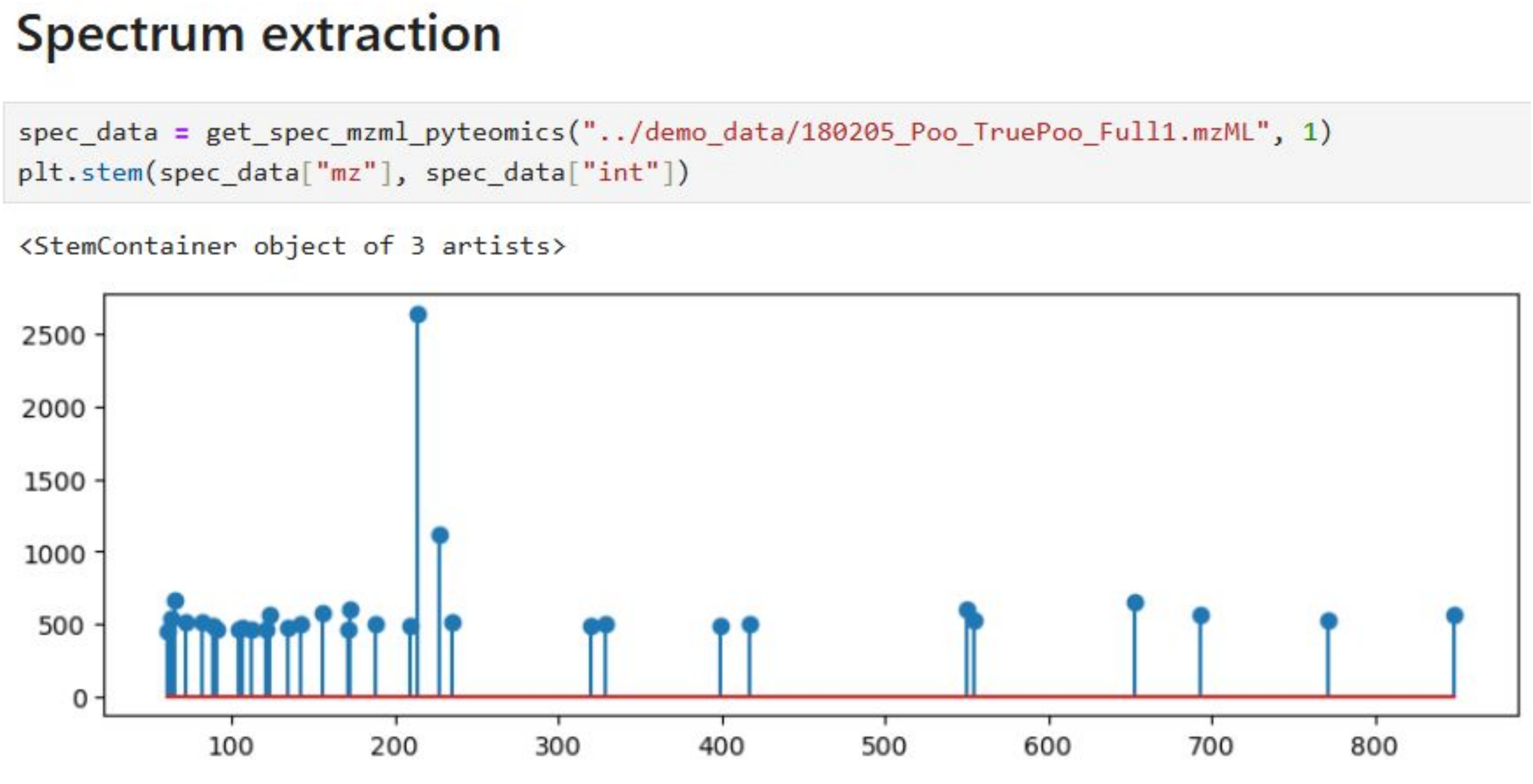
# Mass Spectrometry Data Formats

By Sam LaRue and Will Kumler

# The Problem

- **Raw MS Data Formats Limitations:**
  - Raw mass-spectrometry data is often stored in vendor-specific formats or .mzML files
  - These encode retention time, m/z ratio, and intensity
  - Existing formats lack intuitive, rapid, and easy to use search capabilities
  - Users must understand idiosyncratic file formats, which hinders accessibility and interoperability
- **Challenges with Current Methods:**
  - Difficulty in performing efficient queries and data extractions
  - Inefficient handling of multi-file data aggregation
  - Limited support for storing processed data alongside raw data
  - Reliance on formats not actively supported by larger development communities

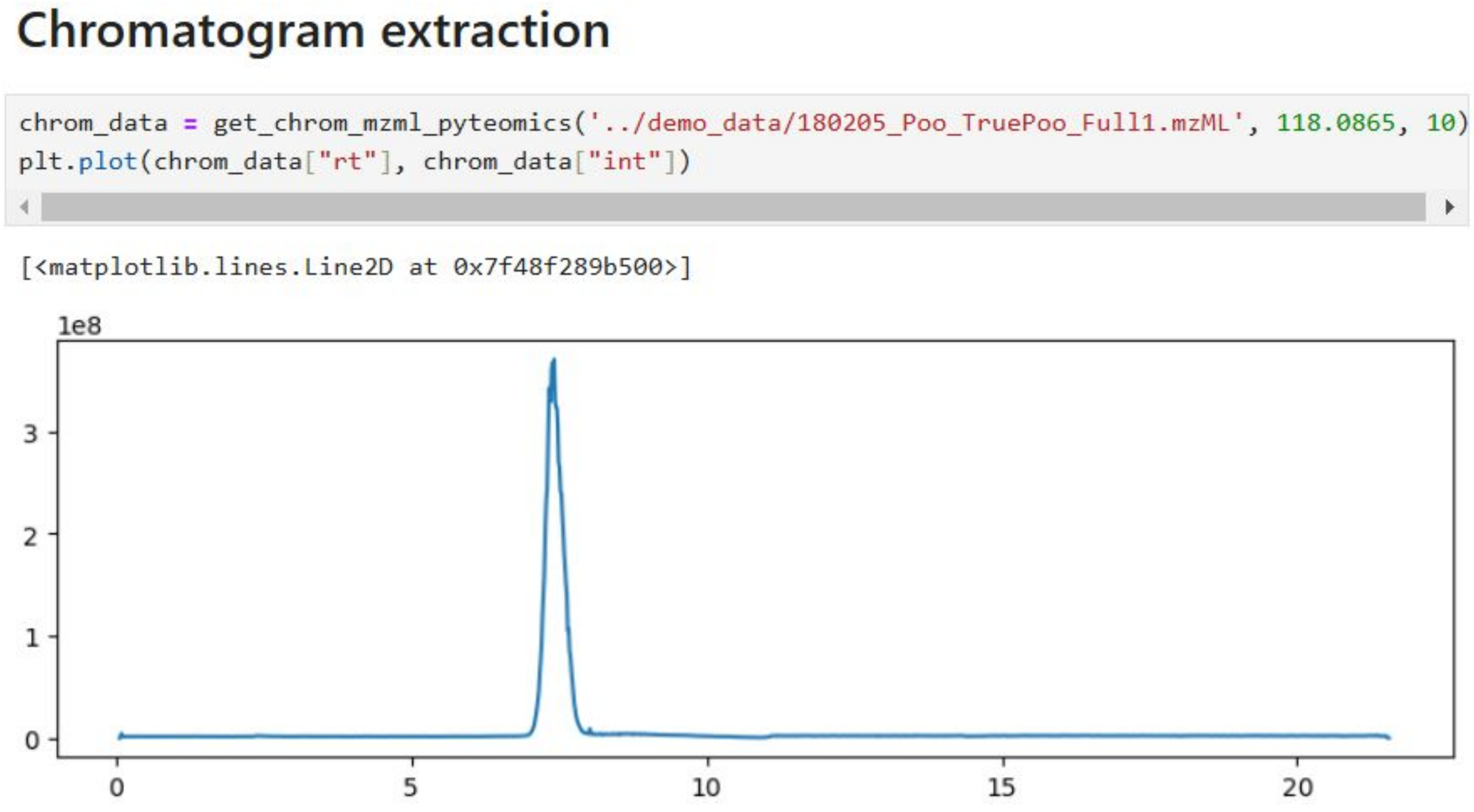


## Spectrum extraction

```
spec_data = get_spec_mzml_pyteomics("../demo_data/180205_Poo_TruePoo_Full1.mzML", 1)
plt.stem(spec_data["mz"], spec_data["int"])
```

```
<StemContainer object of 3 artists>
```

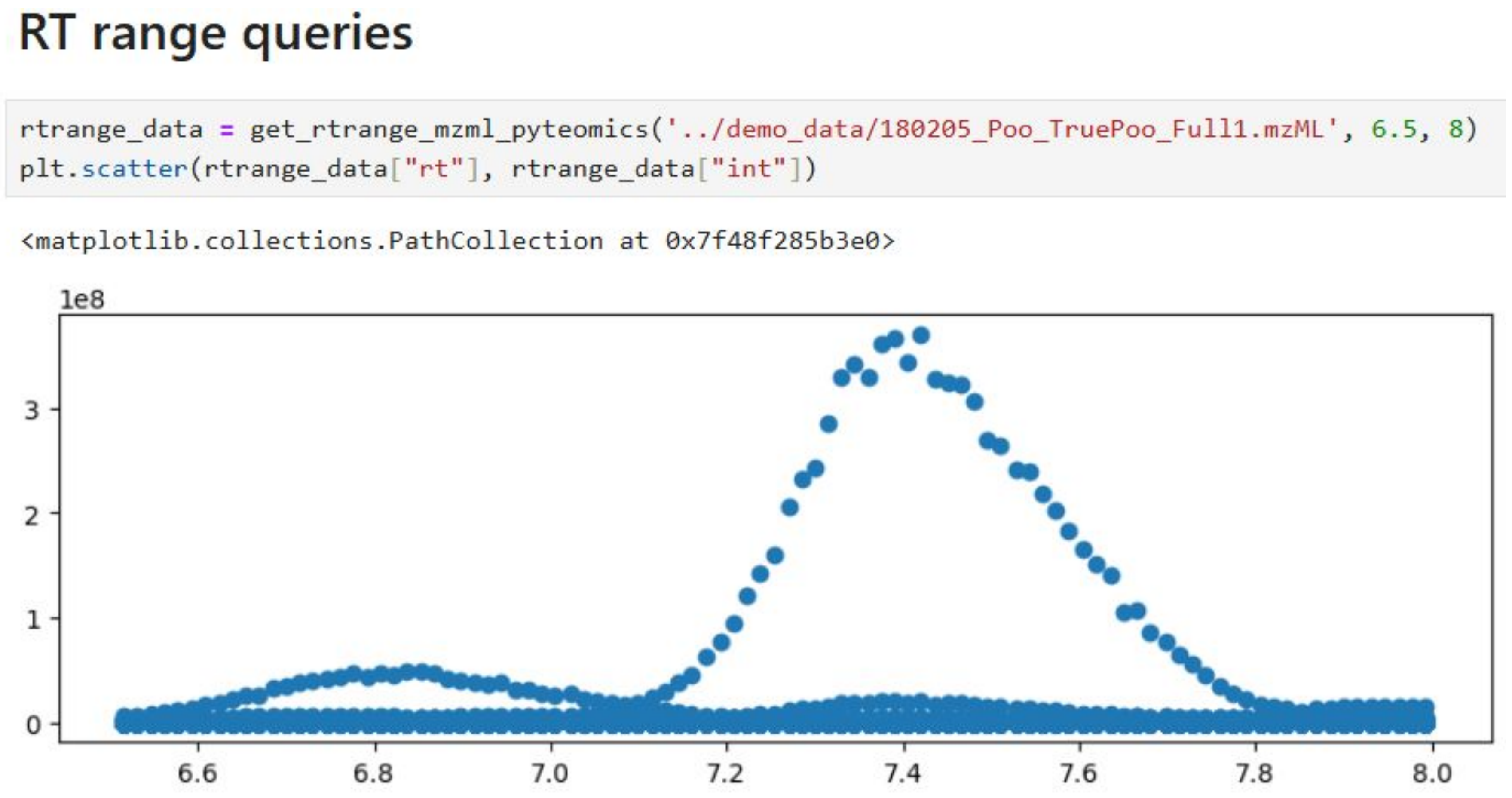
Problematic: only useful if the arbitrary ID number is somehow known in advance



# Chromatogram extraction

```
chrom_data = get_chrom_mzml_pyteomics('../demo_data/180205_Poo_TruePoo_Full1.mzML', 118.0865, 10)
plt.plot(chrom_data["rt"], chrom_data["int"])
```

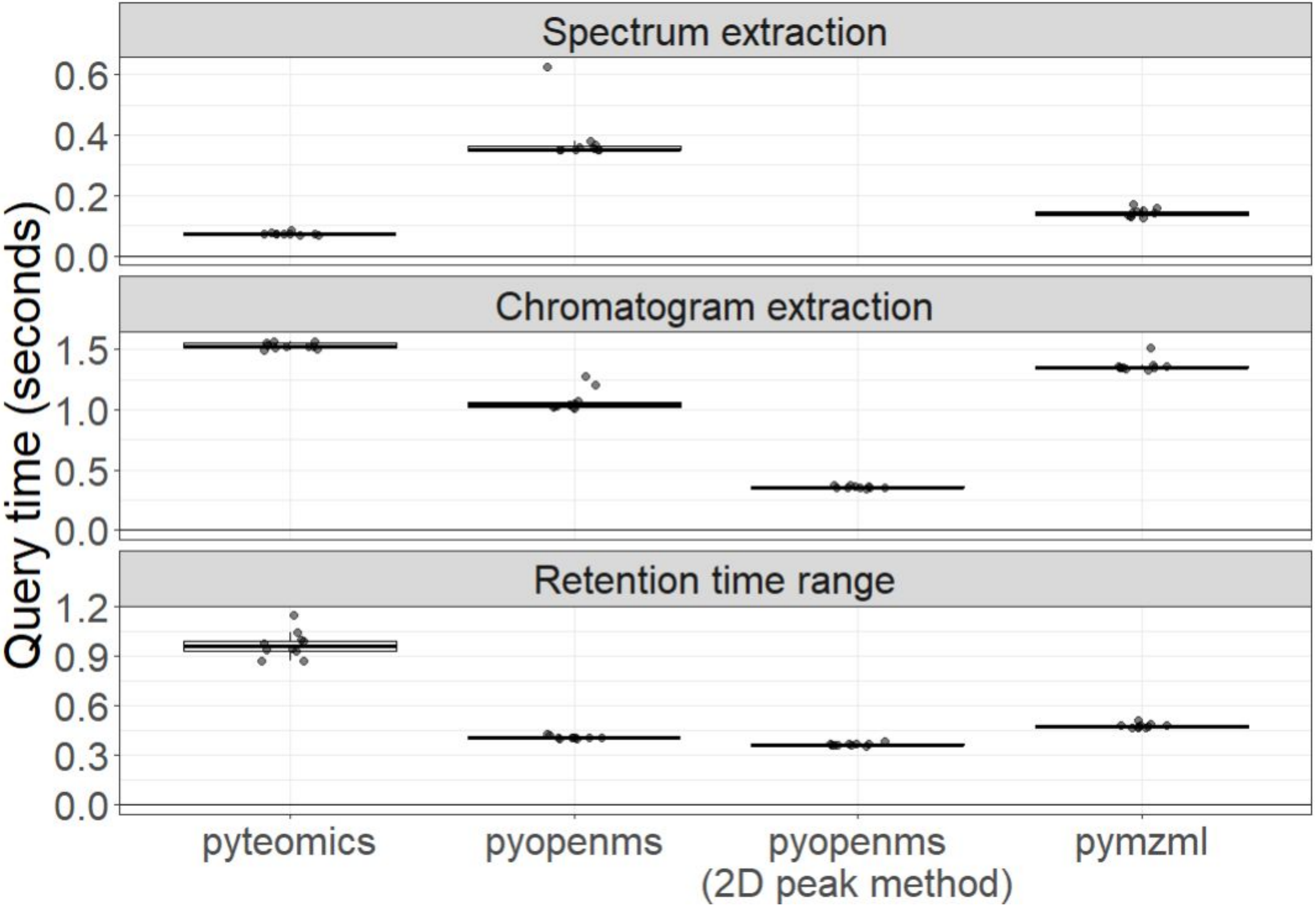
```
[<matplotlib.lines.Line2D at 0x7f48f289b500>]
```

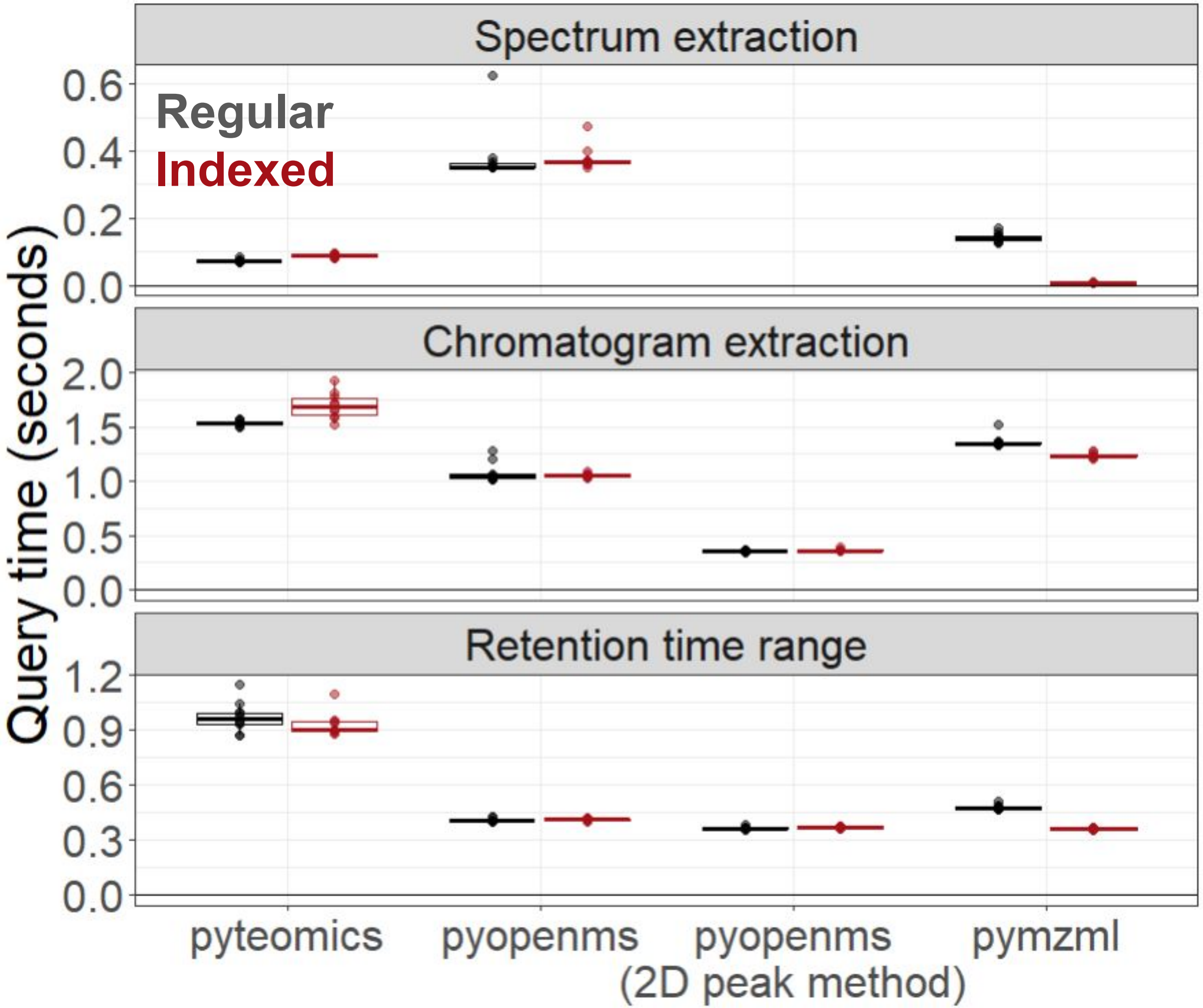


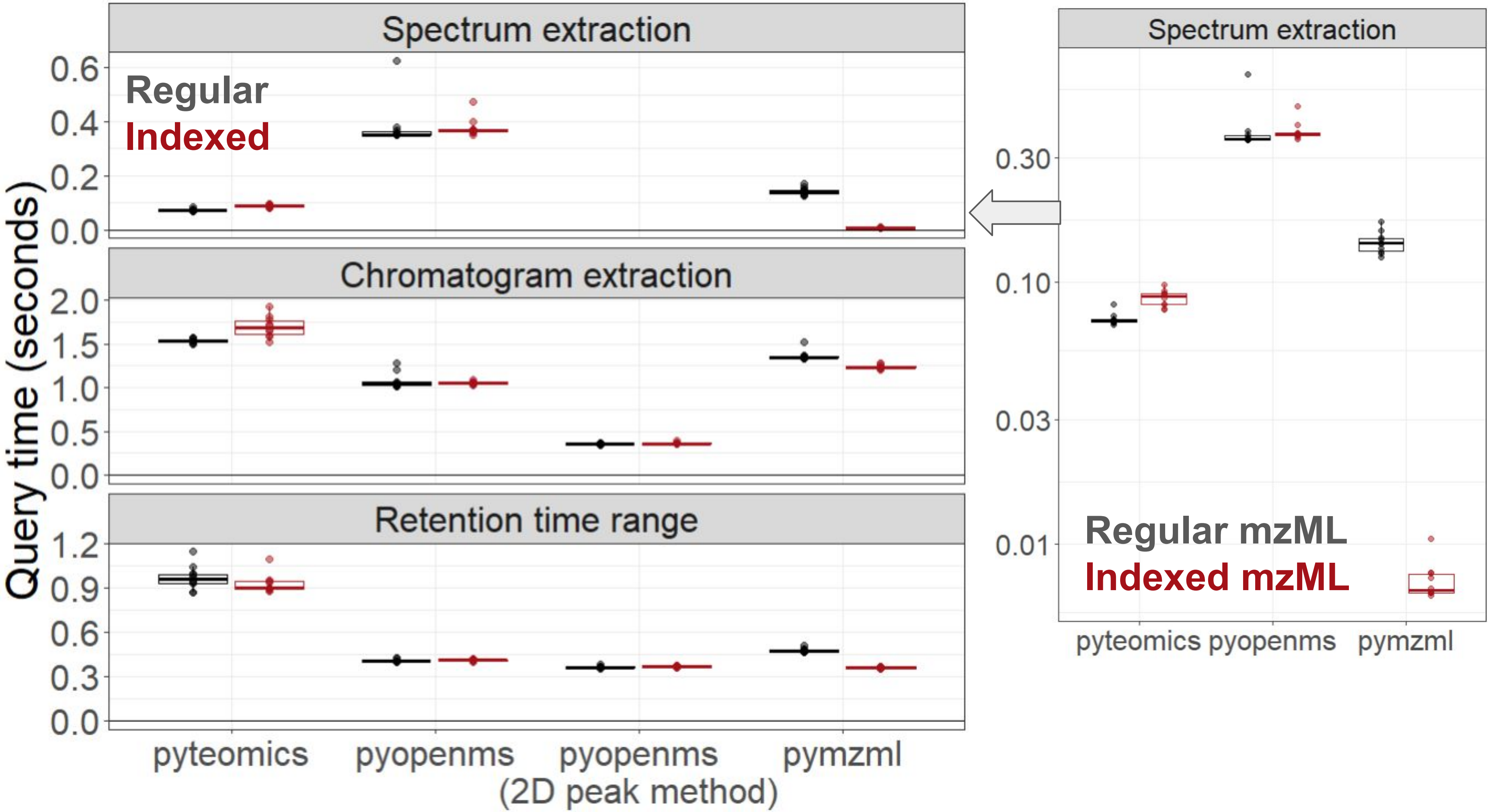
# RT range queries

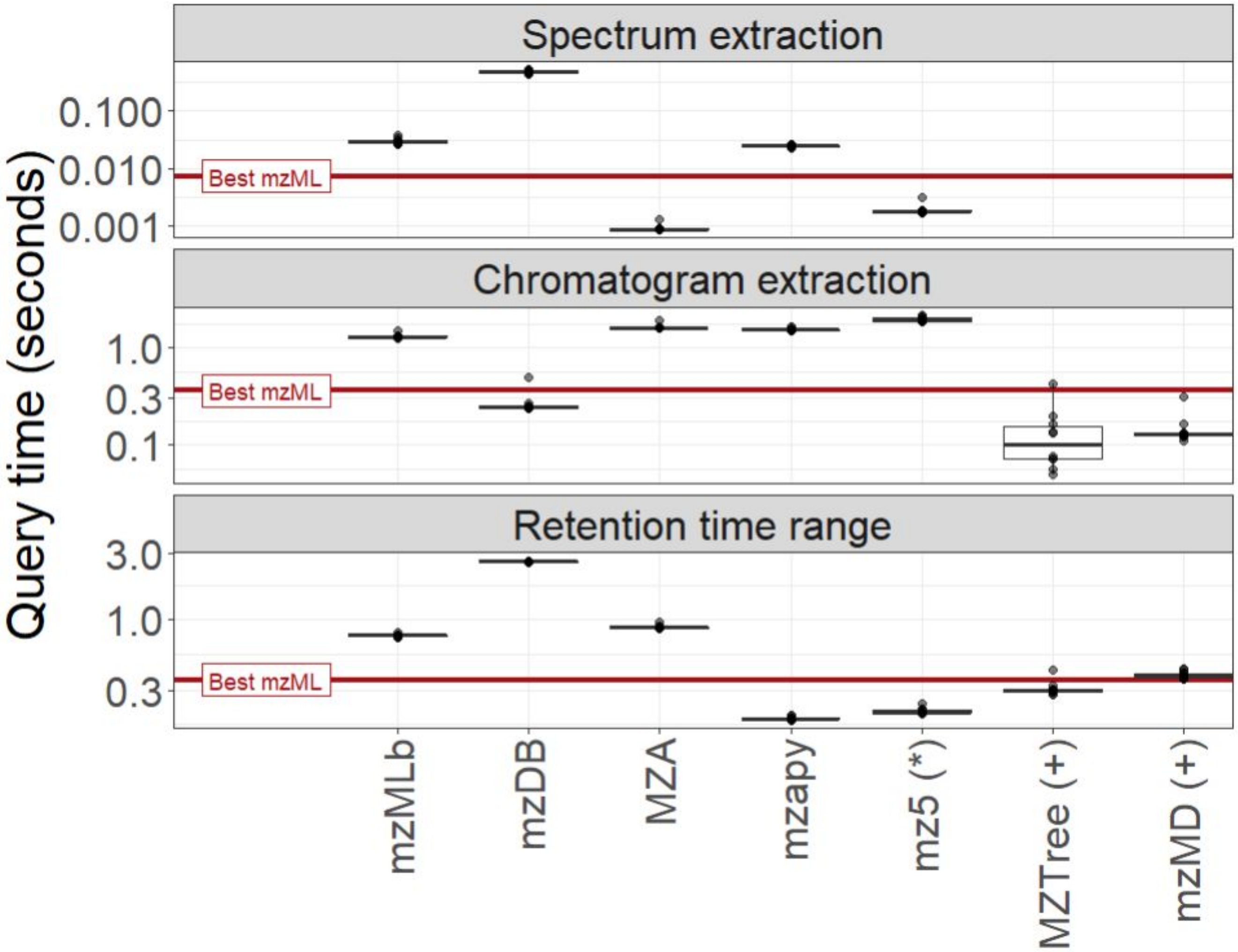
```
rtrange_data = get_rtrange_mzml_pyteomics('../demo_data/180205_Poo_TruePoo_Full1.mzML', 6.5, 8)
plt.scatter(rtrange_data["rt"], rtrange_data["int"])
```

```
<matplotlib.collections.PathCollection at 0x7f48f285b3e0>
```


Spectrum extraction
Chromatogram extraction
Retention time range
Query time (seconds)
0.6
0.4
0.2
0.0
1.5
1.0
0.5
0.0
1.2
0.9
0.6
0.3
0.0
pyteomics
pyopenms
pyopenms
(2D peak method)
pymzml

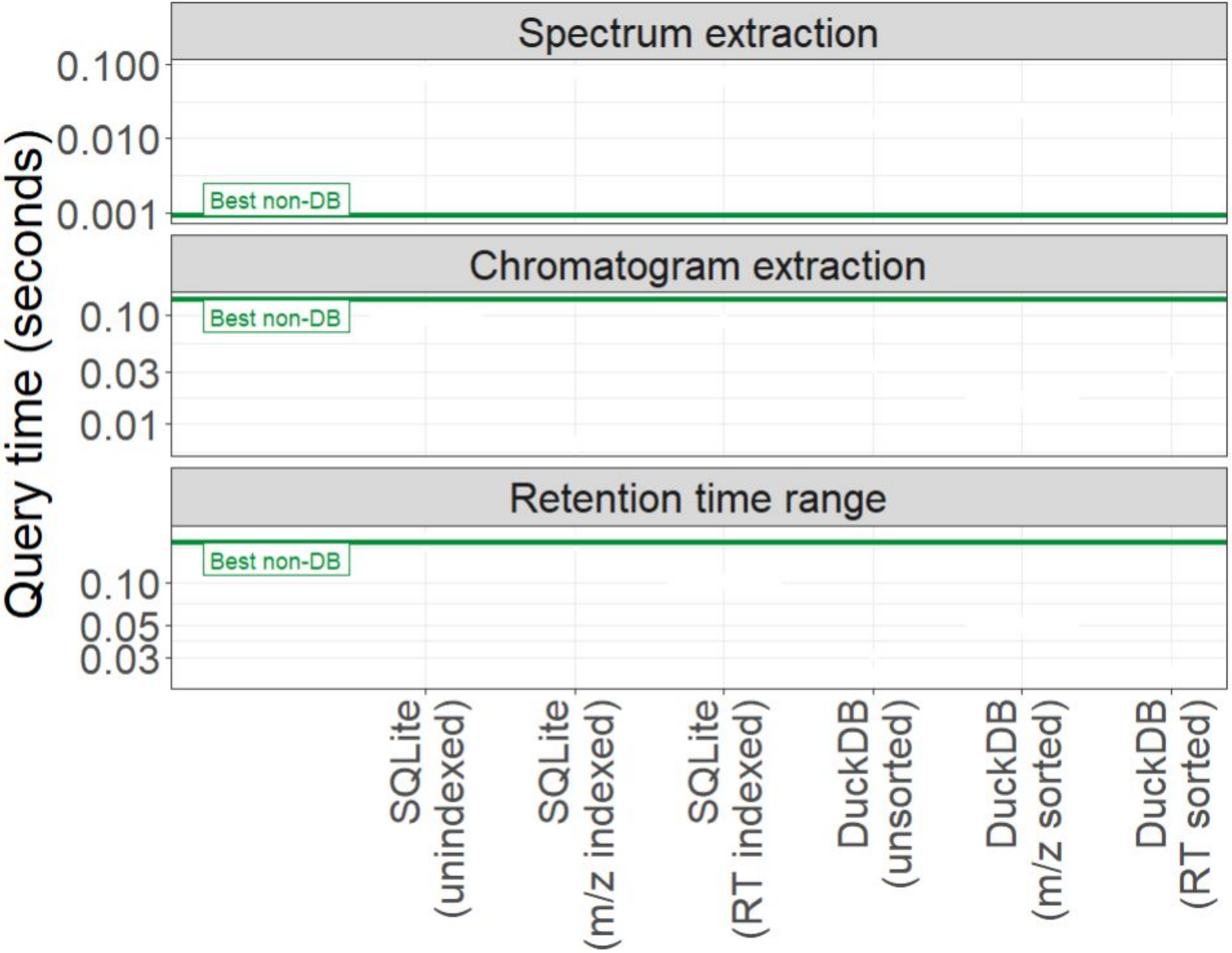

Spectrum extraction
Regular
Indexed
0.6
0.4
0.2
0.0
Chromatogram extraction
2.0
1.5
1.0
0.5
0.0
Retention time range
1.2
0.9
0.6
0.3
0.0
Query time (seconds)
pyteomics
pyopenms
pyopenms
(2D peak method)
pymzml


Spectrum extraction
Regular
Indexed
Chromatogram extraction
Retention time range
Query time (seconds)
0.6
0.4
0.2
0.0
2.0
1.5
1.0
0.5
0.0
1.2
0.9
0.6
0.3
0.0
pyteomics
pyopenms
pyopenms (2D peak method)
pymzml
Spectrum extraction
0.30
0.10
0.03
0.01
Regular mzML
Indexed mzML
pyteomics
pyopenms
pymzml



*mz5 returns slightly different values

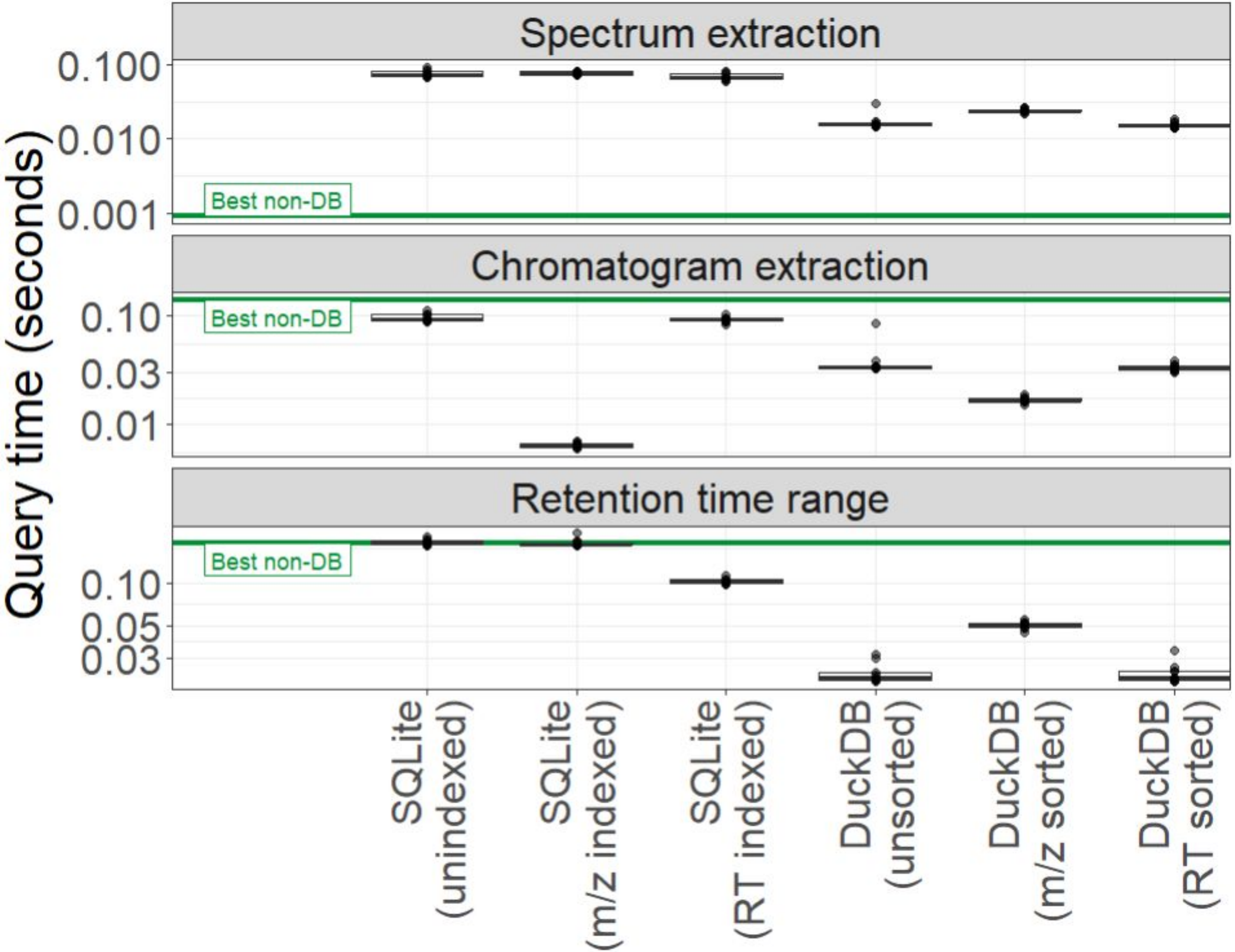
+mzTree/mzMD require manual loading



Do databases perform better than the existing file types?

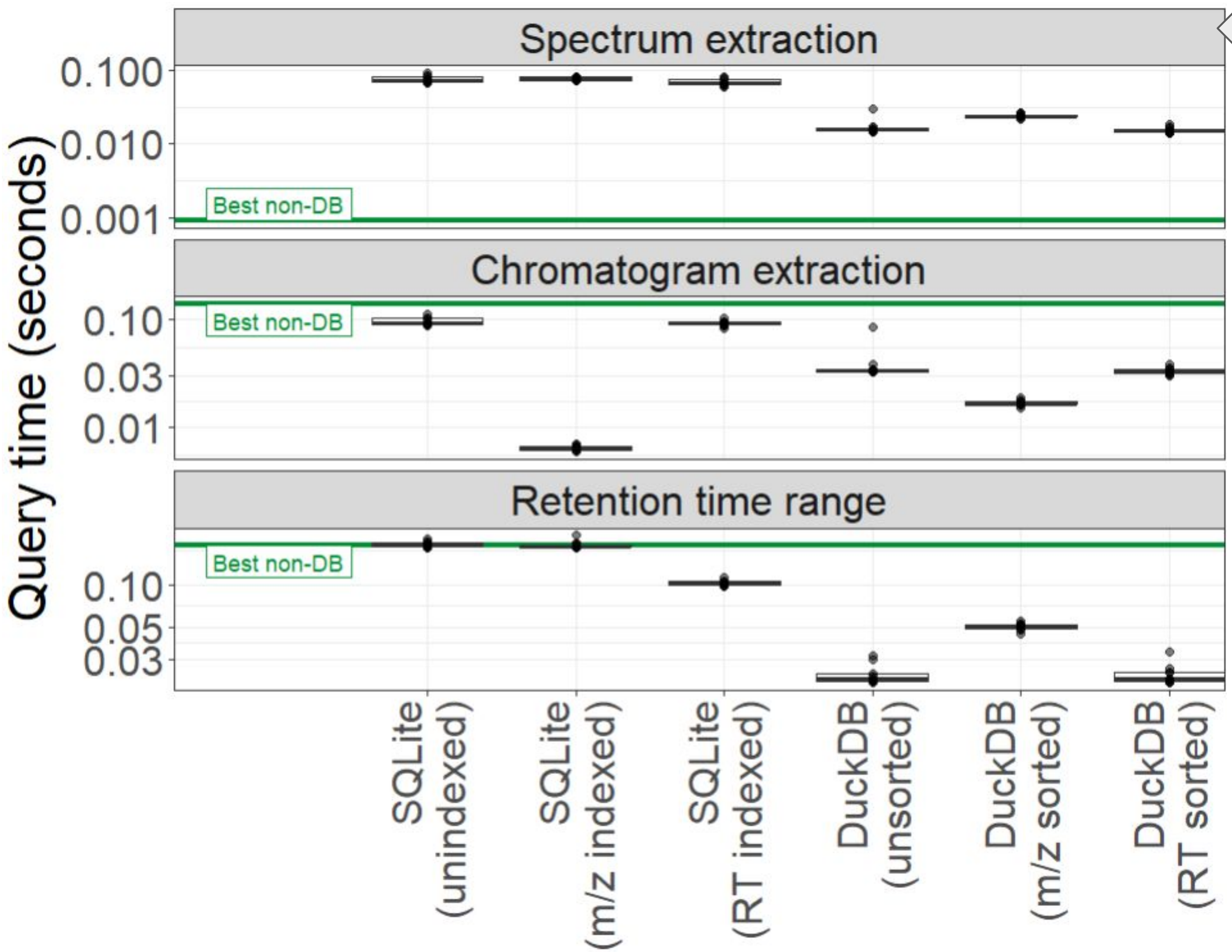
SQLite

DuckDB



SQLite

DuckDB

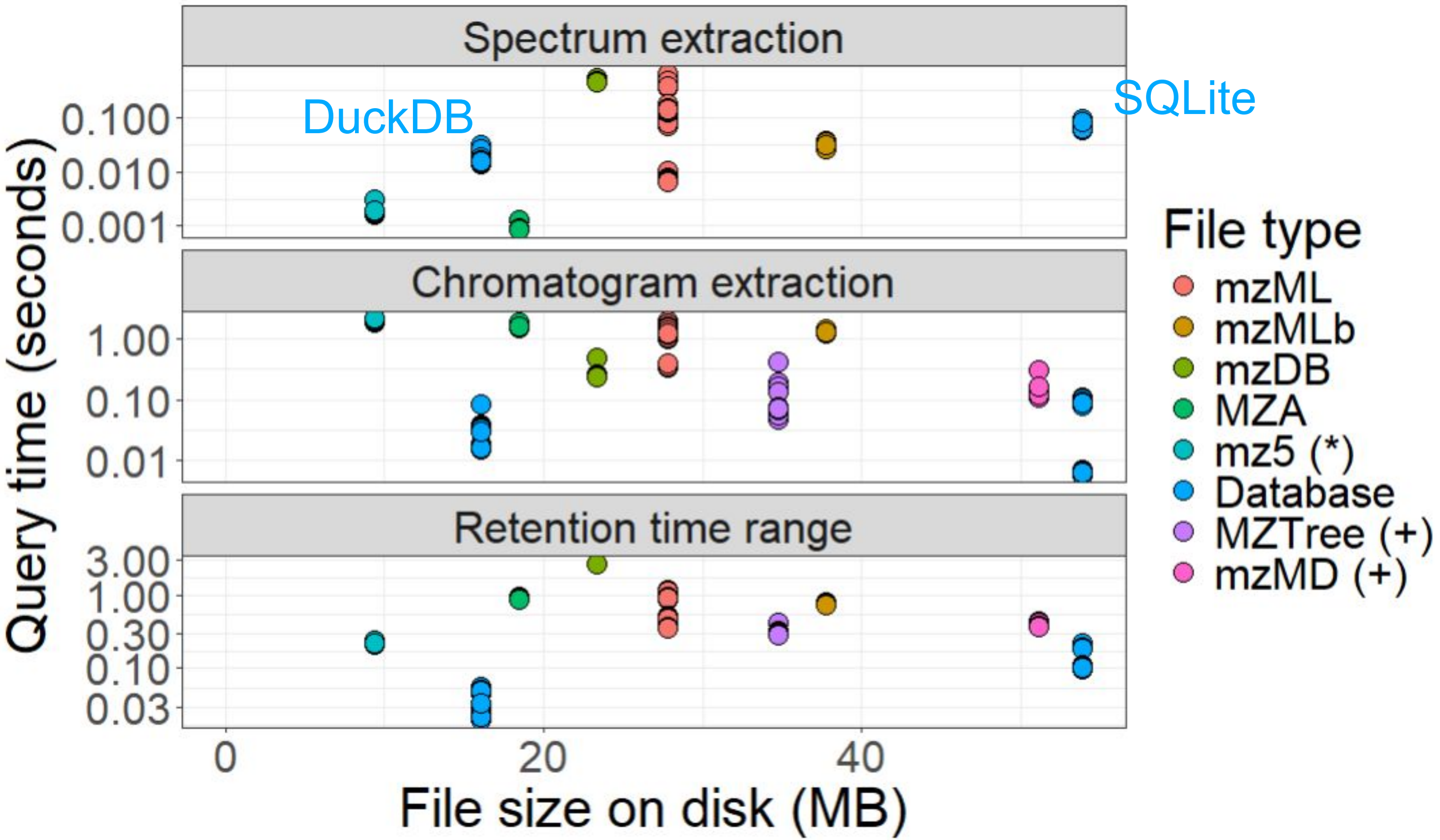


This metric was dumb anyway (so we didn't optimize for it)

SQLite

DuckDB

# Fundamental tradeoff between speed and size with DuckDB as an exception!



# Future Work

- Multi-file comparisons
    - Existing MS data formats preserve the idea of "one sample = one file"
    - Develop methods for managing MS datasets that contain multiple files
    - Create systems to integrate new data files into existing aggregated datasets without reprocessing
    - Current methods have a linear increase every time a file is added
    - Parallel processing techniques???
    - Bottlenecks or other issues???