Human breath contains hundreds of volatile molecules that can provide powerful, non-intrusive spectral diagnosis of a diverse set of diseases and physiological/metabolic states. To unleash this tremendous potential for medical science, we present a robust analytical method that simultaneously measures tens of thousands of spectral features in each breath sample, followed by efficient and detail-specific multivariate data analysis for unambiguous binary medical response classification. We combine mid-infrared cavity-enhanced direct frequency comb spectroscopy (CE-DFCS), capable of real-time collection of tens of thousands of distinct molecular features at parts-per-trillion sensitivity, with supervised machine learning, capable of analysis and verification of extremely high-dimensional input data channels. Here, we present the first application
of this method to the breath detection of Coronavirus Disease 2019 (COVID-19). Using 170 individual samples at the University of Colorado, we report a cross-validated area under the Receiver-Operating-Characteristics curve of 0.849(4), providing excellent prediction performance. Further, this method detected a significant difference between male and female breath as well as other variables such as smoking and abdominal pain. Together, these highlight the utility of CE-DFCS for rapid, non-invasive detection of diverse biological conditions and disease states. The unique properties of frequency comb spectroscopy thus help establish precise digital spectral fingerprints for building accurate databases and provide means for simultaneous multi-response classifications. The predictive power can be further enhanced with readily scalable comb spectral coverage.

Main Text

Introduction

Breath analysis examines the hundreds or thousands of low-concentration molecular species present in exhaled human breath and correlates them with particular health conditions or diseases. Some examples include formaldehyde being correlated with breast cancer, ammonia with asthma, and methane with intestinal problems\textsuperscript{1}. This research field is gaining considerable interest due to its potential for non-invasive, low-cost, and real-time diagnosis. Unleashing the full potential of breath analysis requires the identification of a panel of biomarkers that are associated with a certain medical condition, and how their concentrations co-vary in the presence or absence of a particular condition. Two major roadblocks hindering the adoption of breath analysis to medical diagnosis have been the difficulties in i) reaching the necessary capability to identify breath biomarkers with superior detection specificity and sensitivity, and ii) data analysis to correlate the measurement results with certain medical conditions for unambiguous and reliable predictions.

Cavity-enhanced direct frequency comb spectroscopy (CE-DFCS) offers great advantages for
detecting molecules. It combines ultra-sensitive and high-resolution detection, broadband spectral coverage, and high data acquisition efficiency all in one platform, while also providing both isomer- and isotopologue-specific detection capabilities. The simplicity of in-situ optical absorption together with massively-parallel signal collection provide direct and objective measurements of biomarkers, ideal for building data bases to facilitate cross-field studies.

The application of CE-DFCS to breath analysis was initially demonstrated in the near-infrared spectral region. In 2021, a more than 2-orders-of-magnitude improvement in sensitivity was achieved by extending this technology to the mid-infrared molecular fingerprint region, where strong fundamental vibrational transitions are accessed and detection sensitivities as high as parts-per-trillion ($10^{-12}$) were achieved. We have thus finally reached the full sensitivity requirement for breath analysis where most trace species are present at the $10^{-9}$ to $10^{-12}$ levels. However, the massive density of spectral features carrying richer-than-ever, molecular-specific information poses a serious complexity in data processing to establish robust predictive modeling for medical diagnosis. We overcome this complexity using supervised machine learning dedicated to extremely high-dimensional data analysis and verification. The high-quality comb data has allowed the use of a pattern-based analysis method to establish recognizable digital spectral fingerprints for breath-based disease diagnosis. From these digital fingerprints we can also identify a panel of responsible molecules. This study constitutes the first real-world examination of CE-DFCS’s capability for medical diagnostics.

We target the detection of Coronavirus Disease 2019 (COVID-19), a highly infectious disease with airborne transmission via aerosol and droplets that has tragically caused many deaths globally. In some exploratory COVID-19 breath testing efforts, studies have been conducted with small sample sizes of research subjects of different age groups (mainly above 40) and infection severity using nanomaterial-based sensors, ion-mobility spectrometry, and mass-spectrometry. We have performed breath analysis of COVID infected individuals in a unique manner. We target a group of young people with a high vaccination rate. We focus on collecting highest quality molecular data from breath and performing rigorous cross-validation and baseline confirmation.
From May 2021 to January 2022, we collected breath samples from a total of 170 research subjects with a class distribution for COVID-19 infection of 83 positives (49%) and 87 negatives (51%), as identified by RT-PCR. Ninety-six percent of the research subjects were born after 1980, all above 18 years old and affiliated with the University of Colorado Boulder as a student and/or a university employee. Fifty-nine percent reported 5 days or more of school/work commuting needs per week. The general population on campus is >90% vaccinated. The recruited subjects thereby constitute a young and high transmission potential study group.

Inspired by the DOME recommendations calling for a standard to enable improved cross-study diagnostic comparisons\textsuperscript{14}, we analyze this large sample group and use 10,000 cross-validation runs based on stratified random sampling to obtain converging indicators required for medical diagnostic studies. Excellent cross-validated prediction performance for COVID-19 is achieved with an area under the Receiver-Operating-Characteristic (ROC) curve of 0.849(4). Interestingly, this comb-machine learning breath analysis also identifies significant differences based on a number of variables collected in tandem for this study including male vs. female and smoker vs. nonsmoker. The baseline probability is checked by comparing the COVID-19 and gender results against a binary response classification for parity (even/odd) of birth months.

**Data analysis**

The working principle of the mid-infrared comb breathalyzer for COVID-19 classification is illustrated in Fig. 1a. A research subject provides a breath sample in a standard Tedlar bag. The sample is loaded and analyzed by the breathalyzer to generate high-resolution broadband spectroscopic data that resolves rotational and vibrational molecular absorption features (see sample spectrum in Fig. 1b). The massive amount of data collected, consisting of absorption signals measured by a total of 14,836 frequency comb lines uniformly sampled in a spectral window ranging from 2810 cm\(^{-1}\) to 2945 cm\(^{-1}\), with a frequency uncertainty of \(\sim 0.0000017\) cm\(^{-1}\), is then processed by a machine learning algorithm to predict the COVID-19 infection status (either positive or negative) of the research subject.
The learning algorithm is constructed from a set of observations, each representing the spectroscopic data collected from a single research subject combined with its true class label, also named response. The spectroscopic data used for prediction purposes are called predictor variables. We utilize a Molecular Science (MS) approach, which uses the spectra of 16 known compounds as predictor variables, and a Data Science (DS) approach, which uses all 14,836 frequency comb lines directly as the predictor variables (see Methods). We use the Partial Least-Squares Discriminant Analysis (PLS-DA) algorithm \textsuperscript{15,16} to discriminate between opposing response classes. Linear combinations of the predictor variables are optimized based on maximizing their covariance with the response. For model assessment, a subset of observations is randomly selected and isolated from the algorithm training process and used only for testing the prediction performance (see Fig. 2). To generate the prediction results, the predictor variables for an observation in the testing set are regressed to yield a probability value for grouping into the opposing classes.

We evaluate the prediction performance by using ROC curves \textsuperscript{17,18,19} (see Methods). The entire model assessment, called cross-validation, is repeated 10,000 times to generate an averaged ROC curve obtained from all individual runs. In each run, stratified random sampling is used for the selection of a new training set and testing set so as to preserve the class distribution in the complete data set. The predictive power is quantified by the area under the curve (AUC) of the averaged ROC curve, along with the prediction sensitivity (true positive rate) and specificity (true negative rate) quoted for the data point on the averaged ROC curve giving the highest pair average of the two (see Methods for more details). We investigate binary classification for COVID-19 infection (positive vs. negative) and compare it with a few other response type classifications including birth month (odd vs. even), gender (female vs. male), smoking status, and abdominal pain.

**Results and discussion**

We present a number of important findings. First, the DS approach outperforms the MS approach for classification of COVID-19 infection. We show distributions of the subjects’ data projected on the first three PLS components in Fig. 3. Here, the complete data set (N = 170) is used for
constructing the PLS components and all observations are plotted to provide a complete picture. Viewed from the first three PLS components, the DS approach (Fig. 3b) clearly demonstrates larger separability between the positive and the negative classes over the MS approach (Fig. 3a). The total percentage variance in the response explained by the first three PLS components is 30% for the MS approach and 76% for the DS approach (see Methods for details).

Second, we find the use of 15 PLS components sufficient to saturate the total percentage variance explained for both the MS and the DS approaches, with the values obtained at 31.222% and 99.996%, respectively. We thus use the 15 PLS components to evaluate the relative importance of the predictor variables through the Variable Importance in the Projection (VIP) score\(^{20}\).

Third, based on the VIP score, MS and DS approaches identify their respective important predictor variables. Figures 3c and 3d give their VIP scores. For MS, formaldehyde (\(\text{H}_2\text{CO}\)) is identified as the most important predictor variable for detecting COVID-19 infection owing to its greatest explained variance for the response. In contrast, important predictor variables for the DS approach are distributed nearly uniformly across the entire spectrum, showing no obvious weighting in this spectral region. This broad distribution of discriminating signals across the spectrum illustrates the power of CE-DFCS for distinguishing biological conditions, with future improvements expected from expansion of the spectral range.

Fourth, our method does not generate over-optimistic prediction results. The classification result on birth month (Fig. 4a) examines and tries to predict whether the research subjects were born in the odd (January etc.) or even (February etc.) months. Of course, one would not expect such response type classification to yield a correlated result based on the exhaled human breath. Indeed, we find this a rigorous check to ensure the algorithm does not report unwarranted results. We find the AUC, sensitivity, and specificity to be 0.516(4), 55% and 48% respectively based on the DS approach. For MS, the AUC is 0.487(9). These results confirm that both classifiers are indeed random guessers when relying on the CE-DFCS breath data to distinguish the even/odd parity of birth month.

Strikingly, in a completely unexpected result, our approach is able to distinguish between breath
of men and women. Specifically, we find the DS-based AUC of 0.667(12), sensitivity of 52 %, and specificity of 72 % (Fig. 4b), revealing a significant gender-based difference in exhaled breath. For MS, the AUC is a comparable value of 0.638(9). The gender-based difference may increase when we probe a wider spectral range. Similar analyses on other personal attributes such as smoking and abdominal pain has also revealed significant differences in breath.

Most importantly, our spectra are able to distinguish COVID-19 infected and uninfected individuals. The classification results for COVID-19 (Fig. 4c) give an AUC of 0.849(4), sensitivity of 76 %, and specificity of 82 %. A classifier with an AUC >0.8 is considered to have excellent discriminating capability\textsuperscript{17}. This demonstrates the power of CE-DFCS for breath analysis for COVID-19 diagnosis. Since RT-PCR tests for COVID-19 are not 100 % accurate, it is possible that our diagnostic results are even better than reported. It will be very valuable to conduct future large-scale tests that pair CE-DFCS with other standard COVID-19 testing techniques for daily screening.

For COVID-19 diagnosis, the DS approach clearly outperforms the MS approach, the latter of which produces an AUC of 0.769(7). To understand this result, we note that the CE-DFCS technology is sensitive to the presence of both i) small molecules with their rotational and vibrational structure resolved by absorption signals, and ii) large molecules with unresolved spectroscopic features. The presence of the large molecules will introduce frequency-dependent optical losses inside the cavity, which are encoded in the measured spectroscopic data through alteration in absorption patterns of the resolved molecules. Simple fitting of the spectroscopy features to a set of known molecules can neither guarantee the accuracy of the fitted concentrations nor confirm the exhaustion of all relevant molecules. For breath analysis, this additional sensitivity is actually very valuable as it reveals the presence of more molecular species. The DS approach senses all the variations in the spectroscopic signals related to the presence of unresolved molecules, providing detection sensitivity to both large and small molecules. Thus, in contrast to the MS approach, the variability in the CE-DFCS data measured for different research subjects is entirely preserved by the DS approach. This suggests the superior prediction performance achieved by the DS approach in the COVID-19 diagnosis can be attributed to the presence of unresolved molecules and potentially even the coron-
avirus itself\textsuperscript{21}. This discussion also highlights a key reason for the powerful capability of CE-DFCS for breath analysis: its high detection sensitivity to both resolved and unresolved molecules. Both are important for the goal of providing highly accurate medical diagnosis based on exhaled breath contents.

It is also interesting to note that the classification performance of COVID-19 infection is substantially better than that for gender. This result is likely related to the fact that the former is distinguishing healthy from unhealthy subjects whereas the latter is trying to distinguish an innate biological difference and can thus be more difficult to classify. Comparing the prediction results for the three response types investigated, a natural progression can be seen. First, the random guess birth month check ensures that the model does not generate unwarranted prediction results. A statistically significant prediction of differences in gender reveals the unique capability of CE-DFCS plus learning model. Finally, a much more robust prediction for COVID-19 highlights the remarkable potential of CE-DFCS-based breath analysis paired with supervised learning for reliable medical diagnostics. It is exciting to anticipate future powerful applications of CE-DFCS for other disease state response classifications, by using exhaled breath contents to differentiate between the presence or absence of medical conditions in a non-invasive manner.

For our study group (N = 170), the between-type response covariances evaluated over the complete data set are negligible (see SI for details). The lack of correlations confirms that the classification of one of the three response types is not affected by the natural forces driving the separability for the other two response types. We cannot definitely rule out the possibility of other response types being potentially correlated with a specific response investigated in this work, for example non-COVID-19-related respiratory infections such as the common cold, influenza, or strep throat, which could affect the binary classification results. However, correlations with the response types not monitored in this study can in principle be averaged away statistically with more data collected, as is the case in all machine learning analysis, where a larger data set is always preferred.

It is important to stress the crucial capability of breath analysis for simultaneous multi-response classifications if sufficiently large group sizes are available, together with multiple non-correlated
response types. A massive database can be constructed. To demonstrate this, we present multi-
response analysis for a total of eleven personal attributes in the SI, which includes smoking, ab-
dominal pain, and constipation.

We emphasize the importance of baseline checks to prevent over-interpretation and using the
rigorous cross-validation protocol to ensure a converging AUC value with minimal uncertainty.
These permit accurate and unbiased comparisons of the diagnostic power between non-interfering
response types. There is an exponential growth of machine learning studies in the bio-medical
field and the community has called for standards to be developed to enable meaningful cross-study
assessment and comparisons. The community-proposed recommendations\textsuperscript{14} so far focus on only
supervised learning, although research in breath analysis has broadly employed both supervised
and unsupervised learning, along with many other published studies that provide only correlation
analysis with no diagnostic power assessment\textsuperscript{7}. Vast differences in the cross-validation protocols
and in the characteristics of the research subjects of various studies make it difficult to compare the
diagnostic power of different technologies. With that in mind, our experimental findings together
with other published studies strongly encourage the practical use of breath analysis for COVID-19
testing.

Future outlook

The diagnostic power of our comb breathalyzer can be substantially enhanced by expanding the
spectral coverage to detect more molecular species, based on the recent extension of frequency combs
towards longer wavelengths\textsuperscript{22,23,24,25}. Second, by miniaturization and simplification of the technol-
ogy, CE-DFCS can become transportable\textsuperscript{26,27} and deployed in point-of-care settings with minimal
operator training. Finally, alternative AI approaches that use neural networks\textsuperscript{28} or weighted anal-
yses with non-classifications might improve the predictive power of spectral analyses.

While the study here is limited to binary response classifications, we have also gathered inform-
ation from the research subjects such as the number of COVID-19 symptoms experienced. This
information can be used for designing a continuously varying test score for machine learning regres-
sion analysis as opposed to classification analysis where the response is discrete. In this way, we could assess whether the disease progression could be monitored by the exhaled breath contents. Another future endeavor is to correlate breath analysis data with viral loads quantified by RT-PCR to give more direct interpretation of the COVID-19 infection status and establish the sensitivity limit of CE-DFCS $^{21}$.

**Conclusion**

Supervised machine learning analysis is used to tackle the critical complexity of the extreme high-dimensional frequency comb breath data carrying record-breaking richness in molecular-specific information, allowing the first realization of robust medical diagnosis. With a sample size of 170, the use of 14,836 absorption features as the predictor variables produces an excellent cross-validated prediction performance, yielding AUC of 0.849(4) for COVID-19 classification. At the same time, the same analysis reveals significant breath differences on a number of other personal attributes. The availability of tens of thousands of molecular spectral features for rigorous machine learning and classification provides a powerful new medical capability that promises to surpass traditional techniques of breath analysis including the use of animals, and highlights methodologies for diagnosing a diverse set of disease states, including breast cancer, asthma, and intestinal problems$^4$. With an optical frequency comb serving as a massively parallel data generator, analyzer, and processor, we anticipate other exciting and important applications in the modern scientific era of big data with increased size and complexity$^{29,30}$. 
Figure 1: Comb-based breathalyzer for COVID-19 diagnosis. a, Schematic representation of the working principle of the device. An exhaled human breath sample is collected by a Tedlar bag and then loaded into a high-finesse optical cavity resonantly coupled with a mid-IR frequency comb laser. Each comb line performs an ultra-sensitive measurement of the molecular absorption signals at an isolated optical frequency. The broadband molecular absorption spectrum, containing more than ten thousand molecular spectral features, is then used for supervised machine learning to identify the binary response class for the research subject (either positive or negative). b, Sample absorption spectrum collected from a research subject’s exhaled breath (black). The spectrum is fitted with a total of sixteen molecular species. Inverted in sign and plotted with different colors are the four major species (CH$_3$OH, H$_2$O, HDO, and CH$_4$) that give the most dominant absorption features.
Figure 2: Cross-validation workflow. The complete data set (N = 170) is first divided into a training set and a testing set using stratified random sampling. The predictor variables and the response for each observation in the training set are used for building a classification model using the PLS-DA algorithm. After the model is built, the predictor variables for each subject in the testing set is fed to the machine to yield a predicted class (positive or negative), which is then compared against the true class for assessment of the prediction performance. Note that the machine is blind to the true class of the testing set before all predictions having been made.
Figure 3: Variables construction for the supervised machine learning. Two different predictor variables construction approaches are evaluated: the MS approach (predictor variables: concentrations of the 16 molecular species), and the DS approach (predictor variables: absorption signals measured by 14,836 comb lines). For both approaches, Partial Least-Squares (PLS) regression analysis of the response on the predictor variables leads to a set of PLS components generated from linear combinations of the predictor variables. The magnitudes and signs are determined based on the covariance of each predictor variable with the response. PLS regression allows dimensionality reduction of the prediction variables, maximized separability between the two opposing response classes, and evaluation of the relative variables importance in predictions. Results shown for the (a, c) MS and (b, d) DS approaches are calculated using the complete data set (N = 170) for the response type of COVID-19 (positive or negative). a, b, distribution of the subjects’ data for the first three PLS components, with red and blue triangles representing positive and negative research subjects, respectively. c, d, the VIP scores evaluated over a set of fifteen PLS components. The number of PLS components is chosen to ensure saturation of the total percentage variance explained in the response, which are 31.222% and 99.996% respectively for the MS and the DS approaches. Predictor variables with VIP scores above (or below) unity are plotted in purple (or black) and considered as important (or unimportant) for predictions.
Figure 4: Prediction performance. Cross-validation results analyzed by the DS approach using a set of 15 PLS components are presented in the form of the ROC curves. Shown are the results averaged over 10,000 cross-validation runs from repeated stratified random sampling at the fixed partition ratio of the training and testing set (140 vs. 30). AUC values are reported for the averaged ROC curves. See Methods for details on the averaging. Prediction results using different numbers of PLS components and different partition ratios of the training and testing set are presented in the SI, from which the uncertainties of the AUC values are determined. Three different binary response types are examined: (a) birth month (positive: odd; negative: even), (b) gender (positive: female; negative: male), and (c) COVID-19 (positive: infected; negative: not infected). For birth month and gender, the respective assignment of the response classes to positive and negative is done at random and does not carry any particular meaning. The class distributions for the three response types are 83(49 %):87(51 %), 87(51 %):83(49 %), and 83(49 %):87(51 %), respectively.
Methods

Subject recruitment. The study was approved by the Institutional Review Board (protocol no. 21-0088) of the University of Colorado. Research subjects are all affiliates of the University of Colorado Boulder, at least 18 years old, recruited after taking a saliva-based or nasal swab COVID-19 RT-PCR test provided by the University. After receiving their COVID-19 test results, potential subjects received a recruitment email for this study and were asked to contact the research team within 24 hours if interested in participating. They then reviewed and signed the informed consent form, completed a questionnaire, and scheduled an appointment with the research team members to collect their breath samples. For the COVID-19-positive subjects, the average time delay between the completion of the COVID-19 test and the collection of the breath samples was 2.05 ± 0.95 days (error denotes one standard deviation), where the disease is likely to progress over this period of time. The information gathered through the questionnaire includes date of birth, gender, and commuting needs. Other gathered information includes COVID-19 symptoms (for COVID-19 positive subjects only), abdominal symptoms, commuting behaviors, and typical alcohol consumption and smoking habits. Analysis with respect to these additional attributes is not discussed here and will be presented in a future publication. A total of 170 research subjects were recruited from May 2021 to January 2022. Forty-eight percent of the subjects were born after 1990, and 96% born after 1980. We note that all information gathered through the questionnaire is self-identified. The question about gender in the questionnaire gives three options, “male”, “female” and “other”. Two research subjects selected two options simultaneously, with “other” alongside either “male” or “female”. Since there are no research subjects who selected both “male” and “female” simultaneously, the classification analysis on gender performed in this work uses the subject’s selection for “male” or “female” as the binary true class labels. This results in a gender distribution of 51% “female” and 49% “male”. All data (i.e. informed consent form, questionnaire, and Tedlar bag ID) are collected and managed using the REDCap electronic data capture tool31,32 hosted by the University of Colorado Denver.
**Sample collection and analysis.** Standard Tedlar bags (1-liter, part no. 249-01-PP, SKC Inc.) were used to collect exhaled breath. During the appointment for sample collection, research subjects were asked to hold their noses and breathe in/out through the mouth. They were instructed to inhale to full lung capacity for 1-3 s, followed by exhaling the first half of their breath to the surroundings and the second half into the bag until the latter was above ~80% full. The location for the sample collection was selected to be an outdoor parking lot within the university campus. The participants were not instructed or required to limit or control their smoking, food or alcohol intake prior to the sample collection. Immediately after the sample collection, the breath sample was transported via an air-tight container to the indoor lab where the comb breathalyzer was set up. Here, each sample was heated at a temperature of ~37°C for 20 minutes to mimic body temperature and avoid molecules condensing onto the bag. It was then loaded into a cleaned vacuum chamber held at room temperature (20°C) by steadily flowing the sample gas through the chamber at ~1 liter per minute. When the sample was about to exhaust, the input gas valve to the chamber was closed and the intracavity chamber pressure finely controlled by timely closure of the output valve to reach a static pressure of 50 Torr (67 mbar). The entire gas flowing process took about 1 minute. After collection, the breath sample was pumped out while the Tedlar bag was autoclaved and disposed. The same collection and analysis protocol was used for all samples in an effort to minimize inter-person systematic variations.

**Cavity-enhanced direct frequency comb spectroscopy.** The principle of CE-DFCS is detailed in Ref. 2 and the characterization of the breathalyzer used in this study is reported in Ref. 4. In brief, the mid-infrared frequency comb source used in this study is generated by a singly-resonant optical parametric oscillator, with a repetition rate of 136 MHz and power of ~100 mW used. The Pound-Drever-Hall technique is used to resonantly lock the mid-IR comb light to an optical cavity, which is formed by a pair of high reflectivity mirrors with finesse ranging from 6,000 to 8,000 depending on the wavelength. The cavity transmission light is analyzed with a home-built Fourier-transform infrared spectrometer sampled at a frequency interval that is matched to half the transmitted comb spacing. By utilizing the known instrument function (a sinc
function), the absorption of each comb light is extracted\textsuperscript{33}. Cavity mirror dispersion restricts the spectral bandwidth of the incident comb light coupled into the cavity. Thus, spectra collected at different center wavelengths are concatenated to form a complete spectrum.

**Supervised Machine Learning.** A classification machine is built from learning the correlation of the predictor variables with the response and then constructing a decision metric for prediction of the class labels of new data based on the predictor variables alone. Feeding knowledge of the true class labels to the machine makes the learning “supervised”, as opposed to “unsupervised” where only the predictor variables are fed into it. The Molecular Science (MS) approach and the Data Science (DS) approach are two different predictor variables construction methods employed and compared in this work. The MS approach aims at identifying the panel of molecular species that can be used to discriminate between opposing classes. Absorption spectra are fitted to the High-Resolution Transmission Molecular Absorption (HITRAN) database\textsuperscript{34}, out of which we choose a total of 16 molecular species with cross sectional data available in the probed spectral region. A total of 16 predictor variables for each observation are then formed by concentrations of the 16 molecular species. On the other hand, the DS approach focuses on carrying out only the necessary analysis steps based on the raw spectroscopy data to enable the construction of a prediction model. Absorption signals measured by a total of 14,836 frequency comb lines are directly used as the predictor variables, without the knowledge of the exact molecular species present. In this work, the Partial Least-Squares Discriminant Analysis (PLS-DA) algorithm\textsuperscript{15,16} is adopted based on its good capability for high-dimensional data modeling. We provide the PLS-DA basics in a separate section in the Methods. A model assessment run begins by dividing the complete data set into a training set used for building the model, and a non-overlapping complement set called the testing set and used for the model evaluation. Each predictor variable in the training set is normalized to give a standard deviation of unity. All observations from the training set are used for construction of a set of PLS components. The linear combination coefficients used for building the set of PLS components determines a set of regression coefficients relating the predictor variables and the response for observations in the training set. These
regression coefficients are used for predicting the class labels for new data. To generate the prediction results, the predictor variables for an observation in the testing set are each normalized by the standard deviation calculated for the same predictor variable in the training set, multiplied by a regression coefficient, then summed together to yield a numerical predicted value. We then translate such numerical value into posterior probabilities for the observation to be grouped into the opposing classes, based on how it proportionally compares to the values used for representing the two true classes in the training process. If the predicted value is outside of the range specified by the true classes values, the observation is assigned with 100% posterior probability to be associated with one of the two classes. The posterior probabilities are then compared with a decision threshold value for predicted class assignments, which can later be checked against the true class labels for model evaluation.

**Principle of PLS-DA.** The principle of PLS regression and its usage for discriminant analysis, namely the PLS-DA algorithm, is briefly introduced here. The PLS regression toolbox used in our work is developed by Matlab and implemented using the SIMPLS formulation. We discuss only the univariate response classification, corresponding to what is used in this work, but interested readers may consult Ref. 15 for more details beyond this classification type and how the actual algorithm is implemented. We use bold upper case, bold lower case, and un-bold letters to denote matrices, vectors, and scalars, respectively. Matrix transpose are denoted by primes (\(^\prime\)). Collected data used for the training process are represented by the \(n \times p\) predictor variables matrix \(X_0\) and the \(n \times 1\) univariate response variable vector \(y_0\). Here, \(n\) is the total number of research subjects, \(p\) is the total number of predictor variables. Both \(X_0\) and \(y_0\) have been column-centered so that the covariance of different predictor variables with the response can be expressed by a \(p \times 1\) column vector \(s_0 = X_0^\prime y_0\). PLS regression relates \(X_0\) and \(y_0\) based on \(y_0 = X_0b + e\), where \(b\) is the \(p \times 1\) coefficients estimate, \(X_0b\) is the explained component, \(e\) is the fit residual. In contrast to least squares regression, where the coefficients estimate \(b\) is constructed by minimizing the residual sum of squares \(e'e\), PLS regression constructs it based on the covariance \(s_0 = X_0^\prime y_0\) to get more stabilized values of \(b\) and achieve more reliable predictive power. The formulation begins by
projecting the predictor variables matrix $X_0$ onto a new coordinate system $T = X_0R$ of reduced dimensionality spanned by a total of $A$ (≤ $p$ − 1) PLS components, where $R$ denotes the $p \times A$ weight transfer matrix and $T$ denotes the $n \times A$ projected scores matrix. The construction of $R$ is subject to two constraints: 1) the covariance vector $T'y_0$ is maximized for each entry, meaning each PLS component exhibits the largest possible covariance with the response; 2) the PLS components are orthonormal, i.e., columns of $T$ satisfy $t'_k t_j = \delta_{ij}$ for any $i, j = 1, 2, ..., A$, where $\delta_{ij}$ is the Kronecker delta. The coefficients estimate $b$ can be determined once $R$ is known, since $y_0 = TT'y_0 = X_0RR'X_0'y_0 = X_0b$, and thus $b = RR'X_0'y_0 = RR's_0$. The process of determining $R$ proceeds column by column. For the first iteration step $k = 1$, the maximization of the covariance of the first PLS component ($t_k = X_0r_k$) with the response, $t'_k y_0 = r'_k X_0'y_0 = r'_k s_0 = \text{max}$, constrains the first weight vector $r_k$ ($k = 1$) to be along the direction of $s_0$. For steps $k > 1$, the orthogonality condition, $t'_k t_i = r'_k (X_0't_i) = 0$ for $i = 1, 2, ..., k - 1$, requires the newly constructed $r_k$ to be orthogonal to each of the $p \times 1$ vectors $X_0't_i$ for $i = 1, 2, ..., k - 1$. We define $p_i \equiv X_0't_i$ called the loading vectors. One may use the Gram-Schmidt process to find the orthonormal basis of the subspace $V_{k-1}$ spanned by the loading vectors $p_i$ ($i = 1, 2, ..., k - 1$) and then determine the $p \times p$ projection operator $P^\perp$ for the orthogonal complement space $V_{k-1}^\perp$. This loosely constrains the direction of $r_k$ to be within $V_{k-1}^\perp$, requiring $r_k = P^\perp r_k$. Now, with the covariance maximization criteria, $t'_k y_0 = r'_k (P^\perp's_0) = \text{max}$, the direction of $r_k$ is ultimately determined to be along the direction of the vector $P^\perp's_0$, which is the projection of the covariance vector $s_0$ onto the subspace $V_{k-1}^\perp$. The iteration process proceeds until the directions of all $r_k$ is determined, where the normalization condition $T'T = 1$ governs the magnitudes of $r_k$. Finally, the coefficients estimate is determined and can be used for prediction of the response class for new observation based on $y_0^{pred} = X_0^{new}b$, where the $m \times p$ matrix $X_0^{new}$ is the testing data for a total of $m$ research subjects. The $m \times 1$ predicted values $y_0^{pred}$ are translated proportionally into posterior probabilities and compared with a threshold value for response class assignment.

**Variable importance in the projection scores.** In PLS-DA, assessment of the importance of
the predictor variables needs to take into account 1) the weighting of a given predictor variable to form different PLS components and 2) the importance of different PLS components in explaining the response. Regarding 1), the formation of the $a$-th PLS component ($a = 1, 2, ..., A$) takes the contribution from the $j$-th predictor variable with the normalized weight given by $w_{ja}/\|w_a\|$, where $w_{ja}$ is the $j$-th row $a$-th column element from the $p \times A$ weight matrix $R$, and $\|w_a\| = (\sum_{j=1}^{p} w_{ja}^2)^{1/2}$ is the normalization. Regarding 2), we first note that the variance of the response among all observations $y_0'y_0$ is explained by the total of $A$ PLS components to the extent of $\hat{y}_0'y_0$, where $\hat{y}_0 = X_0b = y_0 - e$. The total percentage variance explained in the response, $(\hat{y}_0'y_0/y_0'y_0) \times 100\%$, can be used for estimating the minimum number of PLS components needed for reliable predictions (see SI for details). The explained variance $\hat{y}_0'y_0 = \hat{y}_0'\hat{y}_0 = \sum_{a=1}^{A}(\hat{y}_0't_a)^2$ is further broken down into a summation of the square of the covariance of all PLS components with $\hat{y}_0$. We can thus evaluate the importance of the $a$-th PLS component by its variance explained $q_a^2 \equiv (\hat{y}_0't_a)^2$, a quantity assigning larger importance to the PLS components that have larger covariance with the explained component, with the total variance explained by the $A$ PLS components given by $\sum_{a=1}^{A} q_a^2$. Taking both 1) and 2) into account, the variable importance for the predictor variable $j$ summing over all the $A$ PLS components is proportional to $[\sum_{a=1}^{A} q_a^2 \cdot (w_{ja}/\|w_a\|)^2]^{1/2}$. From this, we define its VIP score $^{20}$, a metric for characterizing its importance, by:

$$\text{VIP}_j = \sqrt{\frac{p \cdot \sum_{a=1}^{A} [q_a^2 \cdot (w_{ja}/\|w_a\|)^2]}{\sum_{a=1}^{A} q_a^2}} \quad (1)$$

The normalization is taken to ensure the mean square sums of the VIP scores among all predictor variables equals to unity, $p^{-1} \sum_{j=1}^{p} \text{VIP}_j^2 = 1$. Because of this normalization, predictor variables with VIP scores above (or below) unity can be regarded as important (or unimportant) variables.

**Receiver-Operating-Characteristics curve.** The ROC curve is a model assessment tool and is generated by comparing the posterior probabilities with a varying decision threshold value scanned from zero to unity, where an observation is assigned positive if its positive class posterior
probability is above the threshold. This evaluation method allows an overall assessment of the classifier’s performance for all levels of decision making aggressiveness, from totally liberal by assigning all observations to be positive (threshold = 0) to totally conservative by assigning all to be negative (threshold = 1). Prediction performance is quantified by the area under the curve, where a value of 0.5 means random guessing, 0.7-0.8 means the discrimination capability is acceptable, 0.8-0.9 means excellent, and 0.9-1.0 means outstanding. We perform the averaging of the ROC curves using the non-parametric method adapted from Ref. 19. This method ensures that: 1) the AUC of the averaged curve equals the average AUC of individual cross-validation runs, and 2) the averaged AUC for a perfect (or random) classifier is equal to 1 (or 0.5). Proof for statement 1) can be found in the appendix of Ref. 19, while statement 2) can be straightforwardly deduced from 1).

In our work, we average the individual ROC curves vertically in the tilted space formed by rotating the (FP,TP) axes counter-clockwise by an angle $\theta < \pi/2$, where FP and TP denotes false positive rates and true positive rates, respectively. This enables the averaging to be taken over singular functions. Any data point from an individual ROC curve can take its FP values from $\{(0,1,2,...,N)/N\}$, and TP values from $\{(0,1,2,...,P)/P\}$. Since we are using stratified sampling at the fixed testing set size $L_{test} = P + N$, different cross-validation runs preserve the total number of positives $P$ and negatives $N$. Hence, we choose $\theta = \arctan (P/N)$ such that the curve averaging in the tilted space will be performed to yield a total of $(L_{test} + 1)$ sample points for plotting the ROC curve. The $j$-th ($j = 0,1,2,...,L_{test}$) sample point represents the $j$-th observation in the testing set scanned over by the threshold line, and is obtained from the statistical mean over a total of the number of cross-validation runs of the $j$-th observation from each run.

Acknowledgements

We thank Holly Gates-Mayer, Mark T. Hernandez, Aaron Gilad Kusne, and Lee R. Liu for helpful discussions. This work is supported by AFOSR 9FA9550-19-1-0148; DOE DE-SC0002123, NSF CHE-2053117, NSF QLCI OMA–2016244, NSF PHY-1734006, and NIST. J.T. was supported by
the Lindemann Trust in the form of a Postdoctoral Fellowship.

**Authors contributions**

All authors contributed to the experimental design, results interpretation, and manuscript writing. Q.L. and Y.C. collected and analyzed the data.

**Competing interests**

The authors declare no competing interests.

**References**

[1] C. Wang and P. Sahay, “Breath Analysis Using Laser Spectroscopic Techniques: Breath Biomarkers, Spectral Fingerprints, and Detection Limits,” Sensors 9(10), 8230–8262 (2009).

[2] F. Adler, M. J. Thorpe, K. C. Cossel, and J. Ye, “Cavity-Enhanced Direct Frequency Comb Spectroscopy: Technology and Applications,” Annual Review of Analytical Chemistry 3(1), 175–205 (2010).

[3] M. J. Thorpe, D. Balslev-Clausen, M. S. Kirchner, and J. Ye, “Cavity-enhanced optical frequency comb spectroscopy: application to human breath analysis,” Optics Express 16(4), 2387–2397 (2008).

[4] Q. Liang et al., “Ultrasensitive multispecies spectroscopic breath analysis for real-time health monitoring and diagnostics,” Proceedings of the National Academy of Sciences 118(40) (2021).

[5] Y. Liu et al., “Aerodynamic analysis of SARS-CoV-2 in two Wuhan hospitals,” Nature 582(7813), 557–560 (2020).
[6] H. Ritchie et al., “Coronavirus Pandemic (COVID-19),” Our World in Data (2020). Https://ourworldindata.org/coronavirus. Retrieved 10 January 2022.

[7] A. D. Subali, L. Wiyono, M. Yusuf, and M. F. A. Zaky, “The potential of volatile organic compounds-based breath analysis for COVID-19 screening: a systematic review & meta-analysis.” Diagnostic Microbiology and Infectious Disease 102(2), 115,589 (2022).

[8] B. Shan et al., “Multiplexed nanomaterial-based sensor array for detection of COVID-19 in exhaled breath,” ACS nano 14(9), 12,125–12,132 (2020).

[9] B. N. Zamora-Mendoza, L. Díaz de León-Martínez, M. Rodríguez-Aguilar, B. Mizaikoff, and R. Flores-Ramírez, “Chemometric analysis of the global pattern of volatile organic compounds in the exhaled breath of patients with COVID-19, post-COVID and healthy subjects. Proof of concept for post-COVID assessment,” Talanta 236, 122,832 (2022).

[10] D. M. Ruszkiewicz et al., “Diagnosis of COVID-19 by analysis of breath with gas chromatography-ion mobility spectrometry - a feasibility study,” EClinicalMedicine 29-30, 100,609 (2020).

[11] H. Chen et al., “COVID-19 screening using breath-borne volatile organic compounds,” Journal of Breath Research 15(4), 047,104 (2021).

[12] W. Ibrahim et al., “Diagnosis of COVID-19 by exhaled breath analysis using gas chromatography–mass spectrometry,” ERJ Open Research 7(3) (2021).

[13] S. Grassin-Delyle et al., “Metabolomics of exhaled breath in critically ill COVID-19 patients: A pilot study,” EBioMedicine 63, 103,154 (2021).

[14] F. Walsh, I. et al., “DOME: recommendations for supervised machine learning validation in biology,” Nature methods 18(10), 1122–1127 (2021).

[15] S. de Jong, “SIMPLS: An alternative approach to partial least squares regression,” Chemometrics and Intelligent Laboratory Systems 18(3), 251–263 (1993).
[16] L. C. Lee, C.-Y. Liong, and A. A. Jemain, “Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategies and knowledge gaps,” Analyst 143(15), 3526–3539 (2018).

[17] J. N. Mandrekar, “Receiver Operating Characteristic Curve in Diagnostic Test Assessment,” Journal of Thoracic Oncology 5(9), 1315–1316 (2010).

[18] T. Fawcett, “An introduction to ROC analysis,” Pattern Recognition Letters 27(8), 861–874 (2006).

[19] W. Chen and F. W. Samuelson, “The average receiver operating characteristic curve in multi-reader multicase imaging studies,” The British journal of radiology 87(1040), 20140,016 (2014).

[20] I.-G. Chong and C.-H. Jun, “Performance of some variable selection methods when multicollinearity is present,” Chemometrics and Intelligent Laboratory Systems 78(1), 103–112 (2005).

[21] M. Sawano, K. Takeshita, H. Ohno, and H. Oka, “RT-PCR diagnosis of COVID-19 from exhaled breath condensate: a clinical study,” Journal of breath research 15(3), 037,103 (2021).

[22] K. Iwakuni et al., “Phase-stabilized 100 mW frequency comb near 10 µm,” Appl. Phys. B 124, 1289 (2018).

[23] G. Scalari, J. Faist, and N. Picqué, “On-chip mid-infrared and THz frequency combs for spectroscopy,” Appl. Phys. Lett. 114, 150,401 (2019).

[24] H. Guo et al., “Nanophotonic supercontinuum-based mid-infrared dual-comb spectroscopy,” Optica 7, 1181 – 1188 (2020).

[25] D. Lesko et al., “A six-octave optical frequency comb from a scalable few-cycle erbium fibre laser,” Nature Photonics 15(4), 281–286 (2021).

[26] C. Xiang et al., “Laser soliton microcombs heterogeneously integrated on silicon,” Science 373(6550), 99–103 (2021).
[27] A. Fathy, Y. M. Sabry, S. Nazeer, T. Bourouina, and D. A. Khalil, “On-chip parallel Fourier transform spectrometer for broadband selective infrared spectral sensing,” Microsystems & Nanoengineering 6(1), 1–9 (2020).

[28] F. Amato et al., “Artificial neural networks in medical diagnosis,” Journal of Applied Biomedicine 11(2), 47–58 (2013).

[29] X. FXu et al., “11 TOPS photonic convolutional accelerator for optical neural networks,” Nature 589(7840), 44–51 (2021).

[30] J. Feldmann et al., “Parallel convolutional processing using an integrated photonic tensor core,” Nature 589(7840), 52–58 (2021).

[31] P. A. Harris et al., “Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support,” Journal of Biomedical Informatics 42(2), 377–381 (2009).

[32] P. A. Harris et al., “The REDCap consortium: Building an international community of software platform partners,” Journal of Biomedical Informatics 95, 103,208 (2019).

[33] P. Maslowski et al., “Surpassing the path-limited resolution of Fourier-transform spectrometry with frequency combs,” Phys. Rev. A 93, 021,802 (2016).

[34] I. Gordon et al., “The HITRAN2016 molecular spectroscopic database,” Journal of Quantitative Spectroscopy and Radiative Transfer 203, 3–69 (2017).
Supplementary Information

Variance explained by the PLS components.

The total percentage variance explained in the response analyzed by the MS and the DS approaches for the complete data set (N = 170) are given in Figure S1. We find a sharp rise in the variance explained for both the MS and the DS approaches when the number of PLS components constructed lies in the range from unity to five. A total of 15 PLS components are well sufficient to saturate the percentage variance explained for both approaches. The larger saturated value of the percentage variance explained obtained by the DS approach over the MS approach is attributed to the significantly larger amount of predictor variables for the DS approach.

Figure S1: Total percentage variance explained in the response. Results for (a) the MS approach and (b) the DS approach.

Uncertainty in the AUC.

Uncertainty in the AUC for different response types are calculated using different number of PLS components and different partition ratio of the training and testing set (see Figure S2). For each number of PLS components and partition ratio used, an AUC value is calculated from the averaged ROC curve obtained from 1,000 cross-validation runs based on stratified random sampling. Seen
from Figure S2, the AUC values calculated with only one PLS component are found to give worse prediction performance in general for both the MS and the DS approaches. This is understandable because both approaches show limited total percentage variance explained when only one PLS component is constructed (see Figure S1). For this reason, we calculate the mean and standard deviation of the AUC for each plot excluding those obtained using only one PLS component. Obtained values are reported in the title of each plot. The standard deviations are used as the uncertainty of AUC. The means are provided for reference. Note that in the main text the absolute values quoted for the AUC is computed using 15 PLS components, 140:30 training and testing partition ratio, and 10,000 cross-validation runs. We find the computed values using this setting match to the means obtained here to within the uncertainty calculated.

Figure S2: AUC calculated for different number of PLS components and different training and testing set partition ratio. For different partition ratio, we show the testing set size in plotting the results. The training set size can be obtained by subtracting the complete data set size (N = 170) by the testing set size. a, b, c, results for the MS approach, for birth month, gender, and COVID-19, respectively. d, e, f, results for the DS approach, for birth month, gender, and COVID-19, respectively.
Binary classification results for a total of eleven response types.

A summary of binary response classification results for a total of eleven response types is provided in Table 1. The obtained AUC shown for each response type are the mean and standard deviation calculated for the results obtained using 1,000 cross-validation runs based on stratified random sampling, evaluated at 3, 5, 7, ..., 15 PLS components, and at 10, 20, 30, ..., 60 test set size with training set size given by subtracting the complete data set size (N = 170) by the testing set size. We note that the true class labels for the research subjects are obtained from the self-identified information collected using the study questionnaire, with the only exception of COVID-19 which is checked by the RT-PCR test results. See Methods for more details.

| Response type                        | True class assignment (positive/negative) | Class distributions (positive/negative) | Obtained AUC (mean and std. dev.) | Discrimination capability |
|--------------------------------------|------------------------------------------|----------------------------------------|-----------------------------------|--------------------------|
| Birth day                            | Odd/Even                                 | 83(49%)/87(51%)                        | 0.510(21)                         | Random guessing          |
| Birth month                          | Odd/Even                                 | 83(49%)/87(51%)                        | 0.517(4)                          | Random guessing          |
| Work/school commute by driving a car | Yes/No                                    | 70(41%)/100(59%)                       | 0.522(15)                         | Random guessing          |
| Alcohol frequency                    | >0 days per week/0 days per week         | 125(74%)/45(26%)                      | 0.542(16)                         | Random guessing          |
| Lactose intolerance                  | Moderate to very severe/Not at all to mild | 23(14%)/147(86%)                      | 0.574(16)                         | Random guessing          |
| Smoker                               | Yes/No                                    | 31(18%)/139(82%)                       | 0.604(13)                         | Fair                    |
| Abdominal pain                       | Rarely to frequently/Never               | 91(54%)/79(46%)                       | 0.660(15)                         | Fair                    |
| Gender                               | Female/Male                               | 87(51%)/83(49%)                       | 0.673(12)                         | Fair                    |
| Constipation                         | Moderate to very severe/Never to mild     | 11(6%)/159(94%)                       | 0.674(25)                         | Fair                    |
| COVID-19                             | Yes/No                                    | 83(49%)/87(51%)                        | 0.851(4)                          | Excellent               |
| Breath or Air                        | Breath/Air                               | 170(92%)/15(8%)                       | 1.000(0)                          | Perfect                 |

Table 1: Summary of classification results for eleven response types. AUC values are obtained for each response type and used for judgement of the discrimination capability.

In Figure S3, we show the ROC curves for the eleven response types calculated using 10,000 cross-validation runs based on stratified random sampling, evaluated at 15 PLS components, and at 140:30 training and test set partition ratio. Reported on the curves are the AUC calculated for this particular cross-validation setting and the uncertainties are obtained from Table 1. All results
shown in Table 1 and Figure S3 are analyzed by the DS approach.

We identify five response types to be random guessing, four with fair discrimination capability, and two with excellent or perfect discrimination capability. The classification results reveals the discrimination capability for the detectable biomarkers present in the spectral range of 2810 cm$^{-1}$ to 2945 cm$^{-1}$ and show how the biomarkers can be used for predictions of multiple response types in a simultaneous manner. The air samples used for the human vs. air classification are collected on separate days at the assigned meeting location where the research subjects have their breath sample provided. An AUC of unity with zero uncertainty is obtained. VIP score analysis via the MS approach using a saturated total of 15 PLS components shows only two molecules, H$_2$O and HDO, are identified as the important predictor variables, which have their VIP scores above unity.

**Between-type response correlation.**

We evaluate the between-type response correlation for our study group (N = 170) using the Pearson correlation coefficient. For any two univariate response types to be investigated, their Pearson correlation coefficient is calculated from the ratio between their covariance and the product of their standard deviations. The Pearson correlation coefficient ranges from minus unity to plus unity. A value of zero means no between-type response correlations, i.e., the opposing response classes of the first response type each contains an equal amount of opposing response classes of the second response type. Calculation results are shown in Figure S4. We find negligible between-type response correlation for COVID-19, gender, and birth month. The gender, however, shows mild correlation with constipation, abdominal pain, and lactose intolerance. These response types could potentially jointly explain the breath difference between male and female.
Figure S3: ROC curves for eleven response types. Left, middle, and right columns show response types with random guessing (AUC~0.5), fair (0.6≥AUC<0.7), and excellent or perfect (AUC≥0.8) discrimination capability, respectively. See Table 1 for class assignment and class distributions for each response type. Results shown here for the response types of birth month, gender, and COVID-19 are the same as that shown in the Figure 4 of the main text.
**Figure S4: Between-type response correlation.** Pearson correlation coefficients calculated for the univariate binary response types investigated in our study. See text for discussions.