sensors

MDPI

Article

Retention Time Trajectory Matching for Peak Identification in Chromatographic Analysis

Wenzhe Zang [1,2,3], Ruchi Sharma [1,2,3], Maxwell Wei-Hao Li [1,2,3,4] and Xudong Fan [1,2,3,*]

[1] Department of Biomedical Engineering, University of Michigan, 1101 Beal Avenue, Ann Arbor, MI 48109, USA
[2] Center for Wireless Integrated MicroSensing and Systems (WIMS2), University of Michigan, Ann Arbor, MI 48109, USA
[3] Max Harry Weil Institute for Critical Care Research and Innovation, University of Michigan, Ann Arbor, MI 48109, USA
[4] Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109, USA
* Correspondence: xsfan@umich.edu

Abstract: Retention time drift caused by fluctuations in physical factors such as temperature ramping rate and carrier gas flow rate is ubiquitous in chromatographic measurements. Proper peak matching and identification across different chromatograms is critical prior to any subsequent analysis but is challenging without using mass spectrometry. The purpose of this work was to describe and validate a peak matching and identification method called retention time trajectory (RTT) matching that can be used in targeted analyses free of mass spectrometry. This method uses chromatographic retention times as the only input and identifies peaks associated with any subset of a predefined set of target compounds. An RTT is a two-dimensional (2D) curve formed uniquely by the retention times of the chromatographic peaks. The RTTs obtained from the chromatogram of a sample under test and those pre-installed in a library are matched and statistically compared. The best matched pair implies identification. Unlike most existing peak-alignment methods, no mathematical warping or transformation is involved. Based on the experimentally characterized RTT, an RTT hybridization method was also developed to rapidly generate more RTTs and expand the library without performing actual time-consuming chromatographic measurements, which enables successful peak matching even for chromatograms with severe retention time drifts. Additionally, 3.15×10^5 tests using experimentally obtained gas chromatograms and 2×10^{12} tests using two publicly available fruit metabolomics datasets validated the proposed method, demonstrating real-time peak/interferent identification.

Keywords: gas chromatogram; liquid chromatography; retention time; retention time drift; peak identification

Citation: Zang, W.; Sharma, R.; Li, M.W.-H.; Fan, X. Retention Time Trajectory Matching for Peak Identification in Chromatographic Analysis. *Sensors* **2023**, *23*, 6029. https://doi.org/10.3390/s23136029

Academic Editor: Jun Wang

Received: 20 May 2023
Revised: 26 June 2023
Accepted: 27 June 2023
Published: 29 June 2023

1. Introduction

Gas chromatography (GC)-based volatile organic compound (VOC) analysis can be classified into untargeted analysis and targeted analysis. The former involves evaluation of chemical substances in an unknown sample, whereas the latter aims only at a predetermined list of interesting compounds or a subset of those, with all other VOCs treated as interferents. Due to the complexity of sample composition and the lack of pre-existing knowledge, accurate identification in untargeted analysis requires confirmation or cross-validation by at least two parameters, such as chromatographic retention time (RT) and mass spectrometry (MS) fragmentation profile. In contrast, in targeted analysis, oftentimes only the retention time is used for compound identification in order to avoid using bulky and expensive mass spectrometry. Therefore, targeted analysis has broad applications in on-site real-time measurements, such as environmental protection [1–3], workplace environment monitoring [4–8], industries (e.g., petroleum [9–11] and food [12,13]), and metabolomics [14–16].

For targeted analysis, the retention time of each peak in the GC chromatogram is compared with the pre-installed values of all compounds of interest in a library. In any given sample, a positive alarm is reported when the retention time of a peak matches a corresponding time in the library; the lack of any match instead means that a peak would be ignored or reported as an interferent. However, variations in physical factors such as ambient temperature, column temperature ramping profile, and carrier gas flow rate can affect the retention time of each peak from run to run, which hinders identification or triggers false alarms. The inability to correctly identify peaks with only retention times is exacerbated when a sample contains a large number of compounds or when some of the targeted peaks are closely eluted out in a chromatogram. Consequently, proper matching or alignment of chromatographic peaks across different samples is a crucial preprocessing step that must be performed prior to any subsequent analysis.

A simple and popular solution is data binning, which divides the signals into bins (e.g., histograms) and incorporates all data into a recognition profile for each measurement [17,18]. The binning method is easy to use and shows acceptable performance in processing both chromatograms and spectra when the peak drift from sample to sample is much smaller than the distance between two adjacent peaks. However, in the presence of large peak drifts, this approach suffers from reduced resolution and information loss (see illustrations in Figure S1). The time warping technique, such as segment-wise correlation optimized warping (COW) [19], point-wise dynamic time warping (DTW) [20], global polynomial model-based parametric time warping (PTW) [21,22], multiscale peak alignment (MSPA) [23], and other variants [24–27], is one of most commonly adopted methods to correct retention time drifts across chromatograms. It aligns a whole measured chromatogram profile against a reference chromatogram using pattern recognition routines in order to achieve peak identification. While time warping is powerful and works well with samples of various complexities, an accurate warping-based aligning demands the fine tuning of alignment parameters, which can often involve human intervention, thus making automated peak identification less reliable. Moreover, all warping-based methods can suffer, to different degrees, from misalignments and are concentration-sensitive, even with samples of the same compositions. In some cases, warping-based aligning approaches may not be able to yield exactly the same retention time value for the same analyte from different measurements. Consequently, subsequent retention-time-based peak identification or statistical analysis often requires further value correction via data binning or clustering. Machine learning-based aligning approaches, which utilize artificial intelligence systems to acquire knowledge by extracting patterns from data, are also able to achieve positive alignment with decent accuracy [28–32], and they are amenable to automation without relying on human intervention. However, these approaches often employ the mass spectra of the peaks as one the of key subnetworks among the overall network architecture, making it more suitable for bulky mass-spectrometry-based analysis rather than for onsite monitoring where mass spectrometry is not used. Moreover, similar to most other machine learning-based approaches, machine learning-based aligning suffers from high computational cost during parameter training and feature extraction.

While the aforementioned methods can mitigate the peak drift issue across chromatograms, they all suffer from a number of drawbacks that make them unsuitable for compound identification in targeted analysis. First and foremost, the total number of peaks in the chromatogram obtained from the sample under test needs to be exactly the same as in the reference chromatogram. If only a subset or, in an extreme case, a single species of the compounds of interest (target compounds) is present in the sample under test, which is often the case for targeted analysis (e.g., pipeline leakage detection in a chemical plant), chromatogram aligning fails, as the detected peaks in the chromatogram have nothing to align with. Second, foreign interferents cannot be filtered out. If an additional peak is present in a chromatogram of the sample under test, it would be treated as one of the target compounds (misalignment) and/or it may cause failure in alignment of the whole chromatogram profile. Consequently, there has been an unmet need for an MS-free and

chromatogram-based peak identification method that is able to identify any arbitrary subset of the target compounds as well as to report interferents.

This paper proposes a retention time trajectory (RTT) matching method (Figure 1A) for peak identification. With peak retention times as the only input, RTT matching can identify peaks associated with any subset of target compounds, as well as filter interferents outside of said targets. An RTT (Figure 1B) is made up of a series of retention times of all compounds (peaks) in a chromatogram obtained under one set of experimental conditions (such as ambient temperature, column temperature ramping profile, carrier gas flow rate, etc.) and uniquely represents one particular condition. Similarities between the RTT for the sample under test (RTT_{sample}) and those pre-installed in the library (RTT_{lib}) are globally evaluated using a statistical expression, until an RTT_{sample} that best matches an RTT_{lib} is found. In contrast to most existing MS-free chromatogram-aligning algorithms, our RTT algorithm identifies peaks through simple *matching* instead of *chromatogram aligning*, that is, our method does not align the chromatogram of the sample under test against the reference chromatogram. Therefore, no mathematical warping or transformation is involved.

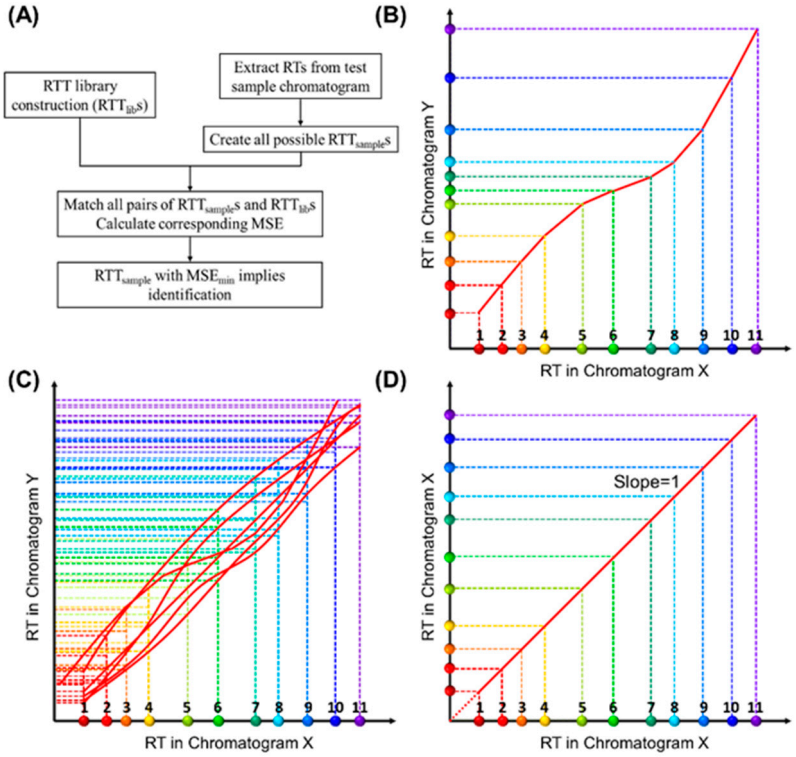

Figure 1. Conceptual illustration of retention time trajectory (RTT) and corresponding library construction. (**A**) Algorithm flow chart. (**B**) RTT of a chromatogram (Chromatogram Y)—RTT_Y. The *X*-axis represents the retention time (RT_X) from one of the chromatograms, which we call Chromatogram X. The colored dots along the *X*-axis represent different compounds and are numerically labelled as 1, 2, 3, . . . , 11. Similarly, the *Y*-axis represents the retention time, RT_Y, in another chromatogram, Chromatogram Y. The entire set of coordinates, ($RT_{X,compound\ i}$, $RT_{Y,compound\ i}$), where "compound *i*" refers to a specific compound, forms a trajectory (red curve) in 2D. (**C**) Conceptual illustration of the RTT_{lib} library, which is composed of multiple chromatograms obtained under various experimental conditions (column temperature ramping profiles, carrier gas flow rate, etc.) using a mixture containing *all* target compounds (and internal standards if needed). (**D**) RTT of Chromatogram X (RTT_X), where the retention time values along the *Y*-axis exactly match those along the *X*-axis (i.e., the slope of RTT_X is unity). Note that the RTT method, as shown in detail later, uses only the discrete peak retention time values. Smooth curves drawn here are for visualization purposes.

2. Retention Time Trajectory Matching Algorithm

A glossary of abbreviations and symbols is summarized in Section S1 in the Supplementary Materials.

2.1. Overview

Between GC analytical runs, the retention time (RT) for a given compound may drift due to perturbation of various physical factors, including ambient temperature, column temperature programing profile, and carrier gas flow rate. The influence of these perturbations on the analytes in a sample can be quite different due to their diverse characteristics (such as volatility, polarity, functional groups, etc.). Consequently, the RT drifts of the analytes in a chromatogram are often non-linear and unpredictable [21,33]. The RT deviation (ΔRT) against RT has been described using quadratic functions in PTW [20,21] or local regression fitting (LOESS) in XCMS [33], which often over-simplifies the diverse and complex nature of RT drifting. These methods are either limited to samples with the same constitutions [21] or require MS-based peak matching before final aligning [33]. In contrast, our RTT matching approach treats the RT drifts of *all* analytes (or peaks) in a chromatogram as a whole cohesive entity rather than independent individuals. Instead of fitting the RT drift with a mathematical formula, our RTT matching approach statistically compares the similarities between the RTT_{sample}s and the RTT_{lib}s globally.

2.2. Construction of a Retention Time Trajectory (RTT)

First, we describe how to construct a library containing many RTT_{lib}s. Let us assume that we obtain multiple chromatograms under various experimental conditions, and each chromatogram includes *all* compounds, i.e., target compounds and internal standard compounds (if needed). The *X*-axis of Figure 1B represents the retention time (RT_X) obtained from one of the measured chromatograms, which we call Chromatogram X. The dots of different colors along the *X*-axis represent different compounds. RTs of any pre-characterized chromatogram can be used as the *X*-axis. Similarly, the *Y*-axis represents the retention time, RT_Y, in another chromatogram. Therefore, the set of coordinates, ($RT_{X,compound\ i}$, $RT_{Y,compound\ i}$), where "compound *i*" refers to a specific compound, forms a trajectory in a two-dimensional (2D) diagram (Figure 1B), which we call retention time trajectory (RTT). Each trajectory corresponds uniquely to a chromatogram obtained under a certain set of conditions (temperature ramping, flow rate, etc.), and the entire set of trajectories in the 2D diagram (i.e., the library of RTT_{lib}s, Figure 1C) captures all chromatograms under various experimental conditions. One special case is that Chromatogram X itself is represented by a straight line at 45° with respect to the *X*-axis (Figure 1D), as the coordinates along the *X*- and *Y*-axis exactly match each other. Ideally, the RTT library should contain an infinite number of RTTs to reflect the infinite number of experimental conditions (for example, the flow rate may vary by only 0.0001 mL/min and ramping temperature rate may vary by 0.001 °C/min). However, as we will discuss later, a library containing a finite (small) number of RTTs is sufficient for us to correctly identify peaks in a chromatogram for the sample under test.

Construction of an RTT_{sample} for a sample under test is similar (Figure 2A). Each RTT_{sample} is made up of a series of coordinates, ($RT_{X,compound\ i}$, $RT_{sample,peak\ j}$), where $RT_{sample,\ peak\ j}$ refers to the retention time of a peak in the chromatogram obtained from the sample under test (i.e., sample chromatogram). Note that the peaks in the sample chromatogram may contain all target compounds plus some interferents or only a subset of target compounds plus some interferents. Because the chemical identities of the detected peaks are unknown before analysis, multiple RTT_{sample}s may form for a given sample chromatogram, each of which corresponds to one set of peak identification results. Our goal is to eliminate all impossible RTT_{sample}s and find the one that best matches one of the RTT_{lib}s in the library.

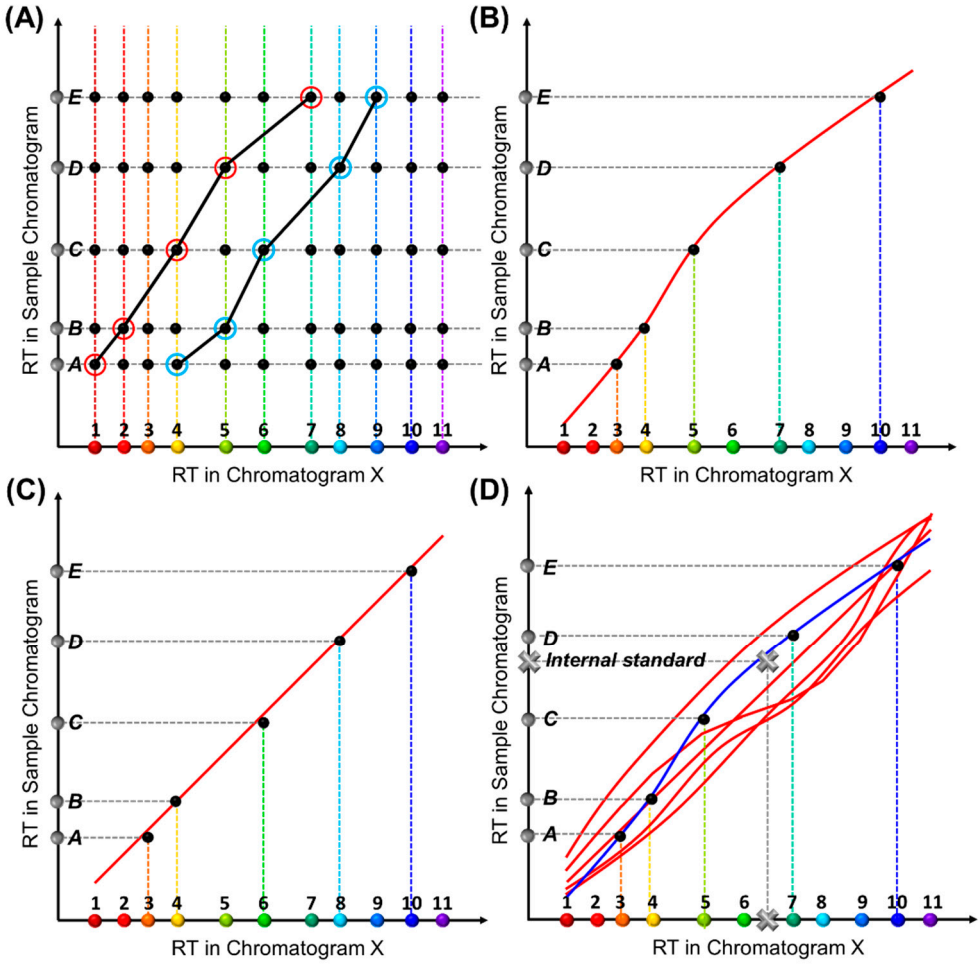

Figure 2. Conceptual illustration of the generation of an RTT for a sample under test, RTT_{sample}. (**A**) A retention time matrix formed by RT_X and RT_{sample} of the same compounds. Suppose there are a total of N_{tgt} target analytes (colored dots along the *X*-axis) and a total of N_{sample} peaks in the sample chromatogram (grey dots denoted from A to E along the *Y*-axis). A matrix of $N_{tgt} \times N_{sample}$ coordinates (black dots) can be formed. The two black curves show two exemplary RTT_{sample}s formed by connecting one dot on each row. (**B**,**C**) Two sets of black dots fall simultaneously on two different RTT_{lib}s, suggesting that two RTT_{sample}s are found to match two different RTT_{lib}s, which leads to two different chemical identification results. Accordingly, Peaks A-E are identified as Compounds 3, 4, 5, 7, and 10 in (**B**), and Compounds 3, 4, 6, 8, and 10 in (**C**), respectively. (**D**) An internal standard (grey cross) anchors one of the RTT_{lib}s (plotted in blue), thus producing unique chemical identification.

2.3. General Description of the RTT Matching Approach

In general, our RTT matching approach involves four steps:

(1) Experimentally generate multiple chromatograms under various experimental conditions (temperature ramping profiles, flow rate, etc.) using a mixture that contains *all* target compounds and internal standards (if needed). Retention times in each chromatogram are extracted, which form the RTT_{lib}s (see Figure 1C) pre-installed in the library. Note that the chemical identities of all peaks in any RTT_{lib} are known.

(2) Create all possible RTT_{sample}s from the chromatogram obtained from the sample under test (see Figure 2). Note that the test sample may contain only a subset of the target compounds along with some interferents.

(3) Eliminate the RTT_{sample}s that violate certain rules, which expedites computation.

(4) Compare all possible RTT_{sampe}s with the RTT_{lib}s in the library one by one in terms of similarities. Find the RTT_{sample} that best matches one of the RTT_{lib}s and then extract the chemical identities for the detected peaks accordingly. Again, our method does

not align the peaks in the sample chromatogram against those in the chromatograms in the library.

In the next section, the following two cases will be considered. (I) Only a subset of compounds of interest (target compounds) and internal standards (if needed) are present in the sample chromatogram. (II) A few chemical interferents are present in the sample chromatogram.

2.4. Case I: Sample Containing Only a Subset of Target Compounds

2.4.1. Construction of Possible RTT_{sample}s

Consider a sample under test that has no interferents present (i.e., all detected peaks are a subset of the target compounds). We assume that there are a total of N_{tgt} target compounds and that N_{sample} peaks are detected in the sample under test. If $N_{sample} = N_{tgt}$, then all N_{sample} peaks can be identified easily by examining the peak elution order.

We now consider only the situation in which $N_{sample} < N_{tgt}$, that is, only a subset of analytes are present in the sample chromatogram. A matrix with $N_{tgt} \times N_{sample}$ intersections (coordinates) is then formed in the 2D diagram (marked as black dots, i.e., coordinates in Figure 2A), because each peak in the sample under test is unknown before final identification and, in principle, can be any of the N_{tgt} target compounds. Consequently, a total of $C\left(N_{tgt}, N_{sample}\right)$ sets of RTT_{sample}s can be formed by connecting one black dot in each of the N_{sample} rows, two of which are exemplified as black lines in Figure 2A. $C\left(N_{tgt}, N_{sample}\right) = \frac{N_{tgt}!}{(N_{tgt}-N_{sample})!N_{sample}!}$ is a combinatorial number and can be extremely large when N_{tgt} is above 20 and N_{sample} is around half of N_{tgt} (for example, $C(20,10) = 184{,}756$).

However, not all $C\left(N_{tgt}, N_{sample}\right)$ RTT_{sample}s are possible, and many need to be eliminated before the comparison with the RTT_{lib}s in the library, which expedites computation and avoids false identifications. The elimination rules are described as follows:

(1) One target compound can only be mapped to one peak in the sample chromatogram. Therefore, any RTT_{sample} with a vertical section between any two consecutive coordinates (or intersections) in the 2D diagram should be eliminated, as illustrated in Figure S2A.

(2) The elution order should be preserved. Therefore, any RTT_{sample} with a section that has a negative slope between two consecutive coordinates (i.e., opposite elution order in library and sample chromatograms) should be eliminated. This is illustrated in Figure S2B.

(3) The RT drifts arise from minor perturbation, and the resulting deviations (ΔRT) should be small values within a certain range. Therefore, only coordinates falling within the cutoff range of $RT_{sample} \pm \Delta t$ should be considered (Figure S2C). The value of Δt can be estimated empirically. For example, sample chromatograms with larger drifts require sufficiently large Δt (e.g., larger than typical RT drifting range in the RTT library). Note that the RTT matching algorithm is still able to effectively identify the peaks even without applying this criterion, but a reasonable estimation of Δt significantly reduces the computational cost by narrowing down possible RTT_{sample}s.

Once all possible RTT_{sample}s are formed (and all impossible RTT_{sample}s are eliminated), each individual RTT_{sample} should be compared with all RTT_{lib}s in the library to find the best-matched RTT_{sample}. In other words, we need to find which set of coordinates in Figure 2A falls on one red RTT_{lib} in Figure 1C. Assuming a total of n_{lib} RTT_{lib}s stored in the library and n_{sample} possible RTT_{sample}s generated from the sample chromatogram, $n_{lib} \times n_{sample}$ pairs of RTT_{lib} and RTT_{sample} are formed and compared with each other. For example, by comparing the trajectories in Figure 1C, two groups of black dots, which are composed of two different RTT_{sample}s, fall simultaneously on two different RTT_{lib}s (Figure 2B,C) and yield different chemical identification results for Peaks C and D. These are either Compounds 5 and 7 (Figure 2B) or 6 and 8 (Figure 2C).

To circumvent this, we introduce internal standard compounds (i.e., internal standards) outside of the list of target compounds to anchor the RTT_{lib}s. In both RTT_{lib} library preparation and actual measurement of the test sample, the internal standard(s) are spiked into the mixture containing *all* target compounds (for RTT_{lib} library preparation) or the sample under test. The peaks corresponding to these standards are identified during the data preprocessing and then used to generate the RTT along with all other peaks. As depicted in Figure 2D, when an internal standard (marked as a grey cross) is introduced, only one of the two RTT_{lib}s can be anchored (blue line) and thus a unique set of identification results is obtained.

Note that while our RTT matching method, as shown later, works well even without using any internal standard, introduction of an internal standard(s) can further increase identification accuracy and significantly expedite computation by narrowing down possible RTT_{sample}s and RTT_{lib}s due to the following reasons. (1) All possible RTT_{sample}s and all RTT_{lib}s must go through the coordinate(s) formed by the internal standard(s). (2) Because the elution order is preserved, the whole 2D diagram can be divided into small regions determined by the internal standards' coordinates, and only the RTT_{sample}s falling within these regions are possible candidates (Figure S3A). (3) When there is only a single analyte in the test sample, identification of this peak in the chromatogram is nearly impossible without internal standard(s), as formation of an RTT requires at least two coordinates. The addition of one or more coordinates resulting from the internal standards allows for the creation of the RTT_{sample}s with more accurate identification of the single peak. (4) Misidentification can be significantly reduced even when the chromatogram-to-chromatogram RT drift of the same compound is greater than the distance between two adjacent peaks (Figure S3B), which has long been the bottleneck of many peak-matching or profile-aligning algorithms [33]. In practice, internal standards can be strategically positioned in the region with more drastic variations to more effectively narrow down RTT_{sample} and RTT_{lib} selection (Figure S3C). Note that, similar to all internal-standard-based chromatographic analysis methods, the addition of internal standards might potentially worsen the co-elution issues with the neighboring target compounds. To avoid this, internal standards whose RTs fall in the chromatogram sections with low peak densities are preferred.

Internal standards have commonly been utilized by many aligning algorithms [34,35], in which RT drift is corrected by firstly dividing the chromatogram into multiple sections delineated by the standards and secondly applying linear stretching/compressing in each section (Figure S4). However, these methods cannot account for the various non-linear drifts that often occur between any two standards. To make the linear stretching/compressing more accurate, more internal standards need to be introduced to reduce the section size to the point that linear approximation within each section is valid. This makes both sample preparation and peak identification much more complicated. More advanced techniques employ polynomial fitting within each section to account for non-linearity. However, polynomial fitting highly depends on experimental conditions and is required for each section in a chromatogram, thus hampering automation in peak identification. In contrast, as discussed in detail later, our approach compares the retention times of the chromatogram (or its corresponding RTT) globally, which automatically takes into account any non-linearities within each section. To demonstrate the advantages of RTT matching, we use the same sample and internal standards to compare the performance of the RTT matching approach and linear warping (see examples in Table S5), as well as correlation optimized warping (COW, see examples in Figures S8 and S9, and Table S4 in the Supplementary Materials).

2.4.2. Statistical Method for RTT Similarity Comparison—Least Mean Squared Residual (MSR)

As mentioned previously, while coordinates in the X-axis are discrete, their variation along the Y-axis is continuous due to continuously varying experimental conditions. In theory, the number of the RTT_{lib}s is infinite. In practice, only a limited number of conditions, and, hence, a limited number of RTT_{lib}s, can be characterized and stored. Consequently,

while the RTT_{sample} for the sample under test may not exactly match any RTT_{lib} stored in the library, the most similar one can still be easily found. To globally compare similarities between an RTT_{sample} and an RTT_{lib}, we calculate the mean squared residuals (*MSR*) of RTs from the same compounds between these two trajectories, as illustrated in Figure 3. A smaller *MSR* indicates higher similarity between two trajectories, as exemplified in Figure 3B, where $RTT_{lib(3)}$ is the most similar to the RTT_{sample} in the figure. The *MSR* is normalized (or scaled) from the sum of squared residuals (*SSR*) by the total number of paired compounds in order to ensure that it does not grow as the number of pairs grows. This is important when we need to compare RTT_{sample}s with different numbers of target compounds; for example, when an RTT_{sample} of six target compounds is compared with an RTT_{sample} of five target compounds plus one interferent (the case of interferent identification will be discussed later).

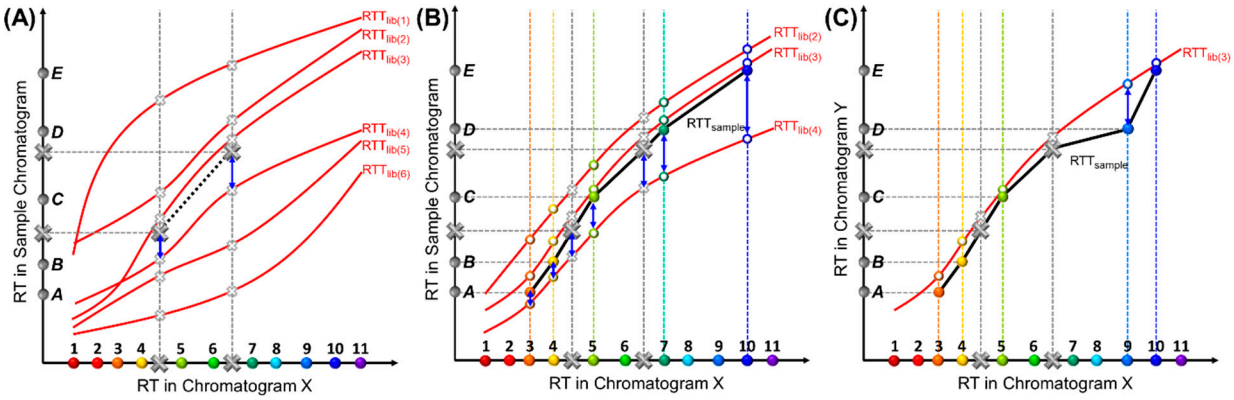

Figure 3. MSR or SSR calculation. (**A**) Illustration of calculating the $SSRst^d$ between an RTT_{sample} and $RTT_{lib(4)}$ using the internal standards for RTT_{libs} screening. All $RTT_{samples}$ must pass through the two large solid grey crosses, which indicate the RTs of the internal standards in the sample chromatogram. The smaller hollow grey crosses indicate the RTs of the internal standards in the RTT_{libs}. The internal standard retention time residuals are marked by blue arrows. Half of the RTT_{libs} with the lowest $SSRst^d$s (i.e., $RTT_{lib(2)}$, $RTT_{lib(3)}$, and $RTT_{lib(4)}$) are kept and used for the next step described in (**B**). (**B**) Calculation of the MSR between an RTT_{sample} and an RTT_{lib} screened in (A) using both target compounds and internal standards. This process quantifies the similarities between the two RTTs under comparison (i.e., one RTT_{sample} and one RTT_{lib}). The RTT_{sample} (black curve) is formed by pairing Peaks A to E with target Compounds 3, 4, 5, 7, and 10. The retention time residual of the same compound is marked by a blue arrow. (**C**) Interferent identification. One RTT_{sample} (black curve) is formed by pairing Peaks A to E with target Compounds 3, 4, 5, 9, and 10, respectively, and is further compared with $RTT_{lib(3)}$ (red curve). The retention time residual between Peak D and Compound 9 is much larger than those of other pairs. The physical interpretation is that the coordinates associated with Peaks A, B, C, and E in one RTT_{sample} closely match those on $RTT_{lib(3)}$. However, Peak D deviates far from $RTT_{lib(3)}$, likely identifying Peak D as an interferent, and identifying Peaks A, B, C, and E as Compounds 3, 4, 5, and 10, respectively.

In order to expedite the computation, it is not necessary to compare each RTT_{sample} with all RTT_{lib}s. Instead, we can first use internal standard RTs to anchor the best matching sets of RTT_{lib}s by calculating the *SSR* of the internal standards ($SSRst^d$), as shown in Figure 3A. Because all possible RTT_{sample}s must go through the coordinates formed by the internal standards, they have the same $SSRst^d$ for a given $RTT_{lib(i)}$, where i refers to a specific pre-characterized RTT_{lib}. Assuming that there are a total of N_{std} internal standards, the $SSRst^d$ between any RTT_{sample} and an $RTT_{lib(i)}$, denoted as $SSR^{std}_{lib(i)}$, can be calculated as

$$SSR^{std}_{lib(i)} = \sum_{k=1}^{N_{std}} \left(RT^{std(k)}_{lib(i)} - RT^{std(k)}_{sample} \right)^2,$$

where $RT_{lib(i)}^{std(k)}$ is the retention time of one internal standard, k, in $\mathrm{RTT_{lib(}}_{i)}$, and $RT_{sample}^{std(k)}$ is the retention time of the same internal standard in the sample chromatogram. The $SSRst^d$ of all $\mathrm{RTT_{lib}}$s are sorted in ascending order, with the top ones, which have the least $SSRst^d$s, giving the potentially matched $\mathrm{RTT_{lib}}$s. All other $\mathrm{RTT_{lib}}$s in the library, which have higher $SSRst^d$s, can be eliminated.

Next, the $\mathrm{RTT_{sample}}$ and $\mathrm{RTT_{lib}}$ are further compared, based on retention times of both internal standards and target compounds, by sorting the *MSR* (Figure 3B). In this step, only the top $\mathrm{RTT_{lib}}$s (e.g., the first half or the top 20 $\mathrm{RTT_{lib}}$s) with the lowest $SSRst^d$s are selected. Assuming that, in addition to N_{std} internal standards, there are N_{sample} peaks to be identified in the test sample, the *MSR* between one $\mathrm{RTT_{lib}}$ (denoted as $\mathrm{RTT_{lib(}}_{i)}$) and one $\mathrm{RTT_{sample}}$ (denoted as $\mathrm{RTT_{sample(}}_{j)}$) can be calculated as

$$MSR = \frac{\mathrm{SSR}_{lib(i),sample(j)}}{N_{std}+N_{sample}} = \left(SSR_{lib(i)}^{std} + \sum_{l=1}^{N_{sample}} \left(RT_{lib(i)}^{compound(l)} - RT_{sample(j)}^{compound(l)}\right)^2\right) \Big/ (N_{std} + N_{sample}),$$

where $RT_{lib(i)}^{compound(l)}$ is the retention time of the Compound l in $\mathrm{RTT_{lib(}}_{i)}$, and $RT_{sample(j)}^{compound(l)}$ is the retention time of a peak in the sample chromatogram that is hypothetically assigned to the same compound (i.e., Compound l) in $\mathrm{RTT_{sample(}}_{j)}$. Note that all $\mathrm{RTT_{sample}}$s have the same retention time for each peak, but each peak can hypothetically be paired with a different compound in different $\mathrm{RTT_{sample}}$s, which has been discussed previously (e.g., Figure 2A). The *MSR* of all $\mathrm{RTT_{lib}}$s are sorted in ascending order. The first in the list has the minimum *MSR* value denoted as $MSR_{sample(j),\,min}$, which is generated by the $\mathrm{RTT_{lib}}$ that best matches $\mathrm{RTT_{sample(}}_{j)}$.

Because each $\mathrm{RTT_{sample}}$ is formed by pairing the detected peaks with one set of target compounds, it represents one set of peak identification results. For any $\mathrm{RTT_{sample}}$, which we call $\mathrm{RTT_{sample(}}_{j)}$, the best-matched $\mathrm{RTT_{lib}}$ can be found by screening all $\mathrm{RTT_{lib}}$s in the library and finding the one that generates $MSR_{\,sample(j),\,\mathrm{min}}$. If there is another $\mathrm{RTT_{sample}}$, denoted as $\mathrm{RTT_{sample(}}_{k)}$, that has $MSR_{sample(k),min}$ smaller than $MSR_{sample(j),min}$, it means that $\mathrm{RTT_{sample(}}_{k)}$ is a better match with one of the $\mathrm{RTT_{lib}}$s in the library. Therefore, the corresponding peak identification results from $\mathrm{RTT_{sample(}}_{k)}$ are more accurate than those from $\mathrm{RTT_{sample(}}_{j)}$.

2.5. Case II: Sample Containing Interferents

In targeted analysis, interferents are the compounds not on the list of the target compounds and need to be filtered out (or reported). In our algorithm, two criteria are used to identify the presence of interferents. First, for a particular peak in the sample chromatogram, if none of the retention time values in the $\mathrm{RTT_{lib}}$s fall in the range of $\mathrm{RT_{peak}} \pm \Delta t$, this peak is identified as an interferent. The value of Δt can be chosen empirically and can be set higher than a typical retention time drift range. Second, for one particular pair of $\mathrm{RTT_{lib(}}_{i)}$ and $\mathrm{RTT_{sample(}}_{j)}$, and if the squared residual between one peak (for example, Peak *D* in Figure 3C) in $\mathrm{RTT_{sample(}}_{j)}$ and its paired compound (Compound 9 in Figure 3C) in $\mathrm{RTT_{lib(}}_{i)}$ is much larger (e.g., twice) than $MSR_{lib(i),sample(j)}$, it is highly likely that this peak is an interferent. The validity of this approach lies in the fact that all other coordinates formed by the detected peaks and their paired target compounds match well with $\mathrm{RTT_{lib(}}_{j)}$, except for the one formed by Peak *D* paired with Compound 9. A new $MSR_{lib(i),sample(j)}$ is then calculated by normalizing $SSR_{lib(i),sample(j)}$, which excludes residuals from all identified interferents, with $N_{std} + N_{sample} - N_{interf}$ (N_{interf} is the total number of identified interferents). Based on this, all possible peak identification results, with or without interferents, can be ranked by *MSR*s. The results with the smallest *MSR* give the highest confidence level.

3. Experimental Section

3.1. Chromatogram Generation

Our RTT approach is validated using nine chromatograms obtained with NovaTest P300 GC provided by Nanova Environmental, Inc., which is equipped with a 6 m long Rtx-VMS column (Restek, Bellefonte, PA, USA) and a microfluidic photoionization detector developed in-house [36]. The chromatograms were generated under the same nominal experimental setting (carrier gas: helium; flow rate: 3.5 mL/min; temperature programming profile: 40 °C held for 5 min, ramped to 70 °C at 30 °C/min, held for 2 min, then ramped to 150 °C at 30 °C/min, and held for 1 min). The injected mixture is part of EPA Method TO-14. Exemplary chromatograms are presented in Figure 4A, and corresponding peak information is listed in Table S1.

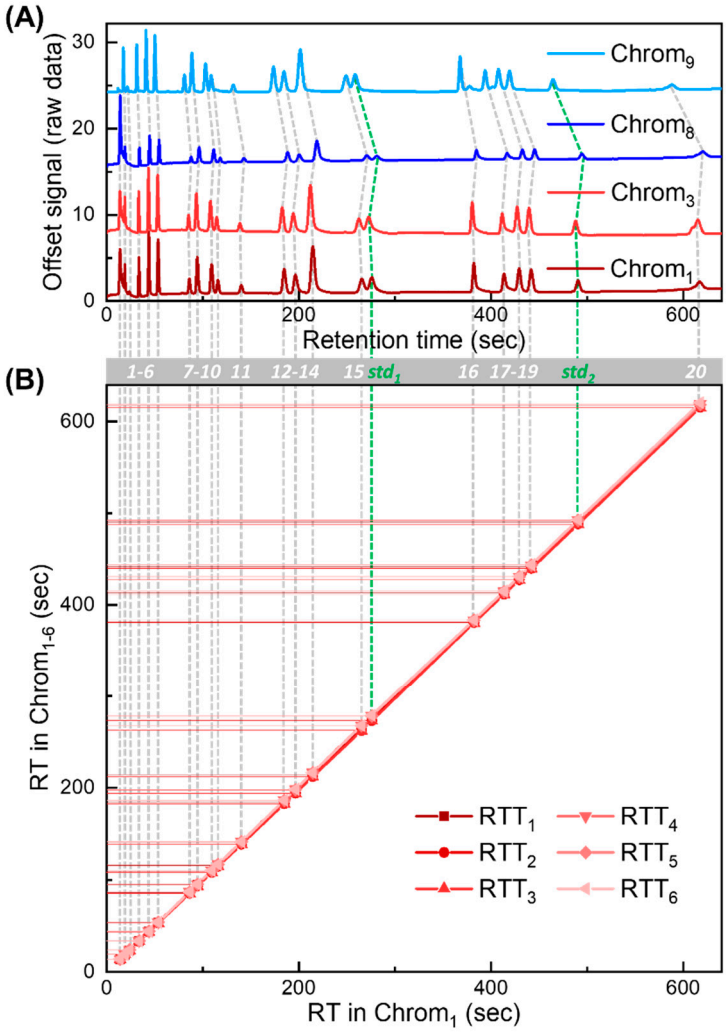

Figure 4. Experimentally generated chromatograms and corresponding RTT_{lib}s for algorithm validation. (**A**) Four exemplary chromatograms ($Chrom_1$, $Chrom_3$, $Chrom_8$, and $Chrom_9$) obtained experimentally. There are a total of 22 detected peaks in each chromatogram. A total of 2 out of the 22 peaks are treated as internal standards (labeled as std_1 and std_2 in green). The remaining peaks are treated as target compounds for peak identification in targeted analysis. Peaks of the same compounds are labeled with compound IDs from 1 to 20 (grey bar). (**B**) RTT_{lib} library formed by RTT_{lib}s ($RTT_{1–6}$) generated from $Chrom_{1–6}$. The *X*-axis represents retention time in $Chrom_1$. The *Y*-axis represents retention times in $Chrom_{1–6}$. The retention time deviation (ΔRT) of $Chrom_{1–9}$ against the RT in $Chrom_1$ is plotted in Figure S5.

3.2. Chromatogram Preprocessing

Detection of a peak in a chromatogram is accomplished by scanning for local maxima and the associated peak apex positions (i.e., retention times) [37]. A series of retention times are extracted, which are used to form the RTT_{lib}s or RTT_{sample}s in the next section. Therefore, the cumbersome chromatographic data (a large 2D array of detection signals) are converted to a simple list of retention times, which significantly reduces data storage and processing workload. Extensive preprocessing (e.g., baseline removal) and broad background variations can also be eliminated because only the local maxima (i.e., peak apexes) are extracted.

4. Validation Results

Out of nine experimentally generated chromatograms, six chromatograms (denoted as $Chrom_{1–6}$) are used in the library, forming RTT_{lib}s (Figure 4B). The remaining three (denoted as $Chrom_{7–9}$) are used to generate tests to validate our approach in various scenarios.

Detailed validation test design is described in Section S2 in the Supplementary Materials. There are a total of 22 peaks in each measured chromatogram, among which 20 are treated as target compounds and the other two are used as the internal standards. The retention times and compound IDs of $Chrom_1$ are summarized in Table S1. The RT deviation (ΔRT) against the RT in $Chrom_1$ for all chromatograms ($Chrom_{1–9}$) is plotted in Figure S5, showing strong non-linear drifting behavior. Note that while the chemical names for most compounds are provided by the vendor and further confirmed by injecting individual compounds, which are given in the third column in Table S1, the chemical identities of the first and the third eluted peaks are unknown (which might result from contamination and are only designated as ID 1 and ID 3, respectively). Nevertheless, the results presented in this work remain the same regardless of whether the chemical names of those compounds are known.

In total, $3 \times \sum_{i=1}^{20} [C(20, i)] = 3.15 \times 10^5$ validation tests were generated from $Chrom_{7–9}$, covering *all* subsets of the 20 target compounds (ranging from single compounds to 20 compounds). Moreover, three additional validation tests were generated, representing samples with a subset of target compounds *and* interferent(s). In all 3.15×10^5 validation tests, peak identifications achieved 100% accuracy. Here, we present detailed results of 11 representative tests, covering various MS-free chromatographic analysis scenarios, i.e., (1) different levels of retention time (RT) drift, (2) different numbers of the target compounds in the sample under test, and (3) samples containing interferents.

4.1. Identification of Target Compounds in Sample under Test without Interferents

To validate our algorithm, we first discuss the scenarios in which no interferents are present and all the detected peaks are a subset of the target compounds. As shown in Table 1, based on the RTs in $Chrom_7$ and $Chrom_8$, six groups of RTs are generated, three from $Chrom_7$ (Table 1A) and three from $Chrom_8$ (Table 1B), for six tests. They represent six different samples under test that contain various subsets of the target compounds (the number of the target compounds ranges from 1 to 13 out of a total of 20 target compounds), along with two internal standards (std_1 and std_2). Note that the retention time deviation in $Chrom_7$ is within the range of the RT deviations in the chromatograms ($Chrom_{1–6}$) stored in the library, whereas the retention time deviation in $Chrom_8$ is slightly out of this range (Figure S5). For each test, the four best peak identification results are enumerated based on the *MSR* ranking (Table 1). The peaks in all six tests are successfully identified with the top result (i.e., smallest *MSR*) producing 100% accuracy. The 2nd-4th best results in each test also correctly identify most of the peaks with the best ones giving 100% accuracy and the worst ones misidentifying only one peak. Note that even a single-species sample (Compound 7 in Test 6) can be correctly identified due to the use of internal standards, despite it being very close to neighboring Compound 8. This is impossible for all warping-based chromatogram-aligning approaches, because the peak has nothing to be aligned with.

Table 1. Algorithm experimental design and peak identification results with a sample containing only the target compounds. The peaks listed in Tests 1–3 are generated from $Chrom_7$; the peaks listed in Tests 4–6 are generated from $Chrom_8$. An asterisk "*" denotes peak misidentification.

(A)																	
Test data generated from $Chrom_7$		**Retention time (s)**			13.8	19	33.5	43.7	85.6	108.5	115	183.6	213.3		412.7		616
		Compound ID			1	2	4	5	7	9	10	12	14		17		20
	Test 1	**Ranking**	**MSR**	**Accuracy**	**Individual peak identification result**												
		1st	0.57	100%	1	2	4	5	7	9	10	12	14		17		20
		2nd	0.71	100%	1	2	4	5	7	9	10	12	14		17		20
		3rd	0.91	100%	1	2	4	5	7	9	10	12	14		17		20
		4th	2.2	90.9%	1	3 *	4	5	7	9	10	12	14		17		20
		Retention time (s)			19	33.5	43.7	85.6	93.8	108.5	115	139.3	195.2	213.3	264	381.3	412.7
		Compound ID			2	4	5	7	8	9	10	11	13	14	15	16	17
	Test 2	**Ranking**	**MSR**	**Accuracy**	**Individual peak identification result**												
		1st	0.67	100%	2	4	5	7	8	9	10	11	13	14	15	16	17
		2nd	0.74	100%	2	4	5	7	8	9	10	11	13	14	15	16	17
		3rd	1.14	100%	2	4	5	7	8	9	10	11	13	14	15	16	17
		4th	2.15	92.3%	3 *	4	5	7	8	9	10	11	13	14	15	16	17
		Retention time (s)			85.6		108.5		195.2			381.3			428.2		
		Compound ID			7		9		13			16			18		
	Test 3	**Ranking**	**MSR**	**Accuracy**	**Individual peak identification result**												
		1st	0.88	100%	7		9		13			16			18		
		2nd	0.94	100%	7		9		13			16			18		
		3rd	1.48	100%	7		9		13			16			18		
		4th	5.74	100%	7		9		13			16			18		
(B)																	
Test data generated from $Chrom_8$		**Retention time (s)**			34.1	44.6	96.2	111.4	142.9	188.4		218.6		384.8		432.6	
		Compound ID			4	5	8	9	11	12		14		16		18	
	Test 4	**Ranking**	**MSR**	**Accuracy**	**Individual peak identification result**												
		1st	2.42	100%	4	5	8	9	11	12		14		16		18	
		2nd	2.86	100%	4	5	8	9	11	12		14		16		18	
		3rd	3.34	100%	4	5	8	9	11	12		14		16		18	
		4th	5.06	88.9%	4	5	8	10 *	11	12		14		16		18	
		Retention Time (s)			87.9		111.4		218.6			384.8			432.6		
		Compound ID			7		9		14			16			18		
	Test 5	**Ranking**	**MSR**	**Accuracy**	**Individual peak identification result**												
		1st	2.99	100%	7		9		14			16			18		
		2nd	3.74	100%	7		9		14			16			18		
		3rd	4.28	100%	7		9		14			16			18		
		4th	7.14	80%	7		10 *		14			16			18		
		Retention time (s)			87.9												
		Compound ID			7												
	Test 6	**Ranking**	**MSR**	**Accuracy**	**Individual peak identification result**												
		1st	3.80	100%	7												
		2nd	5.10	100%	7												
		3rd	5.80	100%	7												
		4th	16.75	100%	7												

4.2. Identification of Target Compounds in Sample under Test with Interferents

Another three validation experiments (Tests 7–9 in Table S2) were generated based on the RTs in $Chrom_8$, which mimic scenarios in which both a subset of target analytes and interferents are present in the sample under test. In Tests 7 and 8, one hypothetical

interferent peak was added at 340 s (for Test 7) and 449 s (for Test 8). In particular, the interferent peak at 449 s was very close to the target Compound 19. In both cases, the top identification result successfully identified all target compounds and singled out interferent related peaks with 100% accuracy. In Test 9, two hypothetical interferent peaks were added at 62 s and 395 s, which were very close to target Compounds 6 and 16, respectively. Both interferents and all target compounds were correctly identified. Like all other RT-based peak identification methods discussed previously, the RTT-matching-based interferent identification works only when $RT_{interferent}$ is sufficiently different from those of target compounds. If $RT_{interferent}$ is the same as or extremely close to any target compound, the interferent cannot be identified. Additionally, the majority of the peaks in the sample under test should be the target compounds. If most peaks are interferents and only a few target compounds are present, the validity of our method may decrease, because the residuals of most peaks are very large. One way to circumvent these issues is to further enrich the RTT library, either experimentally or through the RTT hybridization discussed in Section 4.3. Introducing more internal standards may also be helpful.

It is also worth noting that the RTT matching approach is intended for the scenarios where RT drifts are caused by only minor fluctuations in experimental conditions, and, therefore, the elution order is expected to be preserved among the measurements. When the experimental condition varies drastically (e.g., major changes in the device settings or ambient temperatures), the elution order may switch. Therefore, a new RTT library needs to be constructed under the new experimental conditions to avoid misalignment/misidentification in the RTT matching approach.

The above two issues (i.e., serious co-elution and elution order change) have always been the bottleneck for all existing MS-free chromatogram-aligning algorithms. MS (and other spectroscopic methods such as infrared absorption spectroscopy and Raman spectroscopy) would potentially be needed for peak identification in these cases.

4.3. Identification of Target Compounds in Sample under Test with Severe RT Drifts

An ideal RTT library should contain all possible RTT_{lib}s that cover all possible drift-inducing conditions. If the library has only a limited number of RTT_{lib}s and when the RT drift in a sample chromatogram (or RT deviation) exceeds the RT deviations covered by the RTT_{lib}s, peak misidentification may occur, as exemplified by Tests 10 and 11 in Table S3. One method to expand the library is to experimentally generate as many RTT_{lib}s as possible by varying the experimental conditions around the nominal conditions. However, this is extremely labor-intensive and difficult to realize. Alternatively, new RTT_{lib}s can be generated and added to the library by linearly hybridizing existing experimentally obtained RTT_{lib}s. This method is valid because the retention time drift is caused by minor fluctuations of system's physical factors, and such a small perturbation in the retention time of one particular compound from one state (one RT) to another state (another RT) can be simplified as linear variation. RTT linear hybridization can be performed using two, three, or more existing experimentally obtained RTT_{lib}s, i.e.,

$$C_1 \times RT_{lib(a)}^{compound(i)} + C_2 \times RT_{lib(b)}^{compound(i)},$$

or

$$C_1 \times RT_{lib(a)}^{compound(i)} + C_2 \times RT_{lib(b)}^{compound(i)} + C_3 \times RT_{lib(c)}^{compound(i)},$$

where C_1, C_2, and C_3 are the linear coefficients, and $RT_{lib(a)}^{compound(i)}$ refers to the retention time for Compound i in one of the RTT_{lib}s (i.e., $RTT_{lib(a)}$). The hybridization can easily generate more RTT_{lib}s of the intermediate states that may be difficult to obtain experimentally (due to either time limitations and/or difficulty in realizing the exact experimental conditions), which significantly increases the tolerance to more severe RT drifts. Note that the RT variation of one particular compound from one state to another is simplified to be linear, although ΔRT against RT in one chromatogram is generally non-linear.

In the current work, in order to validate the hybridization method, we use two-RTT hybridization based on the following three formulas: (1) $\left(RT_{lib(a)}^{compound(i)} + RT_{lib(b)}^{compound(i)}\right)/2$, (2) $2RT_{lib(a)}^{compound(i)} - RT_{lib(b)}^{compound(i)}$, and (3) $2RT_{lib(b)}^{compound(i)} - RT_{lib(a)}^{compound(i)}$. These are used to enrich the RTT_{lib}s in the library (see Figure S6), where $RTT_{lib(a)}$ and $RTT_{lib(b)}$ are selected from the $Chrom_{1–6}$ that are experimentally generated. Two tests (Tests 10 and 11) are generated based on $Chrom_9$, which has much more severe RT drift than $Chrom_{7–8}$ and any pre-characterized chromatograms ($Chrom_{1–6}$) in the library (see Figure S5). As shown in Table S3, when the RTT library contained only the experimentally generated RTT_{lib}s, it failed to correctly identify all peaks in either Test 9 or 10. In contrast, with more RTT_{lib}s added through hybridization, all peaks were successfully identified with 100% accuracy. The accuracy of the 2nd–4th identification results in both tests were also increased.

4.4. Comparison with Other Chromatogram-Aligning Approaches

In order to compare peak identification performance with other chromatogram-aligning approaches, parts of the above validation tests are also performed with the whole-chromatogram-based COW [19,20] method and two peak-list-based aligning methods, namely the internal-standard-based linear warping approach and fast PTW [22]. Again, it is worth emphasizing that our RTT algorithm identifies peaks through matching instead of chromatogram aligning.

For a chromatogram-aligning method to work, we use $Chrom_1$ (or its corresponding peak list) as the reference for other chromatograms (i.e., sample chromatograms) to align with. Tests 5 and 7 (based on $Chrom_8$), which represent samples without and with an interferent, respectively, are used to evaluate COW and linear warping. The retention times after COW or linear warping [35] are summarized in Table S4 and Table S5, respectively. Note that in the reference chromatogram (i.e., $Chrom_1$), all the peaks are present, including *all* target compounds and internal standards. The sample chromatograms, which are to be aligned and identified, are reconstructed from $Chrom_8$, but contain only the peaks listed in Tests 5 and 7. The remaining peaks in $Chrom_8$ are replaced with the baseline (see Section S3 in the Supplementary Materials for details of sample chromatogram reconstruction).

The COW aligning results can be highly parameter-dependent [19]. When the peak compositions in the reference and sample chromatograms are the same, optimal parameters can be easily chosen so that the apex positions of the peaks of the same elution order in the sample chromatogram and the reference chromatogram coincide. However, in the presence of only a subset of the target compounds in the sample under test (as in Tests 5 and 7), the peak in the sample chromatogram has no specific target peak to align to, and, therefore, the COW alignment parameter selection becomes dubious. Similarly, the interferent peak (as in Test 7) might be misaligned to one of the target compound peaks in the reference chromatogram. Multiple COW alignments with various tuning parameters (slack and correlation power) were conducted. The identification results based on the RTs extracted from the aligned $Chrom_{sample}$ are summarized in Table S4. Out of the seven peaks in Test 5, the best COW aligning correctly identifies only five peaks (slack = 4 and correlation power = 3) and the worst aligning fails to align any of the peaks (slack = 1 and correlation power = 1; slack = 2 and correlation power = 2). For Test 7, the highest identification accuracy reaches only 50% (slack = 4 and correlation power = 3) and the worst aligning fails with the whole chromatogram (slack = 1 and correlation power = 1; slack = 2 and correlation power = 2).

For the internal-standard-based linear warping method (Table S5), the same internal standards (std_1 and std_2) were employed. After alignment, none of the target compound-associated peaks yields the same RT as in the reference chromatogram ($Chrom_1$). Therefore, peak identification for target compounds fails when RT values are compared (all peaks are identified as interferents, except the interferent peak itself).

Additionally, we compared the peak identification performance of the fast PTW algorithm [22], which is a further development of the original PTW [21], and the algorithm

input was the chromatogram peak list (retention times and peak heights). The peak lists in $Chrom_1$, which was treated as the reference chromatogram, and $Chrom_{8,9}$, which were used to generate Tests 5, 7, 10, and 11, are summarized in Table S6A. The time warping results (order = 2 or 3) using the full peak lists in $Chrom_{8,9}$ are summarized in Table S6A. Although misalignments were greatly reduced after fast PTW aligning, sub-second to seconds of RT difference for the same compound persists. Table S6B summarizes the fast PTW aligning results (order = 2 or 3) of Tests 5, 7, 10, and 11. Tests 5, 10, and 11 represent samples with only a subset of target compounds; Test 7 represents a sample with a subset of target compounds and an interferent. The fast PTW fails to identify the correct peaks in all four tests. In comparison with the reference chromatogram, the RT differences in the same compound in the warped sample peak list fell in the range of sub-seconds (which could be solved by data clustering) to hundreds of seconds (which completely fails in peak identification). We also noticed that the warped RTs of some peaks gave unreasonable negative values (e.g., Compound 1 in Test 10, with order = 2, and Compound 17 in Test 10, with order = 3).

4.5. Further Validation with Fruit Metabolomics Data

Two publicly available liquid chromatography (LC) fruit metabolomics datasets from the Metabolights repository (http://www.ebi.ac.uk/metabolights, accessed on 10 March 2021) with identifiers MTBLS99 and MTBLS85 are used to generate additional validation tests. This demonstrates the application of the RTT matching algorithm to complicated samples and other chromatographic techniques. The first dataset consists of LC measurements of a pooled sample that was injected regularly as a quality control during the measurements of apple extracts (Figure S10A). The second dataset is from LC measurements of carotenoids in grape samples (Figure S10B). The test design and results are summarized in Section S4 in the Supplementary Materials. In all of the two trillion designed validation tests, all peaks associated with the target compounds were correctly identified.

5. Discussion and Conclusions

We have described in detail and validated a new RTT matching approach for the identification of target compounds and the detection of interferents. This approach has the following advantages. First and foremost, the matching is conducted between the entire trajectories (RTT_{sample} and RTT_{lib}) rather than between the individual peaks in the sample and those in the reference chromatogram. Second, simple statistics are used, which avoids time-consuming training or feature extraction used in machine learning [28,31,32]. The *MSR* is adopted to describe the similarities between RTTs, which works well with the chromatograms obtained in this article. Other statistical approaches, such as linear regression, can also be introduced within the framework of RTT matching. Third, hybridization of RTT_{lib}s greatly expands the RTT library, which not only increases the tolerance to more serious drifting but also significantly reduces the cost for RTT_{lib} generation through actual experimentation. In this work, each hybridized RTT_{lib} is generated out of two experimentally obtained RTT_{lib}s. Depending on the complexities of the subjects to be analyzed, a larger number of experimentally obtained RTT_{lib}s could be used, and the linear coefficients could be further tuned. Fourth, the only input variables are the retention times of each peak instead of the whole peak profile or mass spectra, making the RTT approach insensitive to concentration or background variations and eliminating the need for bulky instruments, such as mass spectroscopy. All the above features make RTT matching highly amenable to automation with a low computation cost. Additionally, the method described here can be easily translated to other chromatographic techniques (e.g., ion exchange chromatography) and has great potential to be applied to other spectral data (e.g., nuclear magnetic resonance spectroscopy). Finally, the introduction of internal standards, though contributing to increasing identification accuracy and computation efficiency, are not always necessary. When an increased number of target compounds are present in the sample under test, the contribution of the standard(s) becomes less prominent. To validate this, 10 of the

validation test examples in the above discussions, except the single-species sample in Test 6, have been re-performed with RTT matching without the use of any internal standards. All of the peaks were correctly identified. Our current work is focused on the RTT matching for one-dimensional chromatography. However, the same idea and approach can be extended to multi-dimensional chromatography, such as comprehensive 2D GC.

Finally, it is worth noting that the applications of the RTT matching approach are not limited to just peak identification. The RTT matching approach can help in chromatogram aligning. Briefly, one can choose any RTT_{lib} as the reference chromatogram to extract RTs of the target compounds. Each peak in the sample chromatogram is identified via the RTT matching approach. The peak profile can be fitted by the exponentially modified Gaussian (EMG) model, with apexes shifted to the positions as in the reference. Aligned sample chromatograms can be formed by the summation of individual EMG reconstructed peaks. A detailed description and an illustration are discussed in Section S5 and Figure S11.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/s23136029/s1 [19,20,22,37–40], Figure S1: Conceptual illustration of the binning approach. Figure S2: Rules to eliminate impossible RTT_{sample}s. Figure S3: Use of internal standards. Figure S4: Conceptual illustration of peak aligning with the internal-standard-based linear stretching/compressing approach using the RTT 2D diagram. Figure S5: Retention time deviation (ΔRT) of $Chrom_{1–9}$ against the RT in $Chrom_1$. Figure S6: Demonstration of RTT_{lib}s hybridization. Figure S7: Illustration of algorithm validation test design using $Chrom_8$. Figure S8: Peak identification using COW aligning with a sample containing a subset of target compounds (Test 5). Figure S9: Peak identification using COW aligning with a sample containing a subset of target compounds plus one interferent (Test 7). Figure S10: Fruit metabolomics chromatograms for RTT peak identification verification. Figure S11: Chromatogram aligning enabled by RTT matching. Table S1: Peak retention times, assigned compound IDs, and compound names in $Chrom_1$. Table S2: Algorithm validation tests and the corresponding peak identification results. Table S3: Algorithm validation tests and the corresponding peak identification results when a sample chromatogram has severe RT drift issues. Table S4: Peak identification performance comparison with COW. Table S5: Peak identification performance comparison with internal-standard-based linear warping. Table S6: Peak identification performance comparison with fast PTW.

Author Contributions: Conceptualization, W.Z. and X.F.; methodology, W.Z. and X.F.; software, W.Z.; validation, W.Z.; formal analysis, W.Z. and X.F.; resources, X.F.; data curation, W.Z., R.S., M.W.-H.L. and X.F.; writing—original draft preparation, W.Z. and X.F.; writing—review and editing, W.Z., R.S., M.W.-H.L. and X.F.; supervision, X.F.; project administration, X.F.; funding acquisition, X.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Institute for Occupational Safety and Health (NIOSH) via grant R01 OH011082-01A1 and the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA) via contract FA8650-19-C-9101. The APC was funded by the University of Michigan Richard A Auhll Professorship.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors acknowledge the support from the National Institute for Occupational Safety and Health (NIOSH) via grant R01 OH011082-01A1, the Beijing Institute for Collaborative Innovation via the University of Michigan internal funding, and the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA) via contract FA8650-19-C-9101. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The authors acknowledge NGK Spark Plugs for the gift fund and Nanova Environmental, Inc. for providing some of the chromatograms.

Conflicts of Interest: Xudong Fan is an inventor of the microfluidic photoionization detector (PID) that is licensed to Nanova Environmental Inc. The PID was used in a Nanova portable GC system to generate the exemplary chromatograms analyzed and presented in this article.

References

1. Fischer, M.; Scholz-Böttcher, B.M. Microplastics Analysis in Environmental Samples-Recent Pyrolysis-Gas Chromatography-Mass Spectrometry Method Improvements to Increase the Reliability of Mass-Related Data. *Anal. Methods* **2019**, *11*, 2489–2497. [CrossRef]
2. Santos, F.J.; Galceran, M.T. Modern Developments in Gas Chromatography-Mass Spectrometry-Based Environmental Analysis. *J. Chromatogr. A* **2003**, *1000*, 125–151. [CrossRef]
3. Moularat, S.; Robine, E.; Ramalho, O.; Oturan, M.A. Detection of Fungal Development in a Closed Environment through the Identification of Specific VOC: Demonstration of a Specific VOC Fingerprint for Fungal Development. *Sci. Total Environ.* **2008**, *407*, 139–146. [CrossRef]
4. Leidinger, M.; Sauerwald, T.; Reimringer, W.; Ventura, G.; Schütze, A. Selective Detection of Hazardous VOCs for Indoor Air Quality Applications Using a Virtual Gas Sensor Array. *J. Sens. Sens. Syst.* **2014**, *3*, 253–263. [CrossRef]
5. Szulczyński, B.; Gębicki, J. Currently Commercially Available Chemical Sensors Employed for Detection of Volatile Organic Compounds in Outdoor and Indoor Air. *Environments* **2017**, *4*, 21. [CrossRef]
6. Wu, C.H.; Feng, C.T.; Lo, Y.S.; Lin, T.Y.; Lo, J.G. Determination of Volatile Organic Compounds in Workplace Air by Multisorbent Adsorption/Thermal Desorption-GC/MS. *Chemosphere* **2004**, *56*, 71–80. [CrossRef] [PubMed]
7. Campos-Candel, A.; Llobat-Estellés, M.; Mauri-Aucejo, A. Comparative Evaluation of Liquid Chromatography versus Gas Chromatography Using a β-Cyclodextrin Stationary Phase for the Determination of BTEX in Occupational Environments. *Talanta* **2009**, *78*, 1286–1292. [CrossRef] [PubMed]
8. Wu, C.H.; Lin, M.N.; Feng, C.T.; Yang, K.L.; Lo, Y.S.; Lo, J.G. Measurement of Toxic Volatile Organic Compounds in Indoor Air of Semiconductor Foundries Using Multisorbent Adsorption/Thermal Desorption Coupled with Gas Chromatography-Mass Spectrometry. *J. Chromatogr. A* **2003**, *996*, 225–231. [CrossRef] [PubMed]
9. Nizio, K.D.; McGinitie, T.M.; Harynuk, J.J. Comprehensive Multidimensional Separations for the Analysis of Petroleum. *J. Chromatogr. A* **2012**, *1255*, 12–23. [CrossRef]
10. Mahé, L.; Dutriez, T.; Courtiade, M.; Thiébaut, D.; Dulot, H.; Bertoncini, F. Global Approach for the Selection of High Temperature Comprehensive Two-Dimensional Gas Chromatography Experimental Conditions and Quantitative Analysis in Regards to Sulfur-Containing Compounds in Heavy Petroleum Cuts. *J. Chromatogr. A* **2011**, *1218*, 534–544. [CrossRef]
11. Periago, J.F.; Zambudio, A.; Prado, C. Evaluation of Environmental Levels of Aromatic Hydrocarbons in Gasoline Service Stations by Gas Chromatography. *J. Chromatogr. A* **1997**, *778*, 263–268. [CrossRef] [PubMed]
12. García-Reyes, J.F.; Hernando, M.D.; Molina-Díaz, A.; Fernández-Alba, A.R. Comprehensive Screening of Target, Non-Target and Unknown Pesticides in Food by LC-TOF-MS. *TrAC-Trends Anal. Chem.* **2007**, *26*, 828–841. [CrossRef]
13. Rissato, S.R.; Galhiane, M.S.; de Almeida, M.V.; Gerenutti, M.; Apon, B.M. Multiresidue Determination of Pesticides in Honey Samples by Gas Chromatography-Mass Spectrometry and Application in Environmental Contamination. *Food Chem.* **2007**, *101*, 1719–1726. [CrossRef]
14. Saasa, V.; Beukes, M.; Lemmer, Y.; Mwakikunga, B. Blood Ketone Bodies and Breath Acetone Analysis and Their Correlations in Type 2 Diabetes Mellitus. *Diagnostics* **2019**, *9*, 224. [CrossRef] [PubMed]
15. Buszewski, B.; Ligor, T.; Jezierski, T.; Wenda-Piesik, A.; Walczak, M.; Rudnicka, J. Identification of Volatile Lung Cancer Markers by Gas Chromatography-Mass Spectrometry: Comparison with Discrimination by Canines. Anal. *Bioanal. Chem.* **2012**, *404*, 141–146. [CrossRef]
16. Roach, C.R.; Hall, D.E.; Zerbe, P.; Bohlmann, J. Plasticity and Evolution of (+)-3-Carene Synthase and (-)-Sabinene Synthase Functions of a Sitka Spruce Monoterpene Synthase Gene Family Associated with Weevil Resistance. *J. Biol. Chem.* **2014**, *289*, 23859–23869. [CrossRef]
17. Vu, T.N.; Laukens, K. Getting Your Peaks in Line: A Review of Alignment Methods for NMR Spectral Data. *Metabolites* **2013**, *3*, 259–276. [CrossRef]
18. Sudol, P.E.; Gough, D.V.; Prebihalo, S.E.; Synovec, R.E. Impact of data bin size on the classification of diesel fuels using comprehensive two-dimensional gas chromatography with principal component analysis. *Talanta* **2020**, *206*, 120239. [CrossRef]
19. Nielsen, N.P.V.; Carstensen, J.M.; Smedsgaard, J. Aligning of Single and Multiple Wavelength Chromatographic Profiles for Chemometric Data Analysis Using Correlation Optimised Warping. *J. Chromatogr. A* **1998**, *805*, 17–35. [CrossRef]
20. Tomasi, G.; Van Den Berg, F.; Andersson, C. Correlation Optimized Warping and Dynamic Time Warping as Preprocessing Methods for Chromatographic Data. *J. Chemom.* **2004**, *18*, 231–241. [CrossRef]
21. Eilers, P.H.C. Parametric Time Warping. *Anal. Chem.* **2004**, *76*, 404–411. [CrossRef]
22. Wehrens, R.; Bloemberg, T.G.; Eilers, P.H.C. Fast Parametric Time Warping of Peak Lists. *Bioinformatics* **2015**, *31*, 3063–3065. [CrossRef]
23. Zhang, Z.M.; Liang, Y.Z.; Lu, H.M.; Tan, B.B.; Xu, X.N.; Ferro, M. Multiscale Peak Alignment for Chromatographic Datasets. *J. Chromatogr. A* **2012**, *1223*, 93–106. [CrossRef]

24. Zheng, Q.X.; Fu, H.Y.; Li, H.D.; Wang, B.; Peng, C.H.; Wang, S.; Cai, J.L.; Liu, S.F.; Zhang, X.B.; Yu, Y.J. Automatic Time-Shift Alignment Method for Chromatographic Data Analysis. *Sci. Rep.* **2017**, *7*, 256. [CrossRef] [PubMed]
25. Clifford, D.; Stone, G.; Montoliu, I.; Rezzi, S.; Martin, F.P.; Guy, P.; Bruce, S.; Kochhar, S. Alignment Using Variable Penalty Dynamic Time Warping. *Anal. Chem.* **2009**, *81*, 1000–1007. [CrossRef]
26. Halvorsen, R.C.; Trinklein, T.J.; Warren, C.G.; Rogan, R.D.; Synovec, R.E. Optimizing column-to-column retention time alignment in high-speed gas chromatography by combining retention time locking and correlation optimized warping. *Talanta* **2023**, *254*, 124173. [CrossRef] [PubMed]
27. Kumar, K. Chemometric assisted correlation optimized warping of chromatograms: Optimizing the computational time for correcting the drifts in chromatographic peak positions. *Anal. Methods* **2018**, *10*, 1006–1014. [CrossRef]
28. Li, M.; Wang, X.R. Peak Alignment of Gas Chromatography–Mass Spectrometry Data with Deep Learning. *J. Chromatogr. A* **2019**, *1604*, 460476. [CrossRef]
29. Smolinska, A.; Hauschild, A.C.; Fijten, R.R.R.; Dallinga, J.W.; Baumbach, J.; Van Schooten, F.J. Current Breathomics—A Review on Data Pre-Processing Techniques and Machine Learning in Metabolomics Breath Analysis. *J. Breath Res.* **2014**, *8*, 027105. [CrossRef]
30. Lebanov, L.; Tedone, L.; Ghiasvand, A.; Paull, B. Random Forests Machine Learning Applied to Gas Chromatography—Mass Spectrometry Derived Average Mass Spectrum Data Sets for Classification and Characterisation of Essential Oils. *Talanta* **2020**, *208*, 120471. [CrossRef]
31. Yang, Y. Chromatogram Alignment Algorithm Based on Deep Neural Network and an Application in Bio-aerosol Detection. In Proceedings of the 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), Miami, FL, USA, 14–17 December 2020; pp. 1081–1086.
32. Cao, L.; Zang, W.; Sharma, R.; Tabartehfarahani, A.; Thota, C.; Sivakumar, A.D.; Lam, A.; Fan, X.; Ward, K.R.; Ansari, S. Automated Gas Chromatography Peak Alignment: A Deep Learning Approach using Greedy Optimization and Simulation. In Proceedings of the 45th Annual Conference of IEEE Engineering in Medicine and Biology Society, Sydney, Australia, 24–28 July 2023.
33. Smith, C.A.; Want, E.J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Anal. Chem.* **2006**, *78*, 779–787. [CrossRef]
34. Broeckling, C.D.; Reddy, I.R.; Duran, A.L.; Zhao, X.; Sumner, L.W. MET-IDEA: Data Extraction Tool for Mass Spectrometry-Based Metabolomics. *Anal. Chem.* **2006**, *78*, 4334–4341. [CrossRef] [PubMed]
35. Frenzel, T.; Miller, A.; Engel, K.H. A Methodology for Automated Comparative Analysis of Metabolite Profiling Data. *Eur. Food Res. Technol.* **2003**, *216*, 335–342. [CrossRef]
36. Zhu, H.; Nidetz, R.; Zhou, M.; Lee, J.; Buggaveeti, S.; Kurabayashi, K.; Fan, X. Flow-through Microfluidic Photoionization Detectors for Rapid and Highly Sensitive Vapor Detection. *Lab Chip* **2015**, *15*, 3021–3029. [CrossRef] [PubMed]
37. Morris, J.S.; Coombes, K.R.; Koomen, J.; Baggerly, K.A.; Kobayashi, R. Feature Extraction and Quantification for Mass Spectrometry in Biomedical Applications Using the Mean Spectrum. *Bioinformatics* **2005**, *21*, 1764–1775. [CrossRef]
38. Zhang, Z.M.; Chen, S.; Liang, Y.Z. Baseline Correction Using Adaptive Iteratively Reweighted Penalized Least Squares. *Analyst* **2010**, *135*, 1138–1146. [CrossRef]
39. Cleveland, W.S.; Devlin, S.J. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *J. Am. Stat. Assoc.* **1988**, *83*, 596–610. [CrossRef]
40. Lee, J.; Zhou, M.; Zhu, H.; Nidetz, R.; Kurabayashi, K.; Fan, X. Fully Automated Portable Comprehensive 2-Dimensional Gas Chromatography Device. *Anal. Chem.* **2016**, *88*, 10266–10274. [CrossRef]