Matterworks

EBOOK / PRACTICAL GUIDE

Getting Started with AI-Powered Mass Spec Analysis

A Practical Guide to AI-Powered Biochemical Insight

Matterworks | 2026

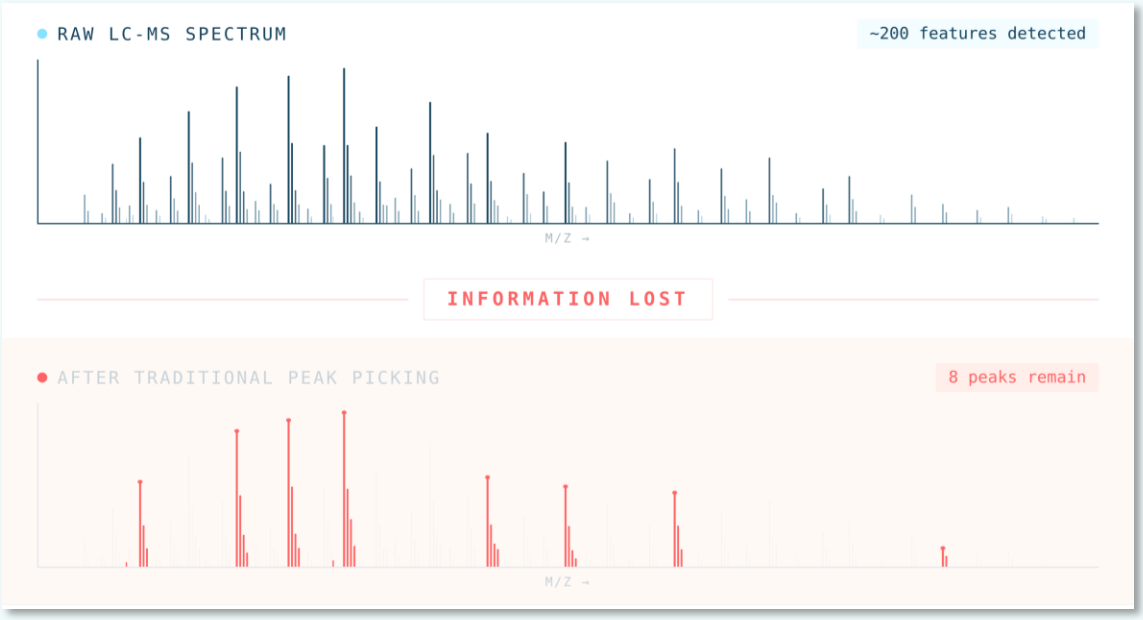
Contents

CHAPTER ONE

Where Identification Breaks Down

Mass spectrometry generates enormous amounts of molecular data. But the vast majority of it goes unidentified. **Across all public MS/MS repositories, less than 10% of spectra have structural annotations.** The rest is biological dark matter: real signal from real molecules, sitting in datasets with no name attached.

This isn't a data collection problem. Modern instruments are sensitive and fast. The bottleneck is interpretation. Traditional identification workflows rely on matching experimental spectra against reference libraries, but those libraries cover a fraction of biochemical space.



Most metabolomics studies report confident identifications for only a small fraction of detected features. Researchers either narrow their focus to a short list of known targets or accept that most of their data will go uninterpreted.

CHAPTER TWO

Why Untargeted Approaches Hit a Ceiling

The Library Matching Problem

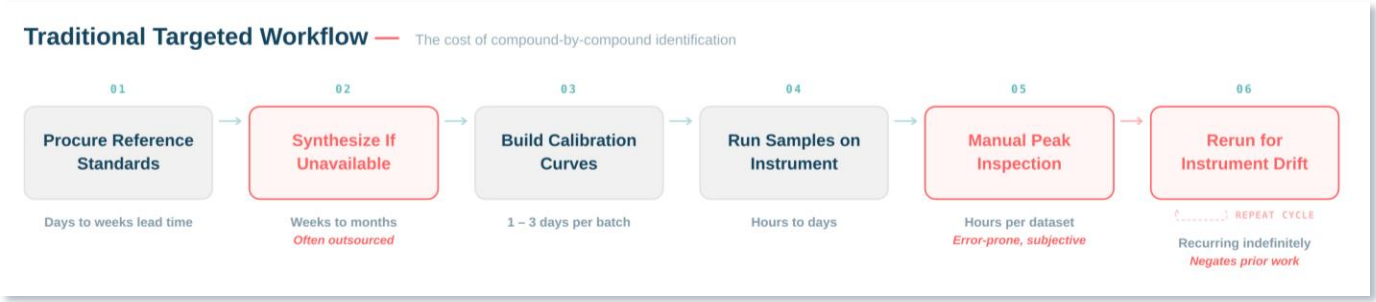
The standard approach to MS2 identification is spectral library matching. You take an experimental fragmentation spectrum, compare it against a database of reference spectra, and look for the closest match.

The problem is coverage. Even the largest public libraries contain reference spectra for only a fraction of known metabolites.

Cosine similarity, the most common scoring method, compares peak positions and intensities directly. It works well for clean, high-abundance spectra with exact library matches. It struggles with noisy spectra, low-abundance compounds, and isomers that produce similar fragmentation patterns.

The Quantitation Gap

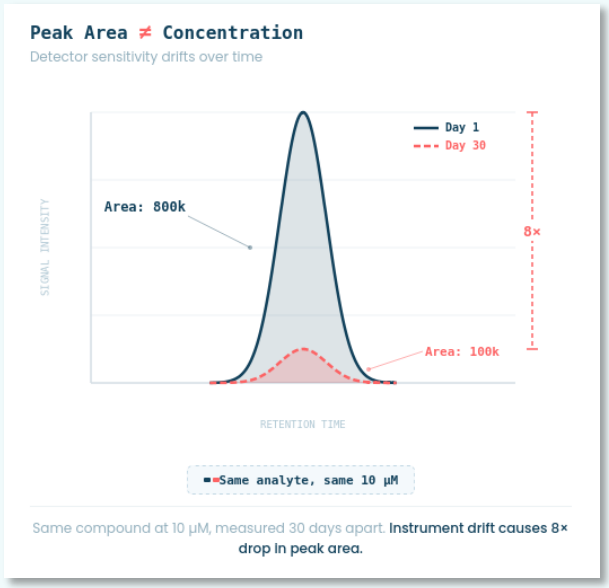
Even when you can identify a compound, measuring how much is present introduces its own set of problems. Targeted quantitation requires isotopically labeled standards for each analyte, calibration curves that drift over time, and manual inspection of every peak. In practice, measuring absolute concentration is limited to narrow panels of pre-selected molecules. It is not possible to quantify analytes not selected a priori for inclusion in an absolute quantitation method.

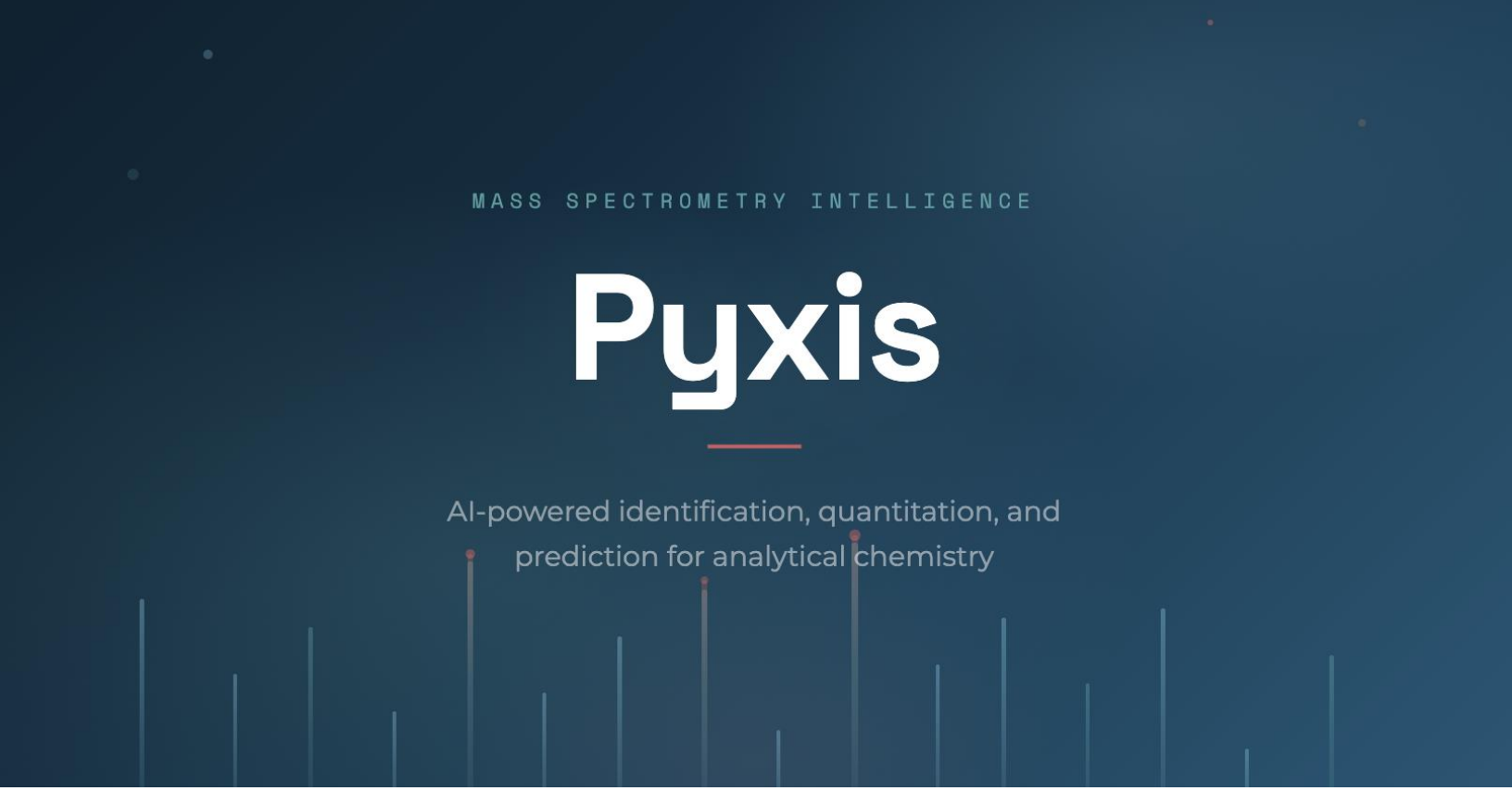


Peak Area ≠ Concentration

For untargeted studies, the common workaround is to compare peak areas as a proxy for concentration. But peak area is not concentration.

The result is a gap between what instruments can detect and what researchers can actually interpret.





CHAPTER THREE

What is Pyxis?

Matterworks builds foundation models for biochemical omics research. Large language models transformed how we work with text. Matterworks has built the equivalent for analytical chemistry. Our models are trained on millions of spectra, allowing them to identify, quantify, and predict biology across workflows.

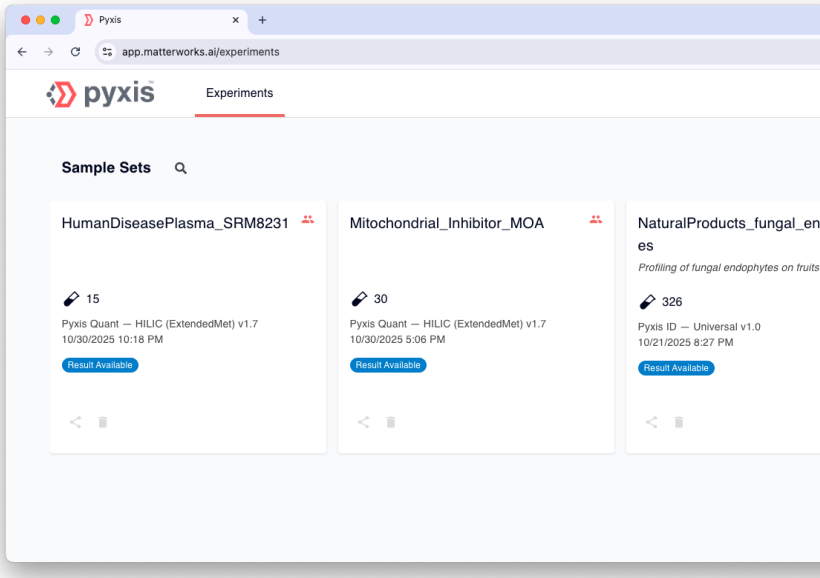
Pyxis is the platform that puts those models to work. Every capability you access through Pyxis, whether it's compound identification, quantitation, differential analysis, or prediction, comes from the same underlying platform.

Pyxis in Five Steps

You can try the full platform without uploading anything. Sign up, pick a demo dataset, and walk through the complete workflow in under ten minutes. Or upload your own MS2 data and see results on your samples.

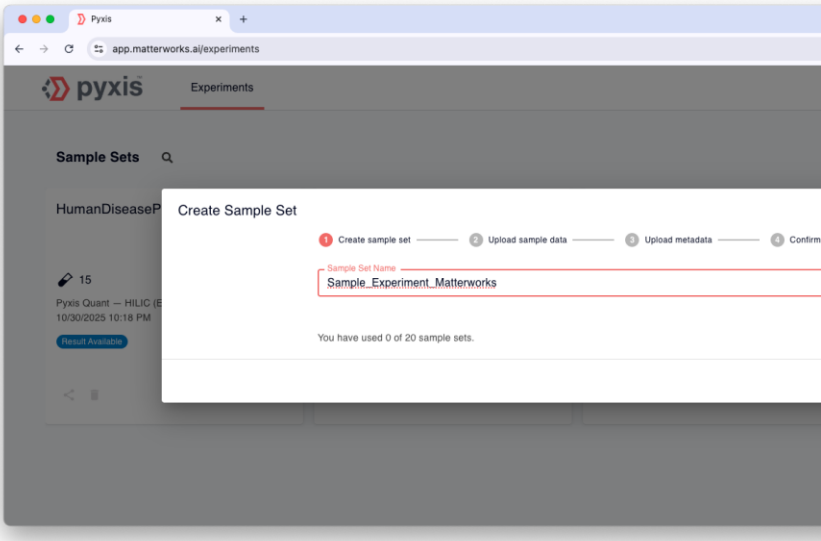
1 Sign up and load your data data

Create a free account at app.matterworks.ai. From the homepage, choose a pre-loaded demo dataset to explore immediately, or drag and drop your own MS2 files. Pyxis accepts .raw, .mzML, .mzXML, and .wiff with no preprocessing required.



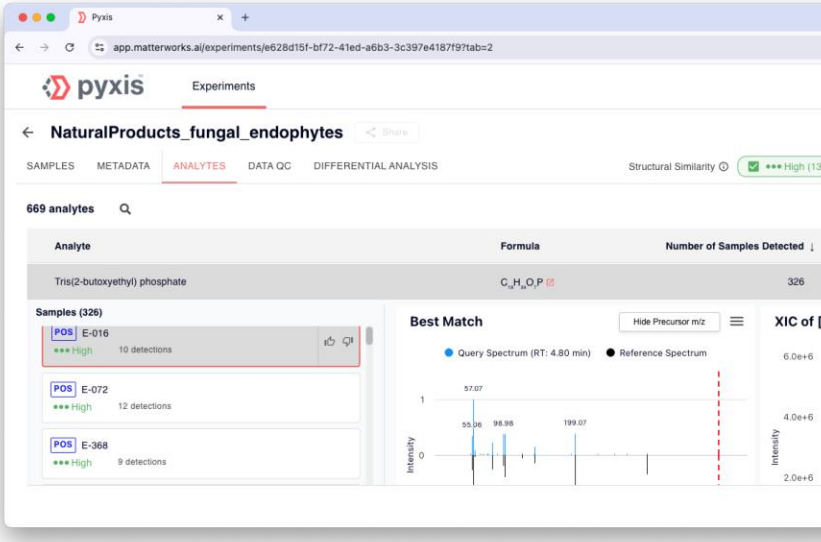
2 Run analysis

Hit analyze and Pyxis processes your data automatically. The platform extracts MS2 features, runs identification through the Large Spectral Model, and returns results. Typical studies with 50–100 samples complete in about 20 minutes.



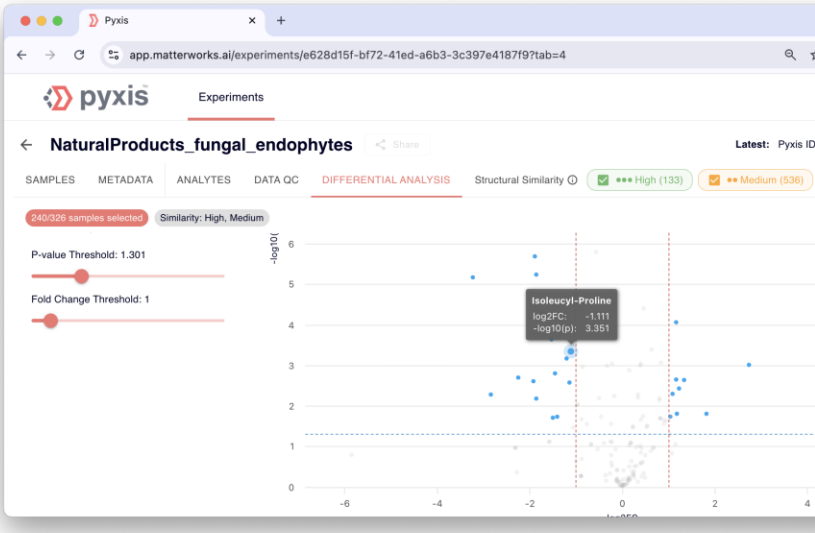
3 See what's in your samples

The analytes tab shows every identification with name, molecular formula, number of samples detected, and structural similarity score. Click any analyte to see its XIC to confirm the underlying signal.



4 Run differential analysis

Upload metadata, select sample groups, and Pyxis generates a volcano plot showing significantly up and downregulated analytes. A heatmap of the top differential analytes shows consistency across samples at a glance.



5 Export everything

Download identification tables, statistical results, and publication-ready figures. All exports use standard formats compatible with downstream tools.

Confident ID at Scale

LSM-MS2 processes MS2 spectra and returns putative identifications with transparent scores. The model excels at distinguishing isomeric analytes, compounds with identical masses but different structural arrangements, which have historically challenged automated identification systems.

> **Structural Similarity Scoring**
>
> Every identification includes a structural similarity score, which reflects how closely a compound matches in latent space. Pyxis distinguishes high, medium, and low matches, allowing you to self-select which ones to investigate further .

Every identification includes a high/medium/low structural similiarity score, allowing researchers to filter results based on their quality requirements—whether prioritizing recall for discovery or precision for targeted validation.

Typical untargeted analyte studies with 50–100 samples complete in approximately 20 minutes, enabling rapid iteration between data collection and interpretation.

How It Works: Large Spectral Models

Large Spectral Models (LSMs) extract information directly from raw MS data without requiring manual feature curation or metabolite annotation. LSMs convert raw MS spectra into compact numerical representations (“embeddings”) that capture the full biochemical information content of samples.

Pre-training

Base models are pre-trained on a massive data set of highly variant raw MS spectra using the techniques of self-supervised machine intelligence.

Fine-tuning

Specific biological prediction problems can then be performed by fine-tuning application-specific models that require tractable numbers of samples.

Traditional approaches require detecting and quantifying specific known metabolites. LSMs use all spectral information including subtle baseline variations, minor unresolved peaks, isotope patterns, and relationships between features. The model learns which spectral patterns are predictive, even if they don’t correspond to currently annotatable metabolites.

MATTERWORKS · PYXIS

LLM ↔ LSM

The same paradigm — applied to a fundamentally new data type.

LANGUAGE

LLM

Large Language Model

INPUT — TEXT

The drug [redacted] caused [redacted] harm

Words and sentences — data that already exists on the internet

MODEL

ARCHITECTURE

Transformer

Pre-trained on trillions of tokens across diverse self-supervised objectives

OUTPUT

Understands Language

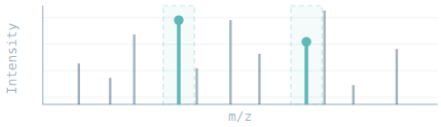
Meaning, grammar, context, reasoning

BIOCHEMICAL SIGNALS

LSM

Large Spectral Model

INPUT — MASS SPECTRUM



Instrument readings — a data type that has never appeared in LLM training data

MODEL

ARCHITECTURE

Transformer

Pre-trained on millions of spectra across diverse self-supervised objectives

OUTPUT

Understands Biochemistry

Metabolites, lipids, proteins — inside cells

Transparent and Verifiable Results

Every identification in Pyxis links back to verifiable spectral evidence. Click any metabolite to inspect the mirror plot comparing your experimental spectrum against the reference, with matching fragment ions highlighted.

You can also export just about any data you see in Pyxis: metabolite identifications with structural similarity scores, statistical analysis results, and publication-ready figures. All exports use standard formats compatible with downstream analysis tools.

Comprehensive ID Coverage

LSM-MS2 achieves high identification rates across diverse compound classes and sample types. Our reference library comprises 1.8 million high-quality spectra corresponding to 100K+ unique analytes.

All entries were curated, quality-controlled, and merged across multiple public and internal sources (NIST, MassBank, MSnLib, MoNA, GNPS, and internally acquired datasets).

1.8M	99K	~20 min
reference spectra	unique analytes	for 50–100 samples

Our Solutions

Everything runs through our Pyxis platform.
Whether you self-serve or partner with us

Self-service

Upload your LC-MS/MS data and run it through the same AI that powers our lab services.

BASIC	Free compound ID and differential analysis, with limited uploads.
STANDARD	Unlimited uploads with dedicated support.
PREMIUM	Micromolar quantitation via universal calibrants. No per-analyte standards.

BEST FOR

Teams with LC-MS/MS instruments that want faster, broader interpretation of their own data.

app.matterworks.ai

Lab Services

Ship samples to our lab. We run method dev, sample prep, acquisition, and analysis. Results returned in Pyxis.

EVERY ANALYSIS INCLUDES

- Untargeted ID across ~137K metabolites
- µM concentration on ~2K analytes
- Biological interpretation report (PDF + PPTX)
- HILIC + reverse-phase chromatography
- One year Pyxis portal access + full data download

BEST FOR

Teams without instrument access, needing end-to-end execution.

info@matterworks.ai

Partnerships

Our models learn deep molecular structure from millions of spectra. We partner to design the right study and deliver through Pyxis.

APPLICATIONS

Bioprocessing QC · Clinical classification
Clone selection · Raw material screening

info@matterworks.ai

WHAT'S SUPPORTED TODAY

Molecules: Metabolites + small lipids (expanded lipids + peptides coming soon) | Methods: HILIC, RP, pos/neg ionization

Matterworks

CHAPTER FIVE

Case Studies: Use Cases in Pyxis

Guillain-Barré Syndrome Biomarker Discovery

To demonstrate what Pyxis can do with existing data, we reanalyzed publicly available LC-MS/MS data from a clinical study comparing plasma samples from 30 Guillain-Barré Syndrome patients against 30 healthy controls.

UMAP dimensionality reduction analysis immediately showed clear separation between GBS patients and healthy controls, with quality control samples clustering tightly together. The biomolecular perturbations in GBS are substantial enough to drive cohort separation without any prior knowledge of disease status, suggesting strong, reproducible biomarker candidates.

Differential expression analysis revealed dozens of significantly altered biomolecules ($p < 0.05$, fold change > 2) between GBS patients and controls. The predominance of oxidized and hydroxylated lipid species points to oxidative stress and inflammation as central features of GBS pathophysiology. Hierarchical clustering confirmed these as robust, reproducible signatures rather than individual variability.

Within 15 minutes of uploading data, Pyxis returned confident identifications with structural validation, statistical analysis, and visualization for immediate biological insight.

Biomarker Discovery in Human Disease

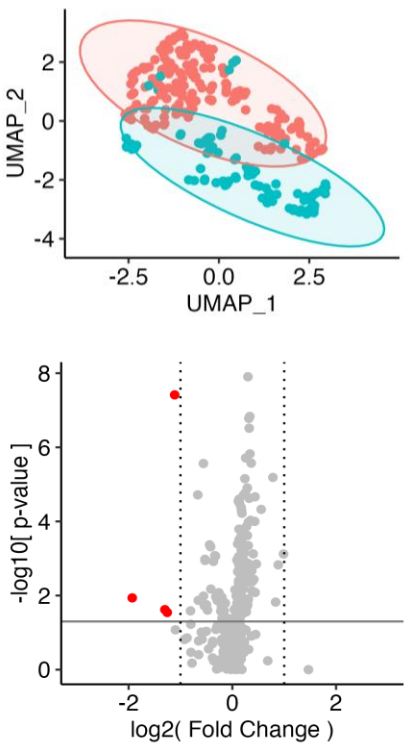
10X
improvement in phenotype discrimination over untargeted methods

4–5X
increase in discovery of statistically significant biomolecules

100X
acceleration to result vs conventional metabolomics

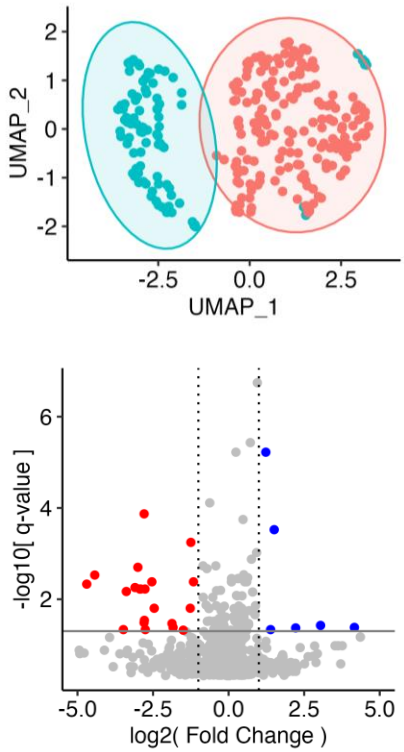
Conventional LC-MS Targeted Metabolomics

500 analytes with concentration
(6 weeks, $400/sample)



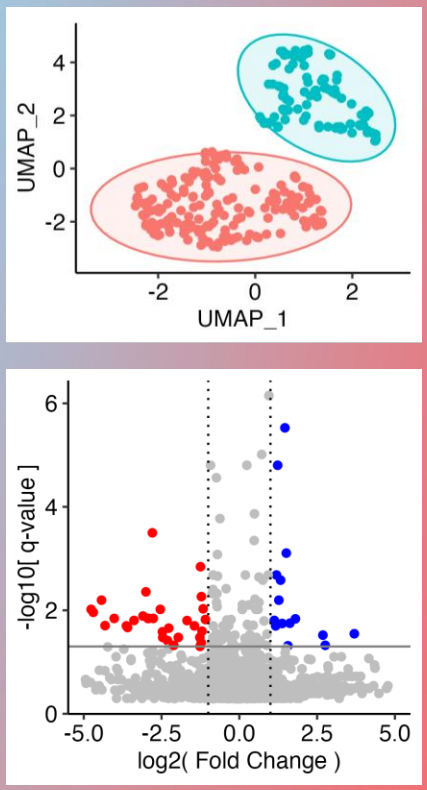
Conventional LC-MS/MS Untargeted Metabolomics

~1000 library matches and area
(14–20 weeks, $1,100/sample)



Pyxis Next-Gen LC-MS/MS Metabolomics

Vector search and concentration
(same day, parity w/ sequencing)



Methods

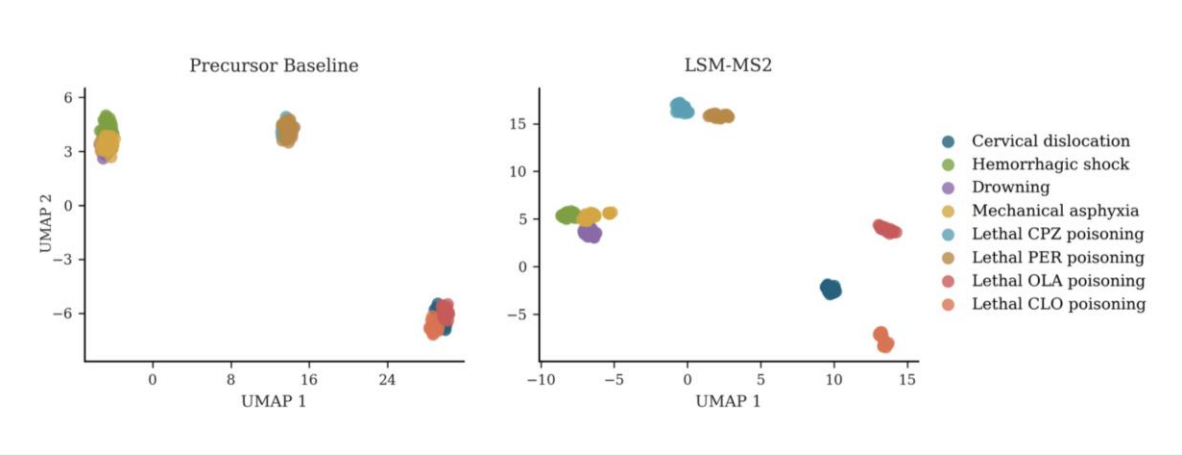
Methods: 300 clinical samples from multiple study sites, dates, and collection methods. For each patient, Healthy and Diseased phenotypes were assigned by a clinician. Samples were analyzed by LC-HRAM/MS2 (HILIC, RP, positive and negative mode) on an Orbitrap Exploris240 (Thermo Fisher Scientific). Biochemical annotation was performed in Pyxis using Universal ID (version 1.0.5) and predicted concentration (version 1.8.3). Clustering and mechanistic targets analyzed by Pyxis.

Antipsychotic Overdose Classification

Fatal intoxication by antipsychotic agents remains a major challenge in forensic toxicology today. Using a dataset of 80 mouse plasma samples, Bai et al. (2022) performed LC–MS–based metabolomic profiling to investigate causes of death. The dataset comprised eight groups: four drug-induced fatalities and four non–drug-related controls. The original study achieved clear separation between (1) overdose and control groups and (2) fatalities from chlorpromazine versus olanzapine

To assess whether spectral representations can recover this missing structure, we construct a simple precursor baseline embedding for comparison with LSM-MS2. As shown below, this baseline reproduces the same lack of separation reported by Bai et al. (2022).

In contrast, clustering based on LSM-MS2 embeddings yields markedly improved resolution across all drug groups. Notably, samples corresponding to drowning, asphyxia, and hemorrhagic shock—conditions sharing hypoxic mechanisms—cluster closely together. These findings suggest that LSM-MS2 captures a more structured and biologically informative representation of metabolic variation than heuristic baselines.



Ovarian Cancer Detection from Serum

We validated LSM performance using publicly available serum lipidome data from 422 Korean women: 208 with ovarian cancer, 117 with other gynecological malignancies, and 97 controls (Sah et al., 2024). Raw UHPLC-HRMS files were processed directly by LSM without peak picking or metabolite identification.

Performance Metric	Conventional	LSM
False Negative Rate	22%	2%
False Positive Rate	24%	5%
Balanced Accuracy	74%	96%
Sensitivity	0.78 ± 0.15	0.98 ± 0.02
Specificity	0.76 ± 0.12	0.95 ± 0.05

The LSM approach represented a 10-fold increase in performance over conventional analysis using 10 manually curated lipid biomarkers. LSM embeddings showed clear clustering by phenotype with distinct separation between cancer types and benign samples. Some samples diagnosed as benign by histopathology clustered with cancer samples, suggesting LSM may detect biochemical signatures of early-stage malignancy.

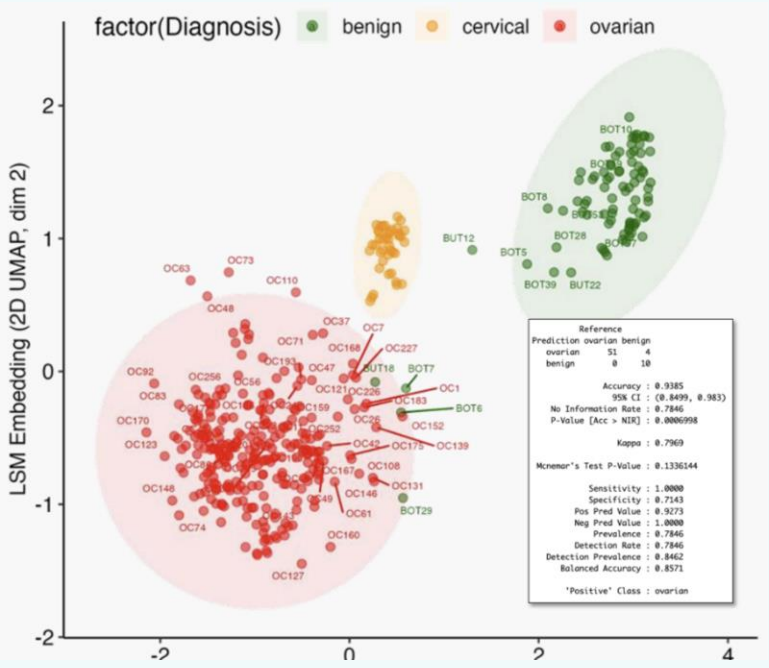
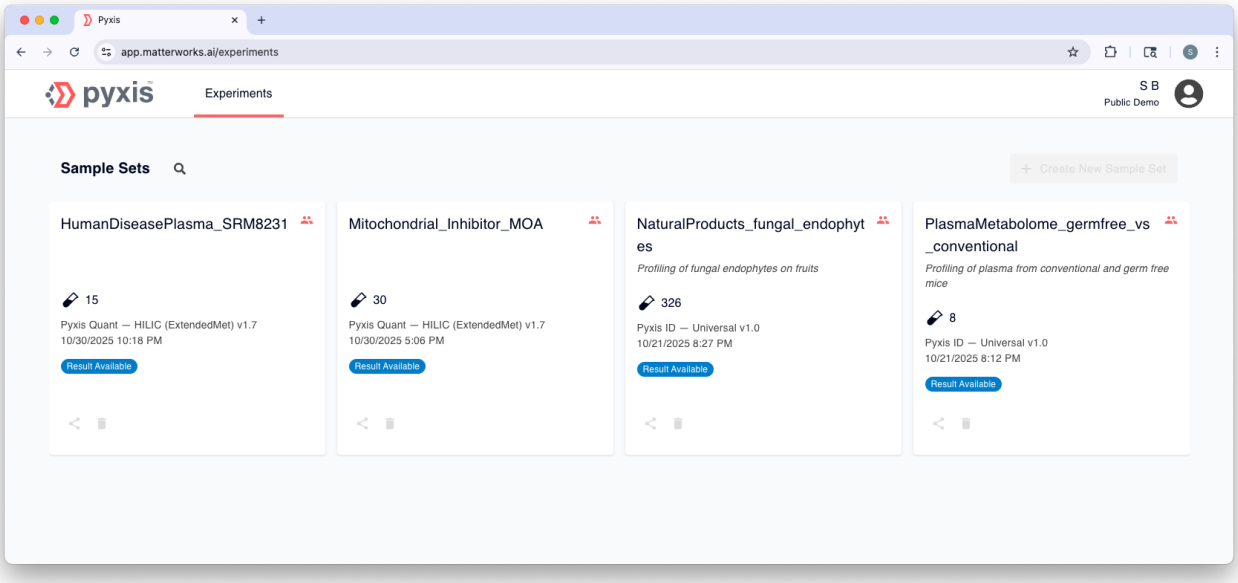


Figure 1: Clustering of clinical subjects using dimensionally reduced LSM embeddings reveals clear phenotypic grouping by histology.

CHAPTER SIX

See Pyxis in Action with Demo Data

Pyxis comes with demo datasets so you can explore the full platform and see results without uploading anything of your own. Get set up and running in 10 minutes. When you're ready, you can upload a dataset you've already analyzed. You'll see similar results, faster, and you may even learn something new.



Two Ways to Get Started

Upload your own data

Sign up and upload LC-MS/MS data directly. Results in minutes.

app.matterworks.ai

Let Matterworks run your samples

Need method dev, sample prep, or instrument time? We run it end-to-end.

info@matterworks.ai

Ready to see what's in your data?

Get started at app.matterworks.ai

Questions? Reach out at info@matterworks.ai

Matterworks