Matterworks™

matterworksbio.com

Absolute quantitation of metabolites using machine learning and StandardCandles as universal calibrators - the second-generation model

Jennifer M Campbell, Timothy Kassis, Ana S. H. Costa, Jeff Pruyne, Joshua D Lauterbach, Luke Ferro, Steven Hooper, Jack Howland, Mimoun Cadosch Delmar, J. M. Geremia

Matterworks Inc., 444 Somerville Ave, Somerville, MA 02143

ASMS 2023, TP493

Introduction

The transformative power of Artificial Intelligence (AI) has become increasingly apparent in both our daily lives (e.g. GPT-X and DALL-E) and in scientific research (e.g. AlphaFold). Recently, some novel use AI methods for the identification of compounds from MS/MS spectra have also been published.(1) The promise of AI in science is two-fold: to find nuanced connections within large datasets, which enable the prediction of hitherto “unseen” phenomena and to “learn” inherently difficult tasks which typically need high levels of expertise. At Matterworks, we are actively working on both capabilities of AI – focusing on novel methods to access hitherto intractable metabolomic information directly from raw spectra.

One of the fundamental challenges of MS is that we are determining the behavior of neutral compounds in the liquid phase through detecting charged analytes in the gas phase. Decades of efforts have failed to produce an accurate mathematical model for complex ionization processes.(2) Consequently, methods have been developed to link the detected gas phase MS signal to the absolute concentration in the gas phase (e.g., matrix matched calibration curves, isotope dilution MS). These methods lack both simplicity and scalability. They are typically developed on an analyte-by-analyte basis and require experts to validate analyte detection and to analyze data.(3) The determination of absolute concentration of small polar metabolites directly from raw spectra is the first challenge Matterworks is addressing using AI with our first product, Pyxis (Figure 1).

The Pyxis platform approaches the determination of absolute metabolite concentrations in a fundamentally different way. One core novelty of the Matterworks solution is the use of universal calibrators, called “StandardCandles” after the celestial objects used to calculate the distance to stars. The StandardCandles were carefully selected to encode structure and composition-dependent ionization physics across a broad swath of chemical space. We have developed a proprietary technology platform to generate spectral data to “feed” our model - incorporating everything from training set design, sample creation, acquisition work list creation, and data extraction. Using this platform we can pretrain the model on a broad concentration range of an ever-growing list of analytes. All training data integrated into the deep learning model is acquired with an optimized LC-MS method and samples incorporating our StandardCandles. Any subsequent experiments that incorporate the StandardCandles and use the same LC-MS method can be submitted for instant absolute concentration determination using the Pyxis model.

Analytical Methods

One of the core improvements in our second-generation model was the development of the method used to collect data. A target group of 160 analytes of key cellular metabolites, were tested using multiple columns, additives, gradients, and pHs, with the goal of maximizing analyte detection in a single method. All data were recorded on three identically configured Thermo Fisher Scientific Orbitrap Exploris 120 mass spectrometers connected to Thermo Fisher Scientific Transcend LX-2/LX-4 with a Vanquish Flex, enabling parallel column equilibration and injection. The method is shown below and has proven robust in providing near identical data.

Parameter	Setting
Source	H-ESI with polarity switching (-2.5kV to 3.5kV)
Transfer Tube	315 °C
Vaporizer	350 °C
RF lens (%)	60 ms
Max Injection	60 m
Max Trap Fill	2E6
Sheath	50
Auxiliary	10
Sweep	1
Scan Range	70-800 m/z
Resolution	60,000
Calibration	EASY-IC at run start

Column: Waters Atlantis Premier BEH Z-HILIC VanGuard FIT Column
(2.5µm, 2.1mm x 50mm)

Solvent A - Water with 20mM Ammonium Carbonate and 0.25 ammonium carbonate (v/v) (pH=9.6).
Solvent B - Acetonitrile

Time (min)	Flow (mL/min)	% Solvent B
0	0.5	95
1	0.5	95
8.5	0.5	20
9.5	0.5	20
10	0.5	95
12	0.5	95

Table 1. *Details of the MS (left) and LC (right) operating parameters which are used in the collection of training set data. Note because we are running on a 2 column system we are continually collecting data – 6.5 minutes on each column.*

Training Set Creation and QC

Remove need for per analyte standard curves	No method development	No peak interpretation	Increased lab automation	Quantitative data can be compared

>90% reduction in cost for standards	10 - 100x increase in productivity	Reduced training & personnel requirements	Scientists can focus on understanding the biology & informing decision making	Data exchange and inter-lab comparisons - globally

Figure 1. Workflow for predicting concentrations with Pyxis. Samples must be prepared using the StandardCandles and acquired using the canned LC-MS acquisition method. Resultant raw data can be uploaded to the cloud for immediate prediction of absolute concentration.

The goal of Matterworks’ in-house training set generation platform is to create a large data set representing the variations of “situations” the model will be required to predict. Using custom software, plates of simulated samples are created for automated preparation and LC-MS analysis. Each sample has a unique distribution of analyte concentrations and different background matrices.

Because the operation of the AI model requires full spectra as input, we are not constrained by traditional requirements of method development. Figure 2 shows the separation of the isomers L-Leucine and L-Isoleucine – although not sufficient for conventional quantitation, the AI accurately quantifies these analytes. Our initial concentration prediction models (shown in this poster) were evaluated by testing concentration accuracy for a set of 70 analytes.

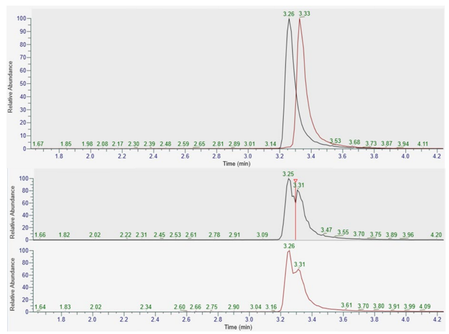

Figure 2. Demonstration of poor-quality separation which can still result in successful predictions from the AI model. XICs for Leucine (black) and Isoleucine (red) injected as separate compounds (top) and from a randomly selected training sample (bottom).

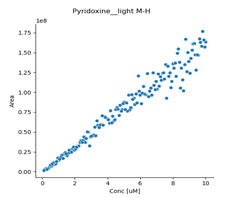

Figure 3. Two examples of ongoing QC completed during training set acquisition. Left is a set-by-set verification that theoretical concentration and peak area (using traditional techniques) are correlated. Right is the mass error from three instruments and over 9 million extracted ion chromatograms.

Performance and Results

In order to validate and track the model performance as training dataset size increases and model architecture changes, we have designed “benchmarking” training sets. The data from these 11 plates incorporate the complete analyte training ranges and five matrices – E. coli lysate, Ham’s medium, Human Plasma-Like Medium (HPLM), RPMI, and the Pyxis “Extraction Solution” (blank matrix). Media dilutions and the RPMI have not been included in the training set. In this way, we can test the ability of the model to predict for unseen matrices. Experimental concentrations are calculated by spiking standards of known concentration above the known concentrations in the defined media.

The current status of model prediction is summarized in Figures 4 and 5. The most critical conclusion from the two figures is that the model accuracy is independent of the matrix. The predictive power, however, is weaker for specific analytes, and degrades near the limits of detection and quantification. The average absolute percent error of all measurements in the benchmark plates was 36% and median was 25%. The worst performing analytes and concentration ranges inform where the training set needs to be augmented in the future.

It is noteworthy that in spite of the poor separation highlighted in Figure 2, the model’s predictions for isomers L-Leucine and L-Isoleucine (highlighted with a red star in Figure 4 and 5) are not substantially different from those of well-separated analytes. Figure 6 further emphasizes this fact by “blowing up” the panels for those analytes.

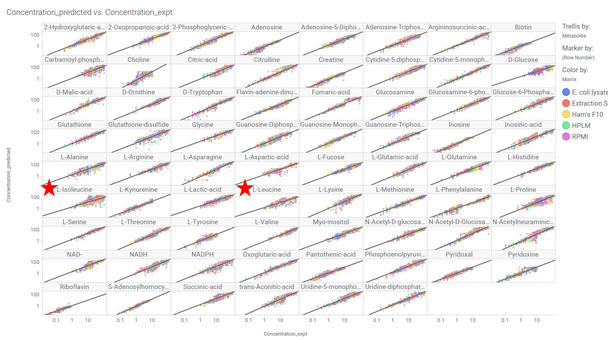

Figure 4. Current predictions from model on benchmark data. Included are ~89,000 known concentrations spanning three orders of magnitude in five different matrices. Note that RPMI (pink) is a matrix which has not been trained on. Log-Log plot.

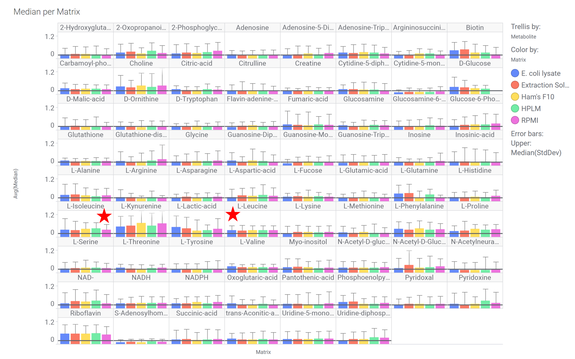

Figure 5. For data plotted in Figure 5, the median of the absolute percent error – parsed by analyte and by matrix. Error bars are standard deviation, and the pink bar is an “unseen” matrix.

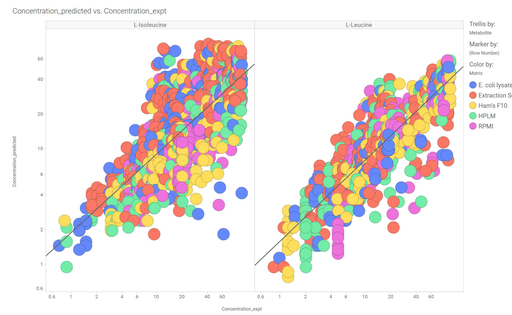

Figure 6. ***Details of the performance on the poorly separated isomers L-Isoleucine (left) and L-Leucine (right). Lack of stratification by color further emphasizes the matrix generalizability of the model.***

Advantages and Conclusions

As our training set increases in size, we expect to improve the accuracy of the concentration prediction to match (and possibly surpass) that of traditional techniques. In the current embodiment there are key advantages for using AI with universal calibrators for absolute concentration prediction.

1) In providing both a turn-key method and lowering the separation requirements for analytes, the barrier to access absolute concentration measurements for a broad swath of metabolites is lowered such that it is easily accessible to the non-expert.

2) By removing the need for post-acquisition data QC and analysis, absolute concentration of large numbers of metabolites can be determined in near real time. The input for the model is the raw data file from the instrument and the output is absolute concentration (with quality metrics).

3) Any analyte which can be detected using the Pyxis method can be trained – providing broad scalability of analytes which can be predicted by the model. Since the absolute concentration is predicted on non-targeted data, when the number of trained analytes increases, concentrations of new analytes can be predicted on previously acquired data.

Pxyis has been shown to deliver accurate absolute concentrations of polar metabolites across three orders of magnitude and a wide range of matrices, including matrices not used for training. Concentration accuracy was assessed for 70 representative polar metabolites by reference to known concentration benchmarks. Data were acquired using a standardized untargeted LC-MS method and AI model that improves in breadth and accuracy through training accretion. These advances in machine learning for quantitative metabolomics demonstrate the power of AI for the interpretation of unstructured, raw MS data.

All authors are employees and shareholders of Matterworks.

References

1. (a) Michael Murphy and Stefanie Jegelka and Ernest Fraenkel and Tobias Kind and David Healey and Thomas Butler, “Efficiently predicting high resolution mass spectra with graph neural networks”, arXiv 2301.11419v1 (2023). (b) Ayumi Kubo, Azusa Kubota, Haruki Ishioka, Takuhiro Hizume, Masaaki Ubukata, Kenji Nagatomo, Takaya Satoh, Mitsuyoshi Yoshida, Fuminori Uematsu, “Construction of a Mass Spectrum Library Containing Predicted Electron Ionization Mass Spectra Prepared Using a Machine Learning Model and the Development of an Efficient Search Method” Mass Spectrometry (Tokyo) , 12 , A0120 (2023).
2. Piia Liigand, Jaanus Liigan, Karl Kaupmees, Anneli Kruve,, “30 Years of research on ESI/MS response: Trends, contradictions and applications”, Analytica Chimica Acta, 1152, 238117 (2021)
3. Gioele Visconti, Julien Boccard, Max Feinberg, Serge Rudaz, “From fundamentals in calibration to modern methodologies: A tutorial for small molecules quantification in liquid chromatography–mass spectrometry bioanalysis”, 1240, 340711 (2023).