Outline

William Kumler

2023-04-17

## Abstract

TBD

## Introduction

LC/MS techniques are neat

* Fast sample processing
* Low detection limits
* Can characterize new molecules

Most of the data produced by them goes unused

* Targeted workflows look at only a fraction of the data
* Untargeted algorithms are still fairly new

Peakpicking is hard

* Smooth transition from “good” into “meh” into “bad” across the limit of detection
* Fundamental tradeoff between false positives and false negatives
* Lots of false positive peaks because attempting to minimize false negative
* Requires manual curation of peak quality which doesn’t scale

Algorithms don’t provide likelihood of “goodness”

* Typical output: m/z, rt, area, maybe SN
* We want/need a probability that this peak is just noise
* Want to control the false positive/negative rate depending on project goals
* We don’t need new algorithms as much as we need a way to quantify success on the existing ones

Multiple advantages to a robust assessment of peak quality

* Reduces manual labor in sorting out good peaks from bad
* Improves statistical power
  + Reduces the number of hypotheses tested for univariate analyses
  + Increases relative power of significant peaks for multivariate analyses
* Allows optimization of peakpicking parameters for untargeted approaches
* “Good” peaks should be independent of software choices while noise is noisy
* Enables more consistent quality control across labs
* Focuses calculation time/energy on best peaks rather than all equally
* Provides consistent, reproducible results via scripting instead of independent expert assessment

A single comprehensive quality parameter is more useful than many independent ones

* Enables sensible threshold selection (50% likely to be a peak)
  + Rather than current arbitrary approaches (SNR > 20, peak height > 10e7)
* Single metric has multivariate advantage over multiple univariate thresholds
  + Good peak can fall be weak in one or two metrics and still be safe
* Relative power of different metrics has never been computed/compared
  + More important to have an isotope? Large area? Good shape? Differ between treatments?

Can be used to check improvements in peakpicking/chromatography

* Traditional method: a few random peaks are checked after making a change to params
* Total number of peaks used as proxy, not total number of GOOD peaks
* Did this change improve overall peak quality?
* Currently impossible to optimize these for untargeted peakpicking

## Methods

### Sample collection

TBD

### Sample processing

TBD

### LC conditions

Get from Laura and MW

### MS conditions

Get from Laura and MW

### Peakpicking with XCMS

Fill from the peakpicking\_and\_prep script or the msnexp\_filled object

### Manual inspection and classification

Gold standard: manual inspection by a human expert

* Decision to view the entire feature rather than a single trace
  + 100x less work
  + More representative of typical MS efforts
* Decision to color by treatments
* Decision to drop missed scans entirely

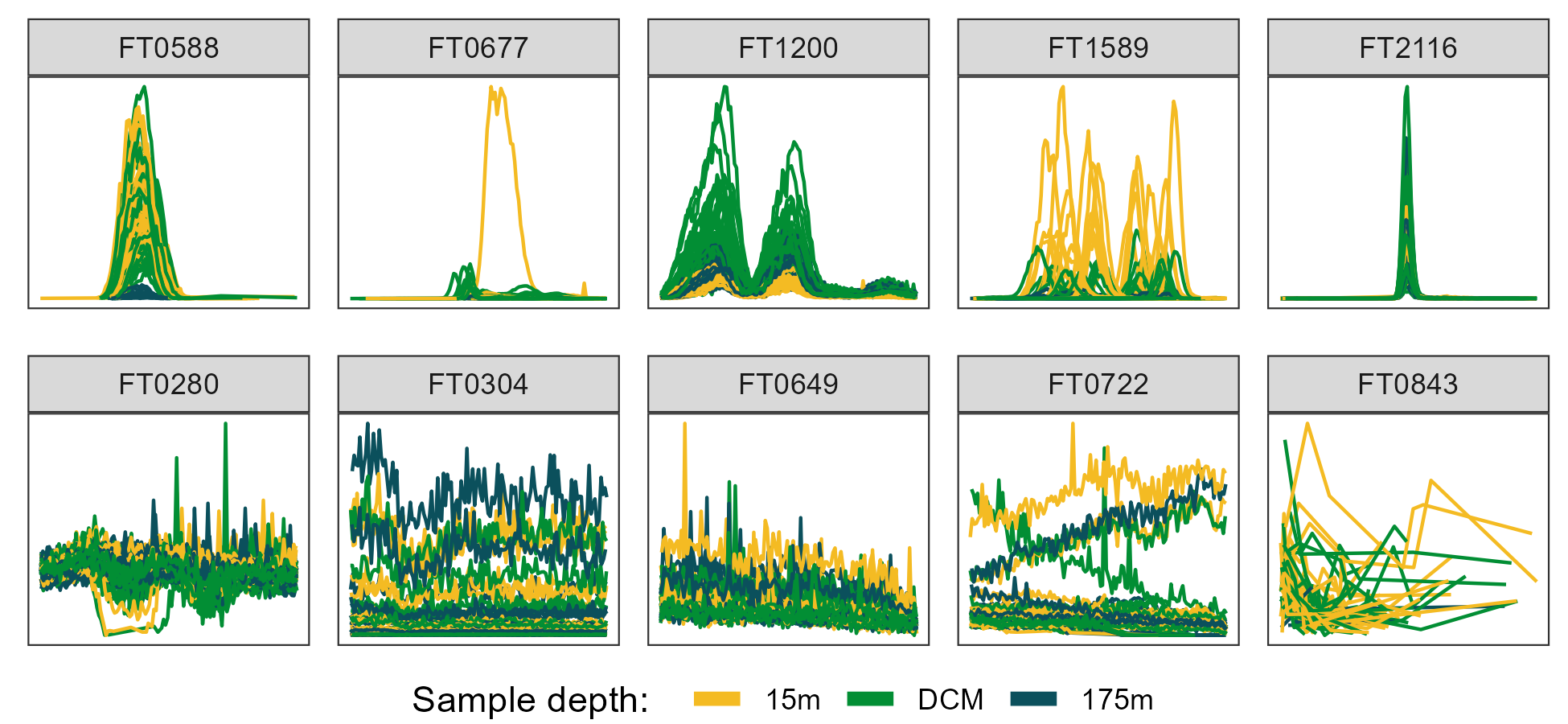
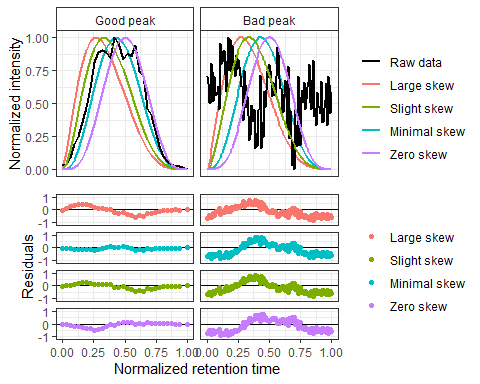


Figure 1: Randomly selected ion chromatograms from both “Good” (top row) and “Bad” (bottom row) manual classifications, colored by the depth from which the biological sample was taken. DCM = deep chlorophyll maximum, approximately 110 meters.

### Feature extraction

Feature extraction - many possible metrics, calculated across features

* Some trivial to calculate
  + Mean m/z, mean RT, number found
* Some require access to the raw data
  + Peak shape, peak SNR, isotope existence, number of missed scans



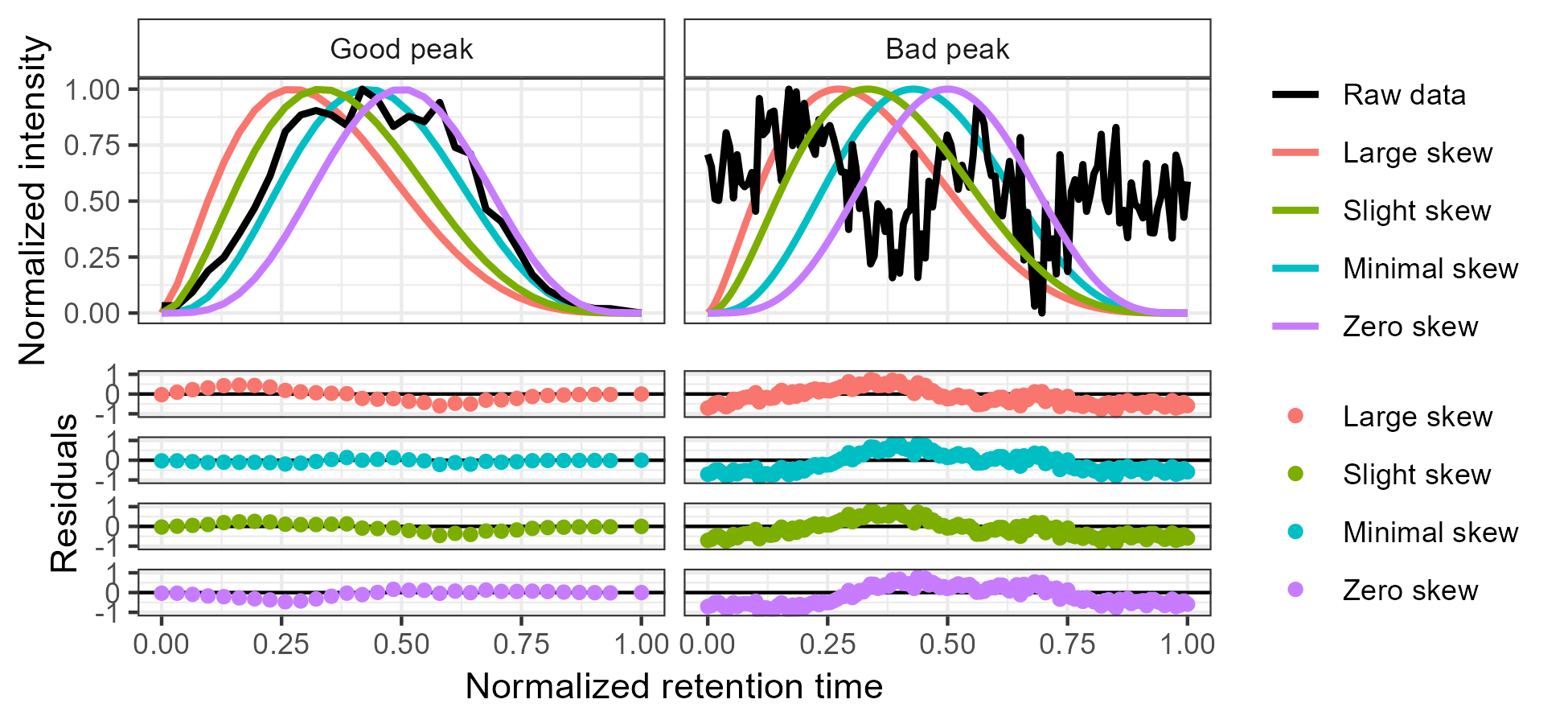


Figure 2: Method used to calculate the metrics for the two-parameter model from the raw data via comparison to an idealized pseudo-Gaussian peak for both manually identified “Good” and “Bad” peaks. Normalization was performed by linearly scaling the raw values into the 0-1 range by subtracting the minimum value and dividing by the maximum. Peak shape similarity was measured with Pearson’s correlation coefficient and the noise level is estimated as the standard deviation of the residuals after the raw data is subtracted from the idealized peak.

## Results

### Predicted probability distributions

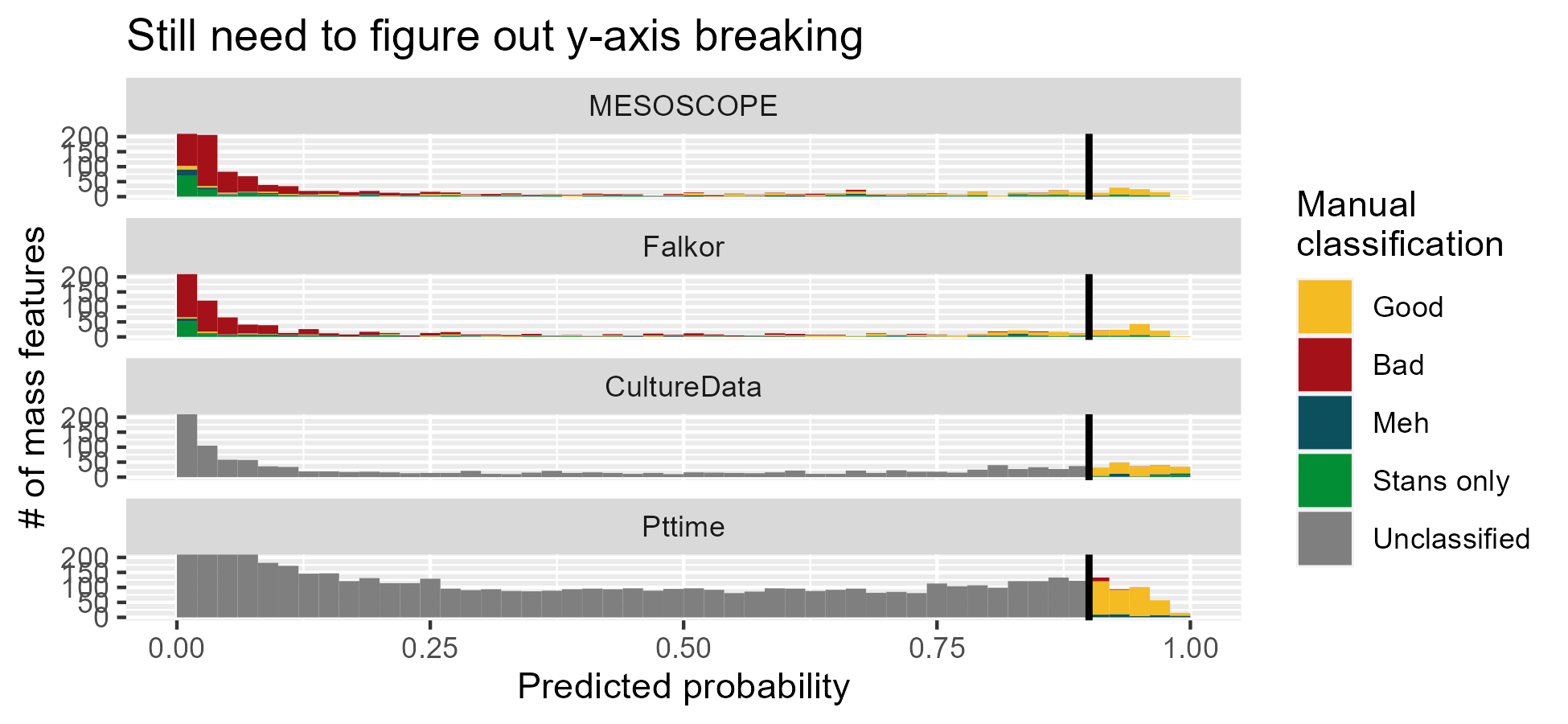


Figure 3: Histograms showing the estimated likelihood of a given mass feature being categorized as “Good” according to the two-parameter logistic model trained on the combined fully-labeled Falkor and MESOSCOPE datasets. Colors indicate the category in which each feature was manually assigned by an expert, with “Stans only” referring to a good peak that was only visible in the standards run alongside the samples. Datasets CultureData and Pttime were manually labeled only for those features with an estimated likelihood above 90% (black vertical line) and were otherwise unclassified.

### Model cross-validation with novel datasets

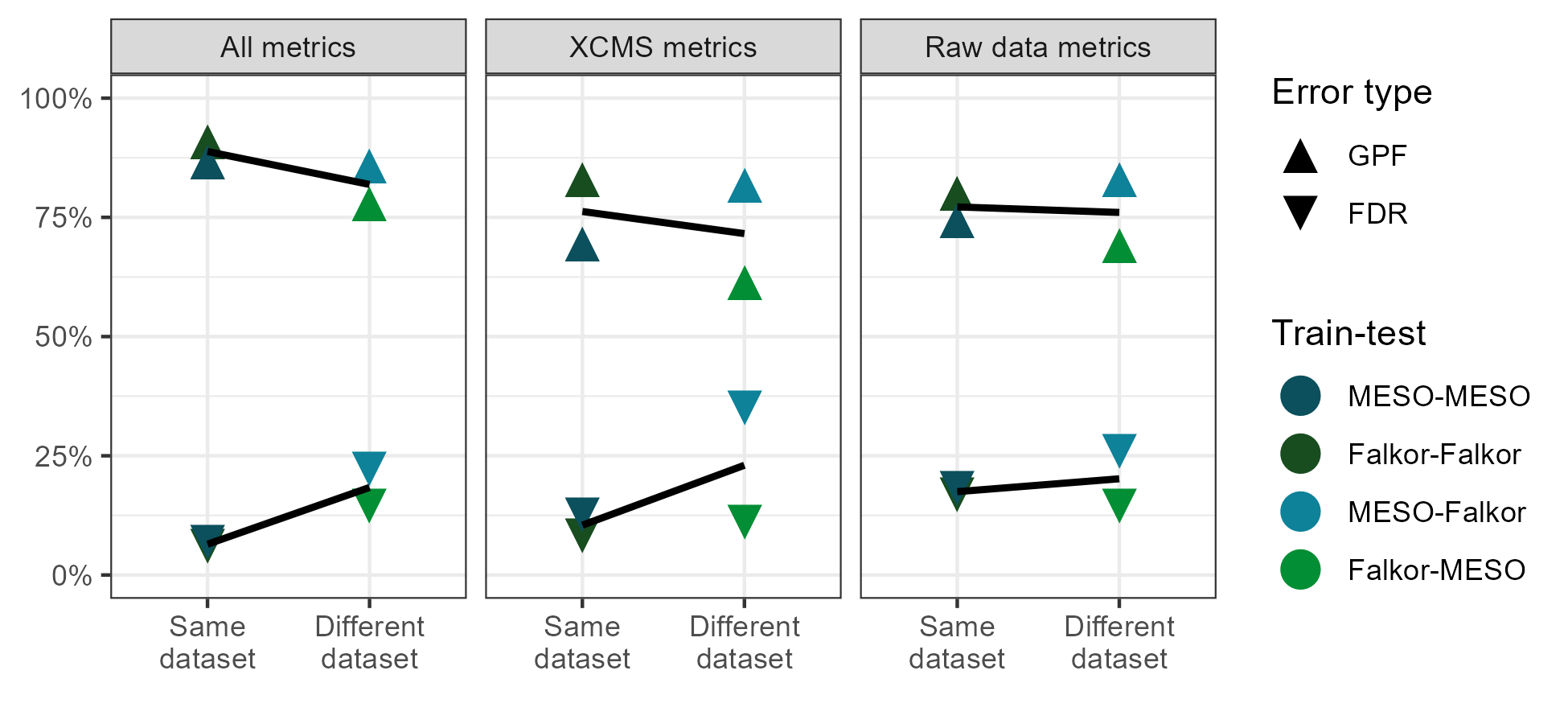


Figure 4: False discovery rate (%FDR) and fraction of good peaks found (GPF) plotted across different subsets of model parameters. Lower %FDR indicates a smaller fraction of false positives among those peaks the model categorized as “Good” using a threshold of 0.5, and higher %GPF indicates a larger fraction of the total good peaks were found using the same threshold. Points are colored by the model used for training and testing, with internal validation (using the same dataset for training as prediction) in the darker colors on the left and external validation (using a different dataset for training than prediction) in the lighter colors on the right of each panel. Lines of best fit have been estimated and plotted in black on top of the data points.

### Full model probability depends strongly on training data, two-parameter model does not

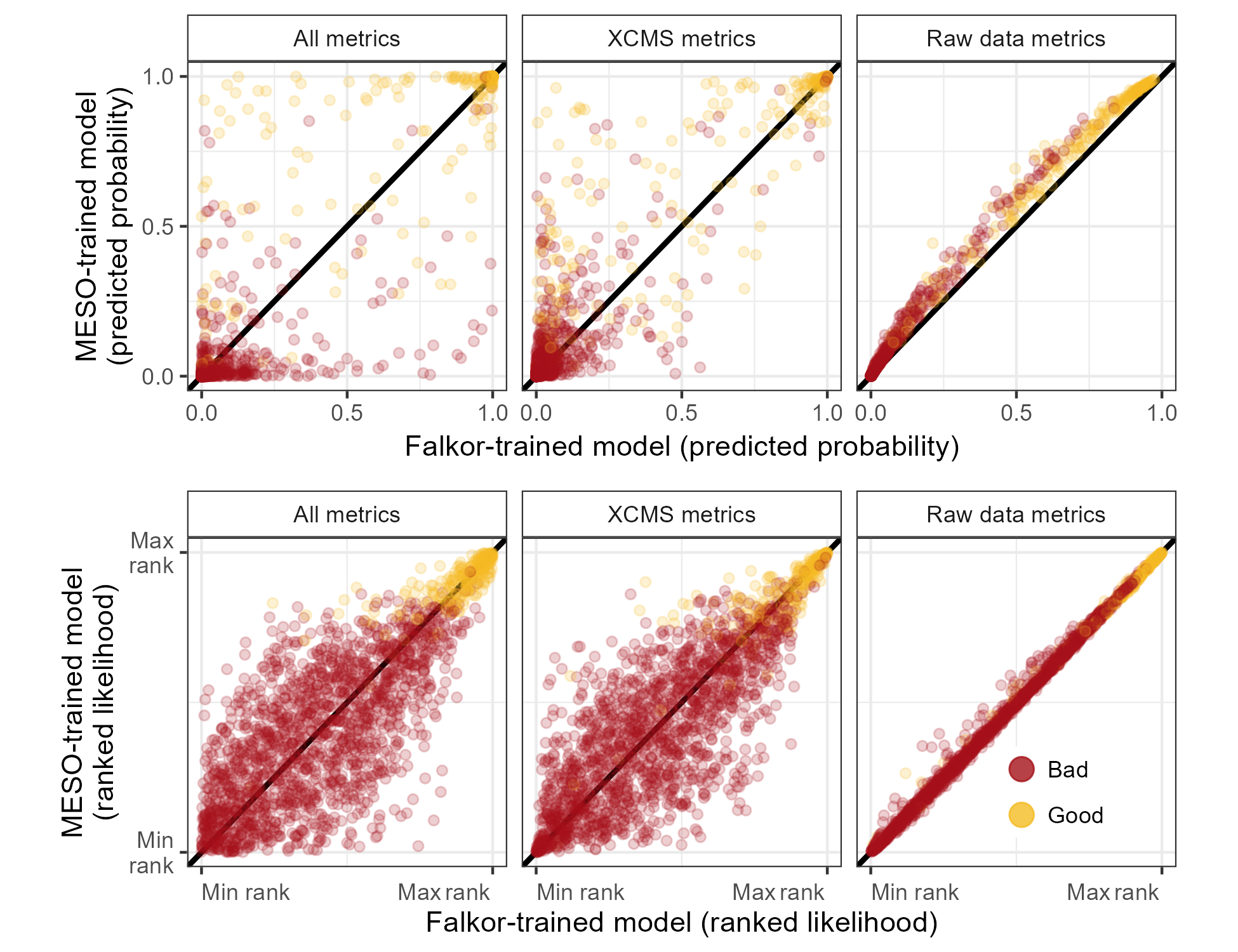


Figure 5: Predicted likelihood of a feature being “Good” according to a model trained on the MESOSCOPE dataset vs a model trained on the Falkor dataset. The top row of plots show the exact likelihood predicted by the logistic model across three different subsets of parameters, while the bottom row shows the estimates ranked from least likely to most likely. Points are colored by their manually-assigned quality according to an expert.

Make sure labels are the same as above - probably stick with XCMS for specificity.

Remember to add Pearson and Spearmen correlation coefficients in the corners!

### Error rates on novel data with stricter likelihood threshold

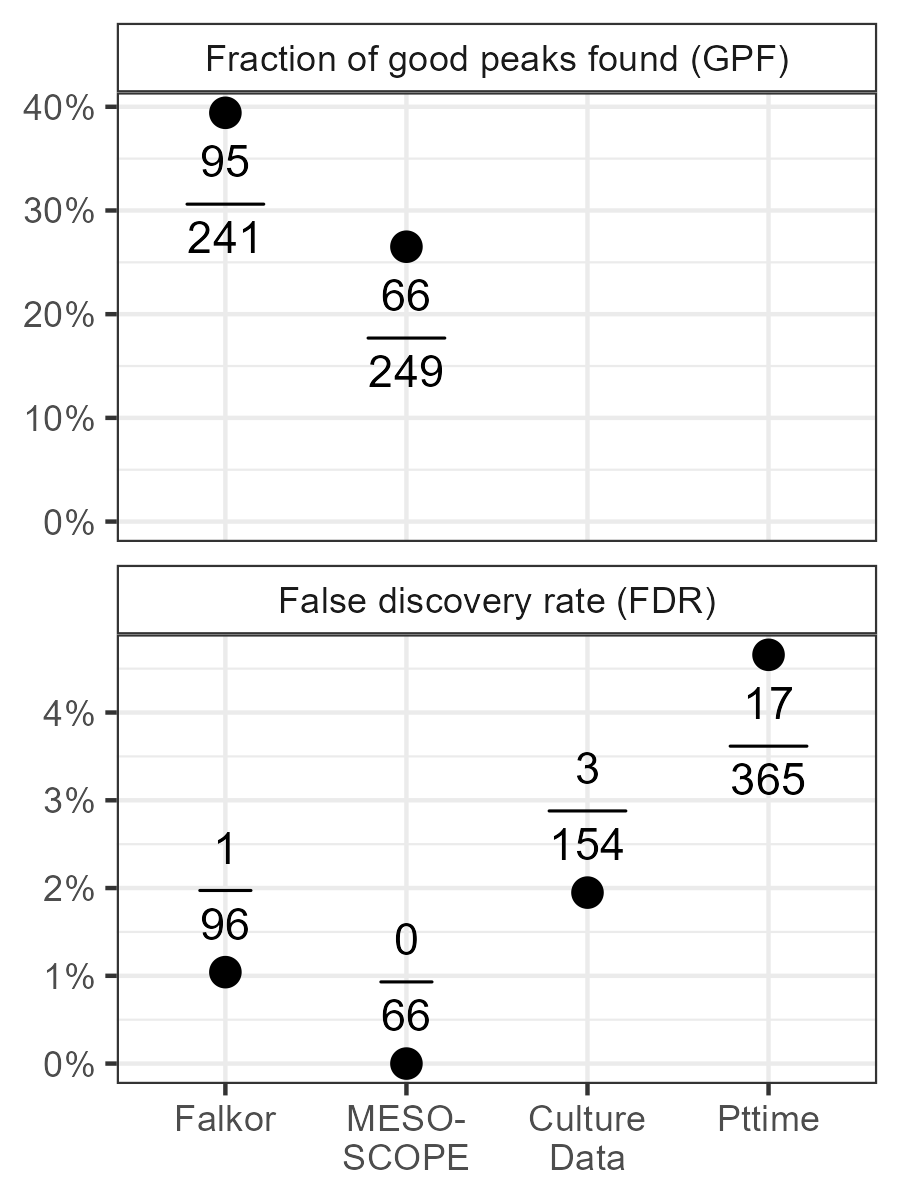


Figure 6: False discovery rate and proportion of total good peaks identified as good by the two-parameter model trained on the combined MESOSCOPE/Falkor dataset and applied to each dataset individually with a threshold of 0.9. FDR is calculated by dividing the number of false positives by the total positives produced by the model and GPF is calculated by dividing the number of true positives by the total number of good peaks as identified manually. Points correspond to the calculated percentage and absolute numbers are provided above/below the point.

Emphasize the use of a different threshold here!

### Implications for biological conclusions

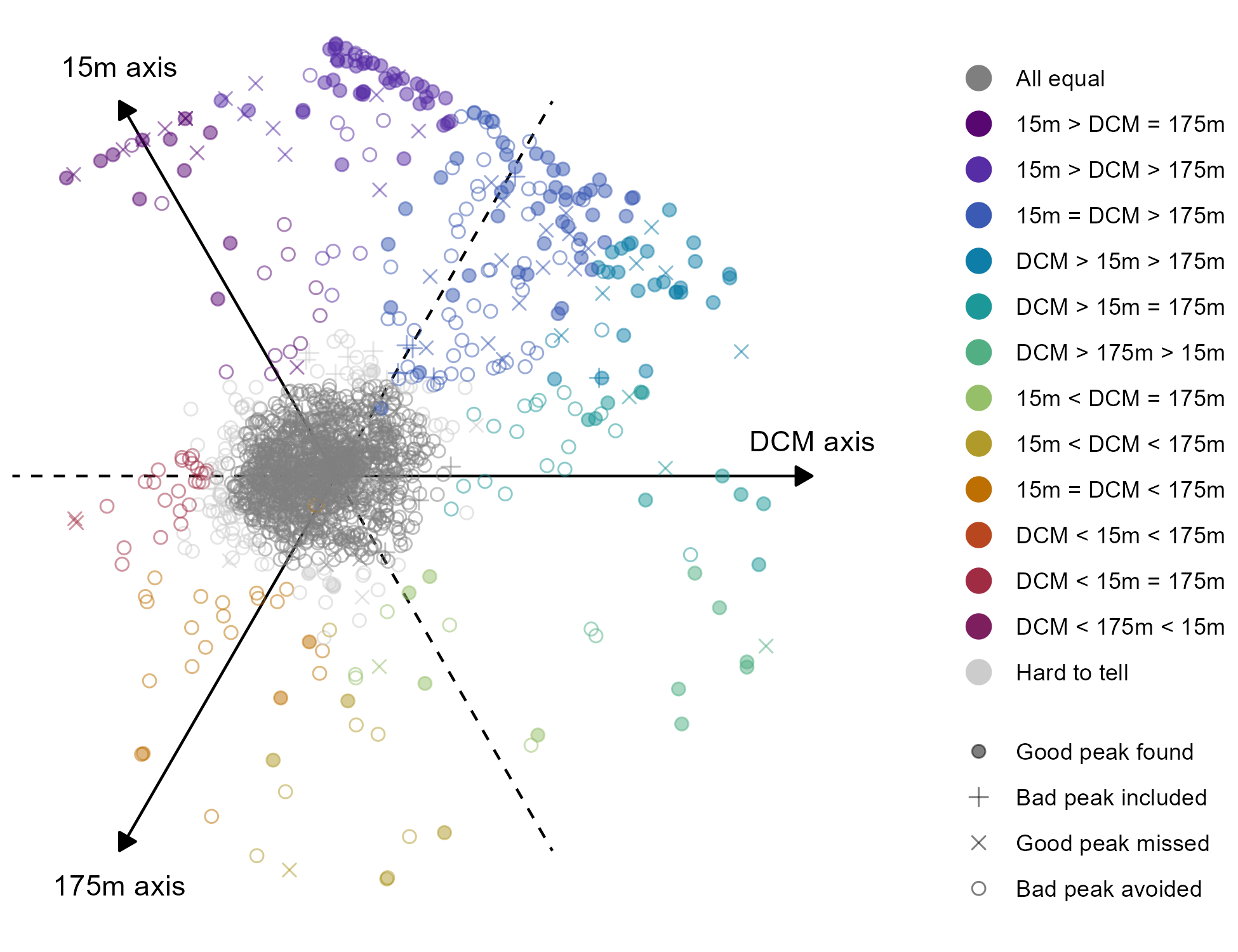


Figure 7: Plot of metabolite response to depth shown across three axes using the rank-normalized median value at each depth as the coordinate for that axis. Each molecular feature corresponds to a point in the plot, and their position on the plot describes the shape of their depth profile. Compounds aligning with the 15m axis correspond to compounds with most of their abundance found in the surface ocean; points far to the right side correspond to compounds that are found only at the deep chlorophyll maximum; points found at the bottom of the plot are those compounds that increased more or less linearly with depth.

Make sure I include the threshold used for the above plot (in the legend?)

Make sure I explain the statistical significance and the “Hard to tell” category

See aobut including ONLY the good peaks in a separate plot (inset figure?)

## Discussion

TBD

## Citations

2020 review and demand for more robust peakpicking and description of noise sources <https://www.sciencedirect.com/science/article/pii/S0165993620302922>

2014 Comparative evaluation of preprocessing freeware on chromatography/mass spectrometry data for signature discovery <https://www.sciencedirect.com/science/article/pii/S0021967314010450?via%3Dihub>

* XCMS is the most cited preprocessing tool currently in the metabolomics literature
* Table comparing parameters across XCMS, MzMine, MetAlign
* Selection by appearance in multiple replicates
* Visual inspection of peaks for quality
* Reports percentage of false positives as between 90% and 50%
* Most pressing improvement needed for all the tested data analysis tools was to reduce the percentage of false peaks

2003 the multivariate advantage <https://www.sciencedirect.com/science/article/pii/S0003267003006810?via%3Dihub>

2017 One Step Forward for Reducing False Positive and False Negative Compound Identifications

## Unstructured thoughts

Pleas: - Report peak boundaries! If not max/min m/z then at least max/min RT - Helps with visualization and raw data feature extraction - Run a pooled/aggregate sample

Brags: - Built the classifier tool in R, very fast, uses RaMS - Did the PCA things, sometimes successful

In a Bayesian framework, the base rate (likelihood of being a real peak) is critically important for any downstream analysis.

## Supplement

Literally ALL the confusion matrices (as a table?)

Include data from other classifiers - Clustering(?) - Random forest - Lasso/ridge/elastic net regression

Figure: feature selection / model selection barplot from powerpoint - Too many numbers for the main text but useful for power users - Colored histograms (Results Fig 1) but for other param options - Red/gold barplot at 0.1, 1, 50, 99, 99.9 thresholds

Figure: model stability under smaller training set? - Both for all params and raw data params?

Figure: model estimate stability errorbar plot?