Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

J Chromatogr A. Author manuscript; available in PMC 2016 September 18.

Published in final edited form as:

J Chromatogr A. 2015 September 18; 1412: 52–58. doi:10.1016/j.chroma.2015.07.113.

Calculation of Retention Time Tolerance Windows with Absolute Confidence from Shared Liquid Chromatographic Retention Data

Paul G. Boswell[a,*], **Daniel Abate-Pella**[a], and **Joshua T. Hewitt**[a]

Daniel Abate-Pella: abate006@umn.edu; Joshua T. Hewitt: hewi0041@umn.edu

[a]Department of Horticultural Science, University of Minnesota, 1970 Folwell Avenue, St. Paul, Minnesota, 55108

Abstract

Compound identification by liquid chromatography-mass spectrometry (LC-MS) is a tedious process, mainly because authentic standards must be run on a user's system to be able to confidently reject a potential identity from its retention time and mass spectral properties. Instead, it would be preferable to use *shared* retention time/index data to narrow down the identity, but shared data cannot be used to reject candidates with an absolute level of confidence because the data are strongly affected by differences between HPLC systems and experimental conditions. However, a technique called "retention projection" was recently shown to account for many of the differences. In this manuscript, we discuss an approach to calculate appropriate retention time tolerance windows for projected retention times, potentially making it possible to exclude candidates with an absolute level of confidence, *without needing to have authentic standards of each candidate on hand*. In a range of multi-segment gradients and flow rates run among seven different labs, the new approach calculated tolerance windows that were significantly more appropriate for each retention projection than global tolerance windows calculated for retention projections or linear retention indices. Though there were still some small differences between the labs that evidently were not taken into account, the calculated tolerance windows only needed to be relaxed by 50% to make them appropriate for all labs. Even then, 42% of the tolerance windows calculated in this study *without standards* were narrower than those required by WADA for positive identification, where standards must be run contemporaneously.

Keywords

Retention projection; Retention prediction; Liquid Chromatography-Mass Spectrometry; Retention database; Retention time tolerance window

*Corresponding author: Paul G. Boswell, 328 Alderman Hall, 1970 Folwell Ave, St. Paul, MN 55108, boswell@umn.edu, Phone: 612-250-5188.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

1. Introduction

The confident identification of known small molecules from complex mixtures continues to be a major challenge. LC-MS has become a principle tool for this purpose, but the confident identification of a chromatographic peak, or “feature”, is still a tedious process. A lab must acquire authentic standards of each potential identity and run them on their LC-MS system to compare the mass spectral and retention properties with that of the unidentified feature [1,2]. Those that do not match are rejected, ideally leaving only one possible identity. Improvements to LC-MS equipment have steadily improved identification rates, but the requirement to have authentic standards on hand is debilitating. When the number of features to be identified is large or the number of potential identities for a set of features is large, it quickly becomes impractical to have standards on hand for all potential identities. The analytical community is in great need of ways to confidently identify small molecules by LC-MS *without* requiring labs to purchase massive libraries of authentic standards.

When authentic standards are not on hand, one is currently forced to rely on exact mass, isotope distribution, libraries of MS/MS spectra, and libraries of retention information. Of these, exact mass and isotope distribution are the only properties that make it possible to reject candidates with an absolute level of confidence [3,4]. Unfortunately, they only narrow down the elemental composition, not the chemical structure; even if the elemental composition is narrowed down to one, there are usually still a large number of candidates with the same elemental composition [5].

Libraries of MS/MS information are usually the next pieces of information used to narrow down the list of potential identities. The MS/MS information is often quite useful, but unfortunately there is no way to use shared MS/MS spectra to assign an absolute level of confidence that a candidate should be rejected. (When using libraries, candidates may only be rejected, not identified, because the true identity may not be in the database.) Spectral matching algorithms assign a score of some kind [6–10], but none of them offer an absolute probability that a candidate is incorrect [11]. This is because product ion spectra collected from different makes and models of MS differ from one another [12–15], so if a measured product ion spectrum differs from that in a library, it is unclear what portion of the differences come from instrument differences and what portion come from actual differences in chemical structure. Ultimately one is left to make a decision whether to reject a potential identity without an absolute probability that the choice to do so is correct.

Libraries of LC retention information are the very last pieces of information one might turn to since, currently, they offer very little help. This is not because retention information is useless – on the contrary, it is extremely valuable for compound identification. LC retention is highly distinguishing and almost entirely unrelated to the information offered by MS and MS/MS [16]. The problem is that it is notoriously irreproducible across labs. It depends on a host of experimental factors, some that are intentionally different between labs (e.g., the gradient, flow rate, column dimensions, etc.) and some that are unintentionally different (e.g., gradient delay, gradient dispersion, solvent misproportioning, etc.). The preceding manuscript [17] explains the inherent problems with existing methods of sharing retention data. In particular, it shows how linear retention indexing (LRI) fails because it assumes that

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the relative retention of compounds is always the same. Figure 1 shows a prime example. It shows chromatograms collected by two different labs (labs A and B of the preceding manuscript), each running precisely the same experimental conditions, but with different makes/models of HPLC systems. In this case, just the unintentional differences in the gradients produced by the two HPLC systems were enough to cause a swap in elution order. In lab A, *N*-allyl aniline eluted *before* 7-amino-4-methylcoumarin, but in lab B, it eluted *after*. Therefore, even when one attempts to rigorously reproduce the conditions used by a different lab to measure a set of retention data, it is still unreliable due to unintentional differences in the experimental conditions.

Overall, two big problems limit the value of retention databases (e.g., linear retention index databases) as a means to share retention information: (1) they are often inaccurate since they cannot properly account for differences in experimental conditions (either intentional or unintentional), and (2) the level of error to expect in LRIs is unpredictable, making it very difficult to use them to reject a candidate on solid statistical grounds. Therefore, it is always unclear how far away a retention time in a library needs to be from the measured retention time to confidently exclude it as a potential identity.

≥1 $\delta\,\delta = \delta t + =$ In the preceding manuscript [17], we showed that a methodology called “retention projection” potentially solves the first problem, effectively accounting for the major intentional and unintentional differences between seven labs. The improved accuracy is welcome, but it is of limited value unless an absolute level of confidence can be assigned to reject a candidate based on comparison with a projected retention time. If the accuracies of RPs are lab- and/or method-dependent in an unpredictable manner, an absolute level of confidence cannot be assigned. On the other hand, if the RP methodology truly accounts for the majority of unintentional differences (between labs) and intentional differences (between methods), it may be possible to calculate the error one should expect in RPs with an absolute level of confidence. Potential compound identities could then be rejected on solid statistical grounds by comparison of observed and projected retention times without the need to acquire and run standards of each candidate, thereby drastically increasing the value of retention information for compound identification.

In this manuscript, we compare two approaches to calculate the level of error to expect in RPs (i.e., to predict the appropriate width of retention time tolerance windows), using data from the seven-lab study [17]. The first approach uses a single, *global* value of error and assumes it is the same regardless of the lab running the sample or the experimental conditions they use (i.e., the gradient and the flow rate). We also compare this to the use of a global value of error for LRIs. The second approach attempts to calculate an appropriate level of error *specific to each RP* assuming the major source of error is lab- and method-independent. We describe how these values are calculated and test their accuracy. Finally, we describe a system suitability check to ensure that a user’s system is in a state such that the calculated tolerance windows can be trusted.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

2. Experimental

2.1 Multi-Laboratory Study

The data used in this work comes from the preceding manuscript [17] describing a multi-laboratory study that tested the ability of the retention projection methodology to account for differences between labs. See ref [17] for experimental details not discussed here. A single test mixture was used throughout the study, which is described in the preceding manuscript. Of the 19 test compounds, the two fully charged compounds (tetrabutylammonium and tetrapentylammonium) were left out of this study. In many cases, months passed between the times the columns were shipped to the labs and when the labs ran the samples. During that time, the column selectivity drifted, disproportionately affecting retention of charged compounds (see [17]). Here we focus on the other 17 test compounds that were unaffected by the change in selectivity.

2.2 Software

The HPLC retention projection software was compiled for compliance with the Java 1.6 (Oracle, Redwood Shores, CA) runtime environment. It includes the Java OpenGL (JOGL) binding library version 2.0-rc11 (JogAmp, http://jogamp.org), the Unidata netCDF library version 4.2 (Unidata®, Boulder, CO), the Savitzky-Golay filter library version 1.2 by Marcin Rze nicki (http://code.google.com/p/savitzky-golay-filter/), the jmzML library, and the jmzReader library. The source code may be downloaded from http://www.retentionprediction.org/hplc/development.

3. Results/Discussion

3.1 Two Approaches Used to Calculate Retention Time Tolerance Windows (i.e., $\sigma_{t_R,expected}$)

The goal of this manuscript is to develop/compare models to predict the appropriate retention time tolerance window for retention time predictions by LRI or RP at a defined, absolute level of confidence. (In the following, notice the distinction between the terms “prediction” and “projection”. Retention “projection” and LRI are the two forms of retention “prediction” we discuss in this work.) Stated another way, the goal here is not to predict retention times, but to predict how accurate each predicted retention time will be. This is important because if the accuracy of a predicted retention time can be predicted, it will be possible to use the difference between an unidentified peak’s measured retention time and a candidate’s predicted retention time to exclude it as a potential identity on solid statistical grounds, without having to run an authentic standard of it.

We begin by defining two variables: $\sigma_{t_R,expected}$ and $\sigma_{t_R,observed}$. The variable $\sigma_{t_R,observed}$ describes the observed standard deviation of the error in retention time predictions (i.e., the standard deviation of the difference between measured retention times and predicted retention times). It is the quantity we would like to predict in this work. The variable $\sigma_{t_R,expected}$ is that prediction, that is, it is the standard deviation of the error in retention time predictions we expect (predict) using the models/approaches described below. Retention

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

time tolerance windows can then be calculated from $\sigma_{t_R,expected}$ at a specific level of confidence (e.g., the 99% confidence interval ≈ ±2.58 $\sigma_{t_R,expected}$).

3.1.1 The First Approach Calculates a Global $\sigma_{t_R,expected}$ for all Retention Predictions—Two different approaches to calculate $\sigma_{t_R,expected}$ were investigated. For the first, we calculated a single, *global* value of $\sigma_{t_R,expected}$ for RPs and another $\sigma_{t_R,expected}$ for LRIs. This simple approach implicitly assumes that the accuracy of RPs and LRIs are completely lab- and method-independent, such that a single retention time tolerance window is appropriate for all conditions. We calculated one for all RPs and one for all LRIs. These values, which from here on we call "*global* $\sigma_{t_R,expected}$", were calculated from the errors in a total of 946 retention predictions (7 labs, each running up to 10 methods, with up to 17 test compounds in each run). We calculated the global $\sigma_{t_R,expected}$ for retention times predicted from LRIs to be ±0.81 min and the global $\sigma_{t_R,expected}$ for RPs to be ±0.032 min. The *global* $\sigma_{t_R,expected}$ for RPs was 25-fold smaller than that for LRIs owing to the ability of the retention projection methodology to account for many of the differences between the labs and methods.

3.1.2 The Second Approach Calculates $\sigma_{t_R,expected}$ Values Specific to Each RP—The second approach to calculate $\sigma_{t_R,expected}$ was more sophisticated and could only be applied to RPs. We observed that the level of accuracy in RPs consistently depended on the experimental conditions. For example, we observed that in long, shallow gradients, error in RPs was greater than in short, steep gradients. This suggests that a global tolerance window, like that described above, is not appropriate for all compounds in all methods. Therefore, we developed a way to calculate values of $\sigma_{t_R,expected}$ for each individual RP (from here on we refer to these as "*specific* $\sigma_{t_R,expected}$") by attempting to take into account the experimental factors that significantly influence error in RPs. We began with two important assumptions: First, we assumed that the RP methodology accounts for the vast majority of experimental factors affecting retention (e.g., that it almost perfectly accounts for the actual gradient produced by the HPLC in each run), such that any unaccounted for factors have a negligible effect on the accuracy of RPs. This assumption leaves only one remaining source of error: the measured k values in the k vs. Φ relationships. Our second assumption was that the relative error in these measured k values was always the same (regardless of the solvent composition they were measured at). We directly measured the RSD in the repeatability of our measurement of k to be approximately 1% (for k values greater than 0.5), but we used an RSD of 3% in the following calculations because it yielded $\sigma_{t_R,expected}$ values that better matched the error observed in RPs.

Therefore, in order to calculate *specific* $\sigma_{t_R,expected}$ values, we propagated the error in k (which, as mentioned above, we assume to be the only source of error) to determine the uncertainty it imparts to each projected retention time. We propagated the error in k by summing uncertainty in the axial position of a compound along the column, σ_x, after each step of numerical integration. (Note that x is unitless; x = 0 at the column inlet and x = 1 at the column outlet.) After propagation of error (and a 3% RSD in k), the relative error in the position of the compound added after each isocratic step $\left(\frac{\delta\sigma_x}{\delta x}\right)$ is determined by:

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\frac{\delta\sigma_x}{\delta x}=\frac{0.03k}{1+k} \quad (1)$$

To determine the amount of uncertainty added to the position of the compound after each step, it must be weighted by the fraction of the column traveled in each step. Recalling eq. 2 in the preceding manuscript, the fraction that a compound moves along the column during each time slice is $\delta t_c/t_{0,\Phi_i\Phi}\,(1 + k_{\Phi i\Phi})$ where t_{0,ϕ_i} is the dead time at the Φ of the isocratic step i, k_{ϕ_i} is the retention factor of the compound at the Φ of the isocratic step i, and δt_c is the time that a solute is under the influence of a particular isocratic step of the gradient as it moves through the column (see eq. 3 of the preceding manuscript). Therefore, the error in the position of a compound (as a fraction of the column length) just as it elutes from the column (i.e., $x = 1$), is given by:

$$\sigma_x=\sum_{i=1}^{n}\left(\frac{\delta t_c}{t_{0,\Phi_i\Phi}\ (1+k_{\Phi_i\Phi})}\right)\left(\frac{0.03k_{\Phi_i}}{1+k_{\Phi_i\Phi}}\right) \quad (2)$$

Note that eq. 6 shows that σ_x will be smaller when most of the length of the column is traveled under the influence of a small k such as, for instance, when a compound elutes in a steep gradient.

Once σ_x is determined, $\sigma_{t_R,expected}$ can be calculated by considering the times at which each side of the position confidence interval (i.e., $+\sigma_x$ and $-\sigma_x$) "elutes" from the column. The difference between their elution times and that of the compound's retention time depends on the final velocity, v_f, of the compound immediately before it elutes from the column, which may be determined from:

$$v_f=\frac{1}{t_{0,f}\,(k_f+1)} \quad (3)$$

where v_f is given as the fraction of the column traveled per unit time as the compound elutes, k_f is the final retention factor of the compound just as it elutes, and $t_{0,f}$ is the final dead time just as the compound elutes from the column. If we approximate that the velocities of the two sides of the confidence interval for the position of the compound are the same as that of the compound when they elute, the expected retention projection error to be calculated from:

$$\sigma_{t_R,expected}=\frac{\sigma_x}{v_f}=\sigma_x t_{0,f}\,(k_f+1) \quad (4)$$

Eq. 8 has some important properties. It predicts that the width of the tolerance window is controlled not only by σ_x, but also by v_f, that is, how fast a compound is moving when it elutes from the column. For example, if a compound is moving at a relatively high velocity when it elutes, it will have a narrow $\sigma_{t_R,expected}$ (both sides of the confidence interval will elute with little time in between them). This means that a compound subjected to a steep gradient will not only have a relatively small σ_x, it will also elute with a smaller k_f (i.e., it

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

will elute at a higher velocity), causing a significantly narrower $\sigma_{t_R,expected}$ than if it were subjected to a shallow gradient. Eq. 8 also predicts that compounds with steeper log k vs. Φ relationships will have narrower $\sigma_{t_R,expected}$ because they will have smaller σ_x and will elute with a smaller k_f. Flow rate and column dimensions also affect $\sigma_{t_R,expected}$, but their change to $t_{0,f}$ is in part cancelled by the associated change in σ_x and k_f. For example, decreasing column length decreases $t_{0,f}$, which would decrease $\sigma_{t_R,expected}$, except that σ_x and k_f become significantly larger because Φ does not have a chance to get as high before compounds elute. One caveat of this approach is that when a test compound is almost completely unretained, unreasonably small $\sigma_{t_R,expected}$ are predicted because σ_x approaches zero. To correct this, we set a lower limit for $\sigma_{t_R,expected}$ of 0.5 s.

Based on the above discussion, one might be tempted to think that steeper gradients are better for retention projections because compounds eluting in them have narrower $\sigma_{t_R,expected}$ values (i.e., more accurate retention projections). Indeed, $\sigma_{t_R,expected}$ values are narrower, but steeper gradients also cause compounds to elute closer to one another, canceling any information that might be gained by the more accurate retention projections.

In this light, it may seem wiser to report $\sigma_{t_R,expected}$ values relative to their retention times since steeper gradients produce smaller $\sigma_{t_R,expected}$ values and also shorter retention times. While in linear gradients it may make more sense to do that, in multi-step gradients there are many situations where it would not. For instance, imagine a run where the solvent composition holds at 5% B for 60 min, then jumps up to 95% B in 1 s. For most compounds eluting under these conditions, $\sigma_{t_R,expected}$ will be extremely small even though their retention times will be quite large (for most, their retention times will be roughly 60 min + t_0).

3.2 Comparison of the Global and Specific Approaches to Calculate $\sigma_{t_R,expected}$

We now compare the *global* and *specific* approaches in terms of how well the calculated $\sigma_{t_R,expected}$ values match the observed errors. We apply the global approach to both RPs and LRIs, but we only apply the specific approach to RPs since LRIs do not provide enough information to calculate *specific* $\sigma_{t_R,expected}$ values. We begin by comparing the accuracy of the two approaches in just one of the seven labs (lab E was chosen as a representative lab). This way, the comparison is restricted to conditions where only the methods are different (i.e., different gradients and flow rates), but the lab is fixed. Later, the accuracy of the approaches will be compared across all 7 labs.

3.2.1 Comparison of the Two Approaches in a Representative Lab—The third and fourth columns of Table 1 show the overall accuracy of RPs and LRIs (i.e., the RMS of the error in retention predictions for the 17 test compounds) in the nine methods that lab E ran (they did not provide results for one of the 10 methods). Global values of $\sigma_{t_R,expected}$ for RPs and LRIs were determined to be ±0.028 min and ±0.68 min, respectively. They were calculated by determining the RMS error in RPs and in LRIs for all test compounds in all methods that Lab E ran. Notice that this *global* $\sigma_{t_R,expected}$ for RPs is 24-fold narrower than for LRIs. The distribution of error across the methods was also much narrower for RPs than for LRIs. For instance, the ratio of the maximum error to the minimum error for RPs was

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

4.8, but for LRIs it was 34. Therefore, even when using global values of $\sigma_{t_R,expected}$, the RP methodology offers two major advantages over LRIs: (1) the global $\sigma_{t_R,expected}$ is considerably smaller, making retention time tolerance windows far narrower, and (2) error in RPs is far less method-dependent than LRIs, making a global $\sigma_{t_R,expected}$ more accurate when it is used across a range of methods.

Although the errors in RPs were less affected by the method conditions than LRIs, they were still method-dependent to a significant extent. For example, method 7 was one of the shallowest gradients and the error was relatively high (±0.057 min). If the global $\sigma_{t_R,expected}$ of ±0.028 min was used to determine tolerance windows for that method, they would be half as wide as they should be. Method 1, on the other hand, had a steep gradient and the error was relatively low (±0.012 min). In that case, the global $\sigma_{t_R,expected}$ would give tolerance windows that are about twice as wide as they should be. In fact, out of the 9 methods, 5 failed a two-tailed chi-square test (at 95% confidence), indicating that the distributions of error in those methods were significantly different from the overall distribution of error. Thus, a single, *global* $\sigma_{t_R,expected}$ was not appropriate for all of the methods; there is a systematic influence of the gradient and flow rate on the error in RPs.

On the other hand, the approach to calculate *specific* $\sigma_{t_R,expected}$ values is intended to take those factors into account, hopefully yielding $\sigma_{t_R,expected}$ values that are appropriate for each RP regardless of the method. To test the accuracy of *specific* $\sigma_{t_R,expected}$ values in the different methods, we first normalized each observed RP error by dividing it by the *specific* $\sigma_{t_R,expected}$ calculated for that RP. If the *specific* $\sigma_{t_R,expected}$ values are accurate, the normalized errors should all come from the same distribution of error regardless of the method they were run in. The rightmost column of Table 1 shows the overall error in the normalized RPs for each method. (The overall normalized error across all methods was less than ±1 because this lab showed somewhat more accurate RPs than the average lab.) Using this normalized error, only method 8 failed a two-tailed chi square test at 95% confidence (i.e., it likely came from a different distribution of error), suggesting that *specific* $\sigma_{t_R,expected}$ values are significantly better predictors of error in RPs across methods than a *global* $\sigma_{t_R,expected}$ value.

Figure 2 further demonstrates the superior ability of *specific* $\sigma_{t_R,expected}$ values to correctly predict RP error under various method conditions. It shows the cumulative distribution of error in all of the retention predictions in all of the methods run by lab E (a total of 146 retention predictions). A good fit to the normal cumulative distribution would indicate that the error in retention predictions are random and drawn from the same distribution of error, regardless of the method used. A poor fit to the normal cumulative distribution would indicate that there is systematic error in the retention predictions, likely due to differences between the methods. Notice that the retention prediction errors in Figure 2 are also normalized. RPs are normalized in one case to the *global* $\sigma_{t_R,expected}$ and in another case to *specific* $\sigma_{t_R,expected}$ values, and LRIs are normalized only to the *global* $\sigma_{t_R,expected}$. In other words, the error in each retention prediction is essentially divided by the error expected for that retention prediction (using the different approaches of calculating expected prediction error). The retention prediction errors are normalized to test whether the methods of calculating the expected error appropriately account for differences between methods. If the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

method of calculating expected error properly accounts for systematic differences between methods, the differences will cancel and the normalized error should follow the normal cumulative distribution.

Not surprisingly, the distribution of error in LRIs deviates the most from the normal distribution because the error is strongly affected by the gradient and flow rate; the global $\sigma_{t_R,expected}$ is too large for LRIs in some methods and far too small for LRIs in other methods. The distribution of error in RPs, normalized to the *global* $\sigma_{t_R,expected}$, fits the normal distribution more closely, but still deviates from it. However, when the errors in RPs are each normalized to *specific* $\sigma_{t_R,expected}$ values, the distribution follows a normal distribution much more closely. This supports that systematic sources of error in RPs (at least, those caused by the gradient and flow rate) are better predicted by the *specific* $\sigma_{t_R,expected}$ values than by a *global* $\sigma_{t_R,expected}$ value.

3.2.2 Comparison of the Approaches to Calculate $\sigma_{t_R,expected}$ in All Seven Labs—Similar trends were observed when data from the rest of the seven labs were included in the comparison. Figure 3 shows a chi-square cumulative distribution of the overall error from all of the methods run by the seven labs (a total of 63 methods with an average of 15 test compounds detected in each run). A good fit to the chi-square distribution can only be obtained if the distributions of normalized error in each method from each lab are indistinguishable from one another—that is, if the normalized error is both lab- and method-independent. As expected, neither the LRI nor the RP distribution of error resemble the chi-square cumulative distribution when normalized by the global $\sigma_{t_R,expected}$ values. To quantitatively test the goodness of fit, we used a chi-square goodness of fit test to compare each curve to the normal chi-square cumulative distribution curve (binned at 4 unit intervals). The LRI distribution, normalized to the global $\sigma_{t_R,expected}$, yields a chi-square statistic that is 284-fold greater than the 95% confidence threshold. The chi-square statistic for the RP distribution, normalized to the global $\sigma_{t_R,expected}$ value, is 77-fold greater than the 95% confidence threshold. However, when RPs are normalized by *specific* $\sigma_{t_R,expected}$ values, the error follows the chi-square cumulative distribution much more closely: The chi-square statistic is only 1.7-fold greater than the 95% threshold. Therefore, it appears that the method of calculating *specific* $\sigma_{t_R,expected}$ values correctly predicts a large amount of the lab- and method-dependent sources of error in RPs, though evidently not quite all of it, since the distribution of normalized error is statistically different from the normal chi-square distribution. The difference can be plainly seen from a visual comparison of the two curves. There appear to be too many methods that gave larger RP errors than expected from *specific* $\sigma_{t_R,expected}$ values, causing the curve to be drawn out at higher χ^2 values.

Indeed, there were relatively small, but significant differences between the labs that were not taken into account by our approach to calculate *specific* $\sigma_{t_R,expected}$ values. Figure 4 shows a cumulative distribution of the RP error (i.e., measured minus projected, normalized to specific $\sigma_{t_R,expected}$ values) in each individual lab along with the normal distribution of error. Five of the seven labs had distributions of error that were statistically different according to a chi-square test at 95% confidence. Of those, labs B and C showed the most error by far, though the reason for this is unclear. Lab B did not have a pre-column heater, which we previously found can affect the accuracy of RPs, but their RPs did not show the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

characteristic signs of temperature inaccuracy. Nor was there anything unusual about lab C. Their HPLC system showed the largest gradient delay and dispersion of all the labs, but the back-calculation algorithm should have been able to account for it as it has on other HPLC systems with large gradient delay and dispersion. We plan to study this more in the future, but regardless, the differences between the labs are relatively small. *The overall normalized error from the least accurate lab is less than twice that of the most accurate lab.*

3.3 Properties of Retention Time Tolerance Windows Calculated from $\sigma_{t_R,expected}$

Though some labs experienced more error than others, we found that if the *specific* $\sigma_{t_R,expected}$ values are relaxed by just 50%, the wider, more conservative tolerance windows are appropriate even for the least accurate lab (lab C) up to 3.3 $\sigma_{t_R,expected}$ (i.e., 99.9% confidence). Since the relaxed tolerance windows are conservative for all but one of the seven labs involved in this study, the confidence associated with them should generally be interpreted as a *minimum* level of confidence.

Even when the tolerance windows are relaxed by 50%, they are still quite narrow. Figure 5 shows a histogram of the 99% confidence windows calculated for all of the 946 RPs in the study. To give a frame of reference, the World Anti-Doping Agency (WADA) specifies that for a positive identification, the retention time of an analyte “shall not differ by more than two (2) percent or ±0.1 min (whichever is smaller)” from that of a reference compound analyzed contemporaneously [18]. Of the 946 retention time tolerance windows, 63% were narrower than the 2% criteria, 50% were narrower than the ±0.1 min criteria, and 42% were narrower than both criteria for identification in WADA labs. This is remarkable because a large fraction fit the criteria *without* requiring a standard on hand to run contemporaneously.

Interestingly, the width of the 99% confidence windows for LRIs calculated in this study is so wide that it would be almost worthless in short gradients. For example, in a 5 min gradient, the conservative, global 99% confidence window was ±3.1 min, which would often bracket the entire chromatogram. On the other hand, using conservative, specific $\sigma_{t_R,expected}$ values, the average 99% confidence window for RPs in a 5 min gradient was ±0.054 min, 57-fold narrower.

3.4 System Suitability Check

For the validity of the tolerance window calculations to extend beyond the test compounds and the labs in this study, two factors must hold true. First, when new k vs. Φ relationships are measured and added to the retention database, the overall accuracy of k measurements must be no worse than the other k measurements already in the database. If the accuracy is worse, the calculated tolerance windows for those compounds may be too narrow. In a future manuscript, we will discuss a practical methodology we are developing to measure the k vs. Φ data that ensures a high level of accuracy.

The second factor is that other labs must have their HPLC systems in a “suitable” state for the tolerance windows to apply. Currently, certain experimental factors must be fixed including the temperature, the stationary phase, and the mobile phase solvents (these requirements are discussed in more detail in a previous article [17]), but even if a lab follows

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

these instructions perfectly, other factors may cause retention projections to lose accuracy, thereby making the tolerance windows too narrow. For example, a column oven may not work properly, a pre-column eluent heater may be ineffective, the mobile phases may be prepared incorrectly, or the column's selectivity may have changed (e.g., due to column aging). Therefore, it is important to test whether a system is in suitable shape to know whether the calculated retention time tolerance windows can be trusted.

We propose a new system suitability check for this purpose, similar to one we developed for gas chromatography [19]. A user spikes their sample with both the instrument calibration standards and the test compounds (or potentially they could run the standards and test compounds in a separate run). Then they run the sample, back-calculate the gradient and t_0 vs. Φ profiles (based on the retention times of the instrument calibration standards), and project the retention times of the test compounds. If the accuracy is under a certain threshold, the system is deemed suitable. Otherwise, the user would need to troubleshoot the system until the test passes. The threshold for a "suitable" system is determined by comparing the measured distribution of error with the expected distribution of error using a chi-square test to see if the distributions are different. If the confidence that they are different is less than 75%, the system passes. If it is between the 75% and 95%, the system suitability is considered questionable, and if it is greater than 95%, it fails.

This system suitability check is now built into the online HPLC retention projection software. Immediately after the gradient and t_0 vs. Φ profiles are back-calculated, the software projects the retention times of the test compounds and compares them with the measured retention times. A numerical rating is then assigned to the user's system by dividing the measured standard deviation of error (among all the test compounds) by the standard deviation of error that would be expected at the chi-square 75% confidence threshold. Therefore, a rating of '1' or lower passes the system suitability check and it falls into the green region of a displayed indicator bar. Yellow and red portions of the indicator bar correspond to the questionable and failed ratings. Following a passed system suitability check, the software then projects the retention times of all of the rest of the compounds in the retention database and calculates the appropriate tolerance window (using conservative values) for each one given a user-specified level of confidence. We must caution, however, that even if a system suitability check passes, it is still possible that the tolerance windows may be too narrow when the column is mass overloaded with one or more components of the sample. As usual, care should be taken to avoid this situation.

4. Conclusions

A new approach was developed to calculate the appropriate retention time tolerance windows for retention times projected using shared k vs. Φ relationships (and back-calculated gradient and dead time vs. Φ profiles), potentially making it possible to use LC retention information to reject candidate identities of a chromatographic feature with an absolute level of confidence *without needing to have authentic standards of each candidate on hand*. The approach assumed that the major source of error comes from the k vs. Φ relationships; propagation of that error enables individual retention time tolerance windows to be calculated for each retention projection. These *specific* tolerance windows were

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

compared with the conventional approach where a single, *global* tolerance window is applied to all retention predictions. In a range of multi-segment gradients and flow rates run among seven different labs, the specific tolerance windows were significantly more appropriate for each retention projection than the global tolerance window.

Nevertheless, the accuracy of the calculated tolerance windows still varied somewhat between labs, suggesting that there were some differences between the labs that were not taken into account. Fortunately, these differences were relatively small; *the accuracy of tolerance windows in the least accurate lab was only 2-fold worse than in the most accurate lab*. The calculated tolerance windows only needed to be relaxed by 50% to make them appropriate for the worst lab. Yet even with 50% wider tolerances, 42% of the tolerance windows calculated in this study at the 99% confidence level were narrower than those required by WADA for positive identification, which is remarkable since the RP methodology does not require authentic standards to be on hand. To ensure the integrity of the calculated tolerance windows, we proposed a simple system suitability check that determines whether a user's system is in a state such that the tolerance windows may be trusted. The system suitability check and the retention time tolerance window calculations are now built into the online retention projection software (www.retentionprediction.org/hplc).

An upcoming challenge will be to ensure that any k vs. Φ relationships added to the database have the same or lower error than other compounds already in the database. We are currently working to develop a methodology to make it relatively fast and easy for other labs to measure k vs. Φ relationships with the required level of accuracy. It will also be important to ensure that the k vs. Φ database is very carefully curated (we are currently building one which is available in its current state at www.retentionprediction.org/hplc/database) as any errors would deteriorate confidence in the tolerance windows. Finally, the calculated tolerance windows will eventually need to be validated in a larger number of labs.

Acknowledgments

We thank the National Institute of General Medical Sciences of the National Institutes of Health [R01GM098290], the Minnesota Agricultural Experiment Station, and we thank Agilent Technologies for generously donating the HPLC columns used in this work.

References

1. Sumner LW, Lei Z, Nikolau BJ, Saito K, Roessner U, Trengove R. Proposed quantitative and alphanumeric metabolite identification metrics. Metabolomics. 2014; 10:1047–1049.
2. Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, et al. Proposed minimum reporting standards for chemical analysis. Metabolomics. 2007; 3:211–221. [PubMed: 24039616]
3. Kind T, Fiehn O. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. BMC Bioinformatics. 2007; 8:105. [PubMed: 17389044]
4. Kind T, Fiehn O. Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm. BMC Bioinformatics. 2006; 7:234. [PubMed: 16646969]
5. Kind T, Fiehn O. Advances in structure elucidation of small molecules using mass spectrometry. Bioanal Rev. 2010; 2:23–60. [PubMed: 21289855]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

6. Mylonas R, Mauron Y, Masselot A, Binz PA, Budin N, Fathi M, et al. X-Rank: A Robust Algorithm for Small Molecule Identification Using Tandem Mass Spectrometry. Anal Chem. 2009; 81:7604–7610. [PubMed: 19702277]
7. Stein SE. Estimating probabilities of correct identification from results of mass spectral library searches. J Am Soc Mass Spectrom. 1994; 5:316–323. [PubMed: 24222569]
8. Pavlic M, Libiseller K, Oberacher H. Combined use of ESI–QqTOF-MS and ESI–QqTOF-MS/MS with mass-spectral library search for qualitative analysis of drugs. Anal Bioanal Chem. 2006; 386:69–82. [PubMed: 16896628]
9. Oberacher H, Pavlic M, Libiseller K, Schubert B, Sulyok M, Schuhmacher R, et al. On the inter-instrument and the inter-laboratory transferability of a tandem mass spectral reference library: 2. Optimization and characterization of the search algorithm. J Mass Spectrom. 2009; 44:494–502. [PubMed: 19152368]
10. Stein SE, Scott DR. Optimization and testing of mass spectral library search algorithms for compound identification. J Am Soc Mass Spectrom. 1994; 5:859–866. [PubMed: 24222034]
11. Neumann S, Böcker S. Computational mass spectrometry for metabolomics: identification of metabolites and small molecules. Anal Bioanal Chem. 2010; 398:2779–2788. [PubMed: 20936272]
12. Gergov M, Weinmann W, Meriluoto J, Uusitalo J, Ojanperä I. Comparison of product ion spectra obtained by liquid chromatography/triple-quadrupole mass spectrometry for library search. Rapid Commun Mass Spectrom. 2004; 18:1039–1046. [PubMed: 15150826]
13. Dresen S, Kempf J, Weinmann W. Electrospray-ionization MS/MS library of drugs as database for method development and drug identification. Forensic Sci Int. 2006; 161:86–91. [PubMed: 16860958]
14. Bristow AWT, Webb KS, Lubben AT, Halket J. Reproducible product-ion tandem mass spectra on various liquid chromatography/mass spectrometry instruments for the development of spectral libraries. Rapid Commun Mass Spectrom. 2004; 18:1447–1454. [PubMed: 15216504]
15. Josephs JL, Sanders M. Creation and comparison of MS/MS spectral libraries using quadrupole ion trap and triple-quadrupole mass spectrometers. Rapid Commun Mass Spectrom. 2004; 18:743–759. [PubMed: 15052556]
16. Boswell PG, Schellenberg JR, Carr PW, Cohen JD, Hegeman AD. A study on retention "projection" as a supplementary means for compound identification by liquid chromatography-mass spectrometry capable of predicting retention with different gradients, flow rates, and instruments. J Chromatogr A. 2011; 1218:6732–6741. [PubMed: 21862024]
17. Abate-Pella D, Freund DM, Ma Y, Simón-Manso Y, Hollender J, Broeckling CD, et al. Retention Projection Enables Accurate Calculation of Liquid Chromatographic Retention Times Across Labs and Methods. 2015 Submitted to J. Chromatogr. A.
18. Identification Criteria for Qualitative Assays Incorporating Chromatography and Mass Spectrometry, WADA Tech. Doc. - TD2010IDCR. (2010).
19. Barnes BB, Wilson MB, Carr PW, Vitha MF, Broeckling CD, Heuberger AL, et al. "Retention Projection" Enables Reliable Use of Shared Gas Chromatographic Retention Data Across Laboratories, Instruments, and Methods. Anal Chem. 2013; 85:11650–11657. [PubMed: 24205931]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Highlights

- Compound identification would benefit from LC retention data
- But standards must currently be run to identify compounds with absolute confidence
- A way to calculate absolute confidence with shared retention data was developed
- It accounts for many lab- and method-derived sources of error
- 99% confidence windows were narrow compared to typical windows *with* standards

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

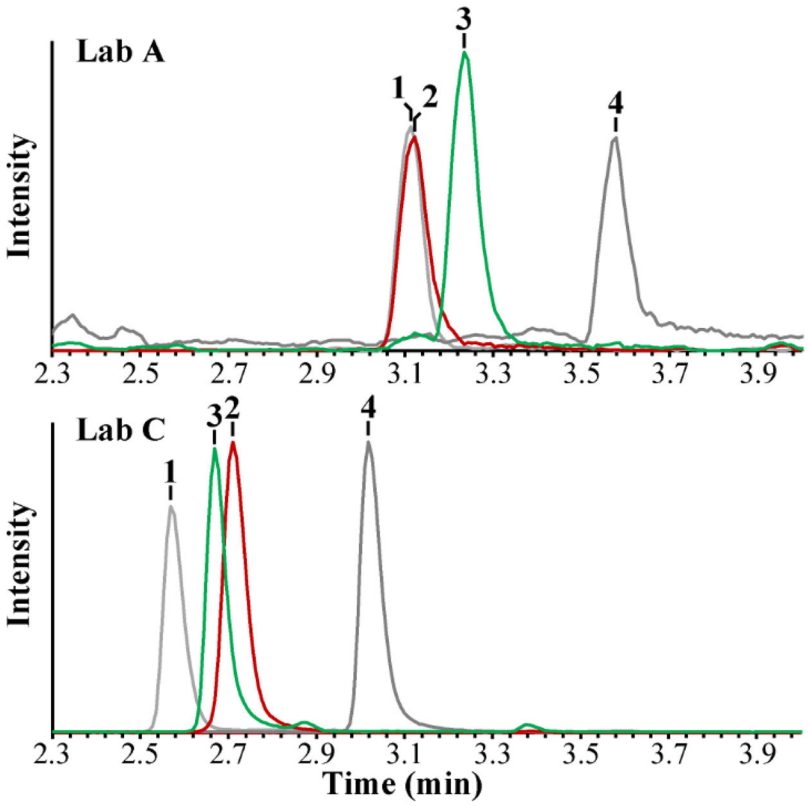

Figure 1.

In gradient elution, even unintentional differences between HPLC systems are enough to cause relative retention to change. Here, *N*-allyl aniline (**2**) and 7-methyl-4-aminocoumarin (**3**) were run under nominally the same conditions by two different labs, but they eluted in a different order. (Compounds **1** and **4** were *N*-ethylbenzamide and *N*-propylbenzamide, respectively.)

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

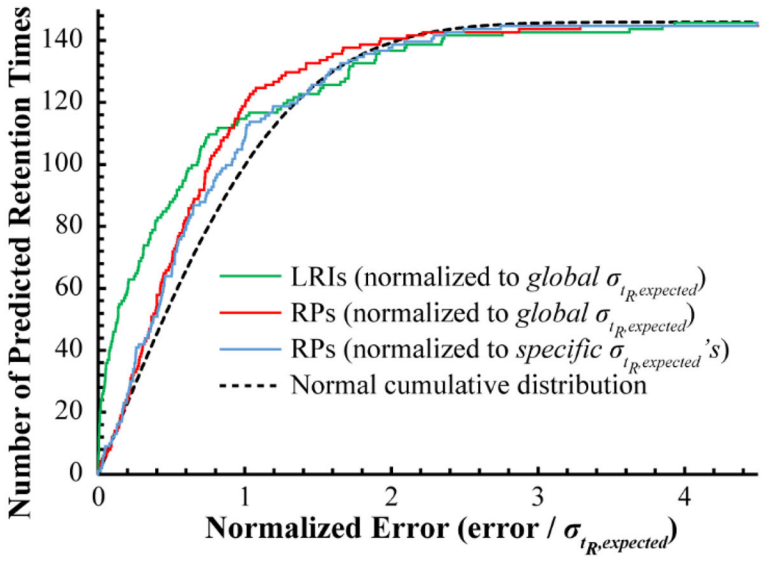

Figure 2.

Cumulative distribution of the normalized error in LRIs and retention projections in nine different methods run by lab E. *Global* values of $\sigma_{t_R,expected}$ do not predict the correct error in LRIs or retention projections under different gradients and flow rates (the normalized error does not follow a normal distribution), but *specific* $\sigma_{t_R,expected}$ values predict the error in retention projections more closely.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

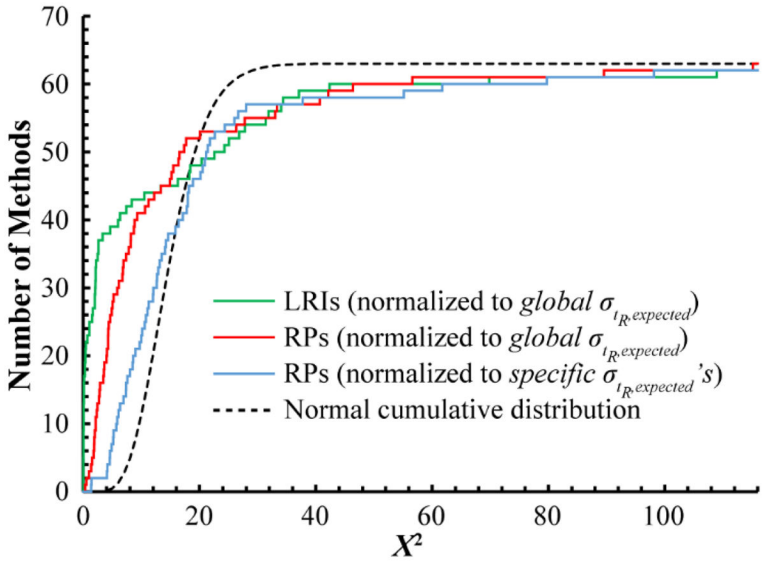

Figure 3.

Chi-square cumulative distribution of the overall normalized errors from each of the 63 methods run among the 7 labs. *Specific* $\sigma_{t_R,expected}$ values predict the error in different labs and methods much more accurately than a *global* $\sigma_{t_R,expected}$ value. This is shown by the better fit of RP error to the normal chi-square cumulative distribution when it was normalized to the *specific* $\sigma_{t_R,expected}$ values than the *global* $\sigma_{t_R,expected}$ value. Still, the curve does not fit the normal distribution perfectly.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

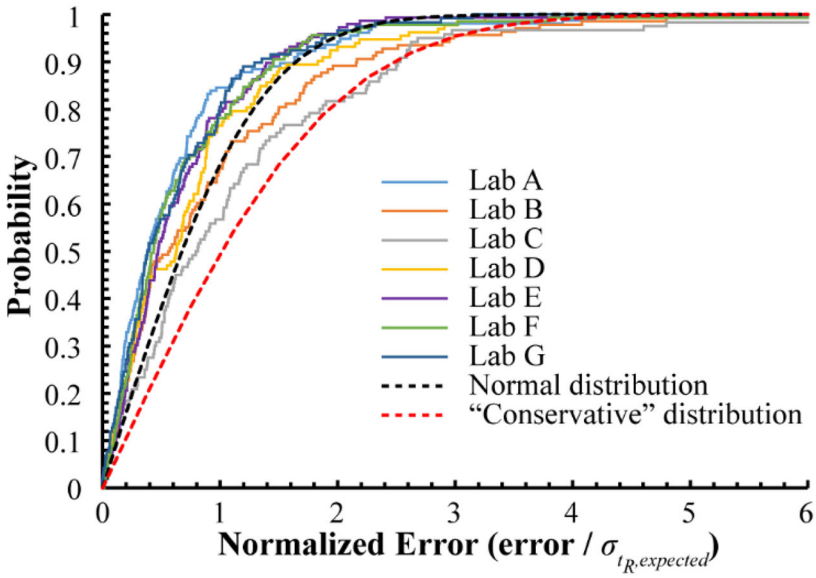

Figure 4.

Cumulative distributions of the error in RPs (normalized to specific $\sigma_{t_R,expected}$ values) in each of the seven labs. Though the error in each lab generally follows a normal distribution, some labs show slightly more or less error than others. To ensure retention time tolerance windows are wide enough, even in the lab with the worst accuracy (lab C), "conservative" tolerance windows may be used that are 50% wider.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

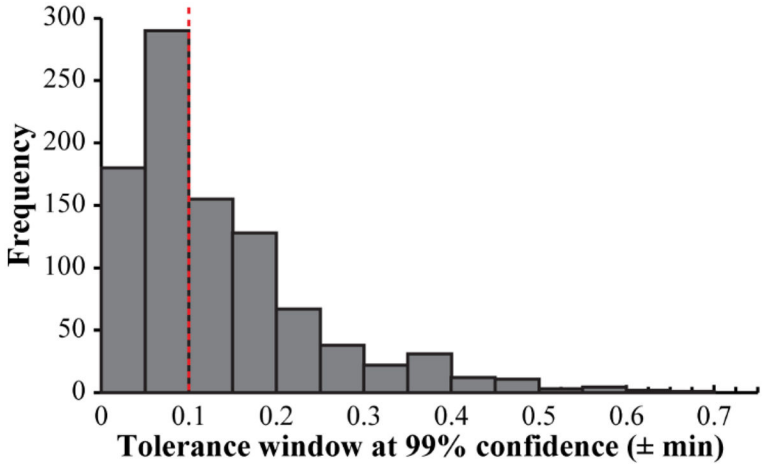

Figure 5.

Histogram of the 99% confidence windows calculated for the 946 retention projections in the seven-lab study. The red dashed line shows the ±0.1 min criteria for positive identification in WADA labs (where standards are run contemporaneously). 42% of the tolerance windows calculated for the runs in this study (at 99% confidence) were narrower than WADA criteria, and they were calculated for each lab and method using only shared retention data, that is, no authentic standards were necessary.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1

Accuracy of RPs and LRIs in the Nine Methods Run by Lab E

Method	Gradient	Flow Rate (mL/min)	RP Error[a] (min)	LRI Error[a] (min)	RP Error Normalized to *specific* $\sigma_{t_{R,expected}}$'s (dimensionless)[b]
1	5 to 95% B in 5 min	0.4	±0.012	±0.038	±1.2
3	5 to 95% B in 10 min	0.2	±0.018	±0.079	±1.0
4	5 to 95% B in 10 min	0.8	±0.017	±0.31	±1.1
5	5 to 70% B in 15 min, 95% B in 1.5 min	0.56	±0.023	±0.74	±0.68
6	5 to 70% B in 15 min, 95% B in 1.5 min	0.7	±0.031	±0.82	±0.91
7	5 to 55% B in 15 min, 70% B in 10 min, 95% in 0.5 min	0.5	±0.057	±1.1	±0.86
8	5 to 55% B in 3 min, 70% B in 6 min, 95% B in 6 min	0.4	±0.018	±0.33	±0.57
9	5 to 25% B in 3 min, 60% B in 6 min, 95% B in 3 min	0.4	±0.022	±1.3	±0.66
10	5 to 25% B in 4 min, 60% B in 2.5 min, 95% B in 13.5 min	0.4	±0.024	±0.30	±0.78
		Overall error:	**±0.028**	**±0.68**	**±0.88**

[a]Errors were calculated from the root mean square (RMS) of errors in the retention predictions (i.e., RP or LRI) for all of the test compounds (up to 17 were detected in each run).

[b]Each RP error (measured minus projected) was normalized by dividing it by the *specific* $\sigma_{t_{R,expected}}$ calculated for it using a relative error in *k* of 3%. Lab E was slightly more accurate than other labs, so the overall error came to ±0.88 rather than ±1. The values shown are the RMS of the normalized errors in each method.