Beyond the Top Hit: Extracting Unknown Structural Information from Hybrid Similarity Search Hit Lists

Brian T. Cooper, UNC Charlotte and NIST (btcooper@uncc.edu)
Tytus D. Mak, NIST (tytus.mak@nist.gov)
Stephen E. Stein, NIST (stephen.stein@nist.gov)

UNC CHARLOTTE

NIST National Institute of Standards and Technology U.S. Department of Commerce

Mass Spectrometry Data Center

MOD pm 3:50

NIST MS Libraries and Search Methods

Electron Ionization

- NIST2020: 350 643 spectra from 306 869 compounds
- NIST17: 306 622 spectra from 267 376 compounds
- Unit-mass resolution

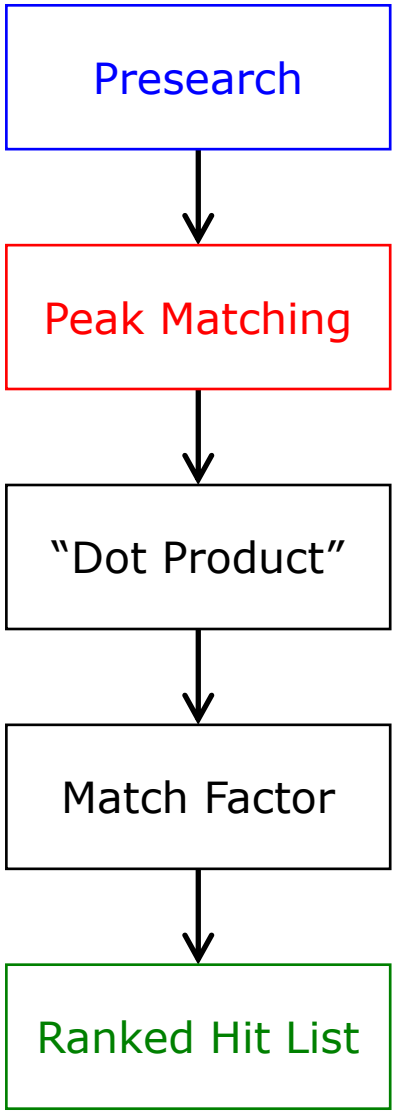
Small-Molecule Tandem

- NIST2020: 1.3 million spectra from 186 000 precursor ions from 31 000 compounds [June 2]
- **NIST17:** 574 826 spectra from 40 266 initial (MS^1) precursor ions (2 680 unique types) from 13 808 compounds
- Resolution to 0.000 1 with relative (ppm) or absolute (m/z) tolerances

Peptide Tandem

- Libraries (over 4.3 million total spectra) freely available at chemdata.nist.gov
- Resolution to 0.000 1 with relative (ppm) or absolute (m/z) tolerances

All NIST searches use the same **five basic steps**:



Select a subset of the library likely to score highly. [Hybrid]

For each candidate spectrum…

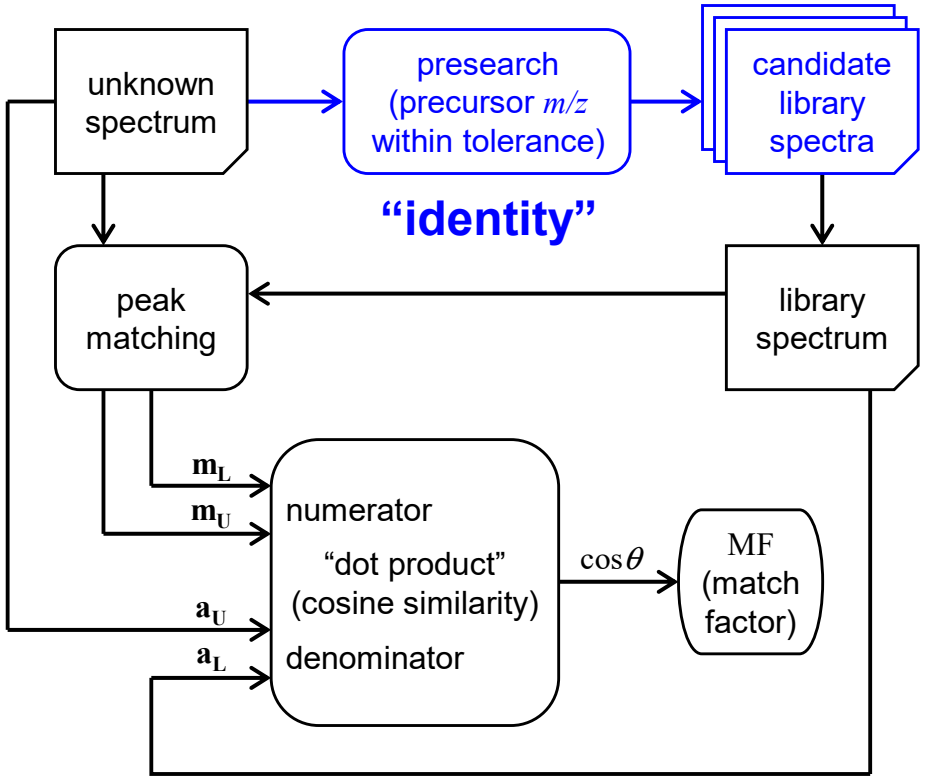
Find library peaks that match the unknown within the specified tolerance. [Hybrid]

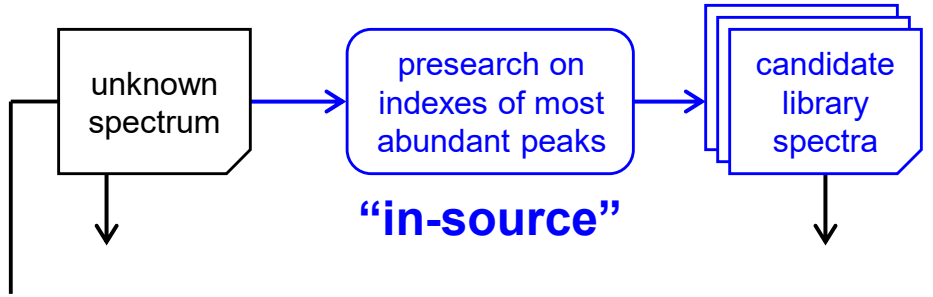
Scoring: calculate weighted cosine similarity…

…then compute the "match factor" by applying various empirical adjustments.

Search method details differ for each type of library!

Tandem Identity and In-Source Searches





The **"in-source"** search was originally intended for cases when the unknown is in the library but only as a different precursor—perhaps due to dissociation in the ESI source. So it eliminates the precursor-matching constraint of the tandem **"identity"** search, and thus requires a more sophisticated presearch. It then proceeds like a conventional identity search.

For ***all*** small-molecule tandem searches, empirical adjustments to $\cos\theta$ reduce the match factor for simple spectra with only a few matching peaks, to reflect the decreased confidence an experienced analyst would have in such hits.

"weights": $w = (m/z)^m I^n = I^{½}$

[small-molecule tandem: $m = 0;\ n = ½$]

$$\cos\theta = \frac{\mathbf{m}_L \cdot \mathbf{m}_U}{\|\mathbf{a}_L\| \|\mathbf{a}_U\|} = \frac{\sum_{\text{match}} w_L w_U}{\sqrt{\sum_{\text{all}} w_L^2} \sqrt{\sum_{\text{all}} w_U^2}} = \frac{\sum_{\text{match}} I_L^{½} I_U^{½}}{\sqrt{\sum_{\text{all}} I_L} \sqrt{\sum_{\text{all}} I_U}}$$

Indexed presearching and dot-product search scoring:
Finnigan Application Report 2, **1978**

Tandem Hybrid Search

Cooper et al., *Anal. Chem.* **2019**, *91*, 13924

The **"hybrid"** similarity search adds ***peak shifting***—the logical equivalent of neutral-loss matching—to the in-source search. The algorithm elevates scores for similar compounds by matching peaks shifted by ***delta mass***:

$$\Delta m = m_{\text{unknown}} - m_{\text{library}}$$

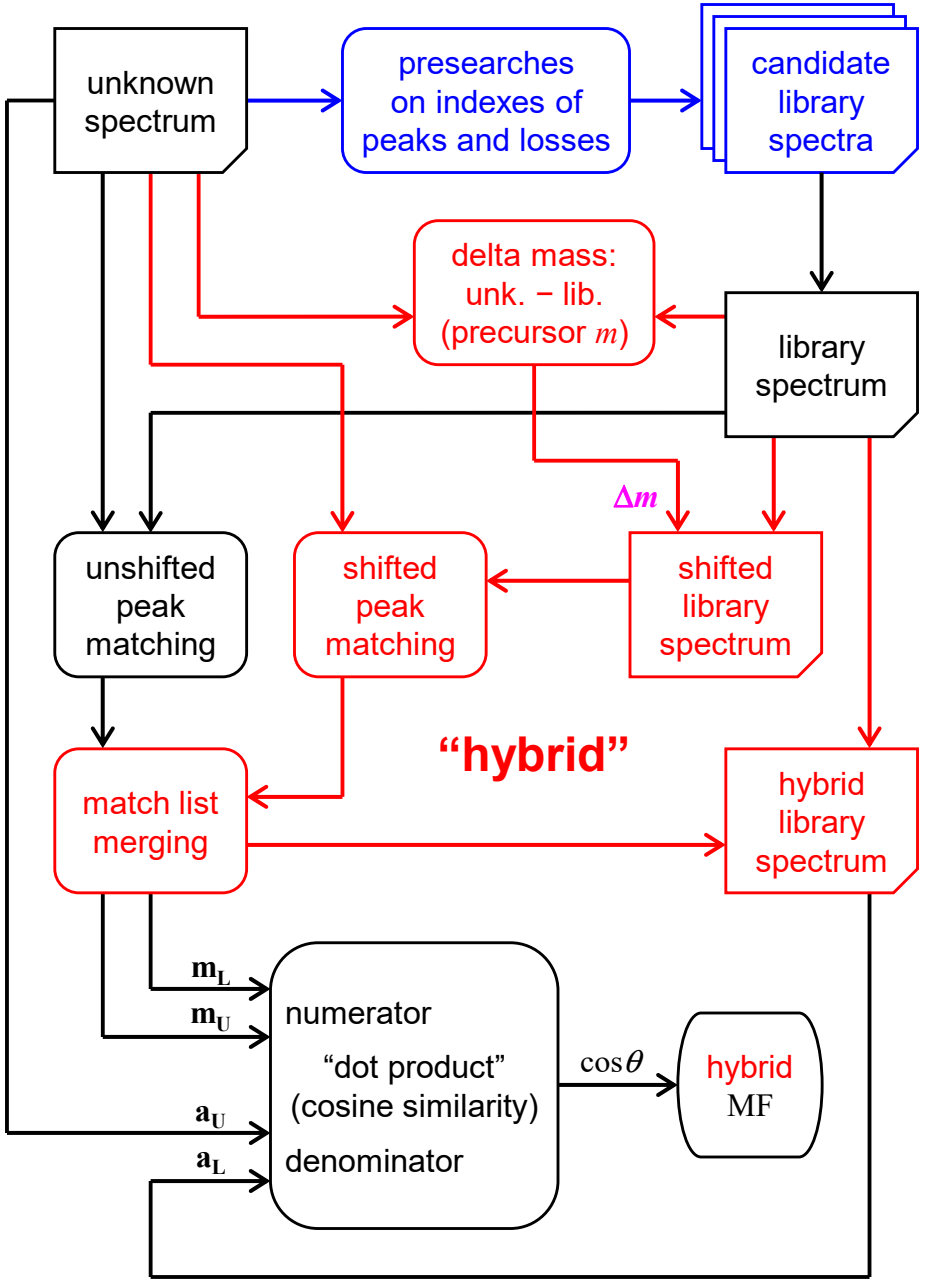
If the shifted and unshifted instances of a library peak both match, its abundance is apportioned between the two. Scoring then proceeds normally.

The hybrid search will return a list of similar compounds if the library contains **"cognates"** of the unknown—compounds for which the structural difference…

- is confined to a single region of the molecule, and…
- does not substantially alter its fragmentation behavior.

Matching shifted peaks from similar compounds:
Biemann, *Tetrahedron Lett.* **1960**, *1(36)*, 9

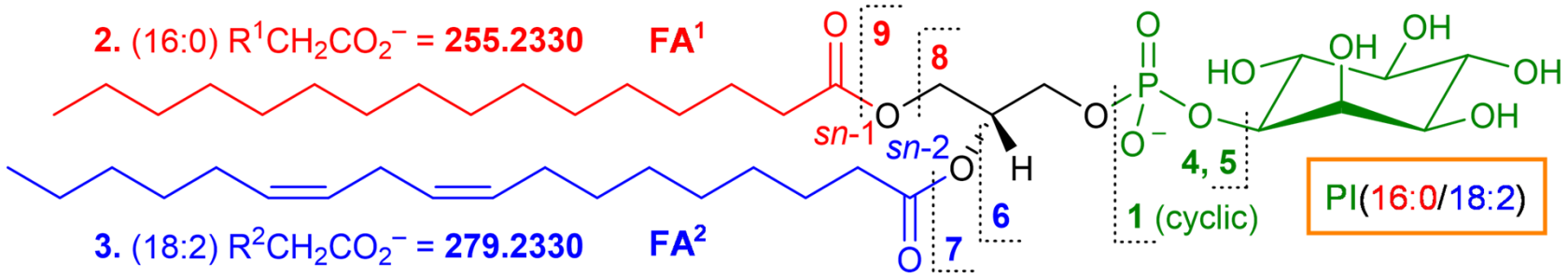
Combining direct and neutral-loss matching:
Stein, *J. Am. Soc. Mass Spectrom.* **1995**, *6*, 644



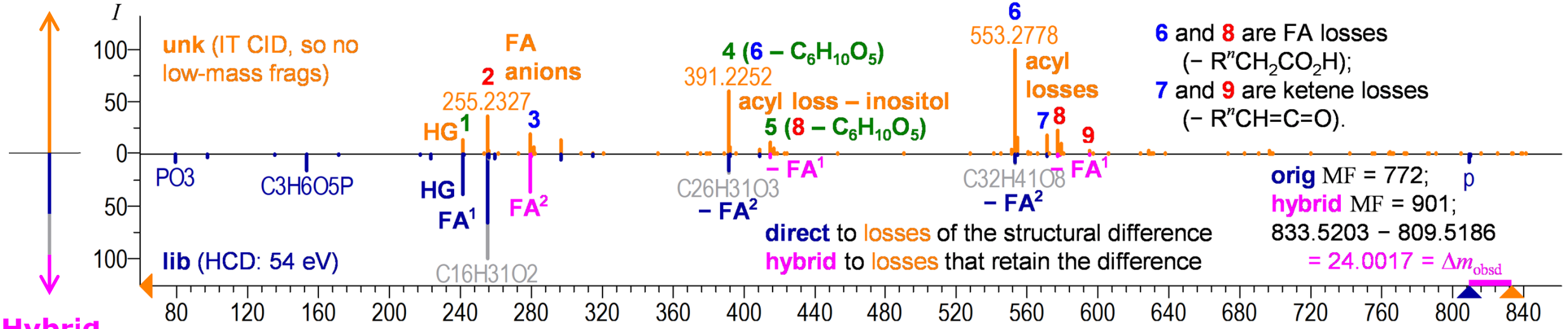
Hybrid Search Example

Recurrent unidentified spectrum (RUS) "cluster_007290" from urine_HR_it_neg_rec library searched against NIST17.

Likely structure and peak assignments deduced by *manual interpretation* of the hybrid search hit list:

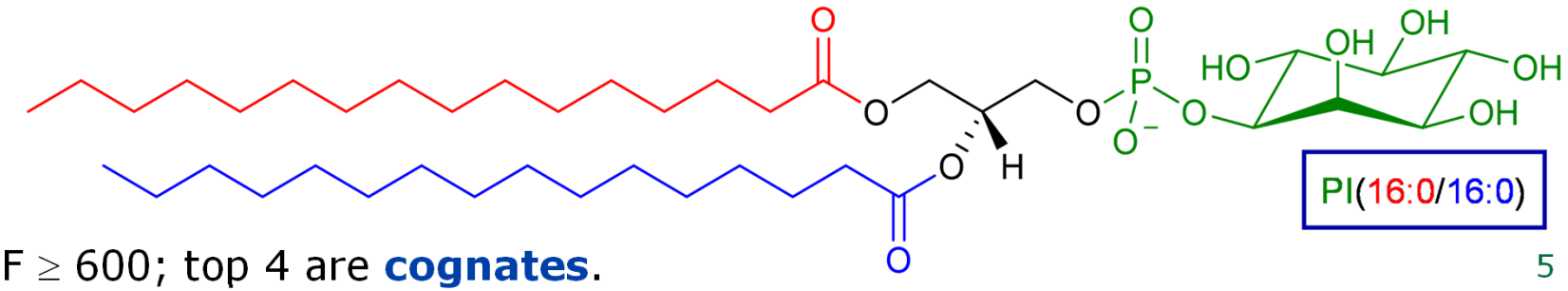


Query spectrum



Hybrid library spectrum (of hit 2)

Δm_{calc} = 24.0000
(unk – lib = C_2)



Hit list: 14 hits (all are GPs); 5 with MF ≥ 600; top 4 are **cognates**.

Excluded-Query-Compound Searches

Accuracy: A similarity search is "accurate" when it returns a *list* of structurally similar compounds that can be usefully interpreted. We used **"excluded-query-compound" (EQC)** searches to investigate the global performance of the hybrid search.

"EQC Library" (subset of NIST17):

- high resolution (product ion m/z to 0.0001)
- MS^2 only (no MS^3 or higher)
- entries with InChIKeys (and thus structures)
- 357 978 spectra from 10 758 unique compounds (10 429 unique connectivities)

InChIKey Format

MYXNWGACZJSMBT-VJXVFPJBSA-N

hash of InChI connectivity layer

hash of other InChI layers (including stereochemistry)

Global EQC Searches (~10.7 million total hits):

- Each spectrum in the EQC library was searched against the rest of the library, excluding all spectra from compounds with the *same connectivity* as the query.
- Up to 100 hits/query were saved, with no minimum score.
- Hit lists were shorter (~30 on average) because only the highest-scoring spectrum from each compound was kept.

Differences from **"leave-one-out"** cross-validation:

- Not just *one*, but *all* spectra from compounds with the same connectivity are left out.
- We are not trying to train or refine a predictive model or even choose a score threshold.
- We do not yet have a cleanly binary, easily automated measure of "success" for the hybrid search. Although we used class membership to define success for *individual hits*, the success of the *search* depends on the usefulness of the structural information throughout the hit list, not merely whether the top hit is "accurate." So we cannot report sensitivity, specificity, or related performance measures commonly used for binary classifiers.

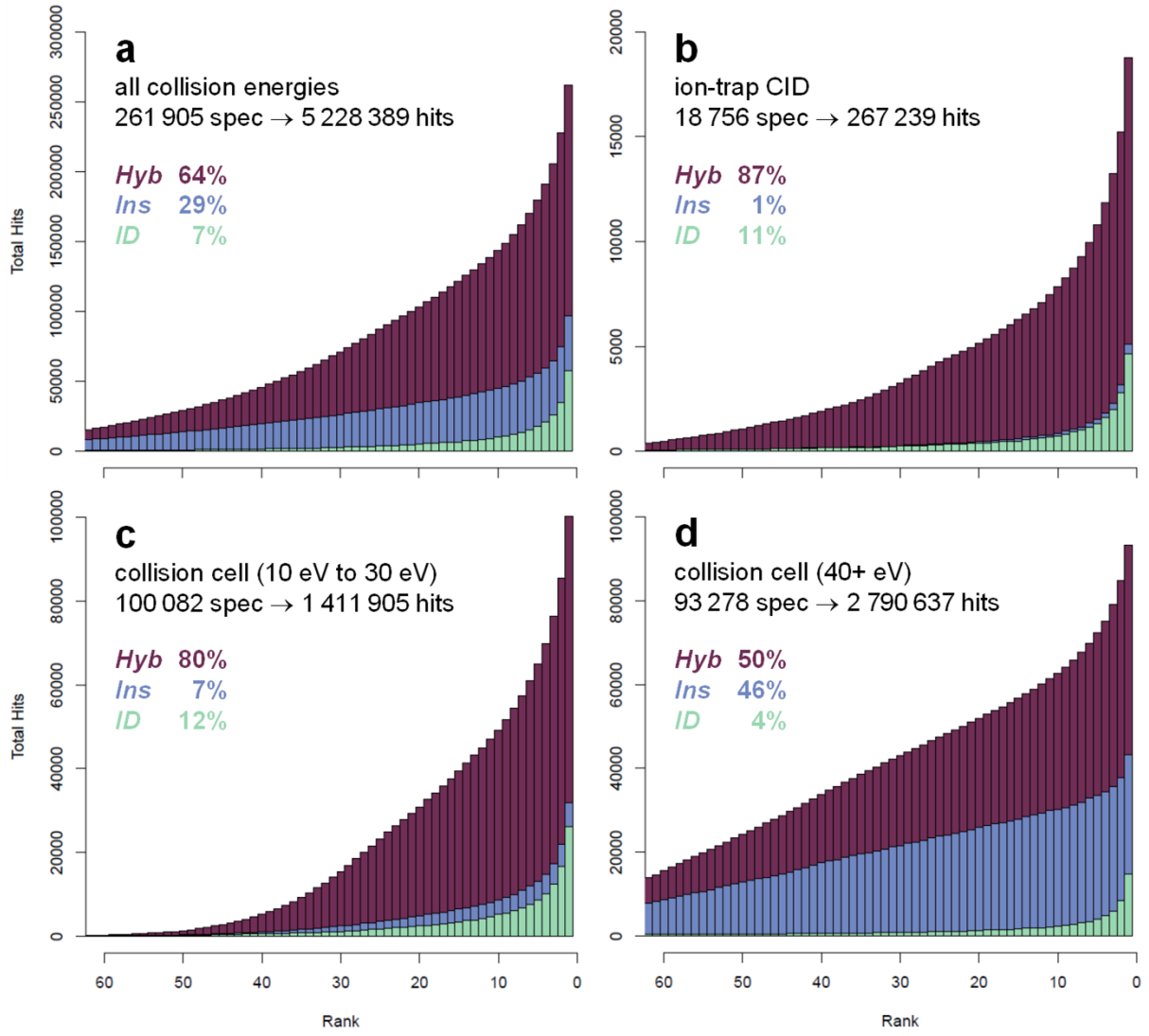
Match Types

The hybrid search algorithm *allows* but does not *require* peak shifting. Many of the hits returned by the hybrid search would also be returned by identity or in-source searches. So we define the following **"match types"** for the hybrid search:

- ***ID*** hits: Δm = 0 (within tolerance) and no shifted peaks; would also appear in an identity search.
- ***Ins*** hits: nonzero Δm but no shifted peaks; would also appear in an in-source search.
- ***Hyb*** hits: nonzero Δm and shifted peaks; only in a hybrid search (but could also appear with their original, unshifted score in an in-source search).

The hybrid search works best for spectra obtained at low collision energies!

Lower energies (**b** and **c**) favor larger fragments that are more likely to contain the structural difference and thus match by the hybrid search. Higher energies (**d**) produce smaller fragments that are more likely to match low-mass peaks from unrelated compounds, generating longer, less-relevant hit lists.



Match types by rank for hybrid EQC searches, excluding low-scoring hits (MF < 600). Percentages are for all ranks.

Hybrid Search Accuracy: Class-Hit Rates

Similar results were obtained for ~20 eV collision-cell spectra.

Molecules in the EQC library that have arbitrary, class-defining substructures

intersection of two classes

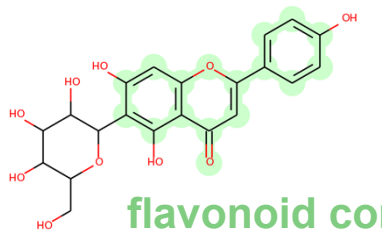
Most hits are to the ***same class***, and most class-hits are ***Hyb*** hits.

Excluded-Query-Compound Hit Rates (Score ≥ 600) for $[M+H]^+$ by Chemical Class

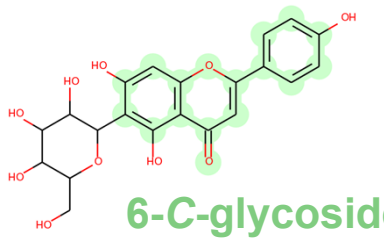
	Query Data			Class-Hit Rate	Distribution of Class-Hits by Match Type		
Query Class	query spectra	avg hits per query	% with > 0 hits	% of hits to the combined class	% ***Hyb*** query / extended	% ***Ins*** combined	% ***ID*** combined
Ion-Trap CID Spectra							***isomers***
Amino acids	974	11.8	84	90	96.7	0.5	2.8
Nucleosides	121	2.6	65	89	91.5	4.6	3.9
Fentanyls	37	12.0	89	81	90.0	3.3	6.7
Flavonoids[a]	353	18.6	90	77	78.4 / 9.1	0.1	12.5
Carnitines	30	8.0	93	92	96.8	0.5	2.7
Sphingolipids	48	4.9	81	61	100	0	0
Glycerolipids[b]	31	12.3	87	55	88.6 / 6.6	0	4.7
Glycerophospholipids[c]	111	2.8	70	96	95.7 / 2.0	0.3	2.0
Hexuronides[d]	56	2.2	54	61	97.3	0	2.7
Steroids	329	22.9	89	77	95.1	1.2	3.7
Glucuronide steroids[e]	8	7.2	75	95	36.4 / 58.2	5.5	0
Overall	**2098**	**13.2**	**83**	**82**	**91.9 / 2.2**	**0.7**	**5.2**

[a] The query class includes isoflavonoids; plus hits to an extended class of anthocyan[id]ins, [iso]flavan[one]s, and phenylcoumarins. [b] Plus hits to glycerophospholipids. [c] Plus hits to glycerolipids. [d] Mostly glucuronides with a few galacturonides. [e] Plus hits to all steroids or hexuronides.

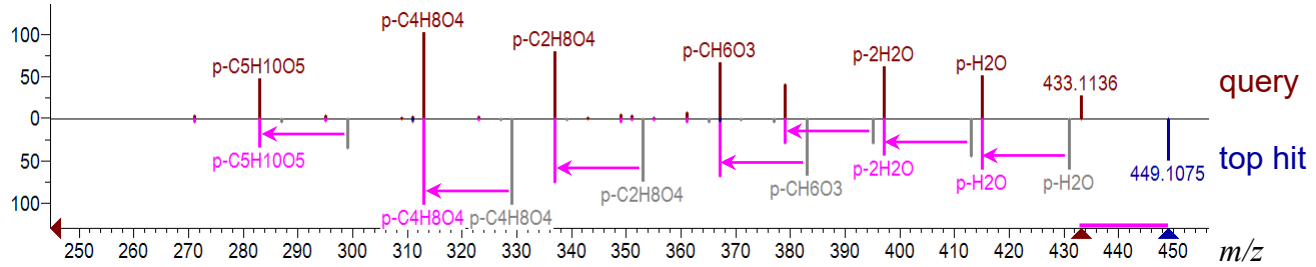
Hybrid EQC Hit List Example: Isovitexin



flavonoid core



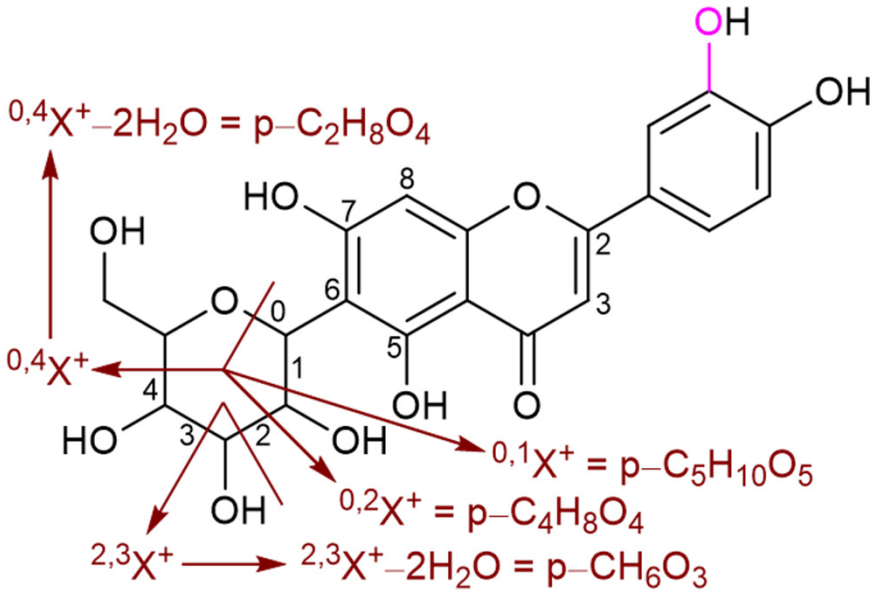
6-*C*-glycoside



Δm = −15.9949
unk − lib = O_{-1}

Hyb / *Ins* / *ID* match types — hits to query class / hits to extended class / misses; -in-source loss

#	Score (MF) hybrid	Score (MF) orig	Matching Pks hybrid	Matching Pks orig	Delta Mass unk − lib	**Unknown**: Isovitexin [M+H]+ HCD 19V P=433.1 **Library**: name, precursor, instrument, collision energy, precursor *m/z*
1	990	0	15	2	-15.9949	Isoorientin [M+H]+ HCD 17V P=449.1
2	986	0	14	1	-14.0157	Swertisin [M+H]+ HCD 17V P=447.1
3	986	0	15	1	-14.0157	Spinosine [M+H-C6H10O5]+ HCD 17V P=447.1
4	982	965	13	13	-162.0528	Saponarin [M+H]+ HCD 29V P=595.2
5	959	0	16	1	10.0207	Mangiferin [M+H]+ HCD 21V P=423.1
6	925	0	14	0	37.9792	Aloesin [M+H]+ HCD 19V P=395.1
7	921	921	15	15	0.0000	Vitexin [M+H]+ IT-FT 35% P=433.1
8	908	4	16	4	-15.9949	Orientin [M+H]+ HCD 26V P=449.1
9	904	904	15	15	0.0000	Vitexin 4-O-glucoside [M+H-C6H10O5]+ IT-FT 35% P=433.1
10	888	0	13	0	-59.9848	Carminic acid [M+H]+ IT-FT 35% P=493.1
11	839	0	11	0	-4.0313	Naringin dihydrochalcone [M+H-C6H10O4]+ HCD 13V P=437.1
12	831	0	11	0	-4.0313	Phlorizin [M+H]+ HCD 8V P=437.1
13	820	820	16	16	-146.0579	Vitexin-2"-O-rhamnoside [M+H]+ HCD 46V P=579.2
14	815	1	11	3	-132.0423	Schaftoside [M+H]+ IT-FT 35% P=565.2
15	814	0	13	0	41.9742	Polydatin [M+H]+ IT-FT 35% P=391.1
16	811	0	14	0	11.9636	Rhaponticin [M+H]+ IT-FT 35% P=421.1
17	794	0	11	0	-152.0321	Neomangiferin [M+H]+ IT-FT 35% P=585.1
18	787	0	11	0	-32.0262	Neohesperidin [M+H-C6H10O4]+ IT-FT 35% P=465.1
19	768	56	11	5	-18.0106	Eriodictyol 7-O-neohesperidoside [M+H-C6H10O4]+ HCD 13V P=451.1
20	758	0	11	0	146.0004	Orcinol β-D-glucoside [M+H]+ HCD 11V P=287.1
21	746	0	7	0	11.9636	4-Deoxyphloridzin [M+H]+ HCD 4V P=421.1



Pairwise Maximum Common Substructures

Pairwise similarity. We have tried three ways of quantifying the structural similarity between a known query compound and individual hits from hybrid EQC searches:

- Chemical **class membership** (but class definitions are arbitrary and often exclude useful structural similarity).
- Molecular **fingerprints** (but these encode local connectivity better than larger structural features).
- Graph-based metrics using pairwise **MCS** calculations (favors larger substructures).

The ***"size"*** $|G|$ of a molecular graph is the sum of its ***vertices*** V (atoms) and ***edges*** E (bonds), excluding hydrogens. The "asymmetric" graph-based similarity (or **overlap coefficient**) C3 is:

$$C3 = \frac{|G_{MCS}|}{\min\left(|G_{hit}|, |G_{query}|\right)};\ 0 \le C3 \le 1$$

Hit-List MCS. For genuine unknowns, structural information from the hit list can suggest at least partial structures. The FindMCS function in RDKit takes a "threshold" argument that lets it omit a fraction of the input structures, generating useful MCS even when some hits are unrelated to the query. When the query is known, we can also find a pairwise **"meta-MCS"** between the hit-list MCS and the query. We then describe the overlap using C3 and the "relative size" HQ:

$$HQ = \frac{G_{\text{hit-list MCS}}}{G_{query}}$$

The hit-list MCS…

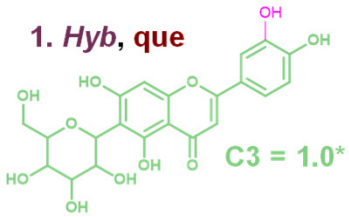
	HQ > 1	HQ = 1	HQ < 1
C3 = 1:	is larger than and completely covers the query.	***is*** the query!	is smaller than and completely included in the query.
C3 < 1:	is larger than but only partly covers the query.	is the same size as but only partly covers the query.	is smaller than but only partly included in the query.

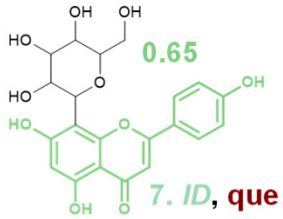
"Thresholded" multiple-MCS calculations:
RDKit: Open-source cheminformatics; http://www.rdkit.org

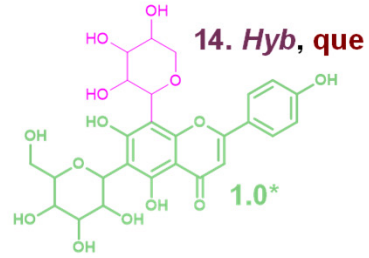
Fingerprint- versus graph-based similarity metrics:
Raymond, Willett, *J. Comput.-Aided Mol. Des.* **2002**, *16*, 59

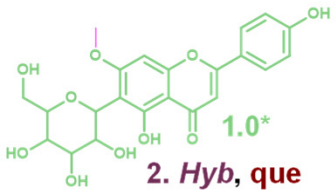
Hit List with Pairwise MCS Similarities

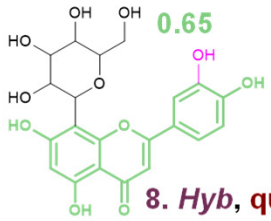
Hyb, *Ins*, and *ID* hits to **query class** or to **extended class**

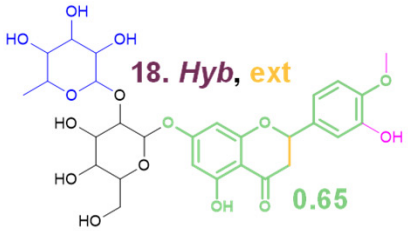


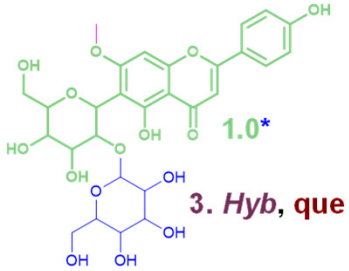


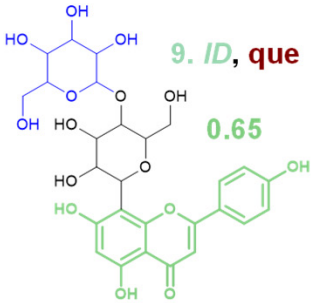


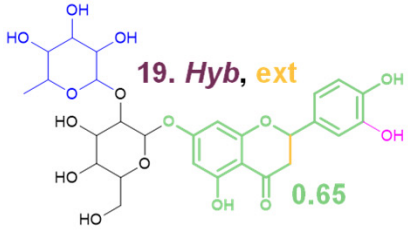


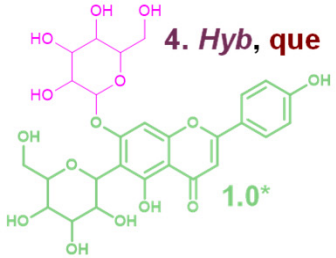


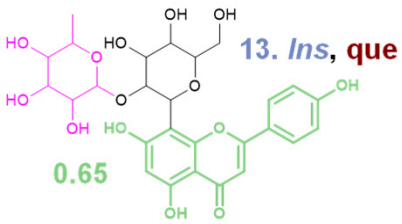






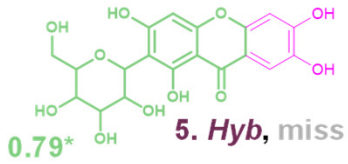


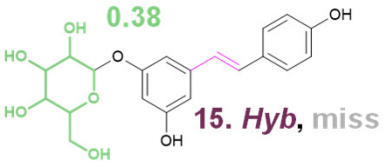


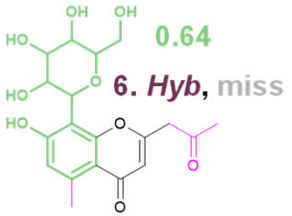


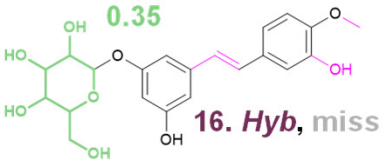
Structure key:
- In-source losses (*O*-linked)
- **Pairwise MCS** (vs query)
- **C3** overlap score (*cognate)
- Extended class differences
- Differences toward Δm

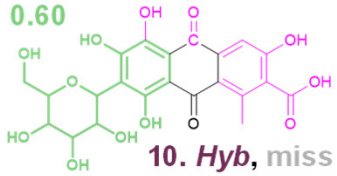
"Misses" (many are still quite structurally similar)

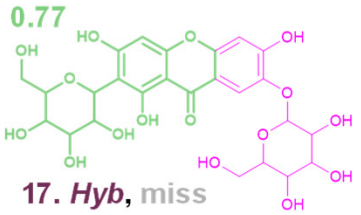


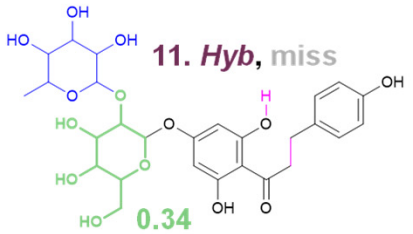


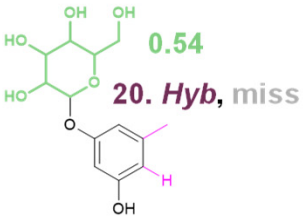


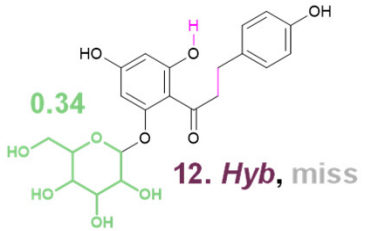


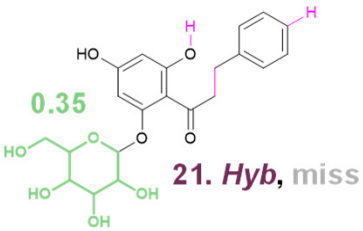












Hit-List Multiple-MCS Calculations

Pairwise similarity. We have tried three ways of quantifying the structural similarity between a known query compound and individual hits from hybrid EQC searches:

- Chemical **class membership** (but class definitions are arbitrary and often exclude useful structural similarity).
- Molecular **fingerprints** (but these encode local connectivity better than larger structural features).
- Graph-based metrics using pairwise **MCS** calculations (favors larger substructures).

The ***"size"*** $|G|$ of a molecular graph is the sum of its ***vertices*** V (atoms) and ***edges*** E (bonds), excluding hydrogens. The "asymmetric" graph-based similarity (or **overlap coefficient**) C3 is:

$$C3 = \frac{|G_{MCS}|}{\min\left(|G_{hit}|, |G_{query}|\right)};\ 0 \le C3 \le 1$$

Hit-List MCS. For genuine unknowns, structural information from the hit list can suggest at least partial structures. The FindMCS function in RDKit takes a "threshold" argument that lets it omit a fraction of the input structures, generating useful MCS even when some hits are unrelated to the query. When the query is known, we can also find a pairwise **"meta-MCS"** between the hit-list MCS and the query. We then describe the overlap using C3 and the "relative size" HQ:

$$HQ = \frac{G_{\text{hit-list MCS}}}{G_{query}}$$

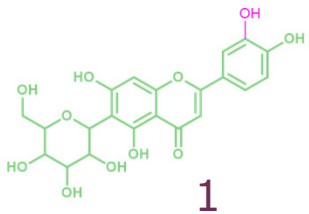
The hit-list MCS…

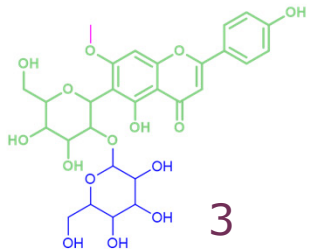
	HQ > 1	HQ = 1	HQ < 1
C3 = 1:	is larger than and completely covers the query.	***is*** the query!	is smaller than and completely included in the query.
C3 < 1:	is larger than but only partly covers the query.	is the same size as but only partly covers the query.	is smaller than but only partly included in the query.

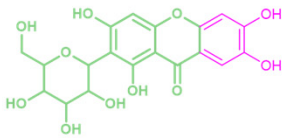
"Thresholded" multiple-MCS calculations:
RDKit: Open-source cheminformatics; http://www.rdkit.org

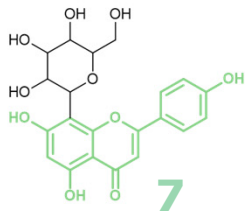
Fingerprint- versus graph-based similarity metrics:
Raymond, Willett, *J. Comput.-Aided Mol. Des.* **2002**, *16*, 59

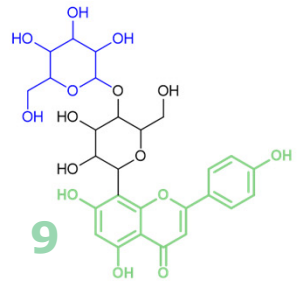
Hit-List MCS (up to 10 hits, *with ID* hits)

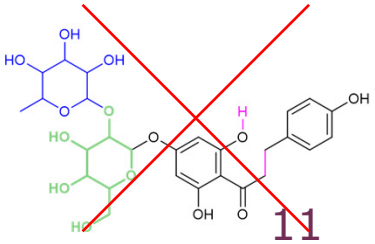


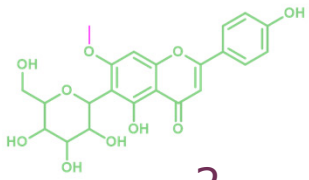


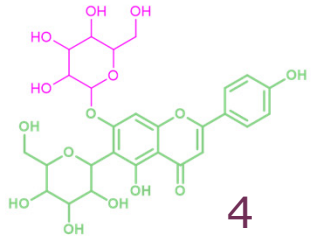


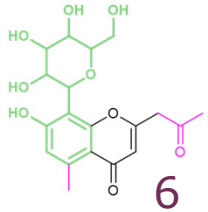


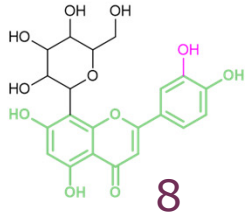


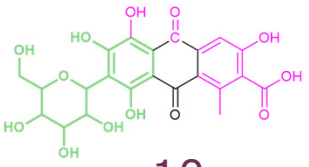


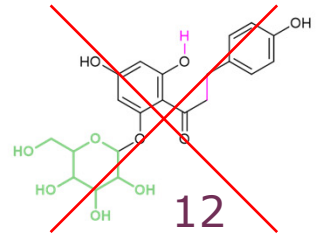




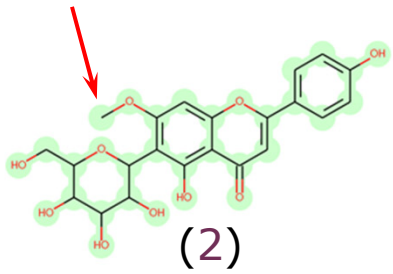




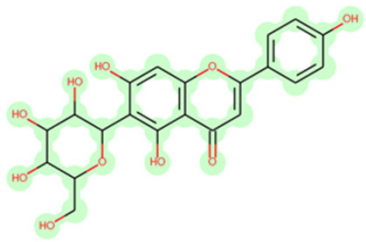




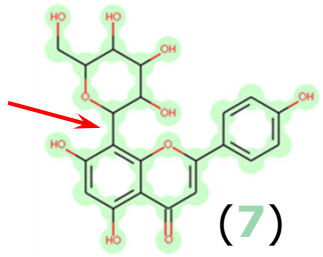
2 of 3–10
HQ = 1.031
C3 = 1.0
Δ*m* = −14.0157



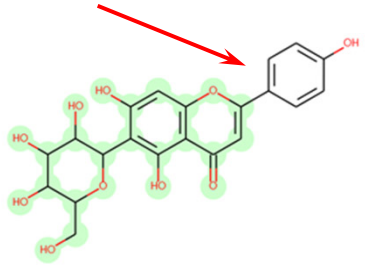
2 of 2
3 of 3–8
4 of 4–10
HQ = 1.0
C3 = 1.0
Δ*m* = 0.0000



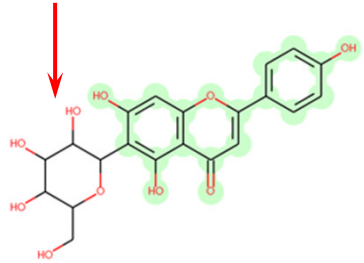
3 of 9–10
HQ = 1.0
C3 = 0.646
Δ*m* = 0.0000



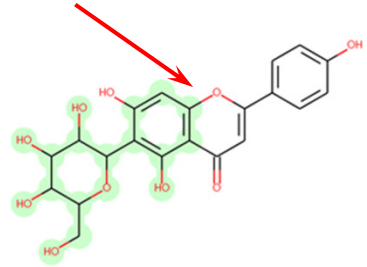
5 of 5–10
HQ = 0.769
C3 = 1.0
Δ*m* = 92.0262



6 of 8–10
7 of 9–10
HQ = 0.646
C3 = 1.0
Δ*m* = 162.0528

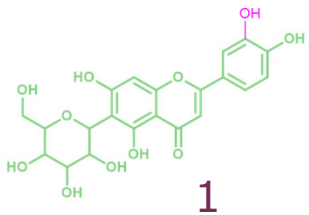


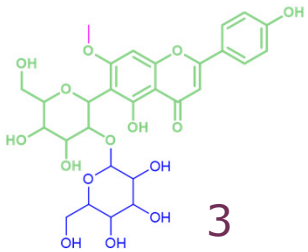
6 of 6–7
7 of 7–8
8–10 of *m*–10
HQ = 0.569
C3 = 1.0
Δ*m* = 176.0109

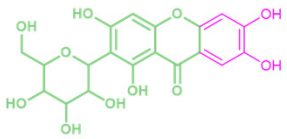


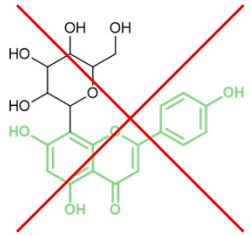
MCS from ***best m* of *top n*** hits overlaid on the query (or on a numbered hit if needed): $\Delta m = m_{query} - m_{MCS}$.

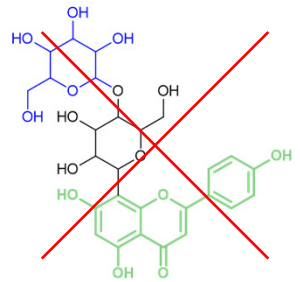
Hit-List MCS (up to 10 hits, *without ID* hits)

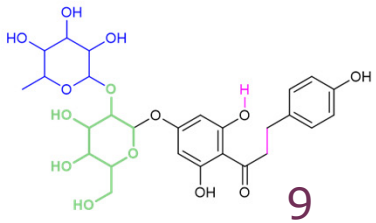

1

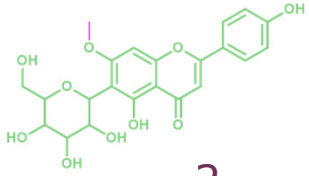

3

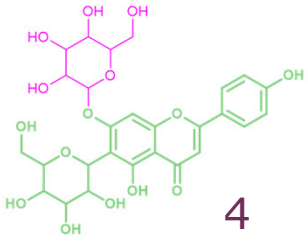

5

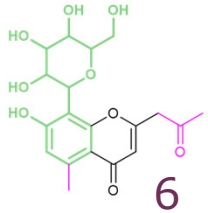


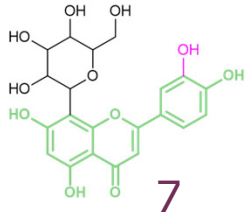


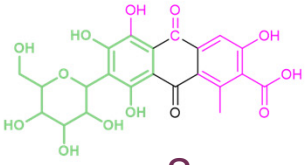

9

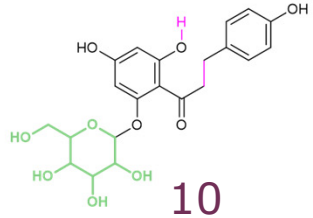

2


4

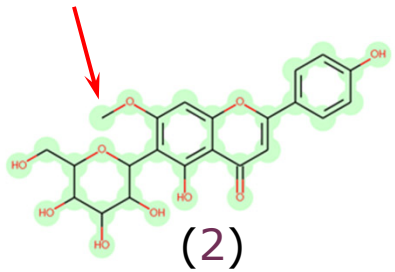

6


7

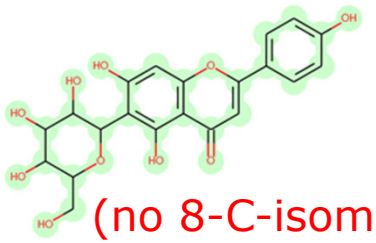

8


10

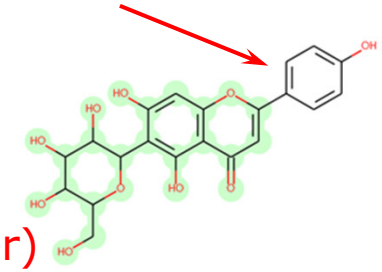
2 of 3–10
HQ = 1.031
C3 = 1.0
Δ*m* = −14.0157


(2)

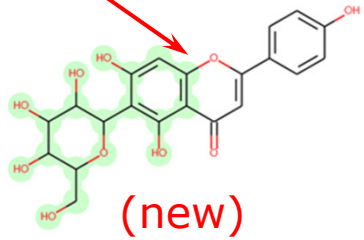
2 of 2
3–4 of *m*–10
HQ = 1.0
C3 = 1.0
Δ*m* = 0.0000


(no 8-C-isomer)

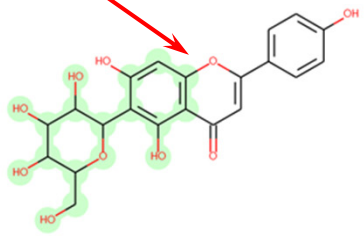
5 of 5–10
HQ = 0.769
C3 = 1.0
Δ*m* = 92.0262



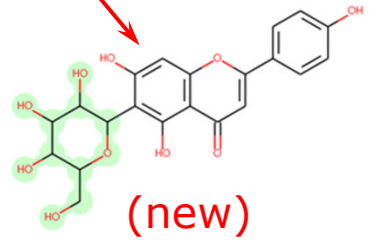
6 of 8–10
HQ = 0.600
C3 = 1.0
Δ*m* = 160.0160


(new)

6 of 6–7
7–8 of *m*–10
HQ = 0.569
C3 = 1.0
Δ*m* = 176.0109

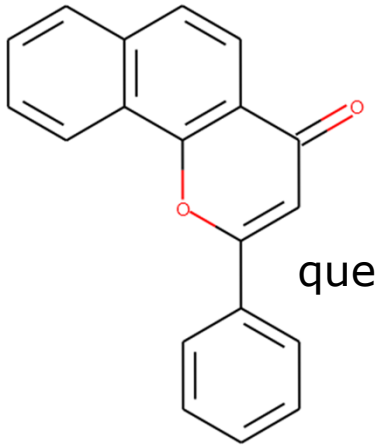


9–10 of *m*–10
HQ = 0.338
C3 = 1.0
Δ*m* = 268.0372


(new)

MCS from ***best m* of *top n*** hits overlaid on the query (or on a numbered hit if needed): $\Delta m = m_{query} - m_{MCS}$.

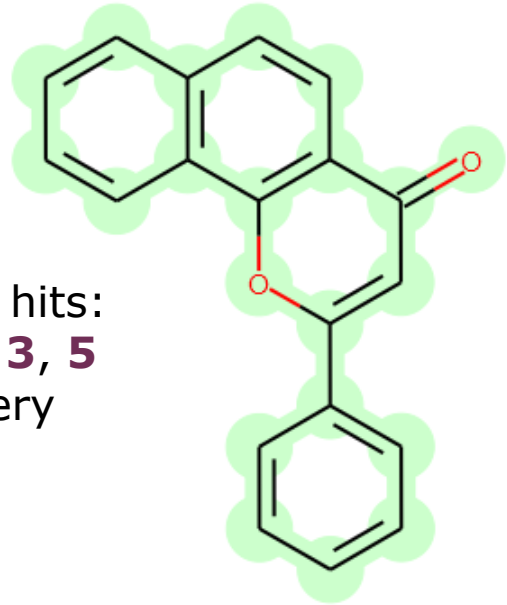
Fused-Ring Positional Isomers



query

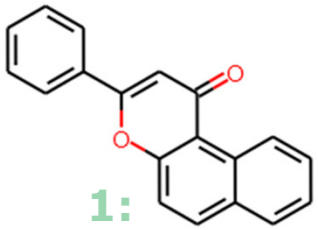
FindMCS settings:

- ring matches ring only;
- complete rings only.

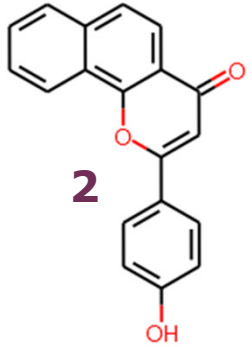


best 3 of top 5 hits:
MCS of hits **2**, **3**, **5**
overlaid on query

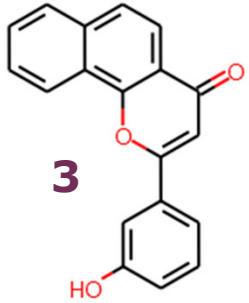
top 8 hits (all but **1** are ***Hyb***)



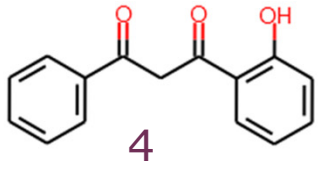
1:
ID



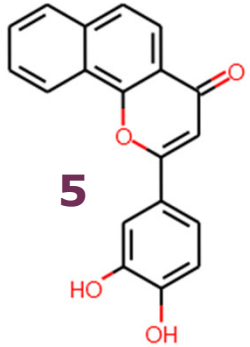
2



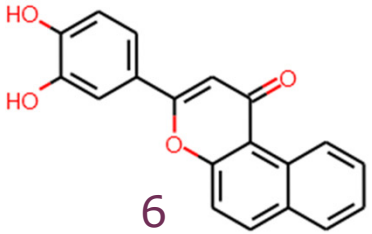
3



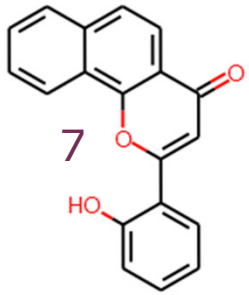
4



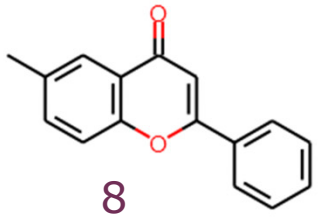
5



6

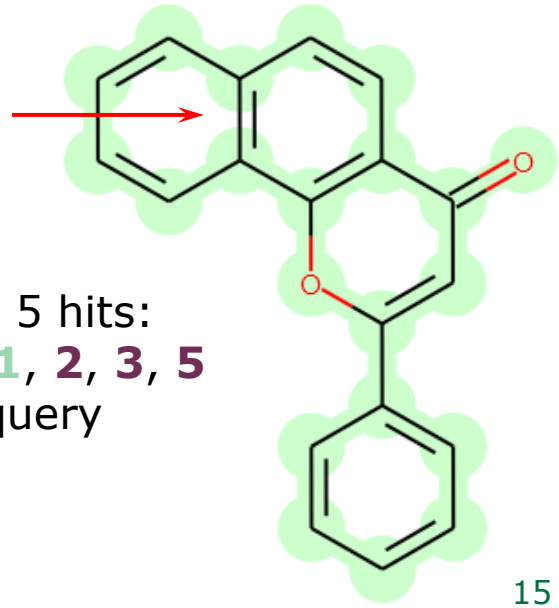


7



8

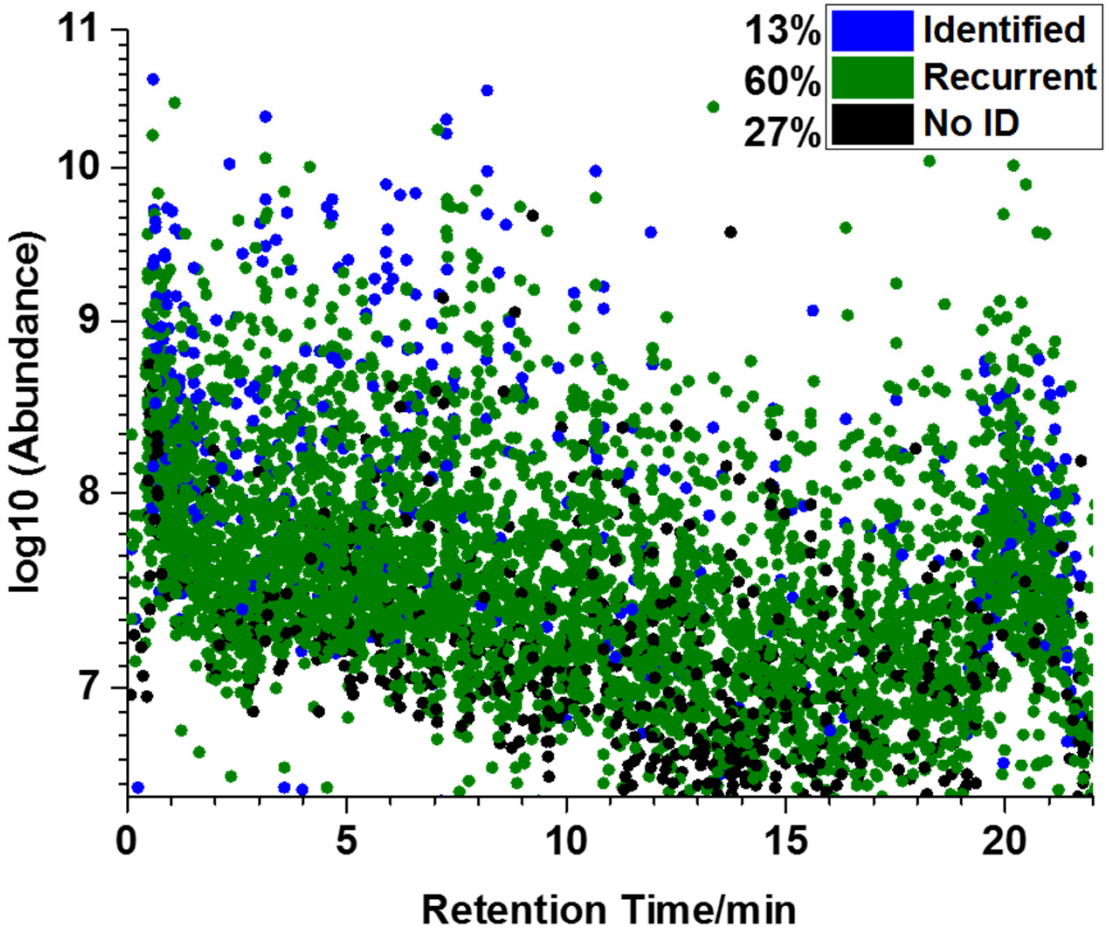
note
gap!



best 4 of top 5 hits:
MCS of hits **1**, **2**, **3**, **5**
overlaid on query

Conclusions and Future Work

- The hybrid search works best at lower collision energies, which produce larger fragments that are more likely to contain the structural difference and thus match after peak shifting.
- Valuable structural information occurs throughout typical hybrid search hit lists, and thresholded multiple-MCS calculations can help extract it.
- Paying attention to match types (especially ***ID*** hits to isomers) and MCS delta mass (using a look-up table of common modifications) can improve MCS-based interpretation.
- The short-term goal of this work is to produce an interactive, MCS-based tool that will help analysts propose likely structures for unknown analytes.
- Our ultimate goal remains to identify recurrent unidentified spectra obtained from biological samples.



NIST SRM 3667 (creatinine in urine)

Acknowledgments:

- Mass Spectrometry Data Center colleagues
- NIST awards 70NANB19H026 and 70NANB18H167