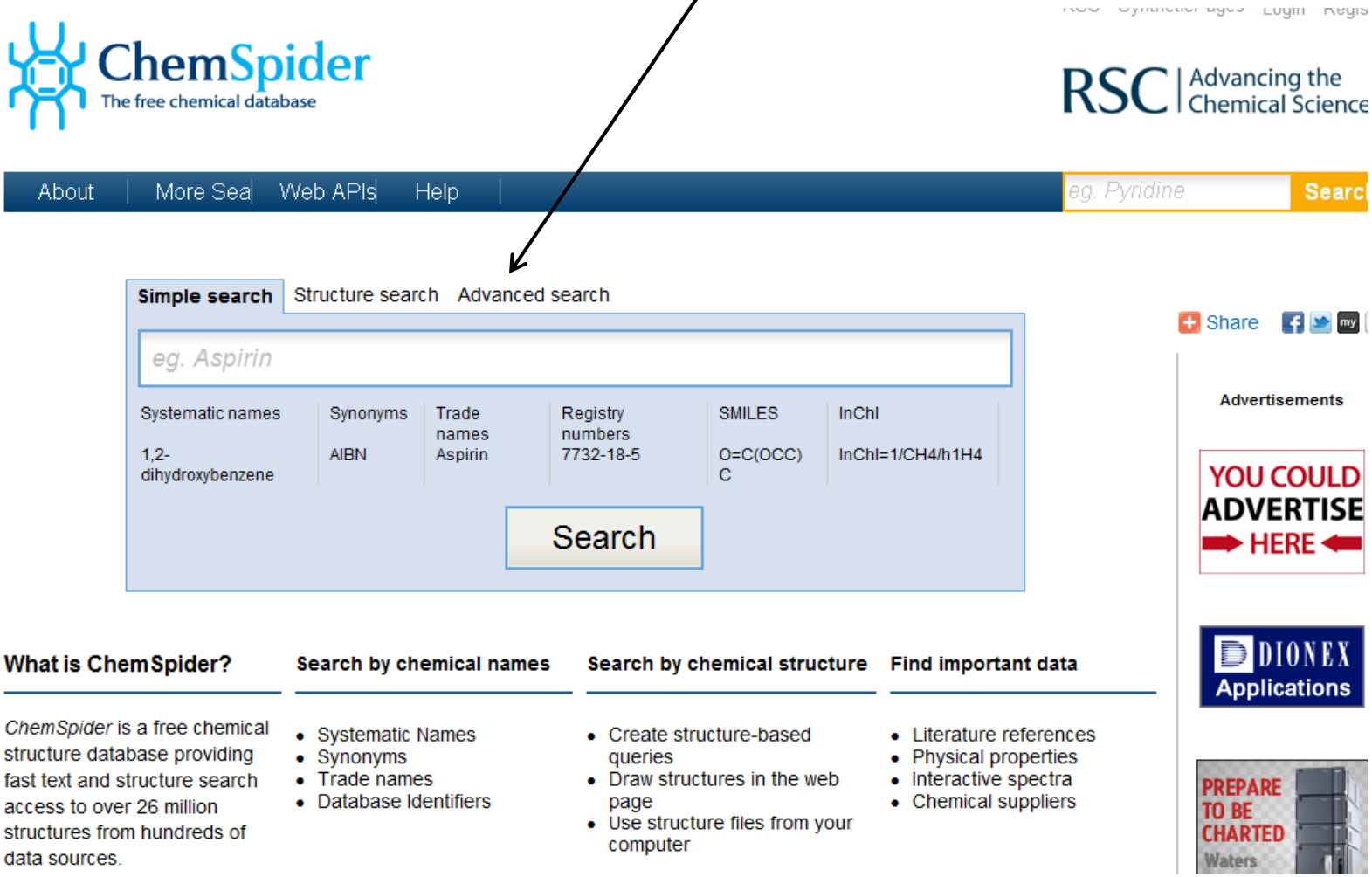
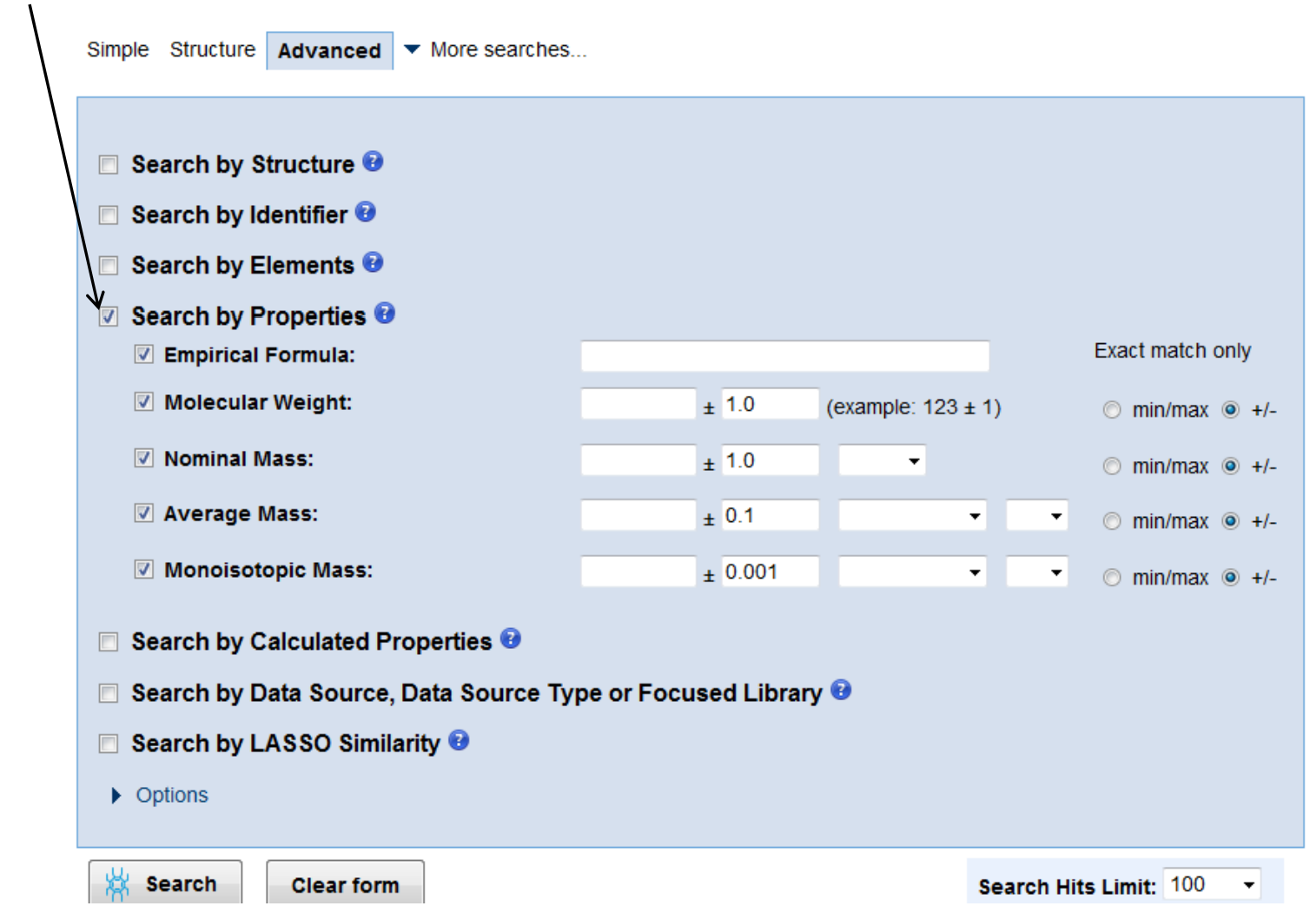
ChemSpider Main Menu

Select "Advanced Search"



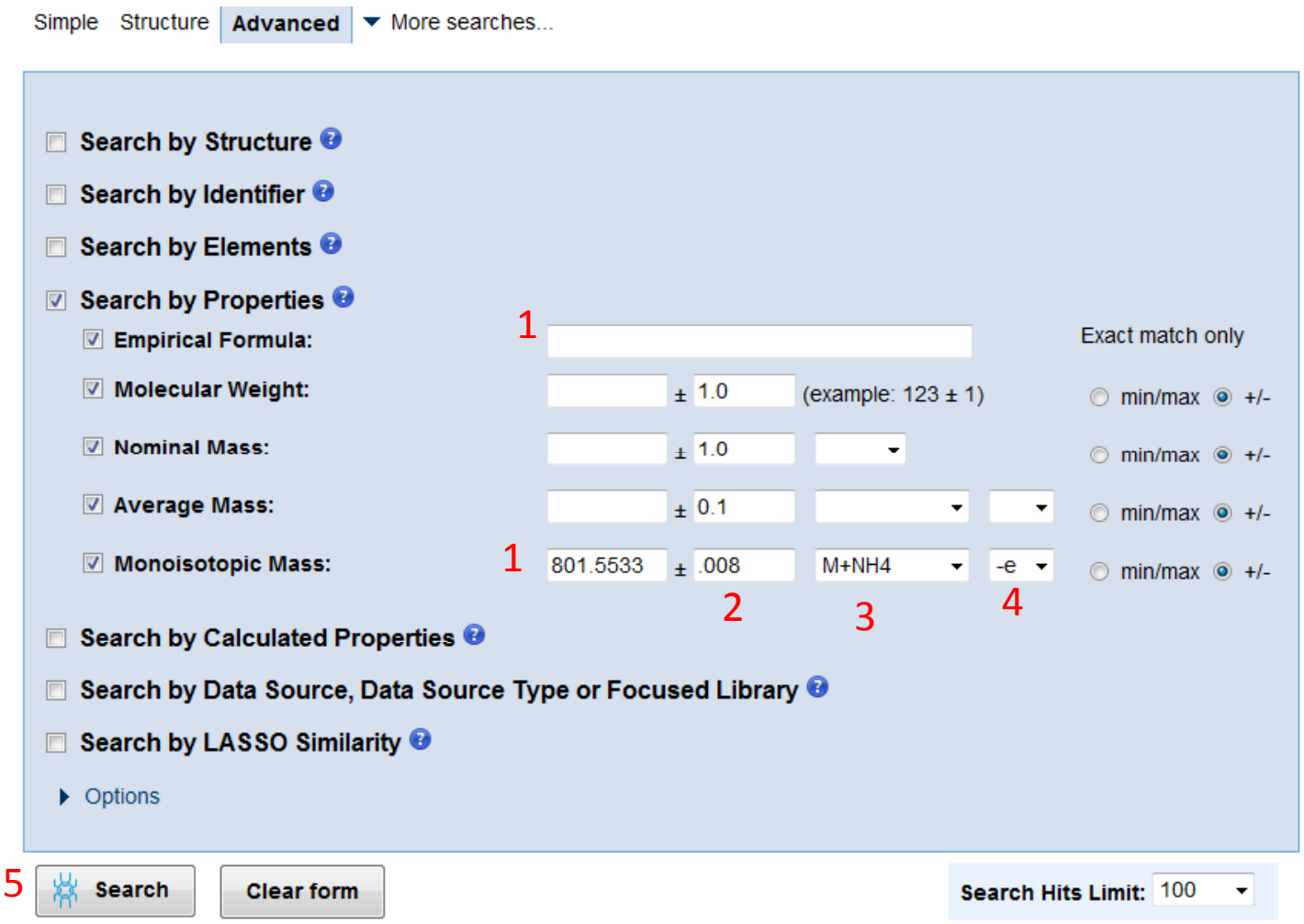
ChemSpider Advanced Menu

Select "Search by Properties"



Search by Formula or Mass

1) *Enter formula or mass 2) for mass enter mass error window 3) select type of ion 4) correct for mass of electron if necessary 5) search*

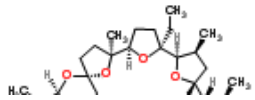


Sort by Desired Column Descending

Sort in descending order by clicking once on desired column

9 hits found in 0.55 seconds.
Search terms: MM >= 783.510925865 AND MM <= 783.526925865

Grid | Tile | Table | Names/Structures | Names

ID	Structure	Empirical Formula	Molecular Weight	# of Data Sources	# of References	# of PubMed	# of RSC	Mass defect
84061		$C_{48}H_{69}N_3O_6$	784.078	17	29	0	0	-0.0003
9940470 - 5/5 defined		$C_{42}H_{69}N_7O_7$	784.04	5	5	0	0	0.0069
23315313								

Accurate Mass Measurements: Identifying "Known Unknowns" Using ChemSpider

James Little, Eastman Chemical Company
Antony Williams, Valery Tkachenko, ChemSpider/Royal Society of Chemistry

Wednesday Poster Session 361
2011 ASMS Conference
Denver, Colorado

ChemSpider
Building community for chemists

/

RSC | Advancing the Chemical Sciences

EASTMAN

Overview

- "Known Unknowns" can be routinely identified
- Known in the literature, unknown to the investigator[1,2]
- Obtain accurate mass data from electrospray LC-MS or EI/CI GC-MS
- ChemSpider searched with molecular formulae (MF) or monoisotopic molecular weights (MW)
- ChemSpider modified to sort structure list by number of associated references
- Evaluate likely structures by sample history, MS/MS, exchangeable protons, derivatization, hydrolysis, etc.

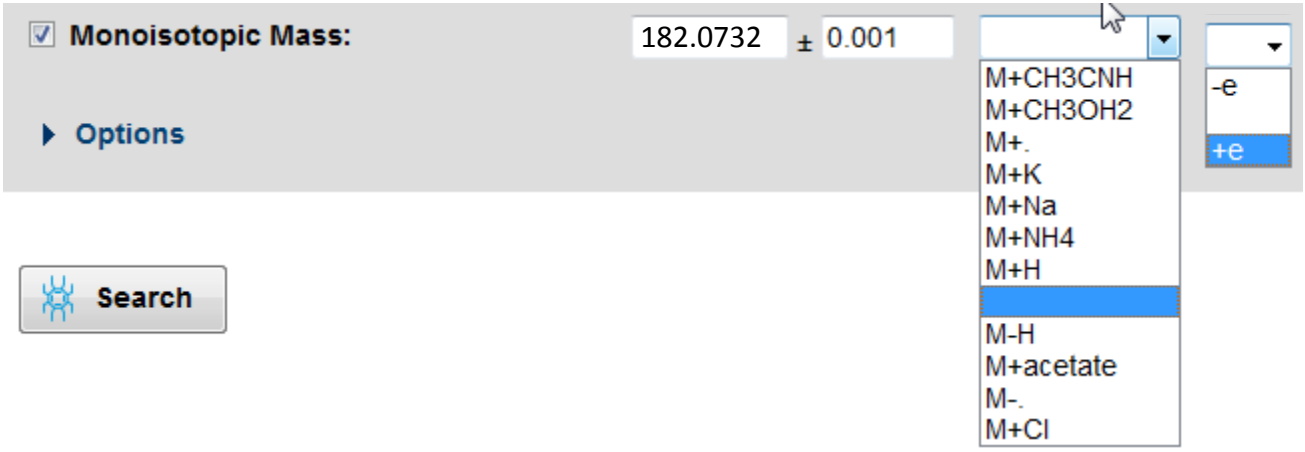
Introduction

An unknown to an investigator, in many cases, is often known in the chemical literature. ChemSpider is a very valuable internet resource of known substances. It contains ~25 million compounds which can be searched by a variety of parameters and includes valuable links to other web-based resources. ChemSpider is owned by the Royal Society of Chemistry (RSC) and provided as a free resource to the community.

Modifications in the web-based interface were made to facilitate the entry of monoisotopic molecular weight data and to sort the search results by the number of associated references. The ability to sort by the number of references was demonstrated to be very useful in bringing the most likely candidates to the top of the list. A similar approach was previously demonstrated utilizing the CAS Registry which was searched by either SciFinder or STN Express.[1,2]

ChemSpider Changes

The interface for entering monoisotopic MW was changed to allow the user to specify a variety of types of ions and ion adducts for searching:

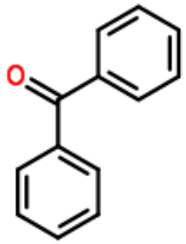
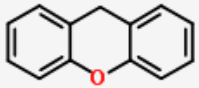


The entered value can be corrected for data systems that ***incorrectly*** calculate the accurate mass by not considering the mass of the electron.[3]

ChemSpider Changes

The user can sort by “clicking” on the desired field including # of References, # of PubMed references, # of RSC references, etc. In the example below, the initial results were sorted in descending order by “# of References.”

1 2 3 4 5

ID	Structure	Empirical Formula	Molecular Weight	# of Data Sources	# of References	# of PubMed	# of RSC	Mass defect
2991		$C_{13}H_{10}O$	182.2179	46	671	6	11	0.0000
6840		$C_{13}H_{10}O$	182.2179	27	552	37	38	0.0000

Evaluation of Approach

A group of 90 compounds was assembled from literature sources, internet sites, and American Society for Mass Spectrometry Conference presentations. The same group of compounds was previously used in evaluating a similar approach with the CAS Registry.[1,2]

The group was searched by molecular formulae (MF) and monoisotopic molecular weights (MW, +/- 3 ppm window) and sorted by descending "# of References." The results in Tables 1 and 2 for the compounds showed that the MF is better for searching for candidates, but that the MW is also a viable approach. The results for MF search using ChemSpider were similar to those noted with the CAS Registry/SciFinder approach.[1,2]

The compounds whose ranking was decreased, e.g. polymer UV stabilizers, had MW's in the range of 300-400 Daltons. This decreased ranking is not surprising since a large number of compounds in the database fall within this MW range (see Figure 1).

Evaluation of Approach

Table 1: Searching ChemSpider by MF and sorting "# of References" descending.

Class of Compounds	# in Class	Position of Compound Sorted in Descending Order by Number of References					
		#1	#2	#3	#4	#5	>#5
Drugs	45	43	1	1			
Pesticides	8	7	1				
Toxins	2	2					
Polymer antioxidants	15	15					
Polymer UV Stabilizers	10	8	1	1			
Polymer Clarifying (Irgaclear DM)	1						1
Polyurethane additives	4	2	1			1	
Natural products	3	2		1			
Herbicide (clofibric acid)	1	1					
Artificial sweetener (Sucralose)	1	1					
Total Compounds	90	81	4	3		1	1

Evaluation of Approach

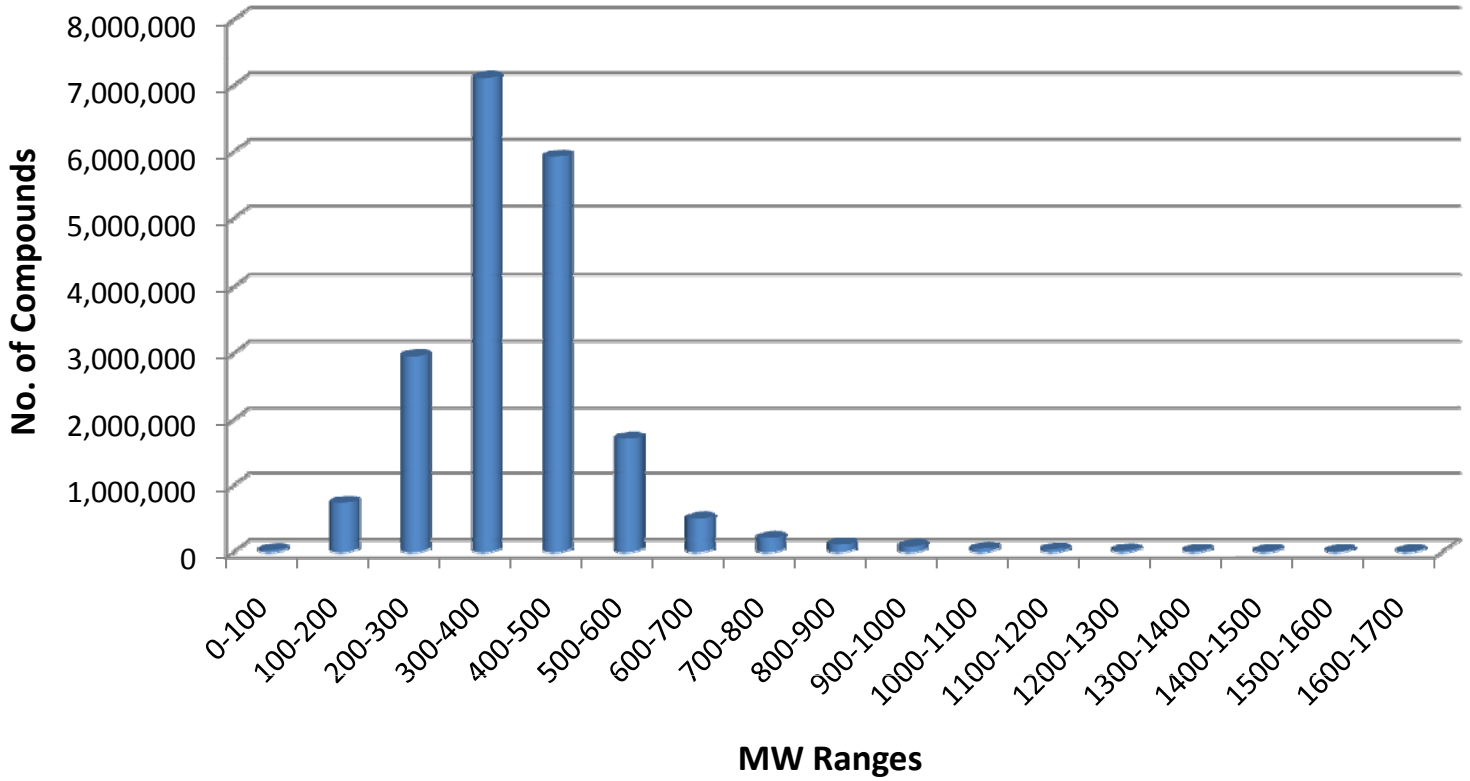
Table 2: Searching ChemSpider by MW and sorting "# of References" descending.

Class of Compounds	# in Class	Position of Compound Sorted in Descending Order by Number of References					
		#1	#2	#3	#4	#5	>#5
Drugs	45	43	1	1			
Pesticides	8	7	1				
Toxins	2	2					
Polymer antioxidants	15	13		1			1
Polymer UV Stabilizers	10	6	1	1		1	1
Polymer Clarifying (Irgaclear DM)	1						1
Polyurethane additives	4	2	1			1	
Natural products	3	2		1			
Herbicide (clofibric acid)	1	1					
Artificial sweetener (Sucralose)	1	1					
Total Compounds	90	77	4	4		2	3

MW Distribution in Database

In theory, the number of possible molecular formulae increase exponentially with MW. However *in practice* the number of known compounds tends to maximize at between 300-600 and falls exponentially with increasing MW. A similar profile is noted in both ChemSpider and the CAS Registry.[1,2]

Figure 1: ChemSpider Entries vs. MW



Compounds MW>600

It is often difficult to determine a *unique* MF for unknowns with MW's > 600 Daltons.[4,5] Either there are too many candidates or one is inadvertently excluded. This is due to the exponential increase in MF's as MW increases and the subjective process of selecting the elements, the range of elements, and double bond equivalents. Since in reality, the number of candidates decrease greatly at MW's>600 in the ChemSpider data base, it is logical to search by MW instead of attempting to obtain a MF for searching.

Of course post facto, it is still extremely useful to employ processing programs that rank isotope intensities of the MF candidates obtained from the ChemSpider search to that of the unknown to further simplify the list.

The ranking noted by number of references for several higher MW compounds noted in the literature[4,5] is compared to a MF versus a MW search below. There is essentially no significant penalty noted for this limited number of examples searching by MW instead of MF.

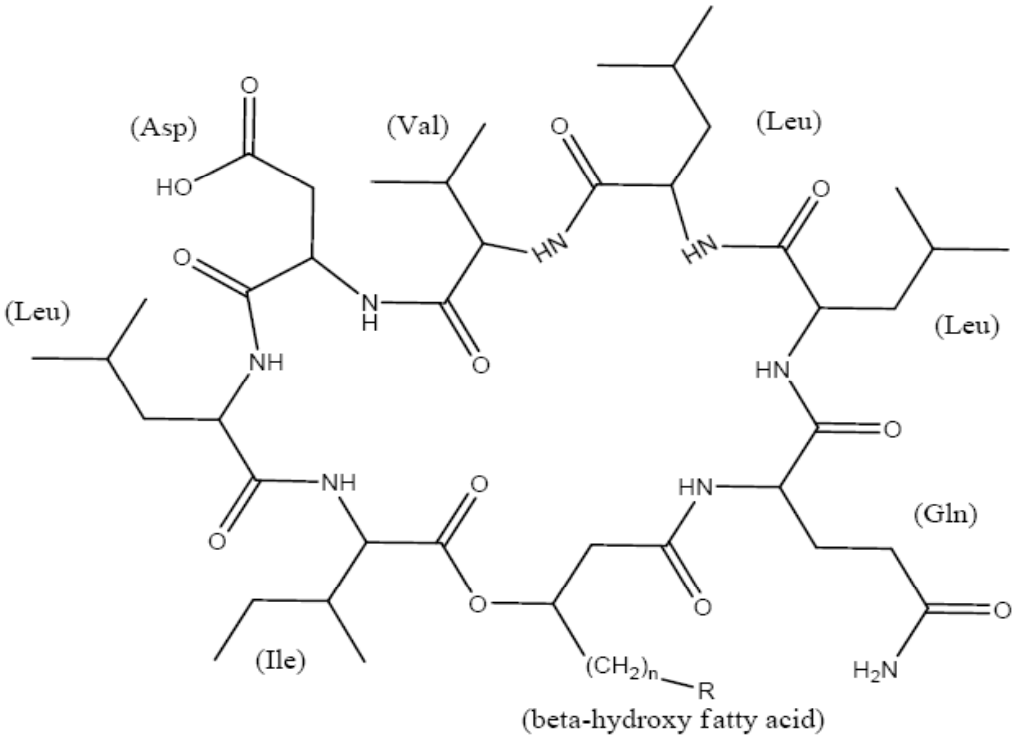
Compounds MW>600

Species	MF	MW	Rank MF	Rank MW using +/- 5 ppm window
Moxidectin[4]	C37H53NO8	639.3771	1 of 5	1 of 39
Erythromycin[4]	C37H67NO13	733.4612	1 of 42	1 of 53
Digoxin[4]	C41H64O14	780.4296	1 of 47	1 of 65
Rifampicin[4]	C43H58N4O12	822.4051	1 of 29	1 of 96
Rapamycin[4]	C51H79N1O13	913.5551	1 of 43	1 of 51
Amphotericin B[4]	C47H73N1O17	923.4878	1 of 33	1 of 42
Gramicidin S[4]	C60H92N12O10	1140.7059	1 of 5	1 of 13
Cereulide[5]	C57H96N6O18	1152.6781	1 of 3	2 of 8
Cyclosporin A[4]	C62H111N11O12	1201.8414	1 of 36	1 of 38
Vancomycin[4]	C66H75Cl2N9O24	1447.4302	1 of 24	1 of 26
Agilent Calibrant Ion[4]	C30H18N3O6P3F48	1520.9642	1 of 1	1 of 1
Thiostrepton[4]	C72H85N19O18S5	1663.4924	1 of 5	1 of 5

Eastman Example

- Identification of material isolated from bacteria
- Electrospray, Waters LCT LC-MS
- Complex mixture, Δ14.015 milli-Daltons for CH_2 repeat groups
- ChemSpider searched with M+H, m/z 1021.6985 +/- 15 ppm
- Best candidate (1 of 12) consistent with a biosurfactant
- Other homologues at m/z 1007.6824 and 1035.7124 in ChemSpider
- Good fits to theoretical isotope patterns
- HCl hydrolysis yielded β-hydroxy fatty acids, GC-MS of bis-trimethylsilyl derivatives confirmed structure
- Amino acid sequence determined[5] by MS/MS
- Amino acid sequence would be required to determine if leucine or isoleucine

Eastman Example



1. n=6, R = $CH_2CH(CH_3)_2$, $C_{51}H_{90}N_8O_{12}$, Exact Mass: 1006.6678, *iso-isomer*
2. n=6, R = $CH(CH_3)CH_2CH_3$, $C_{51}H_{90}N_8O_{12}$, Exact Mass: 1006.6678, *anteiso-isomer*
3. n=7, R = $CH_2CH(CH_3)_2$, $C_{52}H_{92}N_8O_{12}$, Exact Mass: 1020.6835, *iso-isomer*
4. n=7, R = $CH_2(CH_2)_2CH_3$, $C_{52}H_{92}N_8O_{12}$, Exact Mass: 1020.6835, *n-isomer*
5. n=8, R = $CH_2CH(CH_3)_2$, $C_{53}H_{94}N_8O_{12}$, Exact Mass: 1034.6991, *iso-isomer*
6. n=8, R = $CH(CH_3)CH_2CH_3$, $C_{53}H_{94}N_8O_{12}$, Exact Mass: 1034.6991, *anteiso-isomer*

Conclusions

Changes in the ability to sort the results in ChemSpider offered improved capabilities for the identification of “known unknowns.” Furthermore, changes were made to allow easier input of various types of ions and ion adducts for MW searching and to correct for the mass of an electron.

The molecular formula search is the preferred one in the low molecular range (200-600 Daltons), but even the MW search with reasonable error windows is viable in this range. MW searching at higher molecular weights (>600 Daltons) is preferred when it is difficult to determine a unique molecular formula. The resulting candidates should then be ranked by comparing their isotopic data within the manufacturer’s isotope programs to that of the unknown.

In all cases, other data such as sample history, GC-MS, exchangeable protons, MS/MS, hydrolysis, derivatization, NMR, etc. should be used to confirm the proposed structure. For critical applications, a standard should be purchased for confirmation.

Conclusions

Search Approach	Pros	Cons
ChemSpider	-free -automation by instrument manufacturer using ChemSpider Web API (Application Program Interface) -ability to search monoisotopic MW	-smaller database than CAS Registry -not able to search by key words
CAS Registry with SciFinder or STN Exprss	-larger data base of compounds and references -highly curated, well defined abstract process -ability to search with key words for more obscure compounds with fewer references	-expensive -no API available for instrument manufacturer automation -no ability to search by monoisotopic MW, ***only*** average MW with ***complicated*** STN Express Interface

Acknowledgments

Many thanks to Alexey Pshenichnov (ChemSpider/RSC) for programming web interface enhancements and to Tom Barber of East Tennessee State University for isolation of the biosurfactants from bacteria. Also thanks to C. Cleven and S. Brown for the initial work using the CAS Registry.

References

1. Known Unknowns Poster, "Little" Mass Spec and Sailing, ASMS 2010, Salt Lake City, http://littlemsandsailing.wordpress.com/
2. "Known Unknowns Paper," J. Little, C. Cleven, S. Brown, JASMS, **22**, 348-359, (2011).
3. Mamer, O. A.; Lesimple, A. Common Shortcomings in Computer Programs Calculating Molecular Weights. *JASMS., **14,** 626 (2004).*
4. "Spectral Accuracy in Orbitrap," J. Erve, M Gu, Y. Wang, W. DeMaio, R. Talaat, JASMS, 20, 2058-2069 (2009).
5. "Molecules >1100 m/z," W. Lohmann, P. Decker, A. Barsch, Bruker Applications Note # ET-23.
6. S. Yang, D. Wei, and B. Mu, J. Biochem. Biophys. Methods 68 (2006) 69-74.