Pacific Northwest
NATIONAL LABORATORY

PNNL-33977

Agilent AgileBioFoundry CRADA

CRADA 398 (Final Report)

March 2023

Kristin E Burnum-Johnson
Alex Apffel
Anya Tsalenko
Christopher J Petzold
Kunal Poorey

U.S. DEPARTMENT OF ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062;
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@adonis.osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
email: orders@ntis.gov <https://www.ntis.gov/about>

Online ordering: http://www.ntis.gov

Final Progress Report

Federal Agency and Organization	Department of Energy EERE – BETO
Agreement Number:	CRADA No. 398, LBNL FP00006782
WBS Number:	2.5.3.702 NL0033732
Project Title:	Agilent CRADA
Project PI:	Kristin Burnum-Johnson, Sr. Scientist, Kristin.Burnum-Johnson@pnnl.gov and (509) 371-6335
CRADA partner leads	Agilent (Alex Apffel & Anya Tsalenko) LBNL (Chris Petzold) SNL (Kunal Poorey)
Period Covered by Report:	**June 1, 2019 – Sept 1, 2021**
Approved Project Period:	Agilent CRADA: June 1, 2019, to Sept 1, 2021
Project Officer/Technology Manager:	Gayle Bentley
Project Monitor:	Department of Energy EERE – BETO

Abstract

The mission of this CRADA with Agilent was to couple powerful MS platforms (QQQ, IM-QTOF-MS) with Agilent's novel Ultra-High-Performance Liquid Chromatography (UHPLC) fast metabolomic workflows and perform ABF Machine Learning (ML) to generated datasets. Agilent transferred UHPLC methods to PNNL and LBNL and methods were implemented and demonstrated in both labs, achieving total acquisition times of < 10 min. Metabolites analyzed using Agilent's shared methods included metabolites from central carbon metabolism, common across hosts, and metabolites unique to engineered strains. Standards were acquired in an UHPLC-Drift Tube Ion Mobility Mass Spectrometer (DTIMS) system for the first time within the context of ABF and methods were optimized based on Agilent's protocols. Samples from ABF hosts *Pseudomonas putida*, *Aspergillus pseudoterreus*, *Aspergillus niger* and *Rhodosporidium toruloides* were analyzed using the UHPLC-DTIMS platform for a total of 276 runs. A data analysis workflow compatible with the Experimental Data Depot (EDD) and completely shareable was developed for the acquired UHPLC-DTIMS data. Samples were analyzed using a Data Independent Acquisition Approach (DIA), which for most of the standards provided more transitions therefore increasing detection confidence. Using the data acquired by PNNL, LBNL, and Agilent's specifications from previous ML projects, SNL applied an ensemble ML strategy to pick the best performing model for automated LC-method selection. Finally, with the contribution of the participant labs and Agilent, SNL developed an Automated Method Selection (AMS) software tool to predict the best liquid chromatography method for analysis of any new molecules of interest. Samples with novel pathways and new metabolite targets of interest are generated at a high pace in the ABF. Therefore, our accomplishment in this CRADA improved the efficiency and accuracy of strain testing by developing and implementing fast analytical methods, robust processing tools, and software for predicting the best methods for UHPLC analysis.

Technical results/Accomplishments

- High through-put analytical methods that had previously been developed and validated for use on HPLC-QTOF-MS platforms by Agilent were adapted and tested on an UHPLC-QQQ platform in PNNL and LBNL. At LBNL, an MRM assay to target 24 metabolites was developed and validated on a UHPLC-QQQ platform operated in negative ion mode. This developed method leveraged Agilent's

HILIC 3.5-minute UHPLC gradient and successfully detected all 24 metabolites that had been targeted for analysis.

- In PNNL a library of 60 metabolite standards that included metabolites from glycolysis, TCA cycle and from pathways that are specific to the ABF host/bioproduct of interest was acquired in the UHPLC-QQQ instrument. Methods were transferred to the UHPLC-DTIMS instrument, and a library was established for the first time in ABF using this analytical platform with these 60 metabolite compounds. Library information includes (1) metabolite retention time, (2) exact mass, (3) MS/MS fragment information and (4) collision cross section (CCS).

- Software tools with high robustness and reproducibility for processing of UPLC-QQQ and UHPLC-DTIMS data were designed and implemented at PNNL. The LC-IMS-MS-data-independent acquisition (DIA) approach was followed; this approach provided more transitions (fragments) for most of the standards therefore increasing metabolite detection confidence. The data analysis workflow is all shareable (raw data and results) and compatible with EDD.

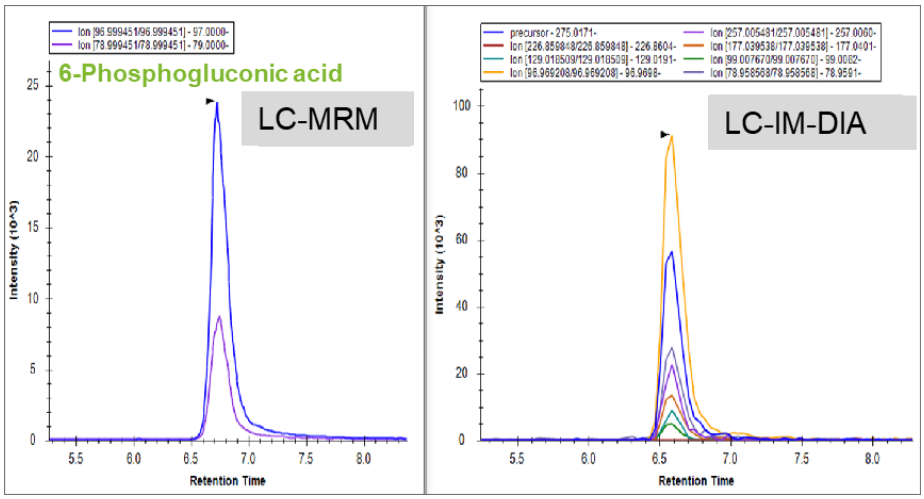


Figure 1. Chromatogram of precursor ions and corresponding transitions of 6-phosphogluconic acid acquired by MRM and IMS.

- Agilent's Automatic Analytical Workflows (AAW) were integrated with EDD from ABF. In LBNL automated method selection, data acquisition, and data processing with the Agilent AAW software and EDD for metabolites from three metabolic pathways (glycolysis, TCA, and the pentose phosphate pathway) was implemented. Input information for the AAW software includes the compound name and the SMILES notation for the chemical structure. The AAW software uses the SMILES information to provide a prediction of what LC-MS method would provide the best ability to detect the compound. The AAW output file lists the scores for each of the four Agilent methods that it can predict for data acquisition (HILIC chromatography in positive- or negative-mode mass analysis and reverse-phase (RP) chromatography in positive- or negative-mode mass analysis) and the recommended method for acquisition. The AAW preferred method information was used as part of the input for the EDD study creation for the strains of interest. The EDD study was then used to generate the data acquisition worklist for the lines (samples) to be analyzed. The worklist was subsequently used to acquire the data on an Agilent UHPLC-QQQ system, the data was processed via Skyline and uploaded to the EDD

study. This process demonstrated rapid, automated method prediction via the Agilent AAW software coupled with data acquisition, data processing, and data sharing.

- In PNNL, samples from four ABF hosts (*P. putida*, *A. pseudoterreus*, *A. niger* and *R. toruloides*) were analyzed using an Agilent UHPLC-DTIMS with LC methods adapted from Agilent and DTIMS methods developed at PNNL. Total run times were < 10 min which is a significant reduction from the usual 40 min required for analysis in the routine gas chromatography (GC-MS) metabolomics platform. Metabolite extracts were analyzed using a library established with 60 metabolite compounds. The panel of compounds included metabolites from central carbon metabolism that are commonly analyzed in ABF studies including glycolysis, TCA cycle intermediates, amino acids, CoAs and some specific to ABF host/bioproduct of interest, like compounds related to engineering of *Aspergillus* for 3-HP production. Metabolite standards that we cannot detect using our routine GC-MS metabolomics platform (i.e., intermediates of the mevalonate pathway) were observed using Agilent's methods on the UHPLC-DTIMS systems. With this, we improved our coverage of metabolic pathways which is critical to inform the optimization of ABF strains. Projects using the MS-software Skyline and EDD studies were generated for the ABF samples acquired using the Agilent UHPLC-DTIMS and shared with the CRADA members.

Table 1. Summary of runs acquired in the UHPLC-DTIMS platform in the Agilent CRADA project

Organism	Acquisition mode (Collision energies)	Number of runs	Acquisition date
Pseudomonas putida	2 (20 V and ramp 10-40 V)	22	August 2020
Aspergillus pseudoterreus	2 (20 V and ramp 10-40 V)	30	August 2020
Aspergillus niger	2 (20 V and ramp 10-40 V)	64	August 2020
Rhodosporidium toruloides	2 (20 V and 40 V).	160	June-July 2021
Total		276	

- Results from the novel IMS analysis of these ABF samples were also analyzed in a biological context. In samples from *P. putida*, the main objective was to examine the impact of insertion of integration location of muconate transporter on host viability, gene expression, and strain performance. The strains consisted of wild type *P. putida* KT2440 (strain CJ019) and CJ019 strains with MucK inserted at different locus (Fig 2). The 2nd batch analyzed were cell extracts from *A. pseudoterreus* and *A. niger* and the objective was to compare 3-HP production in Aspergillus strains, determine other metabolites being produced and pyruvate carboxylase and aspartate aminotransferase overexpression effects (Fig 3). *R. toruloides* were the last samples analyzed under this CRADA, using the UHPLC-DTIMS methods developed and the improved software analysis pipelines. The first batch of *R. toruloides* contained samples engineered in kaurene synthase or geranylgeranyl pyrophosphate synthase. The 2nd *R. toruloides* study (Fig 4) had as aim to investigate the effect of feedstock variability (high or low contents of ash and moisture) on a fermentation for the production of bisabolene.

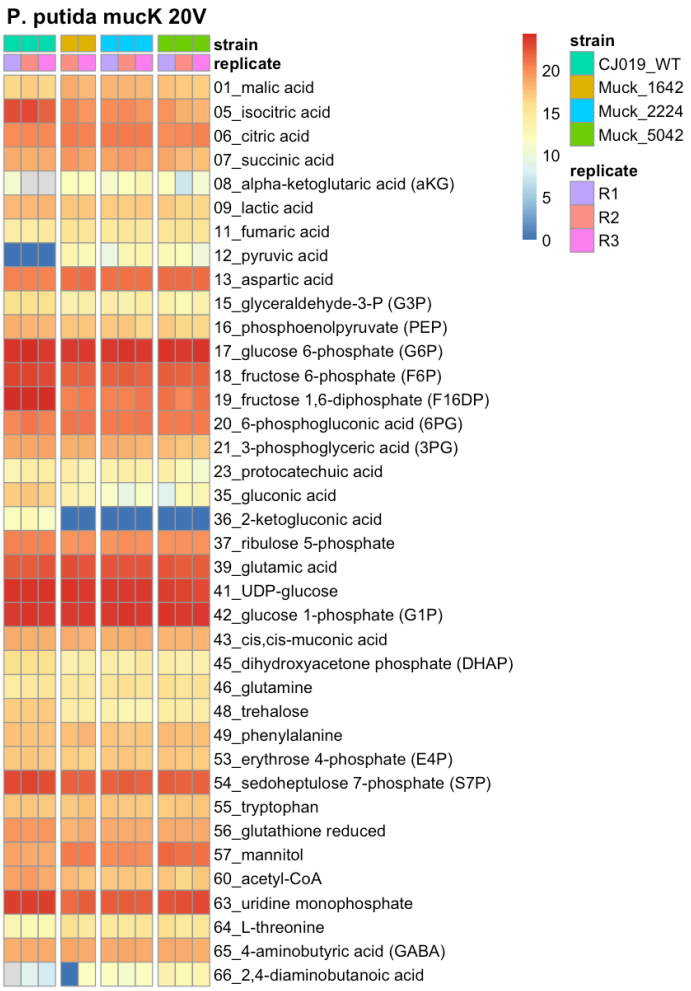
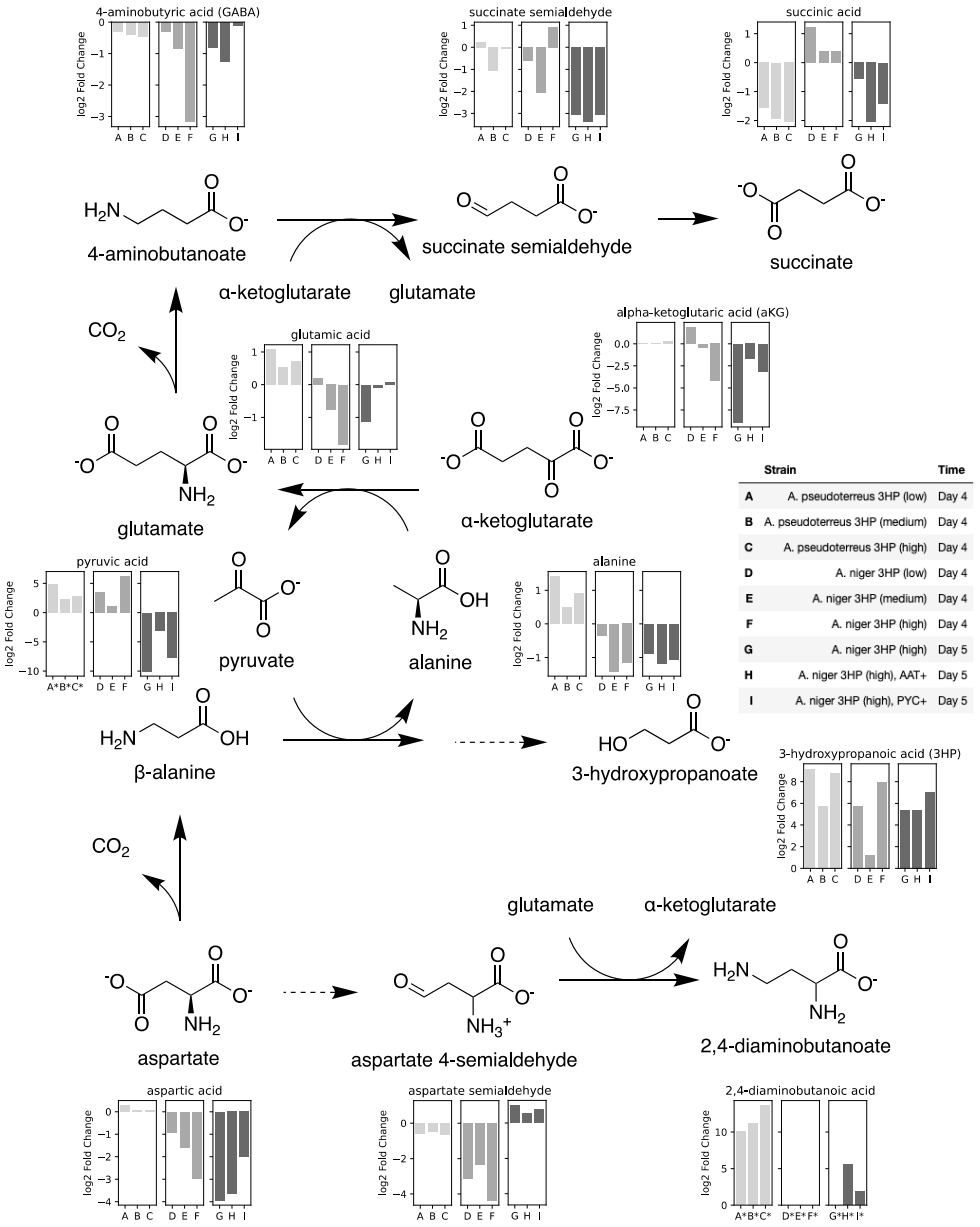


Figure 2. Metabolites detected in *P. putida* samples acquired in the UHPLC-DTIMS platform. In this experiment, *P. putida* WT and strains engineered with a muconate transporter were analyzed. Relative levels of metabolites in the TCA cycle, pyruvic acid and ED-EMP were significantly altered in cells with the MucK inserted compared to WT.

- Previous machine learning (ML) models within the Agilent's Automatic Analytical Workflows (AAW) for selecting the best analytical method based on physio-chemical properties were developed using properties from a commercial product, ChemAxon, which had licensing costs. With the contribution of SNL to this CRADA, this functionality was replaced using an open-source software package, PaDel. New machine learning models were trained using SciKit Learn based on the physio-chemical properties calculated from PaDel. The performance of these models was validated using the IROA metabolite data set, as well as data from additional sets of metabolites from selected pathways generated in this project.



	Strain	Time
A	A. pseudoterreus 3HP (low)	Day 4
B	A. pseudoterreus 3HP (medium)	Day 4
C	A. pseudoterreus 3HP (high)	Day 4
D	A. niger 3HP (low)	Day 4
E	A. niger 3HP (medium)	Day 4
F	A. niger 3HP (high)	Day 4
G	A. niger 3HP (high)	Day 5
H	A. niger 3HP (high), AAT+	Day 5
I	A. niger 3HP (high), PYC+	Day 5

Figure 3. Log2 fold changes (compared to base strain) of peak areas of metabolites that are part of the 3-HP producing pathway built into the *A. pseudoterreus* and *A. niger* analyzed in this CRADA. Species-specific responses to 3-HP production were found, including high accumulation of an undesired byproduct (2,4-diaminobutanoic acid) in *A. pseudoterreus* 3-HP strains.

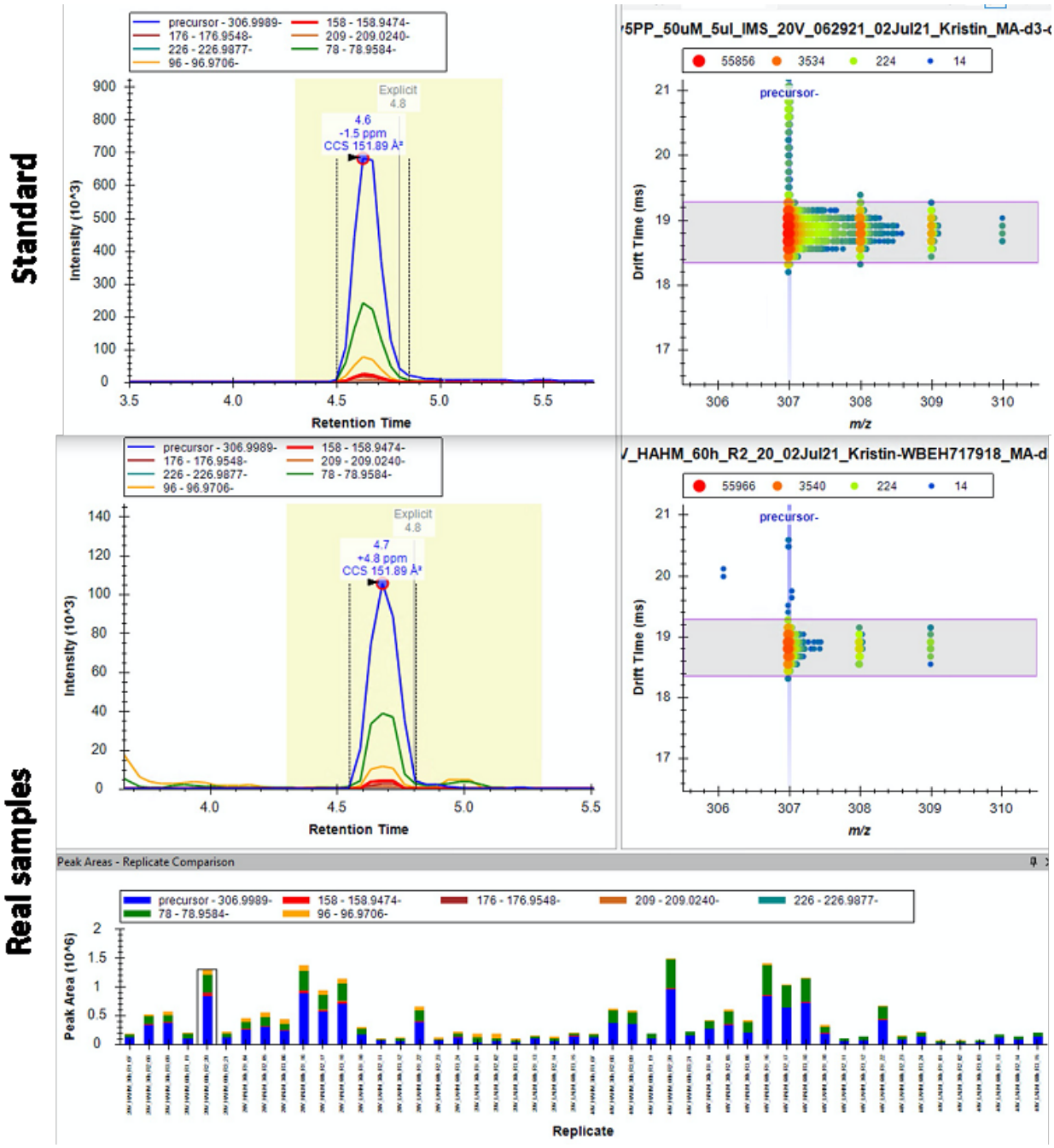


Figure 4. Chromatograms and ion mobility of mevalonate 5-pyrophosphate (mev-5PP) acquired by UHPLC-DTIMS. The detected precursor ions and corresponding transitions (left panels) show the same fragmentation and elution patterns and the precursor arrival times (right panels) show the same mobility (CCS) in the standard and the *R. toruloides* samples, providing a confident metabolite identification. The bottom panel shows the differential abundances detected across the multiple sample replicates.

- Models for the automated LC-MS method selection software were significantly improved during this CRADA. SNL applied the strategy shown on Figure 5 for each of the routine LC-MS methods (HILIC+, HILIC-, RP+, RP-) to develop a new Automated Method Selection (AMS) software. There is a precursor score for each entry in the training dataset (ranging from 0 to 100%) for each method. This marks the

measure of the LC-MS method's performance for the compound. The criterion for method recommendation used was that the precursor score must be higher than 50%. Using this information, four machine learning models, including custom feature engineering for each method were made. The ML model used for prediction was a random forest with 100 bagged trees and utilized feature engineering with the Boruta feature reduction method. The application of Boruta reduced the number of features by two orders of magnitude. As a result of its application, a handful of features was obtained to avoid overfitting and improve the model's performance, as well as performing further exploration and deep dive to understand the science behind those physiochemical and structural properties. Table 2 displays the summary of the feature reduction strategy applied to the training dataset and the ROC (receiver operating characteristic) curves in Figure 6 illustrate the diagnostic ability of the classifiers. The pre-trained models using the dataset provided by Agilent are developed using all the available data and saved for prediction of the new dataset. The AMS software is packaged in a zipped file, available on GitHub, and shared with the group through the SharePoint site.

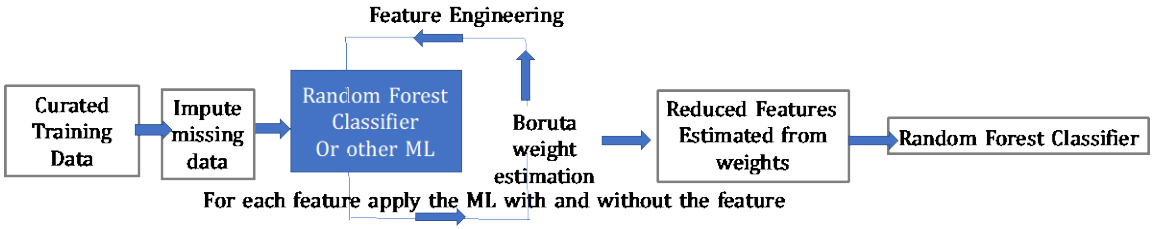


Figure 5. Machine learning strategy with emphasis on feature engineering presently used for the development of the new Automated Method Selection (AMS) software.

Table 2. Machine learning application performance metrics

Model	Model Accuracy	Total extracted features	Boruta Features
HILIC+	84.33 +/- 4.20	7009	81
HILIC-	81.40 +/- .452	7009	95
RP+	76.51 +/- 4.57	7009	56
RP-	77.08 +/- 3.74	7009	94

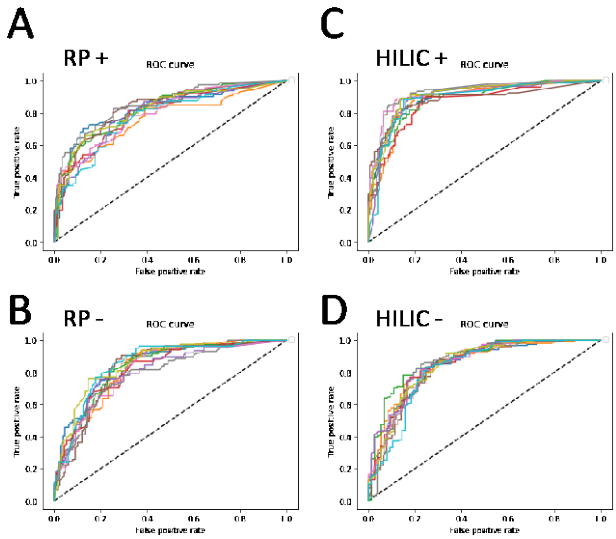


Figure 6. ROC curves for the random forest classifier step for determine the accuracy of the 4 classification models for the new Automated Method Selection (AMS) Software.

- The Automated Method Selection (AMS) software developed by SNL with data and specifications contributed by Agilent, PNNL and LBNL is able to predict the most suitable LC-method for analysis for any new molecule in question. The score can be calculated for each model (HILIC+, HILIC-, RP+, RP-) and reported. Under this CRADA, besides optimization of analytical methods and software for robust processing of data acquired in novel instruments (UHPLC-DTIMS), bioinformatic tools for Automated Method Selection were also developed. These AMS tools will provide capabilities for fast selection of best LC methods for analysis of new metabolites of interest in the ABF. Improvements in analytical methodologies, software for analysis of data and automation have the potential of enabling faster testing of strains generated in the consortium and accelerate the DBTL cycle.

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354
1-888-375-PNNL (7665)

www.pnnl.gov