Reducing the Pain Point of GC Retention Index Calibration for Enhanced GC/MS Compound Identification

Don Kuehl, Stacey Simonoff, Yongdong Wang
Cerno Bioscience, Las Vegas, NV USA

MP 391

Introduction

Retention Index (RI) matching is likely the most important complementary metric for identifying unknown compounds with GC/MS library search. Combining the two can dramatically improve the confidence of unknown compound identification. However, proper RI calibration of GCs can be time-consuming and error-prone, especially for sampling techniques such as SPME and LVI, which can produce significant chemical background interferences. In this paper, we will describe GC calibration strategies to simplify and improve the RI calibration process so that it can be performed routinely, reliably, and automatically.

Method

GC/MS n-alkane calibration runs from over 50 different GC/MS systems, and various sampling techniques (direct injection, split, SPME, LVI, headspace, etc.) were used to evaluate automating RI calibration. Some of the standards included n-alkane ladders with gaps, and many contained significant background interferences, which complicate the calibration process. A number of different algorithms written in C++, Python, or Excel were tested and compared against the data sets. The NIST23 library, which contains extensive RI data and NIST search software, was also used for the evaluation of different approaches.

Results and Discussion

Recent improvements in MS library RI metadata (NIST20, NIST23)[1] have addressed gaps in experimental RI compound coverage using advanced AI modeling[2], allowing comprehensive RI coverage for the entire NIST library of commonly used GC column types. However, to take advantage of the library RI values, one must calibrate the GC for retention index on a regular basis (daily or perhaps every few days or weeks). This is commonly done using a series of n-alkanes such as C6, C7, C8... where C? is the number of carbons in the n-alkane. The series is typically set to Cn, Cn+1...Cn+m where n and m are selected to bracket the retention time (RT) range of the run. The series does not necessarily need to be a continuous ladder, and some C? values may be skipped, especially for the heavier n-alkanes eluting later in the run which tend to "bunch up". An example of a high quality, n-alkane calibration sample is shown in Figure 1.

In the presence of large background interferences (SPME and LVI are notorious for producing large interfering peaks), it may not be very clear where the n-alkane peaks are located, particularly for an automated system. Carry-over, old "dirty" columns, impurities in the n-alkane "standards" and other real-world interferences can also be problematic.

One possible way to automatically identify the calibration compounds is by simply selecting the n largest peaks in the run, where n is the total number of n-alkanes in the calibration, and to assume they are the desired calibration compounds. But, in practice, it is not uncommon for the previously discussed background or contaminate peaks to be present in higher concentrations than the calibration compounds. This is especially true with the peaks of the later, heavier (higher molecular weight) standards that can be broader and less intense than earlier eluting peaks, making them even more susceptible to background interferences. Exacerbating the problem of less intense peaks at the greater retention times is that the background interference may be higher later in the run, typically from column bleed, which can mask the identification of the smaller, heavier n-alkanes peaks (Figure 2).

Another approach to automatically identifying the n-alkane standards is to simply use the library search to help identify the standard peaks. Unfortunately, as the n-alkanes molecular weights get heavier (e.g., particularly above a carbon count of C15-20), conventional search algorithms cannot uniquely identify and differentiate between the n-alkanes.
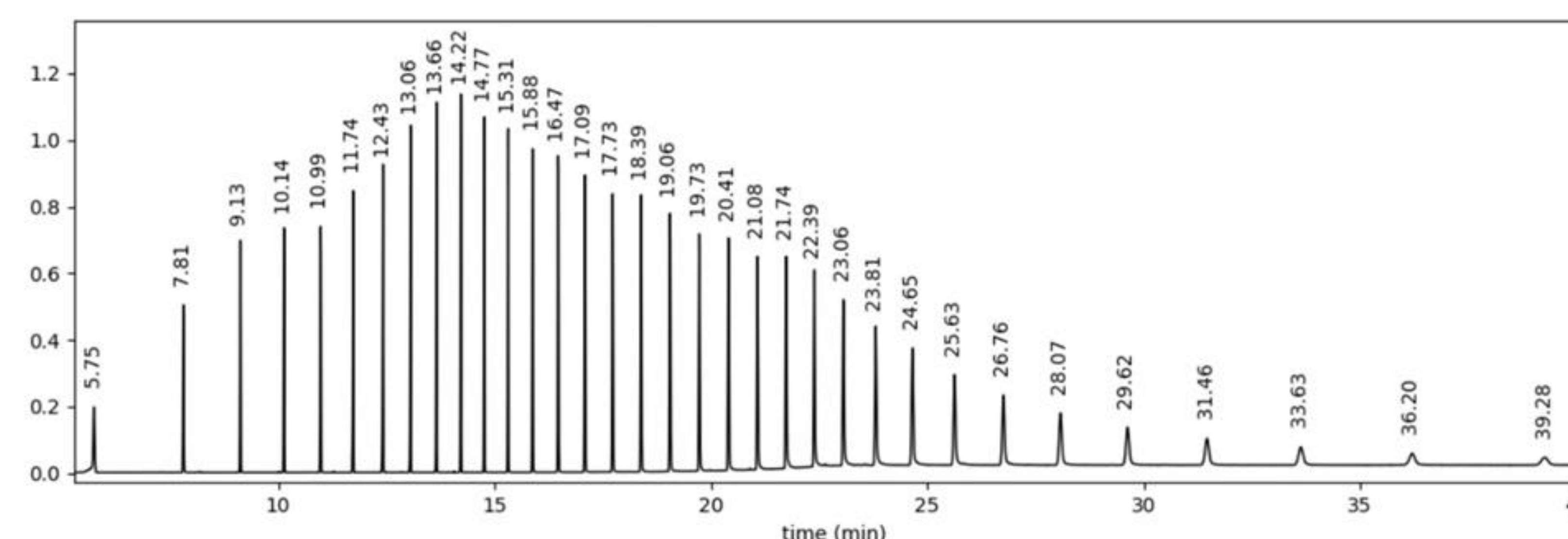

Figure 1. A "clean" alkane calibration containing n-alkanes C7-C39. Note the absence of any significant background or contaminant peaks making the identification of each compound in the series a simple problem of finding the n largest peaks where n is the n-alkane count.
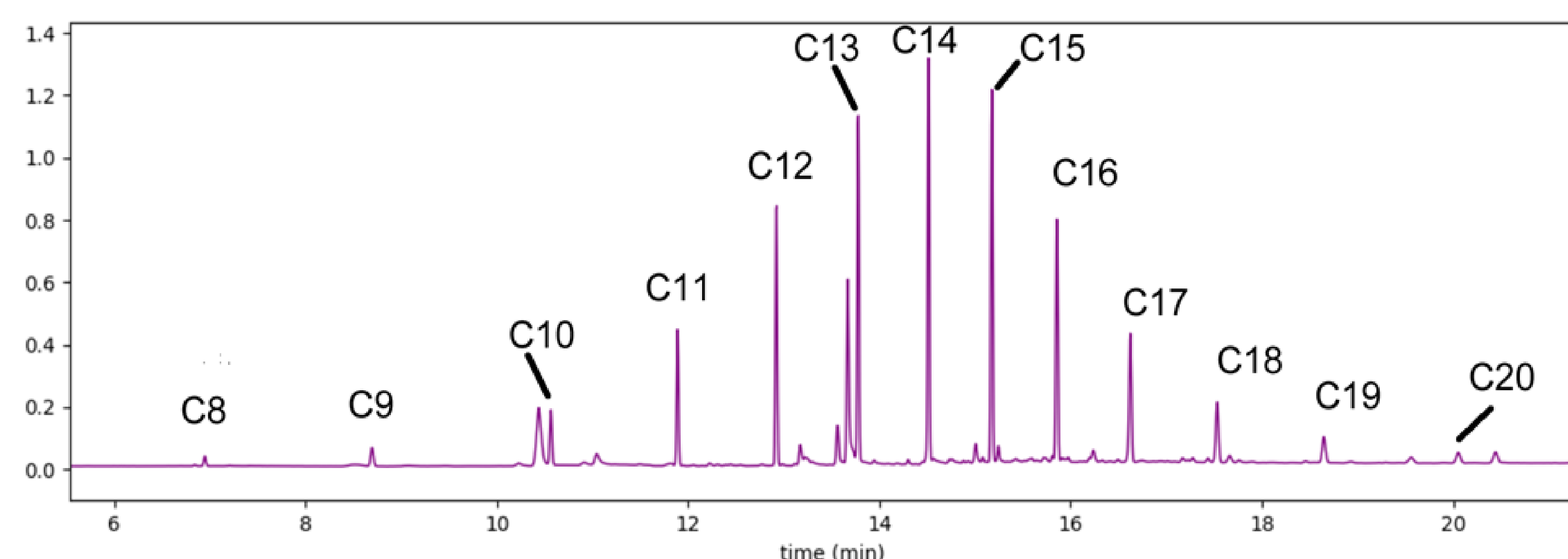

Figure 2. A n-alkane calibration which contains contaminant peaks carried over from SPME. Note that identification of the alkane peaks by intensity would not be possible. In addition, search identification of the n-alkanes greater than C16 fails to find the correct match due to the high background and low signal-to-noise.

The reason for this is that the MS electron impact (EI) fragmentation patterns on n-alkanes all contain the exact same fragments up to the molecular ion m/z of the standard. This is because n-alkanes are a simple linear carbon chain where the characteristic cleaving of the molecule occurs between the carbon bonds. As the molecular weight increases, statistically, the intensity of the heavier fragments is drastically reduced to the point that they contribute a proportionally insignificant unique fragment fingerprint for the compound. In fact, with some of the heaviest n-alkanes, the correct search match sometimes doesn't even show up among the top 20 best search matches. However, while the correct n-alkane may not be correctly identified, the best matches will typically be a compound with an n-alkane formula and almost or nearly all matches will be for saturated alkanes (of the formula $C(n)H(2n+2)$). So, to summarize, while heavier n-alkanes are not uniquely identified, they are usually identified as being in the n-alkane class or "family" of compounds. This helps us identify the compounds that are in the family of n-alkanes, but not the specific n-alkane.

A generalized solution to automatically assigning the correct n-alkane RI standards in a calibration run is to use the identification power of mass spec search combined with the knowledge or estimation of the typical elution spacing between consecutive n-alkane peaks. If one examines a typical n-alkane elution pattern such as exhibited in Figure 1, it shows the alkane spacing starting out wide, becoming narrow, and then becoming wider towards the end. These elution patterns are dictated primarily by the GC oven temperature program used over the GC run as illustrated in Figure 3.
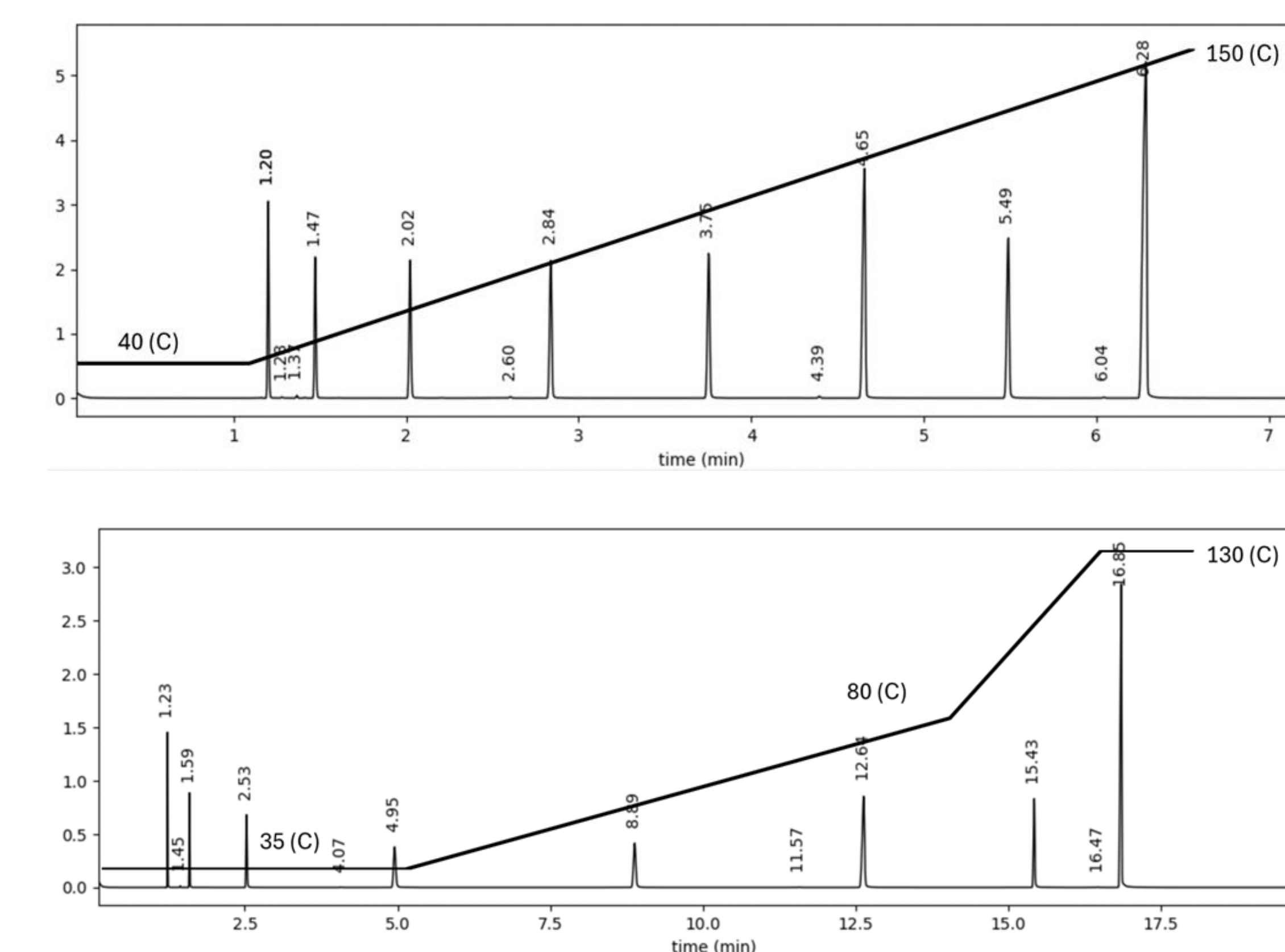

Figure 3. Different oven temperature profiles create dramatic differences in the n-alkane elution pattern.

We know we can easily ID the lighter alkanes, so, we can estimate the elution time of the next alkane peak by using the 3 previous known retention times of the n-alkane ladder. We can think of the delta gap between the 2 previous known n-alkanes peaks as the "velocity" of the elution pattern at that RT. But we also see that the pattern can smoothly change to either a narrower or wider spacing. We can think of this as the "acceleration" of the elution pattern which can be either negative or positive, i.e., smaller or larger spacing of the peak pattern. The acceleration can be calculated by looking at the second derivative, or the delta-delta of the three previous peaks. We can use this to help us identify the location of the next unknown n+1 n-alkane which can help eliminate interfering peaks that fall outside of this estimated position. Another way to use this information is to use search to locate the next n-alkane, which as discussed may not be correctly identified as the correct n-alkane, but in the n-alkane family. Using the calculated velocity and acceleration of the elution profile, we can assign the expected n value for this calibration peak. This is useful as we do not always have a sequential alkane ladder, or possibly an n-alkane is masked from identification as it may coelute with an interfering peak.

This algorithm was implemented in C++ and Python and run against a variety of test cases. In general, the approach could automatically identify the n-alkanes in a run, but in some corner-cases would fail if the estimation error of the RI values was significant. In future work a more accurate RI estimation would help and may be possible by also incorporating the temperature programming information.

Conclusion

Automated RI index calibration is complicated by interferences and the lack of a reliable way to identify heavy n-alkanes. Combining the estimation of the n-alkane RI based on previous n-alkane elution with library search provides a way to calibrate the GC in a more robust fashion.

References

1. NIST20 and NIST23 Mass Spectral Databases, NIST.gov.
2. AIRI: Predicting Retention Indices and Their Uncertainties Using Artificial Intelligence Lewis Y. Geer, Stephen E. Stein, William Gary Mallard, and Douglas J. Slotta Journal of Chemical Information and Modeling Article ASAP