Polly PeakML Accelerates Biological Insight Derivation from Untargeted Metabolomics Analysis

INTRODUCTION

Mass spectrometry-based metabolomics studies can provide crucial insights into alterations in complex biological systems and the systemic effects of diseases and drugs. Even though advancement in high-throughput technologies has enabled large-scale metabolomics studies, key challenges remain in peak detection including manual evaluation of global profile of metabolites for noisy peaks removal in data from multiple biological samples. This step requires considerable time and expertise. From peak curation to analysis- a robust computing environment encompassing all aspects of peak curation is an essential requirement. Polly PeakML enables high-speed mass spectrometry data processing and metabolomic analysis in linear workflows that can be customized.

CUSTOMER DETAILS

Our customer, a renowned biopharmaceutical company pioneering in therapies for rare genetic diseases, wanted to facilitate an initiative for ML-model based classification of untargeted metabolomics data. This approach could then be utilized for investigational therapies in preclinical development.

PROBLEM STATEMENT

An untargeted metabolomics study has advantages over targeted metabolomics study. However, it remains limited in the degree of reliance with which detected signals can be characterized, and there are additional road-blocks during the analysis.

CHALLENGES

- **Noisy Peaks**- with up to 30% of the signals being noisy peaks (adducts, contaminants and artifacts) large populations are rendered unusable.
- Manual peak detection is a **tedious, error-prone process**. Filtering strategies to analyze tractable high confidence peak groups used by experts can result in loss of low abundant signals, inconsistent hypotheses and insights.
- The manual curation process to remove noise is **time-consuming and expertise dependent**.
- **Expert curation can be subjective** making it difficult to classify peak groups as real signal or noise.
- **Data ingestion or processing** could be **computationally intensive or expensive.**

OUR SOLUTION

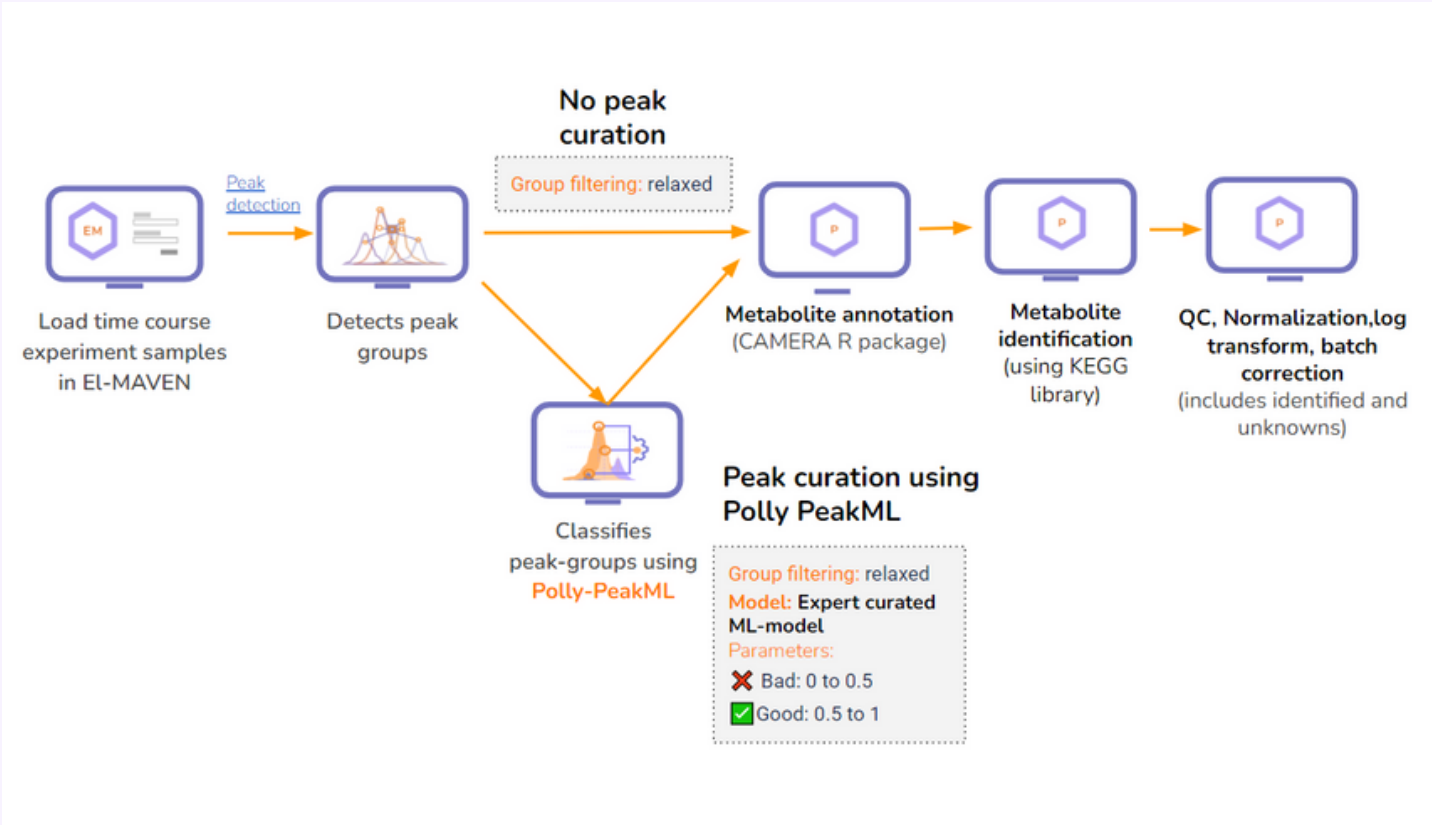
Polly-PeakML is Elucidata's novel machine-learning tool that enables the classification of peak group as **metabolites of interest, real signals** and **noise** accurately and at a much faster rate.

Mass spectrometry data is analyzed in El-MAVEN where signals are detected using an automated compound detection approach--an untargeted approach--that determines peak groups or features with varying intensities. Subsequently, these peaks are annotated and identified as known or unidentified metabolites. Depending on the intensity threshold, the number of detected peaks and noise varies. Peak classification is not straight-forward. The model has to work through different ranges, resolution, and high-throughput.

Using Polly-PeakML, peak-groups are automatically classified into real signal and noise within minutes or hours with an accuracy of 94%. The machine learning model can be customized based on the customer's needs.

Polly PeakML Features

- Trained, Tested & Validated ML algorithms for peak classification
- ~94% accuracy
- Time: 120x faster than an expert
- Enabled identification of novel metabolites: 2.5x more metabolic features



The detected peak-groups are curated in Polly PeakML as compared to un-curated peak-groups in the processing steps

EXPERIMENTAL DETAILS

To study whether PeakML curated data can provide better insights about disease mechanism or drug action, the following cohorts were studied:

- Wild Type (Healthy)
- Wild Type + Doxycycline (Healthy + Dox)
- Disease (Knock Out)
- Drug-Rescue
- Dox-Rescue.

Thereafter, two approaches to understand the metabolic alterations were adopted:

Pathway-centric Approach and Metabolite-centric Approach.

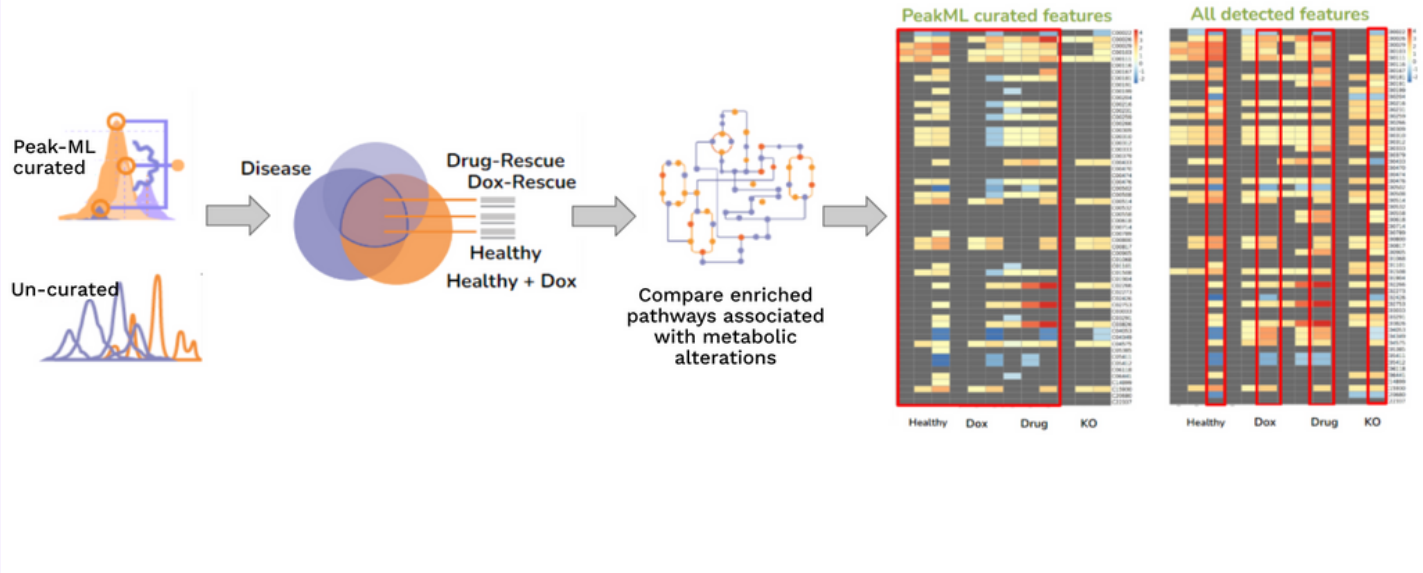
- **Pathway-Centric Approach-**Here, we wanted to study the metabolic alterations that occurred in the different cohorts during the time course of 72 hours to understand which pathways exhibited similar/dissimilar alterations.

HIGHLIGHTS

The Pathway-centric approach led to the identification of many differences between the cohorts in terms of the enriched pathways that were observed. With cohorts that should have a different phenotype, curated data could identify significant differences, and some similarities.

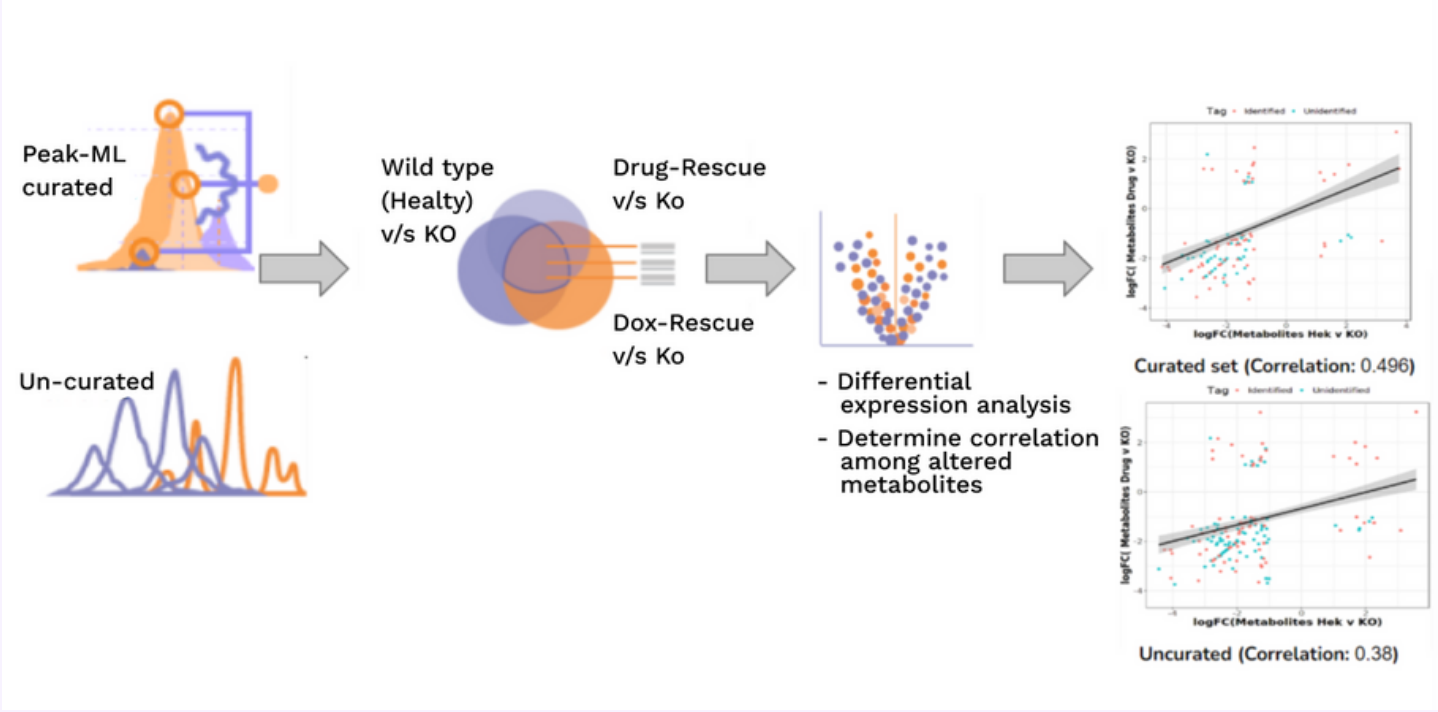
The Metabolite-centric approach helped identify the metabolites that were altered in the Drug-Rescue and Dox-Rescue cohorts; the Healthy-Dox and Healthy cohorts which were not identified in the diseased states, respectively. Also, it was observed that using curated data, the altered metabolites in Dox-Rescue and Drug-Rescue correlated better with the healthy cohorts.

Polly-PeakML pipeline was able to perform the task in just a few minutes and the entire analysis in a few weeks!



Removing bad peak groups allows specific selection of metabolites with significant fold changes

- **Metabolite-Centric Approach-**The correlated metabolites between Healthy, Drug-rescue, and Dox-rescue v/s Disease state were analyzed. This allowed us to determine whether metabolites similar to the healthy condition were recovered.

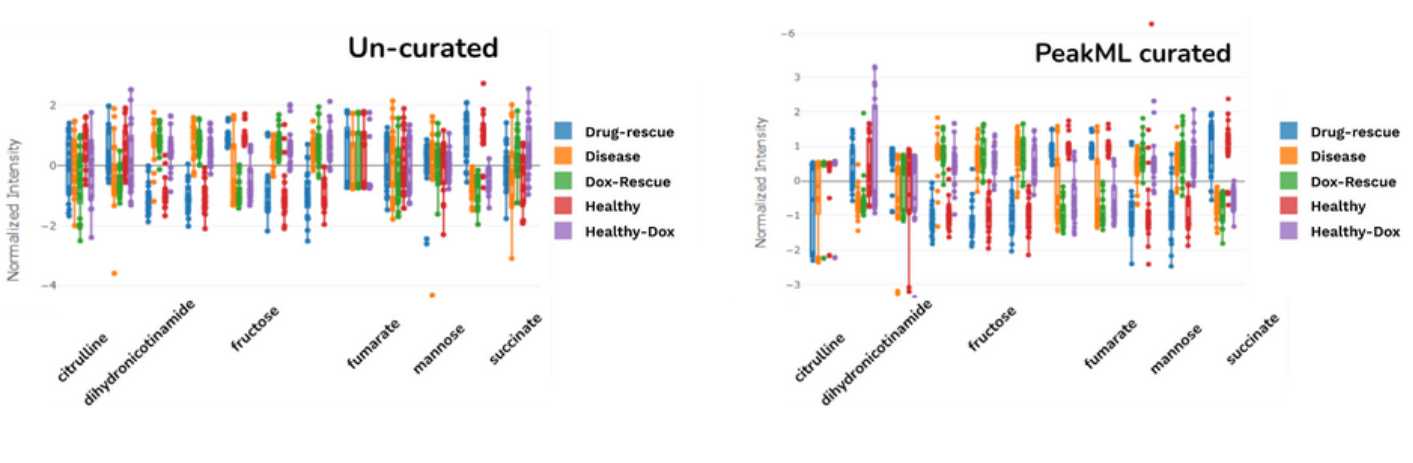


Metabolites corresponding to low quality features introduce non-specific variation

IMPACT

Using Polly-PeakML, we observed that the unidentified and identified metabolites both exhibit ***better correlation*** after peak curation. This procedure also helped in the selection of ***specific biomarkers*** of interest, by weeding out a lot of non-specific metabolites that may have appeared only due to ***noise.***

One primary reason as to why the observations ***varied significantly*** between the two cases is- after the removal of noisy peak groups, the number of differential metabolites and the compound intensities was altered significantly.



Normalized intensity of metabolites

Importantly, we were able to identify True Positive pathways associated with the diseased condition or recovery towards the healthy state distinctly and exclude False Positive, False Negative pathways from the analysis.
Specifically,

1. Pathways associated with Drug-Rescue cohorts identified using un-curated data:
   one pathway was identified as a False Positive,
   four pathways were identified as False Negative.
2. ***Seven pathways*** associated with Drug-Rescue cohort were identified as True Positives using curated data!

The Drug activity or effects were ***clearly beneficial*** based on the observations from the curated data which was not apparent in un-curated or inexpertly curated data!

WHY ELUCIDATA

Elucidata's Polly-PeakML is enabling our customers to identify disease specific pathways and potential biomarkers for drug-rescue state. By diligently removing the inherent noise in the metabolomic datasets, Polly-PeakML has alleviated the challenges of untargeted metabolomics analysis and reduced the manual effort and time involvement in these projects. We also have high-end servers at our end which can additionally reduce the time required for these analyses.

Polly's ML solutions offer a rich ecosystem of machine learning models, multi-omics integration algorithms and statistical tools to enable the bioinformaticians in your team to learn better and faster insights from your data. Leverage Elucidata's expertise in metabolomics data analysis techniques to uncover novel mechanisms. Revitalize your R&D efforts today with Polly. Use our tested and validated models or build models of your own with great ease using ML-ready data from Polly.