Smith, M, Kennedy, R, Rantio-Lehtimaki, A, Albertini, R, Reese, G, Weber, B, Galan, C, Brandao, R, Antunes, C M, Jager, S, Berger, U, Celenk, S, Grewling, L, Jackowiak, B, Sauliene, I, Weichenmeier, I, Pusch, G, Sarioglu, H, Ueffing, M, Behrendt, H, Prank, M, Sofiev, M, Cecchi, L, Grp, H W, 2012. Release of Bet v 1 from birch pollen from 5 European countries. Results from the HIALINE study. Atmos. Environ. 55, 496–505.

Consonni, R, Cagliani, L R, 2018. The potentiality of NMR-based metabolomics in food science and food authentication assessment. Magn. Reson. Chem..

D'amato, G, Cecchi, L, Bonini, S, Nunes, C, Annesi-Maesano, I, Behrendt, H, Liccardi, G, Popov, T, Van Cauwenberge, P, 2007. Allergenic pollen and pollen allergy in Europe. Allergy 62, 976–990.

Dais, P, Hatzakis, E, 2013. Quality assessment and authentication of virgin olive oil by NMR spectroscopy: a critical review. Anal. Chim. Acta 765, 1–27.

Galan, C, Antunes, C, Brandao, R, Torres, C, Garcia-Mozo, H, Caeiro, E, Ferro, R, Prank, M, Sofiev, M, Albertini, R, Berger, U, Cecchi, L, Celenk, S, Grewling, L, Jackowiak, B, Jager, S, Kennedy, R, Rantio-Lehtimaki, A, Reese, G, Sauliene, I, Smith, M, Thibaudon, M, Weber, B, Weichenmeier, I, Pusch, G, Buters, J T, Group, H W, 2013. Airborne olive pollen counts are not representative of exposure to the major olive allergen Ole e 1. Allergy 68, 809–812.

Galan, C, Smith, M, Thibaudon, M, Frenguelli, G, Oteros, J, Gehrig, R, Berger, U, Clot, B, Brandao, R, Group, E Q W, 2014. Pollen monitoring: minimum requirements and reproducibility of analysis. Aerobiologia 30, 385–395.

Gilles, S, Fekete, A, Zhang, X, Beck, I, Blume, C, Ring, J, Schmidt-Weber, C, Behrendt, H, Schmitt-Kopplin, P, Traidl-Hoffmann, C, 2011. Pollen metabolome analysis reveals adenosine as a major regulator of dendritic cell-primed T(H) cell responses. J. Allergy Clin. Immunol. 127, 454–461 e1-9.

Holt, K, Allen, G, Hodgson, R, Marsland, S, Flenley, J, 2011. Progress towards an automated trainable pollen location and classifier system for use in the palynology laboratory. Rev. Palaeobot. Palynol. 167, 175–183.

Kraaijeveld, K, De Weger, L A, Garcia, M V, Buermans, H, Frank, J, Hiemstra, P S, Den Dunnen, J T, 2015. Efficient and sensitive identification and quantification of airborne pollen using next-generation DNA sequencing. Mol. Ecol. Resour. 15, 8–16.

Marcos, J V, Nava, R, Cristobal, G, Redondo, R, Escalante-Ramirez, B, Bueno, G, Deniz, O, Gonzalez-Porto, A, Pardo, C, Chung, F, Rodriguez, T, 2015. Automated pollen identification using microscopic imaging and texture analysis. Micron 68, 36–46.

Mueller, G A, Thompson, P M, Derose, E F, O'connell, T M, London, R E, 2016. A metabolomic, geographic, and seasonal analysis of the contribution of pollen-derived adenosine to allergic sensitization. Metabolomics 12.

Oteros, J, Pusch, G, Weichenmeier, I, Heimann, U, Moller, R, Roseler, S, Traidl-Hoffmann, C, Schmidt-Weber, C, Buters, J T, 2015. Automatic and online pollen monitoring. Int. Arch. Allergy Immunol. 167, 158–166.

Platts-Mills, T A, 2004. Allergen avoidance. J. Allergy Clin. Immunol. 113, 388–391.

Salo, P M, Arbes, S J, Jaramillo, R, Calatroni, A, Weir, C H, Sever, M L, Hoppin, J A, Rose, K M, Liu, A H, Gergen, P J, Mitchell, H E, Zeldin, D C, 2014. Prevalence of allergic sensitization in the United States: results from the National Health and Nutrition Examination Survey (NHANES) 2005-2006. J. Allergy Clin. Immunol. 134, 350–359.

Seifert, S, Merk, V, Kneipp, J, 2016. Identification of aqueous pollen extracts using surface enhanced Raman scattering (SERS) and pattern recognition methods. J. Biophot. 9, 181–189.

Sevillano, V, Aznarte, J L, 2018. Improving classification of pollen grain images of the POLEN23E dataset through three different applications of deep learning convolutional neural networks. PloS One 13, e0201807.

Valencia, J A, Astray, G, Fernandez-Gonzalez, M, Aira, M J, Rodriguez-Rajo, F J, 2019. Assessment of neural networks and time series analysis to forecast airborne Parietaria pollen presence in the Atlantic coastal regions. Int. J. Biometeorol..

Van Der Maaten, L, Hinton, G, 2008. Visualizing Data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605.

Verein-Deutsche-Ingenieure, 2017. Ermittlung von Pollen und Sporen in der Außenluft unter Verwendung einer volumetrischen Methode für ein Messnetz zu allergologischen Zwecken. Kommission Reinhaltung der Luft im VDI und DIN – Normenausschuss KRdl VDI 4252.

Zimmermann, B, 2010. Characterization of pollen by vibrational spectroscopy. Appl. Spectrosc. 64, 1364–1373.