Machine learning to reduce manual correction of LC-MS peak integrations

Lyle Burton and Gillian Brooks
SCIEX, 71 Four Valley Drive, Concord, ON, Canada, L4K 4V8

ABSTRACT

Machine learning improves the correctness of reported peak integrations; this differs from approaches which simply flag peaks needing manual review.

INTRODUCTION

LC-MS and LC-MS/MS are becoming increasingly important in a broad array of application areas. A key step in these workflows is integration of the resulting chromatographic peaks to generate areas which are used for absolute and relative quantitation of the various target analytes. The state of the art is such that is not uncommon for these chromatographic peak integration algorithms to mis-integrate some of the peaks, leading to time consuming manual review and correction. Various rule-based approaches have been used to flag the subset of integrations potentially requiring attention, however this does not reduce the number of integrations needing to be corrected. Here we present a machine learning approach which improves the actual quality of the first-pass integration.

MATERIALS AND METHODS

As discussed above, current LC-MS peak integration algorithms are prone to mis-integrations. These algorithms all expose various settings which control the exact way in which peaks are integrated. It is usually possible to manually adjust these settings to obtain an acceptable integration for problematic cases – only very rarely is fully manual integration by drawing a baseline required.

Based on this observation, we developed an approach which integrates all chromatograms with multiple different combinations of peak-finding settings – enough to fully explore the parameter space. For each of these integrations various features are extracted and a machine learning model was trained using the XGBoost algorithm from the python sk-learn package [1]. Each integration is then scored using the model and the one with highest score is proposed.

The following data sets were used:

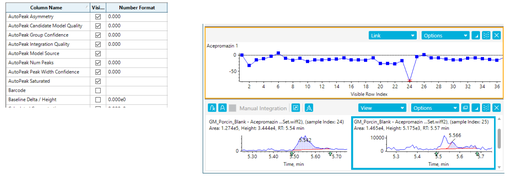
- ‘VitD’ – one batch from a Vitamin D assay with four analytes (Vit D2 and D3 each with one confirmation ion) and containing 142 samples. These chromatograms are generally quite clean with minimal interference.
- ‘MRM pesticide 1’ – a batch of 29 samples screening (scheduled MRM) for 196 pesticides; the batch includes both standards and pesticides spiked at various levels into fruit matrices. Some of the peaks are quite challenging to integrate.
- ‘MRM pesticide 2’’ – similar to the previous data set with 67 samples and 46 target pesticides.
- ‘TOF pesticide’ – a small batch of 18 samples with 184 target pesticides. Quantification was done by TOF MS1 (with MS/MS confirmation); a reasonable number of interferences were present.
- ‘Large panel MRM’ – Similar to the MRM batches above, with 56 samples and 1290 target pesticides. This was the most challenging batch due to the large number of analytes, some of which were in fact below the detection limit.
- ‘ISD’ – An MRM assay for three immunosuppressive drugs. There are 24 batches acquired at roughly equal intervals from March 2016 to June 2016 and five additional batches from February 2021; each batch contained an average of about 30 samples.

In addition to the experimental data listed above, an attempt was made to simulate a Vitamin D assay by generating synthetic chromatograms with additional interferences of various intensity and proximity to the main peak of interest and different noise levels. The idea was that the expected peak areas were exactly known and did not need to be determined by manual curation.

AutoPeak integration algorithm

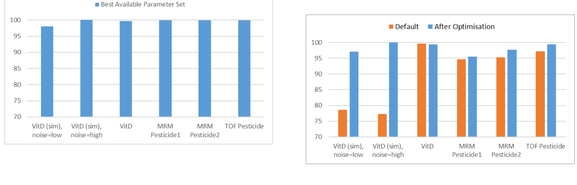
All LC-MS peak integration algorithms report essential information such as the retention time and peak area as well as additional metrics such as the peak width (usually at various height percentages). These metrics can be used to create rules to flag chromatograms with potentially incorrect integration (e.g. all peaks with retention time differing from the expected or average by more than a certain amount, etc.). This enables a ‘review by exception’ workflow where only flagged peaks are manually reviewed and their integrations corrected as needed. This is clearly a useful workflow, however setting up such rules can be quite time consuming since they must be sufficiently sensitive to flag all (or almost all) incorrect integrations without too many false positives. Additionally, these rules do not actually correct the root cause of the problem or correct the integration – this is still done by manually adjusting parameters.

The AutoPeak integration algorithm uses a peak modelling approach whereby an analytical model (i.e. an equation) or peak shape is first determined, usually from a standard sample [2,3]. This model is then fit to the other chromatograms modifying (stretching) it as required, *however* constraints are applied so that the modifications are only within the range expected due to variations in the chromatography. Since the algorithm performs actual peak fitting, it can computer a number of additional metrics compared to more traditional algorithms which basically measure how similar the peak shape is to the original model peak. These metrics are especially useful for evaluating the quality of the integration.



In the figure a list of some of these metrics is shown. The right portion of the figure shows a metric plot (top) of the quality for 36 different samples for the same analyte. In this case, one sample (bottom right) was *intentionally* manually mis-integrated to demonstrate that it appears as an outlier in the plot. This is an extreme example in which a classic ‘rule’ would be likely to find the problem, however in more subtle cases the power of machine learning can be used to find such problems using all available metrics.

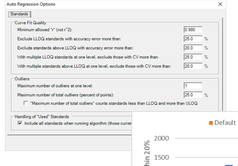
The figure below (left) shows the percentage of correct integrations which are available for each chromatogram in the parameter space explored. This demonstrates that a correct integration is almost always available meaning that it’s potentially discoverable by machine learning (i.e. the necessity for true manual integration by drawing a baseline is very rare).

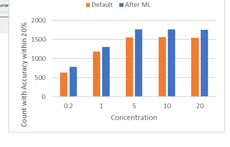


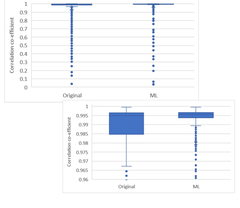
Initial results

The figure above (right) shows the initial default peak integration (orange) and the result of the machine learning optimisation (blue). The y-axis displays the percentage of integrations which were within 20% of the expected value. For the simulated data the ‘expected’ result is the known simulated peak area and for the experimental data is from careful manual curation. Results from six different data sets are shown; the first two are the simulated chromatograms and the others are the experimental data sets.

The most striking observation is the large increase in correct integration for the simulated data sets. In truth, this reflects the fact that these data are unrealistically complex with too many nearby large interferences – nonetheless, it demonstrates the potential improvement for complex data sets. Moderate improvements were seen for the other data sets, although the Vitamin D assay was sufficiently clean that there is in fact minimal room for further improvements.





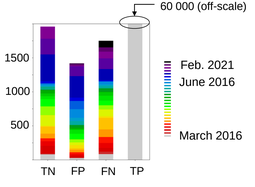


Large panel MRM results

Calibration curves using the peak areas from the default integrations and those after the machine learning procedure were created. This was done automatically using the outlier rejection settings shown in the left above – briefly, a small number of points were allowed to be automatically removed for both standards and (not shown) QCs so that the acceptance criteria were met.

Results were scored for the standards only using the calculated – since the expected concentration is known this allows evaluation without the subjectivity of manual curation of the data. The figure above (middle) shows the numbers of calculated concentrations which were within an acceptable 20% accuracy for these standards. Orange bars show the result for the default integration and blue for the optimised machine learning results. As can be seen at all concentration levels the number of passing integrations was increased.

The right figure shows a box-and-whiskers plot for the correlation co-efficient from these calibration curves for each of the 1290 different analytes using both the original integrations and those from the machine learning results. The solid ‘box’ represents the middle 50% of the points, the ‘whisker’ lines each cover 1.5 times the length of the box and the points are outliers outside this range. The top shows the entire range and the bottom is zoomed. For this larger data set a number of the analytes were unsalvageable (even with best manual integration), however the ML was able to improve the regression in most cases.



ISD results

In order to understand whether a model could be created once and successfully applied to future data sets the multi-batch ISD data was used. There are 24 batches acquired at roughly equal intervals from 2016 and five additional batches from 2021. A model was trained on the first five batches from March and April of 2016 and applied to all batches.

The figure above shows the true positive (TP), true negative (TN) and false positive and negative (FP and FN) results for the different batches. There may be a small degradation applying the model from the early batches to the later ones, but certainly not a significant one. That said, these batches were not the most challenging so a different result might be obtained for other data sets.

CONCLUSIONS

The approach of performing multiple peak integrations using different parameter settings for each chromatogram and then using a machine learning model to pick the highest scoring set is showing considerable promise.

- The increase in correct integrations leads to time savings since fewer peaks need parameters to be manually adjusted
- The ML results required no up-front quantitation method development (*other* than specification of target m/z and approximate retention time), so there is an additional time saving since parameters do not need to be adjusted for each analyte; this is of course most relevant for large panels with many analytes
- Unlike previous work (not shown) the approach does not require standards of known concentration.

Future work is likely to evaluate using data sets acquired over a longer time duration.

REFERENCES

1 The scikit-learn machine learning python library is documented at https://scikit-learn.org.
2 *Evaluation of a new peak integration algorithm for high throughput LC/MS/MS data processing*, Poster Presentation at ASMS 2010 (ThP24 558).
3 *A method for improved LC-MS/MS peak integration by using multiple traces and peak modeling*, Poster presentation at ASMS 2015 (MP05 106).

TRADEMARKS/LICENSING