PeakClimber: A software tool for analyzing biological HPLC data using the exponential Gaussian function

Joshua T. Derrick[1], Steven A. Farber[1,2], William B. Ludington[1,2]

[1] Department of Biology, Johns Hopkins University, Baltimore, Maryland, United States

[2] Department of Embryology, Carnegie Institute for Science, Baltimore, Maryland, United States

*Corresponding author

Abstract

High-performance liquid chromatography (HPLC) is a common medium-throughput technique to analyze metabolic samples. However, analysis of HPLC data is hampered by a lack of tools to accurately determine the precise analyte quantities on a level of precision equivalent to mass-spectrometry approaches. To combat this problem, we developed a tool we call PeakClimber, that uses a sum of exponential Gaussian functions to accurately quantify the peaks in HPLC traces. In this paper we analytically show that HPLC peaks are well-fit by an exponential Gaussian function, that PeakClimber more accurately quantifies known peak areas than standard industry software and utilize PeakClimber to make new discoveries about lipid biology.

Introduction

Liquid chromatography (LC) is a series of techniques to separate individual analytes from a mixture of chemicals using a liquid, or mixtures of various liquids as a solvent[1]. As compared to gas chromatography (GC), that uses inert gases as solvents (usually helium), liquid chromatography can separate particles of larger molecular weight[2]. Original liquid chromatography techniques relied on gravity for solvent flux, which meant that running of individual chromatographs took hours and sometimes days to complete. In the 1960s[3], High-pressure (or performance) liquid chromatography was introduced, speeding up the flow rate by forcing the solvent through an extremely narrow column at high-pressures. The massive decrease in runtime however came with a cost: peaks that had formerly been well-separated now partially overlapped[1], making quantification more difficult. For many HPLC applications, which are more concerned with presence/absence of individual analytes, or with quantification of sparse, relatively pure solvents, simple peak integration is sufficient. For more complicated chemical and biological samples, with overlapping peaks this method is inadequate.

Various solutions have been proposed to this problem. Common industry software, such as ThermoFisher's Chromeleon and Water's MassLynx utilize a method known as valley-to-valley,[4] where a line is dropped from the lowest point between two peaks to the chromatograph's baseline, which is determined by the rolling-ball method[5]. The two peaks are then integrated on either side of the line. This method has the advantage of being neutral to the underlying peak shape, is independent of the surrounding peaks, and has an extremely fast runtime. However, most peaks map to some variation of the Gaussian distribution and are not independent of neighboring peaks that they overlap with. Two more recent open source software packages, HappyTools[6] and hplc.io[7] improve on the valley-to-valley method by fitting each chromatograph to a sum of Gaussian or skew Gaussian curves respectively. Early quantitative models of liquid chromatography show that analytes unbind the column with an exponential decay that is convolved by Gaussian noise based on their physical location along the length of the column and diffusion in the liquid phase before reaching the detector[8–10]

One lingering controversy in the field of chromatography is modeling the shapes of a single peak analytically. While the shape of a peak depends on the amount of sample loaded on the column, however, Langmuir kinetics usually leads to a Gaussian shaped peak with tailing[11]. Peak fronting can occur in rare cases of anti-Langmuir analyte behavior[12,13]; this is a case where PeakClimber is not well-suited for chromatographic analysis. Here, we will argue that an exponential Gaussian function both analytically and empirically fits the distribution of non-overloaded HPLC peaks.

In this manuscript, we show that HPLC analyte peaks are best fit with an exponential Gaussian function. Our tool, PeakClimber, fits chromatographs to a sum of exponential Gaussian curves. We that these curves demonstrate are mathematically and empirically good fits for single analyte peaks, and consistent with extensive literature suggesting that this approach empirically fits for chromatography data[14–16]. PeakClimber also makes iterative

improvements in denoising data, detrending data, and in reducting in runtime. To highlight the utility of PeakClimber, we compare its performance analyzing coinjections of three fatty-acids to the three aforementioned algorithms. Finally, we use PeakClimber to analyze differences in lipid composition between *Drosophila melanogaster* that were reared with and without bacteria.

Results

Traditional chromatography analysis methods fail to accurately quantify complex peaks

Traditional chromatography analysis software relies on integration of the area under the peak to quantify each analyte. This works well when analyte peaks are well separated, but when multiple peaks overlap in the course of a few minutes, a peak-delineation algorithm in needed. Most software, such as ThermoFisher's Chromeleon or Water's MassLynx, relies on a method known as valley-to-valley[4], where a straight line is drawn down from the lowest point between two consecutive peaks to delineate an integration boundary. However, this method often leads to a mismatch between the calculated and true peak area (Figure 1A). To quantify the error of this approach, we conducted a simulation with three-synthetic exponential Gaussian peaks with randomized parameters that overlapped significantly. Our simulations showed that the valley-to-valley method has significant error between the true peak shape and the valley-to-valley integration regions, but this error is especially marked for the first peak in the trace (Figure 1B). This is likely due to the undercounting of the exponential-tail region of the first peak by a valley-to-valley analysis.

Single-analyte HPLC peaks fit an exponential Gaussian distribution

Analytical solution:

Consider a column of finite length containing solute S, initially completely bound to the column at a single location. Solvent U is run over same column. Solute S has affinity k_1 for solvent U. In this case we are assuming that the reverse reaction is negligible, unbounded S flows away in the solvent. Initially, assume there is no diffusion once the solute leaves the column, meaning the solute arrives at the detector a constant amount of time after unbinding from the column. This behavior can be described by the differential equation:

$$\frac{dS}{dt} = -k_1 * S$$

Which we can solve analytically:

$$S(t) = k_1 * e^{-k_1 * t} + C_2$$

This approximates an exponential probability distribution. With an agent-based Monte-Carlo (Figure 2A, histogram) simulation with parameters for S (amount of analyte), k_1 (affinity for solvent U), column length, and flow rate, that are relevant to common HPLC columns, we can recapitulate the analytical solution almost exactly (Figure 2A, red line on blue histogram).

However, once off the column, analyte particles experience both diffusion and flow. We can assume that the column width and height are infinitesimally small compared to its length, and thus the analyte can be considered to follow the advection-diffusion[17] (or diffusion with drift) equation once it leaves the column.

$$\frac{\partial c}{dt} + u * \frac{\partial c}{dx} = D * \frac{\partial^2 c}{\partial x^2}$$

$$c$$

With c(x,t) representing the concentration of the analyte, we can solve this differential equation for c(x,t), obtaining a Gaussian function.

$$c(x,t) = \frac{M}{\sqrt{4\pi Dt}} * e^{-(\frac{x-u*t}{4DT})}$$

With M being the total number of analyte particles. We also only care about the concentration at the detector, thus we can assume x is a constant.

$$c(t) = \frac{M}{\sqrt{4\pi Dt}} * e^{-(\frac{x-u*t}{4Dt})}$$

To remove the flow term from modifying t, we will define a term P=D/u. Dividing the top and bottom inside and outside the exponential by u we get:

$$c(t) = \frac{M/\sqrt{u}}{\sqrt{4\pi Pt}} * e^{-(\frac{x/u-t}{4Pt})}$$

Since the particle is coming off the column with Langmuir kinetics[11], and then is undergoing diffusion-advection, we can represent the union of the two processes with a convolution.

$$Z = S + C$$

The convolution of these two functions is an exponential Gaussian function with the equation.

$$\boldsymbol{Z(t) = \frac{M * k_1}{2 * \sqrt{u}} * e^{(\frac{k_1}{2}*(2(x/u)+4k_1*Pt-2t)} * (-1 + 1/\sqrt{2Pt} + Erfc(\frac{x/u + 4k_1 * Pt - t}{\sqrt{4\pi Pt}})}$$

The ratio between the diffusion rate D and flow rate u, P, is the width or sigma of the exponential Gaussian. As the flow rate increases relative to diffusion this sigma decreases. However, unlike in the conventional exponential Gaussian equation, this sigma will increase over time as it varies directly with t. k_1 is the unbinding coefficient of the analyte from the column and represents the gamma variable of the exponential Gaussian. This determines how large the tails of the function are. The center of the distribution is determined by the mean time to the detector; or the detector distance x divided by the particle velocity u (x/u). Finally, the amplitude of the function is the total number of particles divided by the square root of the flow rate. To verify this analytical equation, we conducted a Monte-Carlo simulation with diffusion and flow rates that were relevant to our HPLC column. There are additional sources of noise in the system. For instance, the analytes are spread over the length of the column, meaning they have different distances to travel to the detector. However, this simulation fit an exponential Gaussian equation almost exactly (Figure 2B, red line on blue histogram), meaning that the exponential Gaussian is theoretically a good distribution to model HPLC peaks.

Empirical Solution

To empirically test our theoretical exponential Gaussian distribution on real data, we injected single, pure fatty acid analytes (either linoleic acid, arachidonic acid, or docosahexaenoic acid) onto a C18 column, at concentrations of either 0.5mg/ml or 1 mg/ml. Analytes were eluted from the column on a 3:1 methanol water to acetonitrile gradient (see methods) adapted from[18]. We then used the Python package lmfit[19] to fit one of four distributions commonly used in chromatography to each of the underlying fatty acid peaks. A representative chromatograph of linoleic acid is shown to be fit to i). a Gaussian distribution[6], ii). an exponential Gaussian distribution, iii). a Voigt distribution[20] and iv). a skew Gaussian distribution[7] (Figure 2C). The error between the model and the underlying data was calculated using root-mean squared error (RMSE) after normalizing to total analyte area. The exponential Gaussian distribution had the significantly lowest RMSE for both concentrations of analytes (Figure 2D).

PeakClimber software package utilizes the exponential Gaussian function to rapidly and accurately identify chromatography peak areas

Current standard industry software from Thermofisher (chromeleon), Waters, and Agilent currently do not quanitify HPLC peak areas using fitting of theoretical peak functions. Other software, such as HappyTools or hplc-py from the Cremer lab at Stanford do not utilize the exponential Gaussian distribution to quantify peaks. Identifying this unmet need, we created PeakClimber, an algorithm and python package that identifies and quantifies individual peaks on a chromatographic trace by fitting a sum of exponential Gaussian functions to the HPLC trace. The workflow for PeakClimber is highlighted in Figure 3.

Taking a text file of the trace as input, PeakClimber will first denoise and detrend the data. Denoising is accomplished using a low pass FFT filter[21], as well as a time-averaging convolution. Detrending is accomplished with a high-pass Whitaker baseline subtraction algorithm that was developed for chromatography called peaked signal's asymmetric least squares algorithm (psalsa)[22]. Exact parameters for these detrending algorithms are inputted by the user. We choose our default values by fitting single peaks of real HPLC data (Figure 3-1). Peaks are then identified on the denoised data using scipy's peak finding algorithm that relies on prominence cutoffs to determine if peak is real or just local variation[23,24]. The prominence cutoff is also user defined, in PeakClimber but we used a value of 0.05 in this paper. Additional shoulder peaks can be optionally identified by identifying local minima and maxima in the first derivative of the HPLC trace that are close to 0 (Figure 3-2).

For each identified peak, an exponential Gaussian function is created with default parameters of the identified peak center, the identified peak height, a sigma of 0.1 minutes, and an exponential decay parameter of 2 using lmfit[19]. These functions are split into regions based on the zeros of the underlying trace (Figure 3-3). These regions of the graph are effectively independent of each other, meaning fits can be performed independently without loss of accuracy. Each group of Gaussians is fit to the appropriate window in the trace using a linear regression to minimize the least-squared distance between the generated sum of functions and the underlying trace (Figure 3-4). The algorithm recombines the fits for the different windows and returns a summary graph of the resulting fit, overlaid with individual peaks, as well as a table with peak number, runtime, and peak area. PeakClimber is currently available as a command line python package on GitHub (see methods for link).

Comparison of PeakClimber to other common HPLC algorithms

To test the utility of PeakClimber, we wanted to compare its performance to commercially available software. To do this, we injected three fatty acids with overlapping retention times: docosahexaenoic acid (12.3 minutes), arachidonic acid (12.5 minutes) and linoleic acid(12.9 minutes) at concentrations of 0.5 mg/ml. We also varied the concentrations of each of the analytes between either 0.5 or 1 mg/ml. Thus, in each injection, the analytes were either or equal concentration or one analyte was double the concentration of the other two (Figure 4A). This was done to test the dynamic range of PeakClimber compared to other algorithms. The raw HPLC trace was then smoothed and normalized before fitting the three peaks by one of four algorithms. PeakClimber is the algorithm previously described in this paper (Figure 4B), hplc.io[7] is a new, python-based, chromatographic fitting software that use skew-Gaussian functions as representative of single peaks (Figure 4C). Happytools is a free, standalone software package that uses Gaussian functions as representative of single peaks[6] (Figure 4D). Finally, valley-to-valley is an abstraction of algorithms[4,5] used by common HPLC-software such as Thermofisher's Chromeleon or Water's MassLynx that integrates the area under the curve of the trace between "valleys", the lowest points between two identified peaks (Figure 4E). Fits (black line Figure 4B-E) were performed on the entire trace (red line Figure 4B-E), but error comparisons were performed with individual peaks for each of the three analytes (lower panel Figure 4B-E, analytes are DHA, ARA, and LA from left to right) using RMSE. Fit peaks were recentered on the canonical single analyte peaks, as run times were shifted later in the co-injections. PeakClimber outperformed all other software regardless of peak position (Figure 4F), and the error that is present is unbiased based on peak position. When analyzing the first peak in the co-injection, the error present in PeakClimber's analysis is lower than the other algorithms and neutral to injection volume (Figure 4G) due to the correct fitting of the tail of the peaks. PeakClimber also performed better for the second and third peaks as well (Supplemental Figure 1 A & B). When error is calculated through percent error of the area, rather than RMSE, this pattern holds as well (Supplement Figure 1 C-F).

Limits of Peak Climber

All algorithms, including Peak Climber, fail to accurately find peak areas for groups of peaks under three separate circumstances: high signal-to-noise ratios, small distance between peaks, and uneven ratio between small and large peaks. To test these bounds specifically for PeakClimber, we ran simulations to generate synthetic data of overlapping exponential Gaussian functions and then attempted to calculate the area of these functions with PeakClimber and a naive valley-to-valley integration function. We simulated three exponential Gaussian peaks, two of which overlap, and the third that serves as a negative control (Supplemental Figure 2). To test the fitting ability of our algorithm, rather than the peak finding, which is already well-tested in other works[23], we provided the peak locations to each algorithm. The RMSE for each of these cases is shown in the respective subpanels of Figure 5.

For single peaks, PeakClimberslightly outperforms manual integration at noise with Gaussian amplitude above 0.3 times the peak size (Figure 5A). This is likely because PeakClimber captures the shape of the underlying Gaussian better than integration after denoising. Accuracy begins to drastically decrease with a 1-sigma distance between peak centers for PeakClimber. For the valley-to-valley method, RMSEs only drop around 10-sigma and significant error is seen with 3-sigma (Figure 5B). PeakClimber also displays no bias between peak position, whereas the first peak has higher-RMSE error in the valley-to-valley

method, even when peaks sizes are the same (Figure 5C). Interestingly, both algorithms display similar error rates across the whole range of peak ratios, suggesting that this is not an important factor driving errors in HPLC data (Figure 5C).

PeakClimber can be used to accurately quantify lipid differences between biological samples

One of the principal reasons we developed the PeakClimber was to analyze complex biological data with many overlapping peaks. We found that although industry software could accurately identify peaks, quantification was highly variable and inaccurate due to aforementioned problems with baseline correction and undercounting of the first peak in a series. To test the utility of PeakClimber on real biological data, we raised female *Drosophila melanogaster* on two different microbial conditions (germ-free or conventionally reared and standard diets. We then performed a lipid extraction, and then ran the isolated lipids on the HPLC, using a two-step methanol:water to acetonitrile, acetonitrile to isopropanol gradient to separate lipid species by polarity and size adapted from[18]. Chromatographs were analyzed using either the Thermofisher Chromeleon Software (Figure 6A) or PeakClimber (Figure 6B). Peaks were identified as significantly different between dietary or microbial conditions using PeakClimber (Figure 6C). These peaks, representing triglyceride compounds are more accurately quantified by PeakClimber as compared to Chromeleon, given their peak ratios to neighboring peak regions, with peak height as a secondary comparison.

Discussion

In this paper we have shown three principal findings. Firstly, the exponential Gaussian function is a good fit for HPLC peaks. We showed this both analytically by developing a model and solving it mathematically and with Monte-Carlo simulations, and empirically by calculating the error of the fit for various common distributions used in chromatography to single analyte peaks, and. Many previous works from the 1970s and 1980s attempt to analytically work out these solutions, and their models, while approximating an exponential Gaussian, do not converge to it[8–10]. Finally, HPLC peaks often do not represent single compounds, but groups of compounds. This means that a single peak is a sum of individual compounds all with behavior as described in **Figure 2**. Due to the central limit theorem, this would suggest that the chromatographic traces that we observe should have more of a Gaussian character, but this is not what we empirically observe.

Secondly, we demonstrated the effectiveness of PeakClimber as compared to other commercially and freely available software tools to quantitatively analyze chromatography data with overlapping peaks. This is due to the ability of our algorithm to capture the tail region of peaks early in a group of peaks, preventing undercounting. The package and documentation for PeakClimber are freely available on GitHub and easy to use with basic command line literacy. We plan to make a graphic user interface available as well to increase accessibility. One limitation of the test data of PeakClimber is that it was only used to test HPLC data from lipid chromatography. Theoretically other biomolecules should have the same kinetic and diffusive behaviors as lipids, and many chromatographic traces present in the literature show single

peaks as exponentially modified Gaussian distributions. However, adapting our algorithm to deal with anti-Langmuir fronted peaks could be a promising next step.

Thirdly, we demonstrated the utility of our algorithm for the analysis of biological data. While mass-spectrometry will always be the gold standard for metabolic analysis[25], HPLC represents a promising medium-throughput option to reduce reliance on mass-spec[26–28]. Consider an experiment similar to one that we've set up here with multiple replicates of different dietary, genetic, or microbial conditions. Rather than run each replicate through a mass-spec machine, one replicate from each group could be run through mass-spec, and the rest on HPLC, where relative changes in the compounds identified by MS could be much more accurately quantified with PeakClimber. The recognition of HPLC as this medium-throughput bridge between MS and high-throughput methods such as colorimetric kits could be one reason for the interest in development of tools to better analyze this type of data.

Limitations and comparison to other algorithms:

The exponential Gaussian function will not perfectly fit some chromatography peaks, as compounds that run with anti-Langmuir kinetics will have peak fronting, which will not fit an exponential Gaussian distribution which always has a tail. Peak fronting can also occur when the peak has been overloaded with analyte. In such cases, we would recommend usage of software tools like hplc.io[7], which uses a skew Gaussian that can account for peak fronting, or reducing injection volume. Additionally, our mathematical model made a number of simplifying assumptions about the geometry and flow rate of common HPLC systems that would lead to a more complicated distribution.

However, despite these limitations, we want to highlight the performance advantages of PeakClimber compared to other available software. As we have shown in this paper, PeakClimber much more accurately quantifies areas of overlapping peaks, due to its ability to consider long peak tails. For the analysis of biological data that contains a). many overlapping and b). non-overloaded peaks, we believe that PeakClimber will prove to be invaluable.

Materials and Methods

Monte-Carlo simulations

A column 1000 units long with 100 analyte particles at position 100 bound to the column is instantiated. At each time step the analyte has a 5% chance of unbinding from the column (representing a k value of 0.05). In the non-diffusive scenario, once unbound the particle arrives at the detector a fixed time later, in this case 900-time steps. In the diffusive case, the particle moves either forward or backward in the column as subject to diffusion before experiencing the fixed flow term. In our simulations the rate of diffusion is approximately ten times slower than the flowrate, and the column is the same length as in the simulations with no diffusion.

Fatty-acid chromatography

Fatty acid aliquots were obtained from Cayman Chemicals: linoleic acid (LA) (CAS 60-33-3), arachidonic acid (ARA) (CAS 506-32-1), and docosahexaenoic acid (DHA) (CAS 6217-54-5)

and resuspended in HPLC isopropanol to a stock concentrations of 10mg/ml. Aliquots were then further diluted to either 0.5mg/ml or 1 mg/ml as individual analytes or as part of one of the four mixtures analyzed (0.5:0.5:0.5, 1:0.5:0.5,0.5:1:0.5,0.5:0.5:1 mg/ml of DHA:ARA:LA respectively. 20ul of each individual analyte or mixture was injected onto the HPLC system. The components of each sample were separated and detected by an HPLC system using a LPG-3400RS quaternary pump, WPS-3000TRS autosampler (maintained at 20°C), TCC-3000RS column oven (maintained at 40°C), Accucore C18 column (150 × 3.0 mm, 2.6 µm particle size), FLD-3100 fluorescence detector (8 µl flow cell maintained at 45°C), and a Dionex Corona Veo charged aerosol detector (all from Thermo Fisher Scientific). Component peaks were resolved over an 30 min time range in a multistep mobile phase gradient as follows: 0–5 min = 0.8 ml/min in 98% mobile phase A (methanol-water-acetic acid, 750:250:4) and 2% mobile phase B (acetonitrile-acetic acid, 1,000:4); 5–30 min = 0.8–1.0 ml/min, 98–30% A, 2–44% B, and 0–3.3% mobile phase C (2-propanol).(HPLC-grade acetic acid and 2-propanol were purchased from Fisher Scientific and HPLC-grade methanol and acetonitrile were purchased from Sigma-Aldrich).

Error tolerance simulations

Noise simulation: a single exponential Gaussian peak was initialized with the following parameters: amplitude between 1 and 5, gamma (skew) between 2.9 and 3, and sigma between 0.1 and 0.2. Noise was added to the peak between 0 and 80% of its amplitude. Peak area was calculated using PeakClimber or manual integration after denoising and compared to the known area of the generated peak.

Proximity and ratio simulations: Three exponential Gaussian peaks were initialized with the following parameters: amplitude between 1 and 3, gamma (skew) between 1 and 3, and sigma between 0.05 and 0.1. In the proximity simulation the distance between peaks 1 and 2 varies as shown, while peak 3 is always ten units away. In the ratio simulation, peak 1 and 2 are a fixed 2.5 sigma apart, but the ratio between their amplitudes varies.

Fly husbandry

Drosophila melanogaster Canton-S flies were initially isolated from long term germ-free stocks kept in lab. The parental generation of flies was inoculated with a 7-species microbiome mixture consisting of *L. plantarum, L. brevis, A. pomorum, A. orientalis, A. cervisea, A. sicerae*, and *A. tropocalis* that recapitulates the microbiome found in a wild fruit fly 5 days after eclosion. Parental Flies were fed a diet consisting of 10% glucose and 5% yeast or 10% glucose, 5% and 5% cocoa butter. These flies were allowed to lay eggs on their food for 3 days. The resulting offspring were raised until 10 days post eclosion before lipid extractions.

Lipid extractions and chromatography for drosophila

Groups of 8 flies were macerated using a bead beater in 500uL of lipid extraction buffer (10mM Tris, 1mM EDTA, 7.8 pH). 400uL of extract was mixed with 1.5 mL 2:1 chloroform:methanol (with 1ng/mL of TopFluor cholesterol as an internal standard) and allowed to sit for 10 minutes. Then 500uL of chloroform followed by 500 uL of extraction buffer was added to the mixture. Mixture was centrifuged at 5000 rpm for 5 minutes and the organic(bottom) phase was separate. This was evaporated to dryness with a SpeedVac and then resuspended in 100uL of HPLC grade isopropanol.

20uL of the sample were injected onto the HPLC system. The components of each sample were separated and detected by an HPLC system using a LPG-3400RS quaternary pump, WPS-3000TRS autosampler (maintained at 20°C), TCC-3000RS column oven (maintained at 40°C), Accucore C18 column (150 × 3.0 mm, 2.6 µm particle size), FLD-3100 fluorescence detector (8 µl flow cell maintained at 45°C), and a Dionex Corona Veo charged aerosol detector (all from Thermo Fisher Scientific). Component peaks were resolved over an 80 min time range in a multistep mobile phase gradient as follows: 0–5 min = 0.8 ml/min in 98% mobile phase A (methanol-water-acetic acid, 750:250:4) and 2% mobile phase B (acetonitrile-acetic acid, 1,000:4); 5–35 min = 0.8–1.0 ml/min, 98–30% A, 2–65% B, and 0–5% mobile phase C (2-propanol); 35–45 min = 1.0 ml/min, 30–0% A, 65–95% B, and 5% C; 45–73 min = 1.0 ml/min, 95–60% B and 5–40% C; and 73–80 min = 1.0 ml/min, 60% B, and 40% C. (HPLC-grade acetic acid and 2-propanol were purchased from Fisher Scientific and HPLC-grade methanol and acetonitrile were purchased from Sigma-Aldrich.)

Code availability

All data and Jupyter Notebooks used to generate figures in this manuscript can be found at github.com/ATiredVegan/PeakClimberManuscriptRepository. The PeakClimber package and use instructions can be found at github.com/ATiredVegan/PeakClimber.

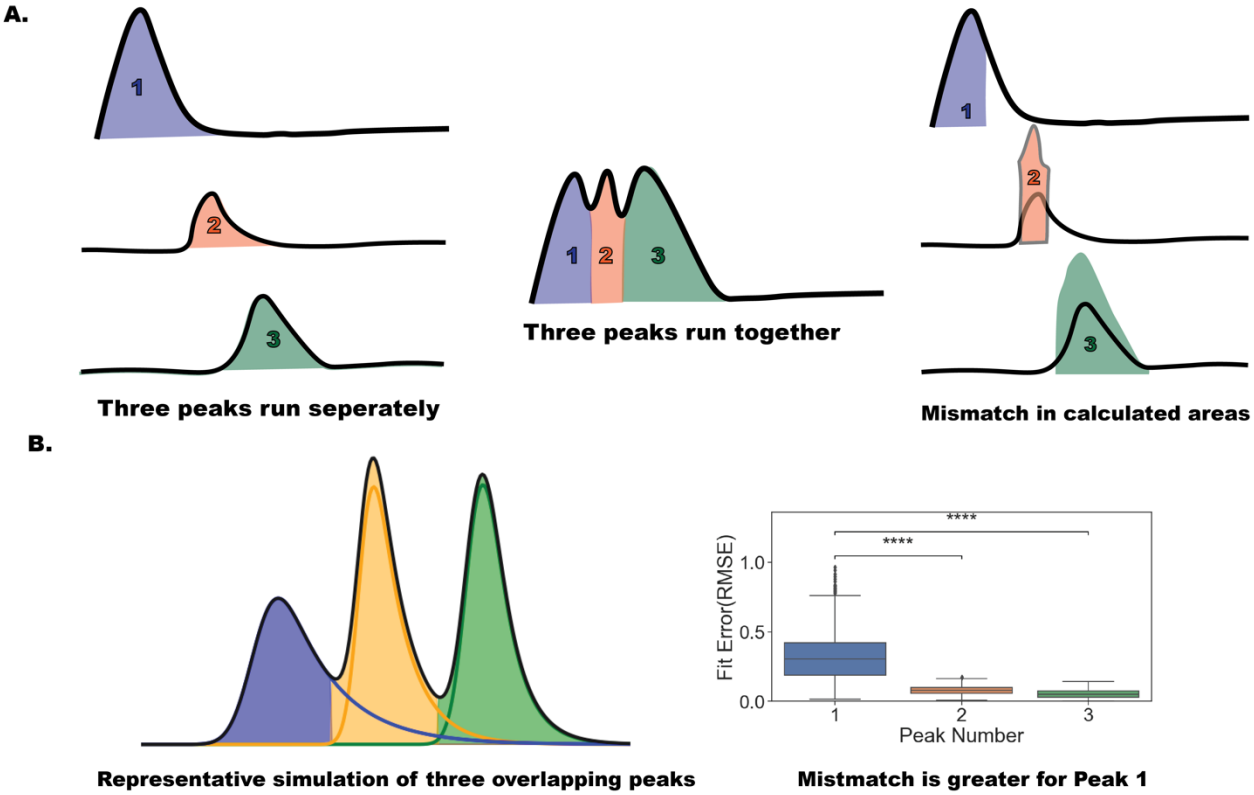

Figure 1: The problem of peak fitting. A). A cartoon depiction of a common error in peak fitting. When three analytes are run separately, their area is accurately calculated by peak integration. When this same operation is performed on the same analytes that are run together, common valley-to-valley area calculation will not accurately determine peak areas. B). A representative simulation of three overlapping peaks in blue, orange, and green. The shaped region represents the integration regions identified by the valley-to-valley algorithm; the solid lines represent the true peaks. The difference between the two is quantified by Root Mean Squared Error (RMSE) (n=1000, Mann-Whitney followed by Wilcoxon Ranked, ****p<10e-4).

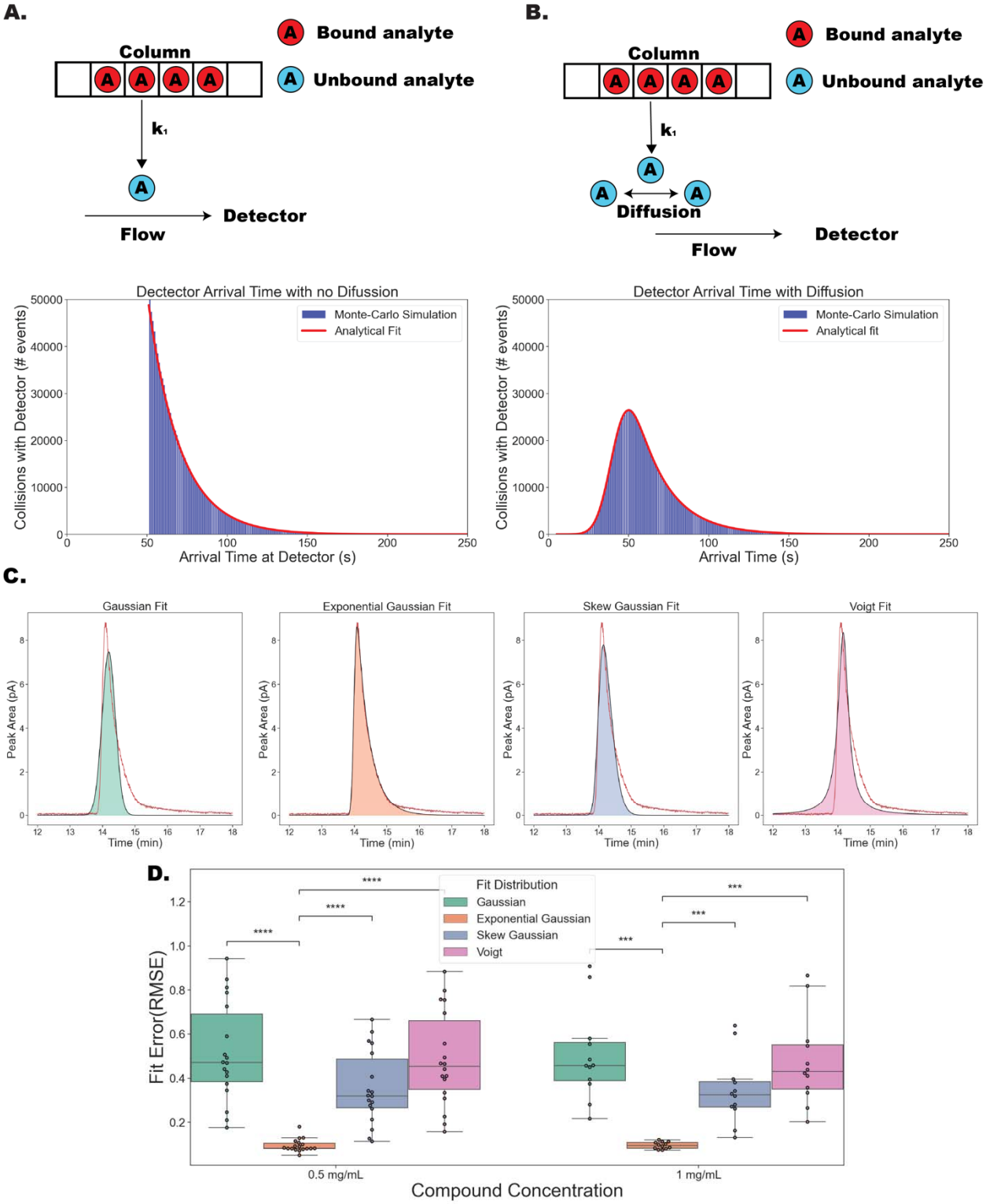

Figure 2: Individual HPLC analyte peaks correspond to an exponential Gaussian Distribution. (**A**). and (**B**). Monte Carlo simulation of solute arrival time at the detector without (**A**) or with (**B**) diffusion (blue) fit to an exponential (**A**) or exponential Gaussian (**B**) distribution (red) (n=1000 simulations consisting of 100 analyte molecules each). (**C**). Example empirical fits of (left to right) Gaussian, exponential Gaussian, skew Gaussian, and Voigt distributions to a single linoleic acid peak (**D**). Root mean squared error (RMSE) of fit of the fit as compared

each underlying chromatograph. Error is split by injection quantity (**: 1.00e-03<=p <= 1.00e-02, ***: 1.00e-04<=p <= 1.00e-03 ****: p <= 1.00e-04, Mann-Whitney U-test with Bonferroni correction, n=3 experimental, 4 technical replicates).

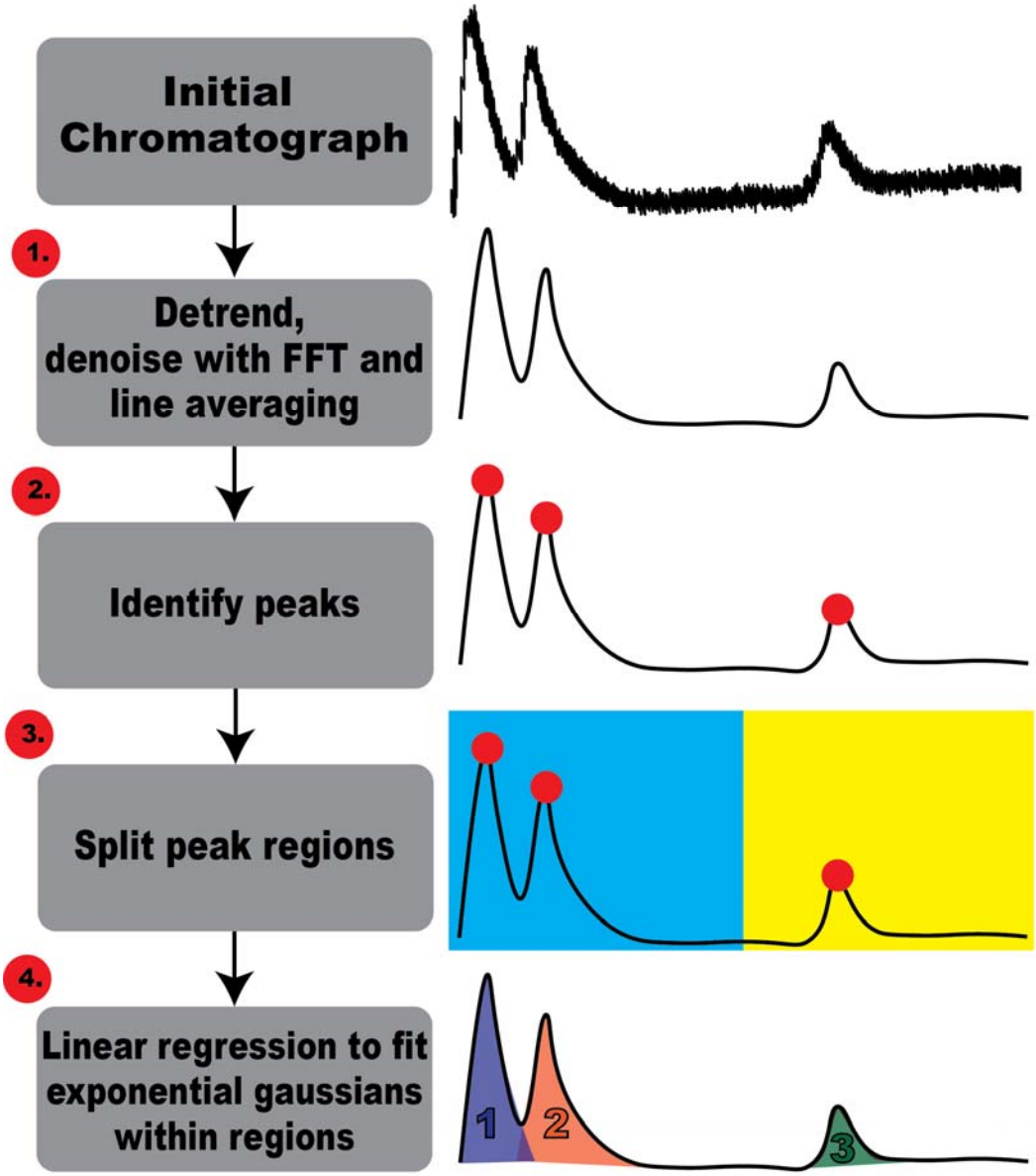

Figure 3: The Peak Climber Workflow Peak Climber will will first denoise (FFT) and detrend (whittaker psalsa) chromatography data (1) before identifying peaks using prominence cutoffs (2). To speed up runtime, peaks are split into regions based on intersections of the trace with the x-axis (3). Within each region, peaks are fit to an exponential Gaussian distribution (4).

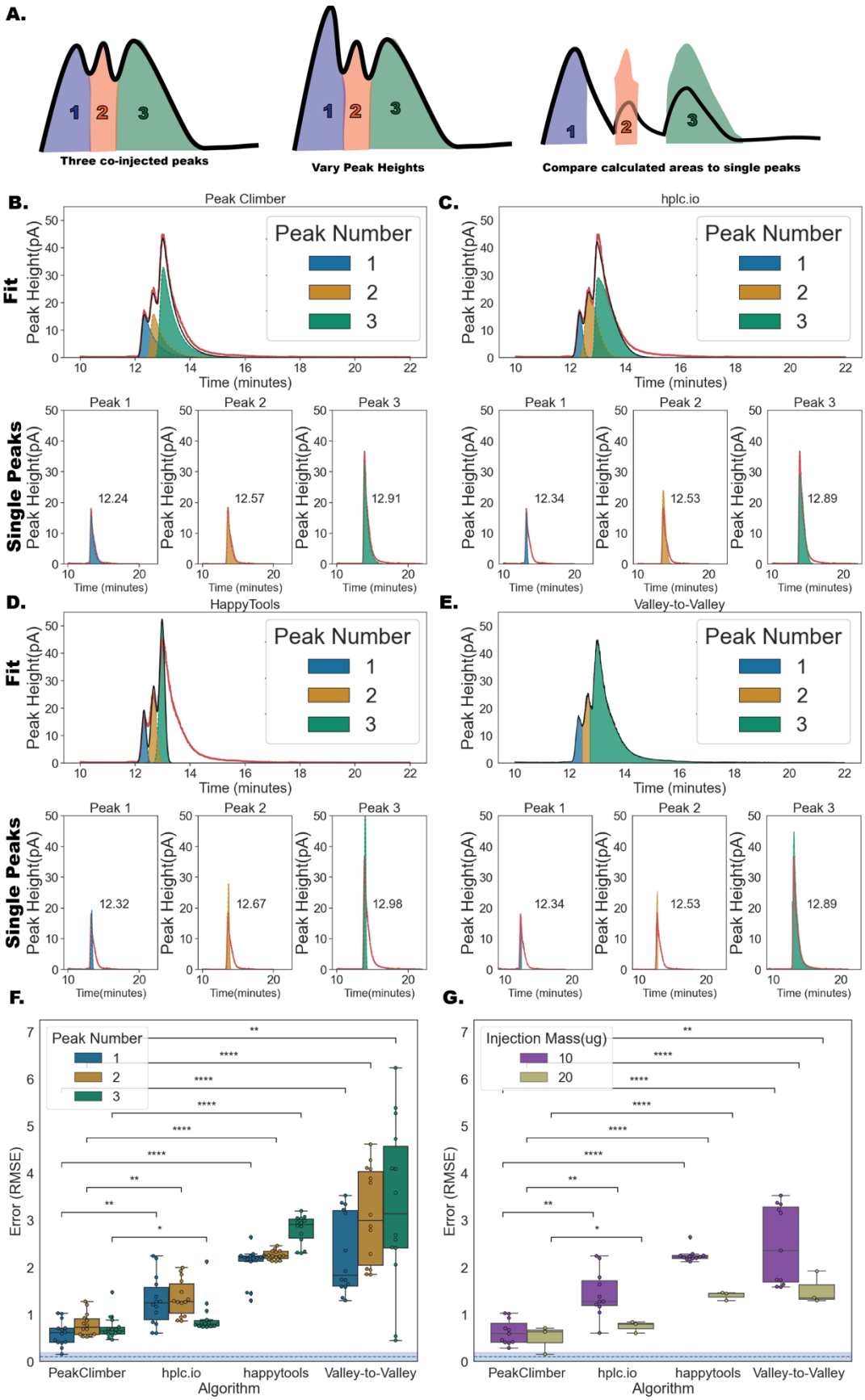
A.
1
2
3
Three co-injected peaks
Vary Peak Heights
Compare calculated areas to single peaks
B.
Peak Climber
C.
hplc.io
D.
HappyTools
E.
Valley-to-Valley
Fit
Single Peaks
Peak Height(pA)
Time (minutes)
Peak Number
Peak 1
Peak 2
Peak 3
12.24
12.57
12.91
12.34
12.53
12.89
12.32
12.67
12.98
F.
G.
Error (RMSE)
Algorithm
PeakClimber
hplc.io
happytools
Valley-to-Valley
Injection Mass(ug)
10
20

Figure 4: PeakClimber Outperforms and is Unbiased Compared to Industry Software: (**A**). Schematic of the experiment performed in this figure. Three overlapping fatty-acid peaks were injected at either a ratio of 1:1:1, 1:2:1, 2:1:1, or 1:1:2. The calculated areas (using the algorithms listed below were compared to the real injected area of a single peak Representative fits of a coinjection of a chromatograph of C18:1, C20:4 and C22:6 elution in a 1:2:1 ratio by Peak Climber (**B**) hplc.io (**C**) Valley-to-Valley (Thermofisher and Waters) (**D**) and HappyTools (**E**). Bottom subpanel shows fit to each individual peak. (**F**). Quantification of error rates by RMSE of pooled coinjections of C18:1, C20:4 and C22:6 depending on peak position by above algorithms. Blue doted line represents minimum RMSE error obtained from single-peak fits (**G**). Quantification of error rates pooled coinjections of C18:1, C20:4 and C22:6 depending on peak size by above algorithms on peak 1 alone. (n=12, 3 experimental replicates with 4 samples per replicate, **: 1.00e-03 < p <= 1.00e-02,***: 1.00e-04 < p <= 1.00e-03, ****: p <= 1.00e-04 Kruskal-Wallis Test with Bonferroni correction for subpanels E and F, n=3 experimental, 4 technical replicates).

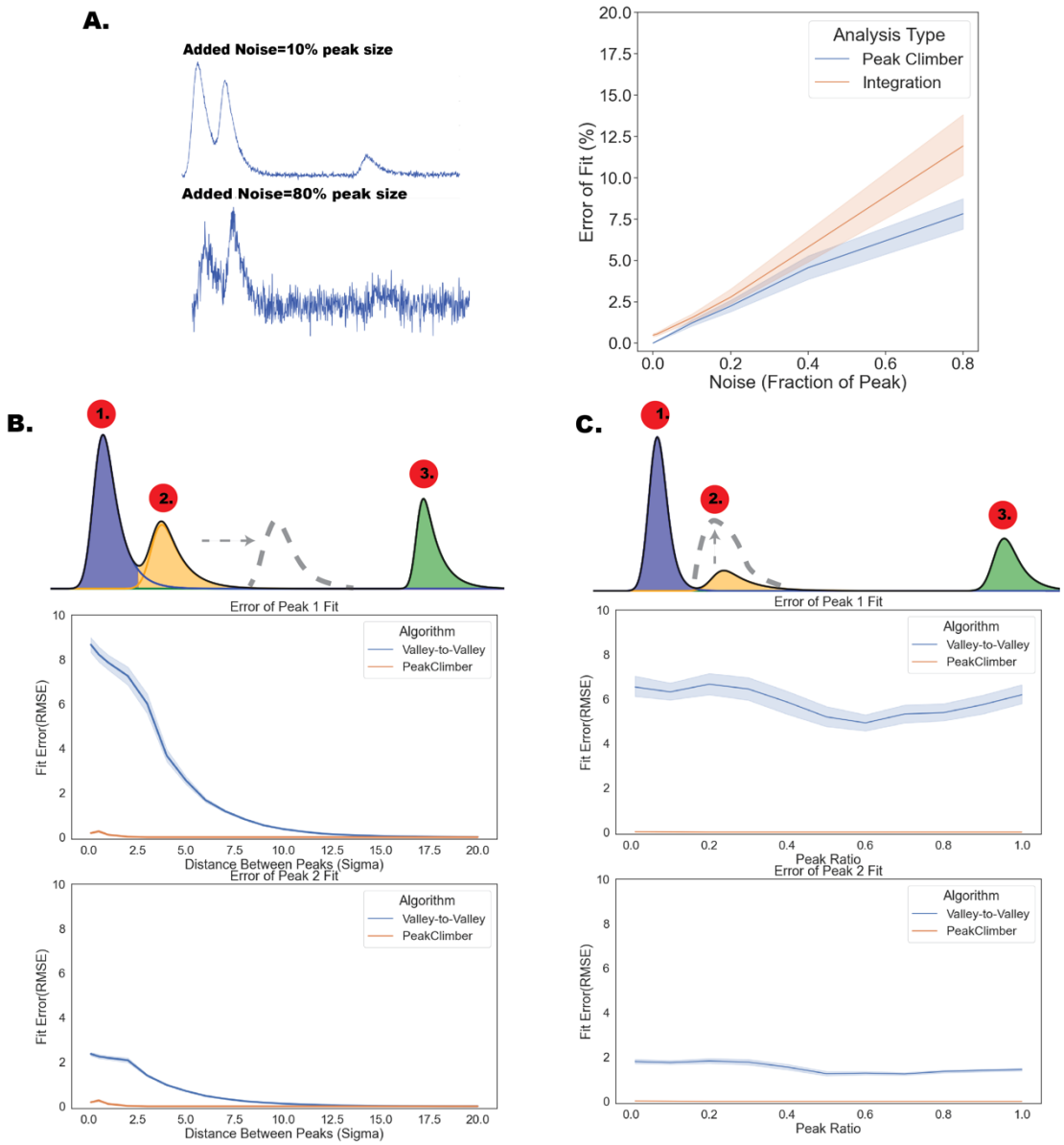

Figure 5: PeakClimber Error Tolerances: (**A**). Gaussian noise as a fraction of the peak amplitude is added to a single synthetically generated peak with amplitude between 1 and 2, sigma between 0.1 and 0.2, gamma between 2.9 and 3 and center 1. This noise is detrended and removed using the PeakClimber algorithm and then the resulting peak area is either found through fitting (PeakClimber) or integration with the valley-to-valley method and compared

to the known area (**B**). Using the same parameters for exponential Gaussians as in subpanel A, three exponential Gaussians are generated. The distance between exponential Gaussian 1 and 2 is shown on the x axis, exponential Gaussian 3 is always 5 units away from exponential Gaussian 2. the resulting peak area is either found through fitting or integration and compared to the known area. (**C**). Using the same parameters for exponential Gaussians as in subpanel A, three exponential Gaussians were generated. The distance between exponential Gaussian 1 and 2 is always 2 sigma, exponential Gaussian 3 is always 5 units away from exponential Gaussian 2. The ratio of areas between exponential Gaussians 1 and 2 is shown on the x-axis, the resulting peak area is either found through fitting or integration and compared to the known area. Exponential Gaussian 3 has the same area as exponential Gaussian. 1.

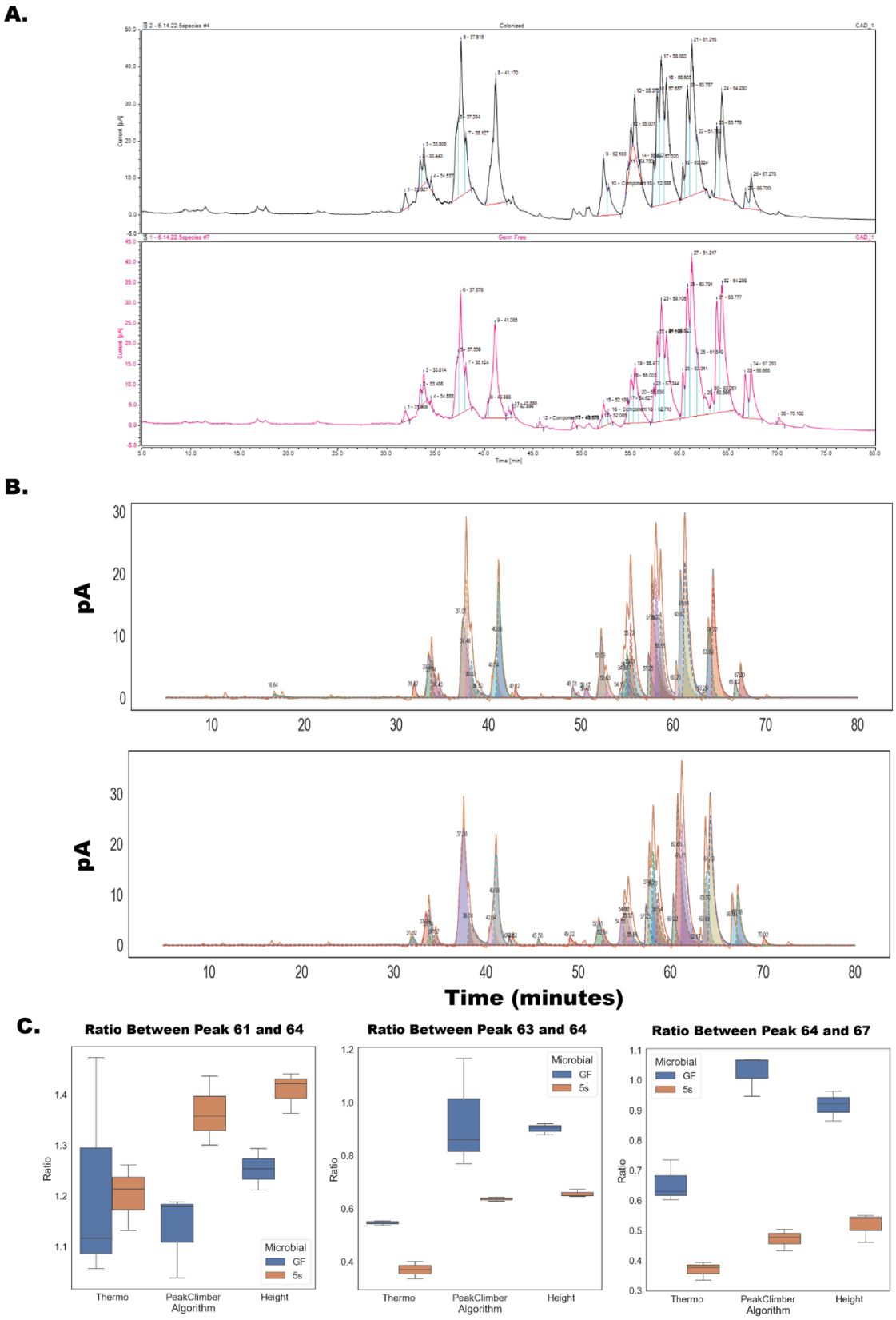

Figure 6: PeakClimber more accurately quantifies biological differences between germ-free and colonized flies: A). Thermofisher Chromeleon quantification of a sample colonized (black) and germ-free

(pink) fly lipid profiles B). Quantitation of the same lipid profiles using PeakClimber. C). Quantitation of peak ratios between peaks that run at 61:64 minutes, 63/64 minutes and 64/67 minutes using Chromeleon, PeakClimber, and peak height between germ-free and colonized flies (n=3,8).

References

(1) Morley, R.; Minceva, M. Liquid–Liquid Chromatography: Current Design Approaches and Future Pathways. *Annu. Rev. Chem. Biomol. Eng.* **2021**, *12* (1), 495–518. https://doi.org/10.1146/annurev-chembioeng-101420-033548.

(2) Ovbude, S. T.; Sharmeen, S.; Kyei, I.; Olupathage, H.; Jones, J.; Bell, R. J.; Powers, R.; Hage, D. S. Applications of Chromatographic Methods in Metabolomics: A Review. *J. Chromatogr. B* **2024**, *1239*, 124124. https://doi.org/10.1016/j.jchromb.2024.124124.

(3) Ito, Y.; Weinstein, M.; Aoki, I.; Harada, R.; Kimura, E.; Nunogaki, K. The Coil Planet Centrifuge. *Nature* **1966**, *212* (5066), 985–987. https://doi.org/10.1038/212985a0.

(4) Dyson, N. A. *Chromatographic Integration Methods*; Royal Society of Chemistry, 1998; Vol. 3.

(5) Sternberg. Biomedical Image Processing. *Computer* **1983**, *16* (1), 22–34. https://doi.org/10.1109/MC.1983.1654163.

(6) Jansen, B. C.; Hafkenscheid, L.; Bondt, A.; Gardner, R. A.; Hendel, J. L.; Wuhrer, M.; Spencer, D. I. R. HappyTools: A Software for High-Throughput HPLC Data Processing and Quantitation. *PLOS ONE* **2018**, *13* (7), e0200280. https://doi.org/10.1371/journal.pone.0200280.

(7) Chure, G.; Cremer, J. Hplc-Py: A Python Utility For Rapid Quantification ofComplex Chemical Chromatograms. *J. Open Source Softw.* **2024**, *9* (94), 6270. https://doi.org/10.21105/joss.06270.

(8) Amundson, N. R. The Mathematics of Adsorption in Beds. III. Radial Flow. Leon,Lapidus. *J. Phys. Colloid Chem.* **1950**, *54* (6), 821–829. https://doi.org/10.1021/j150480a011.

(9) Golshan-Shirazi, Sadroddin.; Guiochon, Georges. Analytical Solution for the Ideal Model of Chromatography in the Case of a Langmuir Isotherm. *Anal. Chem.* **1988**, *60* (21), 2364–2374. https://doi.org/10.1021/ac00172a010.

(10) Frey, G. L.; Grushka, E. Numerical Solution of the Complete Mass Balance Equation in Chromatography. *Anal. Chem.* **1996**, *68* (13), 2147–2154. https://doi.org/10.1021/ac960220o.

(11) Langmuir, I. THE CONSTITUTION AND FUNDAMENTAL PROPERTIES OF SOLIDS AND LIQUIDS. PART I. SOLIDS. *J. Am. Chem. Soc.* **1916**, *38* (11), 2221–2295. https://doi.org/10.1021/ja02268a002.

(12) Seelinger, F.; Wittkopp, F.; Von Hirschheydt, T.; Frech, C. Anti-Langmuir Elution Behavior of a Bispecific Monoclonal Antibody in Cation Exchange Chromatography: Mechanistic Modeling Using a pH-Dependent Self-Association Steric Mass Action Isotherm. *J. Chromatogr. A* **2023**, *1689*, 463730. https://doi.org/10.1016/j.chroma.2022.463730.

(13) Williamson, Y.; Davis, J. M. Modeling of Anti-Langmuirian Peaks in Micellar Electrokinetic Chromatography: Benzene and Naphthalene. *ELECTROPHORESIS* **2005**, *26* (21), 4026–4042. https://doi.org/10.1002/elps.200500245.

(14) Naish, P. J.; Hartwell, S. Exponentially Modified Gaussian Functions—A Good Model for Chromatographic Peaks in Isocratic HPLC? *Chromatographia* **1988**, *26* (1), 285–296. https://doi.org/10.1007/BF02268168.
(15) Grushka, Eli. Characterization of Exponentially Modified Gaussian Peaks in Chromatography. *Anal. Chem.* **1972**, *44* (11), 1733–1738. https://doi.org/10.1021/ac60319a011.
(16) Lan, K.; Jorgenson, J. W. A Hybrid of Exponential and Gaussian Functions as a Simple Model of Asymmetric Chromatographic Peaks. *J. Chromatogr. A* **2001**, *915* (1–2), 1–13. https://doi.org/10.1016/S0021-9673(01)00594-5.
(17) Chandrasekhar, S. Stochastic Problems in Physics and Astronomy. *Rev. Mod. Phys.* **1943**, *15* (1), 1–89. https://doi.org/10.1103/RevModPhys.15.1.
(18) Quinlivan, V. H.; Wilson, M. H.; Ruzicka, J.; Farber, S. A. An HPLC-CAD/Fluorescence Lipidomics Platform Using Fluorescent Fatty Acids as Metabolic Tracers. *J. Lipid Res.* **2017**, *58* (5), 1008–1020. https://doi.org/10.1194/jlr.D072918.
(19) Newville, M.; Stensitzki, T.; Allen, D. B.; Ingargiola, A. LMFIT: Non-Linear Least-Square Minimization and Curve-Fitting for Python, 2014. https://doi.org/10.5281/ZENODO.11813.
(20) Reilly, J. T.; Walsh, J. M.; Greenfield, M. L.; Donohue, M. D. Analysis of FT-IR Spectroscopic Data: The Voigt Profile. *Spectrochim. Acta Part Mol. Spectrosc.* **1992**, *48* (10), 1459–1479. https://doi.org/10.1016/0584-8539(92)80154-O.
(21) Brigham, E. O.; Morrow, R. E. The Fast Fourier Transform. *IEEE Spectr.* **1967**, *4* (12), 63–70. https://doi.org/10.1109/MSPEC.1967.5217220.
(22) Oller-Moreno, S.; Pardo, A.; Jimenez-Soto, J. M.; Samitier, J.; Marco, S. Adaptive Asymmetric Least Squares Baseline Estimation for Analytical Instruments. *2014 IEEE 11th Int. Multi-Conf. Syst. Signals Devices SSD14* **2014**, 1–5. https://doi.org/10.1109/SSD.2014.6808837.
(23) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; Van Der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, İ.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; Van Mulbregt, P.; SciPy 1.0 Contributors; Vijaykumar, A.; Bardelli, A. P.; Rothberg, A.; Hilboll, A.; Kloeckner, A.; Scopatz, A.; Lee, A.; Rokem, A.; Woods, C. N.; Fulton, C.; Masson, C.; Häggström, C.; Fitzgerald, C.; Nicholson, D. A.; Hagen, D. R.; Pasechnik, D. V.; Olivetti, E.; Martin, E.; Wieser, E.; Silva, F.; Lenders, F.; Wilhelm, F.; Young, G.; Price, G. A.; Ingold, G.-L.; Allen, G. E.; Lee, G. R.; Audren, H.; Probst, I.; Dietrich, J. P.; Silterra, J.; Webber, J. T.; Slavič, J.; Nothman, J.; Buchner, J.; Kulick, J.; Schönberger, J. L.; De Miranda Cardoso, J. V.; Reimer, J.; Harrington, J.; Rodríguez, J. L. C.; Nunez-Iglesias, J.; Kuczynski, J.; Tritz, K.; Thoma, M.; Newville, M.; Kümmerer, M.; Bolingbroke, M.; Tartre, M.; Pak, M.; Smith, N. J.; Nowaczyk, N.; Shebanov, N.; Pavlyk, O.; Brodtkorb, P. A.; Lee, P.; McGibbon, R. T.; Feldbauer, R.; Lewis, S.; Tygier, S.; Sievert, S.; Vigna, S.; Peterson, S.; More, S.; Pudlik, T.; Oshima, T.; Pingel, T. J.; Robitaille, T. P.; Spura, T.; Jones, T. R.; Cera, T.; Leslie, T.; Zito, T.; Krauss, T.; Upadhyay, U.; Halchenko, Y. O.; Vázquez-Baeza, Y. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17* (3), 261–272. https://doi.org/10.1038/s41592-019-0686-2.

(24) Kirmse, A.; De Ferranti, J. Calculating the Prominence and Isolation of Every Mountain in the World. *Prog. Phys. Geogr. Earth Environ.* **2017**, *41* (6), 788–802. https://doi.org/10.1177/0309133317738163.
(25) Schrimpe-Rutledge, A. C.; Codreanu, S. G.; Sherrod, S. D.; McLean, J. A. Untargeted Metabolomics Strategies—Challenges and Emerging Directions. *J. Am. Soc. Mass Spectrom.* **2016**, *27* (12), 1897–1905. https://doi.org/10.1007/s13361-016-1469-y.
(26) Kao, D. J.; Lanis, J. M.; Alexeev, E.; Kominsky, D. J. HPLC-Based Metabolomic Analysis of Normal and Inflamed Gut. In *Gastrointestinal Physiology and Diseases*; Ivanov, A. I., Ed.; Methods in Molecular Biology; Springer New York: New York, NY, 2016; Vol. 1422, pp 63–75. https://doi.org/10.1007/978-1-4939-3603-8_7.
(27) Perez De Souza, L.; Alseekh, S.; Scossa, F.; Fernie, A. R. Ultra-High-Performance Liquid Chromatography High-Resolution Mass Spectrometry Variants for Metabolomics Research. *Nat. Methods* **2021**, *18* (7), 733–746. https://doi.org/10.1038/s41592-021-01116-4.
(28) Jacob, M.; Lopata, A. L.; Dasouki, M.; Abdel Rahman, A. M. Metabolomics toward Personalized Medicine. *Mass Spectrom. Rev.* **2019**, *38* (3), 221–238. https://doi.org/10.1002/mas.21548.