Picky with peakpicking: assessing chromatographic peak quality with simple metrics in metabolomics

William Kumler

2023-04-20

## Abstract

Chromatographic peakpicking continues to represent a significant bottleneck in automated LC-MS workflows. Uncontrolled false discovery rates and the lack of manually-calibrated quality metrics require researchers to visually evaluate individual peaks, requiring large amounts of time and breaking replicability. Here, we manually labeled two oceanographic particulate metabolite datasets to assess the performance of individual peak quality metrics and constructed a predictive model calibrated to the likelihood that visual inspection by an MS expert would include a given mass feature in the downstream analysis. A simple logistic regression model built on two metrics calculated from the raw MS data reduced the fraction of false positives in the analysis from 80-90% down to 1-5% and showed minimal overfitting when applied to novel datasets. We then explored the implications of this quality thresholding on the conclusions obtained by the downstream analysis, concluding that the poor performance of peakpicking algorithms significantly reduces the power of both univariate and multivariate statistical analyses in detecting environmental differences.

## Introduction

Liquid chromatography-mass spectrometry (LC-MS) is a powerful tool for exploring the molecular composition of biological samples. Its rapid sample processing (typically <1 hr run time), low limits of detection (pM-nM range), and ability to characterize novel molecules via fragmentation fingerprints make it a common workhorse for metabolomic research. However, much of the data produced by these instruments goes unused because traditional workflows target only a fraction of the compounds detected based on an *a priori* list of known molecules. Significant efforts in the past two decades have established workflows for untargeted metabolomics [@xcms, @mzmine, @msdial], but these data-driven methods are still handicapped by the poor performance of the core peakpicking algorithms and require manual oversight and curation of the datasets after they’re produced by the software. This peakpicking step is typically the very first in the data processing stage, with implications for all downstream analysis and interpretation.

Peakpicking is a well-known problem across many fields of science and especially in LC-MS analysis due to the smooth transition from high-quality peaks into background noise as the limit of detection is approached and crossed, with background signals often sharing many features of high-quality peaks. The imperfection of these algorithms introduces a fundamental tradeoff between false positives and false negatives, where false positives (Type I errors) represent background instrument or chemical noise that has been misclassified as true biological signal and false negatives (Type II errors) indicating biological signals that are misclassified as noise. Existing algorithms tend to be very forgiving of Type I errors and very harsh on Type II, presenting many false positives to the researcher in their attempt to absolutely minimize the number of real molecules missed. This approach makes sense because downstream analyses can always remove additional mass features, but Type II errors cannot be later recovered. However, we would argue that this emphasis on an absolute minimization of the missed peaks does a disservice to scientists for two reasons. First, the abundance of false positives means that peakpicking output cannot be automatically trusted. This mandates manual evaluation of the raw data, a time-consuming and mind-numbing process with a large degree of subjective decision making that cannot be perfectly replicated and scales combinatorially with the number of samples and compounds measured. Second, the presence of false negatives cannot be fully accounted for in the way the false positive rate can. False negatives are introduced in every step of metabolomic analyses, with compounds lost during sample processing, injection, and detection on the instrument. This is a well-known problem in metabolomics and will not be solved by any perfecting of peakpicking algorithms. Instead, emphasis should be placed on allowing the experimenter to set an acceptable threshold for the false discovery rate and accepting that this will inherently add to the number of peaks already lost in the data collection process.

Existing peak-detection softwares do not provide a clear way to do this. Typical outputs consistent across the different implementations consist of the *m/z* ratio, retention time, and area for each molecular feature, with some additional useful information occasionally provided such as the peak’s signal-to-noise ratio or degree of skew which are rarely estimated using a well-documented algorithm. None of these parameters answer the critical question of interest: what is the likelihood that this subset of data corresponds to an actual molecule present in the original sample? This parameter is crucial for downstream analysis because it represents the base rate for error propagation and acceptable thresholds should vary widely by the particular project’s goals. In an exploratory analysis, any molecular feature more than 50% likely to be real is perhaps worth considering, while in a confirmatory study this threshold may need to be above 99%. While significant effort has been recently invested into improvements in the peakpicking algorithms, very little has been done to quantify the accuracy and precision of their outputs on the wide variety of datasets from which they are determined.

A single metric of peak quality calibrated to its likelihood of representing an actual molecule has multiple advantages over the current heuristic approaches. First, providing an intuitive metric of peak goodness allows researchers to focus their energy on those features least likely to be noise and most likely to be found in a repeated analysis. This applies both during the manual curation phase in which the very best or most interesting peaks can be visually inspected while those of poor quality can be discarded, as well as in the compound discovery/characterization steps where time and energy spent attempting to determine the structure or molecular formula of a noise signal is wasted. Second, removing low-quality noise signals from the dataset prior to downstream analysis should improve statistical power by reducing the number of hypotheses tested for univariate analyses and increasing the relative power of significant peaks in multivariate analyses. Third, such a metric would allow optimization of peakpicking and chromatographic parameters for untargeted datasets. Rather than measuring the effect of a change on a small subset of known molecules or by counting the total number of peaks found, one could look at the metric’s distribution to see whether the change was an improvement in the number and absolute quality of the peaks detected. Fourth, high quality peaks should be independent of software implementations and should be found no matter what algorithms are used, relieving the scientist (and reviewers!) of some of the burden of learning new software or programming languages to check their work. Fifth, this metric would enable more consistent quality control between labs and analyists. Rather than discussing a plethora of parameters or making subjective assessments of peak shape, an idealized metric would provide a simple and intuitive threshold upon which to agree for a given project. This metric would be scripted and provide consistent, reproducible results independent of the particular expert reviewing its performance.

A single comprehensive metric also has significant advantages over multiple individual thresholds. First, a metric calibrated to likelihood is easier and more intuitive to use for classification than relatively arbitrary thresholds of signal-to-noise or peak skewness with ambiguous units and unintuitive interpretations. Second, a single compound metric also has the multivariate advantage over independent thresholds by allowing a good peak to be weak in one area which is compensated for by strong performance in other metrics. Sometimes good peaks do have high SNR or small areas, and thresholding on these metrics independently unnecessarily increases the likelihood of missing signals. Finally, the relative power of these many different metrics is difficult to compute and hard to compare. Is it more important to have an isotope, or a Gaussian peak shape? Is it more likely to be a biological signal if it has a large area, or if it differs between treatment? If so, by what degree? These metrics differ in strength depending on the dataset, but many MS experts have some degree of heuristic knowledge about their relative power for the data they’ve worked with and analyzed.

An area particularly ripe for LC-MS analysis is that of the open ocean. Low compound and high salt concentrations make this area difficult to study [@BMIS] but its vast size and direct effect on the Earth’s carbon cycling make it critical that we understand its dynamics of energy and nutrients on a molecular scale. Metabolites are the currency of chemical exchange both intra- and inter-cellularly, serving as building blocks of larger molecules, regulators of osmotic balance and storage of nutrients, as well as chemical signals on their own. These small molecules serve both as signposts for the complex biological landscape in this highly dynamic region and give a sense of not only who is present but also what ecological roles they’re serving and the niches they fill.

In this paper, we use open ocean marine metabolite LC-MS samples to develop and test a variety of chromatographic peak metrics. Here, we construct and validate multiple predictive models of peak quality based on metrics both common in the literature and custom implementations we’ve found useful in our own analysis. This allows us to connect the physical, chemical, and biological measurements taken regularly around the globe to a molecular-scale perspective of particular organic matter in the ocean by linking the chemical currencies that fuel the planet to the environments in which they’re found.