PAPER • OPEN ACCESS

Breath analysis by ultra-sensitive broadband laser spectroscopy detects SARS-CoV-2 infection

View the article online for updates and enhancements.

PAPER

Breath analysis by ultra-sensitive broadband laser spectroscopy detects SARS-CoV-2 infection

Qizhong Liang[1,2,*] , Ya-Chu Chan[1,3], Jutta Toscano[1,2,8] , Kristen K Bjorkman[4] , Leslie A Leinwand[4,5] , Roy Parker[4,6], Eva S Nozik[7], David J Nesbitt[1,2,3] and Jun Ye[1,2,*]

[1] JILA, National Institute of Standards and Technology and University of Colorado, Boulder, CO 80309, United States of America
[2] Department of Physics, University of Colorado, Boulder, CO 80309, United States of America
[3] Department of Chemistry, University of Colorado, Boulder, CO 80309, United States of America
[4] BioFrontiers Institute, University of Colorado, Boulder, CO 80303, United States of America
[5] Department of Molecular Cellular and Developmental Biology, University of Colorado, Boulder, CO 80303, United States of America
[6] Department of Biochemistry and HHMI, University of Colorado, Boulder, CO 80303, United States of America
[7] Cardiovascular Pulmonary Research Laboratories, Departments of Pediatrics and Medicine, and Division of Pediatric Critical Care Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, United States of America
[8] Present address: Department of Chemistry, University of Basel, Klingelbergstrasse 80, 4056 Basel, Switzerland.
[*] Authors to whom any correspondence should be addressed.

E-mail: Qizhong.Liang@colorado.edu and Ye@jila.colorado.edu

Abstract

Rapid testing is essential to fighting pandemics such as coronavirus disease 2019 (COVID-19), the disease caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Exhaled human breath contains multiple volatile molecules providing powerful potential for non-invasive diagnosis of diverse medical conditions. We investigated breath detection of SARS-CoV-2 infection using cavity-enhanced direct frequency comb spectroscopy (CE-DFCS), a state-of-the-art laser spectroscopic technique capable of a real-time massive collection of broadband molecular absorption features at ro-vibrational quantum state resolution and at parts-per-trillion volume detection sensitivity. Using a total of 170 individual breath samples (83 positive and 87 negative with SARS-CoV-2 based on reverse transcription polymerase chain reaction tests), we report excellent discrimination capability for SARS-CoV-2 infection with an area under the receiver-operating-characteristics curve of 0.849(4). Our results support the development of CE-DFCS as an alternative, rapid, non-invasive test for COVID-19 and highlight its remarkable potential for optical diagnoses of diverse biological conditions and disease states.

1. Introduction

The difficulty to rapidly and accurately detect severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection has been a barrier to the response throughout the coronavirus disease 2019 (COVID-19) pandemic [1]. The current gold standard method, reverse transcription polymerase chain reaction (RT-PCR) test to detect viral RNA [2], requires appropriate sample collection and storage for accuracy, and is time-consuming [3]. Sampling is typically invasive (e.g. nasal swab), contributing to test hesitancy. The real-time assessment of community prevalence, implementation of public health protocols, and timely anti-viral intervention for high-risk people

[4, 5], would all benefit significantly from the development of rapid, safe, sensitive, and non-invasive detection methods for SARS-CoV-2 infection, particularly with recent variants showing an increased epidemic growth rate [6].

Exhaled breath analysis is an attractive alternative to RT-PCR detection of SARS-CoV-2 infection as it is non-invasive and can return real-time measurements [7, 8]. Early studies to develop breath-based COVID-19 diagnosis included nanomaterial-based sensors [9, 10], ion-mobility spectrometry [11, 12], and mass spectrometry [13, 14]. A COVID-19 breath diagnostic test based on gas chromatography-mass spectrometry (GC-MS) was recently granted emergency use authorization by the U.S. Food and Drug

Administration after its validation with over 2409 individuals, reporting 91.2% sensitivity and 99.3% specificity [15, 16]. While GC-MS currently represents one of the most powerful techniques for breath analysis due to its superior detection sensitivity and specificity [7, 17], breath molecules present with identical mass-to-charge ratio imposes real analytical challenges for mass spectrometry to discriminate. In addition, unavoidable alteration to breath components via purification, derivatization, and thermal degradation introduced from the use of a pre-concentrator [16] and a high-temperature thermal process [18] can also hinder accurate measurement of breath profiles.

The recently-developed laser spectroscopy-based technique, namely the cavity-enhanced direct frequency comb spectroscopy (CE-DFCS) [19, 20], can help overcome the analytical challenges of mass spectrometry. CE-DFCS rapidly detects and identifies molecules in exhaled breath by ultra-sensitively measuring their structure-specific absorption signals via laser light at numerous optical frequencies. It requires no sample heating or purifying and ensures chemistry-free determinations of breath profiles. Together with the superior parts-per-trillion detection sensitivity [19], and with robust specificity to discriminate between different isomeric, isobaric, and isotopologue compounds [21], this technique offers rapid, accurate, and robust information that can add to diagnosis and mechanistic insight. Recent proof-of-principle studies have demonstrated the use of CE-DFCS to monitor changes in exhaled breath profiles upon fruit intake [19] and smoking [20], showing potential utility for disease diagnostics. To test if this powerful methodology may be useful for non-invasive medical diagnostics, a trial study was carried out for the first time to test its ability to identify SARS-CoV-2 infection in a young, highly vaccinated cohort as a case study.
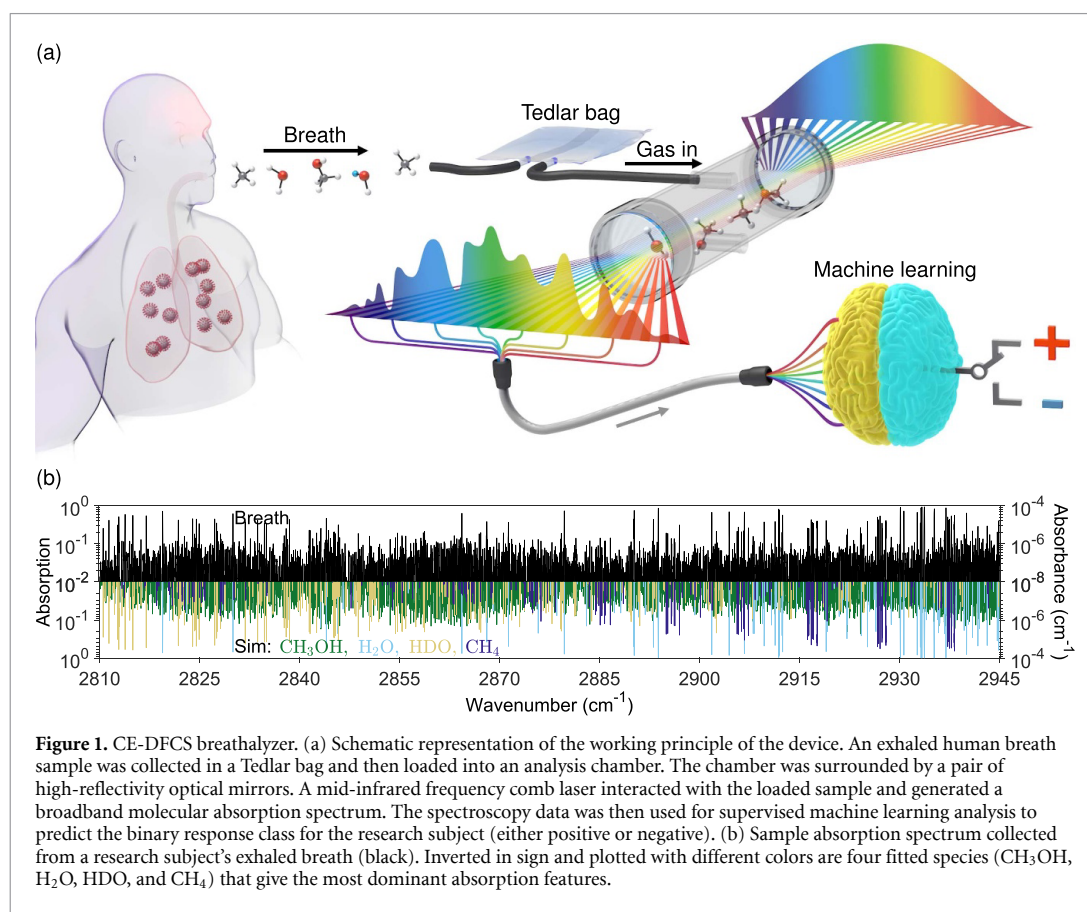
2. Method

2.1. Human subjects
This study was approved by the Institutional Review Board (protocol no. 21-0088) of the University of Colorado Boulder. From May 2021 to January 2022, breath samples from a total of 170 research subjects were collected with a class distribution for SARS-CoV-2 infection of 83 positives (48.8%) and 87 negatives (51.2%). Research subjects were all University of Colorado Boulder affiliates, at least 18 years old, and recruited after taking a university-provided saliva-based or nasal swab COVID-19 RT-PCR test. The general campus population was >90% vaccinated. No participants were severely ill or requiring hospitalization at the time of their sample collection. After receiving their COVID-19 test results, potential subjects received a study recruitment email and were asked to contact the research team within 24 h

if interested in participation. They then reviewed and signed an informed consent form, completed a questionnaire, and scheduled an appointment for the collection of their breath samples. The questionnaire collected self-reported information on sex, age, and race as well as other factors that could impact breath analysis including smoking, alcohol use, and underlying gastrointestinal symptoms. Additional information was collected on acute symptoms experienced by the positive participants. No viral genomes were sequenced, but the Colorado statewide data [22] over our subject recruitment period indicates infection with several viral variants associated with several infection waves (namely, alpha, delta, and omicron) in the community. All data (i.e. informed consent form, questionnaire, and Tedlar bag ID) were collected and managed using the REDCap electronic data capture tool [23, 24] hosted by the University of Colorado Denver.

2.2. Breath sample collection and handling
Standard Tedlar bags (1 l, part no. 249-01-PP, SKC Inc.) were used to collect exhaled breath. During the sample collection appointment, research subjects were asked to hold their nose and breathe through their 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 sample collection location was an outdoor university parking lot. The participants were not instructed to limit or control their smoking, food or alcohol intake prior to sample collection. Right after collection of one breath sample, the Tedlar bag was stored inside an air-tight container at ambient temperature and transported to the indoor lab housing the CE-DFCS setup for immediate data collection and analysis. The breath sample was warmed to 37 °C for 20 min to reduce condensation, then steadily flowed through the cleaned vacuum chamber held at room temperature (20 °C) at a rate of ∼1 l min^{-1}. Just before bag exhaustion, timely closure of the gas valves detained a portion of breath sample inside the chamber and a static pressure of 50 Torr (67 mbar) was reached (without re-condensation) for spectroscopic data collection. After the measurement, the breath sample was pumped out to an exhaust line leading to the building exterior. The used Tedlar bag was autoclaved and disposed of. While direct sampling at atmospheric pressure by our breathalyzer is feasible, off-line sampling and negative pressure were adopted to ensure no SARS-CoV-2 could be introduced into the laboratory air. Spectroscopy data collection for each breath sample was completed in less than 10 min. This can be further reduced to about 1 s when optimized data acquisition and readout are implemented. Overall, from sample collection and transportation to completion of data analysis, the total time was less than an hour. Air samples were

Figure 1. CE-DFCS breathalyzer. (a) Schematic representation of the working principle of the device. An exhaled human breath sample was collected in a Tedlar bag and then loaded into an analysis chamber. The chamber was surrounded by a pair of high-reflectivity optical mirrors. A mid-infrared frequency comb laser interacted with the loaded sample and generated a broadband molecular absorption spectrum. The spectroscopy data was then used for supervised machine learning analysis to predict 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). Inverted in sign and plotted with different colors are four fitted species (CH_3OH, H_2O, HDO, and CH_4) that give the most dominant absorption features.

collected on separate days over the subject's recruitment period at the sample collection location as control specimens.

2.3. CE-DFCS technique

The working principle of the CE-DFCS breathalyzer is illustrated in figure 1(a). A high-resolution broadband absorption spectrum, consisting of a total of 14 836 distinct molecular features each measured ultra-sensitively at individual optical frequencies, was recorded for each breath sample (see sample spectrum in figure 1(b)). The breath spectrum was processed by machine learning analysis for binary response classifications. For additional instrument details, see [19].

2.4. Machine learning analysis

We employed two spectral pre-processing techniques for machine learning analysis: (1) a pattern-based approach that directly used all 14 836 molecular absorption features as the predictor variables; (2) a molecule-based approach that used 16 known small molecule compounds (H_2O, HDO, $^{12}CH_4$, $^{13}CH_4$, OCS, C_2H_4, CS_2, H_2CO, NH_3, CH_3OH, O_3, N_2O, NO_2, SO_3, HCl, and C_2H_6) fitted to the spectra as predictor variables. The former approach identifies all stable patterns that can be used for diagnostics, whereas the latter identifies only the patterns that can

be reduced to known molecular identities, which may result in loss of utilizable chemical information but allows better interpretability into the model details. The 16 compounds were chosen due to their availability from the high-resolution transmission molecular absorption database [25]. While more molecules can potentially be uncovered and fitted, quantitative extraction of their identities requires cross-sectional data at our experimental conditions (20 °C temperature and 50 Torr pressure) to be available. Unfitted species are hence not used in the molecule-based analysis despite being potentially useful to facilitate better predictive power.

To enable binary class assignment, we used partial least squares-discriminant analysis (PLS-DA) [26]. This method allows for the reduction of high-dimensionality data into a one-dimensional scalar number to differentiate between the opposing response classes (positive vs. negative). Variable importance in the projection (VIP) scores [27] were determined for assessing the relative importance of each predictor variable. To assess predictive power, the complete dataset ($N = 170$) was randomly divided into a training set ($n = 140$) and the remaining as a testing set ($n = 30$). Both sets shared the same binary class distributions as the complete data set. The training set was used for model construction (a total of 15 PLS components were constructed) and

the testing set was used for a blind test to obtain a receiver-operating-characteristic (ROC) curve, from which the area under the curve (AUC) value was calculated. Depending on how the complete set was divided, the AUC value obtained can vary to a certain extent. To ensure convergence, we repeated the whole process (i.e. cross-validation) for a total of 10 000 times, and each time a new training set and testing set were randomly re-selected for a new AUC value to be calculated. The ROC curves generated from the total of 10 000 cross-validation runs were averaged together to obtain an averaged ROC curve. The AUC of the averaged curve thus represents the average AUC from all cross-validation runs. To determine the AUC uncertainty, we used different training/testing partition ratios and different numbers of PLS components. All analysis code was written using MAT-LAB, and the PLS-DA was performed using the built-in package based on the SIMPLS algorithm [28]. The supplementary file contains additional details on PLS-DA and VIP score principles, ROC averaging, and AUC uncertainties.

3. Results

3.1. Subject characteristics
One-hundred and seventy participants enrolled in this study, with characteristics summarized in table 1. These included 83 (48.9%) SARS-CoV-2 positive subjects and 87 (51.2%) SARS-CoV-2 negative subjects based on prior RT-PCR tests. The median age was 22 years in the infection-positive and 24 years in the infection-negative groups ($p < 0.05$). Both infection-positive and negative groups were balanced for sex (53.0% female infection-positives, 49.4% female negatives). Race and ethnicity distributions were equivalent between infection-positive and negative groups. A higher number of infection-negative subjects reported a history of rare to occasional abdominal symptoms, though there was no difference in the history of lactose intolerance or constipation between the two groups. SARS-CoV-2-positive subjects were asked additional questions regarding COVID-19-related symptoms, if any (table 2). We found most subjects reported multiple symptoms (figure 2). Of 78 who responded, 50.0% reported 5–7 of the 11 listed symptoms, 5.1% were asymptomatic, and 2.6% reported 10 symptoms.

3.2. Comparable prediction accuracy for SARS-CoV-2 infection by RT-PCR and CE-DFCS
Breath analysis by laser spectroscopy can differentiate between SARS-CoV-2 infection positives and negatives. Using the two spectral pre-processing techniques for machine learning analysis, we found the pattern-based approach yielded an AUC of 0.849 (standard deviation [SD], 0.004) (figure 3(b)) and the molecule-based approach yielded an AUC of 0.769 (SD, 0.007) (figure 3(e)). Both approaches confirmed
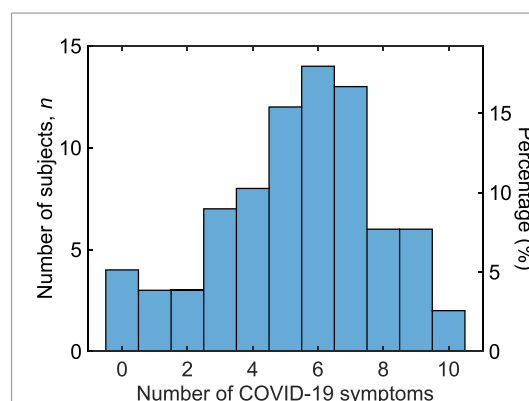

Figure 2. Number of COVID-19 symptoms experienced by the positive participants. See table 2 for a list of COVID-19 symptoms. Only SARS-CoV-2 positive participants with non-missing questionnaire responses were included.

that significant differences in breath contents caused by SARS-CoV-2 infection was successfully detected by CE-DFCS. The classification results on SARS-CoV-2 infection should be interpreted as the co-agreement between the CE-DFCS breath test and the RT-PCR tests employed. As control experiments to validate the analysis methodology, we checked predictions for two cases with known responses: (1) a random guess based on subjects born in even vs. odd months, for which the lowest possible AUC of 0.5 is expected; (2) a perfect discrimination comparing ambient air vs. exhaled breath samples, for which one expects an AUC of 1. Both the pattern-based and molecule-based approaches confirmed expectations for results from a random sampling by birth month (figures 3(a) and (d)), yielding an AUC of 0.516 (SD, 0.004) and 0.488 (SD, 0.009) respectively. With regard to ambient air vs. breath, both approaches yielded AUCs of 1.000 (SD, 0.000) (figures 3(c) and (f)) and confirmed perfect discrimination criterion. These results further support the reliability of our analysis protocol. The AUC of ∼0.5 obtained from predictions of baseline response also suggested that our sample size was large enough to capture sufficient population diversity.

3.3. Pattern-based approach outperforms molecule-based approach
For SARS-CoV-2 infection, we found that the pattern-based approach clearly outperformed the molecule-based approach in prediction performance (AUC of 0.849 (SD, 0.004) vs. 0.769 (SD, 0.007)). To illustrate this result, we made use of the subjects' distribution on the PLS coordinate, which allowed us to visualize which approach can better discriminate opposing response classes. We used the complete data set ($N = 170$) for construction of the PLS coordinate space and plotted subjects' data on the first three PLS components in figures 4(a) (pattern-based) and (b) (molecule-based). The results show significantly

Table 1. Participant characteristics.

Characteristic	Total ($N = 170$)	SARS-CoV-2 positive ($n = 83$; 48.9%)	SARS-CoV-2 negative ($n = 87$; 51.2%)	P^a
Sex				
Female	87 (51.2)	44 (53.0)	43 (49.4)	0.99
Male	83 (48.8)	39 (47.0)	44 (50.6)	
Age, median (IQR), years	23 (8.8)	22 (6)	24 (10)	0.01
Race				
Other/mix	12 (7.0)	4 (4.8)	8 (9.2)	0.53
Asian	20 (11.8)	8 (9.6)	12 (13.8)	
White	138 (81.1)	71 (85.5)	67 (77.0)	
Latino				
Yes	14 (8.2)	7 (8.4)	7 (8.0)	0.84
No	156 (91.8)	76 (91.6)	80 (92.0)	
Alcohol frequency, days week^{-1}				
$d = 0$	45 (26.5)	15 (18.1)	30 (34.5)	
$0 < d \leqslant 3$	114 (67.1)	64 (77.1)	50 (57.5)	0.09
$3 < d \leqslant 7$	11 (6.5)	4 (4.8)	7 (8.0)	
Smoker (Tobacco/Vape/Marijuana)				
Yes	31 (18.2)	11 (13.3)	20 (23.0)	0.05
No	139 (81.8)	72 (86.7)	67 (77.0)	
Abdominal pain				
Never	79 (46.5)	48 (57.8)	31 (35.6)	0.01
Rarely	50 (29.4)	20 (24.1)	30 (34.5)	
$\geqslant$Occasionally	41 (24.1)	15 (18.1)	26 (29.9)	
Lactose intolerance				
Not at all	113 (66.5)	60 (72.3)	53 (60.9)	0.15
Very mild to mild	34 (20.0)	14 (16.9)	20 (23.0)	
Moderate to severe	23 (13.5)	9 (10.8)	14 (16.1)	
Constipation				
Not at all	140 (82.4)	69 (83.1)	71 (81.6)	0.82
Very mild	19 (11.2)	9 (10.8)	10 (11.5)	
$\geqslant$Mild	11 (6.4)	5 (6.0)	6 (6.9)	

Information collected for the total of $N = 170$ participants ($n = 83$ positive; $n = 87$ negative). Unless otherwise indicated, data are presented as n (%). IQR, interquartile range.
[a] P values compare subjects positive and negative for SARS-CoV-2 infection.
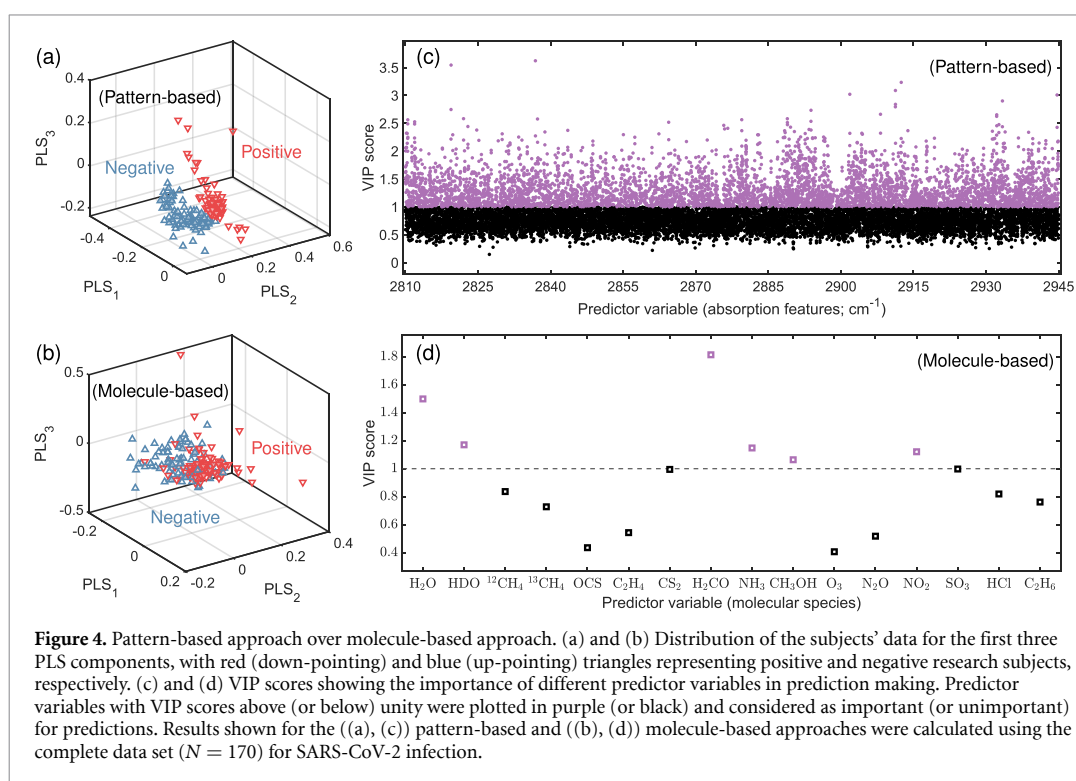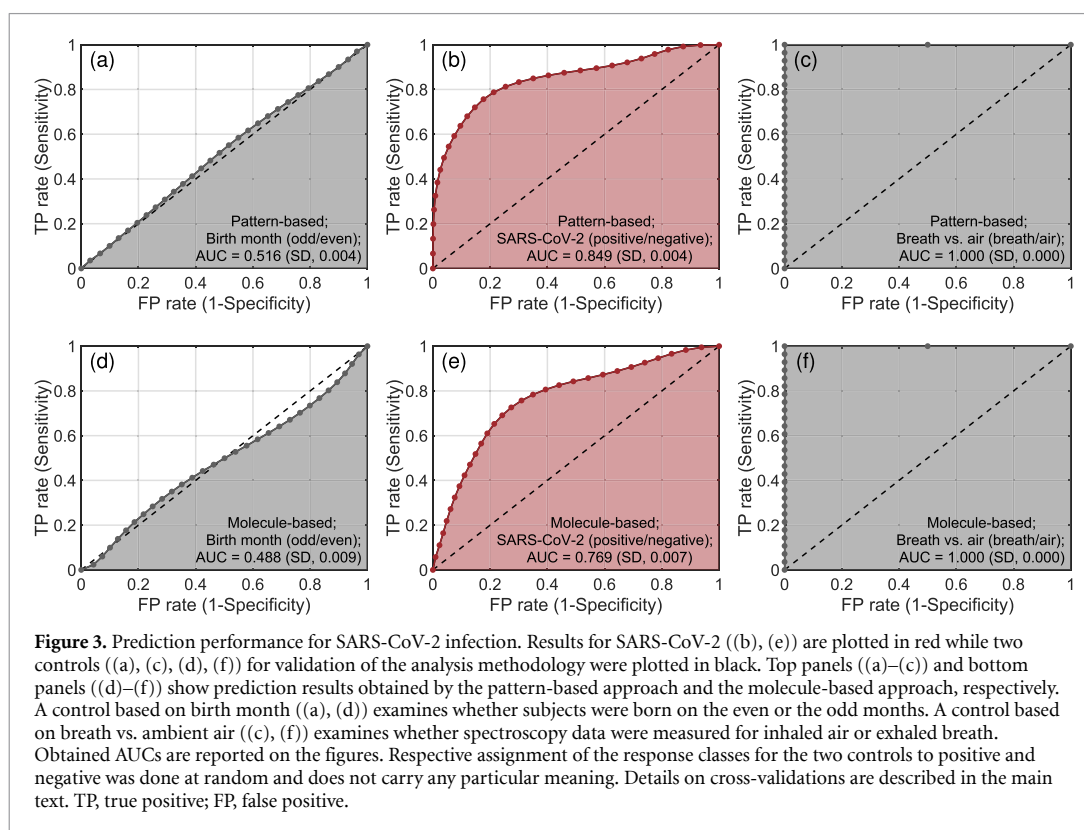
Table 2. COVID-19 symptoms experienced by the positive participants.

Characteristic[a] (positive; $N = 83$)	No, n (%)	Yes, n (%)
Diarrhea	67 (81.7)	15 (18.3)
Fever or chills	44 (53.7)	38 (46.3)
Cough	24 (29.3)	58 (70.7)
Shortness of breath or difficult breathing	64 (78.0)	18 (22.0)
Fatigue	22 (27.2)	59 (72.8)
Muscle or body aches	37 (45.1)	45 (54.9)
Headache	25 (30.5)	57 (69.5)
New loss of taste or smell	59 (72.8)	22 (27.2)
Sore throat	34 (42.0)	47 (58.0)
Congestion or runny nose	12 (14.6)	70 (85.4)
Nausea or vomiting	72 (88.9)	9 (11.1)

[a] Information collected for the COVID-19 positive participants ($N = 83$) only. Statistics n (%) evaluated for those with non-missing values.

better discrimination capability was obtained by the pattern-based approach. The underperformance of the molecule-based approach could potentially be attributed to the exclusion of species with unknown identities in exhaled breath detected by CE-DFCS. As CE-DFCS acquires breath data at extremely high sensitivity, specificity, and dimensionality, applying the pattern-based approach to make full use of the wealth of chemical information collected by CE-DFCS is advantageous in that it bypasses the need for a complete molecular database to directly understand the best possible prediction power.

Figure 3. Prediction performance for SARS-CoV-2 infection. Results for SARS-CoV-2 ((b), (e)) are plotted in red while two controls ((a), (c), (d), (f)) for validation of the analysis methodology were plotted in black. Top panels ((a)–(c)) and bottom panels ((d)–(f)) show prediction results obtained by the pattern-based approach and the molecule-based approach, respectively. A control based on birth month ((a), (d)) examines whether subjects were born on the even or the odd months. A control based on breath vs. ambient air ((c), (f)) examines whether spectroscopy data were measured for inhaled air or exhaled breath. Obtained AUCs are reported on the figures. Respective assignment of the response classes for the two controls to positive and negative was done at random and does not carry any particular meaning. Details on cross-validations are described in the main text. TP, true positive; FP, false positive.

Figure 4. Pattern-based approach over molecule-based approach. (a) and (b) Distribution of the subjects' data for the first three PLS components, with red (down-pointing) and blue (up-pointing) triangles representing positive and negative research subjects, respectively. (c) and (d) VIP scores showing the importance of different predictor variables in prediction making. Predictor variables with VIP scores above (or below) unity were plotted in purple (or black) and considered as important (or unimportant) for predictions. Results shown for the ((a), (c)) pattern-based and ((b), (d)) molecule-based approaches were calculated using the complete data set ($N = 170$) for SARS-CoV-2 infection.

A notable limitation of the pattern-based approach, however, is that it does not reveal which molecules are important for making predictions, but only the optical frequencies at which they are probed.

Variable importance analyzed for the pattern-based approach (figure 4(c)) identified prediction-important optical frequencies (VIP scores > 1 and highlighted in purple) where measured absorption

Figure 5. Prediction performance for a list of potential confounders. Random guessing results (AUC < 0.6) were found for (a) alcohol use, (b) age, and (c) lactose intolerance. Significant differences (0.6 ⩽ AUC < 0.7) were found for (d) smoking, (e) abdominal pain, (f) sex, and (g) constipation. Class assignments for each response type are shown on the figures. For age, a median age of 23 years old was used for class assignment. All results shown were analyzed by the pattern-based approach and details on the cross-validation are described in the main text.

values were strongly discriminative between SARS-CoV-2 positives and negatives. These frequencies are distributed near-uniformly over the entire spectrum. On the other hand, variable importance analyzed for the molecule-based approach (figure 4(d)) identified a panel of indicative molecular species for SARS-CoV-2 infection: water (H_2O), semiheavy water (HDO), formaldehyde (H_2CO), ammonia (NH_3), methanol (CH_3OH), and nitrogen dioxide (NO_2). Being able to identify the molecules provides better clarity to rationalize a possible prediction. To illustrate, variable importance performed for ambient air vs. breath samples based on the molecule-based approach identified water (H_2O) and semi heavy water (HDO) as the only important predictor variables (data not shown). This is easy to understand because water contents were saturated in breath and hence the machine could solely rely on them for prediction. The panel of indicative molecules identified by the molecule-based approach for SARS-CoV-2 infection provides the opportunity for further studies to elucidate the pathophysiology of SARS-CoV-2 infection.

3.4. Prediction performance for a list of potential confounders

We analyzed the prediction performance for a list of subject characteristics and potential factors that could confound the results. For prediction of a specific response, subjects from the complete dataset (N = 170) were divided into opposing classes based on the self-reported questionnaire data. Results obtained using the pattern-based approach are presented in figure 5 and the group assignment criteria for different response types are listed in the panels. A summary for all prediction analyses can also be found in table S1 (see supplementary file). From the results, we found random guessing predictions (AUC <

0.6) for alcohol use, age, and lactose intolerance, but significant prediction capabilities for smoking, sex, abdominal pain, and constipation (0.6 ⩽ AUC < 0.7). On age and abdominal pain, while our subjects had modest correlations with SARS-CoV-2 infection, the significantly better predictive power for SARS-CoV-2 infection suggests that age and abdominal pain do not constitute strong confounders. The superior prediction performance for SARS-CoV-2 infection compared to the list of potential confounders analyzed could potentially be due to SARS-CoV-2 infection eliciting acute and long-term host responses caused by both virus-driven and immune system-associated factors.

4. Discussion

We conducted the first pilot study to evaluate the diagnostic performance of CE-DFCS. Through a case study of SARS-CoV-2 infection detection involving 170 individuals, we found our pattern-based model produced excellent mutual agreement of 0.849 (SD, 0.004) AUC between the CE-DFCS test and the RT-PCR test results. Moreover, using the molecule-based model, we identified the relative importance of different breath molecules in making predictions. Finally, we present preliminary evidence that this technique could be extended to diagnose other conditions.

Our most important finding is that breath analysis by CE-DFCS can differentiate between SARS-CoV-2 infection positives and negatives. This study builds upon our prior works in which we established the use of CE-DFCS for the characterization of exhaled breath molecular profiles upon changes in biological conditions [19, 20]. Here, we have carried out the first trial study for CE-DFCS and employed machine learning analysis to realize robust binary diagnostics. Our study established CE-DFCS as a new

diagnostic tool based on ultra-sensitive broadband laser spectroscopy. Continued assessment of CE-DFCS is important to thoroughly understand its diagnostic utility. Currently, the differences in the study designs make it difficult to compare the performance of CE-DFCS with GC-MS. The GC-MS study that has received FDA approval [15, 16] prospectively conducted RT-PCR tests and collected breath samples within 5 min of each other, restricted eating, drinking, or smoking for the 15 min preceding sample collection and excluded participation from those who had recent exposure to areas of local COVID-19 spread or close contact with COVID-19 positives. By contrast, our study had a much longer time delay from RT-PCR tests to breath sample collections (2.05 (SD, 0.95) days for the positives), and no exclusions based on travel/contact history. The time lag may result in viral clearance, and the more lenient sample collection and recruitment protocols may introduce confounders. These differences preclude a direct comparison of the two techniques. For future studies, examination of CE-DFCS's utility in individuals with severe disease or at higher risk, such as the elderly, the unvaccinated, and those with pre-disposing co-morbidities, will be important.

CE-DFCS may have broader applicability beyond the detection of SARS-CoV-2 infection. It may also (1) serve as a non-invasive tool for evaluation of other health or biological conditions, and (2) provide insights into disease pathogenesis. With respect to (1), our results show that CE-DFCS discriminated between subjects based on smoking history [29, 30], biological sex [31–34], as well as gastrointestinal symptoms [35–37] (recurring abdominal pain and constipation). We were not able to discriminate subjects based on alcohol intake [38] or lactose intolerance [39], but this is not surprising as our subjects had not been specifically challenged with alcohol or lactose ingestion. With respect to (2), it has been recently reported [40] that SARS-CoV-2 virus exhibits strong optical absorption signals within our spectral coverage (2810 cm^{-1}–2945 cm^{-1}). This signal could potentially partly originate from the C–H molecular bonds in the surface-exposed SARS-CoV-2 spike protein [41]. A future measurement of the viral absorption spectrum in the gas phase with proper consideration of protein structure dynamics [42] may allow direct quantification of viral load in exhaled breath with CE-DFCS. This could allow us to examine the correlation between viral burden and other breath biomarkers and to determine the relative contributions of virus and host response to the change in breath molecular profiles. We find our results compelling enough to warrant future investigation into the applicability of CE-DFCS breath analysis to other conditions or diseases, particularly those of respiratory, gastrointestinal, or metabolic origin.

Finally, we note that ongoing rapid developments can further empower CE-DFCS in its use

for medical diagnostics. Spectral range of the current CE-DFCS setup can be expanded to cover more ro-vibrational bands [43–46], thereby probing more discriminative features for stronger predictions. Furthermore, due to the direct measurement capability of CE-DFCS (i.e. no need for chemical treatments, pre-concentrations, and thermal processing), the technique can facilitate the creation of large-scale databases by accumulating breath data from different trial studies. This can promote the construction of deep learning model architectures [47–49] that can outperform traditional machine learning algorithms (e.g. PLS-DA) in predictive power. Recent photonics advances could potentially permit chip-scale miniaturization [50–52] for CE-DFCS and thus the technique could eventually be integrated into portable devices to support low-cost, widespread use and enable daily self-health monitoring on the go.

5. Conclusion

We present the first trial study of laser frequency comb spectroscopy for non-invasive medical diagnostics. Our case study of SARS-CoV-2 infection detection among a total of 170 individuals finds excellent mutual agreement between CE-DFCS and RT-PCR tests and supports the development of CE-DFCS as an alternative and accurate COVID-19 test with non-invasive sampling and rapid turnaround time. While the outstanding prediction performance was achieved using the pattern-based approach, continued enrichment in the molecular absorption database will empower high-resolution comb spectroscopy to employ molecule-based approach providing comparable prediction accuracy but with significantly better model interpretability. The laser spectroscopy-based technique, capable of ultra-sensitive, multi-species, rapid and chemistry-free detection of breath molecular contents with robust isomer-, isobaric-, and isotopologue-specificity opens a complementary approach for the development of breath-based diagnostics research.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Acknowledgment

Authors contributions

All authors contributed to the study design, results interpretation, and manuscript writing. Q L and Y-C C collected and analyzed the data. The authors declare no competing interests.

ORCID iDs

Qizhong Liang ● https://orcid.org/0000-0001-5422-290X
Jutta Toscano ● https://orcid.org/0000-0003-3594-9462
Kristen K Bjorkman ● https://orcid.org/0000-0002-9540-6657
Leslie A Leinwand ● https://orcid.org/0000-0003-1470-4810
Jun Ye ● https://orcid.org/0000-0003-0076-2112

References

[1] Ritchie H *et al* 2020 Coronavirus Pandemic (COVID-19) *Our World in Data*

[2] Center of Disease Control and Prevention 2020 Interim guidelines for collecting, handling, and testing clinical specimens for COVID-19 (available at: www.cdc.gov/coronavirus/2019-ncov/lab/guidelines-clinicalspecimens.html) (Accessed 9 July 2022)

[3] Mei X *et al* 2020 Artificial intelligence–enabled rapid diagnosis of patients with COVID-19 *Nat. Med.* **26** 1224–8

[4] Larremore D B, Wilder B, Lester E, Shehata S, Burke J M, Hay J A, Tambe M, Mina M J and Parker R 2021 Test sensitivity is secondary to frequency and turnaround time for COVID-19 screening *Sci. Adv.* **7** eabd5393

[5] Saravolatz L D, Depcinski S and Sharma M 2022 Molnupiravir and Nirmatrelvir-Ritonavir: oral COVID antiviral drugs *Clin. Infect. Dis.* **76** 165–71

[6] Backer J A *et al* 2022 Shorter serial intervals in SARS-CoV-2 cases with Omicron BA. 1 variant compared with Delta variant, the Netherlands, 13 to 26 December 2021 *Eurosurveillance* **27** 2200042

[7] Wang C and Sahay P 2009 Breath analysis using laser spectroscopic techniques: breath biomarkers, spectral fingerprints and detection limits *Sensors* **9** 8230–62

[8] Arnold C 2022 Diagnostics to take your breath away *Nat. Biotechnol.* **40** 990–3

[9] Shan B *et al* 2020 Multiplexed nanomaterial-based sensor array for detection of COVID-19 in exhaled breath *ACS Nano* **14** 12125–32

[10] Zamora-Mendoza B N, de León-Martínez L D, Rodríguez-Aguilar M, Mizaikoff B and Flores-Ramírez R 2022 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** 122832

[11] Ruszkiewicz D M *et al* 2020 Diagnosis of COVID-19 by analysis of breath with gas chromatography-ion mobility spectrometry - a feasibility study *eClinicalMedicine* **29-30** 100609

[12] Chen H *et al* 2021 COVID-19 screening using breath-borne volatile organic compounds *J. Breath Res.* **15** 047104

[13] Ibrahim W *et al* 2021 Diagnosis of COVID-19 by exhaled breath analysis using gas chromatography–mass spectrometry *ERJ Open Res.* **7** 3

[14] Grassin-Delyle S *et al* 2021 Metabolomics of exhaled breath in critically ill COVID-19 patients: a pilot study *eBioMedicine* **63** 103154

[15] U. S. Food & Drug Administration Coronavirus (COVID-19) Update: FDA Authorizes First COVID-19 Diagnostic Test Using Breath Samples 2022 (available at: www.fda.gov/news-events/press-announcements/coronavirus-covid-19-update-fda-authorizes-first-covid-19-diagnostic-test-using-breath-samples)

[16] U. S. Food & Drug Administration 2022 InspectIR COVID-19 Breathalyzer (for use on PNY-1000) (available at: https://fda.report/media/157723/EUA-InspectIR-Breath-ifu.pdf)

[17] Smith D, Španěl P, Herbig J and Beauchamp J 2014 Mass spectrometry for real-time quantitative breath analysis *J. Breath Res.* **8** 027101

[18] Fang M, Ivanisevic J, Benton H P, Johnson C H, Patti G J, Hoang L T, Uritboonthai W, Kurczy M E and Siuzdak G 2015 Thermal degradation of small molecules: a global metabolomic investigation *Anal. Chem.* **87** 10935–41

[19] Liang Q, Chan Y-C, Changala P B, Nesbitt D J, Ye J and Toscano J 2021 Ultrasensitive multispecies spectroscopic breath analysis for real-time health monitoring and diagnostics *Proc. Natl Acad. Sci.* **118** e2105063118

[20] Thorpe M J, Balslev-Clausen D, Kirchner M S and Ye J 2008 Cavity-enhanced optical frequency comb spectroscopy: application to human breath analysis *Opt. Express* **16** 2387–97

[21] Kranenburg R F, Peroni D, Affourtit S, Westerhuis J A, Smilde A K and van Asten A C 2020 Revealing hidden information in GC-MS spectra from isomeric drugs: chemometrics based identification from 15 eV and 70 eV EI mass spectra *Forensic Chem.* **18** 100225

[22] Colorado Department of Public Health & Environment 2022 COVID-19 Variant Sentinel Surveillance (available at: https://covid19.colorado.gov/data)

[23] Harris P A, Taylor R, Thielke R, Payne J, Gonzalez N and Conde. J G 2009 Research electronic data capture (REDCap)-A metadata-driven methodology and workflow process for providing translational research informatics support *J. Biomed. Inform.* **42** 377–81

[24] Harris P A *et al* 2019 The REDCap consortium: building an international community of software platform partners *J. Biomed. Inform.* **95** 103208

[25] Gordon I E *et al* 2017 The HITRAN2016 molecular spectroscopic database *J. Quant. Spectrosc. Radiat. Transfer* **203** 3–69

[26] Lee L C, Liong C-Y and Jemain A A 2018 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** 3526–39

[27] Chong I-G and Jun C-H 2005 Performance of some variable selection methods when multicollinearity is present *Chemometr. Intell. Lab. Syst.* **78** 103–12

[28] de Jong S 1993 SIMPLS: an alternative approach to partial least squares regression *Chemometr. Intell. Lab. Syst.* **18** 251–63

[29] Kushch I *et al* 2008 Compounds enhanced in a mass spectrometric profile of smokers' exhaled breath versus non-smokers as determined in a pilot study using PTR-MS *J. Breath Res.* **2** 026002

[30] Buszewski B, Ulanowska A, Ligor T, Denderz N and Amann A 2009 Analysis of exhaled breath from smokers, passive smokers and non-smokers by solid-phase microextraction gas chromatography/mass spectrometry *Biomed. Chromatogr.* **23** 551–6

[31] Grasemann H, van's Gravesande K S, Büscher R, Drazen J M and Ratjen F 2003 Effects of sex and of gene variants in constitutive nitric oxide synthases on exhaled nitric oxide *Am. J. Respir. Crit. Care Med.* **167** 1113–6

[32] Taylor D *et al* 2007 Factors affecting exhaled nitric oxide measurements: the effect of sex *Respir. Res.* **8** 82

[33] Good N *et al* 2021 Respiratory aerosol emissions from vocalization: age and sex differences are explained by volume and exhaled CO$_2$ *Environ. Sci. Technol. Lett.* **8** 12

[34] Antufjew H M, A, Antufjew A, Borgmann K, Hempel F, Ittermann T, Welzel S, Weltmann K D, Völzke H and Röpcke J 2011 Influence of age and sex in exhaled breath samples investigated by means of infrared laser absorption spectroscopy *J. Breath Res.* **5** 027101

[35] Pichetshote N and Rezaie A 2018 Breath tests for functional gastrointestinal disorders: when and for what? *NeuroGastroLATAM Rev.* **2** 87–97

[36] De Lacy Costello B, Ledochowski M and Ratcliffe N 2013 The importance of methane breath testing: a review *J. Breath Res.* **7** 024001

[37] Dutta Banik G, De A, Som S, Jana S, Daschakraborty S B, Chaudhuri S and Pradhan M 2016 Hydrogen sulphide in exhaled breath: a potential biomarker for small intestinal bacterial overgrowth in IBS *J. Breath Res.* **10** 026010

[38] Hlastala M 1998 The alcohol breath test–a review *J. Appl. Physiol.* **84** 401–8

[39] Di Costanzo M and Berni Canani R 2019 Lactose intolerance: common misunderstandings *Ann. Nutrition Metab.* **73** 30–37

[40] Barauna V G, Singh M N, Barbosa L L, Marcarini W D, Vassallo P F, Mill J G, Ribeiro-Rodrigues R, Campos L C G, Warnke P H and Martin F L 2021 Ultrarapid on-site detection of SARS-CoV-2 infection using simple ATR-FTIR spectroscopy and an analysis algorithm: high sensitivity and specificity *Anal. Chem.* **93** 2950–8

[41] Soares J *et al* 2021 Diagnostics of SARS-CoV-2 infection using electrical impedance spectroscopy with an immunosensor to detect the spike protein *Talanta* **239** 123076

[42] López-Lorente A I and Mizaikoff B 2016 Mid-infrared spectroscopy for protein analysis: potential and challenges *Anal. Bioanal. Chem.* **408** 2875–89

[43] Iwakuni K, Porat G, Bui T Q, Bjork B J, Schoun S B, Heckl O H, Fermann M E and Ye J 2018 Phase-stabilized 100 mW frequency comb near 10 μm *Appl. Phys.* B **124** 1289

[44] Scalari G, Faist J and Picqué N 2019 On-chip mid-infrared and THz frequency combs for spectroscopy *Appl. Phys. Lett.* **114** 150401

[45] Guo H, Weng W, Liu J, Yang F, Hänsel W, Brès C S, Thévenaz L, Holzwarth R and Kippenberg T J 2020 Nanophotonic supercontinuum-based mid-infrared dual-comb spectroscopy *Optica* **7** 1181–8

[46] Lesko D, Timmers H, Xing S, Kowligy A, Lind A J and Diddams S A 2021 A six-octave optical frequency comb from a scalable few-cycle erbium fibre laser *Nat. Photon.* **15** 281–6

[47] Amato F, López A, Peña-Méndez E M, Vaňhara P, Hampl A and Havel J 2013 Artificial neural networks in medical diagnosis *J. Appl. Biomed.* **11** 47–58

[48] Al Ibrahim E and Farooq A 2021 Prediction of the derived cetane number and carbon/hydrogen ratio from infrared spectroscopic data *Energy and Fuels* **35** 8141–52

[49] Enders A A, North N M, Fensore C M, Velez-Alvarez J and Allen H C 2021 Functional group identification for FTIR spectra using image-based machine learning models *Anal. Chem.* **93** 9711–8

[50] Xiang C *et al* 2021 Laser soliton microcombs heterogeneously integrated on silicon *Science* **373** 99–103

[51] Jin N, McLemore C A, Mason D, Hendrie J P, Luo Y, Kelleher M L, Kharel P, Quinlan F, Diddams S A and Rakich P T 2022 Micro-fabricated mirrors with finesse exceeding one million *Optica* **9** 965–70

[52] Fathy A, Sabry Y M, Nazeer S, Bourouina T and Khalil D A 2020 On-chip parallel Fourier transform spectrometer for broadband selective infrared spectral sensing *Microsyst. Nanoeng.* **6** 1–9