UC Davis

UC Davis Previously Published Works

Title

Denoising Autoencoder Normalization for Large-Scale Untargeted Metabolomics by Gas Chromatography—Mass Spectrometry

Permalink

https://escholarship.org/uc/item/67d1n8f7

Journal

Metabolites, 13(8)

ISSN

2218-1989

Authors

Zhang, Ying

Fan, Sili

Wohlgemuth, Gert

et al.

Publication Date

2023

DOI

10.3390/metabo13080944

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

metabolites

MDPI

Article

Denoising Autoencoder Normalization for Large-Scale Untargeted Metabolomics by Gas Chromatography–Mass Spectrometry

Ying Zhang, Sili Fan, Gert Wohlgemuth and Oliver Fiehn *

West Coast Metabolomics Center, UC Davis, 451 Health Sciences Drive, Davis, CA 95616, USA; yzhang1088@gmail.com (Y.Z.); slfan@ucdavis.edu (S.F.); wohlgemuth@ucdavis.edu (G.W.)
* Correspondence: ofiehn@ucdavis.edu

Abstract: Large-scale metabolomics assays are widely used in epidemiology for biomarker discovery and risk assessments. However, systematic errors introduced by instrumental signal drifting pose a big challenge in large-scale assays, especially for derivatization-based gas chromatography–mass spectrometry (GC–MS). Here, we compare the results of different normalization methods for a study with more than 4000 human plasma samples involved in a type 2 diabetes cohort study, in addition to 413 pooled quality control (QC) samples, 413 commercial pooled plasma samples, and a set of 25 stable isotope-labeled internal standards used for every sample. Data acquisition was conducted across 1.2 years, including seven column changes. In total, 413 pooled QC (training) and 413 BioIVT samples (validation) were used for normalization comparisons. Surprisingly, neither internal standards nor sum-based normalizations yielded median precision of less than 30% across all 563 metabolite annotations. While the machine-learning-based SERRF algorithm gave 19% median precision based on the pooled quality control samples, external cross-validation with BioIVT plasma pools yielded a median 34% relative standard deviation (RSD). We developed a new method: systematic error reduction by denoising autoencoder (SERDA). SERDA lowered the median standard deviations of the training QC samples down to 16% RSD, yielding an overall error of 19% RSD when applied to the independent BioIVT validation QC samples. This is the largest study on GC–MS metabolomics ever reported, demonstrating that technical errors can be normalized and handled effectively for this assay. SERDA was further validated on two additional large-scale GC–MS-based human plasma metabolomics studies, confirming the superior performance of SERDA over SERRF or sum normalizations.

Keywords: GC–MS; data normalization; statistics; primary metabolism; derivatization

check for updates

Citation: Zhang, Y.; Fan, S.; Wohlgemuth, G.; Fiehn, O. Denoising Autoencoder Normalization for Large-Scale Untargeted Metabolomics by Gas Chromatography–Mass Spectrometry. *Metabolites* **2023**, *13*, 944. https://doi.org/10.3390/metabo13080944

Academic Editor: Troy D. Wood

Received: 13 July 2023
Revised: 31 July 2023
Accepted: 8 August 2023
Published: 13 August 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Metabolome is defined as the complete set of low-molecular-mass compounds (<1500 Da) synthesized or modified by a living cell or organism. Metabolomics is the simultaneous measurement of all small molecular metabolites that participate as substrates, reactants, signaling agents, intermediates, and products of enzyme-mediated reactions [1]. Mass spectrometry-based metabolomics has matured as a high-throughput, high-resolution, and high-dimensional technique that identifies multiple metabolite markers present in significantly different abundances between different conditions in large human cohorts or other biomedical and biological studies, enabling the discovery of diagnostic and predictive metabolite levels for disease [2,3].

Metabolomics can be integrated with transcriptomics and proteomics to find biomarkers of diseases or to elucidate biological mechanisms. For both goals, high-quality data mining is needed that removes unwanted (technical) variance. Such technical variance is impacted by various forms of unwanted variations in conducting laboratory experiments, from batch-to-batch differences, variation between different instruments, inter-person variation, and drifts in instrument sensitivity across a specific sequence of samples [4]. To

extract the biologically relevant information, such technical variance needs to be efficiently removed via data normalization methods after raw-data acquisition. Classic quantification strategies in analytical chemistry employ exogenous chemical surrogates as quality controls and for normalization against matrix effects, using either stable isotope-labeled chemicals (deuterium or ^{13}C labeled) or structural analogs of target molecules. Because metabolomics aims to analyze 'all' metabolites, the use of internal standards certainly faces limitations due to the complexity and differences of metabolomics mixtures. Overall, metabolomics normalization has evolved in the past two decades from scaling normalizations [5,6], use of housekeeping metabolites [4], normalization based on internal or external standards [7–9], and quality control samples (QC)-based normalizations [10–13]; specifically, QC-based normalization methods are favored today [14]. Systematic error removal using random forest (SERRF) normalization has been shown to outperform classic QC-normalizations such as locally estimated scatterplot smoothing (LOESS) in large-scale untargeted lipidomics [13,15]. However, no such analysis has been conducted for GC–MS-based untargeted metabolomics. Interestingly, GC–MS-based metabolomics studies typically are much smaller in size than LC–MS-based studies, usually with fewer than 1500 samples [7,16–19]. Untargeted primary metabolomics on gas chromatography–mass spectrometry (GC–MS) suffers technical errors specifically due the need to increase the volatility of metabolites via chemical derivatizations. Here, we show to which extent these errors can be balanced by proper internal standards to normalize this process. In addition, involatile materials may accumulate in the GC injection liners and the beginning of the chromatography columns. Such deposits may alter the local catalysis environment for the delicate balance of derivatization products [20].

GC–MS is an ideal platform from which to detect volatile compounds. For primary metabolites with higher boiling points, a derivatization step reduces boiling points by exchanging acidic hydrogens against derivatization groups. For chemical derivatizations, a wide array of strategies and reagents can be employed, ranging from alkylations and acylations to silylations and others [21–23]. For example, alkylations use boron trifluoride/butanol or dimethylformamide dimethylacetals [24–26], silylation via *N*,*O*-Bis(trimethylsilyl)-trifluoroacetamide, *N*-methyltrimethylsilyl-trifluoroacetamide (MSTFA) or *N*-Methyl-*N*-*tert*.-butyldimethylsilyltrifluoroacetamide (MTBSTFA), acylation/esterification via propyl- or ethylchlorofomate, acetic anhydride or fluorinated anhydrides, or chiral derivatization reactions. Among silylating agents, trimethylsilylations are most frequently used in metabolomics, with MSTFA being the most widely utilized agent [27] due to its ease in handling and wide range of substrates encompassing hydroxyl-, carboxyl-, amino-, or thiol- functional groups. In contrast, other derivatization agents are hampered by less convenient operation and narrower metabolite ranges. For example, boron trifluoride in alkylation and anhydrides in acylation are corrosive, flammable, and highly toxic.

In addition, selective reagents can be used for other functional groups. We use o-methylhydroxylamine (also called methoxyamine) for 25 years in GC–MS-based metabolomics to protect carbonyl groups [28]. A range of other reagents are used in volatile analyses of flavors and odors [29]. Here, we investigate and compare different quality control strategies for metabolomics of human plasma, including derivatization agents, internal standards, external quality control samples, and computational modeling (Figure 1). In this study, we compared three derivatization agents with deuterated internal standards in three different trials across three months. We then applied two different external quality controls for a type 2 diabetes study (T2D) of >4000 human plasma samples: a QC pool made of extracts of the cohort samples and another QC pool that was commercially available. A new modeling tool, called systematic error removal using denoising autoencoder (SERDA), showed an overwhelmingly better performance than SERRF in a large-scale GC–MS-based metabolome dataset. We further compared SERDA with other traditional normalization methods (e.g., mTIC, fTIC, iTIC, metabolite–ISTD ratio) and investigated the performance by combining different normalization methods.

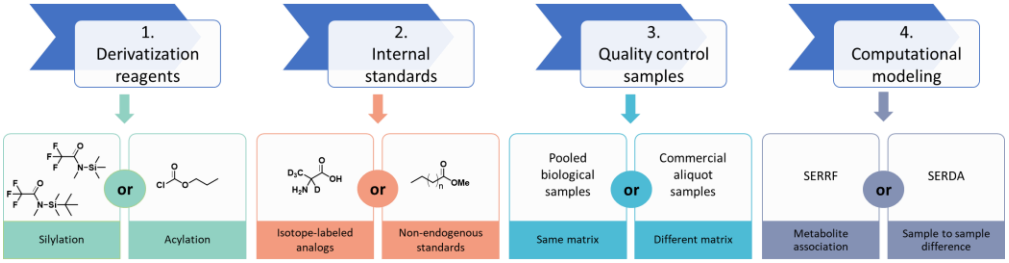

Figure 1. Comparisons of normalizations of large-scale GC–MS human cohort plasma datasets.

2. Materials and Methods

2.1. Reagents

Pooled disodium EDTA plasma was purchased from BioIVT (Westbury, NY, USA), aliquoted into portions of 30 µL, and stored at −80 °C freezer until extraction. The EZ: faast^{TM} amino acid analysis sample testing kit for propyl-chloroformate (PCF) derivatization was purchased from Phenomenex Inc. (Torrance, CA, USA). The 4104 dipotassium EDTA plasma samples were obtained from study participants of a large-scale human cohort for diabetes risk factor analysis. Samples were extracted as published previously [30,31] and aliquoted into analytical samples and backup extracts. The 1032 backup extracts were merged, homogenized, and aliquoted as QC samples. To match the cohort plasma matrix, dipotassium EDTA plasma was purchased from BioIVT (Westbury, NY, USA) and used as validation sample set.

HPLC grade extraction solvents methanol, methyl-tertiary butyl ether (MTBE), and water were obtained from Sigma-Aldrich (Dorset, UK). Twenty-five deuterium-labeled amino acids were purchased from Cambridge Isotope and were used as internal standards in the extraction solutions for the human cohort study. The following concentrations were added to plasma: alanine-d4 * (400 mM), arginine-d7 (110 mM), asparagine-d3 * (100 mM, aspartic acid-d3 * (50 mM), glutamic acid-d5 * (150 mM), glutamine-d5 * (600 mM), glycine-d5 * (400 mM), histidine-d5 (150 mM), homocysteine-d4 (100 mM), isoleucine-d10 * (100 mM), leucine-d10 * (250 mM), lysine-d8 * (200 mM), methionine-d5 * (60 mM, ornithine-d2 (100 mM), phenylalanine-d8 * (100 mM), proline-d7 (200 mM), serine-d3 * (150 mM), threonine-d5 * (200 mM), tryptophan-d8 * (80 mM), tyrosine-d7 * (100 mM), valine-d8 * (400 mM), 2-aminobutyric acid-d6 (40 mM), 2-hydroxybutyric acid-d3 (60 mM), 3-hydroxybutyric acid-d4 (100 mM), and sorbitol-d8 (50 mM). Only 16 of these amino acids (marked by asterisks) were used in the initial derivatization normalization tests of BioIVT human EDTA plasma under different reagents.

2.2. Sample Preparations for GC–MS

For untargeted analyses, plasma samples were extracted using the Matyash liquid–liquid extraction method with cold methanol/MTBE/water [30]. A total of 40 µL of aliquoted plasma was thawed to room temperature and kept on ice during the following steps. Samples were vortexed for 10 s with 225 µL of ice-cold methanol, followed by adding 750 µL ice-cold MTBE. Samples were shaken for 6 min at 4 °C. A total of 188 µL of room temperature water containing the internal standards given above was added. Samples were vortexed for 20 s, followed by centrifugation at 12,210× *g* for 2 min. The lipophilic phase was decanted. The remaining hydrophilic phase was transferred to a new Eppendorf tube, dried down under vacuum and used for derivatization.

For silylations, derivatization started using 10 µL of methoxyamine hydrochloride in pyridine (40 mg/mL, with 5 ug/mL sorbitol-d8), shaken at 30 °C for 90 min. Trimethylsilylation was performed via 90 µL N-methyl-N-(trimethylsilyl) trifluoroacetamide (MSTFA) containing C8-C30 fatty acid methyl esters (FAMEs) at 37 °C for 30 min. For derivatization with tert-butyl-dimethylsilylation, 90 µL of MTBSTFA containing C8-C30 fatty acid methyl esters (FAMEs) was used at 80 °C for 30 min. Samples were centrifuged at 12,210× *g* for 2 min and transferred to crimp top vials for GC–TOF–MS detection.

For targeted derivatization of amino acids in 100 µL plasma sample using propyl-chloroformate (PCF), samples were prepared via a solid-phase extraction method as described previously [32], following the manufacturer's instructions for the EZ: faast^{TM} Amino Acid Analysis sample testing kit.

2.3. Gas Chromatography/Mass Spectrometry Conditions

Each mass spectrometer was coupled to an Agilent 7890 GC system (Santa Clara, CA, USA). For silylated samples, a Restek (Bellefonte, PA, USA) RTX-5Sil MS column was used (30 m length, 0.25 mm i.d, 0.25 µm df, 95% dimethyl/5%diphenyl polysiloxane film) with an additional 10 m guard column. The oven temperature was held at the initial temperature of 50 °C for 1 min, increased from 20 °C/min to 330 °C, and kept isothermal for 5 min. The injection temperature was 275 °C. The injection volume was 0.5 µL in the splitless mode. Silylated samples were measured on a LECO Pegasus IV TOF MS (St. Joseph, MI, USA) at +70 eV, source temperature 250 °C; scan range 85–700 m/z at unit resolution; sampling rate 17 Hz.

For PCF-derivatized amino acids, a Zebron™ ZB-AAA GC column (10 m length, 0.25 mm i.d.) was used. Carrier gas (helium) flow rate was kept constant at 1.5 mL/min (60 kPa). The initial oven temperature was held at 110 °C and then ramped at 30 °C/min from 110° to 320 °C with no final hold. The injection temperature was 250 °C. The injection volume was 2.0 µL at a split ratio of 1:15. PCF-derived amino acids were analyzed by a low-resolution Agilent 5977 single quadrupole MSD (Santa Clara, CA, USA) at +70 eV, source temperature 240 °C; quadrupole temperature 180 °C; scan range 45–450 m/z, sampling rate 4 Hz.

2.4. Data Processing and Data Normalization Scheme

For silylated samples, raw data were deconvoluted via the Leco instrument software ChromaTOF, version 4.5. For silylated samples, deconvoluted data were submitted to the BinBase database for alignment and compound identification [20], including details on signal/noise ratios, missing peak replacements, and data curation. Data files for PCF-derived amino acids were processed using MassHunter Quantitative Analysis B.07.00 version. Data of PCF-derivatized amino acids were normalized to the corresponding internal standards, as described previously [32], and named ISTD normalization. For MTBSTFA-derivatized amino acids, the same ISTD normalization method was used according to internal standards. In addition, three sum-normalization methods were compared: (a) raw amino acid peak intensities were normalized to the sum of all deuterated internal standards, called iTIC; (b) second, data were normalized to the sum of all retention time marker compounds (fatty acid methyl esters), called fTIC; (c) third, data were normalized to the sum of all identified metabolites, as follows: amino acids, hydroxyl acids, and related compounds, called mTIC. The same methods were used to compare the normalization of trimethylsilylated samples in the initial comparison of derivatization methods. In addition, human cohort samples that underwent trimethylsilylation derivatization were normalized via two methods using quality control samples (QC): (1) SERRF (systematic error removal using random forest) [13]; and a new method that we present here: (2) SERDA (systematic error removal using denoising autoencoder; for detailed information, see the Methods section). Statistical analyses were performed via Friedman nonparametric paired tests with adjusted Dunn's significance thresholds of $p < 0.0332$ in GraphPad Prism 8.4.3.

2.5. SERDA Implementation

SERDA is based on denoising autoencoder (dEA), a neural network model, and an extension of autoencoder algorithm. An autoencoder takes an input vector $\mathbf{x} \in \mathbb{R}^d$ and maps it to a hidden representation vector, $\mathbf{y} \in \mathbb{R}^{d'}$, through nonlinear mapping:

$$\mathbf{y} = s(\mathbf{W}\mathbf{x} + \mathbf{b}),$$

where $s()$ is a nonlinear function, in our case, the element-wise exponential linear unit (elu) function; $\mathbf{W}$ is a $d^{'} \times d$ matrix; and $\mathbf{b}$ a $d^{'} \times 1$ bias vector. The hidden representation vector $\mathbf{y}$ is then mapped back to a reconstructed vector, $\mathbf{z} \in \mathbb{R}^d$, by

$$\mathbf{z} = s\left(\mathbf{W}^{'}\mathbf{y} + \mathbf{b}^{'}\right),$$

where $\mathbf{W}^{'}$ is a $d \times d^{'}$ matrix; and $\mathbf{b}^{'}$ a $d \times 1$ bias vector. Each element of training sample $\mathbf{x^{(i)}}$ is mapped to a reconstruction sample $\mathbf{z^{(i)}}$ through the hidden representation $\mathbf{y^{(i)}}$, by minimizing the reconstruction error:

$$\frac{1}{n}\sum_{i=1}^{n} L\left(\mathbf{x^{(i)}}, \mathbf{z^{(i)}}\right),$$

where $L()$ is a loss function, in our case, the absolute error.

The denoising autoencoder, different from the autoencoder, first corrupts the initial input $\mathbf{x}$ to obtain a partially destroyed version, $\widetilde{\mathbf{x}} \in \mathbb{R}^{\boldsymbol{d}}$, by means of a stochastic mapping $\widetilde{\mathbf{x}} \backsim q_{\mathcal{D}}\left(\widetilde{\mathbf{x}} \parallel \mathbf{x}\right)$. In our experiments, we found that the two corruption processes—Gaussian noise and random dropout—performed satisfactorily. For the Gaussian noise, a random vector, $\mathbf{r} \in \mathbb{R}^d$, is drawn from the multivariate normal distribution, $\mathrm{N}\left(0, \sigma^2 \mathrm{I}_d\right)$. For the random dropout, a fixed proportion (ν) of the components are chosen at random, and their values are forced to be 0. These corruptions not only mimic the target sample distribution but also make the algorithm less prone to overfitting. The corrupted input is $\widetilde{\mathbf{x}} = (\mathbf{x} + \mathbf{r})_{-}$, where $()_{-}$ represents the elementwise dropout function. We chose Gaussian noise because metabolomics datasets show a normal distribution after log data transformation. The corrupted input $\left(\widetilde{\mathbf{x}}\right)$ is then mapped to a hidden representation $\mathbf{y} = s(\mathbf{W}\mathbf{x} + \mathbf{b})$, from which we reconstruct $\mathbf{z} = s\left(\mathbf{W}^{'}\mathbf{y} + \mathbf{b}^{'}\right)$. Model parameters $\mathbf{W}$, $\mathbf{b}$, $\mathbf{W}^{'}$, and $\mathbf{b}^{'}$ are trained to minimize the absolute reconstruction error using mini-batch gradient descent algorithm and backpropagation [33,34]. Supplementary Figure S1 shows a schematic representation of the process.

The construction of the SERDA algorithm can be summarized via the following steps:

(1) Take generalized log transformation on training data: $\mathbf{x}_1$ (e.g., QC samples in a compound); and target data:$\mathbf{x}_2$ (e.g., study samples in a compound).
(2) Draw noise from Gaussian distribution, $\mathbf{r} \backsim \mathrm{N}\left(0, \sigma^2 \mathrm{I}_d\right)$. Here, σ is determined by $\sigma_2 - \sigma_1$, where σ_2 is the estimated standard deviation of $\mathbf{x}_2$ and σ_1 is of $\mathbf{x}_1$.
(3) Update training data to obtain corrupted input, $\widetilde{\mathbf{x}}_1$, by adding Gaussian noise corruption, $\mathbf{r}$, to the training data (i.e., $\widetilde{\mathbf{x}}_1 = \mathbf{x}_1 + \mathbf{r}$).
(4) Optionally, oversampling n samples can be applied by adding different random Gaussian noise to each of the training data.
(5) Apply auto-scaling on the training data and target data.
(6) Split the training data, $\widetilde{\mathbf{x}}_1$, into two parts, $\mathbf{x_{1a}}$ and $\mathbf{x_{1b}}$, with proportions of 80% and 20%, respectively.
(7) Initialize $\mathbf{W} \in \mathbb{R}^{d \times d^{'}}$ and $\mathbf{W}^{'} \in \mathbb{R}^{d^{'} \times d}$ with Glorot uniform initializer. Initialize $\mathbf{b} \in \mathbb{R}^{d \times 1}$ and $\mathbf{b}^{'} \in \mathbb{R}^{d^{'} \times 1}$ by zeros.
(8) For each neural network training epoch,
    i. randomly set ν (i.e., the dropout rate) of the elements in $\mathbf{x_{1a}}$ to zero;
    ii. randomly select b samples from $\mathbf{x_{1a}}$ as a mini batch of samples;
    iii. update parameters $\mathbf{W}$, $\mathbf{b}$, $\mathbf{W}^{'}$, and $\mathbf{b}^{'}$ using backpropagation using the Adam algorithm [35] so that the average absolute error of the mini batch samples is reduced as much as possible;
    iv. calculate the average absolute error on $\mathbf{x_{1b}}$ with the updated parameters.

(9) Repeat (7) i–iv until the average absolute error on $\mathbf{x}_{1\mathbf{b}}$ does not decrease for 50 epochs. Mark the number of epochs iteratively processed as n_e.

(10) Apply (7) i–iii on the whole training set, $\widetilde{\mathbf{x}}_1$, n_e epochs. Denote the final trained model as $\Phi(\mathbf{x}) = s\left(\mathbf{W}^{'}s(\mathbf{W}\mathbf{x}+\mathbf{b})+\mathbf{b}^{'}\right)$.

(11) Apply trained model to the target data and obtain the predicted systematic error $\Phi(\mathbf{x}_2)$.

(12) Calculate the normalized values, $\mathbf{x}_2{}^{'}$, by removing the predicted systematic error with subtraction, $\mathbf{x}_2 - \left(\Phi(\mathbf{x}_2) - \overline{\Phi(\mathbf{x}_2)}\right)$, where $\overline{\Phi(\mathbf{x}_2)}$ is the mean average of the predicted systematic error.

(13) Optionally, median normalization can be applied to $\mathbf{x}_2{}^{'}$ to remove leftover inter-batch effect.

(14) Scale and exponentially transform the data $\mathbf{x}_2{}^{'}$ back to the original scale to achieve the final normalized dataset.

Hyperparameters must be provided prior to model building, including the over-sampling number, n; dropout rate, ν; dimension of the hidden representation space, $d^{'}$; mini-batch size, b; and nonlinear function, s. These hyperparameters are determined using 5-fold cross-validation. The SERDA script is publicly available at https://github.com/slfan2013/SERDA (accessed on 11 August 2023).

2.6. Samples and Datasets

For testing and validating SERDA, four types of quality control samples were used. The raw data are given in the Supplementary data file: (a) first, 413 quality control (QC) samples, pooled from a cohort of 4104 human K2EDTA plasma samples (labeled 'pool qc' in the Supplementary data file); (b) second, we added 409 commercial BioIVT K_2EDTA plasma samples as independent secondary quality controls for validation purposes (labeled 'BioIVT validation qc' in the Supplementary data file); (c) third, 104 NIST SRM1950 human plasma QC samples [36] were added as tertiary quality controls (labeled 'NIST validation' in the Supplementary data file); (d) fourth, a total of 102 technical replicate samples were randomly embedded in the human cohort samples in a blinded manner.

3. Results

3.1. GC–MS-Based Metabolomics: Data Normalization for Small Sample Sets

Using BioIVT human EDTA plasma, we first tested the two most common silylation reactions, trimethylsilylation (TMS) and tertiary-butyl dimethylsilylation (TBDMS), and compared these broad-range, untargeted reagents against a commercially available targeted assay for amino acid quantifications via chloroformate reaction. The broad-range silylation agents produced products with unstable ratios for primary amino groups, introducing unwanted variances in untargeted metabolomics. For example, trimethylsilylation usually generates two trimethylsilylated valine products: valine 1TMS with only the carboxyl acidic hydrogen replaced by TMS (Figure 2a, m/z 156); or valine 2TMS with one hydrogen of the amine group and the carboxyl proton replaced by TMS (Figure 2b, m/z 144). In principle, isotope-labeled internal standards should correct for such difficulties in stabilizing reaction conditions and yield exactly the same TMS–derivative ratios for amino acids if ignoring the slight chemical and physical properties between deuterium and hydrogen. We used 16 isotope-labeled metabolite analogs and spiked them into the extraction solution to correct for all technical variations as their corresponding metabolites'—from extraction to derivatization—injection to the gas chromatograph and mass spectrometry. We found that the product ratios of *N*,*O*-TMS-derived amino acids to only O-TMS-derived amino acids varied between pooled QC plasma samples despite all measures of pre-analytical quality controls such as regularly cutting columns, cleaning injectors, or exchanging injector needles [20]. As expected, internal stable isotope standards reduced this technical error. For instance (Figure 2a), the two trimethylsilylated products, valine-d8 1TMS (m/z 164)

and valine-d8 2TMS (*m*/*z* 152), displayed similar ratios between the two QC samples as the endogenous valine 1TMS and valine 2TMS products (Figure 2a,b).

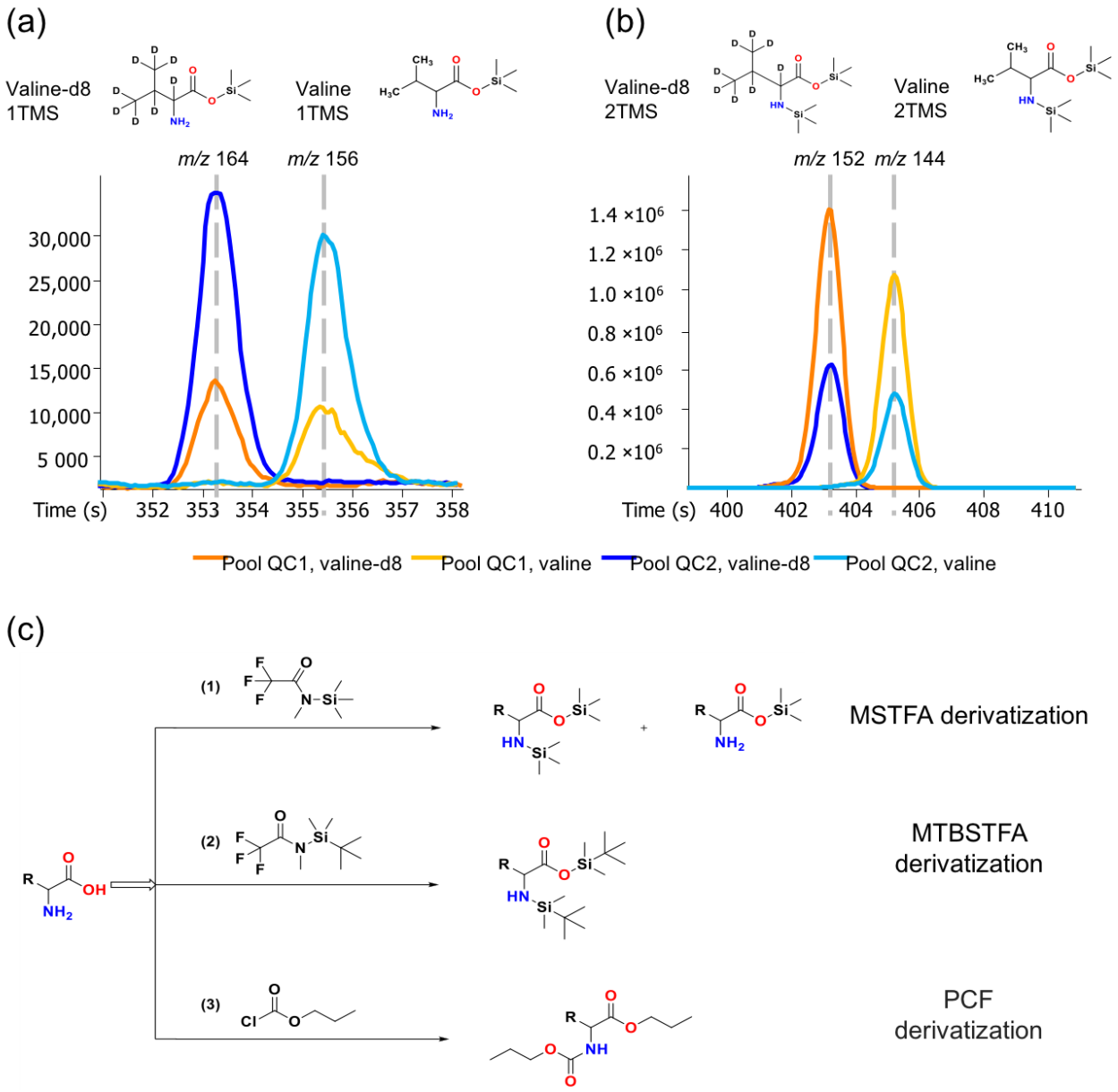

Figure 2. Reaction schemes of MSTFA, MTBSTFA, and PCF derivations, and the chromatography of valine derivatized with MSTFA. (**a**) valine-1TMS and valine-d8-1TMS products; (**b**) valine-2TMS and valine-d8-2TMS products; (**c**) reaction schemes for MSTFA, MTBSTFA, and PCF derivatization.

However, this control for unwanted technical variation did not completely eliminate technical errors when we compared trimethylsilylation to tertiary-butyl dimethylsilylation and targeted chloroformate derivatization (Figure 2c). For this initial comparison, we used 30 commercial plasma samples that were extracted in three independent replicate studies, each conducted one month apart (Figure 3, Supplementary Table S1). We limited the analysis to 16 amino acids that were detectable in all three derivatization methods. For example, arginine was not amenable to any of the methods, while MSTFA did not yield detectable signals for histidine and cysteine in the plasma samples analyzed due to lower sensitivity. Before normalization, the raw data of all three derivatization methods showed significant variance between the three independent analyses (Figure 3). However, average raw data precision worsened from PCF to MTBSTFA to MSTFA, possibly due to

the removal of matrix effects when using PCF derivatization under the Ez:faast protocol, which uses a solid-phase extraction method. Even after normalization to each individual stable-isotope-labeled amino acid, PCF derivatizations gave the lowest precision with 2.7% average relative standard deviation (RSD), followed by MTBSTFA at 8.9% RSD and MSTFA at 9.6% RSD. Although using internal standards for normalization reduced the overall systematic errors for the three tested derivatization agents, residual variance was found across all amino acids, likely due to a random combination of all analytical errors, ranging from pipetting to extraction, moisture during derivatization, and instrument performance.

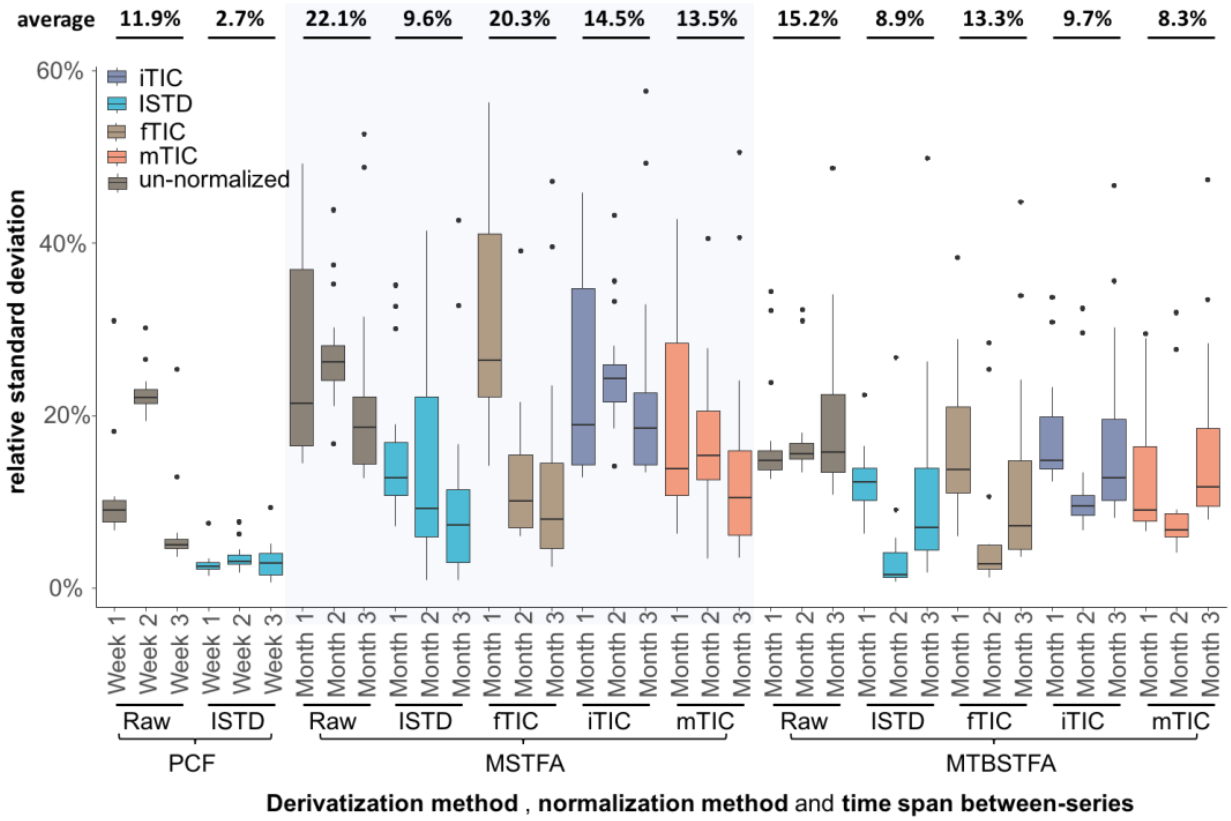

Figure 3. Relative standard deviation (%) of 16 amino acids by three derivatization reagents. Raw: not normalized data; ISTD: normalization of each amino acid to its corresponding internal standard; fTIC: normalization to the sum of 13 fatty acid methyl esters; iTIC: normalization to the sum of all internal isotope-labeled standards; mTIC: normalization to the sum of all identified metabolites.

In metabolomics, such precision values are regarded as acceptable. However, metabolomics aims at analyzing a wide range of compounds, with the coverage of compound classes decreasing from MSTFA to MTBSTFA to PCF derivatization. One cannot include internal standards for all possible small molecule identifications in GC–MS-based metabolomics; therefore, we tested this dataset to see whether other normalization methods might yield acceptable results for MSTFA or MTBSTFA derivatization. To this end, we used three sum-based normalizations: (a) the sum of all internal isotope-labeled standards (total ion chromatogram, iTIC); (b) the sum of all 13 fatty acid methyl esters that are added as retention index markers in our protocol (fTIC); and (c) the sum of all identified metabolites (mTIC) (Supplementary Table S1). Interestingly, the mTIC normalization worked better than iTIC or fTIC for correcting errors for the 16 amino acids for both methods, with 13.5% RSD for MSTFA and 8.3% RSD for MTBSTFA (Figure 3). Both derivatization methods showed little improvement in precision when using fTIC normalization in comparison to the raw data, possibly because fatty acid methyl esters did not undergo any derivatization but only account for random errors during injection. In comparison, iTIC normalizations yielded slightly better precisions for both MSTFA and MTBSTFA than fTIC because the

individual amino acids showed similar error trend as all amino acids as a group. Nevertheless, mTIC should be regarded as the best sum-normalization method for untargeted GC–MS analyses for small datasets such as this, especially for MTBSTFA. Additionally, linearity of instrument responses may not be given for low- and-high abundant peaks, adding complexities to comparisons of RSDs across different derivatization agents.

3.2. GC–MS-Based Metabolomics: Data Normalization for very Large Sample Sets

Most published GC–MS-based metabolomics studies use fewer than 100 samples. Only a single study has been published with almost 1200 samples [16],using relatively matrix-poor tobacco leaf extracts. Apart from instrument drifts, differences in the types and amounts of involatile residues in biological matrices (such as complex lipids or incomplete removal of proteins) may cause additional technical errors in GC–MS-based metabolomics. Here, we used human K_2EDTA plasma samples as an example of a matrix that is highly enriched in fat and protein contents. Such plasma samples are most often used in very large clinical and epidemiological cohort studies, which makes this sample type very relevant with respect to residual technical errors. We used 25 stable-isotope-labeled metabolites during the extraction to investigate if such classic internal standards could be used beyond their corresponding unlabeled endogenous metabolites to correct for drifts during data acquisition and reduce technical (random) errors for the metabolome at large. In addition, we employed four further types of quality control samples to improve analytical precision: (a) From a cohort of 4104 human K_2EDTA plasma samples, we pooled half of the extracts obtained from the first 1032 study samples and aliquoted this pool into 413 cohort-derived quality control (QC) samples. These QC samples were used for data normalization for MSTFA-derivatization-based GC–TOF–MS metabolomics that possess the widest range of metabolite coverage, including amino acids, bioorganic acids, sugars, hydroxyl acids, and fatty acids. Pool QC samples were added after each subset of 10 clinical cohort samples. (b) Secondly, we added one method blank and one commercial BioIVT K_2EDTA plasma sample as independent secondary quality controls for validation purposes. (c) Thirdly, NIST SRM1950 human plasma QC samples [30–34] were added after each set of 40 human cohort samples. (d) Fourthly, we further analyzed a total of 102 technical replicate samples that were used within a single set of 80 samples (Figure 4a).

Data acquisition was conducted across 1.2 years in seven batches with many column cuts and >60 injection liner exchanges in addition to eight column changes and instrument autotunings following the detailed recommendations published earlier [20]. Due to these frequent but necessary interventions, raw pooled QC data showed large technical variations. We first tested the three sum-normalization methods used in the small amino-acid derivatization method sets above (fTIC, mTIC, and iTIC). As expected, none of the classic sum-based normalization methods yielded acceptable precisions for such large-scale studies, with unacceptably high median technical errors between 53–63% for both the cohort pool QC and the commercial plasma QC samples (Table 1).

Table 1. Comparison of relative standard deviations (%RSD) of different normalization methods using pooled T2D plasma extracts (after every 10th sample) as the training QC set and BioIVT as the validation QC set versus normalizations to sum intensities of FAMEs (fTIC), internal standards (iTIC), or all identified metabolites (mTIC).

	SERDA			**SERRF**			**fTIC**		**iTIC**		**mTIC**		**Raw**	
	Pool QC	**Cross-Valid.**	**BioIVT Valid.**	**Pool QC**	**Cross-Valid.**	**BioIVT Valid.**	**Pool QC**	**BioIVT Valid.**	**Pool QC**	**BioIVT Valid.**	**Pool QC**	**BioIVT Valid.**	**Pool QC**	**BioIVT Valid.**
Median	5%	16%	19%	13%	19%	34%	53%	51%	59%	63%	53%	60%	58%	56%
Mean	15%	25%	24%	15%	21%	53%	74%	67%	83%	80%	75%	83%	83%	74%

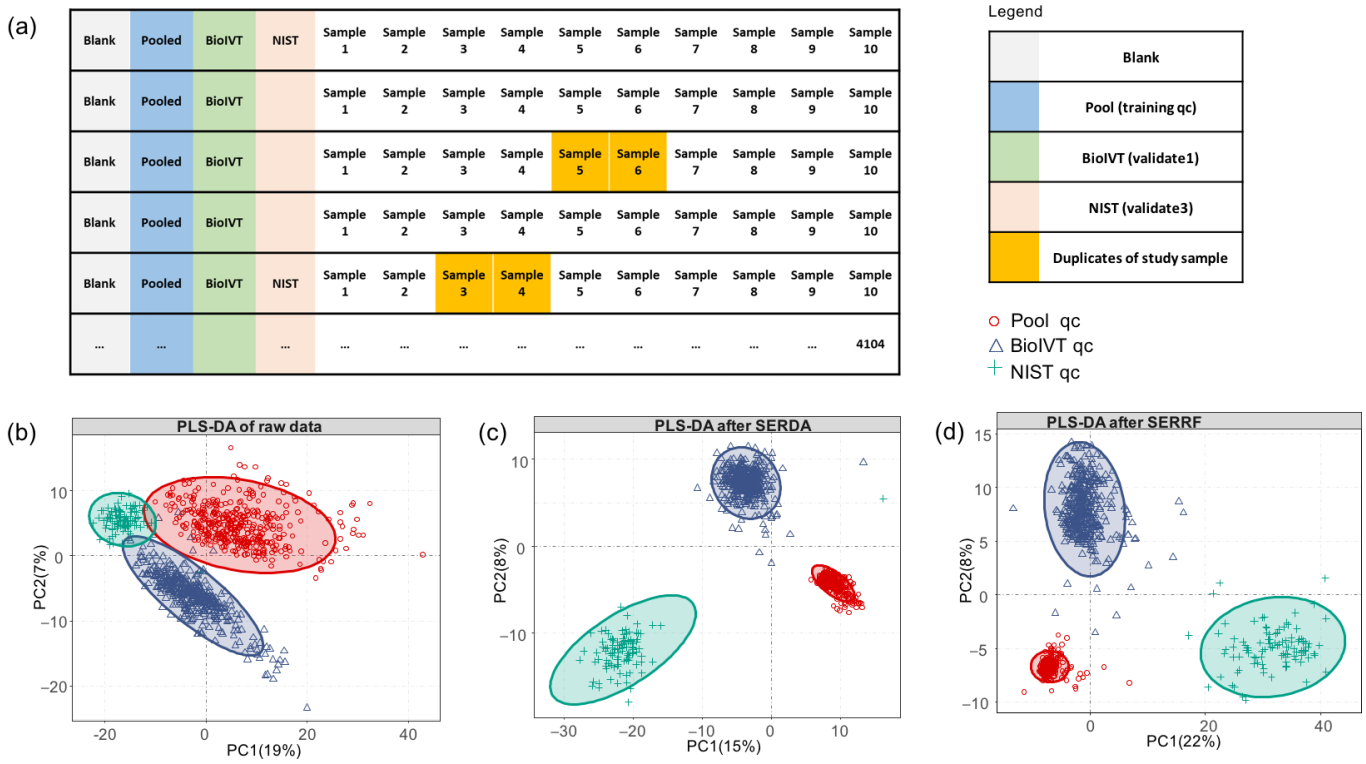

Figure 4. Sequence of sample acquisition and distribution of three sets of quality controls (QC) in large scale GC–MS metabolomics. (**a**) Sequence of injections of blanks, pooled sample quality controls, and BioIVT and NIST external plasma quality controls, plus blinded sample doublets. (**b**–**d**) Partial least square-discriminant analysis plots (PLS-DA) of (**b**) raw data, and effect of normalization by (**c**) SERDA and (**d**) SERRF normalization.

Next, we used a machine-learning-based data normalization method that we previously successfully used for large-scale lipidomics data with more than 5000 samples (systematic error removal by random forest, SERRF) [13]. To avoid overfitting, we applied a five-fold cross-validation on the training QC data to calculate the RSD scores for each compound. SERRF uses correlation patterns of signal drifts of multiple metabolites in QC samples to determine correction factors that are then applied to the biological samples. SERRF avoids overcorrection by using any single metabolite, unlike the classic LOESS algorithm (locally estimated scatterplot smoothing). [37] When applied to GC–TOF–MS cohort pool QC samples, SERRF indeed greatly reduced the median technical error to only 19% RSD (Table 1), clearly below the margin of 30% RSD that had been proposed for metabolomics [38]. Correspondingly, supervised classification of the cohort pool QC, commercial pool QC, and NIST plasma pool QCs showed a large shrinkage of the data dispersion for the training data (cohort pool QC) compared to the raw data (Figure 4b,d). However, when the SERRF model was applied to the primary validation BioIVT commercial plasma QC samples, data still showed considerable dispersion (Figure 4d) and a median 34% RSD (Table 1). In comparison to the success of SERRF in lipidomics, this diminished normalization power in GC–TOF–MS metabolomics may be due to a higher random effect on absolute intensities (trimethylsilylation ratios, Figure 2a,b). In contrast, no chemical derivatization is required in lipidomics; therefore, it only has to be corrected for signal drift patterns due to systematic errors that occur gradually across hundreds of samples in a continuous way.

To better correct for random effects that may be caused by derivatizations or the less controllable splitless injection procedure in gas chromatography, we developed and applied a new normalization method, systematic error removal by denoising autoencoder (SERDA). Denoising autoencoders are used to recognize signals despite large but random noise.

Denoising autoencoders randomly hide some features from input data and automatically generate a new dataset that is similar to the input data [39]. After capturing the useful information interrupted by noise signals via iterative neural network machine learning, the tool generates a reconstructed output with the same shape as the input data in the decoder stage. The denoising autoencoder method (dAE) has the following advantages when applied to data normalization: (1) it fits the structure of metabolomics data where the number of compounds is comparable to or greater than the number of samples; (2) dAE employs a nonlinear model, providing the flexibility of summarizing complex trends of systematic errors in metabolomics data; (3) dAE can tolerate multicollinearity and a high correlation across multiple metabolites, as is frequently observed in metabolomics data [15]; and (4) dAE, coupled with dropout techniques, is robust with respect to outliers and less prone to overfitting [33]. SERDA and SERRF share some similarities, such as nonlinearity and robustness. However, SERRF normalizes each compound sequentially by building a random forest model using correlations to other metabolites in the datasets, while SERDA directly predicts the pattern of systematic error for each biological sample. We can treat SERRF as a compound-by-compound normalization method, while SERDA is sample-by-sample. In LC–MS-based lipidomics, many lipids share strong correlations within each lipid class. Hence, SERRF can successfully reduce the systematic error to 5% RSD [13]. However, in GC–MS metabolomics, correlations of metabolite intensities are not only dependent on classes of chemical structures; they are also based on the stability of derivatization products and injector discrimination based on boiling points. Hence, correlations are expected to be weaker than in LC–MS, causing SERRF to be less powerful, reducing systematic errors significantly. SERDA, on the other hand, does not rely on compound correlation. It captures the data pattern of all metabolites reported by the quality control samples and predicts what the systematic error would be if a QC sample was injected in the sample position. Consequently, we found that SERDA achieved a better performance in untargeted GC–MS metabolomics datasets that show high random variance in addition to systematic drifts. For example, for training cohort QC samples in this study, a residual median technical error of 16% RSD was achieved (Table 1) along with low dispersion in multivariate clustering of the three types of QC samples (Figure 4c). Even more importantly, when the SERDA model was applied to the independent commercial BioIVT QC samples, the data dispersion in the multivariate cluster was smaller than with the SERRF data (Figure 4c,d), and the overall median errors after SERDA modeling were found at 17% RSD, well within the acceptable limits in untargeted metabolomics. Although SERDA requires cross-validation for hyperparameter tuning, we observed that the performance of SERDA was robust against hyperparameter selection. All combinations achieved a consistently lower RSD than SERRF. In addition, when we tested the effect of sum-normalization on of SERDA modeled data, neither mTIC, fTIC, nor iTIC normalizations further improved the residual technical errors of the SERDA data (Table 1).

Next, we analyzed the effect of SERDA on 102 biological duplicate samples that were interspersed into the data acquisition of the 4104 cohort study samples (Figure 5, Supplementary Table S2). A total of 48 biological replicates were measured in adjacent positions within GC autosamplers (Figure 4a), while 54 additional biological replicates were measured apart within a set of 80 samples. Across all metabolites, correlation coefficients for both adjacent and nonadjacent biological replicates were found at excellent r_{xy} = 0.98, with ranges of r_{xy} = 0.81 to 1.0 for nonadjacent replicates that were only slightly worse than adjacent replicates with ranges of r_{xy} = 0.90 to 1.0 (Figure 5, Supplementary Table S2). These data showed that the SERDA algorithm correctly normalized metabolite intensities in biological replicates for both high- and low-abundant compounds.

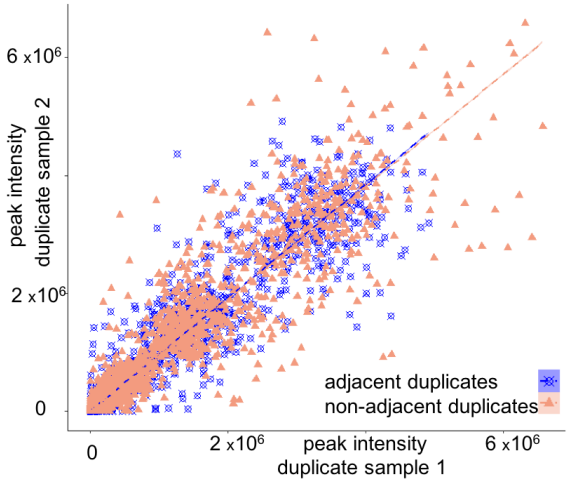

Figure 5. Correlation of blinded T2D cohort sample duplicates after SERDA normalization.

We then investigated how the number of training cohort QC samples affected the overall efficiency of the SERDA algorithm. To this end, we used fewer cohort QC samples for training and then applied the SERDA models on the remaining cohort QC pool samples and the commercial BioIVT QC pools. Median technical errors for commercial BioIVT pools worsened from 17% RSD when using training QC samples after each set of 10 biological samples to 20%, 22%, and 25% RSD when using training QC samples after each set of 20, 40, or 80 biological samples (Figure 6). Technical errors for the remaining cohort QC pool followed the same trend when using SERDA models for each set of 20, 40, or 80 biological samples (Figure 6). Interestingly, technical errors for cohort QC pool samples remained 1–5% RSD lower when used as testing samples than errors obtained via the BioIVT commercial pool samples, and the longer the pooled QC interval, the larger the difference. This observation can be explained by the slight differences in biological matrix compositions between the plasma QC pool of a specific age and ethnicity composition of a human cohort study, as well as the composition of commercial BioIVT samples. Based on these data, we recommend using one cohort pool QC sample for every 10 biological samples as it yielded significantly better precision than using fewer cohort QC pool samples. We also advocate for using secondary-test QC samples—as shown here by commercial BioIVT QC plasma samples—to independently test for the overall effect of the SERDA models.

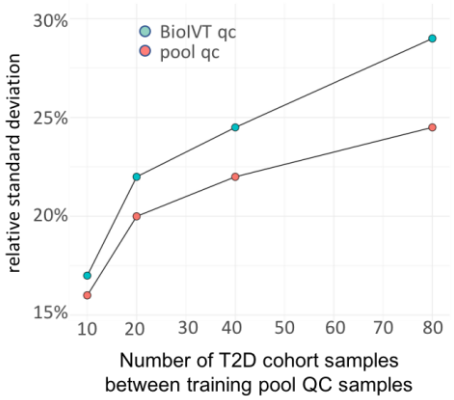

Figure 6. The relative standard deviation (RSD) of pool QC samples (red) and BioIVT qc samples (blue) for each set of 10, 20, 40, and 80 biological samples. With a smaller number of training QC samples, the performance of SERDA decreases, as indicated by the increasing of both the RSD of pool QC traing qc samples and BioIVT validation qc samples.

Furthermore, we investigated how SERDA normalization compared to the use of 25 internal standards (ISTD) that were spiked into the plasma extraction solution. For 23 of

these compounds, the corresponding unlabeled endogenous plasma metabolites levels were detected by GC–TOF–MS metabolomics; two additional internal standards (homocysteine-d4 and sorbitol-d8) were structurally similar to endogenous metabolites (cysteine and blood sugar alcohols). When using the 23 ISTDs to normalize, one-to-one, to their endogenous plasma counterparts, a median 19% RSD was found, with proline and histidine as outliers (Figure 7a). For the same 23 endogenous compounds, SERDA normalization achieved a much lower technical error of 3.8% RSD, and even SERRF outperformed the use the internal standards, with 5% RSD (Figure 7a). The same trends were observed for both cohort pool QC samples and commercial BioIVT plasma QC samples (Figure 7a and Supplementary Figure S2). Use of sum-normalization methods (mTIC, fTIC, iTIC) worsened the performance of the SERRF or SERDA methods alone and did not outperform the one-on-one use of internal standards either (Figure 7a). When we used sorbitol-d8 as the sole internal standard surrogate for a class of 20 detected sugars, sugar alcohols, or disaccharide, this 'one-to-class' normalization failed to improve the technical errors for these 20 sugars, whereas SERDA still outperformed SERRF with respect to carbohydrates for both cohort pool QCs and commercial BioIVT QCs (Figure 7b and Supplementary Figure S2).

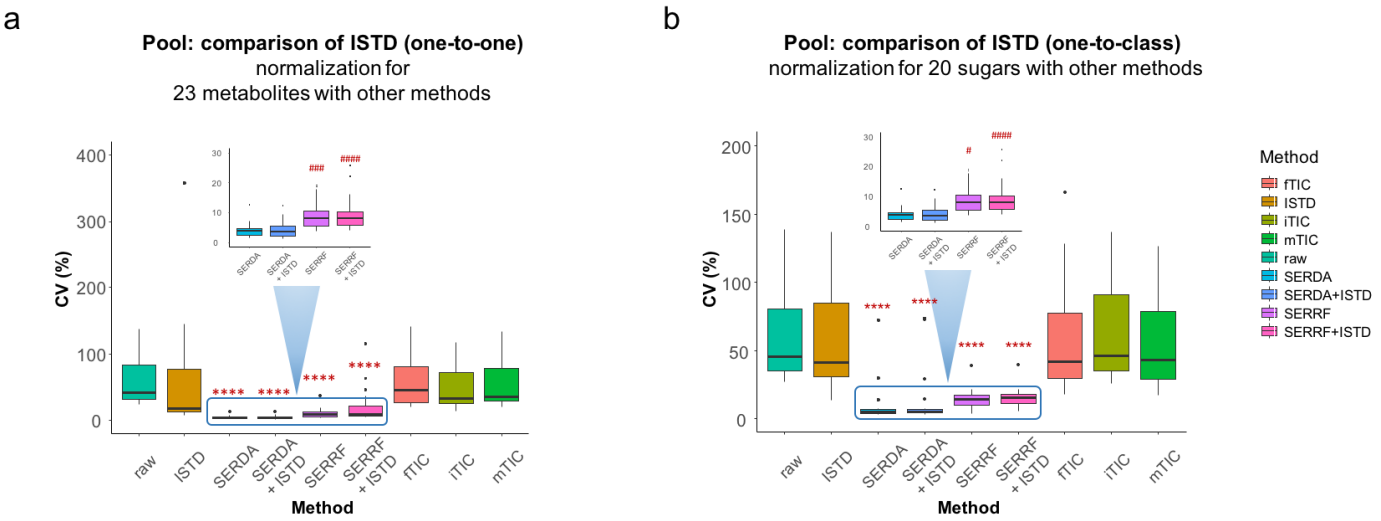

Figure 7. Comparison of ISTD absolute ratio normalization with QC-based and TIC-based normalization methods. The Friedman nonparametric test was used for significance comparison with raw. *p*-value threshold: 0.1234 (ns), 0.0001 (****). The Friedman nonparametric test was used for significance testing compared to SERDA. *P* value threshold: 0.1234 (ns), 0.0332 (#), 0.0002 (###), 0.0001 (####). (**a**) One-to-one: the absolute ratio was calculated by dividing the peak intensity of endogenous metabolite by the corresponding deuterated ISTD; (**b**) One-to-class: the absolute ratio was calculated by dividing the peak intensity of endogenous metabolite by a single deuterated compound as an analog ISTD for the entire class.

Last, we compared the performance of the SERDA method against batchwise-LOESS and SERRF on three GC–MS datasets with respect to the outcome of error residuals using the median RSDs across all compounds [40]. A detailed description of the metabolomics studies can be found in the Supplementary Table S3.

To avoid overfitting, we applied a five-fold cross-validation on the training QCs data to calculate the RSD scores for each compound. For the current T2D study QCs, we added external validation data by employing commercial BioIVT human plasma samples. Such samples were not available for the GeneBank study because at that time, secondary pool QCs were not deemed necessary, and certainly not for the MPA study because the MPA study compared different animals. A successful normalization method should yield small RSDs for both cross-validated training QCs and external-validation QCs. In addition, pool quality controls that were left out of model training provided an excellent opportunity

to evaluate the performance of all models without the risk of overfitting, in addition to the external validation BioIVT plasma QCs. Table 2a,b shows that SERDA normalized datasets achieved significantly lower cross- and external-validation RSD compared with SERRF normalization (Wilcoxon signed-rank test, p-value = 0.02). Residual errors were reduced by SERDA by an additional 3–6% RSD in cross-validated training data compared to SERRF normalization for the three studies, and by more than the additional 17% RSD for the external BioIVT plasma QCs. Such error reduction meant that for the three datasets, the coverage of compounds largely increased that achieved <30% RSD, a typical threshold used for metabolomic datasets [38] (Figure 8).

Table 2. Median % relative standard deviation (RSD) of raw data compared to cross-validated % RSD of SERRF and SERDA normalized data. Median % RSD of raw data compared to externally validated %RSD of SERRF and SERDA normalized data using BioIVT plasma quality control samples.

GC–MS Study	Raw Data	SERRF	SERDA
GeneBank	55%	25%	21%
T2D	58%	19%	16%
MPA	50%	28%	22%
GC–MS study	raw data	SERRF	SERDA
T2D	56%	34%	17%

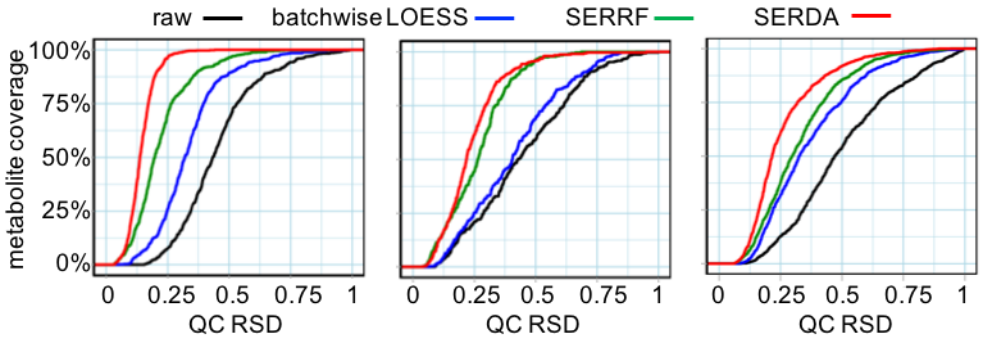

Figure 8. Effect of different normalization methods on residual errors in GC–MS-based metabolomics datasets. Left panel: this study, N = 413 quality control human plasma samples (QC), reporting 661 metabolites. Mid panel: 104 human plasma QC samples of the GeneBank study on 319 metabolites. Right panel: 30 QC plasma samples of the MPA study on 991 metabolites. For each panel, cumulative distributions of cross-validated relative standard deviations (RSD) are given using raw (black), batchwise-LOESS (blue), SERRF (green), and SERDA (red) normalized dataset. The coverage of metabolites achieving specific RSD levels is given as the y-axis.

SERDA yielded the highest coverage of compounds with acceptable residual errors for the three GC–MS datasets, ranging from 72–98% of all metabolites (Table 3). In comparison, SERRF only achieved a coverage of 50–81% of all metabolites reaching that precision threshold (Table 3). Details of the precision coverage for each method is given in Figure 7, which shows the cumulative distribution of the cross-validated RSDs of the quality controls for the raw data and the batchwise-LOESS, SERRF, and SERDA normalized data. SERDA consistently reached a higher portion of smaller RSDs across all the compounds compared with the other two methods. The T2D study was much larger in size compared to the GeneBank and MPA studies, leading to a much larger number of quality control samples employed in the normalization. Hence, difference in improvement of residual errors by the SERDA method may be explained by the number of QCs in the datasets. This finding is consistent with the results of our investigation into using fewer QC samples in the T2D study (Table 2), where the performance of SERDA was found to be increasingly improved with a larger number of QC samples. We conclude that SERDA works well for large-scale GC–MS normalization, while it can still achieve considerable improvement in small-sample settings.

Table 3. Percentage of compounds with cross-validated relative standard deviation < 30% (GC–MS) for raw data, SERRF, and SERDA normalized data. Percentage of compounds with externally validated relative standard. deviation < 30% (GC–MS) for raw data, SERRF, and SERDA normalized data.

Dataset	Raw Data	SERRF	SERDA
GeneBank	27%	61%	76%
T2D	15%	81%	98%
MPA	17%	50%	72%
dataset	raw data	SERRF	SERDA
T2D	12%	67%	91%

4. Conclusions

In summary, we present the results of different normalization methods for GC–MS-based metabolomics studies in human plasma. While MSTFA derivatization showed higher technical errors for amino acids than alternative MTBSTFA or PCF derivatizations, due to the much broader coverage of plasma metabolite annotations for trimethylsilylated compounds, MSTFA is still the reagent of choice for GC–MS profiling. Yet, due to both random and systematic errors in large-scale human plasma cohort studies, the technical errors in GC–MS-based screening of primary (polar) metabolites are harder to control than in LC–MS-based lipidomics [13]. Neither classic internal standards nor sum-based normalizations were sufficient to correct for GC–MS drifts, and even the machine-learning-based SERRF method did not yield satisfying results. However, using denoising autoencoder algorithms implemented in the SERDA random neural networks presented here reduced the median residual errors to less than 20% RSD, making the data usable for epidemiological statistics. This human plasma dataset was acquired over 1.2 years. Date/time stamps in the supplemental dataset showed that effects of breaks in the data acquisition schedule, e.g., for maintenance, were overcome by the SERDA algorithm. To match blood matrix effects in the best way possible, we strongly recommend using quality control samples pooled from the exact same samples as the human cohort study. Commercial plasma QC or NIST SRM1950 QC samples should only be used as independent test samples to estimate the overall residual technical errors. The use of internal standards should be limited to estimating the absolute concentrations of specific plasma metabolites; it should not be used for broad-scale normalization schemes.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/metabo13080944/s1, Supplementary Data File: raw data output from GC-BinBase data processing of Type 2 diabetes cohort samples and pooled QC, BioIVT qc and NIST QC samples. Table S1: Average of the median relative standard deviation of 16 amino acids by three derivatization reagents across three months; Table S2: Correlation coefficients rxy of biological duplicates after normalization by SERDA; Figure S1: A schematic representation of a denoising autoencoder algorithm; Figure S2: Comparison of absolute ratio normalization using internal standards (ISTD) with QC-based and sum-based normalization methods; Table S3: Overview of metabolomics studies used for development and validation of the SERDA algorithm.

Author Contributions: Y.Z. generated data, performed data analyses, and wrote the first draft of the manuscript. S.F. wrote and validated the SERDA algorithm. G.W. performed data processing, analyzed data, wrote and validated code. O.F. conceptualized the study, supervised staff, and performed data analyses. All authors have read and agreed to the published version of the manuscript.

Funding: YZ salary was funded by NIH R01 DK107532 (PI Zhao, J., University of Florida). SF salary was funded by NIH U19 AG023122 (PI Cummings, S-Cal Pacific Medical Center Research Inst.). GW salary was funded by NIH U2C ES030158 (to OF).

Institutional Review Board Statement: Not applicable as no subject information (e.g., T2D status) was used or disclosed to the authors of this technical study.

Informed Consent Statement: Not applicable.

Data Availability Statement: Processed, un-normalized data of the pool quality controls for the large T2D human cohort study are available for download as a supplementary data file. Raw unprocessed data files can be requested from OF.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Sakaguchi, C.A.; Nieman, D.C.; Signini, E.F.; Abreu, R.M.; Catai, A.M. Metabolomics-Based Studies Assessing Exercise-Induced Alterations of the Human Metabolome: A Systematic Review. *Metabolites* **2019**, *9*, 164. [CrossRef]
2. Wishart, D.S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A.C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S.; et al. HMDB: The Human Metabolome Database. *Nucleic Acids Res.* **2007**, *35* (Suppl. S1), D521–D526. [CrossRef] [PubMed]
3. Zeki, Ö.C.; Eylem, C.C.; Reçber, T.; Kır, S.; Nemutlu, E. Integration of GC–MS and LC–MS for Untargeted Metabolomics Profiling. *J. Pharm. Biomed. Anal.* **2020**, *190*, 113509. [CrossRef] [PubMed]
4. De Livera, A.M.; Dias, D.A.; De Souza, D.; Rupasinghe, T.; Pyke, J.; Tull, D.; Roessner, U.; McConville, M.; Speed, T.P. Normalizing and Integrating Metabolomics Data. *Anal. Chem.* **2012**, *84*, 10768–10776. [CrossRef]
5. Scholz, M.; Gatzek, S.; Sterling, A.; Fiehn, O.; Selbig, J. Metabolite Fingerprinting: Detecting Biological Features by Independent Component Analysis. *Bioinformatics* **2004**, *20*, 2447–2454. [CrossRef] [PubMed]
6. Borrego, S.L.; Fahrmann, J.; Datta, R.; Stringari, C.; Grapov, D.; Zeller, M.; Chen, Y.; Wang, P.; Baldi, P.; Gratton, E.; et al. Metabolic Changes Associated with Methionine Stress Sensitivity in MDA-MB-468 Breast Cancer Cells. *Cancer Metab.* **2016**, *4*, 9. [CrossRef]
7. Redestig, H.; Fukushima, A.; Stenlund, H.; Moritz, T.; Arita, M.; Saito, K.; Kusano, M. Compensation for Systematic Cross-Contribution Improves Normalization of Mass Spectrometry Based Metabolomics Data. *Anal. Chem.* **2009**, *81*, 7974–7980. [CrossRef] [PubMed]
8. Sysi-Aho, M.; Katajamaa, M.; Yetukuri, L.; Orešič, M. Normalization Method for Metabolomics Data Using Optimal Selection of Multiple Internal Standards. *BMC Bioinform.* **2007**, *8*, 93. [CrossRef]
9. Boysen, A.K.; Heal, K.R.; Carlson, L.T.; Ingalls, A.E. Best-Matched Internal Standard Normalization in Liquid Chromatography-Mass Spectrometry Metabolomics Applied to Environmental Samples. *Anal. Chem.* **2018**, *90*, 1363–1369. [CrossRef]
10. Dunn, W.B.; Wilson, I.D.; Nicholls, A.W.; Broadhurst, D. The Importance of Experimental Design and QC Samples in Large-Scale and MS-Driven Untargeted Metabolomic Studies of Humans. *Bioanalysis* **2012**, *4*, 2249–2264. [CrossRef]
11. Li, B.; Tang, J.; Yang, Q.; Li, S.; Cui, X.; Li, Y.; Chen, Y.; Xue, W.; Li, X.; Zhu, F. NOREVA: Normalization and Evaluation of MS-Based Metabolomics Data. *Nucleic Acids Res.* **2017**, *45*, W162–W170. [CrossRef] [PubMed]
12. De Livera, A.M.; Sysi-Aho, M.; Jacob, L.; Gagnon-Bartsch, J.A.; Castillo, S.; Simpson, J.A.; Speed, T.P. Statistical Methods for Handling Unwanted Variation in Metabolomics Data. *Anal. Chem.* **2015**, *87*, 3606–3615. [CrossRef] [PubMed]
13. Fan, S.; Kind, T.; Cajka, T.; Hazen, S.L.; Tang, W.H.W.; Kaddurah-Daouk, R.; Irvin, M.R.; Arnett, D.K.; Barupal, D.K.; Fiehn, O. Systematic Error Removal Using Random Forest for Normalizing Large-Scale Untargeted Lipidomics Data. *Anal. Chem.* **2019**, *91*, 3590–3596. [CrossRef] [PubMed]
14. Viant, M.R.; Ebbels, T.M.D.; Beger, R.D.; Ekman, D.R.; Epps, D.J.T.; Kamp, H.; Leonards, P.E.G.; Loizou, G.D.; MacRae, J.I.; van Ravenzwaay, B.; et al. Use Cases, Best Practice and Reporting Standards for Metabolomics in Regulatory Toxicology. *Nat. Commun.* **2019**, *10*, 3041. [CrossRef] [PubMed]
15. Law, K.P.; Han, T.L.; Yang, Y.; Zhang, H. Analytical Challenges of Untargeted GC-MS-Based Metabolomics and the Critical Issues in Selecting the Data Processing Strategy. *F1000Research* **2017**, *6*, 967.
16. Zhao, Y.; Hao, Z.; Zhao, C.; Zhao, J.; Zhang, J.; Li, Y.; Li, L.; Huang, X.; Lin, X.; Zeng, Z.; et al. A Novel Strategy for Large-Scale Metabolomics Study by Calibrating Gross and Systematic Errors in Gas Chromatography-Mass Spectrometry. *Anal. Chem.* **2016**, *88*, 2234–2242. [CrossRef] [PubMed]
17. Duan, L.; Ma, A.; Meng, X.; Shen, G.A.; Qi, X. QPMASS: A Parallel Peak Alignment and Quantification Software for the Analysis of Large-Scale Gas Chromatography-Mass Spectrometry (GC-MS)-Based Metabolomics Datasets. *J. Chromatogr. A* **2020**, *1620*, 460999. [CrossRef]
18. Bijlsma, S.; Bobeldijk, I.; Verheij, E.R.; Ramaker, R.; Kochhar, S.; Macdonald, I.A.; Van Ommen, B.; Smilde, A.K. Large-Scale Human Metabolomics Studies: A Strategy for Data (Pre-) Processing and Validation. *Anal. Chem.* **2005**, *78*, 567–574. [CrossRef]
19. Adeola, H.A.; Papagerakis, S.; Papagerakis, P. Systems Biology Approaches and Precision Oral Health: A Circadian Clock Perspective. *Front. Physiol.* **2019**, *10*, 399. [CrossRef]
20. Fiehn, O. Metabolomics by Gas Chromatography–Mass Spectrometry: Combined Targeted and Untargeted Profiling. *Curr. Protoc. Mol. Biol.* **2016**, *114*, 30.4.1–30.4.32. [CrossRef]
21. Beale, D.J.; Pinu, F.R.; Kouremenos, K.A.; Poojary, M.M.; Narayana, V.K.; Boughton, B.A.; Kanojia, K.; Dayalan, S.; Jones, O.A.H.; Dias, D.A. Review of Recent Developments in GC–MS Approaches to Metabolomics-Based Research. *Metabolomics* **2018**, *14*, 152. [CrossRef] [PubMed]
22. Khodadadi, M.; Pourfarzam, M. A Review of Strategies for Untargeted Urinary Metabolomic Analysis Using Gas Chromatography-Mass Spectrometry. *Metabolomics* **2020**, *16*, 66. [CrossRef] [PubMed]

23. Curtius, H.C.; Wolfensberger, M.; Steinmann, B.; Redweik, U.; Siegfried, J. Mass Fragmentography of Dopamine and 6-Hydroxydopamine: Application to the Determination of Dopamine in Human Brain Biopsies from the Caudate Nucleus. *J. Chromatogr. A* **1974**, *99*, 529–540. [CrossRef]
24. Šťávová, J.; Beránek, J.; Nelson, E.P.; Diep, B.A.; Kubátová, A. Limits of Detection for the Determination of Mono- and Dicarboxylic Acids Using Gas and Liquid Chromatographic Methods Coupled with Mass Spectrometry. *J. Chromatogr. B* **2011**, *879*, 1429–1438. [CrossRef] [PubMed]
25. Rahn, W.; König, W.A. GC/MS Investigations of the Constituents in a Diethyl Ether Extract of an Acidified Roast Coffee Infusion. *J. High Resolut. Chromatogr.* **1978**, *1*, 69–71. [CrossRef]
26. Lamoureux, G.; Agüero, C. A Comparison of Several Modern Alkylating Agents. *Arkivoc* **2009**, *2009*, 251–264. [CrossRef]
27. Liebeke, M.; Puskás, E. Drying enhances signal intensities for global GC–MS metabolomics. *Metabolites* **2019**, *9*, 68. [CrossRef]
28. Fiehn, O.; Kopka, J.; Dörmann, P.; Altmann, T.; Trethewey, R.N.; Willmitzer, L. Metabolite profiling for plant functional genomics. *Nat. Biotechnol.* **2000**, *18*, 1157–1161. [CrossRef]
29. Piergiovanni, M.; Termopoli, V. Derivatization Strategies in Flavor Analysis: An Overview over the Wine and Beer Scenario. *Chemistry* **2022**, *4*, 1679–1695. [CrossRef]
30. Barupal, D.K.; Zhang, Y.; Shen, T.; Fan, S.; Roberts, B.S.; Fitzgerald, P.; Wancewicz, B.; Valdiviez, L.; Wohlgemuth, G.; Byram, G.; et al. A Comprehensive Plasma Metabolomics Dataset for a Cohort of Mouse Knockouts within the International Mouse Phenotyping Consortium. *Metabolites* **2019**, *9*, 101. [CrossRef]
31. Yu, S.; Fan, J.; Zhang, L.; Qin, X.; Li, Z. Assessment of Biphasic Extraction Methods of Mouse Fecal Metabolites for Liquid Chromatography-Mass Spectrometry-Based Metabolomic Studies. *J. Proteome Res.* **2021**, *20*, 4487–4494. [CrossRef] [PubMed]
32. Badawy, A.A.B.; Morgan, C.J.; Turner, J.A. Application of the Phenomenex EZ:FaastTM Amino Acid Analysis Kit for Rapid Gas-Chromatographic Determination of Concentrations of Plasma Tryptophan and Its Brain Uptake Competitors. *Amino Acids* **2008**, *34*, 587–596. [CrossRef] [PubMed]
33. Liang, J.; Liu, R. Stacked Denoising Autoencoder and Dropout Together to Prevent Overfitting in Deep Neural Network. In Proceedings of the 2015 8th International Congress on Image and Signal Processing, CISP 2015, Shenyang, China, 14–16 October 2015; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2016; pp. 697–701.
34. Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; Manzagol, P.A.; Bottou, L. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.* **2010**, *11*, 3371–3408.
35. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
36. Simón-Manso, Y.; Lowenthal, M.S.; Kilpatrick, L.E.; Sampson, M.L.; Telu, K.H.; Rudnick, P.A.; Mallard, W.G.; Bearden, D.W.; Schock, T.B.; Tchekhovskoi, D.V.; et al. Metabolite Profiling of a NIST Standard Reference Material for Human Plasma (SRM 1950): GC-MS, LC-MS, NMR, and Clinical Laboratory Analyses, Libraries, and Web-Based Resources. *Anal. Chem.* **2013**, *85*, 11725–11731. [CrossRef] [PubMed]
37. Ballman, K.V.; Grill, D.E.; Oberg, A.L.; Therneau, T.M. Faster Cyclic Loess: Normalizing RNA Arrays via Linear Models. *Bioinformatics* **2004**, *20*, 2778–2786. [CrossRef] [PubMed]
38. Dunn, W.B.; Broadhurst, D.; Begley, P.; Zelena, E.; Francis-Mcintyre, S.; Anderson, N.; Brown, M.; Knowles, J.D.; Halsall, A.; Haselden, J.N.; et al. Procedures for Large-Scale Metabolic Profiling of Serum and Plasma Using Gas Chromatography and Liquid Chromatography Coupled to Mass Spectrometry. *Nat. Protoc.* **2011**, *6*, 1060–1083. [CrossRef] [PubMed]
39. Lange, S.; Riedmiller, M. Deep Auto-Encoder Neural Networks in Reinforcement Learning. In Proceedings of the 2010 International Joint Conference on Neural Networks (IJCNN), Barcelona, Spain, 18–23 July 2010; pp. 1–8.
40. Parsons, H.M.; Ekman, D.R.; Collette, T.W.; Viant, M.R. Spectral Relative Standard Deviation: A Practical Benchmark in Metabolomics. *Analyst* **2009**, *134*, 478–485. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.