Pacific Northwest
NATIONAL LABORATORY

PNNL-36690

Statistically-driven Experimental Design to Improve Reference-free Quantification of Small Molecules by Liquid Chromatography-Mass Spectrometry

September 2024

Fanny Chu
Jessica Bade
Luke Durell
Matthew Turner
Charlie Doll
Sean Colby
Sydney Schwartz
Anna Hale
Eva Brayfindley

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
www.osti.gov
ph: (865) 576-8401
fox: (865) 576-5728
email: reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: http://www.ntis.gov

PNNL-36690

Statistically-driven Experimental Design to Improve Reference-free Quantification of Small Molecules by Liquid Chromatography-Mass Spectrometry

September 2024

Fanny Chu
Jessica Bade
Luke Durell
Matthew Turner
Charlie Doll
Sean Colby
Sydney Schwartz
Anna Hale
Eva Brayfindley

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

Pacific Northwest National Laboratory
Richland, Washington 99354

Abstract

Non-targeted analysis of small molecules and metabolites in unknown, complex samples using liquid chromatography-tandem mass spectrometry remains challenging. One of the main bottlenecks is the extensive unannotated regions of metabolomics mass spectrometry data, resulting in knowledge gaps. Small molecule annotation in mass spectrometry data has conventionally relied on reference standards and libraries for compound identification and confirmation, which can constrain compound identification to those molecules already known, thus limiting the ability to discover new knowledge and new markers.

Retention time prediction can facilitate and expedite unknown compound identification in non-targeted analysis of complex metabolomics samples. Additionally, accurate retention time predictions can also inform sample mixture design for LC-MS/MS analyses. However, current machine learning-based methods for retention time prediction are typically developed for specific chromatographic platforms and are not generalizable across scales.

And while technologies and methods to improve reference-free metabolite identification for more comprehensive annotation of unknowns has received much attention, development of the same for quantitation without reference standards has been much more limited, despite its importance in toxicological, environmental, food safety, forensics, and clinical applications. We believe that a reference-free quantitation strategy that exploits mass spectrometry data already collected for reference-free identification can provide much more insight on unknowns, and move the metabolomics field for more complete unknowns characterization.

As such, we pursue two efforts to improve upon current state-of-the-art methods in non-targeted analysis: (1) machine learning-based retention time prediction and (2) statistical design of experiments framework for reference-free quantitation.

In this work, we develop and demonstrate (1) a generalizable retention time prediction capability across chromatographic conditions and scales, and (2) a statistical design-based framework for response factor contribution elucidation and reference-free quantitation. Evaluation of our retention time prediction model, PrediToR, showed approximately 24% improvement over current models, and we observed approximately 10X improvement in concentration estimation accuracy from our statistical design-based response factor model over a primarily ionization efficiency-based model. We expect that future efforts to improve upon these new capabilities will further advance non-targeted analysis of small molecules towards truly reference-free metabolomics.

Acknowledgments

The authors would like to thank Dr. Tom Metz and Dr. Robert Ewing for support and fruitful discussions. This research was supported by the m/q Initiative, under the Laboratory Directed Research and Development (LDRD) Program at Pacific Northwest National Laboratory (PNNL). PNNL is a multi-program national laboratory operated for the U.S. Department of Energy (DOE) by Battelle Memorial Institute under Contract No. DE-AC05-76RL01830.

Acronyms and Abbreviations

ACN: acetonitrile
APCI: atmospheric pressure chemical ionization
BRF: bootstrapped response factor
CCF: face-centered central composite design
CDK: Chemistry Development Kit
CEAP: Co-eluting analyte pair
CRF: constant response factor
DoE: design of experiments
EDTA: ethylenediaminetetraacetic acid
ESI: electrospray ionization
GPU: graphics processing units
HCD: higher-energy collisional activated dissociation
HILIC: hydrophilic interaction liquid chromatography
HMDB: Human Metabolome Database, public database for mass spectrometry data
HPLC: high-performance liquid chromatography
IE: ionization efficiency
LC-MS: liquid chromatography-mass spectrometry
LC-MS/MS: liquid chromatography-tandem mass spectrometry
logIE: logarithm-transformed ionization efficiency
MAE: mean absolute error
MeOH: methanol
ML: machine learning
MoNA: MassBank of North America, public database for mass spectrometry data
MPA: mobile phase A
MPB: mobile phase B
MS1: mass spectrometry, first-dimension mass spectrometry acquisition, precursor ion spectra
MSE: mean signed error
MS/MS: tandem mass spectrometry, also MS2, fragmentation spectra
MSMLS: Mass Spectrometry Metabolite Library of Standards
nanoLC: nanoflow liquid chromatography
qNTA: quantitative non-targeted analysis
PI: prediction interval
RF: response factor
RP: reversed phase
RPLC: reversed phase liquid chromatography
RT: retention time

SMILES: Simplified Molecular Input Line Entry System
UPLC: ultra-high-performance liquid chromatography

Contents

Figures

Tables

1.0 Introduction

We describe the development and performance of two new capabilities to advance non-targeted mass spectrometry analysis of small molecules in unknown, complex samples. Though non-targeted analysis has become the predominant approach to data acquisition and analysis in metabolomics, the continued existence of large regions of unannotated mass spectrometry data from non-targeted analysis remains a bottleneck in metabolomics. Further, quantitation of metabolites remains a time-intensive and costly endeavor using the current targeted analysis paradigm. Current capabilities in these areas of metabolomics are insufficient to fully address the underlying needs. As such, we believe that development of new capabilities is critical.

1.1 Gaps and Needs

It is well known that retention time information can facilitate compound identification, towards more complete annotation of unknowns. However, linking retention time information to detected compounds in mass spectrometry data is a non-trivial task. Retention time prediction has been shown to speed up the process of compound identification. But current tools to predict retention times typically show limited applicability across a broad range of chromatographic conditions and scale.

On the other hand, metabolite quantitation has conventionally relied on a targeted analysis and reference standards, but the latter may not be available for many metabolites of interest and this targeted approach can be quite time-intensive, as compound-specific calibration curves are typically needed. Recently, machine-learning based ionization efficiency predictions have emerged as an alternative to bespoke calibration curves for concentration estimation, specifically in replacing the response factor determination. However, this technique has not considered experimental variables, such as sampling and chromatographic conditions, that may represent a not insignificant contribution to a compound's measured response. The current state-of-the-art ionization efficiency models have also demonstrated up to 1-2 orders of magnitude error when estimating concentrations.

1.2 Aims

To address current gaps and needs in non-targeted analysis, we focus on two distinct efforts: (1) to improve accuracy of retention time prediction models for applicability to a broader range of chromatographic conditions, including nanoflow liquid chromatography, and (2) to improve accuracy of reference-free quantification models by accounting for sampling and chromatographic conditions.

We demonstrate the success of a new, generalizable retention time prediction model, PrediToR, that utilizes a novel deep learning architecture and is extensible to a broader range of chromatographic conditions than implemented in current models. Further, we describe the development and evaluation of a new capability to implement reference-free quantitation by characterizing the contributions of experimental conditions to response factor via statistical design of experiments. Contributions from experimental variables to response factor were found to be greater than chemical-specific ionization efficiency, the latter of which has received the most attention, which has implications for quantitation accuracy. We show, on average, approximately 10X improvement in quantitation accuracy over current state-of-the-art methods and without the need to first elucidate compound identity. The successful performance of both our retention time and response factor models demonstrates the promise of these new

capabilities towards improvement of non-targeted analysis, with broad implications in reference-free metabolomics.

2.0 PrediToR: Hypernetwork for Improved and Generalizable Retention Time Prediction

2.1 Introduction

Accurate prediction of retention times for small molecules as applied to liquid chromatography-mass spectrometry (LC-MS) measurements remains challenging, as current methods are typically not generalizable across the wide range of possible chromatographic conditions. A number of applications rely on knowing the retention times of small molecules and/or the capability to predict retention times for compounds without *a priori* empirical measurements.

Retention time information is highly useful for applications such as unknown compound identification and sample mixture design. Knowing the expected retention times of compounds can expedite unknown compound identification, as retention time information provides some degree of specificity to narrow down potential candidates. Sample mixture design can also be facilitated by knowing the retention times of compounds of interest, particularly if the mixture is designed to meet certain chromatographic conditions, e.g., to contain compounds with elution times that span the entire chromatographic run with little coelution. We are interested in characterizing retention times of small molecules precisely for these applications. However, retention times are specific to a set of chromatographic conditions and it can be difficult to assign retention times to compounds if they have not been empirically measured.

Previous efforts have described the development of retention time prediction models to be generalizable to a larger set of chemicals and minimize the need for empirical measurements. There may also be cases in which it may not be feasible to empirically measure the retention times of compounds of interest. In these cases, accurate retention time (RT) prediction models can be a useful alternative. Several different RT prediction machine learning (ML) models are available: Retip[1], METLIN[2], and RT-Ensemble Pred[3]. We briefly describe each model below.

Retip is an R package that enables prediction of retention times for small molecules[1]. It allows for the import of training and testing data in Excel file format, requiring specific variables in the dataset such as the name of the compound, InChIKey, SMILES, and retention time value. Notably, Retip does not consider any chromatographic method parameters in its process. The framework utilizes the Chemistry Development Kit (CDK)[4] to calculate chemical descriptors, and it handles data cleaning by removing descriptors with NA values or near-zero variance. The data are then partitioned into training and testing sets using the caretDataPartition function. This training data is subsequently fed into various model architectures under the Retip umbrella, and the most performant model (as demonstrated on the test set) is then selected for deployment.

The METLIN deep learning model uses a feedforward neural network architecture for retention time prediction[2] [2]. This is implemented using the keras package in R. The model utilizes four hidden layers (1000, 500, 200, 100 nodes, respectively), each with a ReLu activation function. The model is optimized with Adaptive Moment Estimation (Adam)[5], the loss function is mean squared error, the default number of epochs is 20, and the default batch size is 35. This model takes extended connectivity fingerprints (1024 fingerprints per biomolecule) as model feature inputs to represent compounds, are obtained for each biomolecule and treated as predictor variables, and a 75%-25% train-test split is applied prior to model fitting. METLIN model performance was benchmarked using the Small Molecule Retention Time (SMRT) dataset comprising 80,038 molecules[2].

RT-Ensemble Pred employs an ensemble model to predict retention times of compounds under different liquid chromatographic conditions[3]. Temperature, flow rate, elution time, mobile phase compositions, and column descriptors are all utilized in model training. Ensemble sampling is a technique used to address the issue of imbalanced datasets. This method involves creating multiple subsets or “ensembles” from the original dataset, each with a balanced class distribution. In the case of the curated datasets used in the RT-Ensemble Pred model, it was observed that implementing ensemble sampling significantly reduced the model errors. This suggests that the technique was effective in improving the model's performance by addressing imbalance in the data. The RT-Ensemble prediction model uses four types of molecular descriptors calculated with various tools, including ChemoPy[6], Pybel[7], RDKit, and PaDEL.[8] After testing numerous models and evaluating their performance, a subset of 50 molecular descriptors was selected for the final model. This selection was based on their ability to effectively predict retention time, thereby optimizing the model's performance.

Several differences exist among these three models, including model architecture and deployment approach (e.g., single model vs top-performant model), benchmark datasets, and model feature inputs (e.g., inclusion of chromatographic conditions, CDK descriptors vs fingerprints). These differences can make it challenging to compare performance without reimplementation of each model on the same train/test sets.

Further, RT prediction models are typically trained on limited chromatographic conditions; that is, retention time prediction models are trained on data representing limited chromatographic conditions. The data used to train these models typically consists of calculated molecular descriptors, which provide a quantitative representation of the molecular properties and structure. However, in some instances, such as in the case of the RT-Ensemble Pred model, certain chromatographic parameters are also incorporated. These include temperature, flow rate, elution time, mobile phase compositions, and column descriptors. Despite these additions, there is still a notable absence of certain data in the training sets. This missing data can include parameters like column type, pH, and the %B gradient. The lack of these features in the training data can limit the model's ability to accurately predict retention times under a variety of chromatographic conditions. Therefore, efforts to include these parameters could significantly enhance predictive accuracy and applicability.

In this work, we aim to develop a generalizable RT prediction model that is extensible across a broad range of chromatographic conditions, across different acquisition run times, and other collection variables. We envision that this model can be generalized across the range of LC conditions (i.e., from HPLC down to nanoLC), different mobile phase systems, and column types. Second, the desired model would accurately predict retention times agnostic to differences in chromatographic gradients and total run times. To support such an extensible model, we anticipate that model development would require a substantial amount of training data.

While literature data is plentiful and accessible, the current state of available literature retention time data lacks a standardized approach for data reporting. This inconsistency is evident in the various compound labels used across different datasets, which can include SMILES, InChIKey, InChI, PubChem IDs, and a variety of naming schemas. Furthermore, there is a lack of uniformity in the units of measurement used for parameters such as flow rate, retention time values, and temperature. Some datasets provide detailed methodological information, including the temperature of the column, retention times per molecule, column type, elution time and more. However, other publications may not include all these crucial method factors. These factors could significantly influence the resulting retention time of the compound. However,

because of the lack of standardized reporting, many of these parameters cannot be used as a feature in a retention time prediction model. The absence of a standardized documentation method makes it challenging to compile data from different sources. This difficulty can hinder the accumulation of a sufficiently large dataset necessary for accurately predicting LC-MS retention time values.

As we encountered this lack of dataset standardization during curation for model development, we focused on careful curation and harmonization of literature data as the second thrust area in our generalizable RT prediction model development efforts. Information sourced from RepoRT[9], a new open-source data repository containing retention time data, significantly enriched our curated dataset effort for ML model development.

Our deep learning RT prediction model, PrediToR, was developed with the ultimate goal of empowering chemists and cheminformaticists alike to easily predict retention times for their own data with minimal software expertise and with high confidence in accuracy of predictions.

We demonstrate the success of PrediToR for RT prediction of our carefully curated literature data, that is extensible to a wide range of chromatographic conditions. We describe the novel network architecture, which, to our knowledge, is the first report of such a model, and resultant performance. We show that this new model enables accurate retention time prediction and can be generalizable to a number of different chromatographic conditions, allowing this model to be easily deployed on different datasets without retraining.

2.2 Methods

While our efforts herein focused on development of a new, generalizable RT prediction model, our initial motivation to investigate RT prediction models stemmed from the need to efficiently identify compounds in sample mixtures from nanoLC-MS data for subsequent sample design. Section 2.2.1 below describes the sample mixture and acquisition details which produced the empirical nanoLC dataset that seeded this RT prediction effort.

2.2.1 Sample preparation and data acquisition

Samples were created using the Mass Spectrometry Metabolite Library of Standards (IROA Technologies). The Mass Spectrometry Metabolite Library of Standards (MSMLS) is a collection of high-quality small biochemical molecules that span a broad range of primary metabolism. This library of standard reference materials comprises 603 unique compounds present in individual wells, each containing 5 µg of compound. Compounds in the MSMLS library were dissolved according to the manufacturer's instructions using 20 µL of either 5% methanol in water or a 1:1 chloroform:methanol mixture.

For initial method development and acquisition of preliminary data, 31 compounds from the MSMLS library were mixed into a single sample and diluted such that each compound was present at 1 µM in the final sample mixture. This mixture was diluted to a final concentration of 500 nM for analysis.

The single sample was analyzed via nanoflow LC-MS/MS using an UltiMate 3000 RSLCnano system (Thermo Scientific). Separation was achieved using reverse-phase nanoflow LC coupled to an in-house prepared fused silica capillary column packed with Jupiter C_{18} stationary phase that was maintained at 45 °C. Two chromatographic conditions were investigated, using acidic

(formic acid) aqueous mobile phase A (MPA), with acetonitrile or methanol mobile phase B (MPB). The flow rate was 0.3 µL/min, and compound separation was achieved using an 80-minute linear gradient from 2-95% MPB and holding at 95% for 45 minutes. 1 µL of sample was injected onto the column using the µL-injection mode. Spectra were acquired on an Orbitrap Exploris 480 MS (Thermo Scientific) with a resolution setting of 120,000 in MS1. Data-dependent MS/MS mode was used to acquire MS/MS data at 15,000 resolution, with the top 15 most abundant ions selected for MS/MS. Precursor ions were fragmented at normalized collision energies of 20, 40, and 60 using higher-energy collisional activated dissociation (HCD). No dynamic exclusion was used to ensure MS1 spectra were collected at a consistent interval throughout the data acquisition period. Peak detection to identify retention times for each metabolite was performed using Skyline version 23.1 and corroborated using MS/MS data.

2.2.2 Dataset curation

A comprehensive dataset for our ML model training and validation was created by cleaning, preparing, and combining a total of 143,542 compound entries from 43 different data sources. Accurate molecular identifiers and standardized chromatographic condition descriptors were confirmed and harmonized across all data sources. Furthermore, only datasets acquired under chromatographic conditions compatible with LC-MS acquisitions were considered.

RepoRT is a new open-source data repository that offers a standardized format for reporting of retention time data and accompanying experiment metadata[9]. This repository proved to be immensely beneficial in identifying chromatographic parameters that were not adequately documented in the original articles. Adoption of the formats represented in RepoRT has enabled the inclusion of many more instrumental acquisition parameters in our final curated dataset that were initially missing during literature dataset gathering.

The final dataset includes descriptors such as flow rate, temperature, column type, column dimension, column particle size, pH, total elution time, composition of mobile phases A and B, and %B (i.e., the chromatographic gradient). To ensure consistency across the various sources, several conversions were made. Temperature values were standardized to degrees Celsius, particle size to micrometers (µm), flow rate to milliliters per minute (mL/min), and retention time (RT) values to minutes. The %B value, representing the percentage of the organic phase in the mobile phase at a specific time, is documented as a percent value, colon-separated with the corresponding time. The composition of both A and B phases was rearranged such that the full chemical composition and associated quantities could be captured, and all chemical names were standardized to ensure consistent representation across data sources. This meticulous data preparation process resulted in a robust and comprehensive dataset for retention time prediction efforts.

2.2.3 Dataset representation

2.2.3.1 Description of categorical variables

In the dataset, certain features are represented as categorical variables, including pH, column type, and the composition of phases A and B. The decision to represent pH as a categorical variable was driven by the fact that many of the data sources did not provide numerical values for this feature. Categorical variables are particularly useful in such instances as they allow for the classification of data into distinct groups or categories. For pH, this was represented as acidic or basic. Similarly, column type and the composition of phases A and B are also

categorized, providing a structured and simplified way to understand and analyze these variables. This approach to data representation can help in handling missing or incomplete data, and can also aid in the interpretation and prediction processes of the model.

2.2.3.2 Chromatographic column information

Column descriptors included type, dimension, and particle size. The former was represented as one-hot vectors to indicate one of reversed phase or HILIC (the former was additionally split into HPLC and UPLC, though these would conceivably be encoded in the flowrate vector). Dimension was divided into scalars for length and width, and particle size was also represented as a scalar.

With respect to column type, the literature-sourced data had a significant imbalance, overrepresenting reversed phase configurations relative to HILIC. In all, there were 457 HILIC entries and 143,181 reversed phase, further split as 62,669 RPLC, 80,038 HPLC, and 474 UPLC.

We note that though column temperature was curated, the substantial amount of data missingness for this parameter across many data sources led us to drop this parameter from consideration as a model feature input.

2.2.3.3 Vectorization of mobile phase information

Mobile phase was represented as a vector of all unique phase components observed across the dataset. For example, position 1 represents water, position 2 represents ethanol, and so on. The proportion of each component is thus encoded in this vector such that it sums to one. We identified 6 unique mobile phase constituents across the literature dataset (water, methanol, acetonitrile, ethanol, acetone, isopropyl alcohol).

Note however that phase additives were captured separately, as sources did not report these values in a consistent manner (e.g., as percentage, concentration, or simply an indication of presence). To harmonize, we encoded a separate vector of mobile phase additives to indicate a binary presence/absence of a particular compound. While ideally we would be able to represent more precise values, this simplification represents an initial unifying solution. We identified 5 unique additives across the literature dataset (formic acid, EDTA, ammonium acetate, ammonium carbonate, and ammonium formate). In all, there were 39 unique phase composition and additive combinations across literature datasets.

2.2.3.4 Gradient length mapping

To enable comparisons across different experiment (also gradient) lengths, all time-domain values were normalized by dividing by experiment length. The experiment length is defined as the total time to the end of any chromatographic isocratic hold. This hold is typically applied after a gradient ramp of a specific mobile phase to ensure elution of all compounds of interest, including recalcitrant compounds that have a stronger affinity for the chromatographic column's stationary phase. We determined the gradient length for each dataset source using the reported chromatographic gradient and by applying the above definition. The time-domain normalization procedure was applied to individual retention times, flow rate, and gradient (%B) information. Both flow rate and gradient information were encoded as 64-length vectors.

2.2.3.5 Molecule representation

Molecules are initially represented as simplified molecular line input (SMILES) strings. These are converted to Deep Graph Library - LifeSci (DGL-LifeSci) molecular graphs[10], which creates features over nodes (atoms) and edges (bonds). This format is amenable to our selected molecular featurization subnetwork, AttentiveFP[11], which has demonstrated performance in prediction tasks.

2.2.4 Software framework

PrediToR has been organized as a Python package, with each piece of distinct functionality separated into core modules. This enables readability/usability, maintainability, and extensibility. The deep learning architecture was developed in PyTorch, with individual subnetworks responsible for each of the various input sources. These include AttentiveFP[11] for molecular featurization, a series of 1D convolutional layers to process gradient and flow information, and a series of fully connected layers to process phase composition, additives, and column metadata. These are then combined through concatenation and passed through additional dense layers to yield final retention time prediction activated by the sigmoid function, mapping outputs between 0 and 1. An overview of the network architecture is shown in Figure 1.

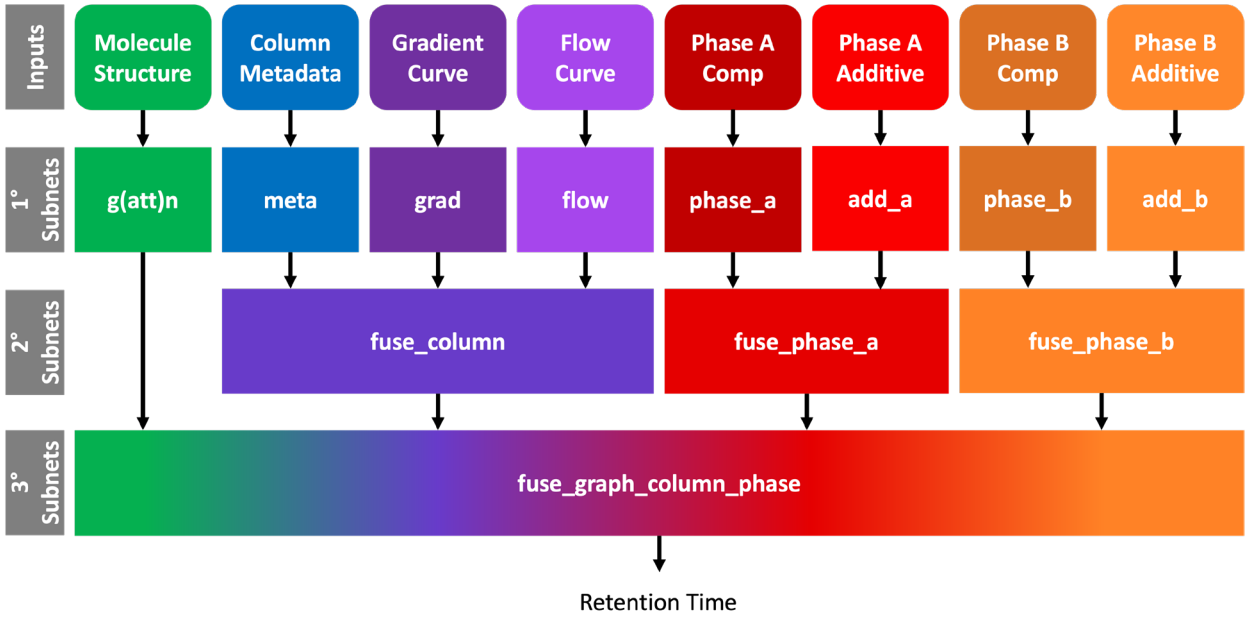

Figure 1. Network schematic. Depicted are model inputs, as well as primary, secondary, and tertiary layers. At each level, model abstractions are combined according to their contextual relatedness. For example, mobile phase A composition and additives are initially processed separately, but later combined to yield an abstraction of all mobile phase A information. Similarly, column metadata, gradient, and flow information are first processed separately, then combined to yield an abstraction of "column information". Finally, all information is unified before making the final prediction.

The training routine was designed to take advantage of graphics processing units (GPUs) on the internal high-performance computing resource Deception. The Adam optimizer[5] was used with default parameters and learning rate 1×10^{-4}. Loss was calculated as mean squared error

between predicted and target retention times, normalized by experiment length (i.e. predictions/targets range between 0 and 1).

2.3 Results & Discussion

Our need to develop a new and more generalizable retention time prediction model, whose architecture was described in Section 2.2.4, was driven by the poor performance of existing models on our empirical nanoLC datasets.

2.3.1 Existing retention time prediction tools are not extensible to nanoLC

We found that current retention time prediction methods performed poorly when applied to our empirical nanoLC-MS/MS small molecules data.

We trained and evaluated two different models, Retip[1] and METLIN[2], for retention time prediction of small molecules, with the goal of applying these models to our compounds of interest from the MS-MLS library. Our initial goal was to use RT prediction tools to expedite the compound identification process from empirical liquid chromatography-mass spectrometry measurements as well as to inform sample mixture design to achieve intended coelution conditions across the entire chromatographic time.

The Retip and METLIN models, respectively, were trained and validated using the Bruderer dataset[12], separated into two datasets containing measurements under acidic and basic conditions, respectively. The Retip models utilize chemical descriptors (CDK descriptors) of compounds as feature input while the METLIN model uses two-dimensional chemical fingerprints (Morgan, radius = 2) to represent compounds as feature input. We reimplemented the Retip approach to RT prediction on these datasets, that is, training and validating 4 different model architectures (random forest, XGBoost, Bayesian regularized neural network, and Keras), and then selecting the architecture that yields the best performance on the validation set for deployment. The train/validation split applied to Retip model architectures was 80/20, following the method outlined in Bonini et al. (2020)[1]. The METLIN model utilizes a feedforward neural network architecture, and a 75/25 train/validation split was implemented, following the method outlined in Domingo-Almenara et al. (2019)[2]. Note that because of the different train/validation splits between these two RT prediction models, these two models were trained on slightly different sets of compounds, which may affect validation performance.

These models were trained to predict one of three different endpoints: retention time (on the minute-time scale), percentage of mobile phase B (%B), or percentage of mobile phase B normalized to 0 – 1 scale (%B normalized). We investigated alternative endpoints to retention time to be able to generalize the trained models to our empirical nanoLC datasets, which were acquired with chromatographic lengths and gradients that are very different from the Bruderer dataset. Instead of utilizing absolute time as the prediction endpoint, prediction of %B allows generalizability to datasets acquired under different chromatographic gradients, and the predicted %B values can then be transformed back into retention time values according to the relevant chromatographic gradient.

In total, accounting for all combinations of model architecture, train/validation dataset, and prediction endpoint, 28 models were trained and validated. Of these, Table 1 below reports the performance of 8 models; note that the top-performing Retip model architecture for each combination of train/validation dataset and prediction endpoint is selected and reported below.

Table 1. Evaluation of retention time prediction models on validation dataset.

Train/validation dataset	Model	Prediction endpoint	Model RMSE	Model R^2	Model mean absolute error
Bruderer Acidic	Best Retip Acidic (Random Forest)	Retention Time	1.52	0.93	1.12
Bruderer Basic	Best Retip Basic (XGBoost)	Retention Time	1.13	0.92	0.71
Bruderer Acidic	METLIN Acidic	Retention Time	3.26	0.78	1.99
Bruderer Basic	METLIN Basic	Retention Time	2.59	0.77	1.31
Bruderer Acidic	Best Retip Acidic (Random Forest*)	%B	7.07	0.93	5.13
Bruderer Basic	Best Retip Basic (Random Forest)	%B	5.24	0.93	3.03
Bruderer Acidic	METLIN Acidic	%B normalized	0.16	0.70	0.12
Bruderer Basic	METLIN Basic	%B normalized	0.10	0.85	0.06

*Random Forest model was not the most performant model, but it was selected because the model fixed the predicted values between the minimum and maximum values of the training data.

We observe better performance (i.e., higher R^2 values) in Retip models compared to the METLIN models, under both acidic and basic conditions (Table 1). For Retip models, similar performance was achieved when predicting either retention time or %B ($R^2 > 0.90$).

Once trained and validated, these models were deployed to predict retention times of small molecules in our empirical nanoLC data, split into datasets acquired with two different organic mobile phases (acetonitrile or methanol). As our empirical data were only acquired under acidic (pH 2) conditions (see Section 2.2.1 for data acquisition parameters), we only used the models trained under acidic conditions. Our empirical dataset, used as the test set, comprises of 18 compounds, out of 31 total compounds in the sample mixture. 18 compounds were detectable in either the acetonitrile and/or methanol mobile phase B conditions; 13 compounds were detectable where acetonitrile was used as mobile phase B, while all 18 were detected with methanol as mobile phase B. Note that of the 18 compounds represented in the empirical

dataset, only 4 are shared with the Bruderer Acidic dataset (total of 228 compounds) split between training and validation sets. Retention times for the empirical dataset were curated for each detected compound, for comparison to predicted retention times.

We find that all retention time prediction models demonstrate terrible performance on our nanoLC dataset. Table 2 below reports the RT prediction results for each model, as compared to the empirical RT values. It is clear that the RMSE, R^2, and mean absolute errors reported for the test set are significantly higher than those observed in Table 1 for the validation set; this large discrepancy is observed across all models for all prediction endpoints. For example, the METLIN Acidic model demonstrated a mean absolute error of 1.99 min on the validation set but showed mean absolute error of 83.86 min for the test set. In practice, a mean absolute error of 1.99 min in a 25-minute run (data acquisition condition for the validation set) is well within acceptable ranges, but a mean absolute error of ~84 minutes in a 130-minute run (data acquisition condition for the nanoLC test set) is entirely unacceptable.

Table 2. Performance of retention time prediction models on empirical test set.

Test dataset	Model	Prediction endpoint	Model RMSE	Model R^2	Model mean absolute error
Empirical nanoLC, acetonitrile, pH 2	Best Retip Acidic (Random Forest*)	%B	55.42	0.10	52.75
Empirical nanoLC, methanol, pH 2	Best Retip Acidic (Random Forest*)	%B	54.79	0.19	49.18
Empirical nanoLC, acetonitrile, pH 2	METLIN Acidic	Retention Time	84.34	0.04	83.86
Empirical nanoLC, methanol, pH 2	METLIN Acidic	Retention Time	86.63	0.09	81.00
Empirical nanoLC, acetonitrile, pH 2	METLIN Acidic	%B normalized	0.55	0.06	0.51
Empirical nanoLC, methanol, pH 2	METLIN Acidic	%B normalized	0.55	0.07	0.48

*Random Forest model was not the most performant model, but it was selected because the model fixed the predicted values between the minimum and maximum values of the training data.

It is clear from these model predictions that the models are not generalizable to our empirical dataset, specifically to the underlying chromatographic conditions that produced this empirical dataset. The training and test set conditions exhibit such dissimilarity that it would be difficult to deploy trained literature retention time prediction models on our nanoLC data without having relevant nanoLC data in the training set, that is, nanoLC data in the training set that were acquired under similar conditions as the empirical dataset. Given this limitation of generalizability across different chromatographic conditions observed in literature RT prediction models, we investigated the feasibility of developing a new type of ML model architecture that could overcome current limitations.

2.3.2 PrediToR performs well on literature data

We describe a new type of deep learning model, named PrediToR (see Section 2.2 for model inputs and architecture details), that aims to predict retention times of small molecules under a broad range of chromatographic conditions. To train and validate this model, we curated literature data from a wide range of available data sources.

Curated literature data (N=143,542) from data sources described in Section 2.2.2 were split into 80% train, 20% validation, based on unique SMILES, and trained with batch size 64 for up to 1000 epochs. The network with the lowest validation loss was saved after each epoch. We were able to achieve a validation mean absolute error (MAE) of 0.032 for normalized retention times (that is, each experiment was normalized to be in the range 0 – 1). For comparison, Retip[1], a widely used ML-based retention time prediction tool and the tool we evaluated in the previous section, achieved ~0.5 minute MAE for experiments with 12-minute run times. In other words, ~0.042 for normalized retention times (conversely, our method should achieve ~0.38 minute MAE for a 12-minute run time). This amounts to an improvement of about 24%, and with the added benefit of generalizing to multiple, rather than just one, experimental conditions.

In an overly simplified extrapolation of PrediToR's MAE to longer experiment times that are much more in line with that expected for nanoLC-MS acquisition, this value would translate to approximately 4.2-minute MAE for a 130-minute run time, which would represent ~95% improvement over the METLIN Acidic model's performance on the empirical nanoLC dataset reported in Table 2. As greater run-to-run variability (on the order of minutes) is typically observed for nanoLC acquisitions in comparison to microflow or capillary LC, this simple-extrapolated MAE value aligns with retention time variability expectations.

A broader representation of performance is shown in Figure 2, a density plot of predicted versus experimental values from the validation set. A high proportion of predictions are distributed tightly around the ideal regressor (y=x). There are only a few low-density regions of significant error, distributed along low experimental retention times. That is, for these cases, the network predicts higher values than expected. Note also, however, that the density of low retention time data is much lower; additional low retention time training data would likely ameliorate predictions here.

We note that, due to the severe imbalance in column type (RP versus HILIC), it is unclear whether performance on HILIC predictions matches those of the average, as the dataset is so heavily dominated by reversed phase entries. To improve generalizability to column type (also type of chromatographic separation), we would have preferred additional training instances from HILIC experiments. This is certainly an area worthy of further attention in follow-on efforts.

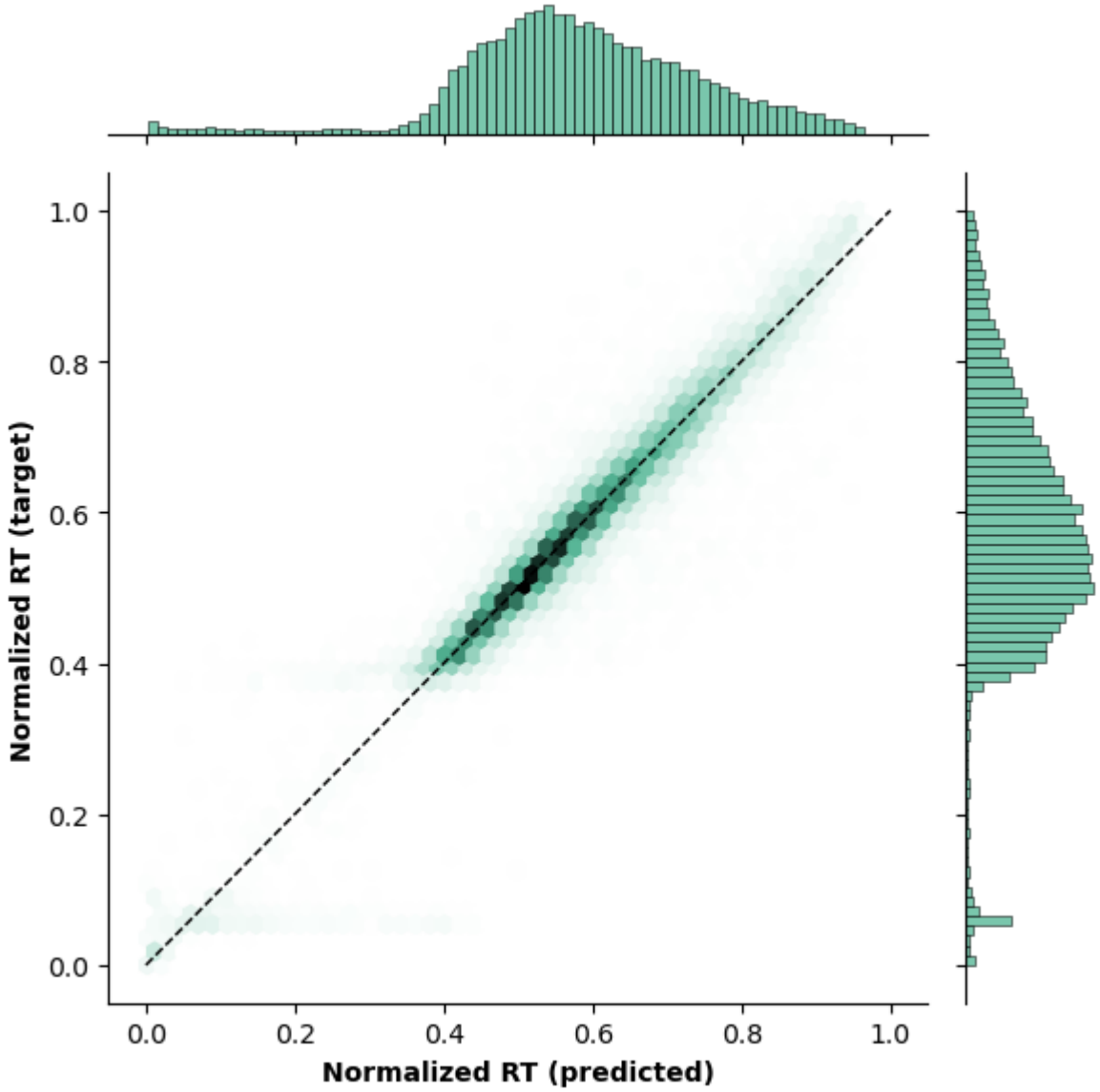

Figure 2. Retention time prediction validation result. Predicted and experimental retention times, normalized between 0 and 1 to compare across differing experiment lengths, are portrayed as a density plot to illustrate network performance. In the margins, the distributions of predicted and experimental values are shown.

Despite the dataset imbalance towards reversed phase chromatographic separation, PrediToR shows promise in its ability to generalize across a broader range of chromatographic conditions compared to those currently described in the literature.

2.4 Conclusions

We successfully developed a new, accurate, and generalizable ML model for retention time prediction and evaluated its performance on curated literature data. The data we utilized in training and validation consider a broader range of chromatographic conditions than observed in

previous models. As such, PrediToR shows promise in generalizing to nanoLC datasets, which will be examined in future efforts.

Additional efforts that have the potential to either boost model performance in a substantive manner or provide insight into model performance include: (1) molecule featurization, (2) bootstrapping and retraining the model on different train/validation/test splits, (3) examining model generalizability to conditions such as pH range and column type by holding out specific data, (4) comprehensive hyperparameter tuning, (5) input sensitivity analysis, and (6) exploration of network architecture variations.

Our original intent for RT predictions was to expedite the compound identification process from nanoLC-MS measurements and to inform sample mixture design. Though we were not able to evaluate PrediToR on our nanoLC-MS data, which still represent a chromatographic condition different from the training and validation data, we expect that the model would perform well under chromatographic conditions similar to the training and validation datasets (e.g., reversed phase HPLC or UPLC), for a diverse set of chemical classes, for these intended applications and more.

3.0 Statistical experimental design for reference-free quantitation of small molecules

3.1 Introduction

Non-targeted analysis using liquid chromatography-tandem mass spectrometry (LC-MS/MS) has become increasingly important as a qualitative analysis tool (that is, non-targeted analysis for compound identification), but its use as a quantification technique has been studied in a limited capacity. Quantification of molecules in complex samples is an essential capability across fields, including clinical[13, 14], environmental[15-17] and food safety[18, 19], forensics[20-22], and more. However, such methods conventionally rely on reference materials to build calibration curves or calculate relative quantitation estimates[23].

Instead, reference-free quantitation capabilities will enable us to better utilize the mass spectrometry (MS) data that we are already collecting for *qualitative analysis*, making past and future data amenable to *quantitative analysis* and concentration determination. Recent studies have both demonstrated the feasibility of a reference-free approach and laid the groundwork for both machine learning considerations and computational analyses in this space. However, concentration predictions are still often 1-2 orders of magnitude away from ground truth[24, 25] and there remain large gaps in understanding how instrumentation parameters affect quantification estimates.

Quantifying the absolute concentration of compounds requires development of calibration curves, which are considered compound-specific. The fit between known absolute concentrations for a specific compound and its measured responses (also known as response factor) is modeled, typically using simple linear regression, which is then deployed to predict unknown concentrations given empirical measurements of a compound's response. Various methods to calibration curve development and deployment have been implemented, though each come with a different set of limitations.

One common approach to calibration curve development uses reference standards and the creation of one curve per compound of interest, but this can be costly, time-intensive, and limited by the availability of reference materials for compounds of interest.

An alternative approach that attempts to overcome the limitations of reference material availability relies on developing calibration curves from structurally similar surrogate chemicals, e.g., derived from a parent-transformation product[26], under the assumption that structurally similar chemicals behave similarly. In the sense that reference materials for the compounds of interest are not needed, this approach can be considered a reference-free quantitation method. However, high structural similarity does not imply a similar compound's response; that is, even structurally similar compounds can exhibit different response factors, resulting in a different measured response. Errors ranging between four-fold and upwards of an order of magnitude when applying this approach to different chemical classes have been observed[26, 27].

A second reference-free quantitation approach implements a variation of the surrogate calibration curve concept to take advantage of neighbor compounds that are known to elute closely to the compounds of interest for quantitation[28, 29]. This approach requires use of chromatography as a separation technique and assumes that compounds eluting at similar retention times exhibit similar response. However, a drawback to this approach is the sensitivity of the method to elution profile changes under different chromatographic conditions. Further, the

potential for significant dissimilarity in compounds' response owing to differences in molecular structure challenges this underlying assumption.

Finally, a third reference-free quantitation approach leverages machine learning (ML) models trained to predict a compound's ionization efficiency[24, 30], which contributes to the relationship between target compounds and their measured response. This approach requires compounds to be ionized during data acquisition, as in mass spectrometry techniques. Ionization efficiency represents a neutral compound's ability to form charged ions in the gas phase and is quantified by the number of gas-phase ions that are generated per mole of compound. This is considered a chemical property intrinsic to a compound but also influenced by experimental conditions[30]. With this ML-based approach, the underlying assumption is that a compound's ionization efficiency is the predominant contribution to a compound's response factor.

In practice, this can be a two-step approach in deployment. Chemical information (e.g., molecule descriptors) of compounds of interest is used in a trained ML model (which can also include mass spectrometry features and solvent information as input) to predict ionization efficiency values[24], which are then corrected to account for instrument-specific contributions—empirically determined using reference data—to derive predicted response factors[24, 31]. Subsequently, concentration values for compounds of interest can then be directly predicted from the measured responses and predicted response factors.

The primary limitation with the ML-based approach is the need to derive an instrument-specific correction factor using reference compounds. Most accurate efforts in literature leverage a "universal ionization efficiency" prediction, which is then translated into an instrument and/or method specific ionization efficiency. These methods rely on a calibration data set of known concentration measured on the instrument and used to generate the translation equation used in samples of unknown concentration.

While this approach could be of interest for less common experimental conditions, there is a heavy reliance on anchor compounds[31] and reference data collected on two instruments. Further, current ionization efficiency (IE) ML models have demonstrated limited examination of effects of sampling and chromatographic conditions, focusing solely on contributions from mass spectrometry acquisition via direct injection[24, 32-34].

As such, we move beyond an approach that relies on iterative and chaining of experiment-experiment comparisons to obtain predicted absolute concentrations. Instead, we aim to improve upon the current reference-free quantitation paradigms by deriving a unified solution to quantified contributions of experimental conditions (including chromatographic conditions such as nanoflow liquid chromatography) on response factor, which can then be directly implemented to obtain predicted concentration values.

In this work, we introduce an alternative approach to reference-free quantitation that does not require reference or anchor compounds nor their ionization efficiency values. Building upon mass spectrometric parameters known to influence compound response, we aim to build a model for concentration estimation that is neither instrument nor experiment specific. Through an experiment design that enables maximizing characterization of sampling and chromatographic conditions, we are able to evaluate the importance of these factors on the measured mass spectrometry signal. Following this elucidation of experimental contributions to response factor, we investigate the inclusion of predicted "universal" ionization efficiency values into our model for a more precise response factor prediction.

We envision that development of such an alternative reference-free quantitation capability and providing the framework in an open format would enable end-users (e.g., chemists with little informatics experience) to:

- Utilize the provided model(s) to obtain more accurate concentration value predictions if using a similar instrumental setup, and/or
- Apply a similar statistical design framework to establish response factor contributions given their instrument's responses for reference-free quantitation if using a different instrumental setup.

As such, we focus on the following areas for investigation:

- Derive a more precise response factor that accounts for sampling and chromatographic conditions using a statistical experimental design approach,
- Develop an ML model to predict concentration values for different compounds given a more precise response factor, and
- Provide a statistical experimental design and ML model framework for reference-free quantitation in an open format for reuse.

We find that sampling and chromatographic conditions contribute substantively to response factor, and some even more so than ionization efficiency. This has implications for quantitation accuracy, especially since previous efforts have focused much more heavily on examining ionization efficiency and any effects of experimental conditions have received little attention. Evaluation of our developed response factor model demonstrated approximately 10X improvement in quantitation accuracy over existing ionization efficiency models. To facilitate sharing of this capability with end-users, we organize our developed products into workflows and scripts that capture an end-to-end pipeline, from the statistical design of experiments framework to model training, and finally to model testing with new data. This new approach has advanced quantitative non-targeted analysis.

3.2 Methods

To derive a more precise response factor for reference-free quantitation, we investigate the role of 8 experimental and sampling parameters in compound response on a nanoLC-MS/MS platform using a fractional factorial experiment design. We additionally investigate the importance of these parameters with logIE predictions. We implement a modified version of literature logIE prediction frameworks. Furthermore, we implement two workflow files for training and testing these models. These workflows support users in training their own models and/or testing new data on any developed models.

3.2.1 Statistical design for study parameters and rationale

A design of experiments (DoE) is needed for a quantitative non-targeted assessment of liquid chromatography-mass spectrometry data. The overarching goal is to collect a high-quality dataset that covers a suitable range of instrumentation factors for predicting compound response. The goal of the data collection is to (1) identify key parameters not yet explored, (2) quantify the "bounds of uncertainty" by collecting at the minimum and maximum of parameter

space, and (3) build a very high-quality data set. Generally, we were interested in an experimental design that is amenable to iterative collection, starting with a screening experiment that could support potential follow-up investigation of a fine-grid collection over levels of significant factors. A wide variety of statistical designs were considered: screening designs (e.g., fractional factorial, plackett burman, definitive screening designs)[35, 36], space-filling designs (latin hypercube)[37], D-optimal designs[35], factorial designs[35, 36, 38], response surface designs (e.g., central composite, box-behnken)[35, 37, 38], and robust parameter designs[37].

To ensure that the selected statistical design was appropriate for our study parameters, we first examined the parameter type and potential levels. The 8 factors considered in our study include both categorical (also discrete) factors and continuous factors. Categorical factors include (1) organic solvent (i.e., mobile phase B) and (2) pH. These parameters are explored at the levels of acetonitrile and methanol at pH 2 and 8. The pH factor is driven by the organic modifier (e.g., ammonium salt, formic acid) and organic modifier quantity. Continuous factors include (1) chromatographic flow rate, the rate at which solvent flows across the chromatographic column, (2) electrospray ionization (ESI) voltage, the voltage applied to the electrospray nozzle for ionization, and (3) sample loading, the volume of sample injected onto the column (which is related to the amount of sample injected). To better understand any effects of sample complexity, we additionally examine (1) the number of compounds in a sample, meant to capture sample level or “global” level of suppression with higher complexity, (2) extent of co-elution, the percentage of compounds in the sample experiencing at least 25% co-elution[39], which captures compound level or “local” ion suppression, and (3) concentration ratio, the factor by which the concentration is modified for half of the compounds in the sample, to capture the magnitude of ion suppression or enhancement observed in a compound pair.

Given the study parameters described above, we elected to apply a fractional factorial experiment with a resolution of 4 and screens 8 factors for initial screening. Of the 8 factors considered, 6 are continuous and 2 are categorical. Extrema are investigated in addition to mid-points collected at each combination of categorical level. An example of exploring the parameter space is shown below in Figure 3. In addition to the categorical factor levels described above, we selected the following values for each parameter for investigation through this statistical design: (1) flow rate: 200, 250, 300 nL/min; (2) ESI voltage: 1500V, 2250V, 3000V; (3) sample loading: 1000 nL, 1750 nL, 2500 nL; (4) number of compounds: 10, 15, 20; (5) extent of co-elution: 0%, 50%, 100%; and (6) concentration ratio: 50% (depression of one in a pair), 100% (no modification in a pair), 150% (elevation of one in a pair).

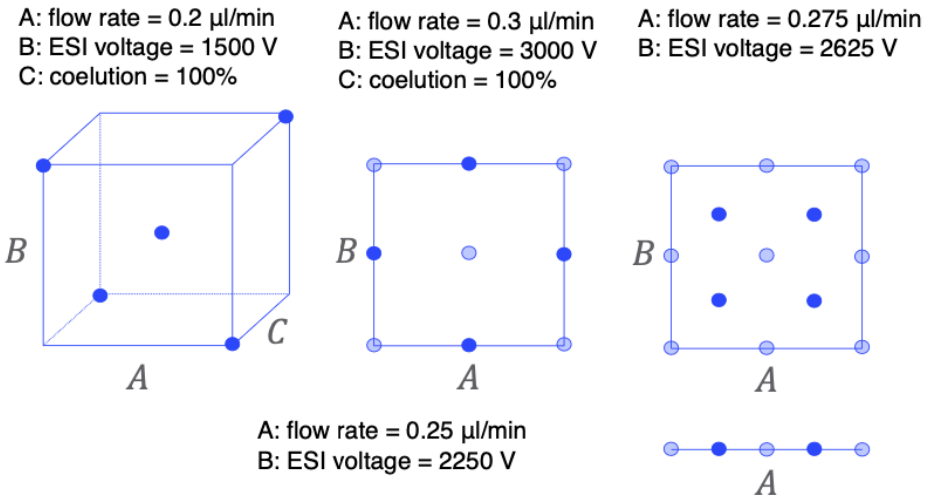

Figure 3. Graphic describing the progression of a factorial design with iterative data collection steps. A fractional factorial design is displayed on the left, where only vertices (representing parameter space extrema) and center-points (midpoints of parameter space) of variables of interest are sampled. With iterative data collection, moving left to right and then down, variables that remain of interest are sampled more finely in a grid format.

Although we focus on the design for initial screening, we confirm that the fractional factorial design is amenable to iterative and deeper investigation of parameters of interest with additional data collection.

We see that following this screening approach, a response surface design (face-centered central composite design, or CCF)[36, 37] for a subset of the parameters can be implemented if curvature is identified. Finer grids of parameters of importance can also be introduced using embedded factorials or direct grid collections. This type of experimental design allows for (1) built-in "replication": even within one replicate, each high/low level for each parameter is measured multiple times, (2) main effects are not confounded with first order interaction effects, (3) the collections are iterative and initial collections are subsets of future collections, and (4) factorial and response surface designs are replete in ESI parameter effects studies[40]. Challenges with this approach are that it (1) assumes all treatment combinations of levels are possible, (2) if we have more than three factors for further study, response surface can require a lot of samples, and (3) categorical factors require repeating draws for each level, which can become costly if the categorical factors are of high interest. If it is determined that interaction effects do not exist or are negligible, a fine grid collection for variables of interest can be implemented instead.

As such, we can have confidence that the initial screening design contains sufficient flexibility for deeper investigation. In addition to the 8 factors and their levels identified above, we collect at 6 distinct concentrations across a total of 20 treatments (16 treatments and 4 center-points) that cover the full parameter space with duplicate acquisitions for a total of 240 runs.

3.2.2 Sample design and instrument method development

Samples were created using the Mass Spectrometry Metabolite Library of Standards (IROA Technologies). The Mass Spectrometry Metabolite Library of Standards (MSMLS) is a collection of high-quality small biochemical molecules that span a broad range of primary metabolism.

This library of standard reference materials comprises 603 unique compounds present in individual wells, each containing 5 µg of compound. Compounds in the MSMLS library were dissolved according to the manufacturer's instructions using 20 µL of either 5% methanol in water or a 1:1 chloroform:methanol mixture.

For initial method development and acquisition of preliminary data, 31 compounds from the MSMLS library were used. Various LC and MS parameters were evaluated to establish maximum and minimum values to be used in this reference-free quantitation study.

Flow rates ranging from 0.2 - 0.43 µL/min were evaluated, and it was determined that exceeding 0.4 µL/min consistently resulted in backpressures exceeding the pressure limits of the pump. The addition of a capillary heater to maintain the column at 45 °C provided consistent column temperature and helped alleviate high backpressures that were consistently problematic when using methanol as the organic phase. Various linear gradients and method times were evaluated to ensure compound separation and successful elution of highly nonpolar compounds. All injection modes possible on the system were evaluated. These include full-loop, partial-loop, and µL-injection modes. The µL-injection mode was ultimately chosen as it provided the best peak shape and most consistent peak area. ESI voltages ranging from 1500 – 3000 V were evaluated.

Once the instrument parameters were evaluated and validated, an initial screening of 566 compounds from the library was performed to evaluate the impact of chromatographic condition on compound retention, to create sample mixtures comprising of compounds whose elution behavior aligned with the intended chromatographic conditions from the fractional factorial statistical design. More specifically, compounds were selected such that some elute independently while others co-elute in analyte pairs consistently across all chromatographic conditions. Compound selection criteria included (1) the exclusion of compounds not retained, (2) exclusion of ultra broad peaks, (3) consistently co-eluting or not with other compounds across the 4 solvent conditions, and (4) maximizing class diversity where possible.

To ensure the sample design adheres to these criteria, a down-selected set of 53 compounds were used to create six mixtures. These samples were used to confirm that the selected compounds eluted independently or in a pair across all chromatographic conditions as expected. This additional screening was used to create the 5 final sample mixtures considered in this study. Four of the 5 mixes have the concentration ratio varied, resulting in 9 total sample mixes.

3.2.3 Sample preparation for response factor model development

To determine instrument response under varying chromatographic conditions and MS parameter settings, samples were prepared using the MSMLS. Nine mixtures containing different chemical compositions were created, containing 10-20 compounds at varying concentrations and six dilutions, for a total of 54 unique samples. These samples cover the concentration range of 0.5 - 750 nM.

3.2.4 Data acquisition for response factor model development

Data were acquired for these 54 samples under four optimized and validated chromatographic conditions, using either acidic (formic acid) or basic (ammonium acetate, pH 8) aqueous mobile phase A (MPA), with acetonitrile or methanol mobile phase B (MPB). The flow rate was set to

either 0.2, 0.25, or 0.3 µL/min. Compound separation was achieved using different linear gradients and run times depending upon flow rate as follows: 0.3 µL/min used 80-minute linear gradient from 2-95% MPB and holding at 95% MPB for 45 minutes with 160 minutes of total acquisition time, 0.25 µL/min used 140-minute linear gradient from 2-95% MPB and holding at 95% MPB for 20 minutes with 180 minutes of total acquisition time, and 0.2 µL/min used 180-minute linear gradient from 2-95% MPB and holding at 95% MPB for 15 minutes with 210 minutes of total acquisition time. 1, 1.75, or 2.5 µL of sample was injected onto the column using the µL-injection mode.

Positive ESI voltages of 1500, 2250, or 3000 V were used, and spectra were acquired on an Orbitrap Exploris 480 MS (Thermo Scientific) with a resolution setting of 120,000 in MS1. MS/MS data were acquired at 15,000 resolution using a targeted inclusion list. Precursor ions were fragmented at 30% normalized collision energy using higher-energy collisional activated dissociation (HCD). No dynamic exclusion was used to ensure MS1 spectra were collected at a consistent interval throughout the data acquisition period. Peak detection to identify retention times for each metabolite was performed using Skyline version 23.1[41] and corroborated using MS/MS.

In total, 240 datasets representing two sets of sample replicates were generated using the fractional factorial experimental design described above in Section 3.2.1.

3.2.5 Mass spectrometry data processing

Follow data acquisition via nanoLC-MS/MS, raw mass spectrometry datafiles were imported into Skyline version 23.1[41], a software tool used to analyze LC-MS data for a set of targeted compounds. Here, Skyline was used to confirm peak detection and to perform peak picking for target compounds. Each of the 120 datasets, collected in 16 batches, were processed separately, and the following adducts were considered for each compound: $[M+H]^+$, $[M+Na]^+$, $[M+K]^+$.

To disambiguate and confirm compound identifications, MS2 spectra were examined using FreeStyle (version 1.8, Thermo Scientific). Reference MS2 spectra that were acquired on a similar MS platform, obtained from sources such as the Human Metabolome Database (HMDB)[42] and MassBank of North America (MoNA), were used in comparison to empirical MS2 spectra.

Once the peak for each detectable target compound was selected, the protonated precursor trace was retained in the Skyline file, while the others were removed. If the peak selection was ambiguous, the other precursors were included in the final report for that compound, to alert to the ambiguous detection.

The final report was then generated, which included the following columns: File Name, Sample Name, Molecule Name, Precursor Mz, Precursor Adduct, Precursor Charge, Isotope Dist Index, Min Start Time, Max End Time, Best Retention Time, Min Best Retention Time, Max Best Retention Time, Total Area MS1, Product Mz, Product Adduct, Product Charge, Retention Time, Start Time, End Time, Area, Height, Points Across Peak, Fwhm, Fwhm Degenerate, Chromatogram Extraction Width, Raw Number of Points, Raw Times, Raw Intensities, Raw Mass Errors, Mass Error PPM, Chromatogram Source, and Coeluting. This comprehensive report was then passed on for further analysis and interpretation.

Further analysis of the comprehensive report, produced for each dilution acquired under a specific experimental treatment, entailed several cleaning and filtering processes to obtain accurate empirical measurements of MS1 area for each target compound, that are used for downstream response factor calculation. For each target compound, only the monoisotopic peak [M] was retained. Ambiguous detections of the target compound, either from inability to assign a chromatographic peak to a compound or owing to low peak signal, resulted in its removal from downstream analysis; the target compound is registered as a non-detect. In practice, to alert to ambiguous detections, all potential adducts for these compounds were retained in the report. For unambiguous peak detections, all protonated precursor traces with precursor m/z ≥ 100 were retained. This minimum m/z threshold aligns with a low m/z cutoff that was applied to MS1 scans during LC-MS/MS acquisition. This cleaned report was then used in response factor calculation efforts.

3.2.6 Response factor calculation

We examined two different methods of calculating response factor from measured mass spectrometry response and known concentration values for each sample mixture under a unique set of chromatographic conditions.

Constant Response: Response factor (RF) calculations of compounds using a least squares regression for an example dilution series using the equation $\boldsymbol{Y_{obs}/Conc = RF}$ is shown in Figure 4A. The issue in calculating an RF with this equation for data with a dilution factor is two-fold: (1) the unequal spacing of concentration values results in elevated influence of higher concentration data points on the regression, and (2) the non-uniformity in absolute error of replicates across the concentration range (heteroscedasticity) violates the least-squares regression assumptions. As suggested in the Groff et al. (2022)[25], performing the regression on log10-transformed values overcomes both issues and enables calculation of RF directly from intercept in the case of the slope equaling 1. An example of the log transformation is shown in Figure 4B, where the equation $\boldsymbol{10^{b'} = 10_{log}(Y_{obs}/Conc) = RF}$ is applicable. This approach to response factor calculation is further referred to as constant response factor or CRF. In the implementation of the CRF calculation, only compounds with at least 3 observations across dilution series and replicates have a value calculated, otherwise a numpy.NaN (not a number) value is returned. This limit is in place due to the degrees of freedom requirements. A drawback of this approach, in assuming a constant response, is that compounds' differing linear dynamic range could result in a divergence from a slope of 1 when looking at many compounds within a specific concentration range.

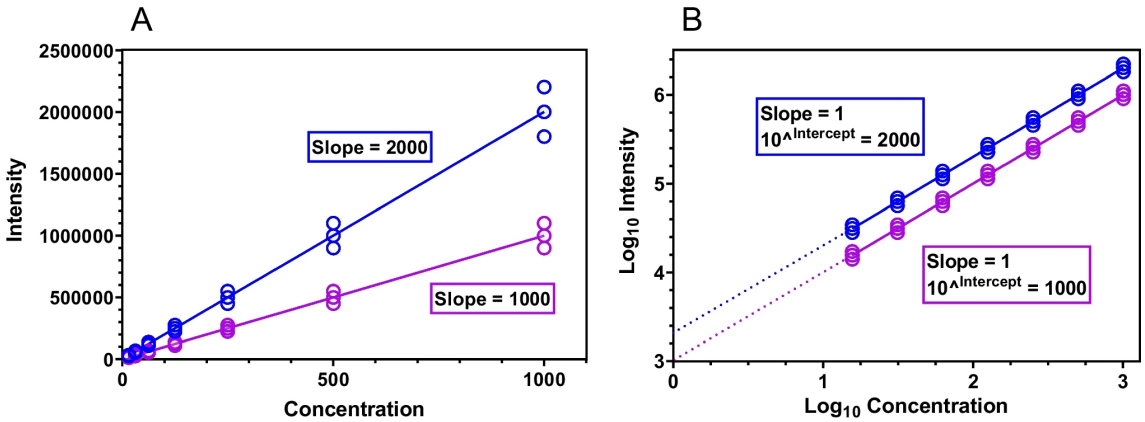

Figure 4. Adapted from Groff et al. (2022)[25]. Linear models fitted to (A) absolute intensity and absolute concentration, and (B) log10-transformed intensity and log10-transformed concentration. Slope values after log-log transformation are 1 for both curves, which meet the underlying assumptions for validity in deriving a chemical-specific response factor.

Bootstrapped Response: As an alternative strategy to CRF, a boot-strapped response factor can be calculated, as suggested in the literature, with the assumption that there is no constant response factor for a compound[25]. Due to this assumption, the boot-strapped response factor is calculated using data from replicates 1 and 2 within treatments (N=20). The resultant RF is calculated as a 95% prediction interval (PI) and is further referred to as the boot-strapped response factor or BRF.

3.2.7 Ionization efficiency

A positive mode ionization efficiency model was re-implemented in Python as described in Liigand et al. (2020)[24] using PaDEL and 5 solvent descriptors (pH_aq, viscosity, polarity index, surface tension, nh4 presence). Using the methods described, 326 PaDEL descriptors remained after pruning. The “universal” logIE values (or comparable to an Agilent XCT) are used to train and predict logIE.

Going beyond the methods described in literature[24], the Python package sklearn was used to perform a 5-fold cross validation using the random forest regressor in our implementation. A grid search was performed to search the hyperparameter space for number of estimators, max tree depth, and min samples split. The parameters that yielded the best model and used in this study are 300 estimators, max depth of 20, and 10 minimum samples to split a node. Trained on a negative mean signed error score, the resultant score for the selected model is 0.25 mean signed error. The mean logIE of the entire training dataset is 2.91 ± 1.14 (s.d.).

The top important features above a threshold of 0.005 are shown in the bar graph below (Figure 5). Note that all 5 solvent descriptors are in the top important features, indicating that experimental conditions are important contributors to a compound’s ionization efficiency.

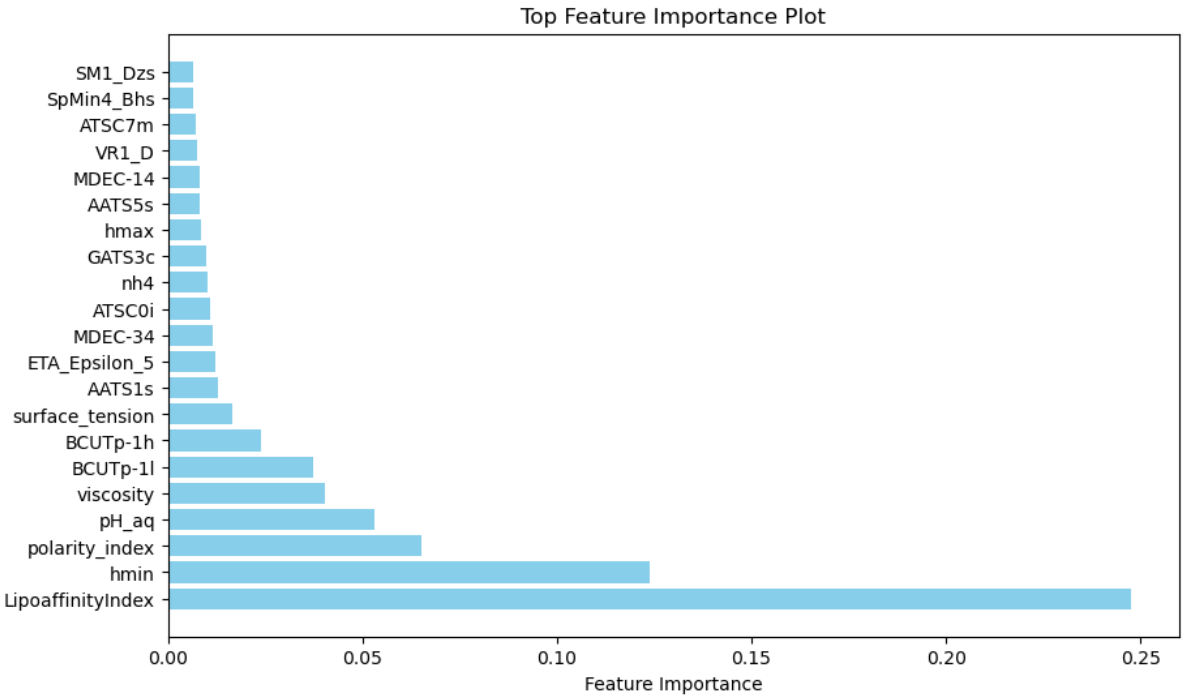

Figure 5. Bar plot displaying feature importance from the ionization efficiency model.

3.2.8 Parameter encoding

For the ionization efficiency predictions of all compounds studied, PaDEL descriptors were calculated using the SMILES supplied by the MSMLS library document. Eluent, additive, and column 2 representing the fraction of mobile phase B are used to calculate viscosity, surface tension, polarity index, and nh4 presence/absence for all conditions of interest[24]. These four descriptors, pH(aq), and the 326 PaDEL descriptors of importance are supplied to the random forest model for a "universal" logIE prediction.

Table 3. Descriptors related to experimental conditions that are calculated and contribute to a "universal" log10-transformed ionization efficiency prediction.

eluent	pH_aq	additive	2	3	viscosity	surface_tension	polarity_index	nh4
methanol	2	formic acid	1	0	0.513	194.922	5.1	0
methanol	8	ammonium acetate	1	0	0.513	194.922	5.1	1
acetonitrile	2	formic acid	1	0	0.280	308.016	5.8	0
acetonitrile	8	ammonium acetate	1	0	0.280	308.016	5.8	1

For training response factor models, the eluent descriptors are encoded as categorical values of 1 and 0 representing acetonitrile and methanol, respectively.

3.2.9 Response factor model design

Four linear models were created using the data acquired from the fractional factorial collection. Linear models offer high interpretability, enable easier investigation into parameter importance

when we only have extremes and mid-points for parameters, and allow us to focus on first and second order interactions of parameters. This approach additionally supports modified parameter encodings, e.g., quadratic terms, based upon learned information from parameter-response relationship investigations.

To develop a framework that explores contribution of the study parameters to chemical response, both CRF and BRF response calculation methods are used in model training. For each, 2 types of models are developed, yielding 4 model variants (Figure 6). The first model returns a sample-specific, rather than a chemical-specific RF prediction. The sample-specific CRF prediction is based upon a linear model that is purely a function of experiment parameters and measurable sample descriptors. The modified version of this model trained on BRF values. Models discussed in this study are trained on the mid-point of the BRF 95% PI. However, future work could leverage an interval regression framework trained on the upper and lower bounds of the PI. The chemical-specific RF models build upon these frameworks with the inclusion of a predicted logIE value.

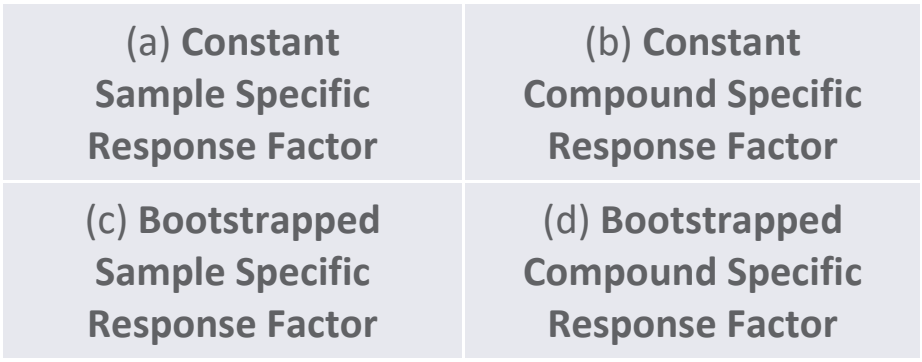

Figure 6. Graphic delineating the four different response factor model variants, using either the constant or bootstrapped response factors, in combination with either the sample-specific or compound-specific response factor models.

Sample level response: Response factor range captures the known and observable study parameters and does not factor chemical information in the prediction. Applicable to non-targeted analyses with features not identified.

Compound specific response: Response factor range captures the known and observable study parameters and does account for chemical information in the prediction. Applicable to studies with identified features.

Constant vs. Bootstrapped response: Two versions of the same sample and chemical response models trained using the different calculation approaches.

3.2.10 Model prediction of absolute concentration

Implementation of the response factor linear model during deployment, to predict concentrations of unknown compounds, uses a slightly modified set of parameters compared to the training of response factor linear model. We describe some of these differences below.

Using the parameter criterion previously described for the model inputs, a 25% feature overlap threshold is used to define if a feature experiences co-elution. The percentage of features (i.e., unidentified/unassigned compounds) in the sample experiencing co-elution is supplied to the model for prediction.

As relative concentration (related to the concentration ratio parameter in our study design) between features cannot be ascertained from an unknown sample, this parameter is set to a value of 1 ($[conc_A]/ [conc_B]$) by default in prediction workflows. By doing so, we do not make any assumptions regarding relatedness of co-eluting compounds. In practice, with this default value, the coefficient associated with this variable is treated as error and incorporated into the error term of the linear model.

For unknown samples or test data, the MS1 areas of detected features (i.e., unidentified/unassigned compounds) are divided by the predicted response factor to yield a concentration estimate on the nM scale.

3.2.11 Blinded sample preparation and data acquisition

Sample Preparation: To evaluate the model, a blinded sample comprising 20 compounds was created. These compounds were prepared from the MSMLS, and used previously in this study, but to varying degrees of scrutiny. Seven compounds from the response factor model (see Section 3.2.3), six compounds from a late-stage screening effort but ultimately not included in the final compound selection (see Section 3.2.2), and seven compounds only used in the initial screening of the MSMLS library (see Section 3.2.2) were used. Seven dilutions of this new sample mixture were prepared, resulting in compound concentrations both above and below the concentration range (0.5 - 750 nM) that was used to develop the response factor model.

Data Acquisition: Data corresponding to two sample replicates for each of the 7 dilutions were acquired under acidic aqueous conditions with acetonitrile as the organic phase. The flow rate was set to 0.3 µL/min, and compound separation was achieved using an 80-minute linear gradient from 2-95% MPB and holding at 95% MPB for 45 minutes with 160 minutes of total acquisition time. 1 µL of sample was injected onto the column using the µL-injection mode. Positive ESI voltage of 2250 V was used, and spectra were acquired as described above (see Section 3.2.3).

Data Processing: Following data acquisition, mass spectrometry datafiles were processed using Skyline, as described in Section 3.2.5.

3.2.12 Software framework

Two snakemake workflows are available: one to train the models and a second one to predict on unknown or test data. Experiment data needs to be modified to match the example input files and the configuration file updated with values reflecting the parameters for the new data.

3.3 Results & Discussion

We demonstrate the ability of our statistical framework to examine the effects of 8 sampling and chromatographic conditions on mass spectrometry response and the extent to which these variables contribute to response factor. We further examine the accuracy of absolute concentration predictions and demonstrate model performance on a blinded sample.

3.3.1 Analysis of fractional factorial collection

We first analyze the mass spectrometry response, that is, the MS1 area, of each detected compound under a specific set of experimental conditions (also known as treatment) and at a specific known concentration, to characterize compound detectability within each treatment.

A full manifest of the 8 factor levels for each of the 20 treatments is shown below in Table 4. Additionally described in this section are the comparisons of the observable sample complexity factors to expected levels. Despite deviations in the expected number of compounds and extent of co-elution observed, expected values in the experiment design are used in all model training. Possible sources contributing to these deviations include ion suppression due to sample complexity factors or a combination of other factors.

Table 4. Description and parameters for each of the 20 experimental treatments investigated for response factor model development.

	Organic phase	pH	Flow rate (nL/min)	ESI voltage (V)	Sample loading (nL)	Conc. ratio (%)	Number of compounds	Co-elution (%)
1	ACN	2	300	1500	2500	150	20	0
2	ACN	2	300	3000	1000	50	20	100
3	ACN	2	200	1500	1000	50	10	0
4	MeOH	2	200	3000	2500	50	20	0
5	MeOH	2	200	1500	1000	150	20	100
6	MeOH	2	300	3000	1000	150	10	0
7	MeOH	8	200	3000	1000	50	10	100
8	MeOH	8	200	1500	2500	150	10	0
9	MeOH	8	250	2250	1750	100	15	50
10	ACN	8	200	3000	1000	150	20	0
11	ACN	8	250	2250	1750	100	15	50
12	ACN	8	300	1500	1000	150	10	100
13	MeOH	8	300	3000	2500	150	20	100
14	MeOH	8	300	1500	1000	50	20	0
15	ACN	2	250	2250	1750	100	15	50
16	ACN	2	200	3000	2500	150	10	100
17	MeOH	2	300	1500	2500	50	10	100
18	MeOH	2	250	2250	1750	100	15	50
19	ACN	8	300	3000	2500	50	10	0
20	ACN	8	200	1500	2500	50	20	100

Our study design includes 3 parameters aimed to capture effects of sample complexity on response factor. The goal in modifying the number of compounds in a sample as a variable is to better understand the importance of global/sample-level ion suppression. The other two variables considered are the percentage of compounds experiencing co-elution and the concentration ratio between compounds that experience overlap. The latter two study parameters are designed to capture local/compound-level ion suppression. Discussed in the next section are observed deviations from expectations given the implemented study design.

3.3.1.1 Number of compounds

Figure 7 depicts some deviations in observed from expected number of compounds in replicates across the 20 different treatment conditions. Small deviations between replicates and higher than expected values can be attributed to the observation of in-source fragments. For example, Treatment 1 samples are expected to contain 20 compounds, but we detect 21 compounds in Replicate 2, owing to the additional detection of an in-source fragment of omega-hydroxydodecanoate, as $[M\text{-}H_2O+H]^+$. Treatments 4 and 10 exhibit large deviations less the expected number of compounds. It is difficult to attribute the source of non-detects and presence of in-source fragments, though ion suppression, low ionization efficiency, or column retention beyond the chromatographic acquisition could individually or in combination result in non-detection of expected compounds, and the intrinsically labile nature of certain ionized compounds in the gas phase could result in the propensity to fragment in the ionization source, thereby producing in-source fragments.

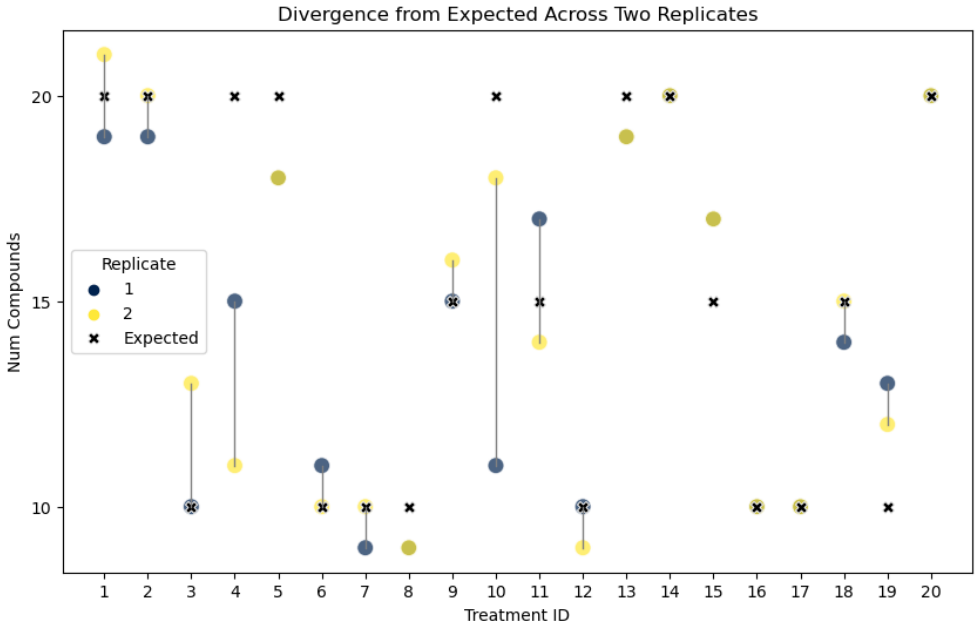

Figure 7. Plot displaying the number of detected compounds in each sample, per replicate, per experimental treatment, and deviation from expected number of compounds in each sample.

3.3.1.2 Extent of co-elution in sample

Mix 2, used in Treatments 1, 4, 10, and 14, consists of 20 compounds expected to have 0% co-elution, where half are sourced from a co-eluting analyte pair and half are not. Treatments 1 and 14 exhibit great reproducibility. Treatments 4 and 10 exhibit poor reproducibility between replicates and a third replicate would be recommended. Two compounds, cytidine 2',3' cyclic

phosphate and 2-methoxyestrone, appear to be suppressed in Treatment 10 such that they are only observed in 3 and 1 dilutions, respectively, out of 6 each, across both replicates. Interestingly and perhaps challenging to reconcile, 2-methoxyestrone has one of the higher predicted logIE values for the compounds in this study. Furthermore, while these samples are expected to experience 0% co-elution, the replicates in fact exhibit co-elution levels of 54% and 58%, respectively. Treatment 1, which similarly contains 20 compounds at 0% co-elution, exhibits much greater reproducibility and co-elution levels below 20%. Of note in Treatments 1, 4, 10 is the inability to separate corticosterone from cortexolone, with the same peaks representing both compounds, thus driving the observed co-elution rate higher. They are constitutional isomers with the sole difference in positioning of a hydroxyl group. As such, it is not surprising that they experience high degrees of co-elution and are difficult to deconvolute using fragmentation patterns.

3.3.1.3 Ion suppression and enhancement

Here, we characterize the extent of ion suppression or enhancement compared to expectations. According to Kruve (2020)[43], it can be difficult to observe ion enhancement. We leverage Equation 1, as described in Remane et al. (2010)[39], where the extent of ion suppression or enhancement can be interpreted from a negative or positive value, respectively. It was previously found that drug-like compounds detected via ESI that experience co-elution exhibited suppressions ranging from 25% to 73%, and at levels greater than observed via APCI[39]. Of the 11 co-eluting analyte pairs identified, 2 elevated compounds were suppressed. Interestingly, the median level of interaction for elevated compounds is 1.435×10^3 % enhancement, which suggests that our compound pairs tended towards ion enhancement. One outlier was observed, which is enhancement of thiopurine s-methylether over methylthioadenosine at the level of 2.204×10^6 %.

$$ion\ suppression\ or\ enhancement = 100 \cdot \left(\frac{Y}{X} - 1\right)$$

Equation 1. Calculation of empirical ion suppression or enhancement in co-eluting pairs.

3.3.2 Ionization efficiency

We reimplemented the ionization efficiency model first described in Liigand et al. (2020)[24] to be included as a contribution to response factor in chemical-specific RF model variants described in Section 3.2.9. “Universal” logIE values,[31] or those comparable to values collected on an Agilent XCT, were calculated using our implementation of a random forest model. For each of the four chromatographic conditions, Table 5 shows the predicted logIE values for all 29 unique compounds investigated in this study. Future efforts could explore inclusion of the PaDEL and solvent descriptors found to be relevant to IE prediction as variable inputs to our linear models instead of a predicted logIE value.

Table 5. List of universal logIE values for each compound under the 4 different chromatographic conditions.

condition	ACN_2	ACN_8	MeOH_2	MeOH_8
Molecule				
2-METHOXYESTRONE	3.162034	3.031174	3.144708	3.015541
3-ALPHA,11-BETA,17,21-TETRAHYDROXY- 5-BETA-PREGNAN-20-ONE	2.674947	2.519326	2.666990	2.519133
3-NITRO-L-TYROSINE	1.683897	1.554420	1.686756	1.559399
4-COUMARATE	1.865913	1.762601	1.885056	1.772273
4-METHOXYPHENYLACETIC ACID	2.017298	1.945033	2.024389	1.951646
BIOTIN	2.236682	2.260226	2.234876	2.267088
CAPRYLATE	2.199648	2.123941	2.204494	2.125833
CORTEXOLONE	2.630393	2.456149	2.629734	2.461303
CORTICOSTERONE	2.710116	2.480873	2.708325	2.484708
CORTISONE	2.892532	2.950235	2.797581	2.848218
CYTIDINE 2',3'-CYCLIC PHOSPHATE	1.948783	1.914400	1.948865	1.915696
DEOXYCHOLATE	2.262065	1.827318	2.198620	1.784280
DEOXYCORTICOSTERONE ACETATE	2.572829	2.148940	2.543054	2.163288
ESTRADIOL-17ALPHA	3.317861	3.051382	3.297456	3.047985
GLUTAMATE	1.601159	1.511776	1.600668	1.514169
HOMOVANILLATE	1.610137	1.462461	1.612542	1.469684
LIOTHYRONINE	2.900313	2.976931	2.762654	2.836028
MELATONIN	2.596909	2.495353	2.590820	2.486731
METHYLTHIOADENOSINE	2.842328	2.915045	2.862486	2.941739
N-ACETYLGLUTAMATE	1.513739	1.397336	1.519916	1.400997
N-ACETYLTRYPTOPHAN	2.240985	2.191408	2.238199	2.195266
N-ETHYL-5-METHYL-2-(1-METHYLETHYL)-CYCLOHEXANECARBOXAMIDE	3.427157	3.340453	3.417660	3.340976
OMEGA-HYDROXYDODECANOATE	2.122435	2.058923	2.137376	2.069099
PANTOTHENATE	1.831218	1.724229	1.832306	1.725108
PHENYLACETALDEHYDE	1.859897	1.779394	1.885383	1.799064
RETINOATE	2.792486	2.550801	2.771151	2.580457
RIBOFLAVIN	2.499297	2.494766	2.504895	2.498269
THIOPURINE S-METHYLETHER	2.543931	2.545766	2.547658	2.557212
THYROXINE	2.908752	2.962731	2.752477	2.807130

3.3.3 Response factor calculation

To evaluate whether the conventionally used constant response factor (CRF) approach is applicable to our experiment, all slope values of log10-log10-transformed 6-point dilution series were calculated and evaluated using a one-sided t-test. If detected, in-source fragments (i.e., $[M\text{-}H_2O+H]^+$ and $[M\text{-}2H_2O+H]^+$) were treated as independent compounds. The null hypothesis is that the log-log slope value is equal to 1, a requirement for the underlying assumptions in this approach. Figure 8 is a histogram of the observed log-log slopes. Across the entire data collection, the mean slope value is 0.94 ± 0.48 (s.d.), which was determined to be statistically different from the expected value of 1 (p-value = 0.0039). As our empirical measurements

deviate from underlying assumptions, the CRF approach is not valid here. Groff et al. (2022)[25] hypothesized that a potential reason for deviation from a slope of 1 could be the measurements are falling outside the chemical linear dynamic range, which would not be unexpected.

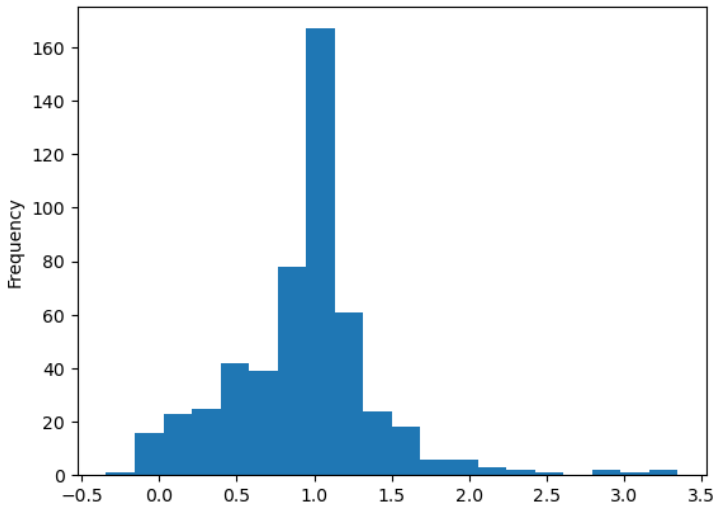

Figure 8. Histogram of the distribution of slope values derived from fitting a linear model to log10-log10 transformed calibration curves.

We generated scripts and workflow files that provide a walkthrough of CRF calculation for collected data, training a linear model, and predicting CRF for new data. While our collected data violates the assumption required for least squares regression to calculate CRF, we use this as example data to demonstrate training the linear model and making new predictions for a blinded or unknown sample.

The BRF approach, assuming response is not constant, which aligns with our empirical data, is used to calculate RF values for all data points. BRF was calculated using data points from both replicates within a treatment condition. While a 95% PI is generated from this approach, the mid-point value of the prediction interval is used to train models. Interval regressions leveraging the full PI are beyond the scope of this work.

3.3.4 Response factor linear model

We move forward with the BRF approach. Shown below in Table 6 are the model coefficients and model importance for the study parameters covering first and second order interactions, from a chemical-specific BRF model. Aliased terms cover unique combinations of the 8 parameters, excluding logIE. We also built a sample-specific BRF model (that is, including only study design variables without logIE), but we found that the coefficients for each study parameter were very similar to those in the chemical-specific BRF, and as such, interpretation in parameter contribution characterization would not change. The sample-specific BRF model showed a larger intercept (i.e., error) term, leading us to believe that logIE can account for a portion of the model error and thus adds value to a response factor model.

Table 6. Model statistics for the chemical-specific bootstrapped response factor linear model, describing the contributions of each variable to response factor. Note that log10-based

ionization efficiency values, which are compound-specific, are included as a contributor to response factor in this model.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	214801217	31210450	6.882	4.39E-11	***
organic	-190996579	43712639	-4.369	1.80E-05	***
ph	10912094	1224804	8.909	< 2e-16	***
flow_rate_nLmin	-793398	82445	-9.623	< 2e-16	***
esi_voltage_V	-16290	5550	-2.935	0.00363	**
sample_loading_nL	94848	5549	17.092	< 2e-16	***
conc_ratio_pct	113290	83271	1.36	0.17485	
num_compounds_n	205337	824538	0.249	0.80353	
sample_coelution_pct	-1358485	83896	-16.192	< 2e-16	***
logIE	2653383	5193823	0.511	0.60987	
Alias 1	-12912893	1729651	-7.466	1.25E-12	***
Alias 2	915705	117232	7.811	1.40E-13	***
Alias 3	-13204	7876	-1.677	0.09482	.
Alias 4	-38226	7880	-4.851	2.12E-06	***
Alias 5	-664094	118196	-5.619	4.95E-08	***
Alias 6	5973444	1172394	5.095	6.71E-07	***
Alias 7	1285789	118143	10.883	< 2e-16	***

N.b. Many of the variables are statistically significant at $p < 0.001$

We evaluate model feature importance by comparing the t-values derived for each parameter from the chemical-specific BRF model (Figure 9). The t-statistic indicates the extent to which the sampled values deviate from the sample mean, with larger deviations demonstrating statistical significance. As such, the magnitude and directionality of the t-statistic can be compared across our study parameters, thereby providing insight into their relative contribution importance to response factor.

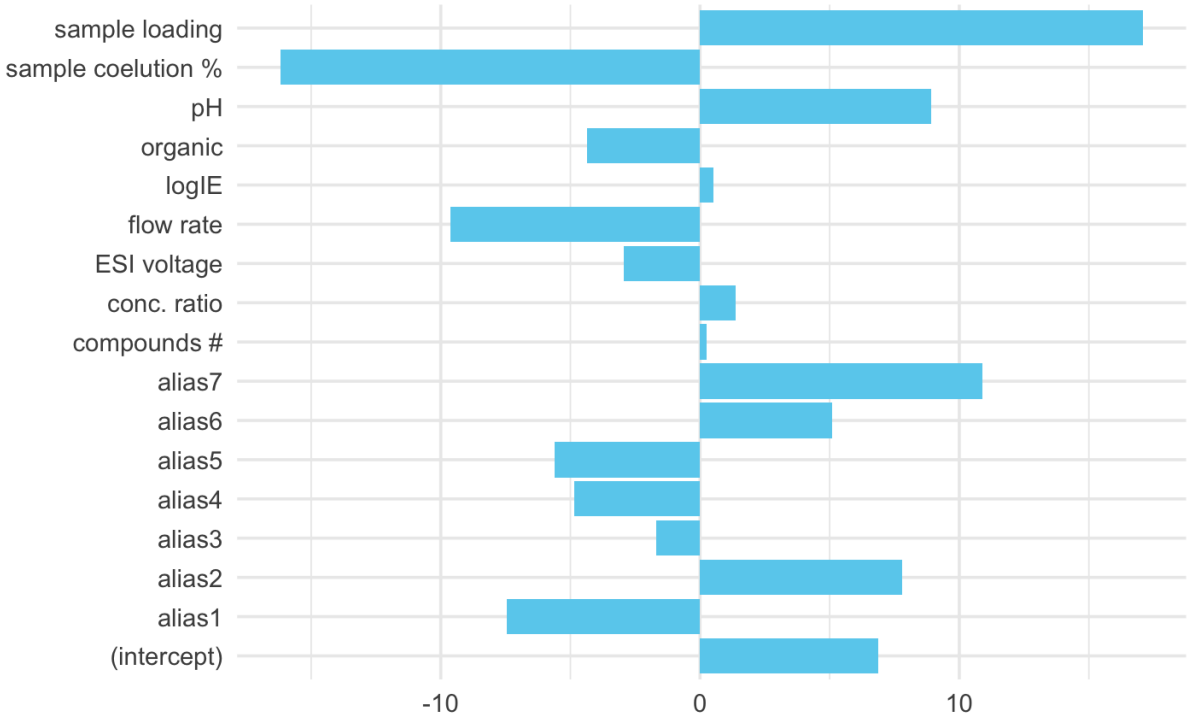

Figure 9. Feature importance as derived from the t-values determined in the chemical-specific bootstrapped response factor linear model.

Some observations on feature importance of study parameters include:

- Sample loading as the most important feature and is a positive contribution to response factor.
- Sample co-elution % also as an important feature in the linear model and exhibits a negative contribution to response factor. This aligns with expectations that greater amount of co-elution will increase potential of ion suppression, thus leading to depressed response.
- LogIE exhibiting one of the smallest contributions to response factor. The positive directionality indicates increase in response with higher ionization efficiency, as expected, though contribution is modest.
- pH demonstrating positive contribution (higher pH with greater response) but organic phase exhibiting negative contribution (depressed response under acetonitrile than methanol mobile phase).
- Flow rate showing negative contribution such that higher flow rates result in depressed response.
- Intercept that represents contributions to response factor that are not accounted for. Though large, other defined study parameters demonstrate greater importance.

3.3.5 Parameter influence on response

To further interrogate parameter relationships with response, we trained a random forest model to predict response factor given the measured responses (i.e., MS1 area) for all compounds in all dilutions under each set of study parameter conditions. Though we did not see additional value provided regarding response factor predictions using the random forest model instead of the linear model, the advantage of the random forest model is the ability to probe the trained

model to provide us with insight into (1) parameter importance, (2) each parameter's relationship to response factor, and (3) effects of each parameter on predicted response factor. Specifically, we investigate the impact on mean signed error (MSE), a quantitative measure of feature importance, due to permutations of study variables and partial dependence to evaluate parameter marginal effect on response factor.

Shown in Figure 10 are the percent increases in mean signed error in response factor predictions when each variable is permuted across their examined range while all other variables are unchanged. For example, the variable sample loading exhibits > 40% increase in mean signed error in response factor predictions when we shuffle sample loading levels, while the other variables (e.g., flow rate, ESI voltage) remain constant. Our expectations are that a larger increase in mean signed error when a parameter's levels are permuted indicate a higher degree of parameter importance to response factor predictions. Following this logic, sample loading appears to be the most important parameter to response factor. This parameter is related to the quantity of the compound, where a higher loading of the same sample results in introduction of a larger quantity of the compound, which typically results in a higher measured response. In this sense, sample loading could be considered another representation of concentration, and these two parameters likely exhibit an interaction effect. Though beyond the scope of this effort, future work could explore the interaction of these two parameters as an additional contributor to response factor.

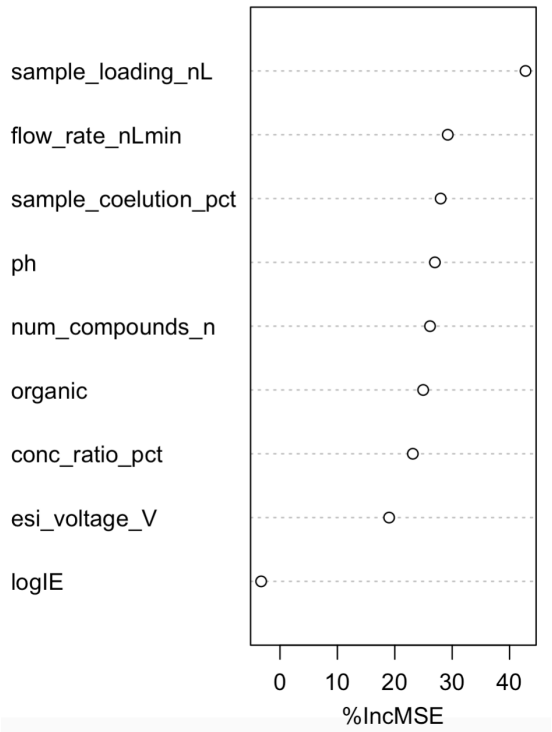

Figure 10. Scatter plot displaying the percent increase in mean signed error of response factor predictions when each variable (y-axis) is permuted across their parameter space while holding constant all other variables. A random forest model was trained to predict response factor

All study parameters exhibit non-zero percent increases in MSE, leading us to conclude that all examined parameters contribute to response factor, and much more so than logIE. Interestingly

and somewhat unexpected, there is minimal change in mean signed error when permuting logIE, which implies that this parameter does not influence response factor predictions. We hypothesize that this unanticipated finding may stem from a more fundamental issue of ionization source configuration differences. While we train on the “universal” logIE, this value actually reflects the expected logIE on an Agilent XCT instrument with an ESI source. However, our LC-MS system (RSLCnano LC coupled to Orbitrap Exploris 480 MS) operates at nanoflow rates, which requires a nanoESI source that is different from the ionization source within the XCT instrument. Taking these observations together, we theorize that the “universal” logIE predictions may not truly be universally applicable across LC-MS configurations.

One potential approach to improve logIE predictions could be to collect data on anchor compounds using our nanoLC system to match those previously collected on an Agilent XCT and derive an instrument correction factor for logIE predictions. However, a translation to our configuration would not provide much value added, as its only impact would be to scaling of features, and this implementation is contrary to the paradigm of our model—a universal response factor prediction. Alternatively, inclusion of chemical descriptors in lieu of the predicted logIE values as contributions to response factor could be explored, as logIE predictions are predicated on chemical descriptor input to the IE model, and IE is an intrinsic property of a chemical. Nevertheless, in its current state, universal logIE values may not be the most appropriate representation of ionization efficiency contributions to response factor, which may have implications in concentration estimation.

We next investigate linearity (or lack thereof) of each parameter’s relationship to response factor. Partial dependence plots are generated to understand the marginal effect that each parameter has on the predicted response factor (also partial dependence). Figure 11 and Figure 12 below display the predicted response factors as an effect of each level of each parameter, for categorical and continuous parameters, respectively. Note that the parameter ranges are normalized to categorically indicate low, medium, and high, in these plots to represent them on the same set of axes rather than displaying each parameter by their true ranges. For example, the low, medium, and high levels for flow rate are: 200 250, and 300 nL/min. These plots can provide insight into parameter relationships, including whether parameters are accurately reflected as linear terms, as we assumed them to be during model development. A sign of divergence from accurate parameter representation could be a “U” or inverted “U” shape; this observation could be indicative of a parameter’s non-linear effect on response.

Figure 11. Partial dependence plot as displayed for categorical factors: organic phase and pH. Response factor (or partial dependence) as an effect of the predictor at different levels.

It is obvious that no conclusions regarding linearity between parameter and response factor can be drawn for the categorical factors given the limitation to two levels (Figure 11), but such divergence behaviour is observed for ESI voltage and concentration ratio % (Figure 12). We hypothesize that these parameters' contributions to response factor may not be linear, though further exploration to elucidate more accurate representations is beyond the scope of this effort.

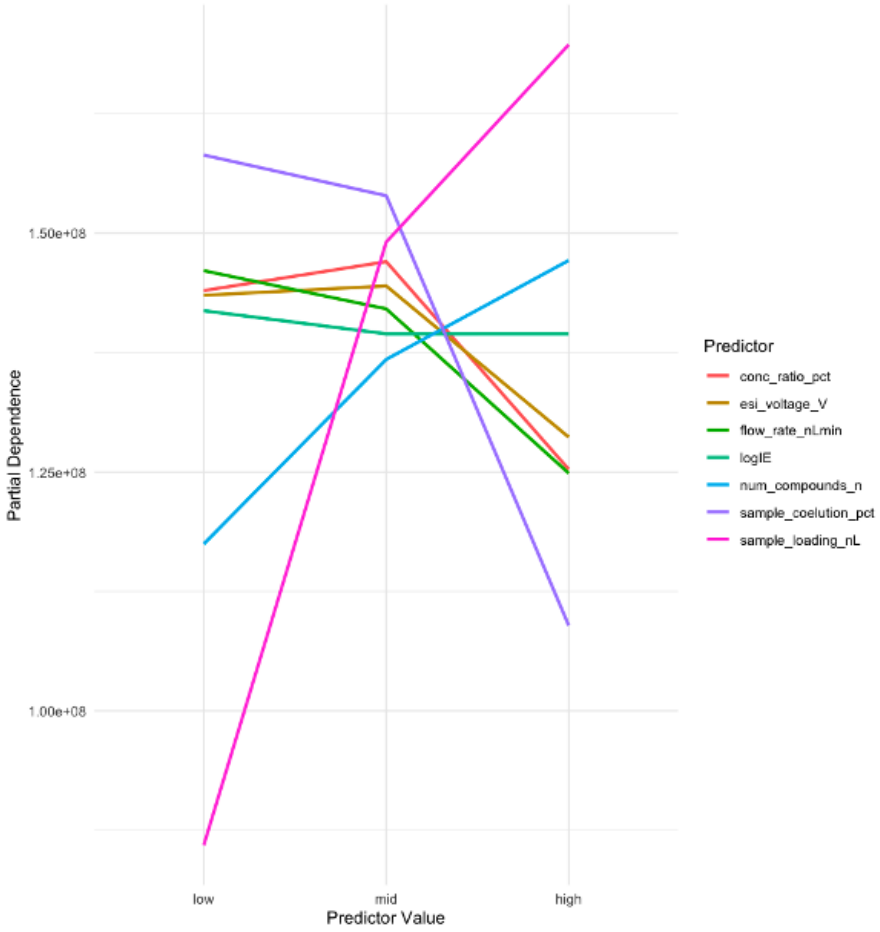

Figure 12. Partial dependence plot as displayed for the 6 continuous factors (concentration ratio %, ESI voltage, flow rate, number of compounds, sample co-elution %, and sample loading) and logIE. Response factor (also partial dependence) as an effect of the predictor at different levels.

The directionality (or net difference between high and low predictor values) of each curve in Figure 11 and Figure 12 should align with that of the coefficients represented in Table 6 and as feature importance in Figure 9. Interestingly, we observe that sample loading exhibits the largest influence on response factor across its parameter space (i.e., from low to high) as the largest net difference in response factor is captured for this parameter out of all examined factors. This finding aligns with the parameter importance results shown in Figure 10, where sample loading was found to have the largest increase in mean signed error. Conversely, the net difference between high and low predictor values for logIE is minimal, which suggests that logIE merely provides a small contribution to response factor. This aligns with the observation in Figure 10, where logIE was shown to have a near 0% increase in MSE.

Also of note is the response factor pattern for the various levels in sample co-elution % (Figure 12). The net negative difference between high and low predictor values indicates that response factor decreases as sample co-elution % increases. This aligns with our expectations, given that compound co-elution typically results in ion suppression, thus depressing response factor(s) for the co-eluting pair(s). However, what is somewhat unexpected is the inconsistent decrease in response factor as we move from low to middle, and then to high predictor values. We posit that the similarity in response factor between the low (0%) and middle (50%) levels of sample co-elution % is owing to the inability to achieve 0% sample co-elution across all experiments. We

observed unexpected compound pair co-elution not infrequently, and in some cases, near levels of ~50%. As such, it is then not surprising to see such similar response factors at these two parameter levels.

We developed a response factor model from the results of our experimental design and characterized the contribution of each study parameter to response factor. We observe that sample loading, flow rate, and extent of sample co-elution are most important to response factor. We see that logIE does contribute to response factor, but its contribution does not appear to be as critically important as other experimental variables. This suggests that accounting for experimental variables will likely improve concentration estimation over current models that rely mostly on logIE. Though logIE accounts for some experimental conditions (e.g., pH of aqueous mobile phase), the difference in magnitude of contribution between logIE and experimental parameters, such as chromatographic flow rate, likely means that the set of experimental descriptors included in prediction of logIE may not be sufficient. Further efforts to more deeply probe sampling and chromatographic contributions to response factor would likely prove beneficial towards elucidating an even more precise response factor, under the assumption that this could result in more accurate concentration predictions, but would require a finer grid acquisition that is beyond the scope of this initial effort.

3.3.6 Performance of blinded sample

We apply the BRF model to elucidate predicted response factors for a blinded sample. For the sample-specific model variant, we obtained predicted response factor by updating the model with experimental conditions and estimated concentrations are produced by dividing empirically measured response for each unknown feature in a sample by the predicted response factor. We examine and compare prediction accuracy of the sample-specific model to that of the chemical specific model. For the chemical-specific model variant, we additionally unblind to obtain compound identity for each measured response, as this model variant requires such information for ionization efficiency predictions that contribute to predicted response factor. Similarly, we

obtain estimated concentrations for each unknown feature by dividing their measured response by predicted response factor.

Here we first examine the raw data to assess reproducibility in measured response across replicates. We observe high reproducibility between replicates (Figure 13).

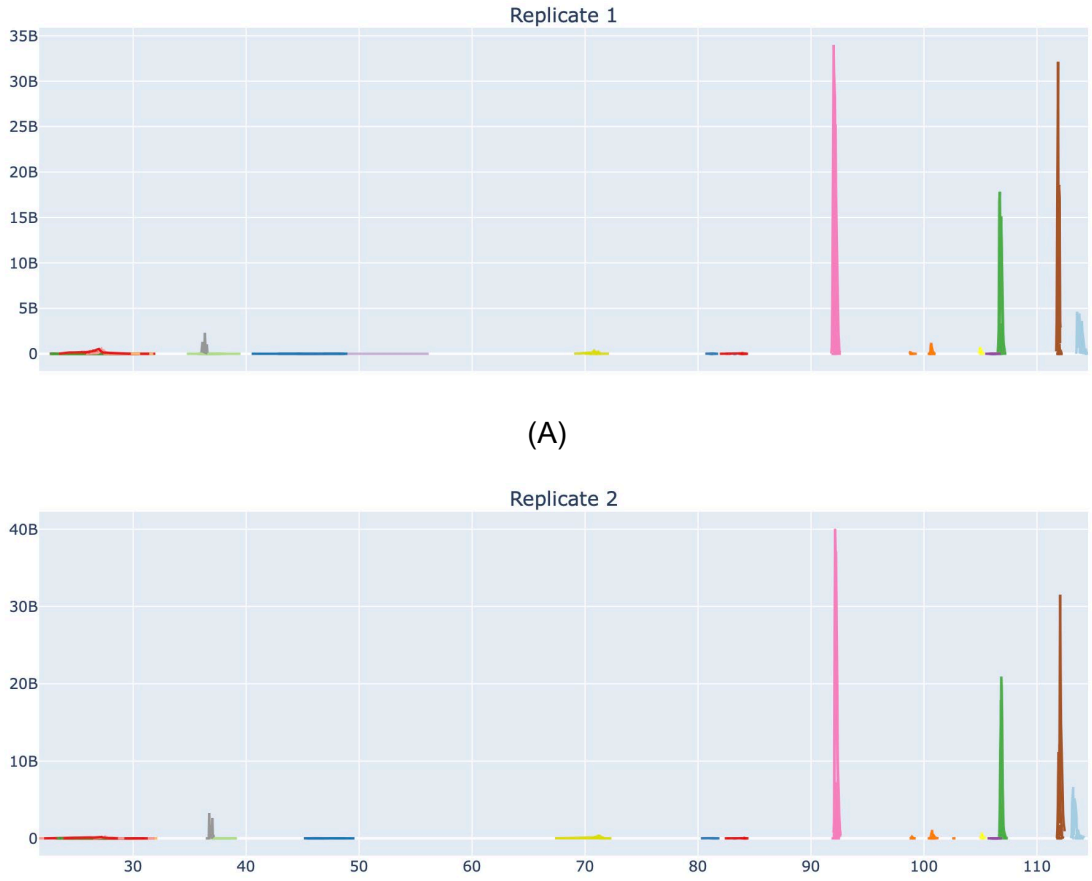

Figure 13. Overlaid extracted ion chromatograms for all detected compounds across the dilution series in (A) replicate 1 and (B) replicate 2 of the blinded sample. A high degree of reproducibility between replicates is observed.

Pseudo-calibration curves for each unknown feature (also Molecule X) in the blinded sample are generated and visualized in Figure 14 below. Note that concentration values are replaced by dilution identifiers and that these identifiers are arranged from least concentrated to most concentrated. Though we arrange the empirical data in this manner for ease of visualization, the relationship between dilution identifiers and relative concentration is not known to the response factor model. We generally observe larger area with more concentrated samples for most molecules, which is as expected. Though we see what appears to be a non-linear response at the lower concentration range, few conclusions can be drawn regarding whether a feature lies within its linear dynamic range, given the blinded relationship between measured response and dilution identifier.

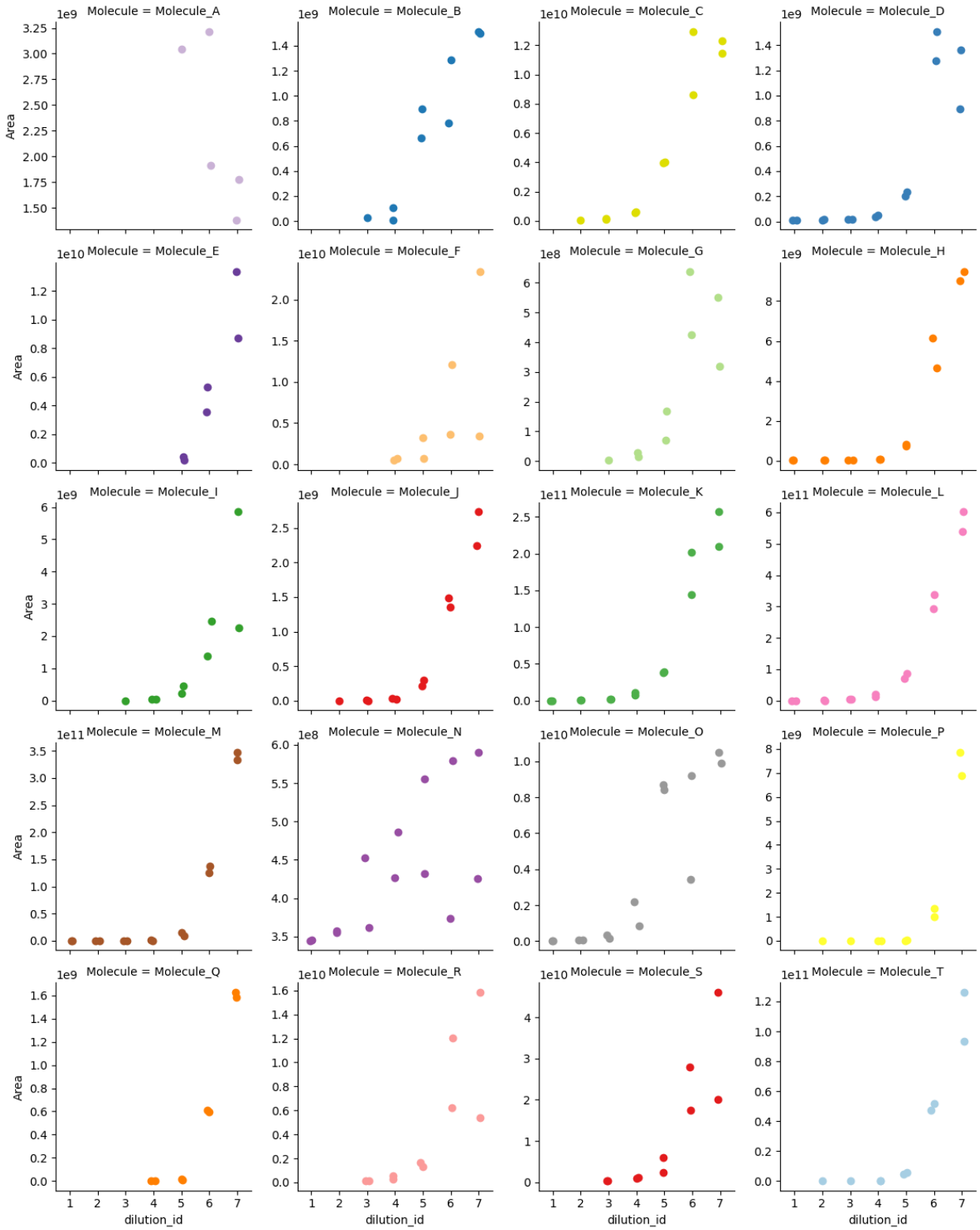

Figure 14. Pseudo-calibration curves generated for each compound, denoted with a unique identifier, in the blind sample across the dilution range, denoted with a dilution identifier.

Measured response of unknown features, observed blinded sample parameters, and experimental conditions were supplied to the sample-specific BRF model. The N peak observations is 20 and the extent of co-elution was found to be 45%. A relative concentration of 1 is supplied, as the true concentration values are not known and no assumptions can be made regarding the relationship between any two dilutions.

We demonstrate results of the blinded sample test by comparing concentration estimation accuracy, as represented by magnitude of error. Other concentration estimation studies have reported results in terms of magnitude of error or $Conc_{pred}$ / $Conc_{actual}$. The validation study by Liigand et al. (2020) achieved a magnitude of error on cereal samples of 5.4X[24]. The sample-specific BRF model, without logIE predictions, achieves a mean magnitude error of 0.5X (need standard deviation), with a maximum error of 18.7X. As 1X magnitude of error represents full accuracy, our slightly lower magnitude of error shows a tendency towards underestimation of concentration values. When we include predicted logIE values as contributors to the response factor model, which relies on elucidating compound identity, we observe slightly better performance (i.e., on average, 0.7X order of magnitude error in concentration estimation). We

demonstrate model performance for each compound in the blinded sample by comparing predicted concentrations to their true concentrations (Figure 15).

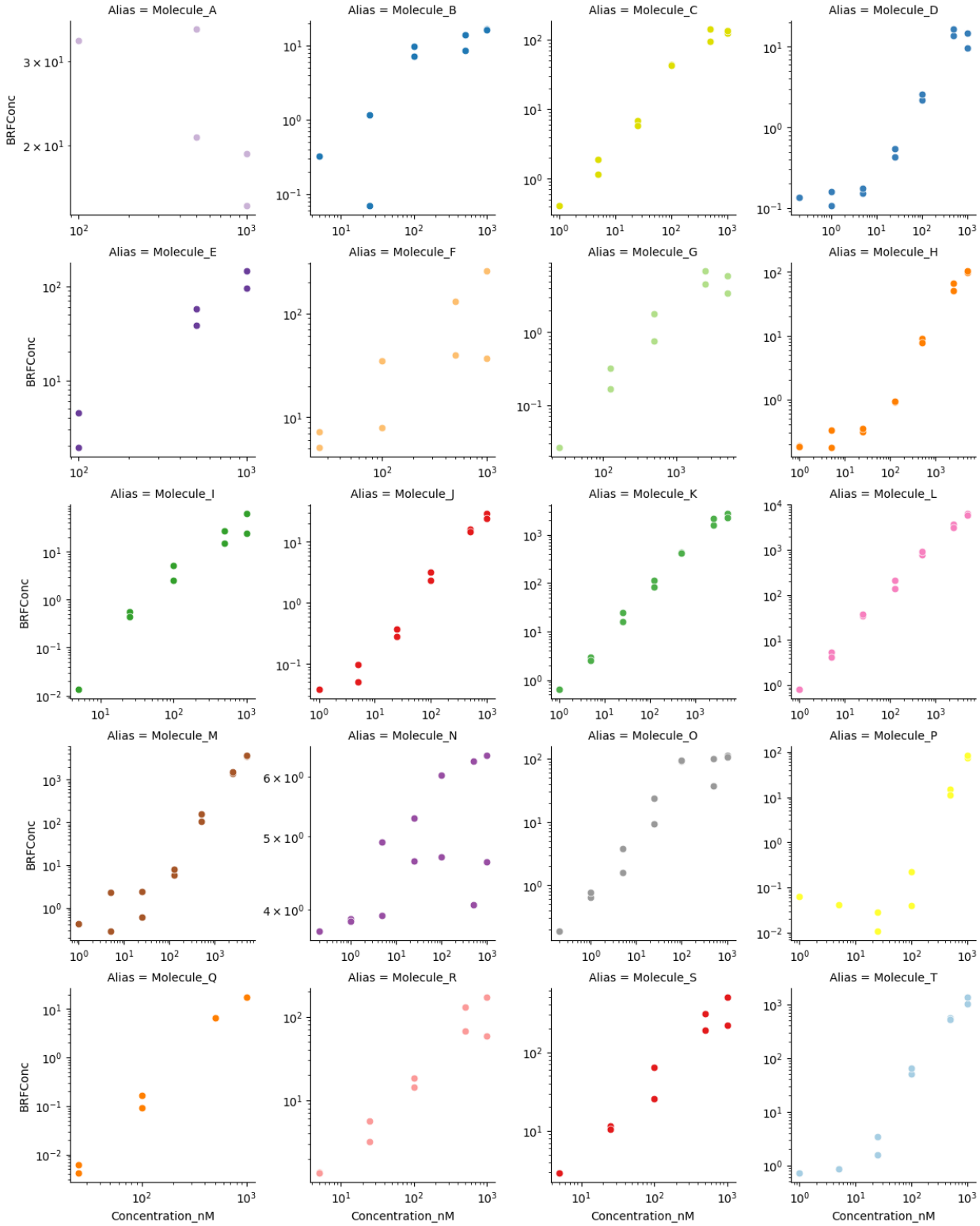

Figure 15. Predicted concentration values (in nM) using the sample-specific bootstrapped response factor linear model for each compound, denoted by unique identifier, in the blind

sample, compared to the true concentration values. Data values are shown on a log scale. Many of the curves show at least a linear relationship between predicted and true concentration values, despite differences between predicted and true values.

The manifest of the blinded sample is shown in Table 7, with 7 dilutions for each compound all shown in the nM range. It should be noted some concentration levels in dilutions 6 and 7 are beyond the concentration ranges used to train the models, with the upper limit of the models being 750 nM.

Table 7. List of compounds included in the blinded sample and their true concentration values in each dilution (in nM). Some concentration values at the extrema of the dilution series are outside the range of the trained model.

Molecule	Alias	Dil.7	Dil.6	Dil.5	Dil.4	Dil.3	Dil.2	Dil.1
PARAXANTHINE	Molecule A	1000	500	100	25	5	1	0.2
L-TRYPTOPHANAMIDE	Molecule B	1000	500	100	25	5	1	0.2
METHYLTHIOADENOSINE	Molecule C	1000	500	100	25	5	1	0.2
DETHIOBIOTIN	Molecule D	1000	500	100	25	5	1	0.2
ADENOSINE 2',3'-CYCLIC PHOSPHATE	Molecule E	1000	500	100	25	5	1	0.2
NICOTINATE	Molecule F	1000	500	100	25	5	1	0.2
3-NITRO-L-TYROSINE	Molecule G	5000	2500	500	125	25	5	1
OMEGA-HYDROXYDODECANOATE	Molecule H	5000	2500	500	125	25	5	1
CYTIDINE 2',3'-CYCLIC PHOSPHATE	Molecule I	1000	500	100	25	5	1	0.2
N-ACETYLTRYPTOPHAN	Molecule J	1000	500	100	25	5	1	0.2
N-ETHYL-5-METHYL-2-(1-METHYLETHYL)-CYCLOHEXANECARBOXAMIDE	Molecule K	5000	2500	500	125	25	5	1
CORTISONE	Molecule L	5000	2500	500	125	25	5	1
DEOXYCORTICOSTERONE ACETATE	Molecule M	5000	2500	500	125	25	5	1
4-HYDROXYBENZALDEHYDE	Molecule N	1000	500	100	25	5	1	0.2
GLUTARYLCARNITINE	Molecule O	1000	500	100	25	5	1	0.2
GLYCOCHENODEOXYCHOLATE	Molecule P	1000	500	100	25	5	1	0.2
GLYCOCHOLATE	Molecule Q	1000	500	100	25	5	1	0.2
N8-ACETYLSPERMIDINE	Molecule R	1000	500	100	25	5	1	0.2
N-ALPHA-ACETYLLYSINE	Molecule S	1000	500	100	25	5	1	0.2
LAUROYLCARNITINE	Molecule T	1000	500	100	25	5	1	0.2

We note that seven of the compounds in the blind sample were used in model development. Shown in Figure 16 are the magnitude of error for each compound in the blind sample. Orange colouring is used to denote compounds that were used in the development of the model. The

model is not better at estimating the concentration of these compounds over the other compounds of the test set, as we observe that these two sets of compounds exhibit the same range of error.

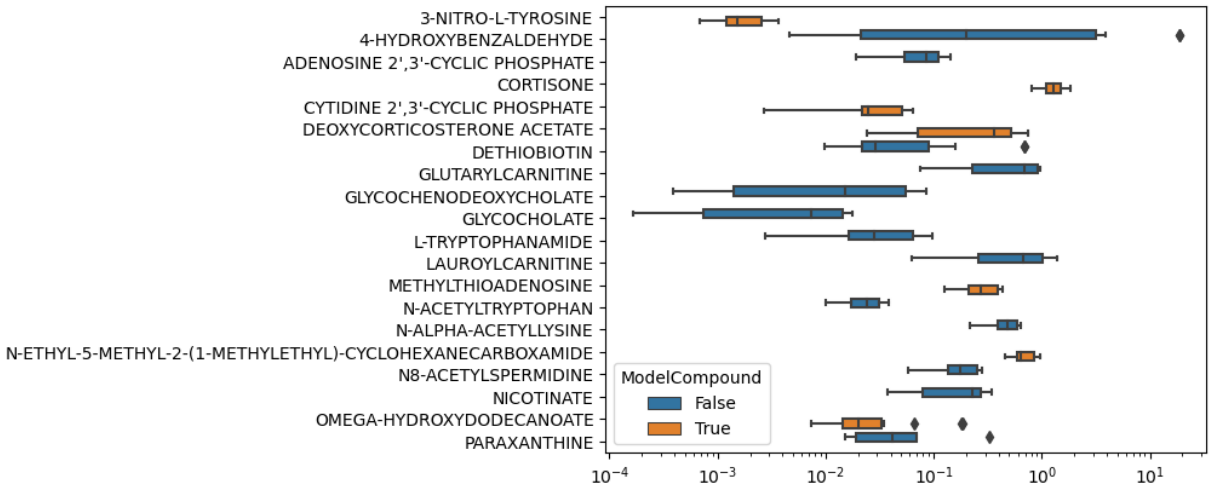

Figure 16. Boxplots displaying the order of magnitude of error from each compound's estimated concentration values in the blinded sample, from the sample-specific BRF model, grouped by compound. A binary label indicates whether the compound was included in the response factor model's training set.

Accuracy of concentration estimation varies among compounds. For example, the compound 4-hydroxybenzaldehyde exhibits a large range of error (Figure 16), indicating poor concentration estimation accuracy, and we observe that its signal begins to saturate at the higher concentration values, and at different points between the two replicates (Figure 14). As such, it is very likely that the irreproducibility and assumption violations of operating within the linear dynamic range can lead to deviations from expected response factors that apply when operating within the linear dynamic range. On the other hand, we observe very low range of error and error close to 0.5X for N-alpha-acetyllysine (Figure 16). Its pseudo-calibration curve depicts a well-behaved compound that exhibits fairly reproducible measured signal under the same experimental conditions and the concentration range considered appears to be within its linear dynamic range (Figure 14). Thus, it is not surprising that concentration estimations for this compound are more accurate than not and show decent precision.

We additionally examine for any differences in error across the different concentration levels. Are there concentration levels in which the model is more or less performant? Figure 17 below depicts the results of this investigation. From examination of magnitude of error at the different true concentration levels, we see a strong tendency towards underestimation, except for the lowest concentration value of 0.2 nM, where error appears to be proximal to 1X or that the model overestimates (mean = 3.2X, median = 0.7X). Note that this concentration value is outside the range of the trained model. The magnitude of error for the actual concentration values greater than 1 nM span the same range, implying that performance is comparable across all these concentration values; no systematic error exists in predicting to any particular concentration level. The wide range of error across concentration levels likely comes from variable performance by compound. Note that performance for the two highest concentration levels are comparable to more diluted levels, which shows promise for model generalizability, as these highest concentration values are also outside the range of the trained model. This suggests that our model could be applicable to concentration ranges outside its training.

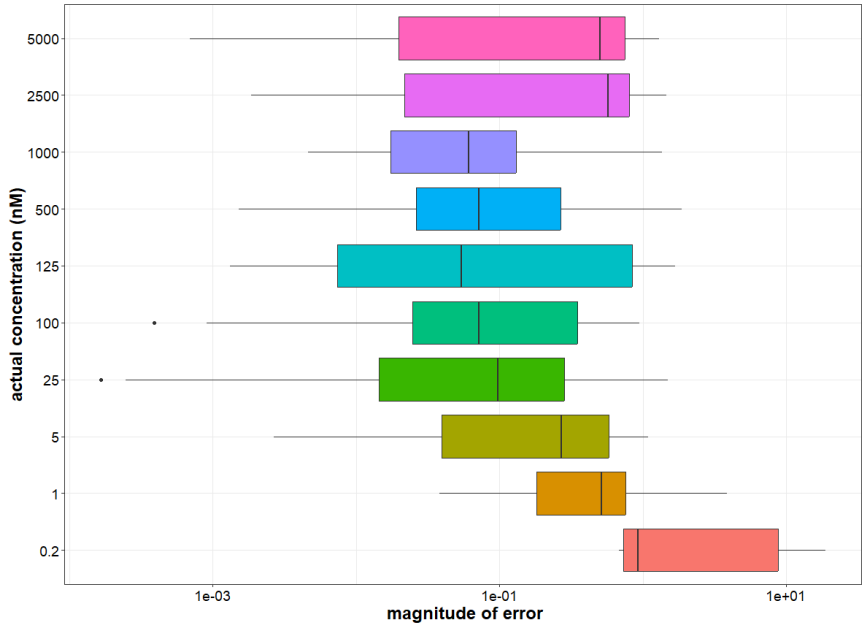

Figure 17. Boxplots displaying the order of magnitude of error from each compound's estimated concentration values in the blinded sample, from the sample-specific BRF model, grouped by true concentration value.

Though we observe better performance (i.e., higher concentration estimation accuracy) by including logIE in the response factor model, the disadvantage of chemical-specific BRF model is the reliance on compound identity elucidation, which may not always be possible in unknown samples. Further, improvement using the chemical-specific BRF model over the sample-specific variant does not appear to be substantial. Our sample-specific BRF approach shows promise as it achieves, on average, under an order of magnitude difference in concentration estimation, and solely by representing the relevant experimental condition; little-to-no compound information is required. As such, we believe that our reference-free quantitation approach that leverages the sample-specific BRF model is much more suited for non-targeted analysis where confident identification may not always be achievable.

3.4 Conclusions

We successfully developed the capability to estimate concentrations using a reference-free quantitation approach. Our approach relies on deriving a more precise response factor model that accounts for contributions from experimental conditions. The novelty in our method is the investigation and characterization of sampling and chromatographic conditions' effects on response factor, which have previously received little attention, and we find that some variables have a larger effect on response factor than others. Notably, we found that our experimental parameters, including chromatographic parameters, demonstrated larger contributions to response factor prediction than ionization efficiency, which would impact concentration estimation. As such, we recommend the inclusion of experimental condition contributions when performing reference-free quantitation where possible.

With our quantitative non-targeted approach, we demonstrate improved reference-free quantitation over existing methods using a blinded sample. Using experimental condition information alone, we observe an average of 0.5X magnitude of error, which represents approximately 10X improvement over current state-of-the-art. While we do see an improved

accuracy when including logIE predictions in the response factor model, accuracy is not substantially higher and its reliance on confident compound identification is not practical for non-targeted analysis.

To further improve upon our framework, future efforts should (1) include additional data collection using a fine grid approach from our statistical design strategy to refine important experimental parameters, (2) investigate automated alternatives to manual peak detection to obtain measured response, (3) more thoroughly examine model performance and generalizability to a larger set of small molecules and instrument configurations, and (4) investigate additional contributors to continue to elucidate a more precise response factor.

While this effort represents an initial proof-of-concept, we demonstrate the promise of such a capability to advance quantitative non-targeted analysis in metabolomics for applicability in a number of disciplines. We envision that further improvements to this framework would enable broad utility in biological, clinical, and chemical applications.

4.0 Concluding Remarks and Outlook

We demonstrate advancement of non-targeted analysis for small molecules through the development and evaluation of two new capabilities: (1) generalizable retention time prediction across chromatographic scales and conditions via novel deep learning model and (2) improved reference-free concentration estimations via sample-specific bootstrapped response factor model using solely experimental condition information from a design of experiments.

Though successful, the work described here represents initial proof-of-concept efforts. Future efforts that build upon these capabilities could focus on algorithm improvement by including more datasets, optimization and validation of model architecture, and more extensive testing and evaluation for applicability to a broader set of application spaces. We believe that continued research and development investments in these areas and this space will push the boundaries of non-targeted analysis of small molecules towards more complete reference-free characterization of unknowns.

5.0 References

(1) Bonini, P.; Kind, T.; Tsugawa, H.; Barupal, D. K.; Fiehn, O. Retip: Retention Time Prediction for Compound Annotation in Untargeted Metabolomics. *Analytical Chemistry* **2020**, *92* (11), 7515-7522. DOI: 10.1021/acs.analchem.9b05765.

(2) Domingo-Almenara, X.; Guijas, C.; Billings, E.; Montenegro-Burke, J. R.; Uritboonthai, W.; Aisporna, A. E.; Chen, E.; Benton, H. P.; Siuzdak, G. The METLIN small molecule dataset for machine learning-based retention time prediction. *Nature Communications* **2019**, *10* (1), 5811. DOI: 10.1038/s41467-019-13680-7.

(3) Chen, B.; Wang, C.; Fu, Z.; Yu, H.; Liu, E.; Gao, X.; Li, J.; Han, L. RT-Ensemble Pred: A tool for retention time prediction of metabolites on different LC-MS systems. *Journal of Chromatography A* **2023**, *1707*, 464304. DOI: https://doi.org/10.1016/j.chroma.2023.464304.

(4) Willighagen, E. L.; Mayfield, J. W.; Alvarsson, J.; Berg, A.; Carlsson, L.; Jeliazkova, N.; Kuhn, S.; Pluskal, T.; Rojas-Chertó, M.; Spjuth, O.; et al. The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *Journal of Cheminformatics* **2017**, *9* (1), 33. DOI: 10.1186/s13321-017-0220-4.

(5) Kingma, D. P. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* **2014**.

(6) Cao, D.-S.; Xu, Q.-S.; Hu, Q.-N.; Liang, Y.-Z. ChemoPy: freely available python package for computational biology and chemoinformatics. *Bioinformatics* **2013**, *29* (8), 1092-1094. DOI: 10.1093/bioinformatics/btt105 (acccessed 9/17/2024).

(7) O'Boyle, N. M.; Morley, C.; Hutchison, G. R. Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chemistry Central Journal* **2008**, *2* (1), 5. DOI: 10.1186/1752-153X-2-5.

(8) Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry* **2011**, *32* (7), 1466-1474. DOI: https://doi.org/10.1002/jcc.21707 (acccessed 2024/09/17).

(9) Kretschmer, F.; Harrieder, E.-M.; Hoffmann, M. A.; Böcker, S.; Witting, M. RepoRT: a comprehensive repository for small molecule retention times. *Nature Methods* **2024**, *21* (2), 153-155. DOI: 10.1038/s41592-023-02143-z.

(10) Li, M.; Zhou, J.; Hu, J.; Fan, W.; Zhang, Y.; Gu, Y.; Karypis, G. DGL-LifeSci: An Open-Source Toolkit for Deep Learning on Graphs in Life Science. *ACS Omega* **2021**, *6* (41), 27233-27238. DOI: 10.1021/acsomega.1c04017.

(11) Xiong, Z.; Wang, D.; Liu, X.; Zhong, F.; Wan, X.; Li, X.; Li, Z.; Luo, X.; Chen, K.; Jiang, H.; et al. Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism. *Journal of Medicinal Chemistry* **2020**, *63* (16), 8749-8760. DOI: 10.1021/acs.jmedchem.9b00959.

(12) Bruderer, T.; Varesio, E.; Hopfgartner, G. The use of LC predicted retention times to extend metabolites identification with SWATH data acquisition. *Journal of Chromatography B* **2017**, *1071*, 3-10. DOI: https://doi.org/10.1016/j.jchromb.2017.07.016.

(13) Chen, C.-J.; Lee, D.-Y.; Yu, J.; Lin, Y.-N.; Lin, T.-M. Recent advances in LC-MS-based metabolomics for clinical biomarker discovery. *Mass Spectrometry Reviews* **2023**, *42* (6), 2349-2378. DOI: https://doi.org/10.1002/mas.21785 (acccessed 2024/09/19).

(14) Sens, A.; Rischke, S.; Hahnefeld, L.; Dorochow, E.; Schäfer, S. M. G.; Thomas, D.; Köhm, M.; Geisslinger, G.; Behrens, F.; Gurke, R. Pre-analytical sample handling standardization for reliable measurement of metabolites and lipids in LC-MS-based clinical research. *Journal of Mass Spectrometry and Advances in the Clinical Lab* **2023**, *28*, 35-46. DOI: https://doi.org/10.1016/j.jmsacl.2023.02.002.

(15) McCord, J. P.; Groff, L. C.; Sobus, J. R. Quantitative non-targeted analysis: Bridging the gap between contaminant discovery and risk characterization. *Environment International* **2022**, *158*, 107011. DOI: https://doi.org/10.1016/j.envint.2021.107011.

(16) Sobus, J. R.; Wambaugh, J. F.; Isaacs, K. K.; Williams, A. J.; McEachran, A. D.; Richard, A. M.; Grulke, C. M.; Ulrich, E. M.; Rager, J. E.; Strynar, M. J.; et al. Integrating tools for non-targeted analysis research and chemical safety evaluations at the US EPA. *Journal of Exposure Science & Environmental Epidemiology* **2018**, *28* (5), 411-426. DOI: 10.1038/s41370-017-0012-y.

(17) Frigerio, G.; Mercadante, R.; Polledri, E.; Missineo, P.; Campo, L.; Fustinoni, S. An LC-MS/MS method to profile urinary mercapturic acids, metabolites of electrophilic intermediates of occupational and environmental toxicants. *Journal of Chromatography B* **2019**, *1117*, 66-76. DOI: https://doi.org/10.1016/j.jchromb.2019.04.015.

(18) Sulyok, M.; Stadler, D.; Steiner, D.; Krska, R. Validation of an LC-MS/MS-based dilute-and-shoot approach for the quantification of > 500 mycotoxins and other secondary metabolites in food crops: challenges and solutions. *Analytical and Bioanalytical Chemistry* **2020**, *412* (11), 2607-2620. DOI: 10.1007/s00216-020-02489-9.

(19) Domínguez, I.; Garrido Frenich, A.; Romero-González, R. Mass spectrometry approaches to ensure food safety. *Analytical Methods* **2020**, *12* (9), 1148-1162, 10.1039/C9AY02681A. DOI: 10.1039/C9AY02681A.

(20) Aldubayyan, A. A.; Castrignanò, E.; Elliott, S.; Abbate, V. A Quantitative LC–MS/MS Method for the Detection of 16 Synthetic Cathinones and 10 Metabolites and Its Application to Suspicious Clinical and Forensic Urine Samples. In *Pharmaceuticals*, 2022; Vol. 15.

(21) Cox, J.; Train, A.; Field, A.; Ott, C.; DelTondo, J.; Kraner, J.; Bailey, K.; Gebhardt, M.; Arroyo-Mora, L. E. Quantitation of Fentanyl and Metabolites from Liver Tissue Using a Validated QuEChERS Extraction and LC–MS-MS Analysis. *Journal of Analytical Toxicology* **2020**, *44* (9), 957-967. DOI: 10.1093/jat/bkaa006 (acccessed 9/19/2024).

(22) Proença, P.; Teixeira, H. M.; Martinho, B.; Monteiro, C.; Franco, J.; Corte-Real, F. LC–MS-MS-MS3 for the determination and quantification of ∆9-tetrahydrocannabinol and metabolites in blood samples. *Journal of Analytical Toxicology* **2023**, *47* (7), 606-614. DOI: 10.1093/jat/bkad046 (acccessed 9/19/2024).

(23) Rischke, S.; Hahnefeld, L.; Burla, B.; Behrens, F.; Gurke, R.; Garrett, T. J. Small molecule biomarker discovery: Proposed workflow for LC-MS-based clinical research projects. *Journal of*

Mass Spectrometry and Advances in the Clinical Lab **2023**, *28*, 47-55. DOI: https://doi.org/10.1016/j.jmsacl.2023.02.003.

(24) Liigand, J.; Wang, T.; Kellogg, J.; Smedsgaard, J.; Cech, N.; Kruve, A. Quantification for non-targeted LC/MS screening without standard substances. *Scientific Reports* **2020**, *10* (1), 5808. DOI: 10.1038/s41598-020-62573-z.

(25) Groff, L. C.; Grossman, J. N.; Kruve, A.; Minucci, J. M.; Lowe, C. N.; McCord, J. P.; Kapraun, D. F.; Phillips, K. A.; Purucker, S. T.; Chao, A.; et al. Uncertainty estimation strategies for quantitative non-targeted analysis. *Analytical and Bioanalytical Chemistry* **2022**, *414* (17), 4919-4933. DOI: 10.1007/s00216-022-04118-z.

(26) Dahal, U. P.; Jones, J. P.; Davis, J. A.; Rock, D. A. Small Molecule Quantification by Liquid Chromatography-Mass Spectrometry for Metabolites of Drugs and Drug Candidates. *Drug Metabolism and Disposition* **2011**, *39* (12), 2355. DOI: 10.1124/dmd.111.040865.

(27) Hatsis, P.; Waters, N. J.; Argikar, U. A. Implications for Metabolite Quantification by Mass Spectrometry in the Absence of Authentic Standards. *Drug Metabolism and Disposition* **2017**, *45* (5), 492. DOI: 10.1124/dmd.117.075259.

(28) Cech, N. B.; Krone, J. R.; Enke, C. G. Predicting Electrospray Response from Chromatographic Retention Time. *Analytical Chemistry* **2001**, *73* (2), 208-213. DOI: 10.1021/ac0006019.

(29) Pieke, E. N.; Granby, K.; Trier, X.; Smedsgaard, J. A framework to estimate concentrations of potentially unknown substances by semi-quantification in liquid chromatography electrospray ionization mass spectrometry. *Analytica Chimica Acta* **2017**, *975*, 30-41. DOI: https://doi.org/10.1016/j.aca.2017.03.054.

(30) Palm, E.; Kruve, A. Machine Learning for Absolute Quantification of Unidentified Compounds in Non-Targeted LC/HRMS. In *Molecules*, 2022; Vol. 27.

(31) Liigand, J.; Kruve, A.; Liigand, P.; Laaniste, A.; Girod, M.; Antoine, R.; Leito, I. Transferability of the Electrospray Ionization Efficiency Scale between Different Instruments. *Journal of the American Society for Mass Spectrometry* **2015**, *26* (11), 1923-1930. DOI: 10.1007/s13361-015-1219-6.

(32) Oss, M.; Kruve, A.; Herodes, K.; Leito, I. Electrospray Ionization Efficiency Scale of Organic Compounds. *Analytical Chemistry* **2010**, *82* (7), 2865-2872. DOI: 10.1021/ac902856t.

(33) Liigand, J.; Laaniste, A.; Kruve, A. pH Effects on Electrospray Ionization Efficiency. *Journal of The American Society for Mass Spectrometry* **2017**, *28* (3), 461-469. DOI: 10.1007/s13361-016-1563-1.

(34) Liigand, J.; Kruve, A.; Leito, I.; Girod, M.; Antoine, R. Effect of Mobile Phase on Electrospray Ionization Efficiency. *Journal of the American Society for Mass Spectrometry* **2014**, *25* (11), 1853-1861. DOI: 10.1007/s13361-014-0969-x.

(35) Lee, R. Statistical Design of Experiments for Screening and Optimization. *Chemie Ingenieur Technik* **2019**, *91* (3), 191-200. DOI: https://doi.org/10.1002/cite.201800100 (acccessed 2024/09/18).

(36) Heckert, N. A.; Filliben, J. J.; Croarkin, C. M.; Hembree, B.; Guthrie, W. F.; Tobias, P.; Prinz, J. Handbook 151: Nist/sematech e-handbook of statistical methods. **2002**.

(37) Arboretti, R.; Ceccato, R.; Pegoraro, L.; Salmaso, L. Design of Experiments and machine learning for product innovation: A systematic literature review. *Quality and Reliability Engineering International* **2022**, *38* (2), 1131-1156. DOI: https://doi.org/10.1002/qre.3025 (acccessed 2024/09/18).

(38) Montgomery, D. C. *Design and analysis of experiments*; John wiley & sons, 2017.

(39) Remane, D.; Meyer, M. R.; Wissenbach, D. K.; Maurer, H. H. Ion suppression and enhancement effects of co-eluting analytes in multi-analyte approaches: systematic investigation using ultra-high-performance liquid chromatography/mass spectrometry with atmospheric-pressure chemical ionization or electrospray ionization. *Rapid Communications in Mass Spectrometry* **2010**, *24* (21), 3103-3108. DOI: https://doi.org/10.1002/rcm.4736 (acccessed 2024/09/17).

(40) Hecht, E. S.; Oberg, A. L.; Muddiman, D. C. Optimizing Mass Spectrometry Analyses: A Tailored Review on the Utility of Design of Experiments. *Journal of the American Society for Mass Spectrometry* **2016**, *27* (5), 767-785. DOI: 10.1007/s13361-016-1344-x.

(41) MacLean, B.; Tomazela, D. M.; Shulman, N.; Chambers, M.; Finney, G. L.; Frewen, B.; Kern, R.; Tabb, D. L.; Liebler, D. C.; MacCoss, M. J. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **2010**, *26* (7), 966-968. DOI: 10.1093/bioinformatics/btq054 (acccessed 9/17/2024).

(42) Wishart, D. S.; Guo, A.; Oler, E.; Wang, F.; Anjum, A.; Peters, H.; Dizon, R.; Sayeeda, Z.; Tian, S.; Lee, Brian L.; et al. HMDB 5.0: the Human Metabolome Database for 2022. *Nucleic Acids Research* **2022**, *50* (D1), D622-D631. DOI: 10.1093/nar/gkab1062 (acccessed 9/17/2024).

(43) Kruve, A. Strategies for Drawing Quantitative Conclusions from Nontargeted Liquid Chromatography–High-Resolution Mass Spectrometry Analysis. *Analytical Chemistry* **2020**, *92* (7), 4691-4699. DOI: 10.1021/acs.analchem.9b03481.

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354

1-888-375-PNNL (7665)

www.pnnl.gov