Volatile organic compounds (VOCs) breath signatures of children with cystic fibrosis by real-time secondary Electrospray Ionization High Resolution Mass Spectrometry (SESI-HRMS). (Supplementary Material)

Weber Ronja[1], Haas Naemi[1], Baghdasaryan Astghik[1,2], Bruderer Tobias[1,3], Inci Demet[1], Micic Srdjan[1], Perkins Nathan[4], Spinas Renate[1], Zenobi Renato[3], Moeller Alexander[1]

[1]Division of Respiratory Medicine and Childhood Research Center, University Children's Hospital Zurich, Switzerland, [2]Joint Medical Center Arabkir, Division of Pulmonology, Yerevan, Armenia, [3]ETH Zurich, Department of Chemistry and Applied Bioscience, Zurich, Switzerland, [4]Division of Clinical Chemistry and Biochemistry, University Children's Hospital Zurich, Switzerland

For the Paediatric Exhalomics Group at the University Children's Hospital Zurich, Switzerland
Baghdasaryan Astghik, Berger Christoph, Bieli Christian, Bruderer Tobias, Haas Naemi, Hersberger Martin, Heschl Katharina, Inci Demet, Jung Andreas, Kohler Malcolm, Micic Srdjan, Moeller Alexander, Müller Simona, Perkins Nathan, Schürmann Tina, Singer Florian, Spinas Renate, Streckenbach Bettina, Usemann Jakob, Weber Ronja, Zenobi Renato.

Table of contents

Clinical data

Clinical data of the CF patients was extracted from their clinical records and included lung function testing, sex, height, weight, current medication and bacterial colonization of the lung

based on analysis of sputum or throat swab. In most patients clinical data was obtained at the same day as the study visit. Due to organisational issues, the clinical visit was apart from the study visit in some of the patients, but not exceeding more than 14 days. For healthy children, a questionnaire on respiratory health was applied and clinical data was collected during the study visit. In order to exclude any undiagnosed lung disease or allergy, healthy participants performed spirometry, Fractional exhaled Nitric Oxide (FeNO) measurement and a skin-prick allergy test including the 6 most frequent allergens.

Breath analysis

During the breathing maneuver, the pressure of exhalation was indicated on a manometer and was to be kept as constant as possible at around 4-5 mbar. The range of pressure variation was dependent on the participant's age and more fluctuation was accepted for younger children. The electrospray fluid was composed of ultrapure water with 0.1 % (v/v) formic acid (LiChrosolv®, Sigma-Aldrich Chemie GmbH, Buchs Switzerland). The sampling line, core and curtain gas for the ionization source were heated to 130 °C. The flow at the exhaust of the ionization source was measured by a mass flow controller (F-201EV, Bronkhorst, Ruurlo, Netherlands).

Data Pre-Processing

Data files were recorded in .wiff format by Analyst (Version TF 1.7, Applied Biosystems Sciex, Toronto, ON, Canada). The mass spectra in each data file were aligned in PeakView (Version 2.2, Applied Biosystems Sciex, Toronto, ON, Canada) with respect to the exact masses m/z = +77.05971, +100.07569, +371.10124 and +445.12004 in the positive mode and m/z = -255.23295 and -283.26425 in the negative mode. The files were subsequently converted into .mzXML format with MSConvert (Version 3, ProteoWizard Tools, Palo Alto, CA, US) from which the mass spectra were imported into Matlab (Version R2017b, The MathWorks Inc., Natick, MA, US). The mass spectra were resampled per subject using piecewise cubic Hermite interpolation [1] onto a linearly spaced m/z-axis with resolution of 0.0005 Da (9x10^8 data points, 50-500 m/z range).

For each subject, a feature list was created by detecting peaks of the spectrum by taking the maximum intensity for each m/z value over all scans. The resulting m/z features were then combined over all subjects into a single list from which kernel density estimate (Gaussian kernel and smoothing bandwidth parameter = 0.003) was computed. Subsequently, local maxima of the smoothed density function were used to define the feature list representative for all subjects. The feature list was then reduced by applying the following steps: For each subject, a subset of features was selected by running linear regression with standardized TIC over all scans as a predictor and removing those features with slope in linear regression < 0. In this way, only features with the same pattern as breath strokes were selected. Additionally, deviation of the sum of squares over all scans of predicted values of linear regression to the sum of squares of standardized TIC was used for further reduction. Furthermore, the deviation of the sum over all scans of absolute feature intensities from the sum of absolute standardized TIC values allowed

an even further decrease. As a final filter, m/z features which appear in at least 30% of the subjects were selected for later analysis.
For each subject, scans corresponding to exhaled breath were extracted using standardized TIC values greater than 0. Every feature was then integrated over a predefined m/z window (+/- 0.0025) in each scan corresponding to exhaled breath, divided by the integrated spectrum (i.e. normalized to the TIC) and averaged over the number of the corresponding scans. Finally, the normalized features were then arranged into an n x k intensity matrix where n is the number of subjects and k is the number of features. In our case, the above feature selection procedure resulted in k = 3468 features found across n=101 samples.

Data Analysis

Data analysis was performed in R (version 3.4.4). Prior to statistical analysis batch correction was applied using proposed algorithm in [2]. Two known batches were defined based on the exchange of the team who was performing measurements. The Mann–Whitney U-test [3] was performed on the batch corrected subject-feature matrix using the normalized intensities contained in the columns and the labels as outcome in the rows. The Mann–Whitney U-test was chosen because it could not be assumed that the distributions of intensities were in general Gaussian. Since multiple tests were performed (for each of the more than 3000 features) the Benjamini–Hochberg method was used for multiple testing corrections. The test was considered significant if the corrected p-values (FDR adjusted p-values) were below the threshold of 0.05. The (two-sample) Hodges–Lehmann estimator [4] was used to estimate the difference between the healthy and CF population. The significant m/z features can be found in the table S2 (for better readability the intensities were scaled using distance to median divided by the median absolute deviation). Principal component analysis (PCA) was performed on the set of the top 9 features (ordered by the corrected p-values).
The data processing pipeline consisting of 1. data preprocessing (see above), 2. removal of the known batch effects, 3. stability selection [5] in conjunction with the Mann-Whitney U-test, was chosen as a feature selection procedure prior to predictive analysis. More precisely, in step 3, variable selection was defined as the selection of all features with p-value (Mann-Whitney U test) below a given quantile of all obtained p-values. Technically, following the outlined algorithm by Meinshausen and Bühlmann [5], we used the p-value quantiles as regularization parameters of variable selection. The range of regularization parameters for stability selection was set to the single element equal to the 0.2-quantile of obtained p-values arising from each subsample (see *pointwise control* in [5]). The number of random subsamples was set to 50 with variable selection applied to each subsample. The selection probability threshold was set to 0.9 so that the features which appear in at least 90% of the subsamples were collected. For more details on stability selection the reader is referred to the seminal work of Meinshausen and Bühlmann [5].
Linear support-vector machines (SVM) [6] was chosen as a supervised algorithm since it is less sensitive to the number of dimensions of the predictor set. To assess the performance of the prediction and to optimize the soft-margin constant of SVM we decided to follow the method proposed in [7]. That is, we used nested cross-validation, where the inner loop (10-fold cross-validation) was used to search for the best soft-margin constant of SVM and the outer loop (10-

fold cross-validation) was used to estimate the generalization error. The procedure was repeated 25 times (25 times repeated 10-fold cross-validation). It is imperative to remark that in any kind of splitting of samples into training and testing sets, as done in cross-validation, feature selection has to be performed every time on each training set and hence in each step of cross-validation [7-9]. Also, when applying batch correction to any hold-out set (i.e. testing data set) the batch labels of the batch corrected training data set have to be frozen and treated as a reference batch for the new batch correction of the merged training and testing data set.
The procedure above resulted in an average accuracy of 72.1% with an average sensitivity of 77.2% and an average specificity of 67.7%. The final model, applicable to unseen data, was trained on the complete data set where the hyperparameter of SVM was tuned in a 10-fold cross-validation. The number of selected variables arising from the feature selection used for the final model was 81 (see table S2 for their selection probability). The distribution of the prediction accuracies over all cross-validation rounds can be found in figure S2 and the ROC curves including the average ROC curve of all cross-validation rounds in the figure S3.

Formula Annotation

A database with possible molecular formulas was created. The exact mass value for the different elements were taken from the literature [10]. The mass range was matched to the acquired mass region from *m/z* = 50 - 500. The selected elements and their maximal numbers were based on the seven golden rules recommendations from Kind and Fiehn [11]. The recommended elements were modified to allow for water clusters and adjusted for carbon/hydrogen ratios >= 0.2 - 4.0 and carbon/oxygen ratios <= 2.0 and Ring Double Bond Equivalents (RDBE) from -4 to 40. The elements were further constrained to only allow for elemental numbers commonly detected with SESI-MS resulting in the following elemental constraints: 39 carbon, 72 hydrogen, 20 oxygen, 4 nitrogen and 2 sulfur. Two databases were created, the first excluding C13 isotopes and the second database with C13 isotopes. The allowed overlap between the measured m/z feature and the exact mass of the proposed formula was +/- 10 ppm. If more than three formulae were possible for any certain m/z feature, the last formula was annotated with a “*”. The second database including C13 isotopes was only consulted when no formula could be annotated within the first formula database. It is important to keep in mind that the annotated formula corresponds to the m/z feature and not the molecule itself. The most common species for SESI-HRMS is either a protonated or deprotonated adduct species resulting in a formula with either a singular plus or minus hydrogen element. Using our automated workflow, only 9 out of the 171 CF features could not be annotated with a putative molecular formula (see table S2). These compounds were investigated in further detail. The m/z feature -74.7773 has been proven to be a satellite peak of +75.0084 based on the MS/MS fragments with identical mass differences. The m/z features -79.9395, -95.9520 and +460.6872 are also most likely satellite peaks due to their unusual mass defects. However, their intensity was too low for MS/MS confirmation. The Pearson’s correlation coefficient between -75.0182 and -75.0084 is 0.87. Furthermore the peak of m/z value -75.0182 has a much lower intensity than -75.0084 which make it prone to overlap issues. It might be the same molecule as the feature -75.0182. The masses -94.0260, -108.0060, +81.0175, +248.0075 could not be annotated with a molecular formula with the selected elemental constraints. It might be possible

to annotate them if more rare elements or higher elemental number constraints are allowed. They might be present as a different adduct species or as a loss. For the mass -122.0195 a formula with a C13 isotope is likely with $C(C13)H_6O_5$ related to -121.0143 with $C_3H_6O_5$. The Pearson's correlation coefficient between -122.0195 and -121.0143 is 0.83 and supports this assumption. A higher mass error of 14.8 ppm was accepted for this case.

Table S1: Distribution of mutations in CF population

Mutation	Mutation	CFTR-function*	Number
F508del	F508del	mf/mf	27
F508del	3272-26 A>G	mf/rf	1
F508del	G542X	mf/mf	2
F508del	S549R(T>G)	mf/mf	2
F508del	Y1092X	mf/rf	2
F508del	R553X	mf/mf	1
F508del	R1066C	mf/mf	1
F508del	R1158X	mf/mf	2
F508del	2789+5G>A	mf/rf	2
F508del	621+1G>T	mf/mf	1
F508del	C524X	mf/mf	2
F508del	S341P	mf/rf	1
F508del	N1303K	mf/mf	1
F508del	2183AA->G	mf/mf	1
405+1G-A	3905insT	mf/mf	1
G542X	R347P	mf/mf	1
Q525X	Q525X	mf/mf	1
L1040P	L1040P	mf/mf	1
15251G>A	1525-1G->A	mf/mf	1
1717-G<A	711+5G<A	mf/mf	1

* mf= minimal CFTR-function; rf = residual CFTR function

Table S2: CF specific m/z features

171 CF specific m/z features order by FDR-adjusted p-value (q-value) including annotated molecular formula. Intensities were previously scaled to difference to the median divided by median absolute deviation for better readability. m/z features which were not selected by the feature selection procedure have the selection probability value replaced by “-”. If more than 3 formulae were assigned, the third formula was annotated with “*”. 9 m/z features could not be annotated with a molecular formula with the automatic workflow. These values were moved to the end of the table. H.L. est. = Hodges-Lehmann estimator. S.P. = Selection Probability (%).

m/z value	p-value	q-value	H.L. est.	95% CI	S.P.	Putative molecular formula / mass error (ppm)

-151.0247	1.4E-08	1.2E-05	1.785	[1.174 , 2.624]	100%	C4H8O6	-0.7	C5H12OS2	-6.5	C5H4N4O2	-9.5
-75.0085	1.8E-08	1.2E-05	0.991	[0.692 , 1.557]	100%	C2H4O3	-3.5				
-121.0143	2.0E-08	1.2E-05	1.216	[0.803 , 1.874]	100%	C3H6O5	0.4	C4H10S2	-6.7		
-122.0195	3.5E-08	1.5E-05	1.283	[0.774 , 1.888]	100%	C(C13)H6O5	14.8				
+297.0825	1.1E-06	3.9E-04	1.110	[0.655 , 1.643]	100%	C11H20O5S2	0.0	C10H16O10	3.0	C11H12N4O6*	-1.6
+445.1200	1.8E-06	5.9E-04	1.197	[0.627 , 2.033]	100%	C16H20N4O11	-0.3	C16H28O10S2	0.7	C17H24N4O6S2*	-2.3
+359.0462	4.0E-06	1.1E-03	0.819	[0.502 , 1.289]	100%	C11H18O9S2	-0.8	C10H14O14	1.6	C11H10N4O10*	-2.1
+445.0985	4.4E-06	1.1E-03	1.162	[0.584 , 1.968]	100%	C19H24O8S2	-0.1	C18H20O13	1.9	C19H16N4O9*	-1.1
-93.0195	4.7E-06	1.1E-03	1.074	[0.637 , 1.623]	100%	C2H6O4	1.8				
+357.0490	7.2E-06	1.5E-03	1.014	[0.561 , 1.615]	98%	C11H16O11S	1.1	C24H8N2S	2.5	C12H12N4O7S*	-2.7
+447.1420	9.6E-06	1.9E-03	0.932	[0.476 , 1.541]	100%	C9H26N4O16	0.8	C30H22O2S	1.5	C10H30N4O11S2*	-1.2
+332.1202	1.2E-05	2.0E-03	1.153	[0.639 , 1.754]	100%	C11H25NO6S2	1.8	C10H21NO11	4.4	C20H17N3S*	-4.2
-105.0188	1.2E-05	2.0E-03	0.917	[0.506 , 1.364]	100%	C3H6O4	-5.0				
+429.0880	1.3E-05	2.0E-03	0.922	[0.507 , 1.472]	100%	C15H24O10S2	-0.9	C14H20O15	1.2	C15H16N4O11*	-1.9
+445.1483	1.3E-05	2.0E-03	0.809	[0.457 , 1.338]	100%	C28H20N4S	0.3	C27H24O4S	3.4	C15H28N2O11S*	-0.8
+188.1645	1.5E-05	2.3E-03	-0.799	[-1.153 , -0.443]	100%	C10H21NO2	0.0				
+447.0983	2.3E-05	3.3E-03	0.889	[0.464 , 1.515]	100%	C28H18N2S2	-0.3	C14H22O16	0.5	C15H26O11S2*	-1.4
+299.0797	2.5E-05	3.3E-03	0.930	[0.548 , 1.492]	98%	C10H18O8S	0.6	C14H10N4O4	7.4	C11H14N4O4S*	-3.9
+359.0285	3.0E-05	3.8E-03	0.872	[0.463 , 1.436]	100%	C10H14O12S	1.7	C23H6N2OS	3.2	C11H10N4O8S*	-2.0
+175.0435	3.2E-05	3.9E-03	1.105	[0.591 , 1.735]	98%	C7H10O3S	6.7				
+144.1380	4.0E-05	4.7E-03	-0.739	[-1.206 , -0.389]	98%	C8H17NO	-2.0				
+481.1563	5.2E-05	5.6E-03	1.025	[0.476 , 1.769]	96%	C20H24N4O10	-0.5	C20H32O9S2	0.5	C21H28N4O5S2*	-2.3
+299.0620	5.4E-05	5.6E-03	0.781	[0.39 , 1.299]	96%	C10H10N4O7	-0.8	C10H18O6S2	0.8	C11H14N4O2S2*	-3.7
+463.1205	5.4E-05	5.6E-03	1.004	[0.484 , 1.833]	96%	C18H26N2O8S2	0.4	C17H22N2O13	2.2	C26H22O6S*	-1.1
-133.0860	5.5E-05	5.6E-03	-0.829	[-1.312 , -0.49]	100%	C6H14O3	-7.6				
+430.1075	6.4E-05	6.3E-03	0.854	[0.454 , 1.282]	90%	C28H15NO4	0.3	C20H19N3O6S	1.8	C16H19N3O11*	-4.0
+148.0967	7.6E-05	7.2E-03	-0.787	[-1.174 , -0.413]	98%	C6H13NO3	-0.8				
+225.0428	1.0E-04	8.9E-03	0.769	[0.372 , 1.202]	96%	C7H12O6S	0.3	C11H4N4O2	9.4	C16H4N2	-8.6
+301.0575	1.0E-04	8.9E-03	0.819	[0.404 , 1.369]	98%	C14H12N4S2	-0.4	C13H8N4O5	2.5	C9H16O9S*	-4.3
-359.2798	1.0E-04	8.9E-03	0.678	[0.257 , 1.507]	96%	C20H40O5	-1.4	C24H40S	5.6	C16H44N2O2S2	7.4
+232.1905	1.1E-04	8.9E-03	-0.772	[-1.165 , -0.373]	98%	C12H25NO3	-1.0				
+225.0610	1.1E-04	8.9E-03	0.955	[0.497 , 1.475]	98%	C8H16O3S2	-1.6	C7H12O8	2.3	C8H8N4O4	-3.7
+108.1020	1.1E-04	8.9E-03	0.895	[0.417 , 1.537]	96%	C4H13NO2	0.9				
+465.1268	1.3E-04	9.7E-03	0.864	[0.424 , 1.5]	92%	C28H20N2O3S	0.1	C20H24N4O5S2	1.5	C15H28O14S*	-1.0
+342.9960	1.3E-04	1.0E-02	0.777	[0.37 , 1.253]	92%	C9H10O12S	-1.7	C14H6N4O3S2	1.7	C10H6N4O8S*	-5.6
-137.0090	1.6E-04	1.2E-02	0.719	[0.351 , 1.082]	94%	C3H6O6	-1.2	C4H10OS2	-7.5		
+60.0808	1.6E-04	1.2E-02	-0.755	[-1.128 , -0.369]	98%	C3H9N	0.4				
+190.1438	1.6E-04	1.2E-02	-0.818	[-1.22 , -0.445]	94%	C9H19NO3	0.2				
+237.0450	1.7E-04	1.2E-02	-0.938	[-1.59 , -0.415]	96%	C17H4N2	1.2	C9H8N4O2S	3.9	C5H16O6S2*	-4.7
+361.0070	1.9E-04	1.3E-02	0.829	[0.41 , 1.308]	92%	C9H12O13S	-0.4	C14H8N4O4S2	2.8	C10H8N4O9S*	-4.1
+176.1275	2.0E-04	1.3E-02	-0.642	[-1.017 , -0.324]	96%	C8H17NO3	-3.5				
+480.1563	2.0E-04	1.3E-02	0.879	[0.38 , 1.469]	92%	C29H25N3S2	0.1	C15H29NO16	0.8	C16H33NO11S2*	-1.0
-149.0093	2.0E-04	1.3E-02	0.811	[0.38 , 1.311]	90%	C4H6O6	0.9	C5H10OS2	-4.9	C5H2N4O2	-8.0

+192.1595	2.2E-04	1.4E-02	-0.745	[-1.137 , -0.358]	96%	C9H21NO3	0.4				
+148.1330	2.3E-04	1.4E-02	-0.662	[-1.021 , -0.326]	96%	C7H17NO2	-1.4				
+247.0118	2.4E-04	1.4E-02	0.870	[0.408 , 1.402]	92%	C5H10O9S	-0.1	C10H6N4S2	4.6	C6H6N4O5S*	-5.6
-165.0043	2.4E-04	1.4E-02	0.810	[0.39 , 1.261]	90%	C4H6O7	1.3	C5H10O2S2	-3.9	C5H2N4O3	-6.7
+226.0410	2.5E-04	1.4E-02	0.803	[0.372 , 1.263]	92%	C15H3N3	4.6				
+160.0965	2.5E-04	1.4E-02	-0.716	[-1.054 , -0.326]	98%	C7H13NO3	-2.0				
-171.1028	2.6E-04	1.4E-02	-0.818	[-1.256 , -0.39]	94%	C9H16O3	0.8				
+202.1075	2.7E-04	1.4E-02	-0.712	[-1.07 , -0.345]	96%	C9H15NO4	0.6				
+175.1148	2.7E-04	1.4E-02	-0.696	[-1.073 , -0.327]	92%	C9H18OS	-1.8				
+340.2482	2.8E-04	1.5E-02	-0.939	[-1.609 , -0.39]	98%	C19H33NO4	-0.1	C11H37N3O6S	1.8	C15H37N3OS2	9.2
+174.1123	2.9E-04	1.5E-02	-0.697	[-1.103 , -0.321]	-	C8H15NO3	-1.0				
-125.0105	3.2E-04	1.6E-02	0.844	[0.385 , 1.291]	-	C3H10OS2	3.7				
+344.9755	3.3E-04	1.6E-02	0.790	[0.329 , 1.296]	94%	C8H8O13S	-1.0	C13H4N4O4S2	2.4	C9H4N4O9S*	-4.9
+256.1902	3.5E-04	1.7E-02	-0.820	[-1.295 , -0.372]	100%	C14H25NO3	-2.0				
+204.1595	3.7E-04	1.8E-02	-0.712	[-1.125 , -0.337]	92%	C10H21NO3	0.4				
+286.2010	3.8E-04	1.8E-02	-0.761	[-1.184 , -0.362]	90%	C15H27NO4	-1.0				
+146.1175	3.9E-04	1.8E-02	-0.653	[-1.007 , -0.299]	-	C7H15NO2	-0.4				
+276.1807	4.0E-04	1.8E-02	-0.734	[-1.062 , -0.326]	96%	C13H25NO5	0.5	C17H25NS	9.6	C14H29NS2	-2.6
+163.0965	4.0E-04	1.8E-02	0.779	[0.371 , 1.317]	90%	C7H14O4	0.1	C8H10N4	-8.2		
+246.2063	4.3E-04	1.8E-02	-0.727	[-1.112 , -0.347]	94%	C13H27NO3	-0.3				
+464.1230	4.3E-04	1.8E-02	0.659	[0.279 , 1.035]	-	C18H25NO11S	1.9	C31H17N3S	3.0	C27H17N3O5*	-2.4
+202.1620	4.6E-04	1.9E-02	-0.716	[-1.111 , -0.318]	-	C11H23NS	-2.0				
+371.1237	4.6E-04	1.9E-02	0.733	[0.342 , 1.136]	92%	C19H18N2O6	-0.2	C11H22N4O8S	1.6	C20H22N2OS2*	-2.5
+204.1230	4.8E-04	2.0E-02	-0.666	[-1.034 , -0.314]	-	C9H17NO4	-0.2				
+232.1540	5.0E-04	2.0E-02	-0.712	[-1.17 , -0.267]	92%	C11H21NO4	-1.4				
+162.1485	5.0E-04	2.0E-02	-0.719	[-1.076 , -0.334]	-	C8H19NO2	-2.2				
+176.1640	5.0E-04	2.0E-02	-0.762	[-1.193 , -0.339]	-	C9H21NO2	-2.9				
+346.1862	5.1E-04	2.0E-02	-0.753	[-1.157 , -0.34]	94%	C16H27NO7	0.5	C8H31N3O9S	2.4	C17H31NO2S2*	-2.0
+193.1242	5.1E-04	2.0E-02	-0.670	[-1.061 , -0.312]	92%	C9H20O2S	-7.7	C12H16O2	9.9		
+230.1392	5.4E-04	2.0E-02	-0.746	[-1.135 , -0.345]	96%	C11H19NO4	2.2				
-389.0755	5.8E-04	2.2E-02	0.693	[0.308 , 1.113]	90%	C25H14N2OS	0.2	C17H18N4O3S2	1.9	C12H22O12S*	-1.1
-115.0763	6.1E-04	2.2E-02	-0.620	[-0.955 , -0.291]	96%	C6H12O2	-1.3				
+132.1015	6.1E-04	2.2E-02	-0.623	[-0.992 , -0.265]	-	C6H13NO2	-3.1				
+234.1335	6.2E-04	2.2E-02	-0.651	[-1.063 , -0.288]	-	C10H19NO5	-0.4	C11H23NS2	-4.2		
+247.1722	6.2E-04	2.2E-02	-0.682	[-1.057 , -0.315]	-	C13H26O2S	-1.7				
-164.0205	6.7E-04	2.4E-02	0.756	[0.316 , 1.329]	90%	C5H11NOS2	-2.6	C4H7NO6	2.7		
+200.1285	6.9E-04	2.4E-02	-0.627	[-1.033 , -0.284]	-	C10H17NO3	1.9				
+246.1490	7.2E-04	2.5E-02	-0.660	[-1.038 , -0.299]	-	C15H19NO2	0.6	C7H23N3O4S	3.2		
+176.0913	7.2E-04	2.5E-02	-0.684	[-1.084 , -0.289]	92%	C7H13NO4	-2.5				
+216.1235	7.4E-04	2.5E-02	-0.610	[-0.964 , -0.274]	92%	C10H17NO4	2.2				
+262.2010	7.4E-04	2.5E-02	-0.693	[-1.061 , -0.297]	90%	C13H27NO4	-1.1				
+234.1697	7.8E-04	2.5E-02	-0.638	[-1.038 , -0.265]	-	C11H23NO4	-1.2				
+162.1123	7.8E-04	2.5E-02	-0.576	[-0.92 , -0.242]	94%	C7H15NO3	-1.1				

+245.0275	8.2E-04	2.6E-02	0.833	[0.365 , 1.408]	-	C13H8O3S	3.3	C5H12N2O5S2	6.0	C9H8O8	-6.9
+244.1540	8.4E-04	2.6E-02	-0.673	[-1.024 , -0.296]	-	C12H21NO4	-1.4				
+190.1075	8.4E-04	2.6E-02	-0.644	[-1.017 , -0.27]	92%	C8H15NO4	0.6				
+462.1462	8.6E-04	2.7E-02	0.595	[0.265 , 1.018]	-	C16H31NO10S2	0.0	C15H27NO15	1.9	C25H23N3O4S*	-4.3
+232.1180	8.8E-04	2.7E-02	-0.719	[-1.086 , -0.345]	-	C10H17NO5	0.2	C11H21NS2	-3.5		
+218.1750	8.8E-04	2.7E-02	-0.657	[-0.995 , -0.266]	-	C11H23NO3	-0.3				
+212.2010	9.0E-04	2.7E-02	-0.700	[-1.219 , -0.239]	-	C13H25NO	0.5				
+226.1430	9.2E-04	2.7E-02	-0.602	[-0.919 , -0.254]	90%	C12H19NO3	-3.4				
+274.2380	9.2E-04	2.7E-02	-0.660	[-1.081 , -0.276]	-	C15H31NO3	1.2				
+212.1635	9.9E-04	2.9E-02	-0.692	[-1.085 , -0.293]	90%	C12H21NO2	-4.8				
+228.1598	1.0E-03	2.9E-02	-0.674	[-1.052 , -0.29]	-	C12H21NO3	1.7				
+373.0807	1.0E-03	2.9E-02	0.622	[0.252 , 1.04]	-	C13H16N4O7S	-1.5	C12H20O11S	2.1	C21H12N2O5*	-3.2
+206.1750	1.1E-03	3.0E-02	-0.703	[-1.127 , -0.298]	90%	C10H23NO3	-0.3				
+260.1855	1.1E-03	3.0E-02	-0.699	[-1.145 , -0.294]	-	C13H25NO4	-0.5				
+260.2220	1.1E-03	3.0E-02	-0.731	[-1.237 , -0.302]	-	C14H29NO3	-0.1				
+232.0788	1.1E-03	3.1E-02	-0.646	[-1.047 , -0.253]	-	C13H13NOS	-1.1	C5H17N3O3S2	1.7	C4H13N3O8	5.4
+178.1068	1.2E-03	3.2E-02	-0.621	[-1.003 , -0.251]	-	C7H15NO4	-3.3				
+162.0757	1.2E-03	3.2E-02	-0.731	[-1.282 , -0.293]	-	C6H11NO4	-2.4				
-154.0507	1.2E-03	3.2E-02	-0.665	[-1.097 , -0.263]	-	C7H9NO3	-1.7				
+362.9865	1.2E-03	3.2E-02	0.706	[0.184 , 1.145]	-	C8H10O14S	0.3	C13H6N4O5S2	3.5	C9H6N4O10S*	-3.4
+354.2632	1.3E-03	3.2E-02	-1.048	[-2.292 , -0.352]	-	C12H39N3O6S	-0.1	C16H39N3OS2	7.0	C20H35NO4*	-1.9
+276.2167	1.3E-03	3.2E-02	-0.804	[-1.241 , -0.331]	-	C14H29NO4	-0.9				
+172.1693	1.4E-03	3.5E-02	-0.632	[-1.064 , -0.241]	-	C10H21NO	-1.7				
+217.1958	1.4E-03	3.5E-02	0.713	[0.244 , 1.318]	-	C16H24	3.3	C8H28N2O2S	6.4		
+245.1527	1.4E-03	3.5E-02	-0.643	[-1.006 , -0.27]	-	C8H24N2O4S	-1.0	C16H20O2	-3.7		
+218.1388	1.5E-03	3.5E-02	-0.586	[-1.002 , -0.217]	-	C10H19NO4	0.5				
+286.1653	1.5E-03	3.5E-02	-0.664	[-1.075 , -0.274]	-	C14H23NO5	1.4	C15H27NS2	-1.6		
+290.1603	1.5E-03	3.5E-02	-0.593	[-1.007 , -0.271]	-	C14H27NOS2	-1.3	C13H23NO6	1.7	C10H27NO6S	-10
+462.1768	1.5E-03	3.5E-02	0.625	[0.254 , 1.008]	-	C24H31NO4S2	0.2	C23H27NO9	2.0	C11H31N3O16*	-2.0
+391.1095	1.5E-03	3.5E-02	0.786	[0.273 , 1.5]	-	C13H18N4O10	-0.2	C13H26O9S2	1.0	C14H22N4O5S2*	-2.4
+260.1492	1.5E-03	3.5E-02	-0.595	[-0.953 , -0.213]	-	C12H21NO5	-0.2	C16H21NS	9.5	C13H25NS2	-3.6
+246.1700	1.5E-03	3.5E-02	-0.736	[-1.236 , -0.314]	-	C12H23NO4	0.1				
+233.1537	1.5E-03	3.5E-02	-0.644	[-1.021 , -0.295]	-	C15H20O2	0.4	C7H24N2O4S	3.2		
+304.2115	1.5E-03	3.5E-02	-0.649	[-1.048 , -0.274]	-	C15H29NO5	-1.2	C19H29NS	7.1	C16H33NS2*	-4.0
+206.1870	1.5E-03	3.5E-02	-0.625	[-0.993 , -0.264]	-	C9H23N3O2	3.4				
+177.0895	1.6E-03	3.7E-02	-0.588	[-0.958 , -0.224]	-	C11H12O2	-8.6				
+274.1653	1.6E-03	3.7E-02	-0.641	[-1.043 , -0.259]	-	C13H23NO5	1.5	C14H27NS2	-1.7		
+334.1517	1.8E-03	3.9E-02	-0.578	[-0.927 , -0.223]	-	C15H27NO3S2	3.6	C14H23NO8	6.2	C11H27NO8S*	-3.9
+160.1330	1.8E-03	3.9E-02	-0.670	[-1.057 , -0.277]	-	C8H17NO2	-1.3				
+259.1717	1.9E-03	4.2E-02	-0.771	[-1.299 , -0.298]	-	C14H26O2S	-3.6	C17H22O2	9.5		
+336.1655	2.0E-03	4.3E-02	-0.715	[-1.208 , -0.25]	-	C14H25NO8	0.6	C18H25NO3S	8.1	C15H29NO3S2	-2.0
+152.0677	2.0E-03	4.3E-02	-0.628	[-1.044 , -0.204]	-	C3H9N3O4	7.4				
+106.0858	2.0E-03	4.3E-02	0.704	[0.248 , 1.257]	-	C4H11NO2	-4.3				

+214.0697	2.1E-03	4.4E-02	-0.711	[-1.208 , -0.308]	-	C13H11NS	5.6	C5H15N3O2S2	8.7	C9H11NO5	-6.1
+283.0337	2.1E-03	4.4E-02	0.661	[0.242 , 1.113]	-	C6H10N4O7S	-2.1	C18H6N2S	4.4	C14H6N2O5	-4.4
+242.1382	2.2E-03	4.4E-02	-0.602	[-1.012 , -0.242]	-	C12H19NO4	-2.0				
+270.1700	2.2E-03	4.4E-02	-0.592	[-0.98 , -0.213]	-	C14H23NO4	0.1				
+120.1017	2.2E-03	4.4E-02	-0.559	[-0.921 , -0.227]	-	C5H13NO2	-1.7				
+196.0948	2.2E-03	4.4E-02	-0.527	[-0.895 , -0.18]	-	C6H17N3S2	5.8				
+344.1683	2.2E-03	4.4E-02	-0.601	[-0.996 , -0.2]	-	C20H25NO2S	1.2	C12H29N3O4S2	3.1	C8H29N3O9S*	-4.2
+360.1802	2.2E-03	4.4E-02	-0.593	[-0.987 , -0.237]	-	C12H29N3O7S	0.8	C24H25NS	6.0	C20H25NO5*	-1.0
+318.2210	2.2E-03	4.4E-02	-0.755	[-1.213 , -0.274]	-	C15H31N3O2S	0.1	C23H27N	-2.0	C11H31N3O7	-7.8
+364.2095	2.2E-03	4.4E-02	-0.793	[-1.308 , -0.319]	-	C24H29NS	0.4	C16H33N3O2S2	2.2	C12H33N3O7S*	-4.7
+218.1028	2.3E-03	4.5E-02	-0.633	[-1.106 , -0.252]	-	C10H19NS2	-1.7	C9H15NO5	2.3		
+220.1180	2.3E-03	4.5E-02	-0.568	[-0.947 , -0.201]	-	C9H17NO5	0.2	C10H21NS2	-3.7		
+230.2117	2.4E-03	4.6E-02	-0.775	[-1.256 , -0.255]	-	C13H27NO2	1.1				
+298.1645	2.4E-03	4.6E-02	-0.563	[-0.935 , -0.235]	-	C7H27N3O7S	0.8	C19H23NS	7.1	C15H23NO5*	-1.3
+149.0598	2.4E-03	4.7E-02	-0.638	[-1.079 , -0.232]	-	C9H8O2	0.6				
+211.1688	2.5E-03	4.7E-02	-0.760	[-1.207 , -0.25]	-	C13H22O2	-2.2				
+146.0813	2.5E-03	4.7E-02	-0.597	[-1.03 , -0.231]	-	C6H11NO3	0.9				
+290.1960	2.5E-03	4.7E-02	-0.624	[-1.033 , -0.201]	-	C14H27NO5	-0.7	C18H27NS	8.0	C15H31NS2	-3.7
+244.2270	2.5E-03	4.7E-02	-0.563	[-0.983 , -0.211]	-	C14H29NO2	-0.4				
+188.1280	2.5E-03	4.7E-02	-0.580	[-0.96 , -0.229]	-	C9H17NO3	-0.6				
+182.1170	2.6E-03	4.8E-02	-0.588	[-0.929 , -0.218]	-	C10H15NO2	-3.1				
+332.2448	2.6E-03	4.8E-02	-0.617	[-1.029 , -0.229]	-	C18H37NS2	2.4	C17H33NO5	5.0	C14H37NO5S*	-5.2
+201.1268	2.6E-03	4.8E-02	-0.621	[-1.01 , -0.229]	-	C6H20N2O3S	0.3	C14H16O	-3.0		
+202.1440	2.7E-03	4.9E-02	-0.631	[-1.046 , -0.254]	-	C10H19NO3	1.1				
+262.1648	2.7E-03	4.9E-02	-0.608	[-0.98 , -0.22]	-	C12H23NO5	-0.4	C16H23NS	9.2	C13H27NS2	-3.7
+227.0308	2.8E-03	5.0E-02	0.556	[0.213 , 0.951]	-	C9H10N2OS2	0.3	C8H6N2O6	4.1		
-199.1340	2.8E-03	5.0E-02	-0.555	[-0.85 , -0.181]	-	C11H20O3	0.2				
+362.1825	2.8E-03	5.0E-02	-0.614	[-0.991 , -0.211]	-	C17H31NO3S2	1.9	C16H27NO8	4.3	C13H31NO8S*	-5.0
+248.9913	2.8E-03	5.0E-02	0.686	[0.241 , 1.127]	-	C5H4N4O6S	-4.6	C9H4N4OS2	5.5		
+144.1022	2.8E-03	5.0E-02	-0.555	[-0.967 , -0.173]	-	C7H13NO2	2.1				
+246.1337	2.9E-03	5.0E-02	-0.590	[-0.99 , -0.201]	-	C11H19NO5	0.4	C12H23NS2	-3.1		
+178.0708	2.9E-03	5.0E-02	-0.670	[-1.146 , -0.224]	-	C6H11NO5	-1.1	C7H15NS2	-6.0		
+261.1835	2.9E-03	5.0E-02	-0.562	[-0.934 , -0.195]	-	C9H28N2O4S	-2.9	C17H24O2	-5.4		
-75.0182	8.8E-09	1.2E-05	1.041	[0.724 , 1.518]	100%	unassignable / probably identical with -75.0085					
-74.7773	1.2E-08	1.2E-05	1.083	[0.721 , 1.597]	100%	unassignable / satellite peak					
-108.0060	3.6E-08	1.5E-05	1.187	[0.776 , 1.773]	100%	unassignable / outside of elemental constraints					
-94.0260	2.1E-06	6.3E-04	0.960	[0.582 , 1.309]	96%	unassignable / outside of elemental constraints					
+248.0075	3.7E-04	1.8E-02	0.821	[0.375 , 1.343]	92%	unassignable / outside of elemental constraints					
-95.9520	9.5E-04	2.8E-02	-0.876	[-1.795 , -0.354]	-	unassignable / probably satellite peak					
+460.6872	1.2E-03	3.2E-02	0.575	[0.209 , 0.946]	-	unassignable / probably satellite peak					
-79.9395	1.6E-03	3.7E-02	-0.722	[-1.304 , -0.275]	-	unassignable / probably satellite peak					
+81.0175	1.7E-03	3.8E-02	-0.772	[-1.574 , -0.289]	-	unassignable / outside of elemental constraints					

Table S3: R squared scores - FEV1 & FVCs in relation with CF related features

The following table contains coefficients of determination (i.e. R^2 values) of the intensities of m/z features predicted from FEV1 and FVCs. All of the R^2 lie between 0 and 0.186 indicating rather low linear relationship.

m/z value	R squared	m/z value	R squared	m/z value	R squared	m/z value	R squared	m/z value	R squared
-75.0182	0.046	+225.0610	0.064	+371.1237	0.084	+260.1855	0.017	+120.1017	0.100
-74.7773	0.002	+108.1020	0.007	+204.1230	0.002	+260.2220	0.085	+196.0948	0.026
-151.0247	0.050	+465.1268	0.003	+162.1485	0.006	+232.0788	0.013	+242.1382	0.002
-75.0085	0.000	+342.9960	0.043	+176.1640	0.147	+178.1068	0.052	+270.1700	0.057
-121.0143	0.021	-137.0090	0.053	+232.1540	0.045	+162.0757	0.061	+344.1683	0.057
-122.0195	0.045	+60.0808	0.113	+193.1242	0.004	+460.6872	0.031	+318.2210	0.001
-108.0060	0.021	+190.1438	0.000	+346.1862	0.066	-154.0507	0.012	+360.1802	0.152
+297.0825	0.016	+237.0450	0.039	+230.1392	0.011	+362.9865	0.006	+364.2095	0.041
+445.1200	0.026	+361.0070	0.010	-389.0755	0.036	+276.2167	0.066	+218.1028	0.021
-94.0206	0.107	+176.1275	0.004	+132.1015	0.131	+354.2632	0.146	+220.1180	0.050
+359.0462	0.018	+480.1563	0.017	-115.0763	0.008	+172.1693	0.041	+230.2117	0.025
+445.0985	0.033	-149.0093	0.028	+234.1335	0.066	+217.1958	0.080	+298.1645	0.116
-93.0195	0.022	+192.1595	0.010	+247.1722	0.001	+245.1527	0.001	+149.0598	0.054
+357.0490	0.005	+148.1330	0.074	-164.0205	0.013	+218.1388	0.050	+146.0813	0.031
+447.1420	0.122	+247.0118	0.051	+200.1285	0.031	+286.1653	0.022	+211.1688	0.094
+332.1202	0.018	-165.0043	0.023	+176.0913	0.076	+290.1603	0.013	+188.1280	0.007
-105.0188	0.016	+226.0410	0.003	+246.1490	0.037	+462.1768	0.049	+244.2270	0.051
+429.0880	0.013	+160.0965	0.027	+216.1235	0.057	+391.1095	0.086	+290.1960	0.036
+445.1483	0.006	-171.1028	0.015	+262.2010	0.004	+206.1870	0.014	+182.1170	0.044
+188.1645	0.039	+175.1148	0.025	+162.1123	0.020	+233.1537	0.009	+332.2448	0.060
+447.0983	0.182	+202.1075	0.125	+234.1697	0.005	+246.1700	0.001	+201.1268	0.025
+299.0797	0.008	+340.2482	0.186	+245.0275	0.155	+260.1492	0.053	+202.1440	0.027
+359.0285	0.023	+174.1123	0.127	+190.1075	0.102	+304.2115	0.002	+262.1648	0.044
+175.0435	0.085	-125.0105	0.007	+244.1540	0.026	+177.0895	0.065	+227.0308	0.028
+144.1380	0.112	+344.9755	0.007	+462.1462	0.009	+274.1653	0.007	+144.1022	0.069

+481.1563	0.024	+256.1902	0.061	+218.1750	0.001	-79.9395	0.047	+248.9913	0.034
+299.0620	0.042	+204.1595	0.109	+232.1180	0.016	+81.0175	0.028	+362.1825	0.049
+463.1205	0.002	+248.0075	0.027	+212.2010	0.095	+160.1330	0.022	-199.1304	0.058
-133.0860	0.017	+286.2010	0.040	+226.1430	0.125	+334.1517	0.034	+178.0708	0.059
+430.1075	0.018	+146.1175	0.003	+274.2380	0.007	+259.1717	0.018	+246.1337	0.061
+148.0967	0.013	+163.0965	0.003	-95.9502	0.114	+336.1655	0.040		
+225.0428	0.043	+276.1807	0.019	+212.1635	0.034	+106.0858	0.056		
+301.0575	0.015	+246.2063	0.110	+228.1598	0.020	+152.0677	0.136		
-359.2798	0.027	+464.1230	0.030	+373.0807	0.034	+214.0697	0.065		
+232.1905	0.040	+202.1620	0.061	+206.1750	0.013	+283.0337	0.031		

Figure S1

Histogram of the prediction accuracies across all 250 cross-validation rounds (25 times repeated 10-fold cross-validation). The average score of 72.1% is given by the dashed line (black). The highest counts appear around the accuracy of 70%.

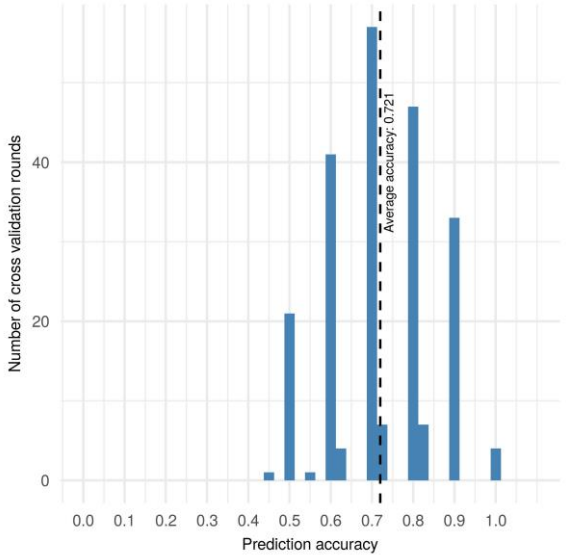
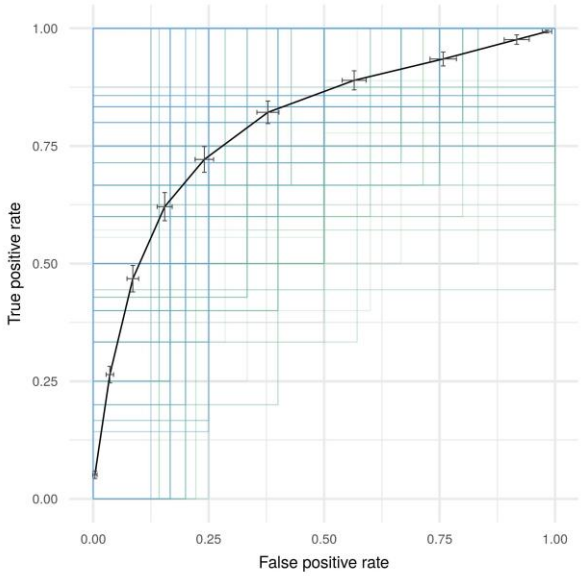


Figure S2

ROC curves (from green to blue, green indicating lower prediction accuracy, blue indicating higher prediction accuracy) were plotted for each one of the 250 cross-validation rounds. Since 250 curves are plotted, transparency of the colors is used to emphasize the overlap of the ROC curves. The more overlaps appear the less transparent the colors are. The average ROC curve (threshold averaging (TA), [12]) of all 250 cross-validation rounds is given in black. The vertical and horizontal bars on the selected points represent the 95% CI of the false positive rate (horizontal) and the true positive rate (vertical).



Bibliography

1. Fritsch FN, Carlson RE. Monotone Piecewise Cubic Interpolation. SIAM Journal on Numerical Analysis. 1980;17(2):238-46.
2. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics (Oxford, England). 2007;8(1):118-27.
3. Mann HB, Whitney DR. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. Ann Math Statist. 1947;18(1):50-60.
4. Hodges JL, Lehmann EL. Estimates of Location Based on Rank Tests. Ann Math Statist. 1963;34(2):598-611.
5. Meinshausen N, Bühlmann P. Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2010;72(4):417-73.
6. Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995;20(3):273-97.
7. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC bioinformatics. 2006;7:91-.
8. Bermingham ML, Pong-Wong R, Spiliopoulou A, Hayward C, Rudan I, Campbell H, Wright AF, Wilson JF, Agakov F, Navarro P, Haley CS. Application of high-dimensional feature selection: evaluation for genomic prediction in man. Scientific reports. 2015;5:10312.
9. Iizuka N, Oka M, Yamada-Okabe H, Nishida M, Maeda Y, Mori N, Takao T, Tamesa T, Tangoku A, Tabuchi H, Hamada K, Nakayama H, Ishitsuka H, Miyamoto T, Hirabayashi A, Uchimura S, Hamamoto Y. Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection. Lancet. 2003;361(9361):923-9.
10. Rumble JR, Lide DR, Bruno TJ. CRC handbook of chemistry and physics : a ready-reference book of chemical and physical data2018.
11. Kind T, Fiehn O. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. BMC bioinformatics. 2007;8:105-.

12. Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2006;27(8):861-74.