Improved compound identification in non-target screening

David Cox, Craig Butt, Janna Anichina, Adrian Taylor | SCIEX

ASMS 2022

SCIEX
The Power of Precision

Searching chemical databases by m/z or molecular formula often generates a **very long list** of possible compounds. The **correct answer** is often the one that is **easiest to read**.

Compound ID difficulty: targeted, suspect, unknown

Targeted

Compound is in library, has been measured by this lab

- Mass error
- Retention time
- Ion ratio
- Fragment mass error
- Isotope match
- Spectral library match

Suspect

Compound is in library, has **never** been measured by this lab before

- Mass error
- Fragment mass error
- Isotope match
- Spectral library match

Unknown

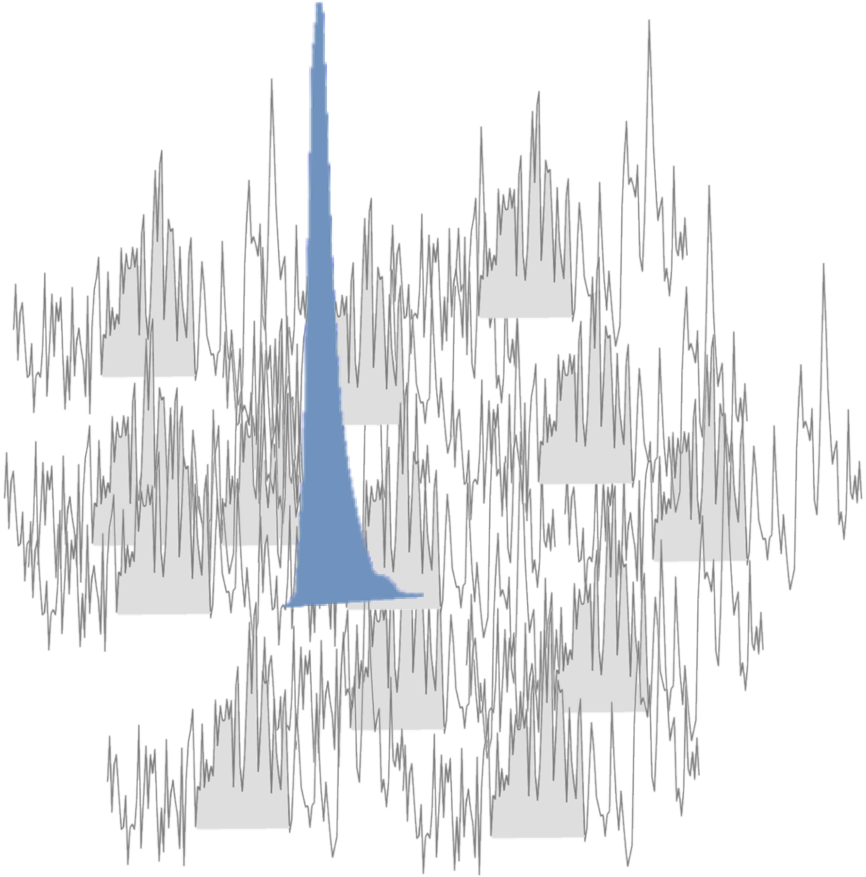
Compound is in PubChem database, no spectrum ever measured

- Mass error
- Isotope match
- Predict formula

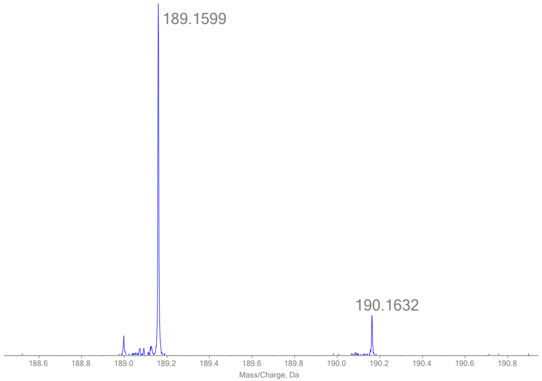
General Unknown

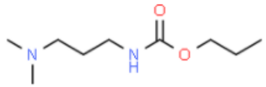
Not in PubChem database

Finding “real” features



Identifying a compound

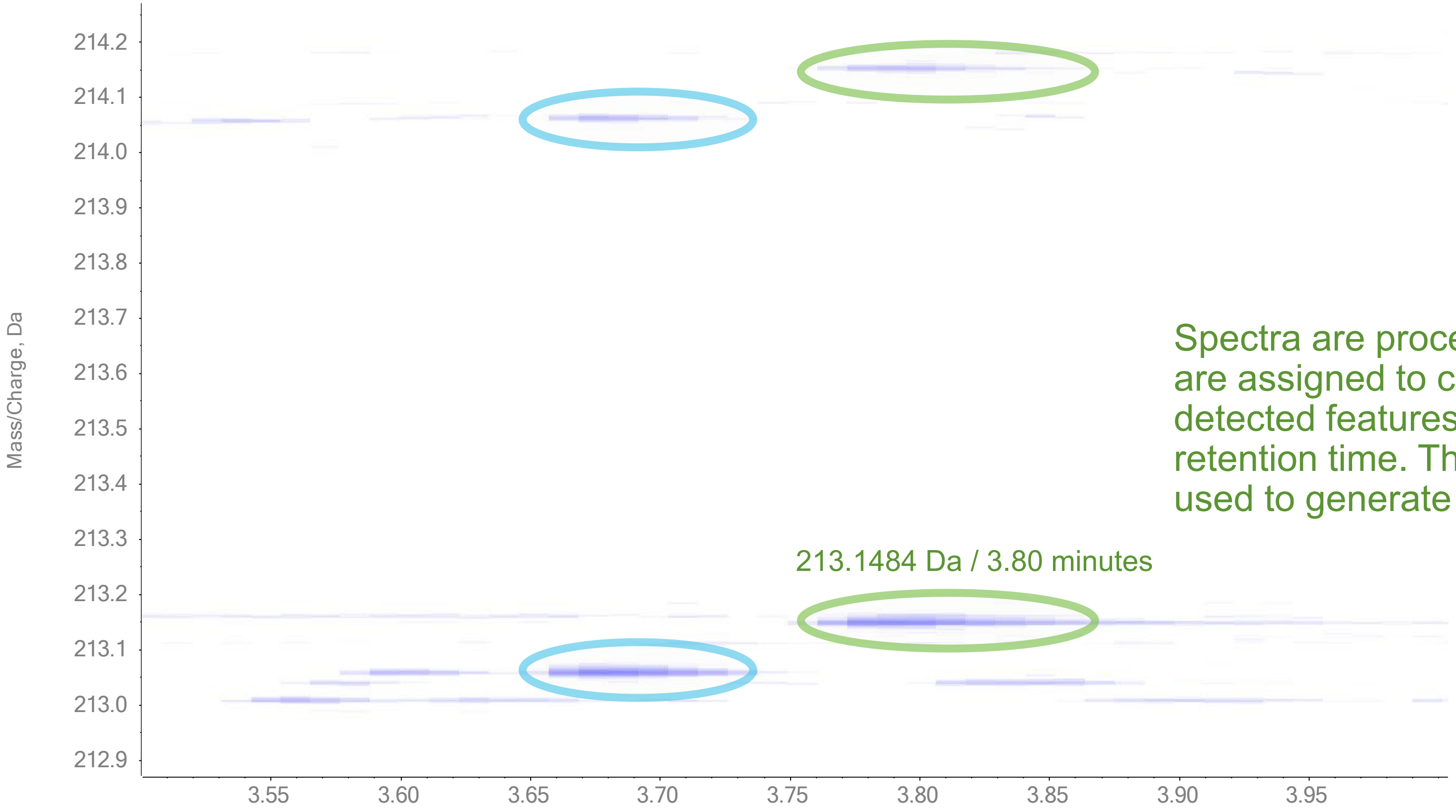




Propamocarb

Molecular Formula	$C_9H_{20}N_2O_2$
Average mass	188.267 Da
Monoisotopic mass	188.152481 Da
ChemSpider ID	30114

Feature finding

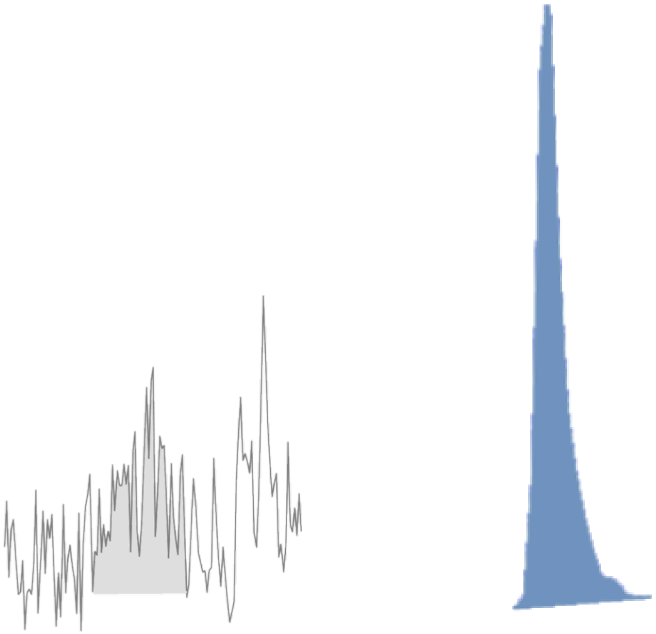


213.1484 Da / 3.80 minutes

Spectra are processed and m/z are assigned to clusters. The detected features are m/z at a retention time. This feature is used to generate XICs.

What is a “good” peak?

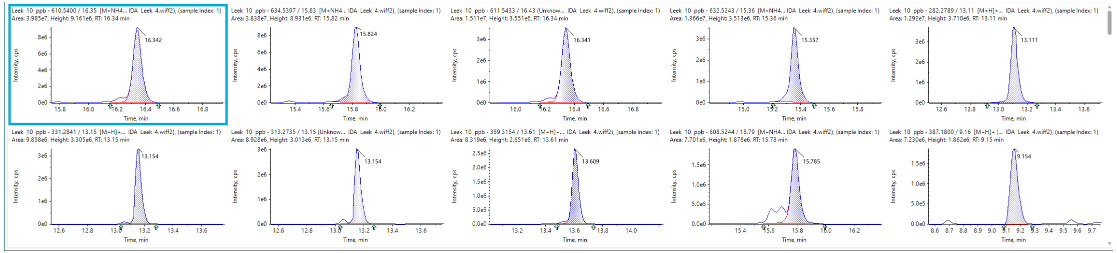
I can’t describe it, but I know it when I see it …



Using intensity to sort “good” from “bad” peaks

While features with a high intensity are typically real, and features with a low intensity are often junk, where do you draw a cut-off?

1st



Leek 10 ppb - 610.5400 / 16.35 [M+NH4... IDA Leek 4.wiff2), (sample Index: 1)
Area: 3.985e7, Height: 9.161e6, RT: 16.34 min

Leek 10 ppb - 634.5397 / 15.83 [M+NH4... IDA Leek 4.wiff2), (sample Index: 1)
Area: 3.838e7, Height: 8.931e6, RT: 15.82 min

Leek 10 ppb - 611.5433 / 16.43 (Unknow... IDA Leek 4.wiff2), (sample Index: 1)
Area: 1.511e7, Height: 3.551e6, RT: 16.34 min

Leek 10 ppb - 632.5243 / 15.36 [M+NH4... IDA Leek 4.wiff2), (sample Index: 1)
Area: 1.366e7, Height: 3.513e6, RT: 15.36 min

Leek 10 ppb - 282.2789 / 13.11 [M+H]+... IDA Leek 4.wiff2), (sample Index: 1)
Area: 1.292e7, Height: 3.710e6, RT: 13.11 min

Leek 10 ppb - 331.2841 / 13.15 [M+H]+... IDA Leek 4.wiff2), (sample Index: 1)
Area: 9.858e6, Height: 3.305e6, RT: 13.15 min

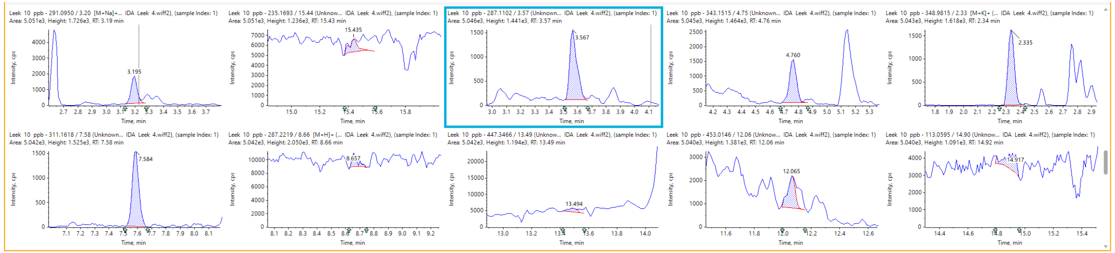
Leek 10 ppb - 313.2735 / 13.15 (Unknow... IDA Leek 4.wiff2), (sample Index: 1)
Area: 8.928e6, Height: 3.013e6, RT: 13.15 min

Leek 10 ppb - 359.3154 / 13.61 [M+H]+... IDA Leek 4.wiff2), (sample Index: 1)
Area: 8.319e6, Height: 2.651e6, RT: 13.61 min

Leek 10 ppb - 608.5244 / 15.79 [M+NH4... IDA Leek 4.wiff2), (sample Index: 1)
Area: 7.701e6, Height: 1.878e6, RT: 15.78 min

Leek 10 ppb - 387.1800 / 9.16 [M+H]+ (... IDA Leek 4.wiff2), (sample Index: 1)
Area: 7.230e6, Height: 1.862e6, RT: 9.15 min

8 971



Leek 10 ppb - 291.0950 / 3.20 [M+Na]+... IDA Leek 4.wiff2), (sample Index: 1)
Area: 5.051e3, Height: 1.726e3, RT: 3.19 min

Leek 10 ppb - 235.1693 / 15.44 (Unknow... IDA Leek 4.wiff2), (sample Index: 1)
Area: 5.051e3, Height: 1.236e3, RT: 15.43 min

Leek 10 ppb - 287.1102 / 3.57 (Unknown... IDA Leek 4.wiff2), (sample Index: 1)
Area: 5.046e3, Height: 1.441e3, RT: 3.57 min

Leek 10 ppb - 343.1515 / 4.75 (Unknown... IDA Leek 4.wiff2), (sample Index: 1)
Area: 5.045e3, Height: 1.464e3, RT: 4.76 min

Leek 10 ppb - 348.9815 / 2.33 [M+K]+ (... IDA Leek 4.wiff2), (sample Index: 1)
Area: 5.043e3, Height: 1.618e3, RT: 2.34 min

Leek 10 ppb - 311.1618 / 7.58 (Unknown... IDA Leek 4.wiff2), (sample Index: 1)
Area: 5.042e3, Height: 1.525e3, RT: 7.58 min

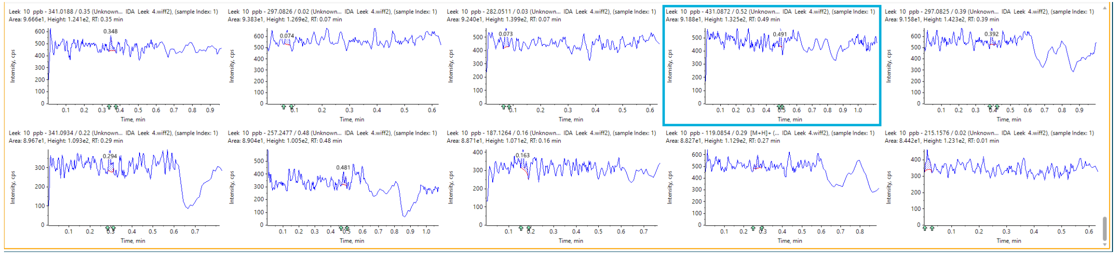
Leek 10 ppb - 287.2219 / 8.66 [M+H]+ (... IDA Leek 4.wiff2), (sample Index: 1)
Area: 5.042e3, Height: 2.050e3, RT: 8.66 min

Leek 10 ppb - 447.3466 / 13.49 (Unknow... IDA Leek 4.wiff2), (sample Index: 1)
Area: 5.042e3, Height: 1.194e3, RT: 13.49 min

Leek 10 ppb - 453.0146 / 12.06 (Unknow... IDA Leek 4.wiff2), (sample Index: 1)
Area: 5.040e3, Height: 1.381e3, RT: 12.06 min

Leek 10 ppb - 113.0595 / 14.90 (Unknow... IDA Leek 4.wiff2), (sample Index: 1)
Area: 5.040e3, Height: 1.091e3, RT: 14.92 min

13 394



Leek 10 ppb - 341.0188 / 0.35 (Unknown... IDA Leek 4.wiff2), (sample Index: 1)
Area: 9.666e1, Height: 1.241e2, RT: 0.35 min

Leek 10 ppb - 297.0826 / 0.02 (Unknown... IDA Leek 4.wiff2), (sample Index: 1)
Area: 9.383e1, Height: 1.269e2, RT: 0.07 min

Leek 10 ppb - 282.0511 / 0.03 (Unknown... IDA Leek 4.wiff2), (sample Index: 1)
Area: 9.240e1, Height: 1.399e2, RT: 0.07 min

Leek 10 ppb - 431.0872 / 0.52 (Unknown... IDA Leek 4.wiff2), (sample Index: 1)
Area: 9.188e1, Height: 1.325e2, RT: 0.49 min

Leek 10 ppb - 297.0825 / 0.39 (Unknown... IDA Leek 4.wiff2), (sample Index: 1)
Area: 9.158e1, Height: 1.423e2, RT: 0.39 min

Leek 10 ppb - 341.0934 / 0.22 (Unknown... IDA Leek 4.wiff2), (sample Index: 1)
Area: 8.967e1, Height: 1.093e2, RT: 0.29 min

Leek 10 ppb - 257.2477 / 0.48 (Unknown... IDA Leek 4.wiff2), (sample Index: 1)
Area: 8.904e1, Height: 1.005e2, RT: 0.48 min

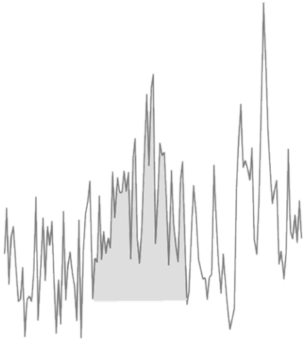
Leek 10 ppb - 187.1264 / 0.16 (Unknown... IDA Leek 4.wiff2), (sample Index: 1)
Area: 8.871e1, Height: 1.071e2, RT: 0.16 min

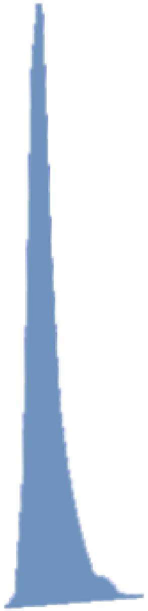
Leek 10 ppb - 119.0854 / 0.29 [M+H]+ (... IDA Leek 4.wiff2), (sample Index: 1)
Area: 8.827e1, Height: 1.129e2, RT: 0.27 min

Leek 10 ppb - 215.1576 / 0.02 (Unknown... IDA Leek 4.wiff2), (sample Index: 1)
Area: 8.442e1, Height: 1.231e2, RT: 0.01 min

What is a “good” peak?

We can use other measures.
But how do we decide on cut-offs? How to keep track of many measures at once?



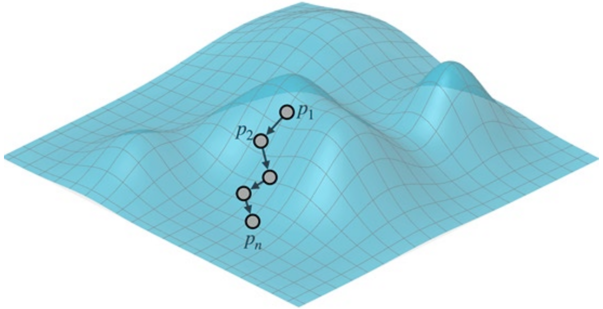


Intensity

S:N, Area Height, Region Height, Quality, Retention Time Delta min_, Total Width, Width at 50 , Baseline Delta Height, Width at 5, Width at 10, Slope of Baseline, Tailing Factor, Asymmetry Factor, Points Across Baseline, Points Across Half Height, AutoPeak Asymmetry, AutoPeak Candidate Model Quality, AutoPeak Group Confidence, AutoPeak Integration Quality, AutoPeak Num Peaks, AutoPeak Peak Width Confidence

Enabling machine learning with Analytics prototype in SCIEX OS software

- Modified version of processing software
- Used Microsoft's open source ML.NET framework
- Implemented a classifier that uses the Stochastic Dual Coordinate Ascent (SDCA) method
  - Trained using a custom column in the results table, where a user enters classifications ("good" or "bad")
  - Results of classification are output to new columns (predicted classification and score)



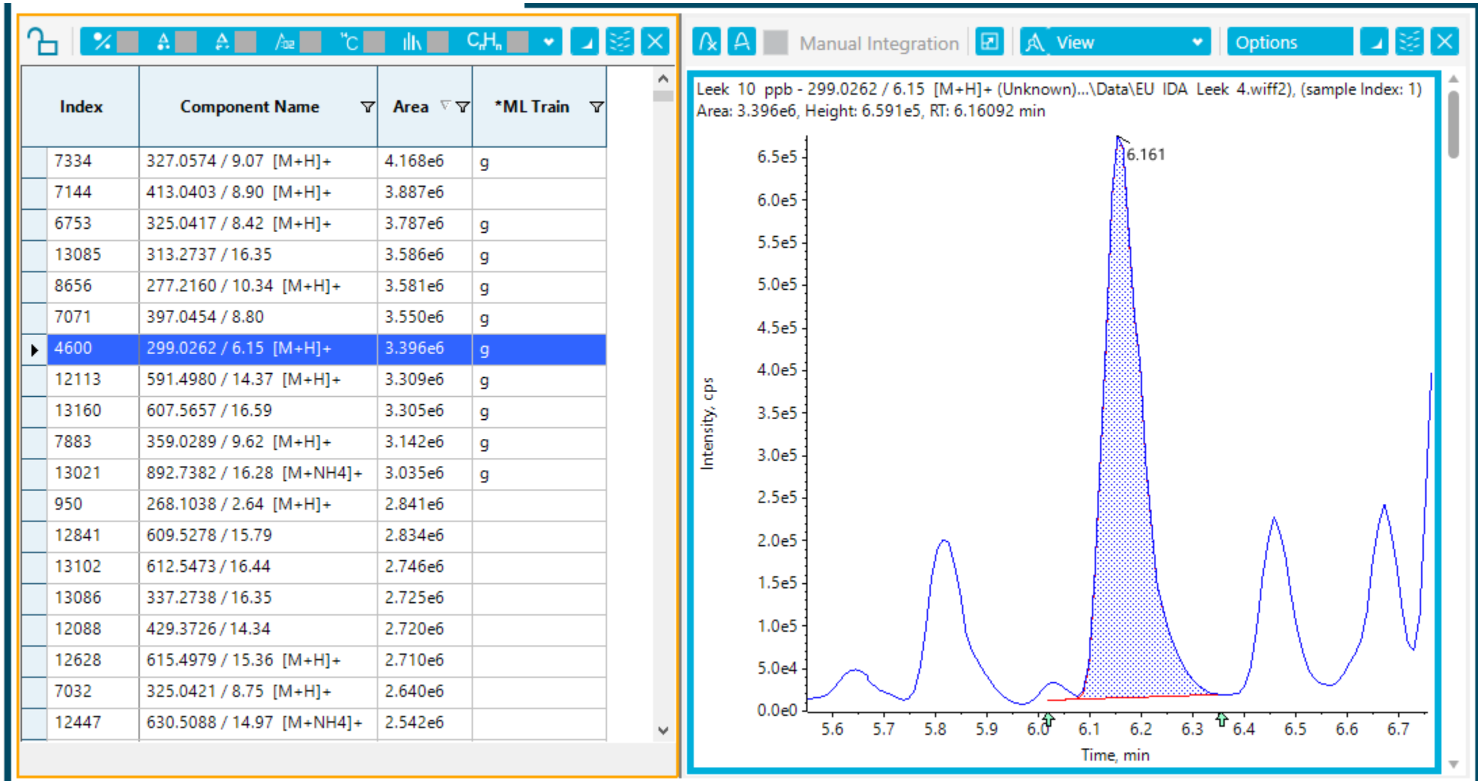
https://livebook.manning.com/book/grokking-machine-learning/appendix-b/v-15/29

- For more details on ML.NET and machine learning in general:

  https://dotnet.microsoft.com/en-us/apps/machinelearning-ai/ml-dotnet
  https://rubikscode.net/2021/02/01/machine-learning-with-ml-net-ultimate-guide-to-classification/
  https://www.kdnuggets.com/2020/05/5-concepts-gradient-descent-cost-function.html
  https://www.jmlr.org/papers/v14/gonen13a.html

Review a few peaks to train the model

With minimal keystrokes (↑ ↓ g b), classification of a **few hundred** peaks is fast and easy.
This set of data is used to train the model so it can be used to automatically classify the remaining ~13 000 features.

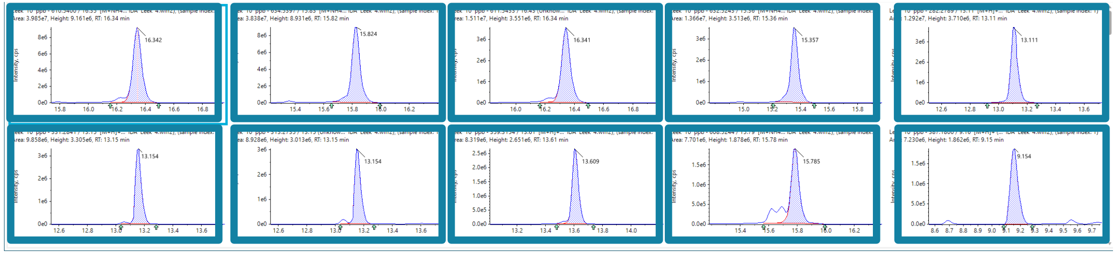


Index	Component Name	Area	*ML Train
7334	327.0574 / 9.07 [M+H]+	4.168e6	g
7144	413.0403 / 8.90 [M+H]+	3.887e6	
6753	325.0417 / 8.42 [M+H]+	3.787e6	g
13085	313.2737 / 16.35	3.586e6	g
8656	277.2160 / 10.34 [M+H]+	3.581e6	g
7071	397.0454 / 8.80	3.550e6	g
4600	299.0262 / 6.15 [M+H]+	3.396e6	g
12113	591.4980 / 14.37 [M+H]+	3.309e6	g
13160	607.5657 / 16.59	3.305e6	g
7883	359.0289 / 9.62 [M+H]+	3.142e6	g
13021	892.7382 / 16.28 [M+NH4]+	3.035e6	g
950	268.1038 / 2.64 [M+H]+	2.841e6	
12841	609.5278 / 15.79	2.834e6	
13102	612.5473 / 16.44	2.746e6	
13086	337.2738 / 16.35	2.725e6	
12088	429.3726 / 14.34	2.720e6	
12628	615.4979 / 15.36 [M+H]+	2.710e6	
7032	325.0421 / 8.75 [M+H]+	2.640e6	
12447	630.5088 / 14.97 [M+NH4]+	2.542e6	

Leek 10 ppb - 299.0262 / 6.15 [M+H]+ (Unknown)...\Data\EU IDA Leek 4.wiff2), (sample Index: 1)
Area: 3.396e6, Height: 6.591e5, RT: 6.16092 min

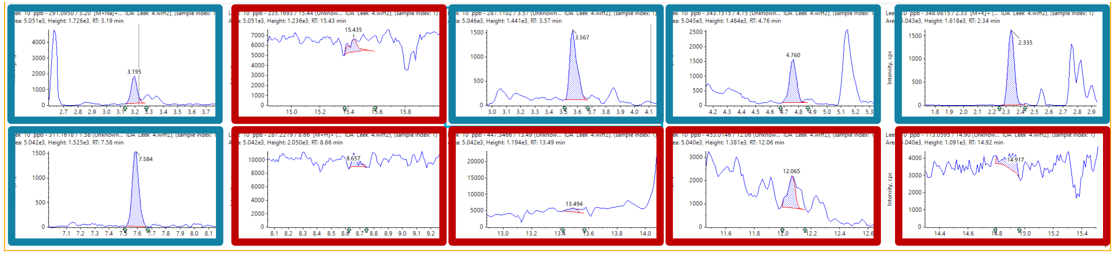
Machine learning classification finds more “real” features

1st

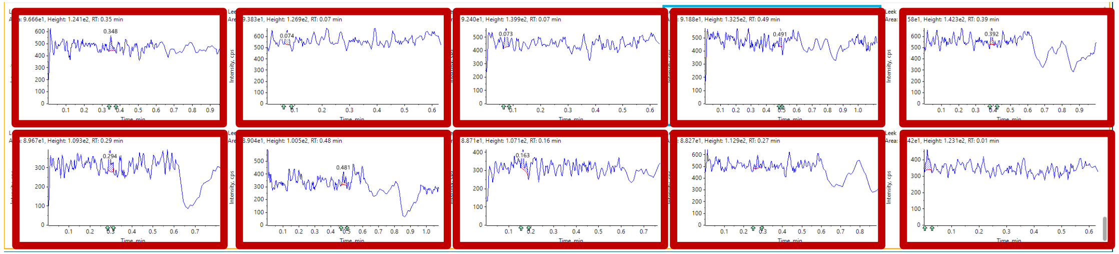


Deciding what to focus our attention on is much easier now. We can still find true features, even at low intensities.

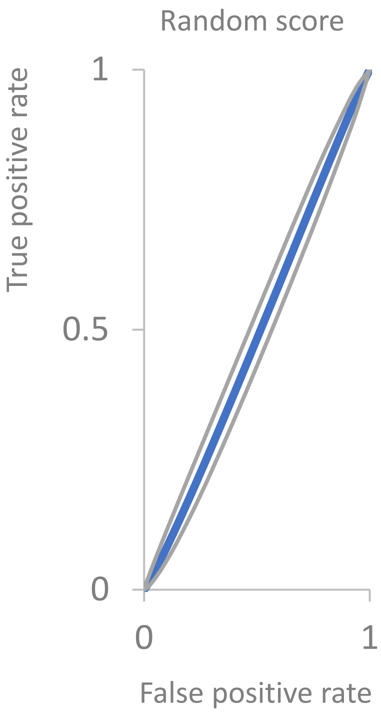
8 971



13 394

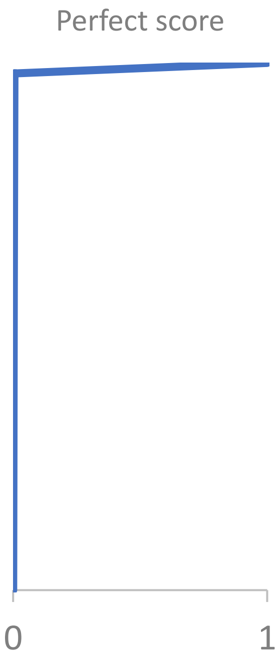


How well does machine learning perform compared to other scoring?

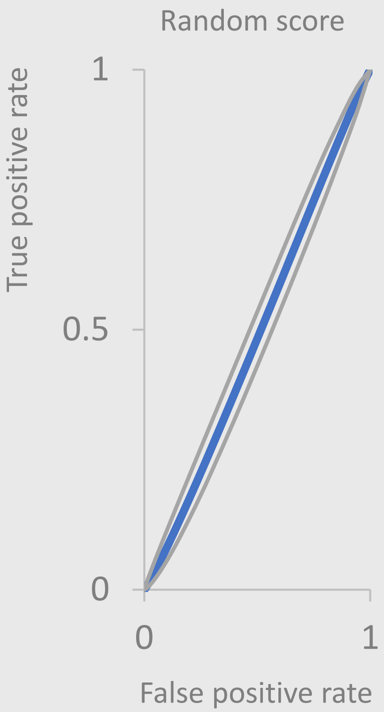


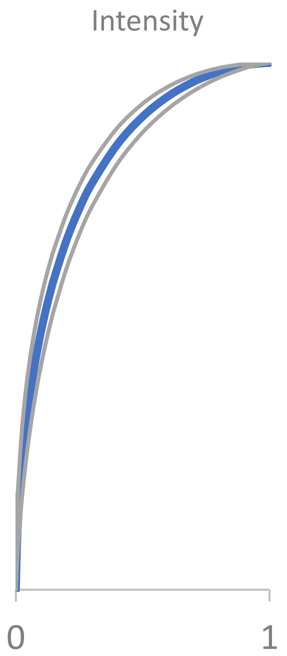
A **randomly** generated number has **no predictive** power. It generates as many true positives as false positives.

A (mythical) perfect scoring algorithm would obtain all true positives before any false positives occur.

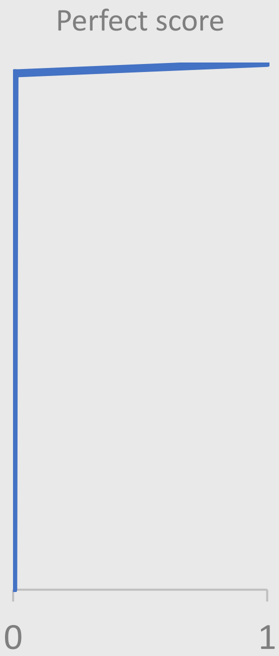


Intensity sorting is better than a random score ¯_(ツ)_/¯

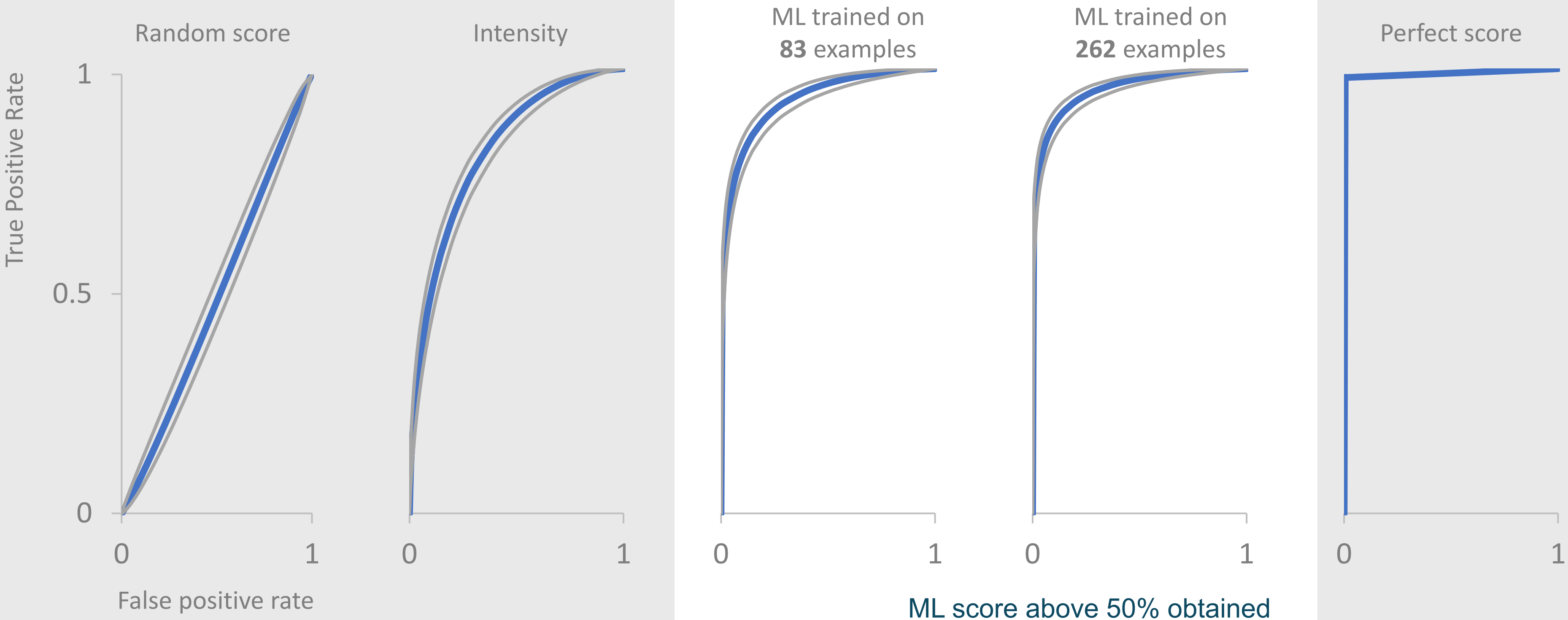




For example, classifying peaks above 0.004% relative intensity obtained **90% of the 559 true peaks** in the test set, but generated **222 false positives**.

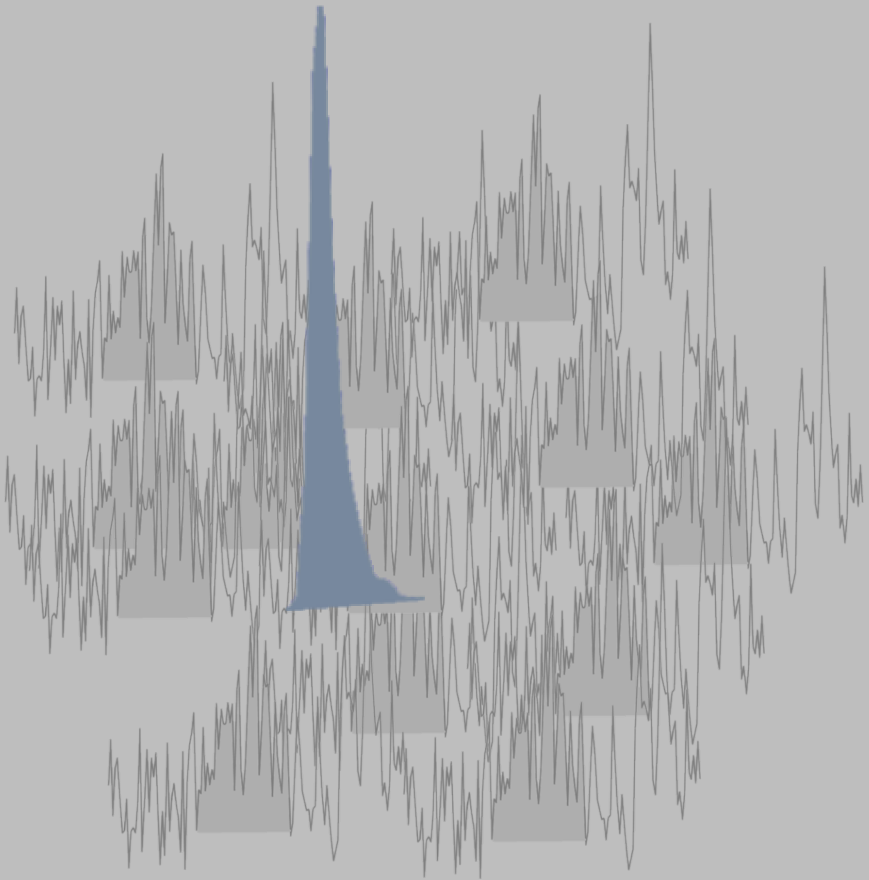


Machine learning scoring performs better than intensity scoring

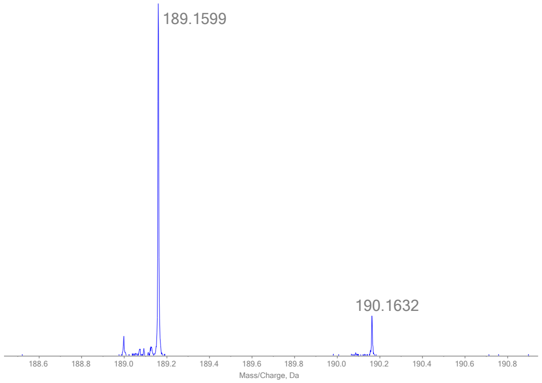


ML score above 50% obtained **90% of the 559 true peaks** in the test set, and only generated **56 false positives**.

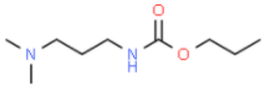
Finding “real” features



Identifying a compound



?

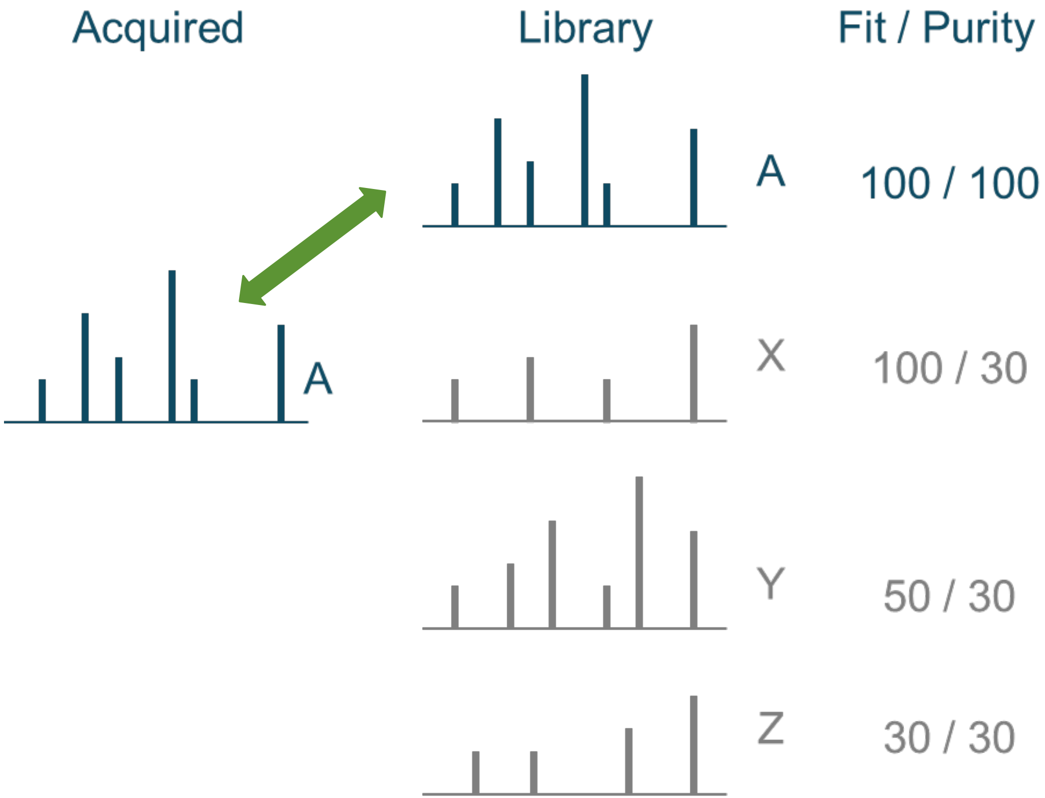


Propamocarb

Molecular Formula	$C_9H_{20}N_2O_2$
Average mass	188.267 Da
Monoisotopic mass	188.152481 Da
ChemSpider ID	30114

Library searching is common for identification

This only works if the MS/MS for the compound has been seen before, and has been entered into the spectral library you have.



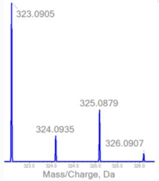
Acquired

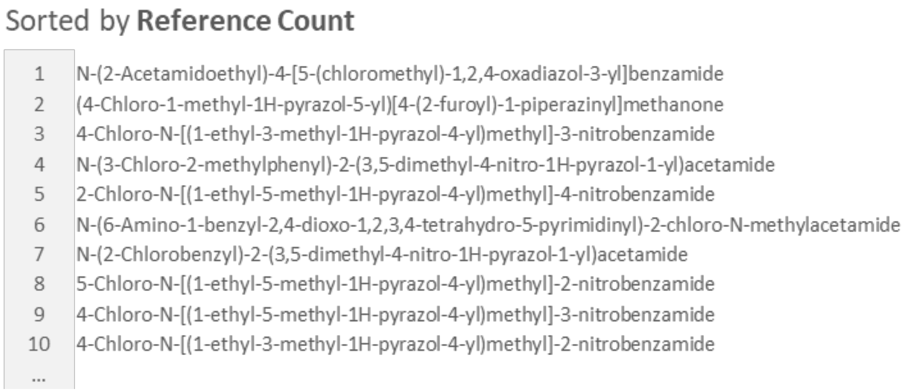
A

Library	Fit / Purity
A	100 / 100
X	100 / 30
Y	50 / 30
Z	30 / 30

Searching a formula in online databases

$C_{14}H_{15}ClN_{4}O_{3}$

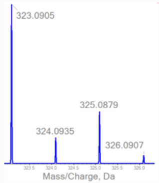



Sorted by **Reference Count**

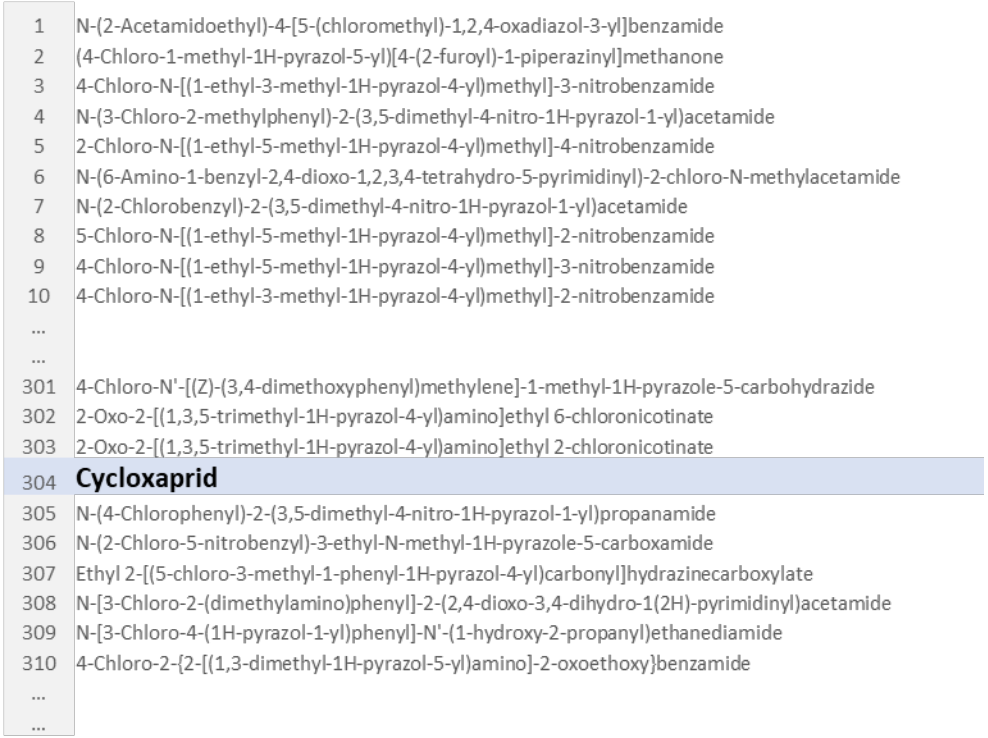
1	N-(2-Acetamidoethyl)-4-[5-(chloromethyl)-1,2,4-oxadiazol-3-yl]benzamide
2	(4-Chloro-1-methyl-1H-pyrazol-5-yl)[4-(2-furoyl)-1-piperazinyl]methanone
3	4-Chloro-N-[(1-ethyl-3-methyl-1H-pyrazol-4-yl)methyl]-3-nitrobenzamide
4	N-(3-Chloro-2-methylphenyl)-2-(3,5-dimethyl-4-nitro-1H-pyrazol-1-yl)acetamide
5	2-Chloro-N-[(1-ethyl-5-methyl-1H-pyrazol-4-yl)methyl]-4-nitrobenzamide
6	N-(6-Amino-1-benzyl-2,4-dioxo-1,2,3,4-tetrahydro-5-pyrimidinyl)-2-chloro-N-methylacetamide
7	N-(2-Chlorobenzyl)-2-(3,5-dimethyl-4-nitro-1H-pyrazol-1-yl)acetamide
8	5-Chloro-N-[(1-ethyl-5-methyl-1H-pyrazol-4-yl)methyl]-2-nitrobenzamide
9	4-Chloro-N-[(1-ethyl-5-methyl-1H-pyrazol-4-yl)methyl]-3-nitrobenzamide
10	4-Chloro-N-[(1-ethyl-3-methyl-1H-pyrazol-4-yl)methyl]-2-nitrobenzamide
...	

Searching a formula in online databases

$C_{14}H_{15}ClN_4O_3$



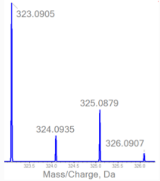
Sorted by **Reference Count**



1	N-(2-Acetamidoethyl)-4-[5-(chloromethyl)-1,2,4-oxadiazol-3-yl]benzamide
2	(4-Chloro-1-methyl-1H-pyrazol-5-yl)[4-(2-furoyl)-1-piperazinyl]methanone
3	4-Chloro-N-[(1-ethyl-3-methyl-1H-pyrazol-4-yl)methyl]-3-nitrobenzamide
4	N-(3-Chloro-2-methylphenyl)-2-(3,5-dimethyl-4-nitro-1H-pyrazol-1-yl)acetamide
5	2-Chloro-N-[(1-ethyl-5-methyl-1H-pyrazol-4-yl)methyl]-4-nitrobenzamide
6	N-(6-Amino-1-benzyl-2,4-dioxo-1,2,3,4-tetrahydro-5-pyrimidinyl)-2-chloro-N-methylacetamide
7	N-(2-Chlorobenzyl)-2-(3,5-dimethyl-4-nitro-1H-pyrazol-1-yl)acetamide
8	5-Chloro-N-[(1-ethyl-5-methyl-1H-pyrazol-4-yl)methyl]-2-nitrobenzamide
9	4-Chloro-N-[(1-ethyl-5-methyl-1H-pyrazol-4-yl)methyl]-3-nitrobenzamide
10	4-Chloro-N-[(1-ethyl-3-methyl-1H-pyrazol-4-yl)methyl]-2-nitrobenzamide
...	
...	
301	4-Chloro-N'-[(Z)-(3,4-dimethoxyphenyl)methylene]-1-methyl-1H-pyrazole-5-carbohydrazide
302	2-Oxo-2-[(1,3,5-trimethyl-1H-pyrazol-4-yl)amino]ethyl 6-chloronicotinate
303	2-Oxo-2-[(1,3,5-trimethyl-1H-pyrazol-4-yl)amino]ethyl 2-chloronicotinate
304	**Cycloxaprid**
305	N-(4-Chlorophenyl)-2-(3,5-dimethyl-4-nitro-1H-pyrazol-1-yl)propanamide
306	N-(2-Chloro-5-nitrobenzyl)-3-ethyl-N-methyl-1H-pyrazole-5-carboxamide
307	Ethyl 2-[(5-chloro-3-methyl-1-phenyl-1H-pyrazol-4-yl)carbonyl]hydrazinecarboxylate
308	N-[3-Chloro-2-(dimethylamino)phenyl]-2-(2,4-dioxo-3,4-dihydro-1(2H)-pyrimidinyl)acetamide
309	N-[3-Chloro-4-(1H-pyrazol-1-yl)phenyl]-N'-(1-hydroxy-2-propanyl)ethanediamide
310	4-Chloro-2-{2-[(1,3-dimethyl-1H-pyrazol-5-yl)amino]-2-oxoethoxy}benzamide
...	
...	

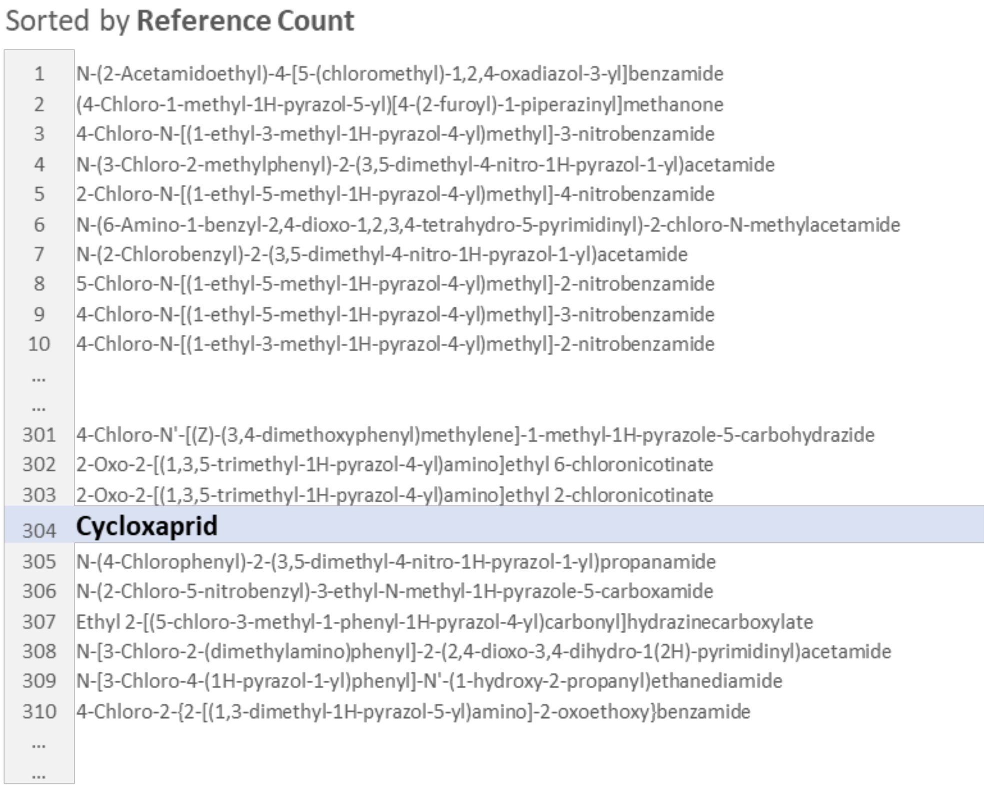
Searching a formula in online databases

$C_{14}H_{15}ClN_4O_3$



GREETINGS, STRANGER.
WHATEVER QUEST DRIVES YOU, ABANDON IT.
YOU SHALL FIND NO ANSWERS IN THESE DESOLATE WASTES.
I KNEW I WOULDN'T. I GUESS I ... JUST HAD TO SEE.
I HATE FEELING DESPERATE ENOUGH TO VISIT THE SECOND PAGE OF ~~GOOGLE~~ RESULTS.
chemical

Randall Munroe:
https://xkcd.com/1334/



Sorted by **Reference Count**

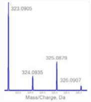
1	N-(2-Acetamidoethyl)-4-[5-(chloromethyl)-1,2,4-oxadiazol-3-yl]benzamide
2	(4-Chloro-1-methyl-1H-pyrazol-5-yl)[4-(2-furoyl)-1-piperazinyl]methanone
3	4-Chloro-N-[(1-ethyl-3-methyl-1H-pyrazol-4-yl)methyl]-3-nitrobenzamide
4	N-(3-Chloro-2-methylphenyl)-2-(3,5-dimethyl-4-nitro-1H-pyrazol-1-yl)acetamide
5	2-Chloro-N-[(1-ethyl-5-methyl-1H-pyrazol-4-yl)methyl]-4-nitrobenzamide
6	N-(6-Amino-1-benzyl-2,4-dioxo-1,2,3,4-tetrahydro-5-pyrimidinyl)-2-chloro-N-methylacetamide
7	N-(2-Chlorobenzyl)-2-(3,5-dimethyl-4-nitro-1H-pyrazol-1-yl)acetamide
8	5-Chloro-N-[(1-ethyl-5-methyl-1H-pyrazol-4-yl)methyl]-2-nitrobenzamide
9	4-Chloro-N-[(1-ethyl-5-methyl-1H-pyrazol-4-yl)methyl]-3-nitrobenzamide
10	4-Chloro-N-[(1-ethyl-3-methyl-1H-pyrazol-4-yl)methyl]-2-nitrobenzamide
...	
...	
301	4-Chloro-N'-[(Z)-(3,4-dimethoxyphenyl)methylene]-1-methyl-1H-pyrazole-5-carbohydrazide
302	2-Oxo-2-[(1,3,5-trimethyl-1H-pyrazol-4-yl)amino]ethyl 6-chloronicotinate
303	2-Oxo-2-[(1,3,5-trimethyl-1H-pyrazol-4-yl)amino]ethyl 2-chloronicotinate
304	**Cycloxaprid**
305	N-(4-Chlorophenyl)-2-(3,5-dimethyl-4-nitro-1H-pyrazol-1-yl)propanamide
306	N-(2-Chloro-5-nitrobenzyl)-3-ethyl-N-methyl-1H-pyrazole-5-carboxamide
307	Ethyl 2-[(5-chloro-3-methyl-1-phenyl-1H-pyrazol-4-yl)carbonyl]hydrazinecarboxylate
308	N-[3-Chloro-2-(dimethylamino)phenyl]-2-(2,4-dioxo-3,4-dihydro-1(2H)-pyrimidinyl)acetamide
309	N-[3-Chloro-4-(1H-pyrazol-1-yl)phenyl]-N'-(1-hydroxy-2-propanyl)ethanediamide
310	4-Chloro-2-{2-[(1,3-dimethyl-1H-pyrazol-5-yl)amino]-2-oxoethoxy}benzamide
...	
...	

… tell me when you see something you can **pronounce**.

The correct answer is usually the one that is easiest to read

When compounds are sorted by how easy their name is to read, using a common children's book readability score, the correct answer is often in the first few rows.

$C_{14}H_{15}ClN_4O_3$



Sorted by **Readability**

1	**Cycloxaprid**
2	(5s,8r)-cycloxaprid
3	avadomide hydrochloride (usan)
4	uracil, 1-{p-[3-(2-chloroethyl)ureido]benzyl}-
5	5-chloro-n-(oxan-2-yloxy)-2-(1h-1,2,4-triazol-1-yl)benzamide
6	8-chloro-4-(2-methyl-1,4-oxazepan-4-yl)-6-nitroquinazoline
7	n-(1-tert-butyl-1h-pyrazol-3-yl)-4-chloro-2-nitrobenzamide
8	1-(5-chloro-2-methoxyphenyl)-3-(6-ethoxypyrimidin-4-yl)urea
9	1-(5-chloro-2-methoxyphenyl)-3-(2-(6-oxopyridazin-1(6h)-yl)ethyl)urea
10	2-(7-chloro-5-nitro-1h-indazol-1-yl)-1-(piperidin-1-yl)ethan-1-one
…	
…	

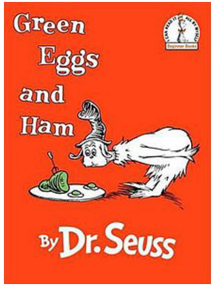
Sorted by **Reference Count**

1	N-(2-Acetamidoethyl)-4-[5-(chloromethyl)-1,2,4-oxadiazol-3-yl]benzamide
2	(4-Chloro-1-methyl-1H-pyrazol-5-yl)[4-(2-furoyl)-1-piperazinyl]methanone
3	4-Chloro-N-[(1-ethyl-3-methyl-1H-pyrazol-4-yl)methyl]-3-nitrobenzamide
4	N-(3-Chloro-2-methylphenyl)-2-(3,5-dimethyl-4-nitro-1H-pyrazol-1-yl)acetamide
5	2-Chloro-N-[(1-ethyl-5-methyl-1H-pyrazol-4-yl)methyl]-4-nitrobenzamide
6	N-(6-Amino-1-benzyl-2,4-dioxo-1,2,3,4-tetrahydro-5-pyrimidinyl)-2-chloro-N-methylacetamide
7	N-(2-Chlorobenzyl)-2-(3,5-dimethyl-4-nitro-1H-pyrazol-1-yl)acetamide
8	5-Chloro-N-[(1-ethyl-5-methyl-1H-pyrazol-4-yl)methyl]-2-nitrobenzamide
9	4-Chloro-N-[(1-ethyl-5-methyl-1H-pyrazol-4-yl)methyl]-3-nitrobenzamide
10	4-Chloro-N-[(1-ethyl-3-methyl-1H-pyrazol-4-yl)methyl]-2-nitrobenzamide
…	
…	
301	4-Chloro-N'-[(Z)-(3,4-dimethoxyphenyl)methylene]-1-methyl-1H-pyrazole-5-carbohydrazide
302	2-Oxo-2-[(1,3,5-trimethyl-1H-pyrazol-4-yl)amino]ethyl 6-chloronicotinate
303	2-Oxo-2-[(1,3,5-trimethyl-1H-pyrazol-4-yl)amino]ethyl 2-chloronicotinate
304	**Cycloxaprid**
305	N-(4-Chlorophenyl)-2-(3,5-dimethyl-4-nitro-1H-pyrazol-1-yl)propanamide
306	N-(2-Chloro-5-nitrobenzyl)-3-ethyl-N-methyl-1H-pyrazole-5-carboxamide
307	Ethyl 2-[(5-chloro-3-methyl-1-phenyl-1H-pyrazol-4-yl)carbonyl]hydrazinecarboxylate
308	N-[3-Chloro-2-(dimethylamino)phenyl]-2-(2,4-dioxo-3,4-dihydro-1(2H)-pyrimidinyl)acetamide
309	N-[3-Chloro-4-(1H-pyrazol-1-yl)phenyl]-N'-(1-hydroxy-2-propanyl)ethanediamide
310	4-Chloro-2-{2-[(1,3-dimethyl-1H-pyrazol-5-yl)amino]-2-oxoethoxy}benzamide
…	
…	

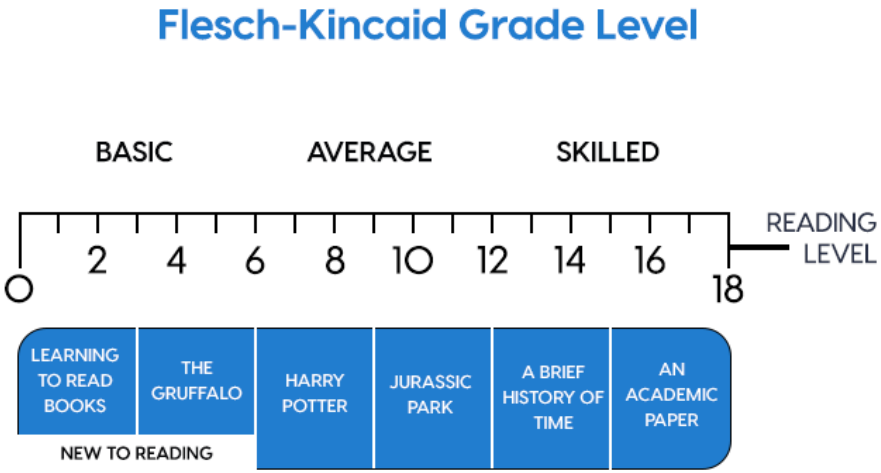
While sorting by how many references a compound has will often reveal the correct answer, for newly registered pesticides, this does not work. In this example it is row 304, and it is on page 16 of a chemical compound search.

How do we measure readability?

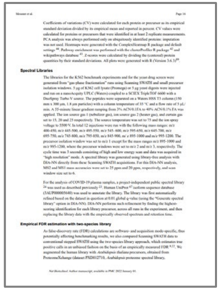
- Simple algorithms (some dating back to 1923) for estimating how difficult it is to read the text in a book



https://en.wikipedia.org/wiki/Green_Eggs_and_Ham#/media/File:Green_Eggs_and_Ham.jpg



https://readable.com/readability/flesch-reading-ease-flesch-kincaid-grade-level/



Messner et al. Page 16

Coefficients of variations (CV) were calculated for each protein or precursor as its empirical standard deviation divided by its empirical mean and reported in percent. CV values were calculated for proteins or precursors that were identified in at least 2 replicate measurements. PCA analysis was always performed only on ubiquitously identified proteins: imputation was not used. Heatmaps were generated with the ComplexHeatmap R package and default settings [44]. Pathway enrichment was performed with the clusterProfiler R package [42] and wikipathways database [43]. Z-scores were calculated by dividing the (centered) protein quantities by their standard deviations. All plots were generated with R (Version 3.6.3)[89].

Spectral Libraries

The libraries for the K562 benchmark experiments and for the yeast drug screen were generated from "gas-phase fractionation" runs using Scanning SWATH and small precursor isolation windows. 5 µg of K562 cell lysate (Promega) or 5 µg yeast digests were injected and run on a nanoAcquity UPLC (Waters) coupled to a SCIEX TripleTOF 6600 with a DuoSpray Turbo V source. The peptides were separated on a Waters HSS T3 column (150 mm x 300 µm, 1.8 µm particles) with a column temperature of 35 °C and a flow rate of 5 µL/min. A 55-minute linear gradient ramping from 3% ACN/0.1FA to 40% ACN/0.1% FA was applied. The ion source gas 1 (nebulizer gas), ion source gas 2 (heater gas), and curtain gas set to 15, 20 and 25 respectively. The source temperature was set to 75 and the ion spray voltage to 5500 V. In total 12 injections were run with the following mass ranges: m/z 400-450, m/z 445-500, m/z 495-550, m/z 545- 600, m/z 595-650, m/z 645-700, m/z 695-750, m/z 745-800, m/z 795-850, m/z 845-900, m/ z 895-1000 and m/z 995-1200. The precursor isolation window was set to m/z 1 except for the mass ranges m/z 895-1000 and m/z 995-1200, where the precursor windows were set to m/z 2 and m/z 3, respectively. The cycle time was 3 seconds consisting of high and low energy scan and data was acquired in "high resolution" mode. A spectral library was generated using library-free analysis with DIA-NN directly from these Scanning SWATH acquisitions. For this DIA-NN analysis, MS2 and MS1 mass accuracies were set to 25 ppm and 20 ppm, respectively, and scan window size set to 6.

For the analysis of COVID-19 plasma samples, a project-independent public spectral library [24] was used as described previously [23]. Human UniProt [87] isoform sequence database (3AUP000005640) was used to annotate the library. The library was first automatically refined based on the dataset in question at 0.01 global q-value (using the "Generate spectral library" option in DIA-NN). DIA-NN performs such refinement by finding the highest-scoring identification for each library precursor, across all runs in the experiment, and then replacing the library data with the empirically observed spectrum and retention time.

Empirical FDR estimation with two-species library

As false-discovery rate (FDR) calculations are software- and acquisition mode-specific, thus potentially affecting benchmarking results, we also compared Scanning SWATH data to conventional stepped SWATH using the two-species library approach, which estimates true positive calls in an unbiased fashion on the basis of an empirically measured FDR [8,27]. We augmented the human library with *Arabidopsis thaliana* precursors, obtained from ProteomeXchange (dataset PXD012710, *Arabidopsis* proteome spectral library.

Nat Biotechnol. Author manuscript; available in PMC 2022 January 01.

Using readability scores on a larger data set

Customer sample

- 242 compounds relevant to food safety
  - 53 of these not in available spectral library

Can the compounds that are **not** in the spectral library be identified?

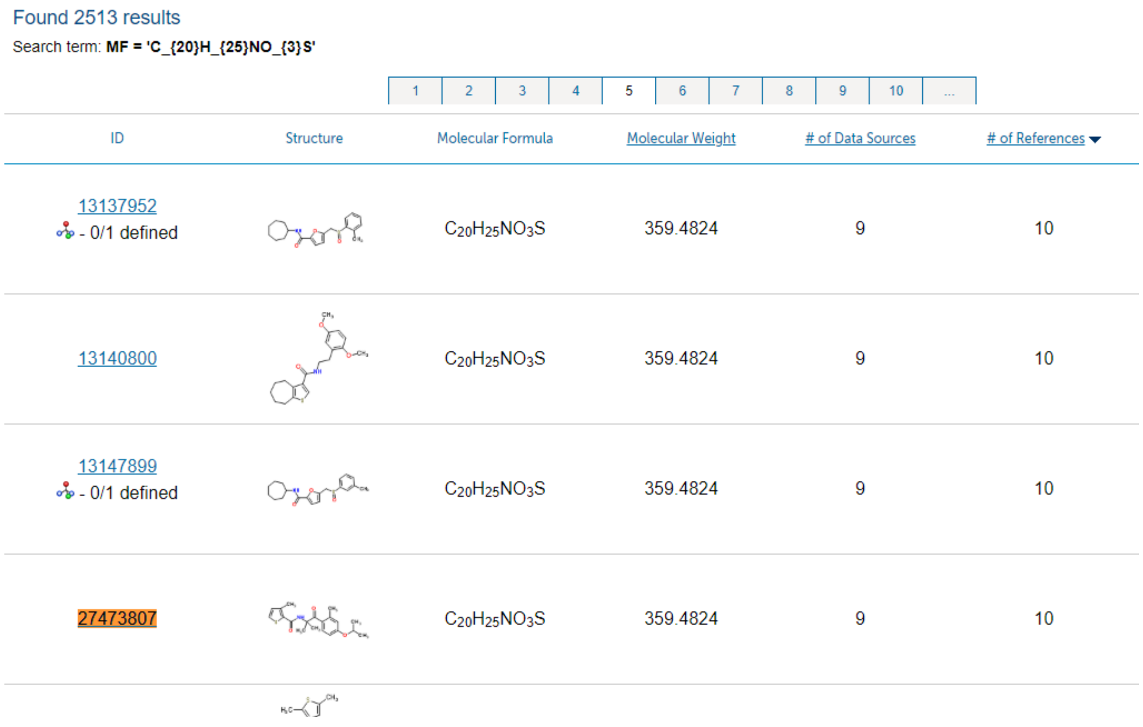
Automate the workflow using Python

- Pre-calculate
  - Use textstat to score readability of all compound name synonyms from PubChem database
- For a given m/z or predicted chemical formula
  - Find all possible PubChem matches using PubChemPy
- Lookup the readability of the synonyms for all matches
  - Keep anything that is readable (from infant up to my thesis supervisor)
- Search anything remaining in Google along with search terms relevant to the sample
  - Using Google custom search API

Using reference counting, Isofetamid is found on page 5

Isofetamid

On the fifth page of results, it finally shows up, buried amongst other structures with similar reference counts



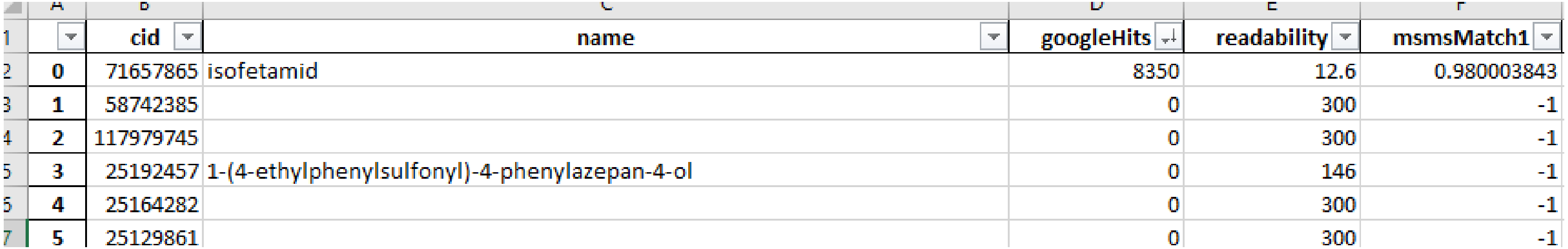
Found 2513 results

Search term: **MF = 'C_{20}H_{25}NO_{3}S'**

ID	Structure	Molecular Formula	Molecular Weight	# of Data Sources	# of References
13137952 - 0/1 defined		$C_{20}H_{25}NO_3S$	359.4824	9	10
13140800		$C_{20}H_{25}NO_3S$	359.4824	9	10
13147899 - 0/1 defined		$C_{20}H_{25}NO_3S$	359.4824	9	10
27473807		$C_{20}H_{25}NO_3S$	359.4824	9	10

Using readability, Isofetamid is successfully identified

Isofetamid

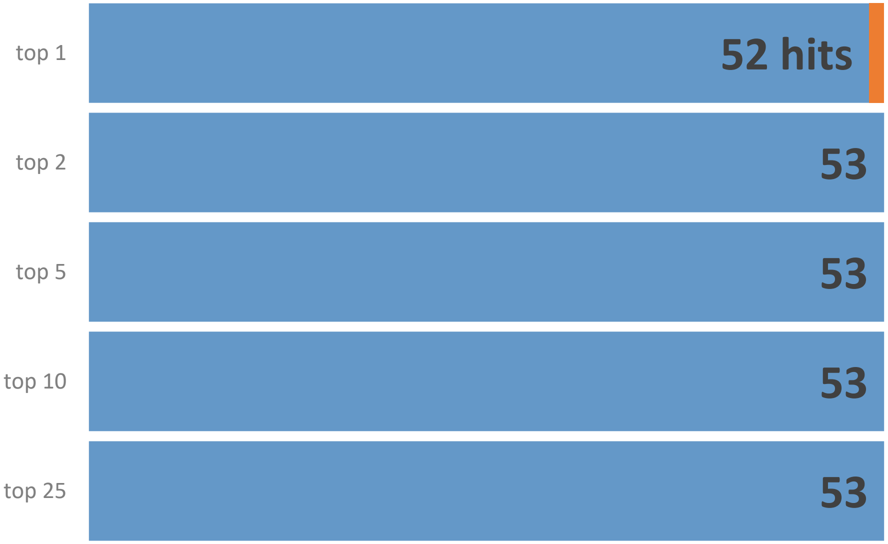


	cid	name	googleHits	readability	msmsMatch1
0	71657865	isofetamid	8350	12.6	0.980003843
1	58742385		0	300	-1
2	117979745		0	300	-1
3	25192457	1-(4-ethylphenylsulfonyl)-4-phenylazepan-4-ol	0	146	-1
4	25164282		0	300	-1
5	25129861		0	300	-1

Using readability on its own, or combined with Google searching or theoretical MS/MS matching, enables successful identification of the correct compound

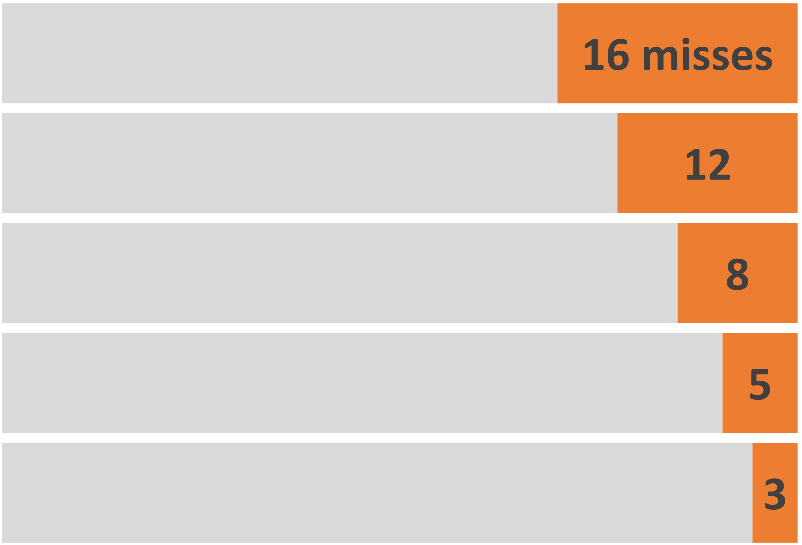
More unknowns identified using readability

Identified using **readability** and number of **Google page hits**



Scoring the number of **Google page hits** for only those compound names that are **readable** gave the **correct** identification within the **top 2 compounds**.

Identified using **reference counts**



While using **reference counts** does work for most unknowns, this technique did **not** work for several compounds. For 3 of these examples, the correct identification was **not** in the **top 25 compounds**. Purchasing this many compounds to confirm a **single** compound would be prohibitively **expensive**.

Trademarks / Licensing

The SCIEX clinical diagnostic portfolio is For In Vitro Diagnostic Use. Rx Only. Product(s) not available in all countries. For information on availability, please contact your local sales representative or refer to www.sciex.com/diagnostics. All other products are For Research Use Only. Not for use in Diagnostic Procedures.

Searching chemical databases by m/z or molecular formula often generates a **very long list** of possible compounds. The **correct answer** is often the one that is **easiest to read**.

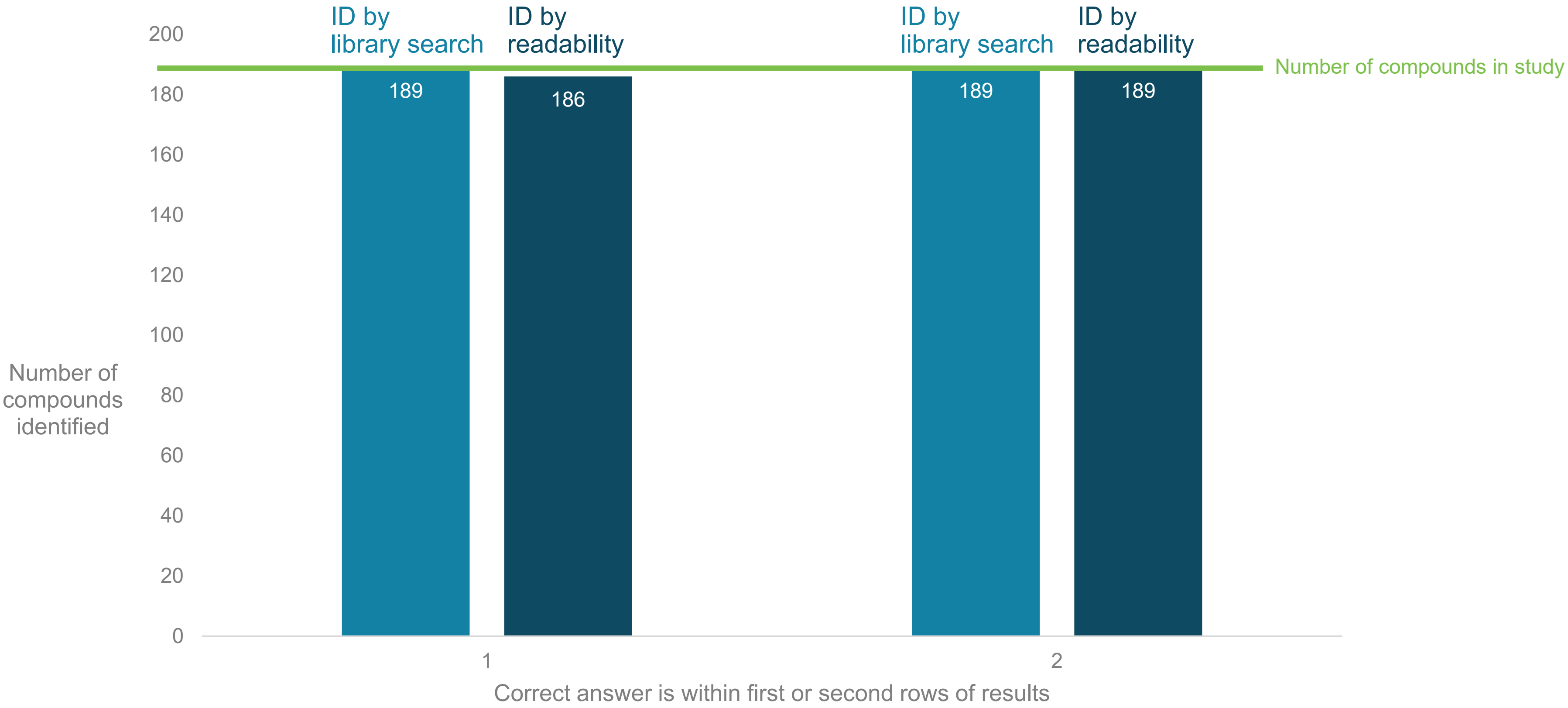
SCIEX

The Power of Precision

Questions and answers

Does **readability** have anything to do with **spectral library** matching?

Readability scoring is almost as effective as library searching



Trademarks / Licensing

The SCIEX clinical diagnostic portfolio is For In Vitro Diagnostic Use. Rx Only. Product(s) not available in all countries. For information on availability, please contact your local sales representative or refer to www.sciex.com/diagnostics. All other products are For Research Use Only. Not for use in Diagnostic Procedures.