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

[2]Department of Physics, University of Colorado, Boulder, CO 80309

[3]Department of Chemistry, University of Colorado, Boulder, CO 80309

[4]BioFrontiers Institute, University of Colorado, Boulder, CO 80303

[5]Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, CO 80303

[6]Department of Biochemistry and HHMI, University of Colorado, Boulder, CO 80303

[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

[8]Present address: Department of Chemistry, University of Basel, Klingelbergstrasse 80, 4056 Basel, Switzerland

[*]Corresponding authors: Qizhong.Liang@colorado.edu, Ye@jila.colorado.edu

Supplementary Information

Partial least squares-discriminant analysis (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 was 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.[1] for more details beyond this classification type and how the actual algorithm is implemented. We use bold upper case to denote matrices, bold lower case for vectors, and un-bold for scalars, with primes (′) denoting a matrix or vector transpose. Collected data used for the training process are represented by the $n \times p$ predictor variables matrix $\mathbf{X}_0$ and the $n \times 1$ univariate response variable vector $\mathbf{y}_0$. Here, n is the total number of research subjects, p is the total number of predictor variables. Both $\mathbf{X}_0$ and $\mathbf{y}_0$ are column-centered so that the covariance of different predictor variables with the response can be expressed by a $p \times 1$ column vector $\mathbf{s}_0 = \mathbf{X}_0'\mathbf{y}_0$. PLS regression relates $\mathbf{X}_0$ and $\mathbf{y}_0$ based on $\mathbf{y}_0 = \mathbf{X}_0\mathbf{b} + \mathbf{e}$, where $\mathbf{b}$ is the $p \times 1$ coefficients estimate, $\mathbf{X}_0\mathbf{b}$ is the explained component, and $\mathbf{e}$ is the fit residual. In contrast to least squares regression, where the coefficients estimate $\mathbf{b}$ is constructed by minimizing the residual sum of squares $\mathbf{e}'\mathbf{e}$, PLS regression constructs it based on the covariance $\mathbf{s}_0 = \mathbf{X}_0'\mathbf{y}_0$ to get more stabilized values of $\mathbf{b}$ and achieve more reliable predictive power. The formulation begins by projecting the predictor variables matrix $\mathbf{X}_0$ onto a new coordinate system $\mathbf{T} = \mathbf{X}_0\mathbf{R}$ of reduced dimensionality spanned by a total of A ($\leq p-1$) PLS components, where $\mathbf{R}$ denotes the $p \times A$ weight transfer matrix and $\mathbf{T}$ denotes the $n \times A$ projected scores matrix. The construction of $\mathbf{R}$ is subject to two constraints: 1) the covariance vector $\mathbf{T}'\mathbf{y}_0$ 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 $\mathbf{T}$ satisfy $\mathbf{t}_i'\mathbf{t}_j = \delta_{ij}$ for any $i, j = 1, 2, ..., A$, where δ_{ij} is the Kronecker delta. The coefficients estimate $\mathbf{b}$ can be determined once $\mathbf{R}$ is known, since $\mathbf{y}_0 = \mathbf{T}\mathbf{T}'\mathbf{y}_0 = \mathbf{X}_0\mathbf{R}\mathbf{R}'\mathbf{X}_0'\mathbf{y}_0 = \mathbf{X}_0\mathbf{b}$, and thus $\mathbf{b} = \mathbf{R}\mathbf{R}'\mathbf{X}_0'\mathbf{y}_0 = \mathbf{R}\mathbf{R}'\mathbf{s}_0$. The process of determining $\mathbf{R}$ proceeds column by column. For the first iteration step $k = 1$, the maximization of the covariance of the first PLS component ($\mathbf{t}_k = \mathbf{X}_0\mathbf{r}_k$) with the response, $\mathbf{t}_k'\mathbf{y}_0 = \mathbf{r}_k'\mathbf{X}_0'\mathbf{y}_0 = \mathbf{r}_k'\mathbf{s}_0 = \max$, constrains the first weight vector $\mathbf{r}_k$ ($k = 1$) to be along the direction of $\mathbf{s}_0$. For steps $k > 1$, the orthogonality condition, $\mathbf{t}_k'\mathbf{t}_i = \mathbf{r}_k'(\mathbf{X}_0'\mathbf{t}_i) = 0$ for $i = 1, 2, ..., k-1$, requires the newly constructed $\mathbf{r}_k$ to be orthogonal to each of the $p \times 1$ vectors $\mathbf{X}_0'\mathbf{t}_i$ for $i = 1, 2, ..., k-1$. We define $\mathbf{p}_i \equiv \mathbf{X}_0'\mathbf{t}_i$ as 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 $\mathbf{p}_i$ ($i = 1, 2, ..., k-1$) and then determine the $p \times p$ projection operator $\mathbf{P}^{\perp}$ for the orthogonal complement space $V_{k-1}^{\perp}$. This loosely constrains the direction of $\mathbf{r}_k$ to be within $V_{k-1}^{\perp}$, requiring $\mathbf{r}_k = \mathbf{P}^{\perp}\mathbf{r_k}$. Now, with the covariance maximization criteria, $\mathbf{t}_k'\mathbf{y}_0 = \mathbf{r}_k'(\mathbf{P}^{\perp'}\mathbf{s_0}) = \max$, the direction of $\mathbf{r}_k$ is ultimately determined to be along the direction of the vector $\mathbf{P}^{\perp'}\mathbf{s_0}$, which is the projection of the covariance vector $\mathbf{s}_0$ onto the subspace $V_{k-1}^{\perp}$. The iteration process proceeds until the directions of all $\mathbf{r}_k$ are determined, where the normalization condition $\mathbf{T}'\mathbf{T} = 1$ governs the magnitudes of $\mathbf{r}_k$. Finally, the coefficients estimate is determined and can be used for prediction of the response class for new observations based on $\mathbf{y}_0^{pred} = \mathbf{X}_0^{new}\mathbf{b}$, where the $m \times p$ matrix $\mathbf{X}_0^{new}$ is the testing data for a total of m research subjects. The $m \times 1$ predicted values $\mathbf{y}_0^{pred}$ are translated proportionally into posterior probabilities and compared with a threshold value for response class assignment.

Variable importance in the projection (VIP) scores

In PLS-DA, assessment of the importance of the predictor variables needs to consider 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 $\mathbf{R}$, 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 $\mathbf{y}_0'\mathbf{y}_0$ is explained by the total of A PLS components to the extent of $\hat{\mathbf{y}}_0'\hat{\mathbf{y}}_0$, where $\hat{\mathbf{y}}_0 = \mathbf{X}_0\mathbf{b} = \mathbf{y}_0 - \mathbf{e}$. The total percentage variance explained in the response, $(\hat{\mathbf{y}}_0'\hat{\mathbf{y}}_0/\mathbf{y}_0'\mathbf{y}_0) \times 100\,\%$, can be used for estimating the minimum number of PLS components needed for reliable predictions. The explained variance $\hat{\mathbf{y}}_0'\hat{\mathbf{y}}_0 = \hat{\mathbf{y}}_0'\mathbf{T}\mathbf{T}'\hat{\mathbf{y}}_0 = \sum_{a=1}^{A}(\hat{\mathbf{y}}_0'\mathbf{t}_a)^2$ is further broken down into a summation of the square of the covariance of all PLS components with $\hat{\mathbf{y}}_0$. We can thus evaluate the importance of the a-th PLS component by its variance explained $\mathbf{q}_a^2 \equiv (\hat{\mathbf{y}}_0'\mathbf{t}_a)^2$, 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 $\sum_{a=1}^{A}\mathbf{q}_a^2$. 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, one can define its VIP score[2], a metric for characterizing its importance, by:

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

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

Variance explained by the PLS components.

For SARS-CoV-2 infection classification, the total percentage variance explained in the response analyzed by the molecule-based and the pattern-based approaches for the complete data set (N = 170) are given in Figure S1. We found a sharp rise in the variance explained for both the molecule-based and the pattern-based approaches when the number of PLS components constructed lies in the range from unity to five. A total of 15 PLS components were sufficient to saturate the percentage variance explained for both approaches. The lower variance explained obtained by the molecular species-based approach suggests fitting the spectroscopy data with more molecular species can better explain the response.

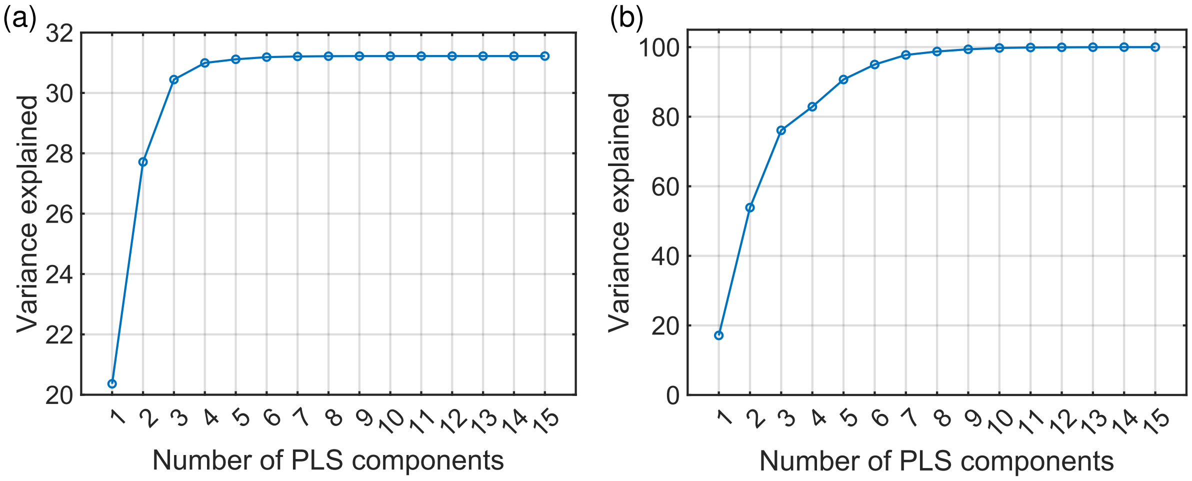


Figure S1: Total percentage variance explained in the response. Results for (**a**) the molecule-based approach and (**b**) the pattern-based approach.

Averaging of the Receiver-Operating-Characteristics curves

We performed averaging of the ROC curves using the non-parametric method adapted from Ref.[3]. This method ensured that: 1) the AUC of the averaged curve equaled the average AUC of individual cross-validation runs, and 2) the averaged AUC for a perfect (or random) classifier was equal to 1 (or 0.5). Proof for statement 1) can be found in the appendix of Ref.[3], while statement 2) can be straightforwardly deduced from 1). In our work, we averaged 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 enabled the averaging to be taken over singular functions. Any data point from an individual ROC curve could take its FP values from $\{(0, 1, 2, ..., \mathrm{N})/\mathrm{N}\}$, and TP values from $\{(0, 1, 2, ..., \mathrm{P})/\mathrm{P}\}$. Since we were using stratified sampling at the fixed testing set size $\mathrm{L}_{\mathrm{test}} = \mathrm{P} + \mathrm{N}$, different cross-validation runs preserved the total number of positives P and negatives N. Hence, we chose $\theta = \arctan{(\mathrm{P}/\mathrm{N})}$ such that the curve averaging in the tilted space would be performed to yield a total of $(\mathrm{L}_{\mathrm{test}} + 1)$ sample points for plotting the averaged ROC curve. The j-th ($j = 0, 1, 2, ..., \mathrm{L}_{\mathrm{test}}$) sample point represented the j-th observation in the testing set scanned over by the threshold line, and was obtained from the statistical mean over a total of the number of cross-validation runs of the j-th observation from each run.

Uncertainty in the AUC

Uncertainty in the AUC for different response types was calculated using different numbers of PLS components and different partition ratios of the training and testing set (see Figure S2). For each number of PLS components and partition ratio used, an AUC value was calculated from the averaged ROC curve obtained from 1,000 cross-validation runs based on stratified random sampling. As seen in Figure S2, the AUC values calculated with only one PLS component were found to give worse prediction performance in general for both the molecule-based and the pattern-based approaches. This is understandable because both approaches showed limited total percentage variance explained when only one PLS component was constructed (see Figure S1). For this reason, we calculated 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 were used as the uncertainty of AUC. The means were provided for reference. Note that in the main text the absolute values quoted for the AUC were computed using 15 PLS components, 140:30 training and testing partition ratio, and 10,000 cross-validation runs. We found the computed values using these settings matched the means obtained here to within the calculated uncertainty.

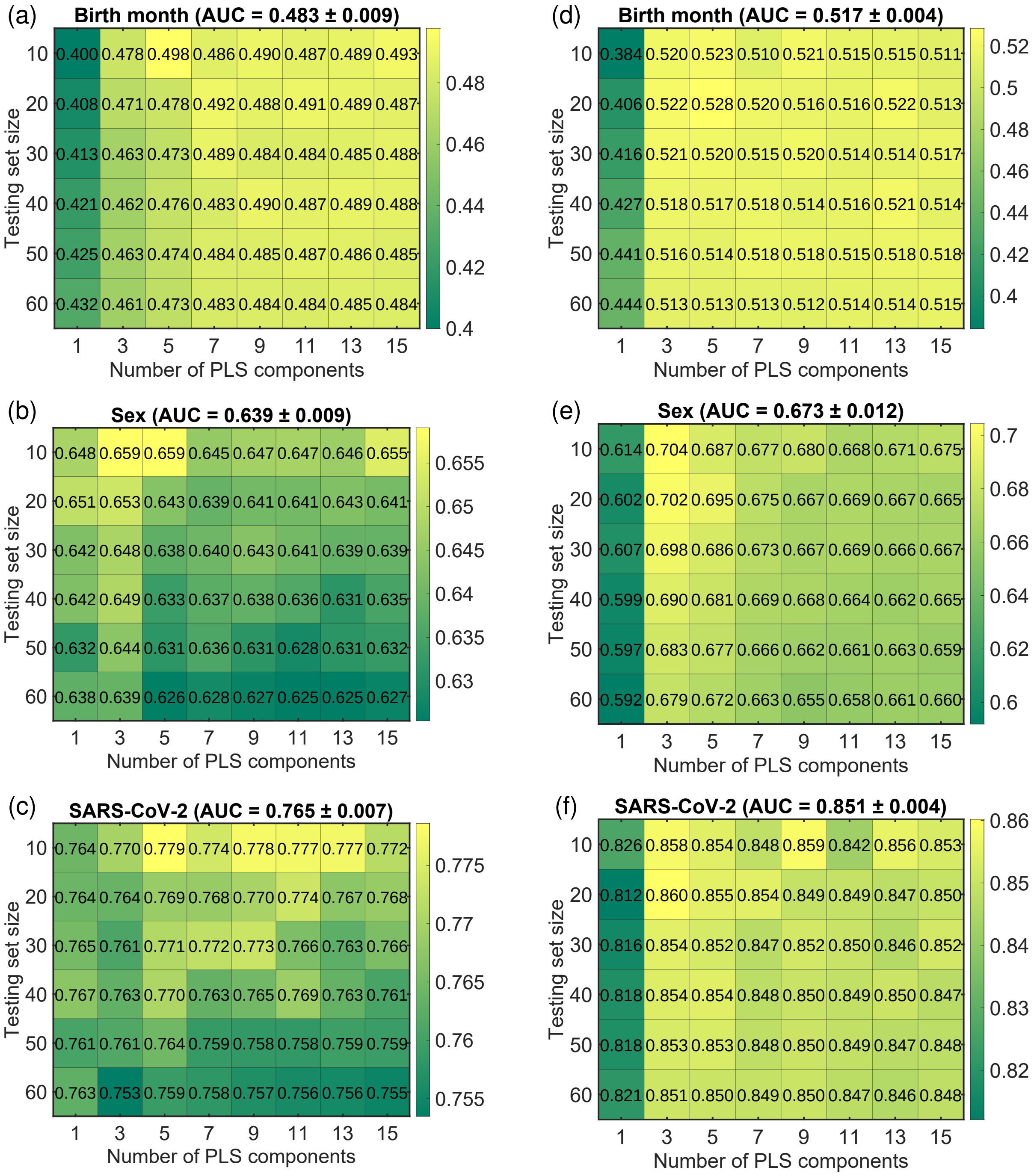
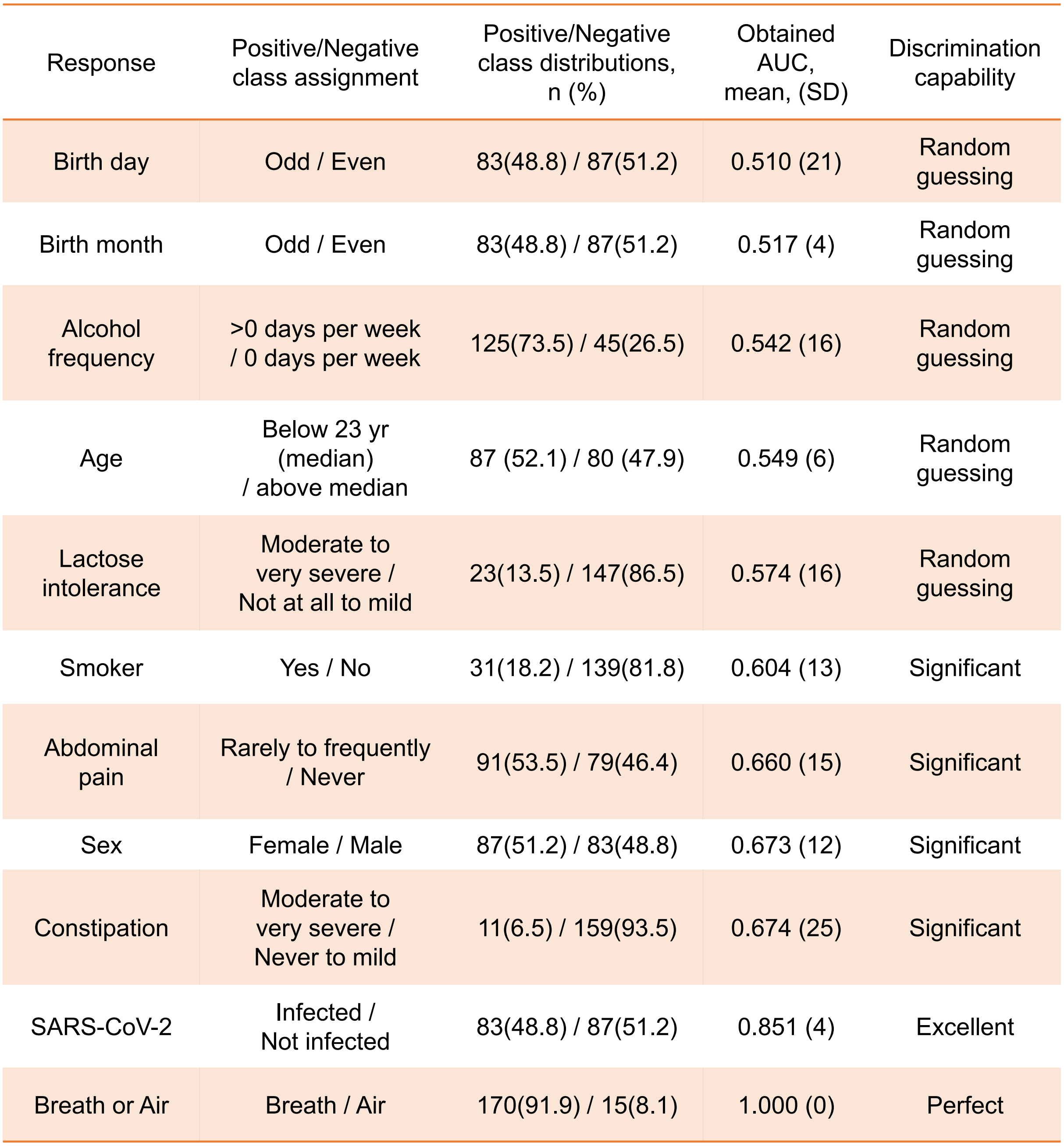


Figure S2: AUC calculated for different numbers of PLS components and different training and testing set partition ratios. For different partition ratios, we show the testing set size in plotting the results. The training set size can be obtained by subtracting the testing set size from the complete data set size (N = 170). **a**, **b**, **c**: results for the molecule-based approach, for birth month, sex, and SARS-CoV-2, respectively. **d**, **e**, **f**: results for the pattern-based approach, for birth month, sex, and SARS-CoV-2, respectively.

Prediction performance summary.

A summary of binary response classification results for various response types is provided in Table S1. The obtained AUC shown for each response type were 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 testing set size from the complete data set.



Response	Positive/Negative class assignment	Positive/Negative class distributions, n (%)	Obtained AUC, mean, (SD)	Discrimination capability
Birth day	Odd / Even	83(48.8) / 87(51.2)	0.510 (21)	Random guessing
Birth month	Odd / Even	83(48.8) / 87(51.2)	0.517 (4)	Random guessing
Alcohol frequency	>0 days per week / 0 days per week	125(73.5) / 45(26.5)	0.542 (16)	Random guessing
Age	Below 23 yr (median) / above median	87 (52.1) / 80 (47.9)	0.549 (6)	Random guessing
Lactose intolerance	Moderate to very severe / Not at all to mild	23(13.5) / 147(86.5)	0.574 (16)	Random guessing
Smoker	Yes / No	31(18.2) / 139(81.8)	0.604 (13)	Significant
Abdominal pain	Rarely to frequently / Never	91(53.5) / 79(46.4)	0.660 (15)	Significant
Sex	Female / Male	87(51.2) / 83(48.8)	0.673 (12)	Significant
Constipation	Moderate to very severe / Never to mild	11(6.5) / 159(93.5)	0.674 (25)	Significant
SARS-CoV-2	Infected / Not infected	83(48.8) / 87(51.2)	0.851 (4)	Excellent
Breath or Air	Breath / Air	170(91.9) / 15(8.1)	1.000 (0)	Perfect

Table S1: Prediction performance summary. AUC values were obtained for each response type and used for judgement of the discrimination capability.

References

[1] S. de Jong, "SIMPLS: An alternative approach to partial least squares regression," Chemometrics and Intelligent Laboratory Systems **18**(3), 251–263 (1993).

[2] 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).

[3] W. Chen and F. W. Samuelson, "The average receiver operating characteristic curve in multireader multicase imaging studies," The British journal of radiology **87**(1040), 20140,016 (2014).