

Quantum Chemical Electron Impact Mass Spectrum (QCEIMS) prediction for de novo structure elucidation – assessment against experimental reference data and comparison to Competitive Fragmentation Modelling (CFM)

• Peter R. Spackman*, Björn Bohman, Amir Karton, Dylan Jayatilaka

August 4, 2017

Abstract

We investigate the success of the Quantum Chemical Electron Impact Mass Spectrum (QCEIMS) method in predicting the electron impact mass spectra of a diverse test set of 61 small molecules selected to be representative of common fragmentations and reactions in electron impact mass spectra. Comparison with experimental spectra is performed using the standard matching algorithms, and the relative ranking position of the actual molecule matching the spectra within the NIST-11 library is examined. We find that the correct spectrum is ranked in the top two matches from structural isomers in more than 50% of the cases. QCEIMS thus reproduces the distribution of peaks sufficiently well to identify the compounds, with the RMSD and mean absolute difference (MAD) between appropriately normalized predicted and experimental spectra being at most 9% and 3% respectively, even though the most intense peaks are often qualitatively poorly reproduced. We also compare the QCEIMS method to competitive fragmentation modeling for electron ionization (CFM-EI), a training-based mass spectrum prediction method, and remarkably we find the QCEIMS performs equivalently or better. We conclude that QCEIMS will be very useful for those who wish to identify new compounds which are not well represented in the mass spectral databases.

1 Introduction

Mass spectrometry (MS) is of major importance for identifying the presence of small quantities of a compound in a mixture, particularly because it is several orders of magnitude more sensitive than equivalent structure determination methods²⁸. Indeed, MS is often the only feasible identification method^{24,18,19} in contexts like the identification of semiochemicals^{8,6,7}, where abundances of critical compounds may be in the ng range.

*School of Molecular Sciences, University of Western Australia, Crawley

Electron-impact mass spectrometry (EIMS) coupled to some chromatographic separation techniques is, at present, likely to be the predominant form of MS practiced for small molecules. The typical protocol for identifying unknown compounds using EIMS involves matching against a known mass spectrum (MSp), using spectral libraries in conjunction with software to identify matches^{15,30,31,29,35,22}. For example, the National Institute of Standards and Technology (NIST) mass spectral library (version 11) contains electron impact (EI) spectra of more than 2×10^5 compounds, and may be readily searched using the NIST MS Search program. However large, such a library may only contain a tiny fraction of the total number of small molecular species found in the universe, which for molecules less than 500 Daltons in weight is estimated to be in excess of 10^{60} ¹⁰. Simply put, the huge number of possible structural isomers for each molecular formula renders it practically impossible for a library to be exhaustive. If the identification of a compound whose spectra has yet to be characterized (i.e. its spectra is not in a library) is desired, much of the value provided by MS matching/library methods is lost (though the structure of the close matches may still provide insight).

In this paper, we are concerned with the problem of identifying the structures of unknown compounds present in low abundances which have not been previously characterized by MS. We will use the term *de novo* structure determination from the mass spectrum (MSp) to describe this. In this case, the only way to confirm the structure is to synthesize several putative candidate compounds in significant amounts, characterize their structure independently (say using nuclear magnetic resonance (NMR), or X-ray crystallography), then compare the candidates MSp with the original unknown spectrum to see which (if any of them) agree. It should come as no surprise that this is a time- and labor-intensive process.

One way to ease the process of *de novo* compound identification is to predict the MSp for each candidate using computational methods. If such computed MSp were accurate enough, only one candidate compound need be synthesized and characterized in order to confirm the structure. Indeed, if the computed MSp were always reliably accurate it would not even be necessary to confirm the structure by synthesis and alternative means; MS may then become competitive with NMR or crystallography as a structure determination method.

Computational methods for the prediction of the MSp have been well reviewed by Bauer and Grimme⁵. They may be broadly separated into two categories:

1. *Expert systems.* Expert systems produce a MSp using rules derived from experimental spectra, for example fitted to experimental MSp using a model such as a neural network^{38,17}. The 'system' is trained or based on a library of MSps for which the structure is known. Indeed, the prototypical example of an expert computer system based on rules²⁷ concerned the identification of molecular structure from MSp: the DENDRAL⁹ project. Regardless of the successes (or failures^{11,28}) of the original DENDRAL project, recently learning-based methods such as competitive fragment modeling (CFM)^{1,2} for predicting EIMS spectra (CFM-EI) have gained more traction, and proved more successful. It should be noted, though, that within any such system there will always exist some bias, both in the model it represents and the data used to 'train' it; such bias may lead to unquantifiable inaccuracies.
2. *First principles simulation techniques.* First principles simulation techniques aim to actively simulate the underlying physical processes involved in MS. Such methods are

not in principle dependent on an existing library of known spectra, but are limited by the accuracy with which one is able to solve the time-dependent Schrödinger equation (or more correctly, the time-dependent Liouville equations for the density matrix²⁵). Unfortunately such calculations were, in general, time consuming to the point of being impractical. However, recently the QCEIMS program from Grimme et al.¹² has allowed practical calculation of MSps using *ab initio* techniques. By directly modeling the time-dependent fragmentation process via Born-Oppenheimer molecular dynamics (BOMD) finite-temperature (fractional orbital occupied) quantum chemical (QC) methods^{12,13,5}, coupled with knowledge of isotopic distributions³³, QCEIMS has proven remarkably accurate. Other quantum chemical methods based on bond indices have also been proposed²⁰, and are examined by us in related work.

The purpose of this paper is to investigate the performance of, QCEIMS, which, although having been applied to many molecules^{12,4,5,13,3}, has not yet been tested by the conventional MSp matching algorithms used to identify compounds from the MSp; an essential task if it is to be used for *de novo* compound identification. For this purpose, we make use of the standard programs and databases from the National Institute of Standards (NIST). We also provide an estimate of the errors associated with their predicted spectra, using counting statistics. Our goal is to identify patterns or trends with any errors in the prediction, particularly those seen to be associated with a chemical class of compounds; such information is invaluable for practical application of QCEIMS.

Despite expert systems and first principles methods having radically different philosophical approaches, for the purposes of *de novo* structure determination these differences mean little. As such, we also assess the performance of QCEIMS method against the CFM-EI model-based spectrum calculator, both as a contemporary reference point and to investigate the differences in accuracy between the two techniques.

2 Methodology

All data analysis and processing of calculations for this paper was performed using the Python programming language. Notably, the numpy³⁴, pandas²¹ packages, and matplotlib¹⁶ for the generation of figures.

2.1 Selection of compounds for analysis

When selecting a set of molecules for investigation, the following criteria were considered: they must have an experimental MSp and structural isomers with experimental MSp, they must be representative of common/important fragmentations in EIMS i.e. made up of a variety of functional groups, they must be relatively small molecules ($n_{atom} < 25$). Further, it is desirable to have a mixture of easy and difficult cases for differentiating between structural isomers. As such, a set of 61 small molecules were chosen, made up of alcohols (6), aldehydes (3), alkenes (7), amines (6), alkylbenzenes (8), carboxylic acids (5), esters (16), ketones (2), alkylphenols (8). The complete list is shown in Table 1, and the structures are available in Figure 1. The primary motivation for the selection of small molecules is the identification of failed predictions and their underlying cause. Both of these aspects are significantly

more straightforward in the case of small molecules. Further, the reduced computational time requirements associated with quantum-chemical predictions with small molecules render them ideal for such a study. Additionally, when establishing the viability for such methods in *de novo* structure elucidation, having several structural isomers distinguishable via their MSp is vitally important.

2.2 QCEIMS and CFM-EI calculations

Initial geometries were drawn by hand in a standard molecular builder program Avogadro¹⁴ and then optimized using the generalized Amber force field³⁶. MSp prediction calculations were then performed using the QCEIMS program version 2.26, as described by Grimme and coworkers^{12,5} using the default parameters, namely: using the OM2-D3³⁷ semi-empirical quantum mechanical wavefunction method from MNDO99³² version 7.0. The default QCEIMS parameters were used, with an initial temperature of 500K, at 70 eV and the number of trajectories set to $25 \times n_{heavyatoms}$. The resulting spectra were normalized in the typical manner: such that the highest peak had an intensity of 100. The time taken for calculations varied with molecular size (see ESI Table S2 for full details), with the shortest being methyl formate (roughly 1.5 CPU hours) and the longest being 4-propan-2-ylphenol (roughly 80 CPU hours). It should be noted that these calculations are readily parallelized across computing clusters, drastically reducing the wall-clock time required. All calculations were performed on the local Linux computing cluster, which consists of 44 nodes (with a variety of configurations) containing Intel Xeon CPUs.

CFM-EI Calculations were performed using the method described by Allen *et al.*² with the provided trained parameters for EI-MS, using the Windows executable 'cfm-predict', version 2.2.

2.3 Spectrum matching

Rankings between standard-normalized calculated and experimental MSp were made using the NIST MS Search v2.0g software coupled with the NIST11 database. Search was performed with the 'identity' matching method where the square root of intensity of the MSp was taken and weighted by the mass-to-charge ratio squared. The difference between this and the available 'similarity' metric (which differs only in the mass weighting) is negligible for our purposes – see³¹ for a comprehensive study of the impact of matching algorithms on the accuracy of matches). Searches were carried out by limiting to molecules to have the same molecular formula, using only the two most abundant isotopes.

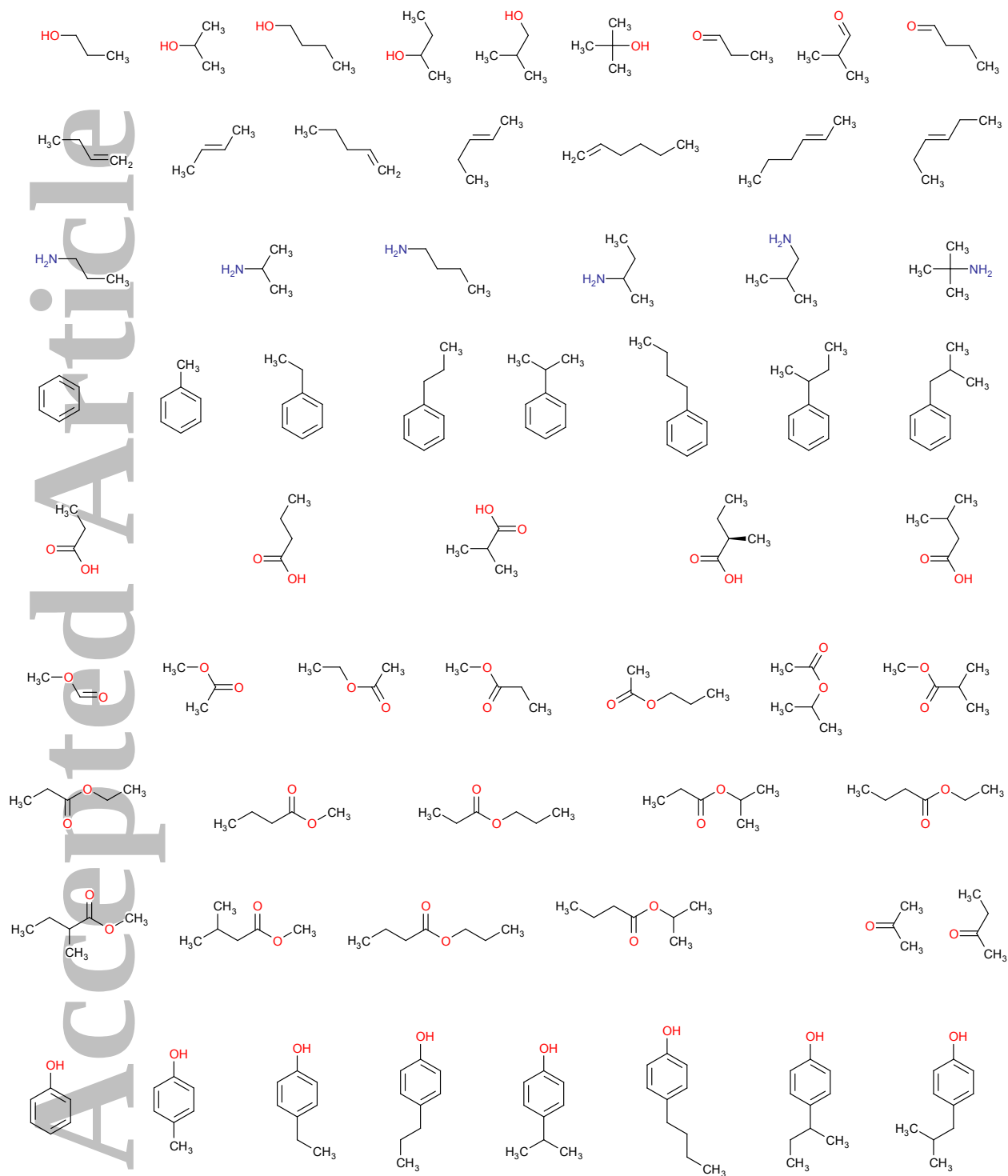
2.4 Ranking comparison

It is not meaningful to directly compare ranks when the number of possible candidates differs. To overcome this, the relative ranking position^{18,2} has been used:

$$RRP = \frac{r - 1}{n - 1} \quad (1)$$

where r is the rank, and n is the total number of candidates with the same molecular weight (this expression is equivalent to that presented in Kerber *et al.*¹⁸). Thus a rank of one maps

Figure 1: Structures of the molecules chosen for analysis grouped into chemical classes



to zero, and a rank of n maps to 1. The RRP is not defined for a class with only one member. This quantity constitutes the metric for direct comparison between compounds with different molecular formulae, as they clearly have different possible numbers of structural isomers and therefore differing numbers of entries in the database.

2.5 Spectrum normalization for statistical comparisons

It is standard to normalize the mass spectra such that the highest peak has an intensity with value 100. However, such a normalization makes direct comparison relative between error statistics such as root mean square deviation (RMSD) across different compounds meaningless. As such, these metrics (see Table 3) have been calculated with spectral intensities normalized such that the sum of the intensities is unity for both the spectra being compared. A byproduct of this normalization scheme is that the mean signed error between two spectra is always zero. Such a normalization scheme has been used previously, by Rasmussen and Isenhour²⁶ who demonstrated that this kind of probability distribution normalization (which they call ‘total ion current normalization’) does not lead to a significant effect when tested for the purposes of identifying a known mass spectrum.

2.6 Estimates of the errors in the calculated peak heights

Since the spectra are built up by counting independent fragmentation processes we assume that the number of counts in a particular peak of the mass spectrum has a Poisson distribution. Then the absolute error in a particular peak of with N_i counts is $\sqrt{N_i}$. If $N_{\text{total}} = \sum_i N_i$ is the total number of counts in the whole mass spectrum the absolute error in a particular probability-normalized mass spectrum with peak heights N_i/N_{total} is $\sqrt{N_i}/N_{\text{total}}$. It is important to keep in mind that the reproducibility error in a particular mass spectrum is about 5-10% in the peak heights, and in some cases even up to 30%¹².

3 Results

Table 1 shows the predicted RRP for the QCEIMS method for the 61 molecules considered in this study; rankings for CFM-EI (Table S0), along with predicted and experimental MSP for each compound (Figures S1-S122) are shown in the electronic supplementary material (ESI) and selected spectra are presented and discussed later.

3.1 The QCEIMS method

Overall success rate. We see from the table that using the QCEIMS method we are able to identify the correct structure from the NIST database (using the standard spectrum matching algorithm for compounds with the same chemical formula) in 24 of the 61 cases. The correct spectrum is ranked in the top two in 32 of the cases (greater than 50% success rate) and in the top three in 42 of the cases (69% success rate). Thus if the QCEIMS method were to be used to actually identify the structure of an unknown compound similar to those in the test set one might expect to get the correct result in the top three matches two-thirds of the

Table 1: Relative ranking position (RRP) of the matching mass spectra found by the QCEIMS program produced by the NIST MS Search program (version 2.0g). Ranking is relative to the total number isomers n_{isomers} . NM indicates that no match was found in the database.

No.	Name	Chemical Formula	RRP / n_{isomers}
Alcohols			
1	propan-1-ol	C ₃ H ₈ O	1/3
2	propan-2-ol	C ₃ H ₈ O	1/3
3	butan-1-ol	C ₄ H ₁₀ O	1/8
4	butan-2-ol	C ₄ H ₁₀ O	1/8
5	2-methylpropan-1-ol	C ₄ H ₁₀ O	1/8
6	2-methylpropan-2-ol	C ₄ H ₁₀ O	5/8
Aldehydes			
7	propanal	C ₃ H ₆ O	1/7
8	2-methylpropanal	C ₄ H ₈ O	3/21
9	butanal	C ₄ H ₈ O	3/21
Alkenes			
10	but-1-ene	C ₄ H ₈	4/6
11	but-2-ene	C ₄ H ₈	1/6
12	pent-1-ene	C ₅ H ₁₀	3/13
13	pent-2-ene	C ₅ H ₁₀	3/13
14	hex-1-ene	C ₆ H ₁₂	6/30
15	hex-2-ene	C ₆ H ₁₂	1/30
16	hex-3-ene	C ₆ H ₁₂	9/30
Amines			
17	propan-1-amine	C ₃ H ₉ N	1/4
18	propan-2-amine	C ₃ H ₉ N	1/4
19	butan-1-amine	C ₄ H ₁₁ N	1/8
20	butan-2-amine	C ₄ H ₁₁ N	1/8
21	2-methylpropan-1-amine	C ₄ H ₁₁ N	2/8
22	2-methylpropan-2-amine	C ₄ H ₁₁ N	1/8
Benzenes			
23	benzene	C ₆ H ₆	1/5
24	toluene	C ₇ H ₈	3/13
25	ethylbenzene	C ₈ H ₁₀	8/23
26	propylbenzene	C ₉ H ₁₂	2/42
27	cumene	C ₉ H ₁₂	16/42
28	butylbenzene	C ₁₀ H ₁₄	3/74
29	butan-2-ylbenzene	C ₁₀ H ₁₄	6/74
30	2-methylpropylbenzene	C ₁₀ H ₁₄	1/74

No.	Name	Chemical Formula	RRP / n_{isomers}
Carboxylic acids			
31	propanoic acid	C ₃ H ₆ O ₂	1/9
32	butanoic acid	C ₄ H ₈ O ₂	7/27
33	2-methylpropanoic acid	C ₄ H ₈ O ₂	3/27
34	2-methylbutanoic acid	C ₅ H ₁₀ O ₂	3/47
35	3-methylbutanoic acid	C ₅ H ₁₀ O ₂	5/47
Esters			
36	methyl formate	C ₂ H ₄ O ₂	1/3
37	methyl acetate	C ₃ H ₆ O ₂	2/9
38	ethyl acetate	C ₄ H ₈ O ₂	1/27
39	methyl propanoate	C ₄ H ₈ O ₂	1/ 27
40	propyl acetate	C ₅ H ₁₀ O ₂	NM/47
41	propan-2-yl acetate	C ₅ H ₁₀ O ₂	4/47
42	methyl 2-methylpropanoate	C ₅ H ₁₀ O ₂	1/47
43	ethyl propanoate	C ₅ H ₁₀ O ₂	1/47
44	methyl butanoate	C ₅ H ₁₀ O ₂	5/47
45	propyl propanoate	C ₆ H ₁₂ O ₂	1/80
46	propan-2-yl propanoate	C ₆ H ₁₂ O ₂	3/80
47	ethyl butanoate	C ₆ H ₁₂ O ₂	4/80
48	methyl 2-methylbutanoate	C ₆ H ₁₂ O ₂	2/80
49	methyl 3-methylbutanoate	C ₆ H ₁₂ O ₂	7/80
50	propyl butanoate	C ₇ H ₁₄ O ₂	2/83
51	propan-2-yl butanoate	C ₇ H ₁₄ O ₂	NM/83
Ketones			
52	acetone	C ₃ H ₆ O	1/7
53	butan-2-one	C ₄ H ₈ O	1/21
Phenols			
54	phenol	C ₆ H ₆ O	2/4
55	4-methylphenol	C ₇ H ₈ O	2/17
56	4-ethylphenol	C ₈ H ₁₀ O	16/43
57	4-propylphenol	C ₉ H ₁₂ O	NM/112
58	4-propan-2-ylphenol	C ₉ H ₁₂ O	16/112
59	4-butylphenol	C ₁₀ H ₁₄ O	NM/163
60	4-butan-2-ylphenol	C ₁₀ H ₁₄ O	3/163
61	4-(2-methylpropyl)phenol	C ₁₀ H ₁₄ O	2/163

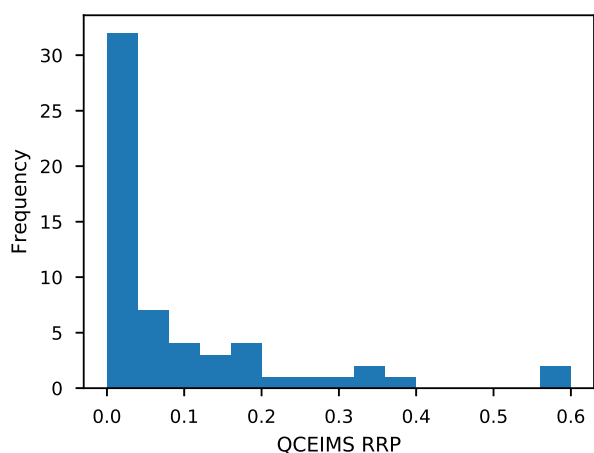


Figure 2: The number of molecules predicted correctly at a given rank for the QCEIMS method. There were four molecules which were not matched to any molecule in the NIST database.

time. However one must qualify this statement in at least two respects. First, the test set is rather small; and even with this small set there were four cases providing unacceptable matches (shown as NM in the table). Also, as we previously remarked, comparing actual ranks is problematic because different chemical classes have different numbers of isomers. Therefore, in Figure 2 we present a histogram of the number of molecules given a particular rank. Our conclusions based on actual rank are essentially the same as those based on RRP.

Dependence of the quality of the rankings on the chemical class of the compound. Some of the classes have too few candidates to make any clear statements, but we may still make the following observations:

- In the case of the alcohols, all the primary alcohols are correctly ranked, but the tertiary alcohol **6** was not correctly predicted (rank 5!).
- Some of the alkenes are very poorly ranked, but there is no clear correlation of the results with the position of the double bond.
- Rankings for the amines were remarkably good, with all with the exception of **21** being matched correctly.
- For the substituted benzenes, poorer results are obtained for candidates with multiple alkyl substituents. This trend is exacerbated for the phenols where two of the compounds **57** and **59** were not matched at all.
- Substituted esters are generally well ranked, but tend to be more poorly predicted when (like the substituted benzene) one of the groups becomes more branched.

Dependence of the quality of the results on the mass of the molecules. There is a weak trend that increasing size of molecule produces worse matches perhaps best exemplified by the esters (Pearson's correlation coefficients $r = 0.40$). It should be kept in mind that

Table 2: Mean Relative ranking position (RRP) for the QCEIMS and CFM-EI methods for different classes of compounds, with the number of compounds in each class. Overall mean RRP's are given in the final line. Only molecules in which both methods matched are compared.

Class	Mean RRP		No. of Compounds
	QCEIMS	CFM-EI	
Alcohols	0.11	0.21	5
Aldehydes	0.07	0.02	3
Alkenes	0.20	0.30	7
Amines	0.02	0.08	6
Benzenes	0.12	0.12	8
Carboxylic acids	0.10	0.14	3
Esters	0.03	0.04	14
Ketones	0.00	0.00	2
Phenols	0.18	0.14	5
Overall	0.09	0.12	53

matches were only considered in the database between molecules of the same molecular formula. It should also be noted that the size of the molecules is very small, and one might expect a strong dependence with molecular weight.

Mean RRP's for each chemical class. Table 2 summarizes some of the discussion above in a more quantitative way, presenting the mean RRP's for each chemical class. Although the results for ketones and aldehydes seem very good, there are too few systems to consider the mean RRP's meaningful for these two classes of compounds. However, we see that the esters and amines are very well predicted, with the means RRP's for carboxylic acids, substituted benzenes and phenols all being quite similar and low. The alcohols seem not well predicted but are in fact skewed by one bad result. The alkenes are ranked the worst. It also deserves to be noted that in the cases where there are a large number of isomers the matching algorithm ranks the QCEIMS spectra very near the top e.g. for **30** we have 1/74 and for **61** we have 2/163.

Comparison of spectra with experiment. Table 3 gives the mean root-mean-square deviation (RMSD) and mean absolute error (MAD) in the probability normalized spectra. We observe that the spectra which agree the worst on the RMSDs (which highlight outlier differences) are in order: the alcohols (with one outlier already noted in the previous section), aldehydes, carboxylic acids, esters and alkenes. With regard to MADs, the worst classes are again the alcohols, alkenes, and ketones. In the ESI we present the mass spectra produced for all the compounds. Most informative are the differences between the predicted and observed spectra: peaks are coloured green if they were present in both spectra, yellow if present only in the experiment, and red if present only in the QCEIMS model spectra. Positive differences indicate that the calculated spectra overestimate the peak intensities, and error bars on the QCEIMS spectra are presented. We see that the errors due to counting statistics are very small, the maximum always less than 8% and usually much smaller, and

Table 3: Mean root mean square deviation (RMSD) and mean absolute deviation (MAD) for the QCEIMS and CFM-EI methods for different classes of compounds, . Overall mean RMSDs and MADs are given in the final line.

Class	Mean RMSD		Mean MAD	
	QCEIMS	CFM-EI	QCEIMS	CFM-EI
Alcohols	0.09	0.05	0.03	0.03
Aldehydes	0.08	0.03	0.03	0.01
Alkenes	0.05	0.05	0.02	0.03
Amines	0.04	0.03	0.02	0.02
Benzenes	0.04	0.04	0.01	0.02
Carboxylic acids	0.06	0.04	0.02	0.02
Esters	0.06	0.04	0.02	0.02
Ketones	0.04	0.05	0.02	0.03
Phenols	0.04	0.04	0.01	0.02
Overall	0.05	0.04	0.02	0.02

then only appearing on the peaks which have the largest counts. Therefore the difference in intensity between the experimental and QCEIMS peaks for the larger peaks to be attributed to counting statistics. Also, the differences in the larger peaks far exceeds the sum of the counting statistics error and any reproducibility error associated with the EIMS experiment. A typical example showing error bars is presented in Figure 4, ethylbenzene, discussed in more detail later.

3.2 The CFM-EI method

Overall success rate. The CFM-EI has a very similar first match rate compared to the QCEIMS method, with 24 matching in first position, 35 in the top two positions.

Dependence of the quality of results on the chemical class. The full set of rankings for the CFM-EI method is shown in Table S1 of the ESI. The following observations can be made.

- The alcohols are well ranked except for one outlier (**3**) which is placed in the 5th position (this is not the same compound ranked 5th in the QCEIMS method).
- The alkenes are rather poorly ranked: both **12** and **16** are ranked at position 10.
- The amines are remarkably well ranked: all are ranked in the first position except **18** and **22**, which are ranked in second place.
- For the substituted benzenes and phenols there are some very poor rankings, which are a bit worse than for the QCEIMS method. For example, **28** is in position 20, **59** in position 35, and **61** had no match.

Mean RRP for each chemical class. Table 2 shows that, as for the QCEIMS method, the alkenes are worst predicted, followed by the alcohols. The results for the alcohols and

alkenes are also significantly worse than in the QCEIMS method (RRPs respectively 0.21 and 0.30 versus 0.11 and 0.20). The next worst grouping was among the benzenes, phenols and carboxylic acids, again generally worse ranked than compared to the QCEIMS method. Based on the individual class results, it is therefore not surprising that the overall ranking from CFM-EI is worse than from the QCEIMS method (mean RRP's respectively 0.12 versus 0.09).

Dependence of the quality of the results on the mass of the molecules. There is a weak trend that results are worse predicted as the mass of the molecule increases ($r = 0.47$) and this trend is a bit stronger than for the QCEIMS method.

Comparison of QCEIMS and CFM-EI. Table 3 shows rather uniformly low RMSDs and MADs for the predicted and observed spectra: in both cases the MADs were less than 3%. However, based on the RMSDs the CFM-EI seems to produce a slightly more accurate i.e. the spectrum shows fewer outlying peaks relative to experiment. Nevertheless, a qualitatively better MS_p does not translate into a superior ranking of the MS_p (see discussion on the RRP's).

4 Discussion

4.1 Detailed analysis of the predicted spectra

Alcohols. QCEIMS underestimates the abundance of most small peaks it predicts. As the peak intensities are relative, this indicates that QCEIMS is systematically overestimating the abundance of the primary fragmentation relative to the others.

On the other hand, CFM-EI seems to be underestimating the primary fragmentation with respect to others.

Further, there exists a tendency for CFM-EI to overestimate the abundance of the molecular ion. This over-abundance is reflected in the poor results in predicting both propan-2-ol and butan-1-ol (see Figure 3) by CFM-EI.

It is noteworthy that both methods generally underestimate the loss of water ($M - 18$) from primary alcohols, which is reflected in the MS of butan-1-ol with the low $m/z = 56$ peak.

Both methods predict the existence of few peaks not in the experimental spectra, (shown in red in Fig 3). CFM-EI particularly in the cases of 2-methylpropan-1-ol and, to a lesser extent, butan-1-ol predicts significant peaks non-existent in the experimental MS_p.

Aldehydes and ketones. Importantly, QCEIMS clearly demonstrates the capacity to 'discover' McLafferty rearrangements. An illustrative example, the case of butanal, shows in the $m/z = 44$ ($M - 28$) peak, from loss of ethene via a McLafferty rearrangement. While the relative abundance of this peak was underestimated by QCEIMS, its existence is vitally important. In the case of 2-methylpropanal QCEIMS seems to be missing or underestimating characteristic $m/z = 27$ and $m/z = 41$ peaks.

The predictions by CFM in the case of 2-methylpropanal is closer to experiment, but still systematically underestimated. Again, as was the case in the alcohols, CFM-EI seems to overestimate the abundance of the molecular ion in some cases (particularly 2-propanone).

Neither MS prediction technique displays any egregious false peaks in the aldehydes or

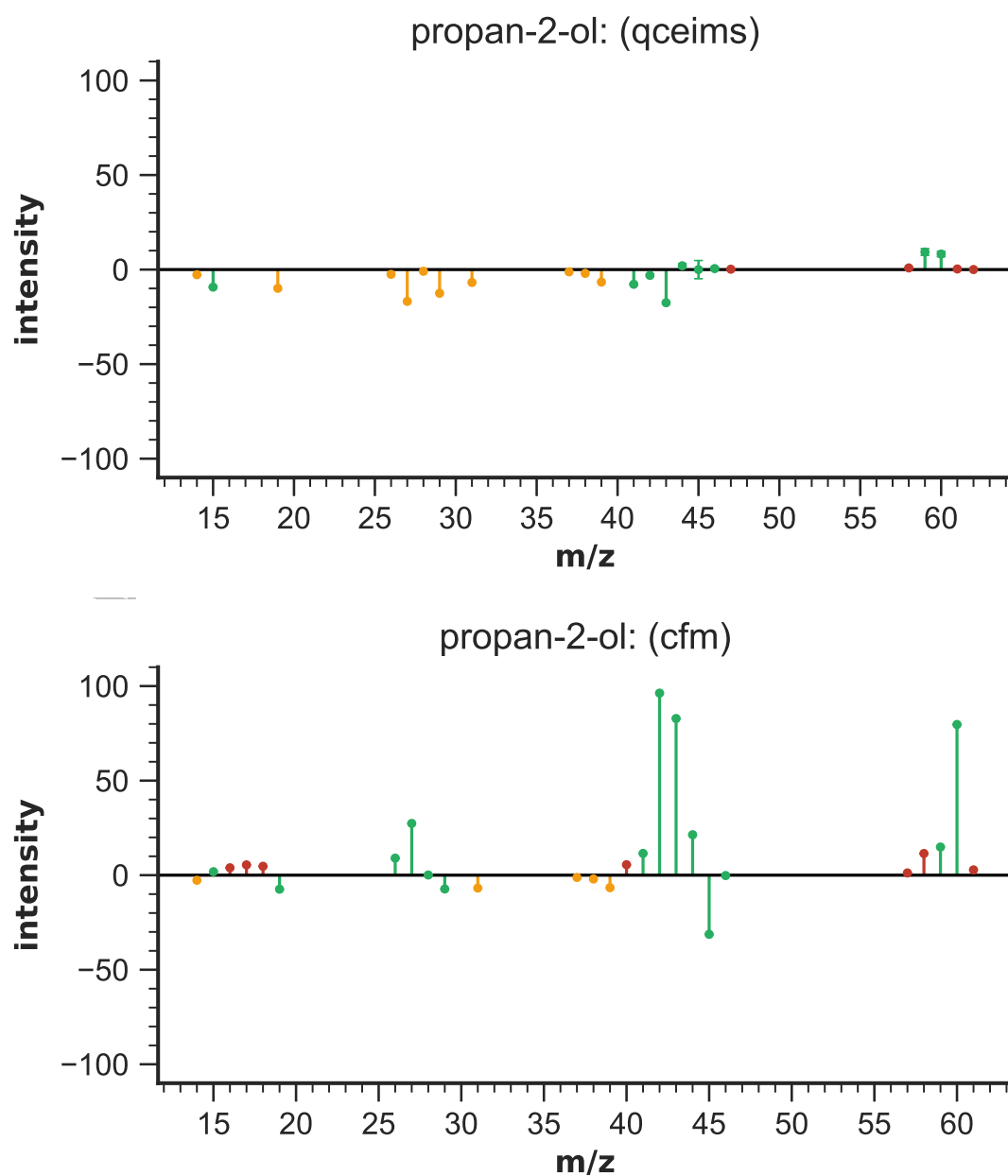


Figure 3: Difference between predicted (QCEIMS/CFM-EI) and experimental (NIST) spectra. Green bars indicate the presence of a correct peak, yellow indicates a missing peak, and red indicates a peak which is incorrect i.e. should not have been predicted.

ketones, they simply tend to underestimate the relative occurrence of certain rearrangements or cleavages.

In the case of the aldehydes, both methods seem to consistently fail to match experimental spectra, and this is reflected in the rankings (2s/3s for both butanal and 2-methylpropanal). Contrast this with the case of the two ketones, where the predictions are almost perfectly on the mark. This might be associated with the decreased importance of α -cleavages in the case of the aldehydes as compared to the ketones, strong conclusions require more studies.

Esters. It is evident for both techniques that McLafferty type rearrangements are absent or underestimated for all relevant esters – particularly in methyl butanoate, where $m/z = 74$ is virtually absent in the predicted spectrum. This is not to say that these methods are not capable of predicting rearrangement fragmentations in-principle, as may be seen upon the examination of the the prominent $m/z = 88$ in ethyl butanoate.

Amines. QCEIMS underestimates or fails to predict some portions of the MSp, notably in the $25 < m/z < 30$ region of propan-2-amine and 2-methylpropan-2-amine.

Intriguingly, despite the QCEIMS MS for propan-2-amine overall having many missing peaks and generally incorrect intensities (i.e. looking qualitatively worse) than that produced by CFM-EI, the spectra proves to be a better match. In this instance, CFM clearly overestimates (once again) the abundance of the molecular ion, which may have resulted in this poor match against experiment. The cases of 2-methylpropan-1- and 2-methylpropan-2-amine proved more difficult for both prediction techniques, though the spectra were very close to experiment. Both cases yield significantly different spectra, something reflected in both predictive techniques.

A significant error in both methods appears to be the overestimated abundance of loss of a proton ($M - 1$), highlighted particularly in the $m/z = 72$ peak of the CFM-EI MSp for 2-methylpropan-2-amine, but also in the $m/z = 72$ and $m/z = 58$ peaks for the QCEIMS spectra for butan-1-amine and propan-2-amine, respectively.

Alkenes. The case of alkenes, in particular those chosen here, requires precise predictions in order to achieve the closest match with experiment. The differences between structural isomers their MSp tend to be independent of the double bond position (unless highly substituted or conjugated)²³. This, coupled with the tendencies of isomerization through migration of the double bond in these low molecular weight, straight chain alkenes makes discrimination between candidates a much more difficult task.

In general, both techniques overestimate the abundance of the molecular ion, further exaggerated in the hexene series where the relative abundance of the molecular ion is much lower than in the pent- and butenes. The regions $25 < m/z < 30$ and $40 < m/z < 45$, and the $m/z = 55$ peak for pent- and hexene are the discriminating regions of these spectra. It is notable that in these regions, both QCEIMS and CFM come close to the experimental spectra in their predictions. The poor rankings (as seen in Table 2) over the alkene set are a contrast with the relatively good error statistics (see Table 3). This is consistent with the rather subtle differences in MSps where the only difference in the molecule is the double bond position. It is not the case, then that the two techniques are failing to accurately predict the spectra of alkenes, merely that their predictions are not accurate enough for the purposes of matching within sets of structural isomers.

Carboxylic Acids. From the results presented in Table 2, it would appear that performance for the majority of carboxylic acids was far from satisfactory. Both techniques showed failures

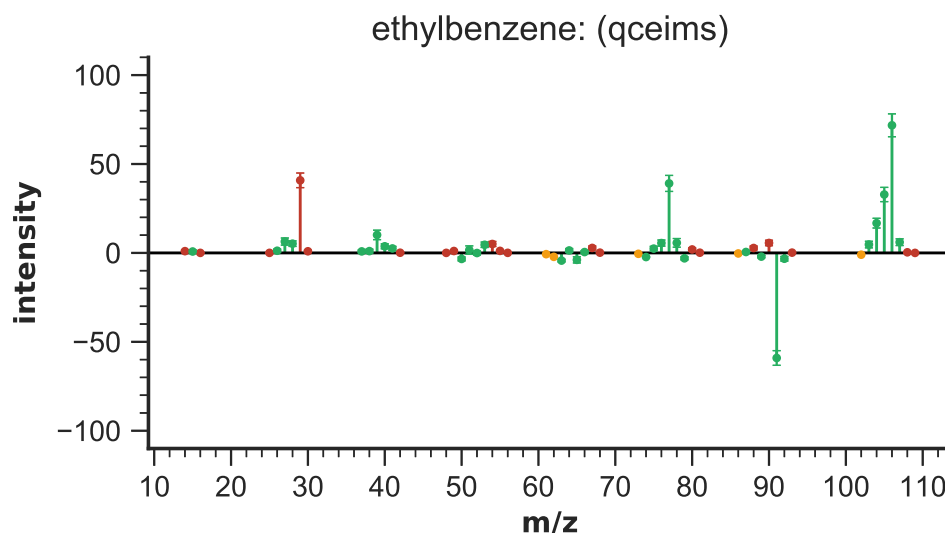


Figure 4: Difference the between predicted (QCEIMS) and experimental (NIST) mass spectrum for ethylbenzene.

in their spectral predictions, but it would seem that in this case QCEIMS outperforms CFM in producing an EIMS matching more closely to experiment.

For longer chain carboxylic acids, QCEIMS seems to severely underestimate crucial McLafferty rearrangements, giving rise to inadequate prediction of mass spectra of these compounds. An illustrative example is the virtual absence of $m/z = 74$, the base peak of 2-methylbutanoic acid. These issues seem to be even more significant for CFM-EI with several carboxylic acids given rise to very inaccurate spectra.

Benzenes. The only significant discrepancy from experiment for QCEIMS is shown in the case of (1-methylethyl)benzene, where this method seems to make a charged fragment of the isopropyl moiety, which does not occur in the real spectrum.

As many alkylbenzenes show similarities in the mass spectra, errors in approximations of abundance of fragments can mean that the rank is relatively low although the correct fragments are all predicted. This is evident, for example, in ethylbenzene (see Figure 4). A general overestimation of the loss of a methyl moiety ($M - 15$) is seen with both methods in most alkyl benzenes.

Phenols. Errors were present in the cases of propyl- and butylphenol, where the matches are very low. The typical aromatic fragmentations of $m/z = 77$ (phenyl cation) and $m/z = 91$ (propylium ion) are often lost in both QCEIMS and CFM-EI leading to problems in matching the MSp.

5 Conclusion

We have demonstrated in this paper that the *ab initio* quantum chemical electron impact mass spectrum (QCEIMS) method of Grimme and coworkers produces simulated mass spectra which are of a sufficient quality to match the correct spectra in the first or second rank

from a field of isomers (using the standard mass spectral search algorithms in the NIST-11 database using standard NIST software) in greater than 50% of the cases for a test set of 61 molecules. Remarkably, the method performs slightly better than the recently published competitive fragmentation model for electron impact (CFM-EI) which presumably has in its training set some of the spectra actually used in the test set. From this we conclude that the QCEIMS method will be very useful for those who seek to identify the structure of unknown chemical structures from the mass spectrum of small compounds, for example, those who seek to identify semiochemicals, where one often has to synthesize several putative candidates to confirm a particular structure. The QCEIMS method will be particularly useful in cases where the new compound is not well represented in the search library (a fact which unfortunately cannot be established before the structure is known).

We have also established that the counting statistics errors on the peaks produced in QCEIMS are much smaller than the experimental reproducibility error in the mass spectrum; and that the spectra are often reproduced better than the corresponding rankings relative to isomers. This demonstrates that the QCEIMS method achieves its successful matches not because of an accurate match in the peak heights, but rather that the peak heights and the distribution of the peak positions as a function of mass-to-charge ratio is *sufficiently different* for different isomers to achieve a high ranking.

As regards the performance of QCEIMS for different chemical classes, we find that the alcohols have the best matching spectra, with a mean MAD of about 0.03. The normalized spectra are therefore reproduced rather well. On the other hand, the worst ranked spectra were the alkenes and substituted phenols with mean relative ranking positions above 0.18. It should be noted, though, that even in these cases the correct candidate is placed in the top fifth of isomers.

The problem with QCEIMS, recognized by the authors, is that it is a time consuming calculation to perform, often taking days or weeks for a single spectrum; by comparison the CFM-EI method can be performed in the browser. We do not think this is necessarily a problem for those working to identify semiochemicals, where in a worst case situation it takes years to identify the structure of an unknown: in this case it is more useful for the information to be correct rather than timely. Still, the length of time required for QCEIMS means that it is best used sparingly after other methods have been used to narrow the field of candidates.

One of the great difficulties with EIMS prediction is that it is not only about simple fragmentations, but frequently involves rearrangements within the molecule.

Both methods tended to suffer from the same problems: underestimation or failure to predict specific rearrangements (e.g. McLafferty rearrangements), and overestimation of the molecular ion. These failures are especially salient given the radically different manner in which both techniques function. Perhaps this indicates that reassessment of some of the fundamental abstractions (namely the notion that it is simply a matter of predicting the fragmentation) involved need to be re-evaluated for general EIMS prediction.

There remains work to be done – ideally the identification protocol would start from an observed MS and predict the molecular geometry, *not* the current prediction of MS from a geometry. While numerous existing tools function quite well at predicting the molecular formula from the observed spectra (see Scheubert et al.²⁸), the ‘brute force’ approach of generating all possible isomers is simply infeasible for even slightly large molecules, due to

the exponential relationship between M and the number of possible isomers. Clearly, the goal of the DENDRAL project remains yet to be realized.

5.1 Acknowledgements

We gratefully acknowledge financial support from the Australian Research Council (grant DE160101313) to BB, and the financial support of the Danish National Research Foundation (Center for Materials Crystallography, DNRF-93) to PRS. This research was undertaken with the assistance of computational resources from the Pople high-performance computing cluster of the Faculty of Science at the University of Western Australia.

Accepted Article

References

- [1] Allen F, Greiner R, Wishart D. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics* 2015;.
- [2] Allen F, Pon A, Greiner R, Wishart D. Computational Prediction of Electron Ionization Mass Spectra to Assist in GC/MS Compound Identification. *Analytical Chemistry* 2016 Jul;88(15):7689–7697.
- [3] Asgeirsson V, Bauer CA, Grimme S. Quantum chemical calculation of electron ionization mass spectra for general organic and inorganic molecules. *Chemical Science* 2017;52(5):6306.
- [4] Bauer CA, Grimme S. Elucidation of Electron Ionization Induced Fragmentations of Adenine by Semiempirical and Density Functional Molecular Dynamics. *The Journal of Physical Chemistry A* 2014 Dec;118(49):11479–11484.
- [5] Bauer CA, Grimme S. How to Compute Electron Ionization Mass Spectra from First Principles. *The Journal of Physical Chemistry A* 2016 May;120(21):3755–3766.
- [6] Bohman B, Flematti GR, Barrow RA. Identification of hydroxymethylpyrazines using mass spectrometry. *Journal of Mass Spectrometry* 2015 Aug;50(8):987–993.
- [7] Bohman B, Phillips RD, Flematti GR, Barrow RA, Peakall R. The chemical web of a spider orchid - Sulfurous semiochemicals seduce male wasp pollinator. *Angewandte Chemie International Edition in English* 2017 May;.
- [8] Bohman B, Phillips RD, Menz MHM, Berntsson BW, Flematti GR, Barrow RA, et al. Discovery of pyrazines as pollinator sex pheromones and orchid semiochemicals: implications for the evolution of sexual deception. *New Phytologist* 2014 Aug;203(3):939–952.
- [9] Buchanan BG, Lederberg J, Feigenbaum E, The Heuristic DENDRAL program for explaining empirical data; 1971.
- [10] Dobson CM. Chemical space and biology. *Nature* 2004 Dec;432(7019):824–828.
- [11] Gasteiger J, Hanebeck W, Schulz KP. Prediction of mass spectra from structural information. *Journal of Chemical Information and Modeling* 1992 Jul;32(4):264–271.
- [12] Grimme S. Towards First Principles Calculation of Electron Impact Mass Spectra of Molecules. *Angewandte Chemie International Edition* 2013 Apr;52(24):6306–6312.
- [13] Grimme S, Bauer CA. First principles calculation of electron ionization mass spectra for selected organic drug molecules. *Org Biomol Chem* 2014 Sep;12(43):8737–8744.
- [14] Hanwell MD, Curtis DE, Lonie DC, Vandermeersch T, Zurek E, Hutchison GR. Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *Journal of Cheminformatics* 2012 Aug;4(1):17.

- [15] Hertz HS, Hites RA, Biemann K. Identification of mass spectra by computer-searching a file of known spectra. *Analytical Chemistry*; 1971.
- [16] Hunter JD. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* 2007;9(3):90–95.
- [17] Jalali-Heravi M, Fatemi MH. Simulation of mass spectra of noncyclic alkanes and alkenes using artificial neural network. *Analytica Chimica Acta* 2000 Jun;415(1-2):95–103.
- [18] Kerber A, Meringer M, Rücker C. CASE via MS: Ranking Structure Candidates by Mass Spectra. *Croatica Chemica Acta* 2006 Nov;79(3):449–464.
- [19] Kind T, Fiehn O. Advances in structure elucidation of small molecules using mass spectrometry. *Bioanalytical reviews* 2010 Dec;2(1-4):23–60.
- [20] Mayer I, Knapp-Mohammady M, Suhai S. Bond orders and energy components in polymers. *Chemical Physics Letters* 2004;389(1-3):34–38.
- [21] McKinney W. Data Structures for Statistical Computing in Python. In: van der Walt Se, Millman J, editors. *Proceedings of the 9th Python in Science Conference*; 2010. p. 51–56.
- [22] McLafferty FW, Stauffer DA, Loh SY, Wesdemiotis C. Unknown identification using reference mass spectra. Quality evaluation of databases. *Journal of the American Society for Mass Spectrometry* 1999 Dec;10(12):1229–1240.
- [23] McLafferty FW, Turecek F. *Interpretation Of Mass Spectra*. Organic chemistry series, University Science Books; 1993.
- [24] Mun IK, McLafferty FW. Computer Methods of Molecular Structure Elucidation from Unknown Mass Spectra. In: *Supercomputers in Chemistry*; p. 117–124.
- [25] N N Bogolubov J. *Introduction to Quantum Statistical Mechanics*. World Scientific; 2010.
- [26] Rasmussen GT, Isenhour TL. The evaluation of mass spectral search algorithms. *Journal of Chemical Information ...* 1979;.
- [27] Russell SJ, Norvig P. *Artificial Intelligence. A Modern Approach*, Prentice Hall; 2010.
- [28] Scheubert K, Hufsky F, Böcker S. Computational mass spectrometry for small molecules. *Journal of Cheminformatics* 2013;5(1):12.
- [29] Stein SE. An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *Journal of the American Society for Mass Spectrometry* 1999 Aug;10(8):770–781.
- [30] Stein SE. Estimating probabilities of correct identification from results of mass spectral library searches. *Journal of the American Society for Mass Spectrometry* 1994 Apr;5(4):316–323.

- [31] Stein SE, Scott DR. Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry* 1994 Sep;5(9):859–866.
- [32] Thiel W, MNDO99, version 7.0; 2004.
- [33] Valkenborg D, Mertens I, Lemièrre F, Witters E, Burzykowski T. The isotopic distribution conundrum. *Mass Spectrometry Reviews* 2012 Jan;31(1):96–109.
- [34] van der Walt S, Colbert SC, Varoquaux G. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering* 2011;13(2):22–30.
- [35] Wan KX, Vidavsky I, Gross ML. Comparing similar spectra: from similarity index to spectral contrast angle. *Journal of the American Society for Mass Spectrometry* 2002 Jan;13(1):85–88.
- [36] Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and testing of a general amber force field. *Journal of computational chemistry* 2004 Jul;25(9):1157–1174.
- [37] Weber W, Thiel W. Orthogonalization corrections for semiempirical methods. *Theoretical Chemistry Accounts* 2000;103(6):495–506.
- [38] Zupan J, Gasteiger J. Neural networks: A new method for solving chemical problems or just a passing phase? *Analytica Chimica Acta* 1991 Jul;248(1):1–30.

Computational prediction of electronic impact mass spectra - supplementary information

Peter R. Spackman,* Björn Bohman, Amir Karton, and Dylan Jayatilaka

*School of Chemistry and Biochemistry, University of Western Australia, Crawley, 6009,
WA, Australia*

E-mail: spackp01@student.uwa.edu.au

Contents

1	Tables	4
2	Plots of EIMS and differences in EIMS	7
2.1	Alcohols	8
2.1.1	1: propan-1-ol	8
2.1.2	2: propan-2-ol	10
2.1.3	3: butan-1-ol	12
2.1.4	4: butan-2-ol	14
2.1.5	5: 2-methylpropan-1-ol	16
2.1.6	6: 2-methylpropan-2-ol	18
2.2	Aldehydes	20
2.2.1	7: propanal	20
2.2.2	8: 2-methylpropanal	22
2.2.3	9: butanal	24

2.3	Alkenes	26
2.3.1	10: but-1-ene	26
2.3.2	11: but-2-ene	28
2.3.3	12: pent-1-ene	30
2.3.4	13: pent-2-ene	32
2.3.5	14: hex-1-ene	34
2.3.6	15: hex-2-ene	36
2.3.7	16: hex-3-ene	38
2.4	Amines	40
2.4.1	17: propan-1-amine	40
2.4.2	18: propan-2-amine	42
2.4.3	19: butan-1-amine	44
2.4.4	20: butan-2-amine	46
2.4.5	21: 2-methylpropan-1-amine	48
2.4.6	22: 2-methylpropan-2-amine	50
2.5	Benzenes	52
2.5.1	23: benzene	52
2.5.2	24: toluene	54
2.5.3	25: ethylbenzene	56
2.5.4	26: propylbenzene	58
2.5.5	27: cumene	60
2.5.6	28: butylbenzene	62
2.5.7	29: butan-2-ylbenzene	64
2.5.8	30: 2-methylpropylbenzene	66
2.6	Carboxylic acids	68
2.6.1	31: propanoic acid	68
2.6.2	32: butanoic acid	71

2.6.3	33: 2-methylpropanoic acid	74
2.6.4	34: 2-methylbutanoic acid	77
2.6.5	35: 3-methylbutanoic acid	80
2.7	Esters	83
2.7.1	36: methyl formate	83
2.7.2	37: methyl acetate	86
2.7.3	38: ethyl acetate	89
2.7.4	39: methyl propanoate	92
2.7.5	40: propyl acetate	95
2.7.6	41: propan-2-yl acetate	98
2.7.7	42: methyl 2-methylpropanoate	101
2.7.8	43: ethyl propanoate	104
2.7.9	44: methyl butanoate	107
2.7.10	45: propyl propanoate	110
2.7.11	46: propan-2-yl propanoate	113
2.7.12	47: ethyl butanoate	116
2.7.13	48: methyl 2-methylbutanoate	119
2.7.14	49: methyl 3-methylbutanoate	122
2.7.15	50: propyl butanoate	125
2.7.16	51: propan-2-yl butanoate	128
2.8	Ketones	131
2.8.1	52: acetone	131
2.8.2	53: butan-2-one	133
2.9	Phenols	135
2.9.1	54: phenol	135
2.9.2	55: 4-methylphenol	137
2.9.3	56: 4-ethylphenol	139

2.9.4 57: 4-propylphenol 141

2.9.5 58: 4-propan-2-ylphenol 143

2.9.6 59: 4-butylphenol 145

2.9.7 60: 4-butan-2-ylphenol 147

2.9.8 61: 4-(2-methylpropyl)phenol 149

1 Tables

Table S1: Relative ranking position (RRP) of the matching mass spectra found by the QCEIMS and CFM-EI programs produced by the NIST MS Search program (version 2.0g). Ranking is relative to the total number isomers n_{isomers} . NM indicates that the actual compound in to match by the program (i.e. it was not presented with a probability).

No.	Name	Chemical Formula	RRP QCEIMS	RRP CFM-EI	n_{isomers}
Alcohols					
1	propan-1-ol	C ₃ H ₈ O	1	2	3
2	propan-2-ol	C ₃ H ₈ O	1	NM	3
3	butan-1-ol	C ₄ H ₁₀ O	1	5	8
4	butan-2-ol	C ₄ H ₁₀ O	1	1	8
5	2-methylpropan-1-ol	C ₄ H ₁₀ O	1	1	8
6	2-methylpropan-2-ol	C ₄ H ₁₀ O	5	1	8
Aldehydes					
7	propanal	C ₃ H ₆ O	1	1	7
8	2-methylpropanal	C ₄ H ₈ O	3	1	21
9	butanal	C ₄ H ₈ O	3	2	21
Alkenes					
10	but-1-ene	C ₄ H ₈	4	3	6
11	but-2-ene	C ₄ H ₈	1	2	6
12	pent-1-ene	C ₅ H ₁₀	3	10	13
13	pent-2-ene	C ₅ H ₁₀	3	5	13
14	hex-1-ene	C ₆ H ₁₂	6	3	30
15	hex-2-ene	C ₆ H ₁₂	1	1	30
16	hex-3-ene	C ₆ H ₁₂	9	10	30
Amines					
17	propan-1-amine	C ₃ H ₉ N	1	1	4
18	propan-2-amine	C ₃ H ₉ N	1	2	4
19	butan-1-amine	C ₄ H ₁₁ N	1	1	8
20	butan-2-amine	C ₄ H ₁₁ N	1	1	8
21	2-methylpropan-1-amine	C ₄ H ₁₁ N	2	2	8
22	2-methylpropan-2-amine	C ₄ H ₁₁ N	1	1	8
Benzenes					
23	benzene	C ₆ H ₆	1	1	5
24	toluene	C ₇ H ₈	3	3	13
25	ethylbenzene	C ₈ H ₁₀	8	1	23
26	propylbenzene	C ₉ H ₁₂	2	1	42
27	cumene	C ₉ H ₁₂	16	16	42
28	butylbenzene	C ₁₀ H ₁₄	3	20	74
29	butan-2-ylbenzene	C ₁₀ H ₁₄	6	5	74
30	2-methylpropylbenzene	C ₁₀ H ₁₄	1	12	74
Carboxylic acids					
31	propanoic acid	C ₃ H ₆ O ₂	1	1	9
32	butanoic acid	C ₄ H ₈ O ₂	7	11	27
33	2-methylpropanoic acid	C ₄ H ₈ O ₂	3	2	27
34	2-methylbutanoic acid	C ₅ H ₁₀ O ₂	3	NM	47
35	3-methylbutanoic acid	C ₅ H ₁₀ O ₂	5	NM	47
Esters					
36	methyl formate	C ₂ H ₄ O ₂	1	1	3
37	methyl acetate	C ₃ H ₆ O ₂	2	1	9
38	ethyl acetate	C ₄ H ₈ O ₂	1	7	27
39	methyl propanoate	C ₄ H ₈ O ₂	1	1	27
40	propyl acetate	C ₅ H ₁₀ O ₂	NM	2	47
41	propan-2-yl acetate	C ₅ H ₁₀ O ₂	4	2	47
42	methyl 2-methylpropanoate	C ₅ H ₁₀ O ₂	1	1	47
43	ethyl propanoate	C ₅ H ₁₀ O ₂	1	1	47
44	methyl butanoate	C ₅ H ₁₀ O ₂	5	3	47
45	propyl propanoate	C ₆ H ₁₂ O ₂	1	13	80
46	propan-2-yl propanoate	C ₆ H ₁₂ O ₂	3	7	80
47	ethyl butanoate	C ₆ H ₁₂ O ₂	4	1	80
48	methyl 2-methylbutanoate	C ₆ H ₁₂ O ₂	2	1	80
49	methyl 3-methylbutanoate	C ₆ H ₁₂ O ₂	7	3	80
50	propyl butanoate	C ₇ H ₁₄ O ₂	2	1	83
51	propan-2-yl butanoate	C ₇ H ₁₄ O ₂	NM	3	83
Ketones					
52	acetone	C ₃ H ₆ O	1	1	7
53	butan-2-one	C ₄ H ₈ O	1	1	21
Phenols					
54	phenol	C ₆ H ₆ O	2	2	4
55	4-methylphenol	C ₇ H ₈ O	2	4	17
56	4-ethylphenol	C ₈ H ₁₀ O	16	2	43
57	4-propylphenol	C ₉ H ₁₂ O	NM	2	112
58	4-propan-2-ylphenol	C ₉ H ₁₂ O	16	16	112
59	4-butylphenol	C ₁₀ H ₁₄ O	NM	35	163
60	4-butan-2-ylphenol	C ₁₀ H ₁₄ O	3	3	163
61	4-(2-methylpropyl)phenol	C ₁₀ H ₁₄ O	2	NM	163

Table S2: Total computation time in CPU hours for each QCEIMS prediction ordered by computation time

Name	Chemical Formula	n_{traj}	Time (CPU hours)	Name	Chemical Formula	n_{traj}	Time (CPU hours)
methyl formate	C ₂ H ₄ O ₂	200	1.4	methyl butanoate	C ₅ H ₁₀ O ₂	425	13.0
methyl acetate	C ₃ H ₆ O ₂	275	2.3	2-methylbutanoic acid	C ₅ H ₁₀ O ₂	425	13.5
propan-1-ol	C ₃ H ₈ O	300	2.6	3-methylbutanoic acid	C ₅ H ₁₀ O ₂	425	15.3
ethyl acetate	C ₄ H ₈ O ₂	350	3.5	hex-1-ene	C ₆ H ₁₂	450	15.5
butan-2-one	C ₄ H ₈ O	325	4.4	phenol	C ₆ H ₆ O	325	15.7
propyl acetate	C ₅ H ₁₀ O ₂	425	4.5	ethyl butanoate	C ₆ H ₁₂ O ₂	500	15.8
propanoic acid	C ₃ H ₆ O ₂	275	4.5	2-methylpropanal	C ₄ H ₈ O	325	17.9
propan-1-amine	C ₃ H ₉ N	325	4.5	hex-3-ene	C ₆ H ₁₂	450	17.9
propan-2-ol	C ₃ H ₈ O	300	4.6	toluene	C ₇ H ₈	375	18.0
2-methylpropan-1-amine	C ₄ H ₁₁ N	400	4.8	butan-1-amine	C ₄ H ₁₁ N	400	18.1
2-methylpropan-1-ol	C ₄ H ₁₀ O	375	4.9	hex-2-ene	C ₆ H ₁₂	450	18.7
methyl propanoate	C ₄ H ₈ O ₂	350	5.0	propan-2-yl butanoate	C ₇ H ₁₄ O ₂	575	19.7
propan-2-yl acetate	C ₅ H ₁₀ O ₂	425	5.2	methyl 2-methylbutanoate	C ₆ H ₁₂ O ₂	500	22.6
2-methylpropanoic acid	C ₄ H ₈ O ₂	350	5.3	4-methylphenol	C ₇ H ₈ O	400	22.7
butan-2-ol	C ₄ H ₁₀ O	375	5.4	propyl butanoate	C ₇ H ₁₄ O ₂	575	23.0
butan-1-ol	C ₄ H ₁₀ O	375	5.8	methyl 3-methylbutanoate	C ₆ H ₁₂ O ₂	500	23.1
ethyl propanoate	C ₅ H ₁₀ O ₂	425	6.9	ethylbenzene	C ₈ H ₁₀	450	24.9
butan-2-amine	C ₄ H ₁₁ N	400	7.1	cumene	C ₉ H ₁₂	525	31.6
but-2-ene	C ₄ H ₈	300	8.0	4-ethylphenol	C ₈ H ₁₀ O	475	32.0
2-methylpropan-2-ol	C ₄ H ₁₀ O	375	8.2	propylbenzene	C ₉ H ₁₂	525	34.1
but-1-ene	C ₄ H ₈	300	8.3	propanal	C ₃ H ₆ O	250	37.8
butanoic acid	C ₄ H ₈ O ₂	350	8.5	4-propylphenol	C ₉ H ₁₂ O	550	44.7
methyl 2-methylpropanoate	C ₅ H ₁₀ O ₂	425	8.8	2-methylpropylbenzene	C ₁₀ H ₁₄	600	46.3
butanal	C ₄ H ₈ O	325	8.9	butylbenzene	C ₁₀ H ₁₄	600	47.5
propan-2-yl propanoate	C ₆ H ₁₂ O ₂	500	9.1	butan-2-ylbenzene	C ₁₀ H ₁₄	600	53.8
pent-1-ene	C ₅ H ₁₀	375	9.1	4-butylphenol	C ₁₀ H ₁₄ O	625	59.2
propyl propanoate	C ₆ H ₁₂ O ₂	500	10.0	4-butan-2-ylphenol	C ₁₀ H ₁₄ O	625	60.9
propan-2-amine	C ₃ H ₉ N	325	10.6	4-(2-methylpropyl)phenol	C ₁₀ H ₁₄ O	625	63.9
2-methylpropan-2-amine	C ₄ H ₁₁ N	400	10.8	acetone	C ₃ H ₆ O	250	71.1
pent-2-ene	C ₅ H ₁₀	375	11.4	4-propan-2-ylphenol	C ₉ H ₁₂ O	550	82.0
benzene	C ₆ H ₆	300	11.9				

2 Plots of EIMS and differences in EIMS

The following subsections contain plots for the experimental (NIST) and predicted (CFM-EI, QCEIMS) electron impact mass spectra (EIMS) for all compounds examined in this paper. Peaks are normalised as is typical, such that the largest peak has a relative intensity of 100, and all other peaks maintain their proportion to the largest peak.

Further, they contain plots of the differences between predicted and experimental spectra. Peaks are coloured according to if they were in both (green), only in experiment (yellow) or only present in predicted spectra (red). The relative intensities of the peaks on these plots indicates the degree to which they were over- or underestimated in the predictions.

Since the spectra are built up from counts of the occurrences of fragments, the error bars were calculated by the $\sqrt{N}$ where N is the number of trajectories.

All raw spectra data are available upon request.

2.1 Alcohols

2.1.1 1: propan-1-ol

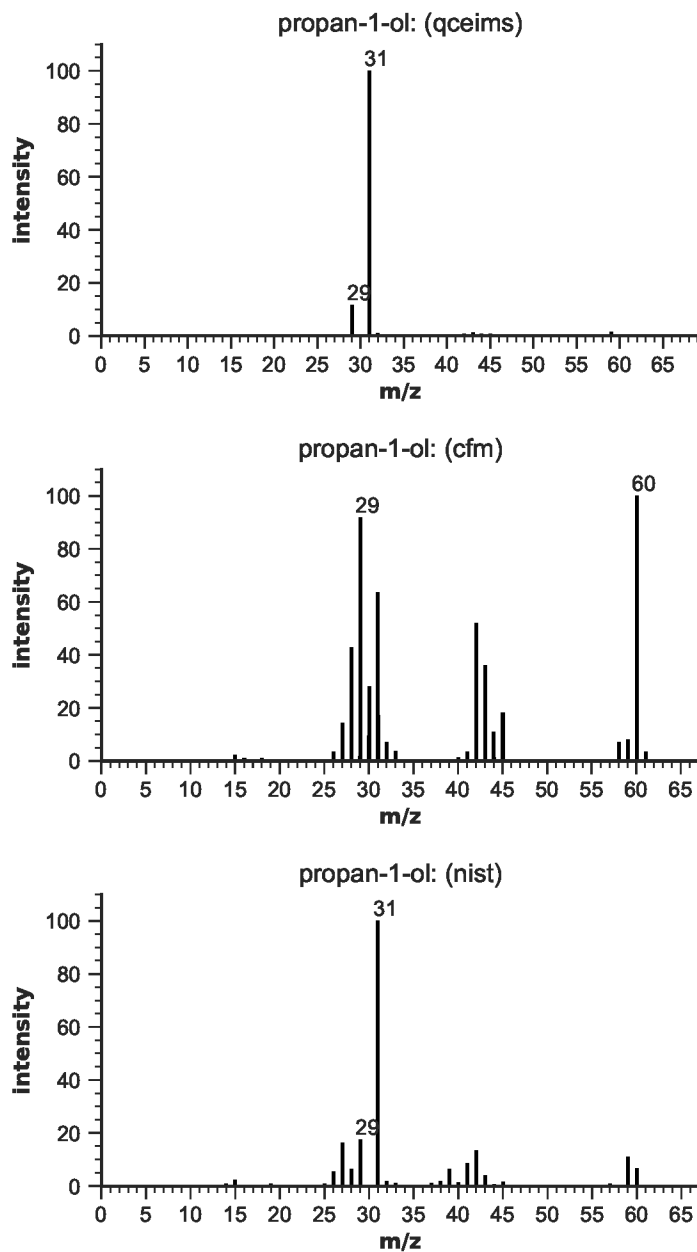


Figure S1: EIMS for propan-1-ol

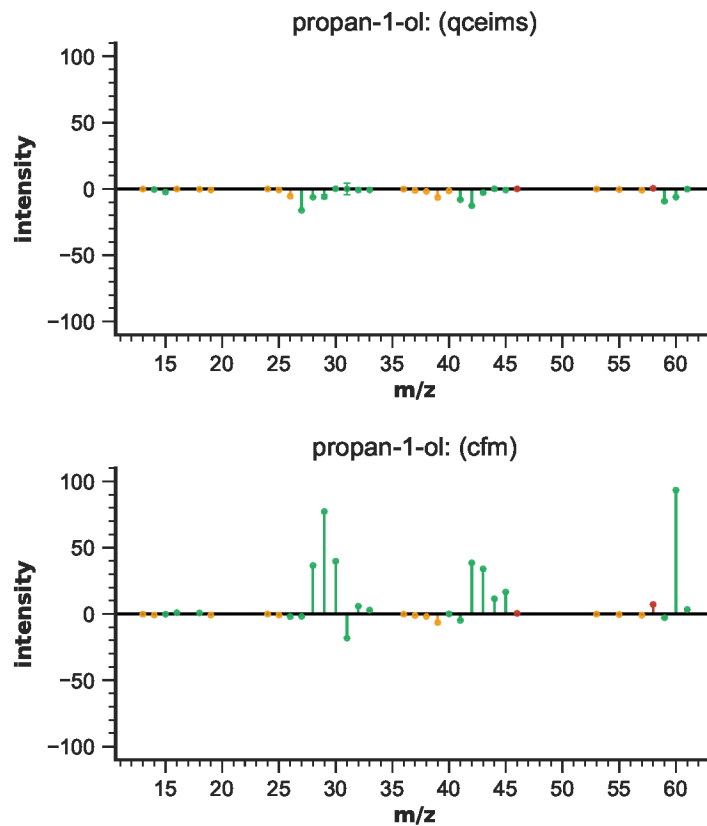


Figure S2: EIMS differences for propan-1-ol

2.1.1.2 2: propan-2-ol

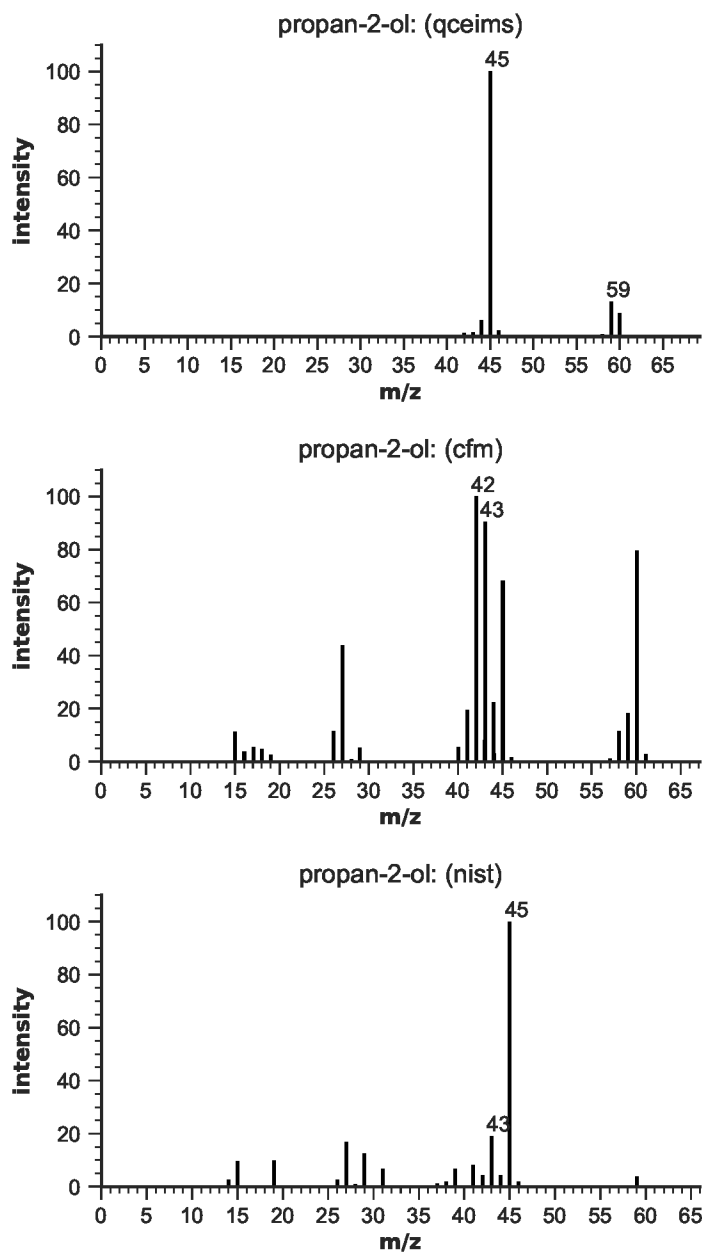


Figure S3: EIMS for propan-2-ol

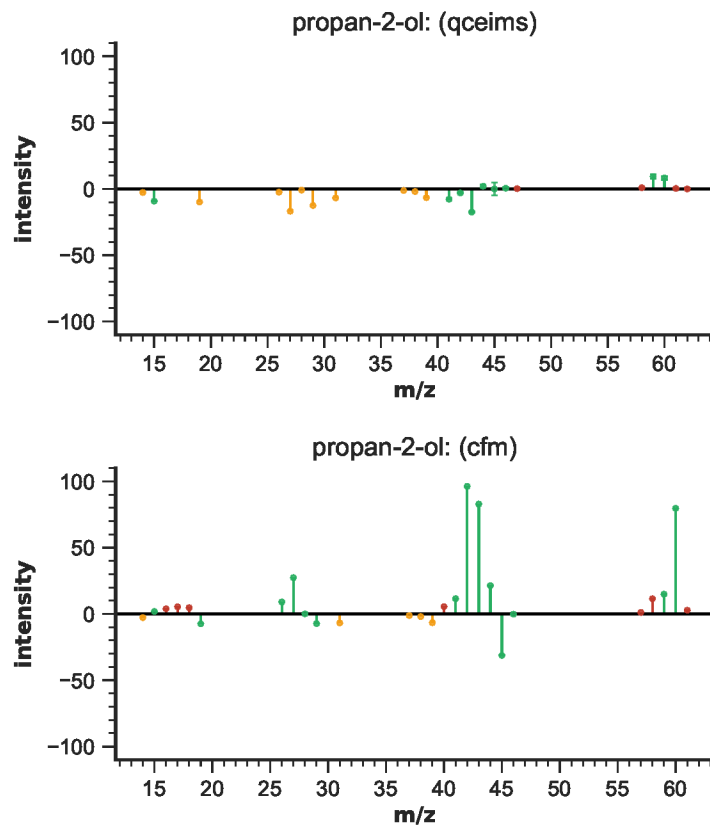


Figure S4: EIMS differences for propan-2-ol

2.1.3 3: butan-1-ol

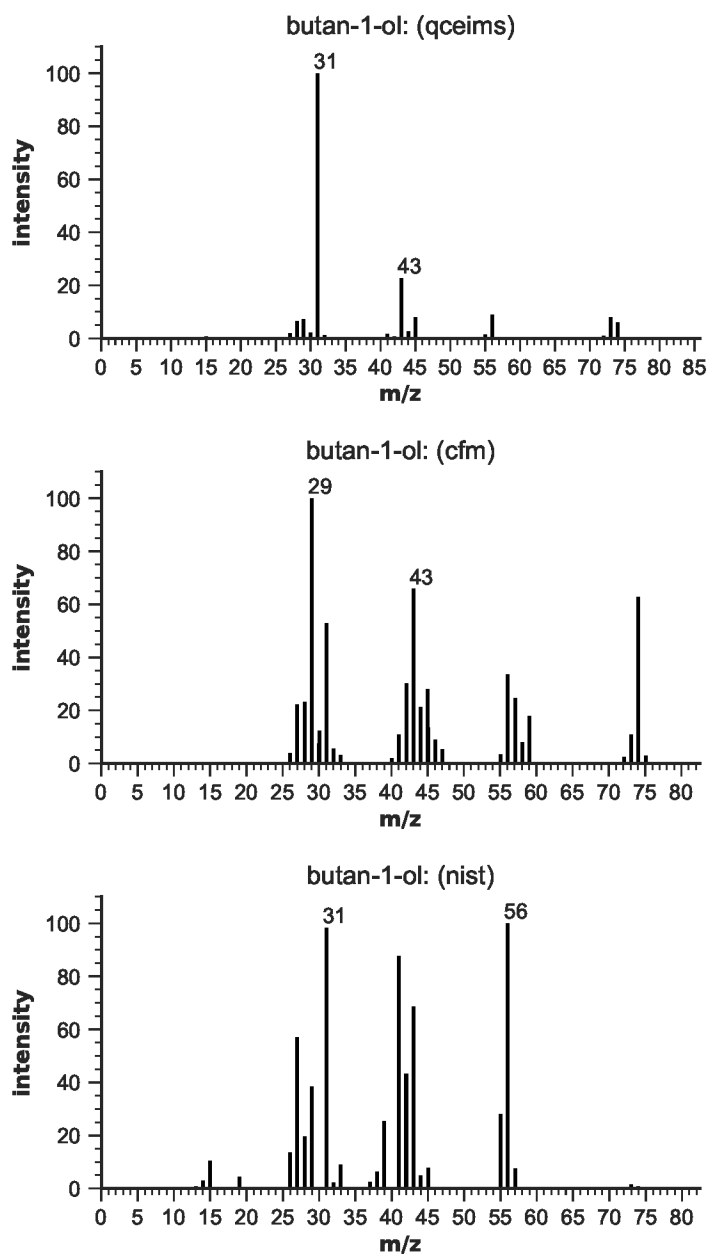


Figure S5: EIMS for butan-1-ol

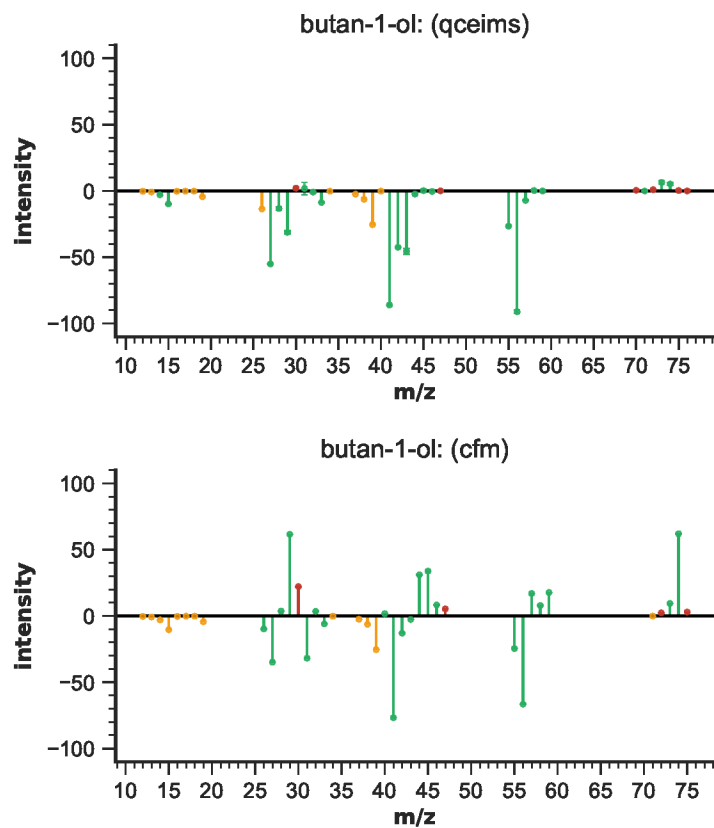


Figure S6: EIMS differences for butan-1-ol

2.1.4 4: butan-2-ol

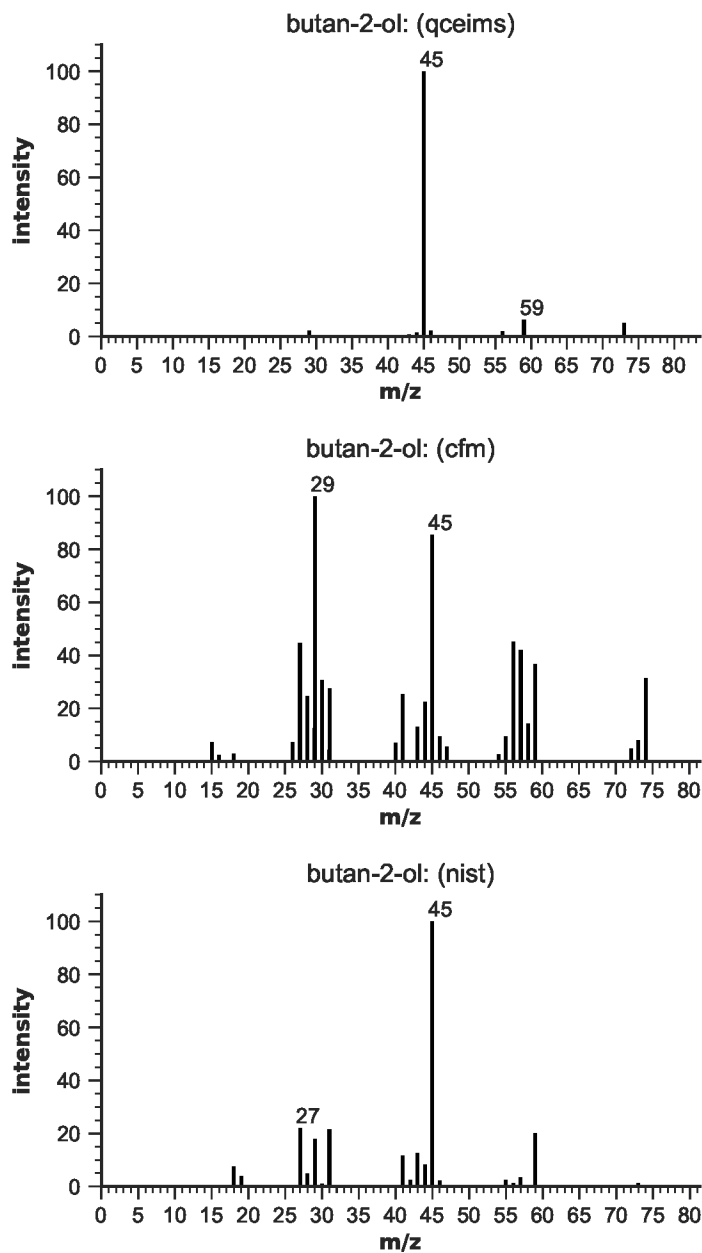


Figure S7: EIMS for butan-2-ol

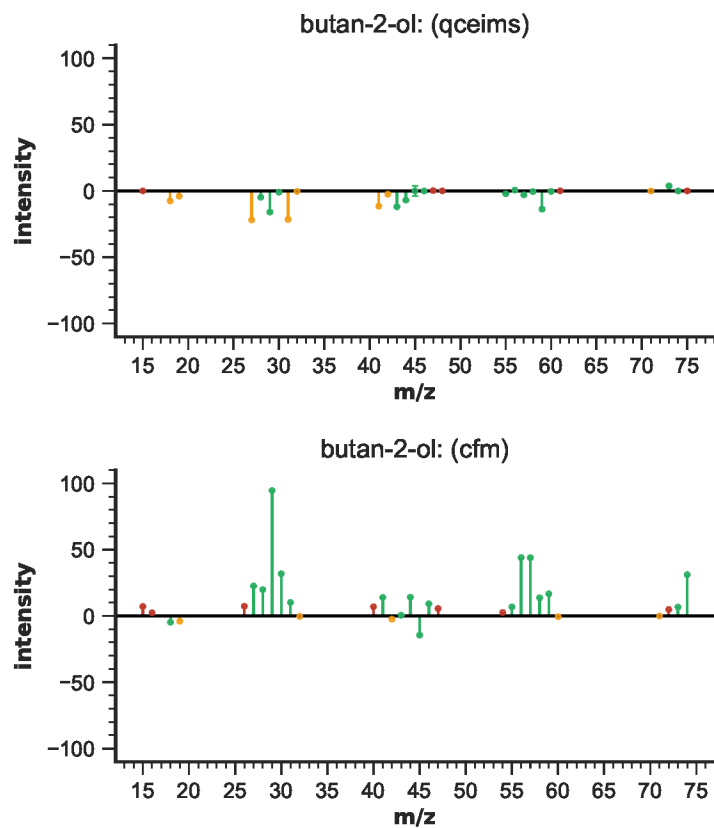


Figure S8: EIMS differences for butan-2-ol

2.1.5 5: 2-methylpropan-1-ol

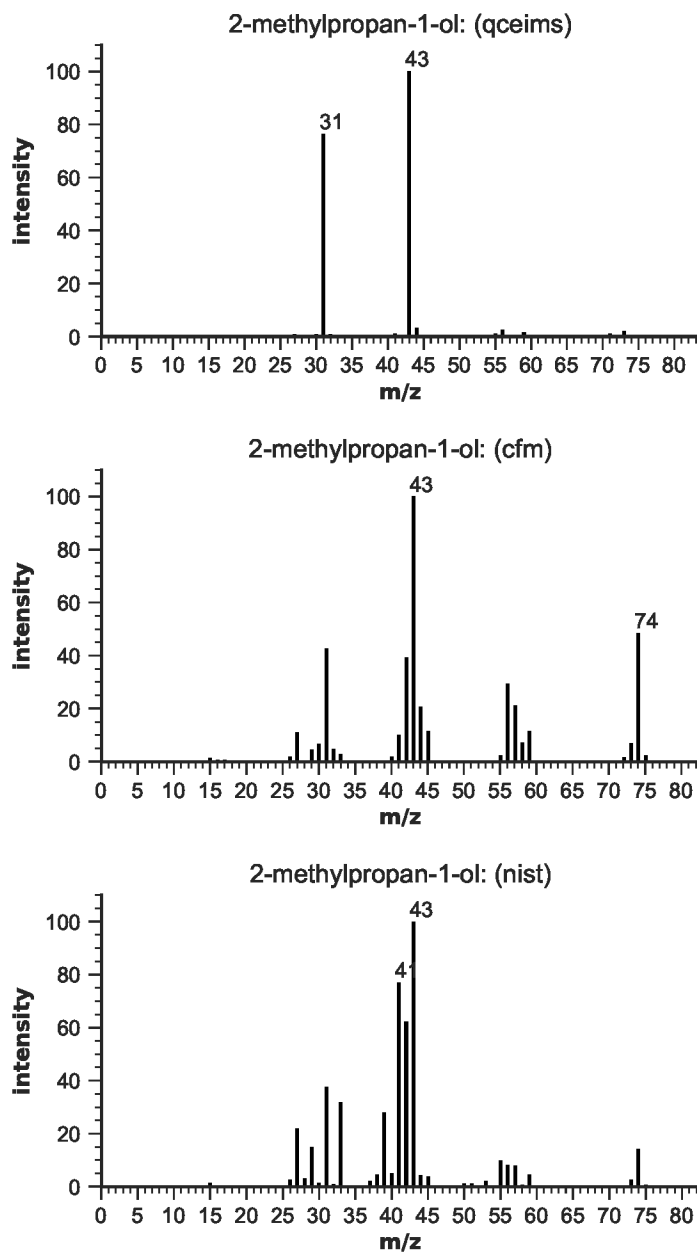


Figure S9: EIMS for 2-methylpropan-1-ol

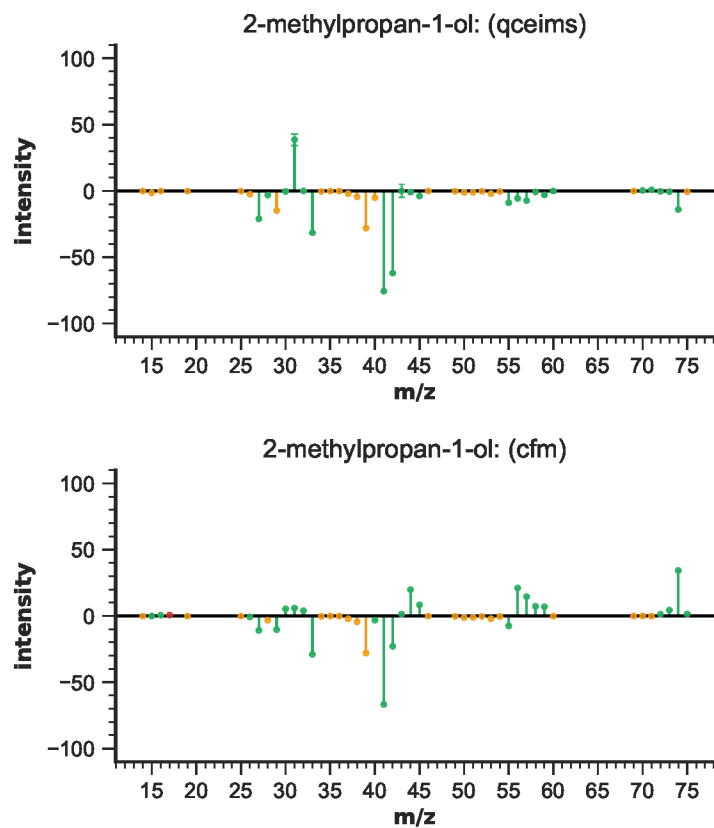


Figure S10: EIMS differences for 2-methylpropan-1-ol

2.1.6 6: 2-methylpropan-2-ol

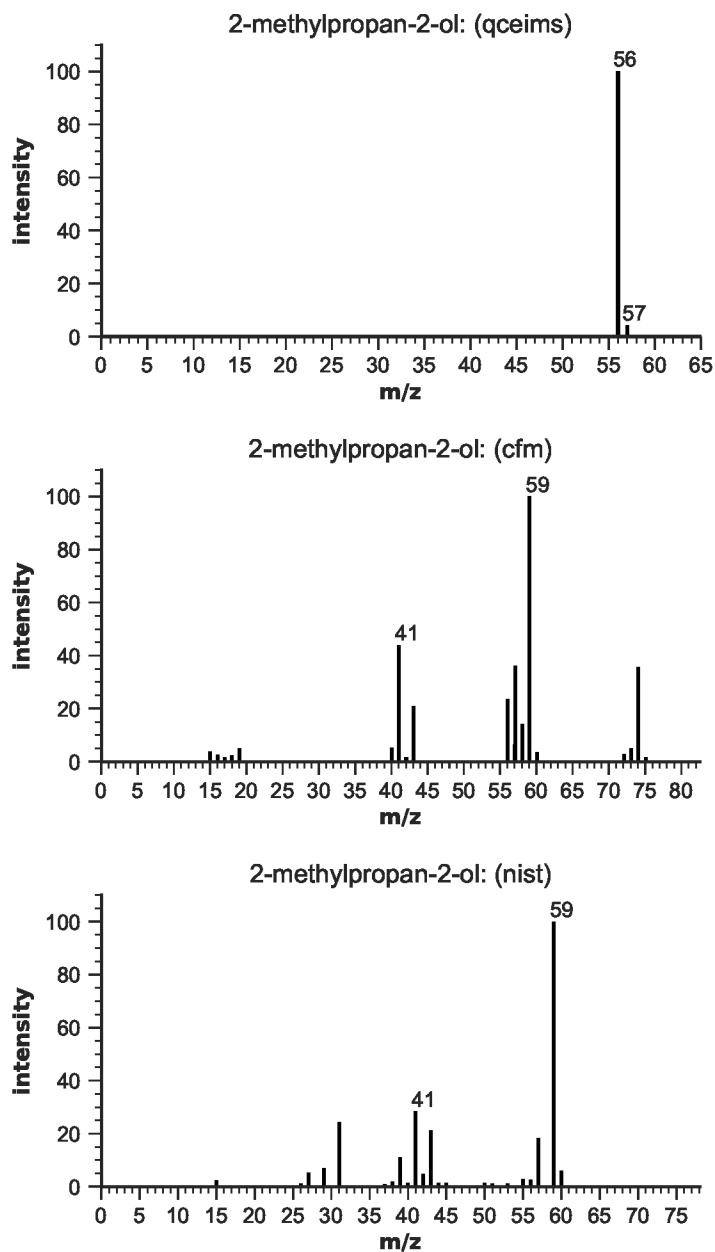


Figure S11: EIMS for 2-methylpropan-2-ol

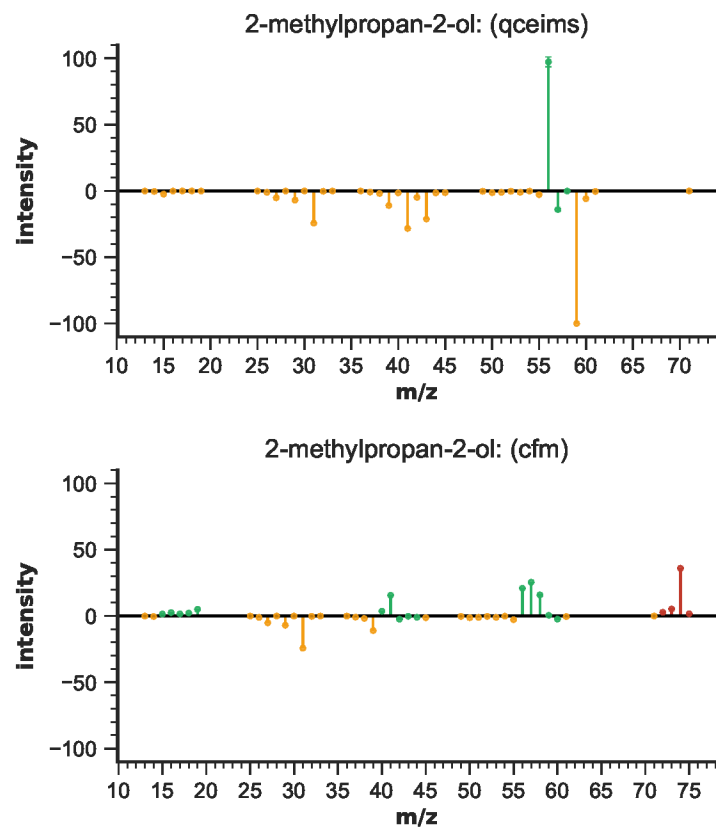


Figure S12: EIMS differences for 2-methylpropan-2-ol

2.2 Aldehydes

2.2.1 7: propanal

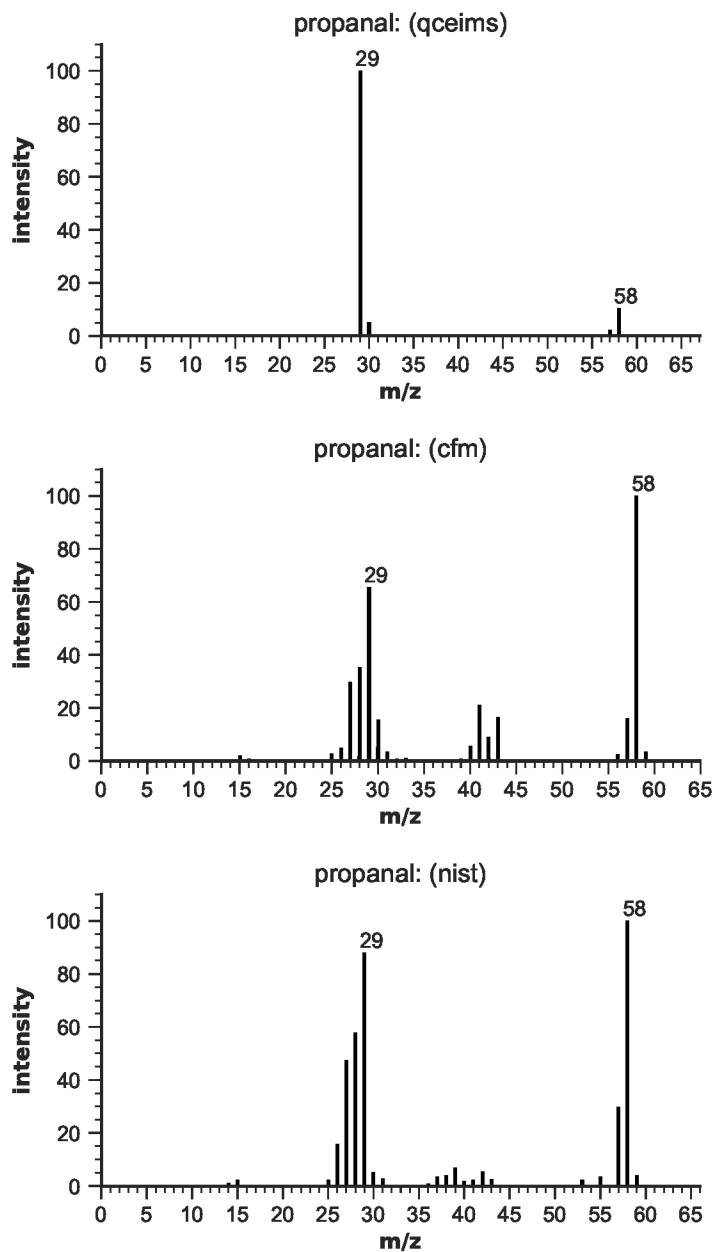


Figure S13: EIMS for propanal

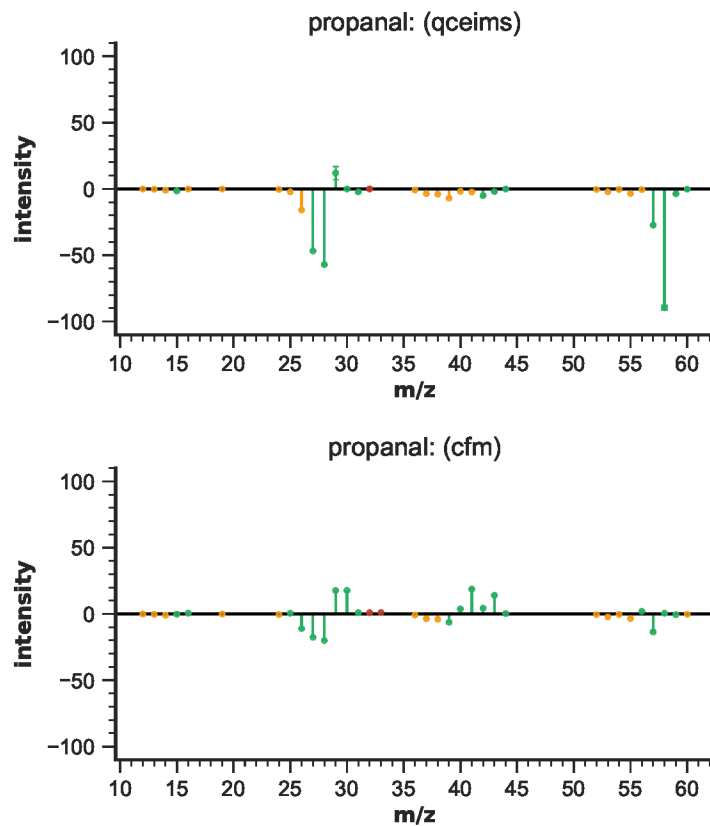


Figure S14: EIMS differences for propanal

2.2.2 8: 2-methylpropanal

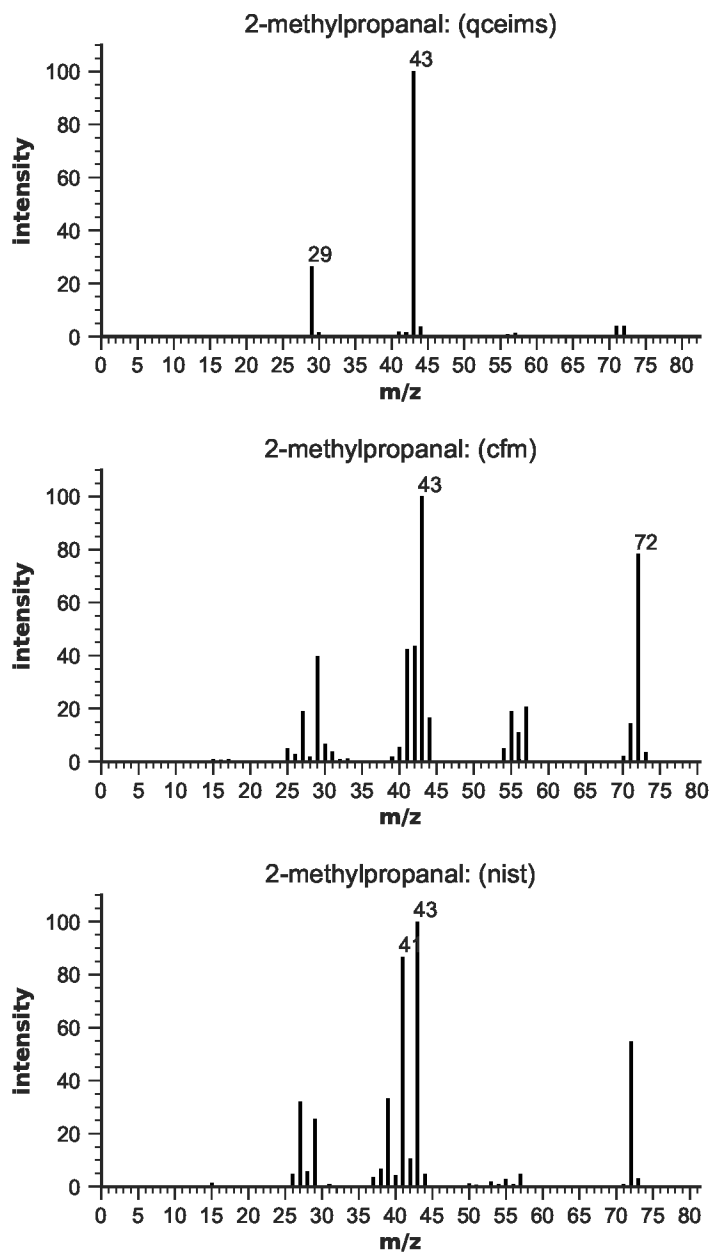


Figure S15: EIMS for 2-methylpropanal

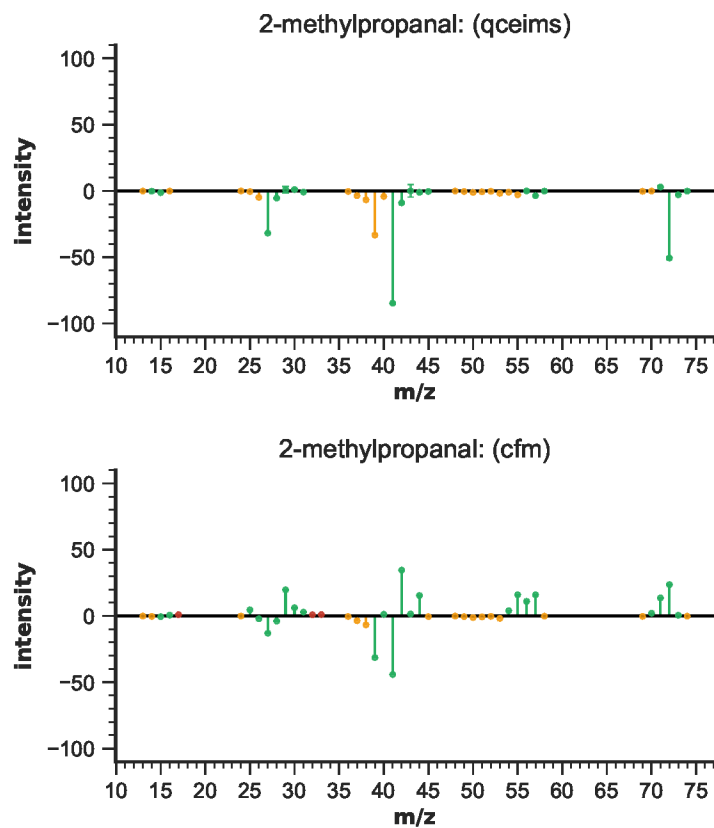


Figure S16: EIMS differences for 2-methylpropanal

2.2.3 9: butanal

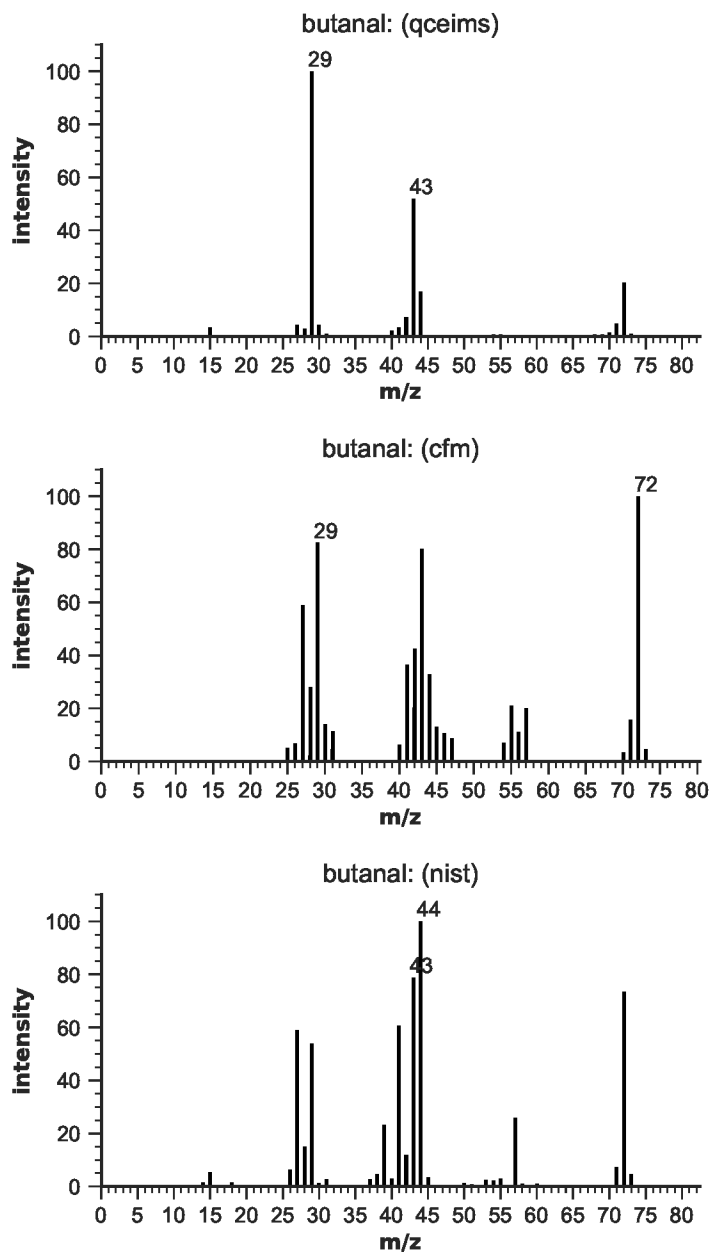


Figure S17: EIMS for butanal

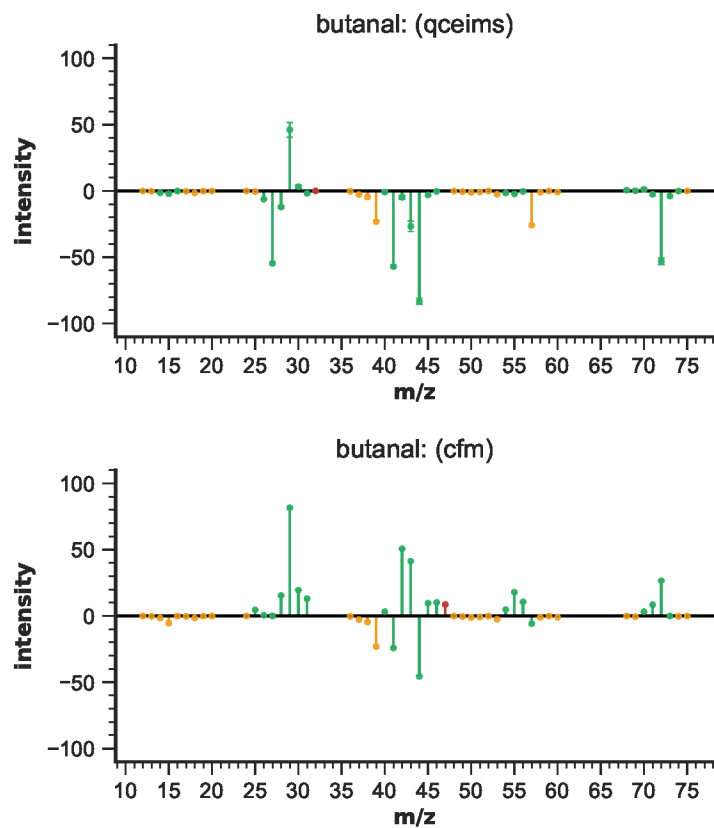


Figure S18: EIMS differences for butanal

2.3 Alkenes

2.3.1 10: but-1-ene

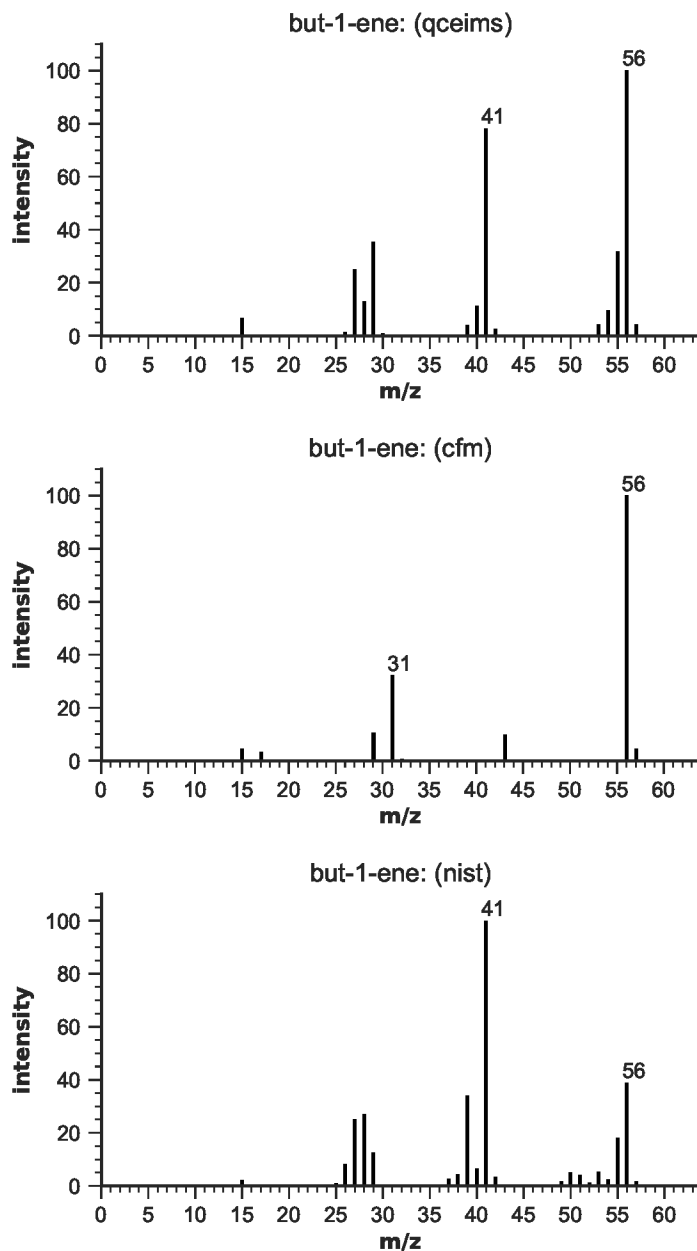


Figure S19: EIMS for but-1-ene

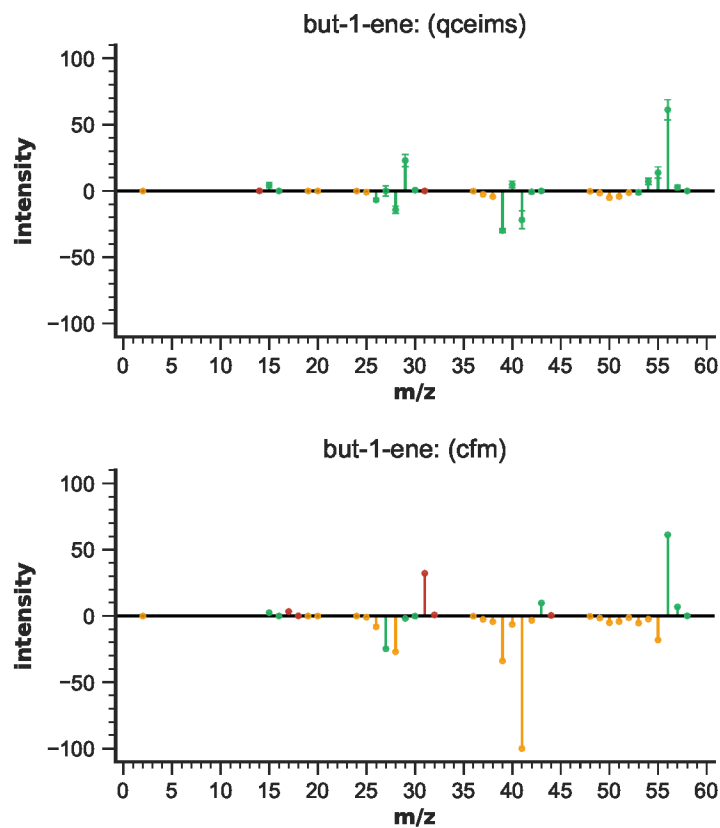


Figure S20: EIMS differences for but-1-ene

2.3.2 11: but-2-ene

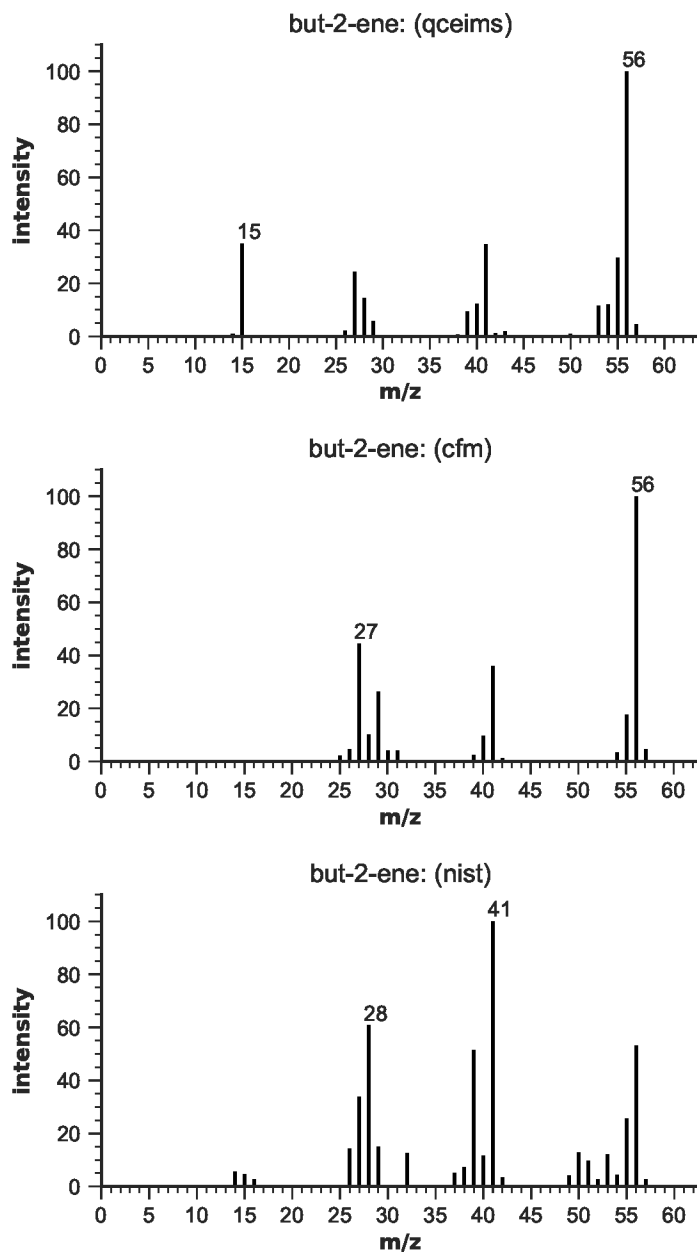


Figure S21: EIMS for but-2-ene

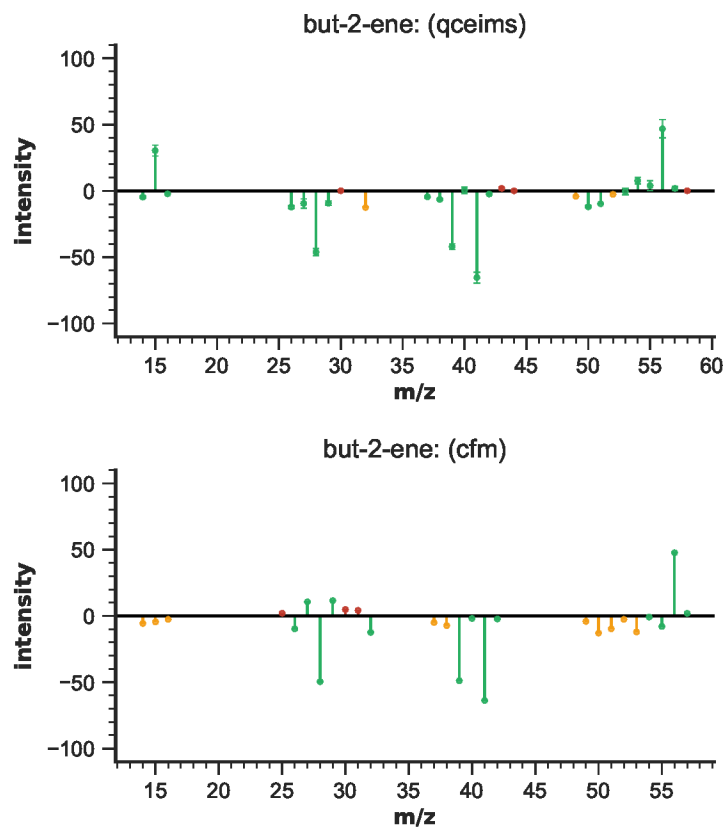


Figure S22: EIMS differences for but-2-ene

2.3.3 12: pent-1-ene

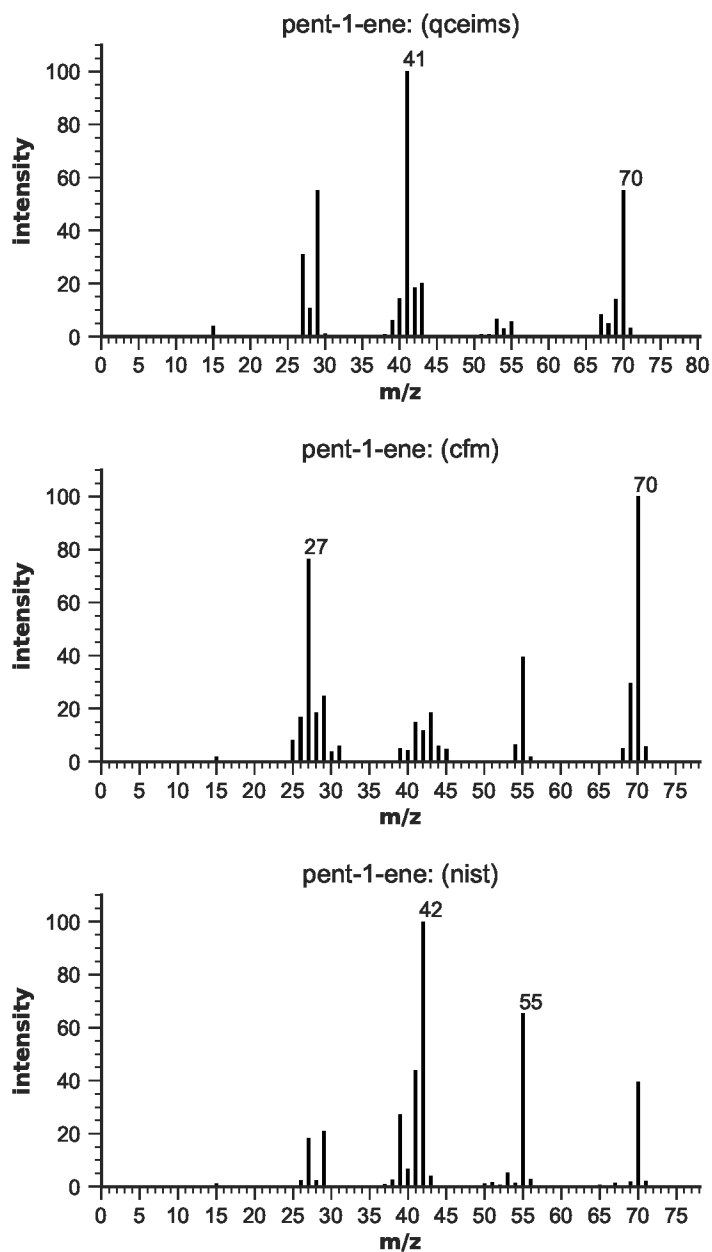


Figure S23: EIMS for pent-1-ene

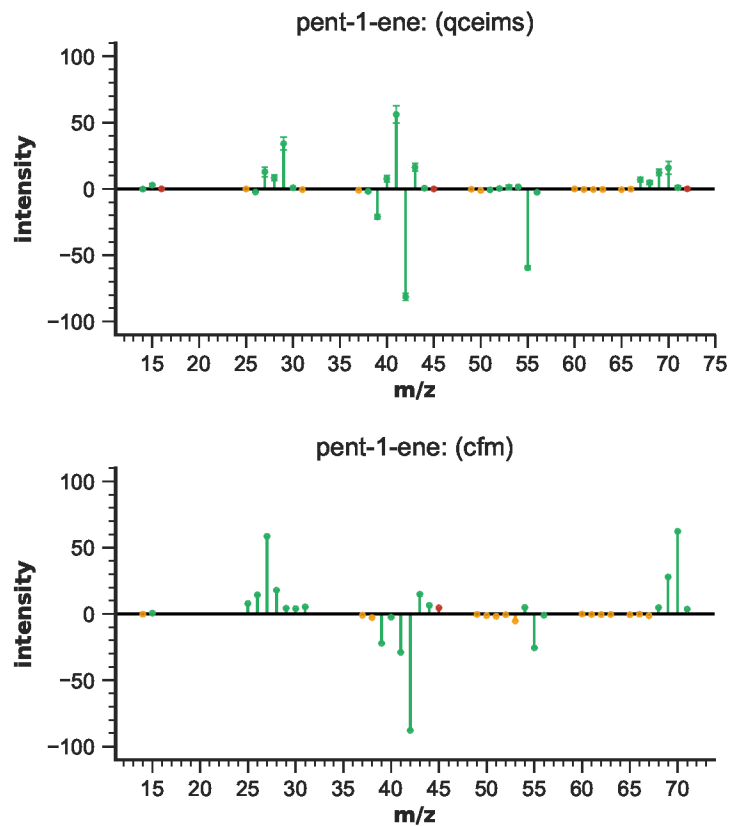


Figure S24: EIMS differences for pent-1-ene

2.3.4 13: pent-2-ene

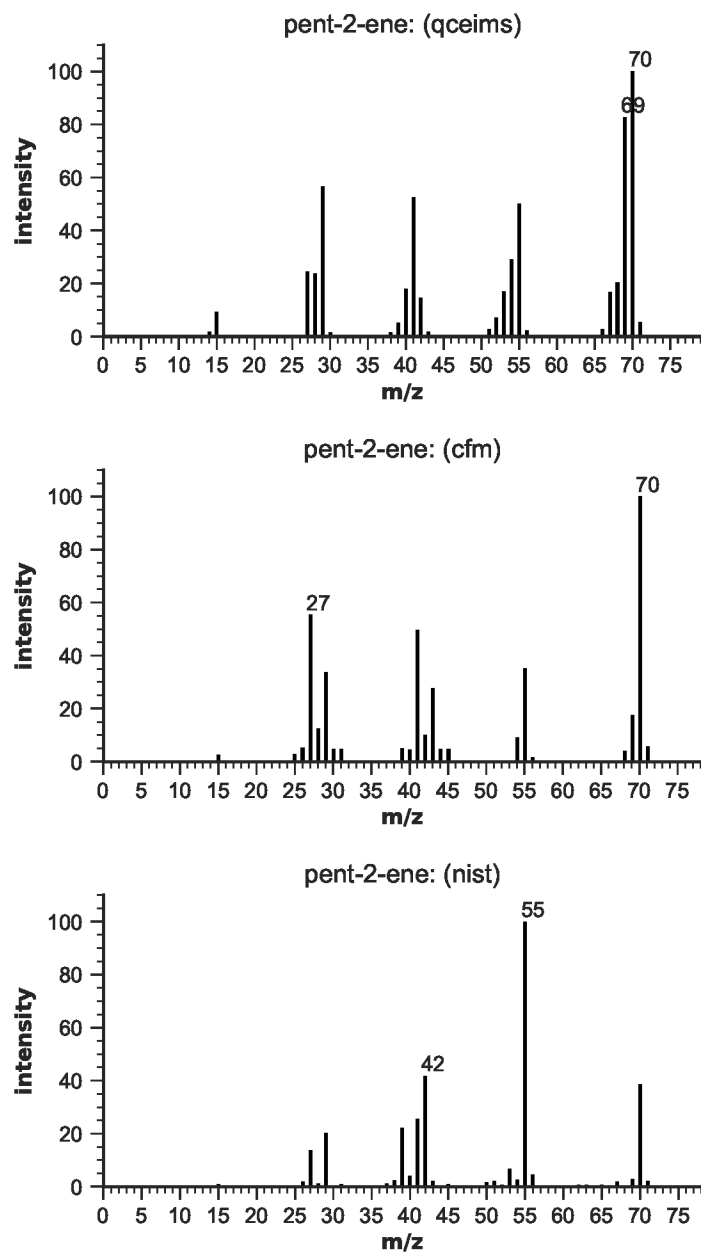


Figure S25: EIMS for pent-2-ene

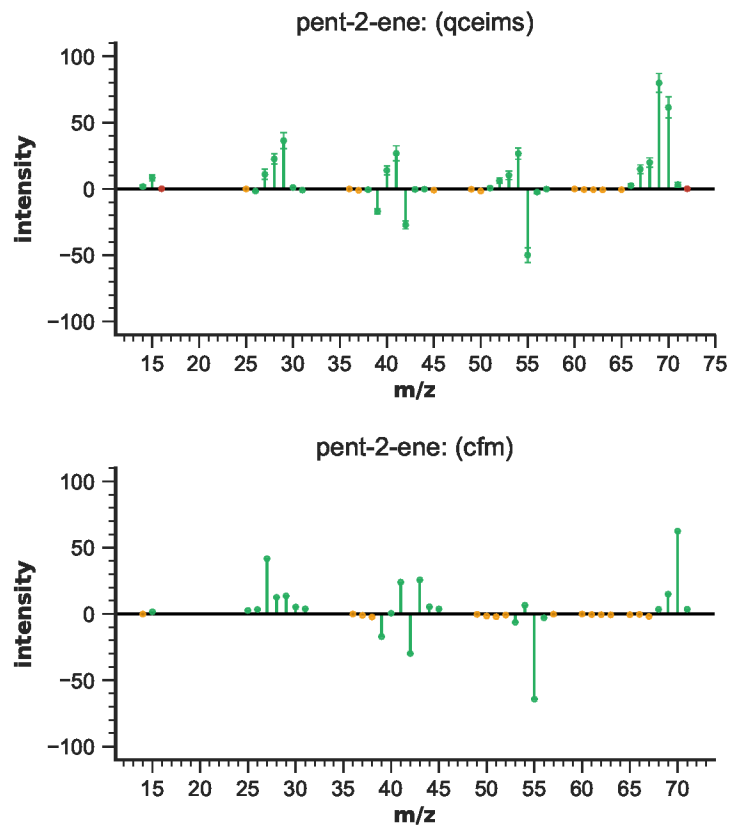


Figure S26: EIMS differences for pent-2-ene

2.3.5 14: hex-1-ene

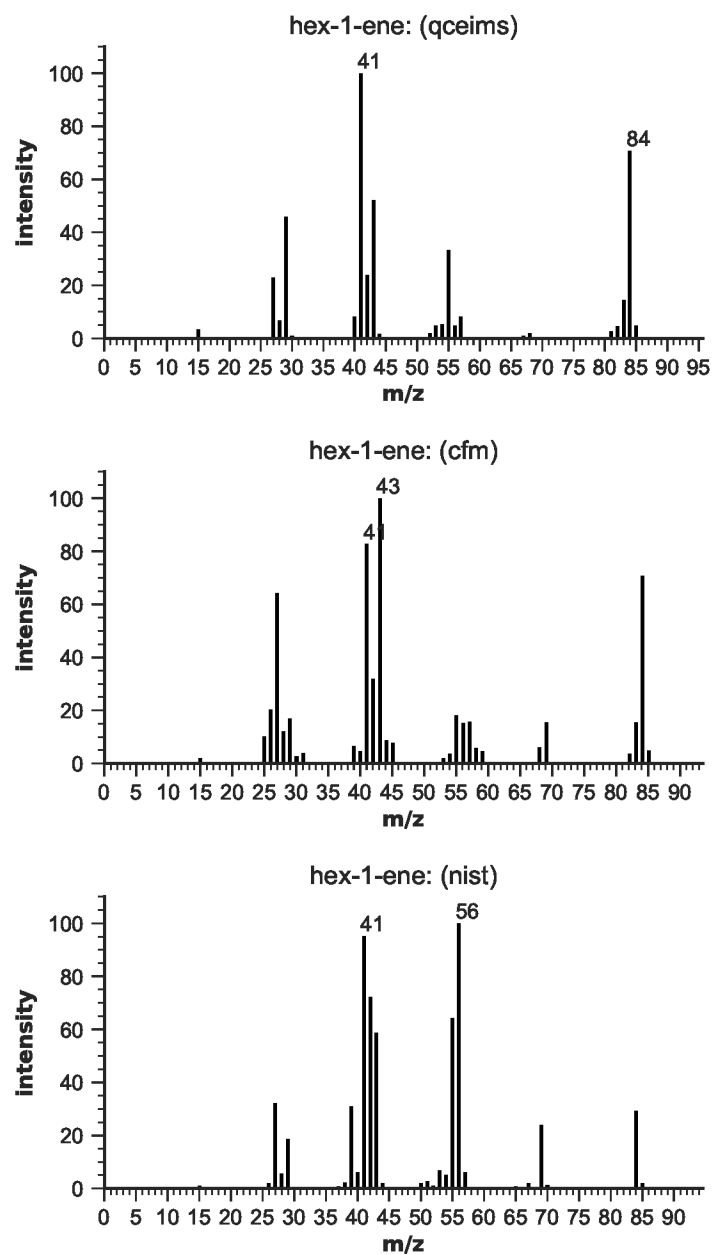


Figure S27: EIMS for hex-1-ene

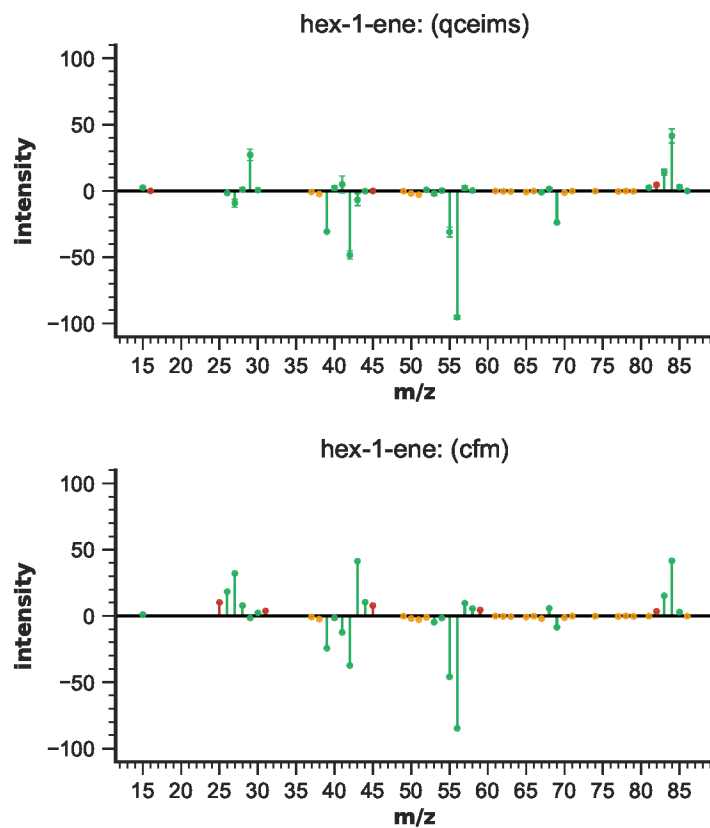


Figure S28: EIMS differences for hex-1-ene

2.3.6 15: hex-2-ene

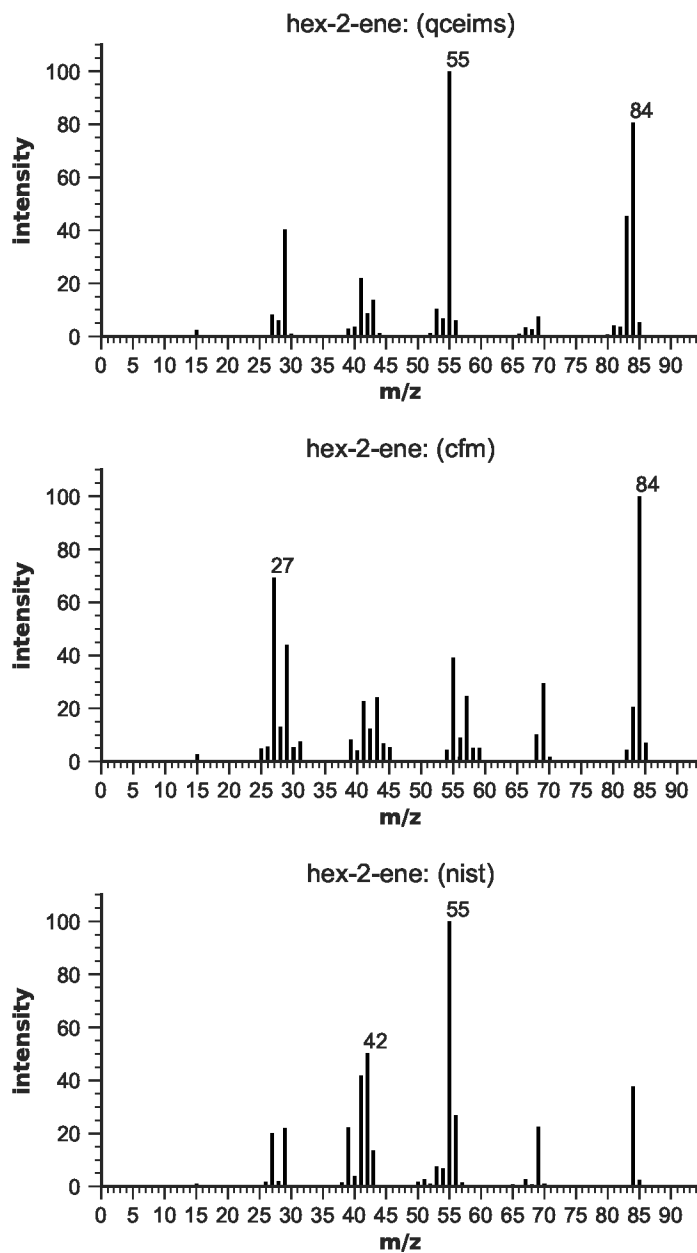


Figure S29: EIMS for hex-2-ene

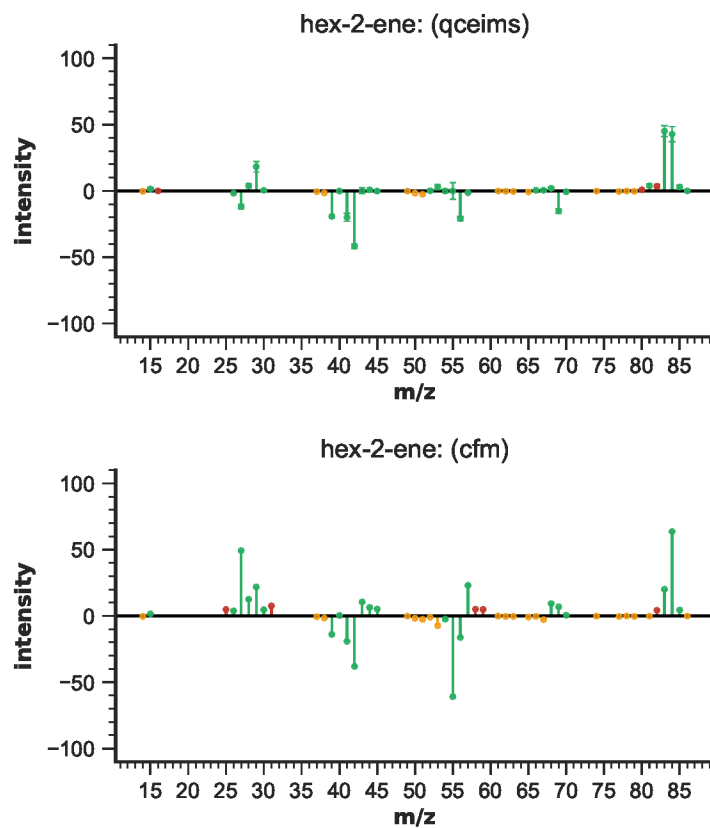


Figure S30: EIMS differences for hex-2-ene

2.3.7 16: hex-3-ene

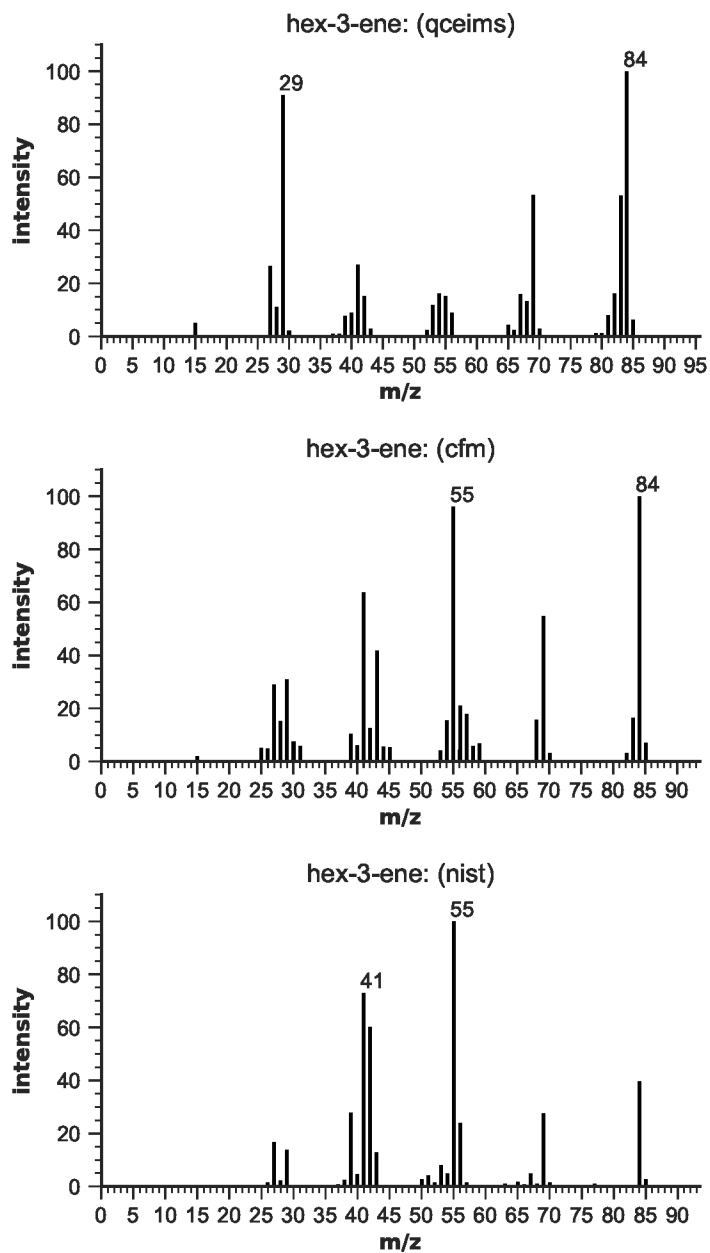


Figure S31: EIMS for hex-3-ene

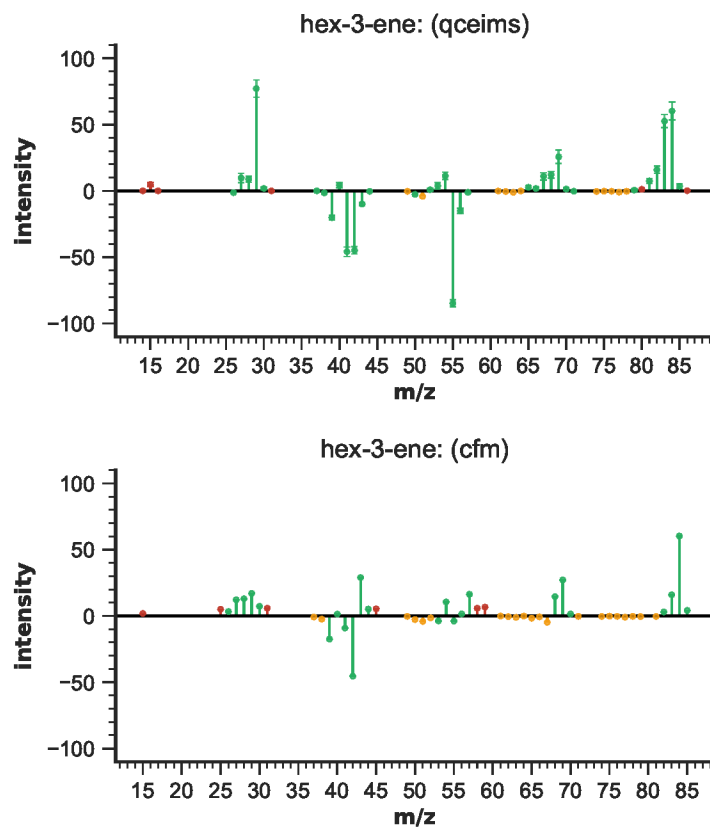


Figure S32: EIMS differences for hex-3-ene

2.4 Amines

2.4.1 17: propan-1-amine

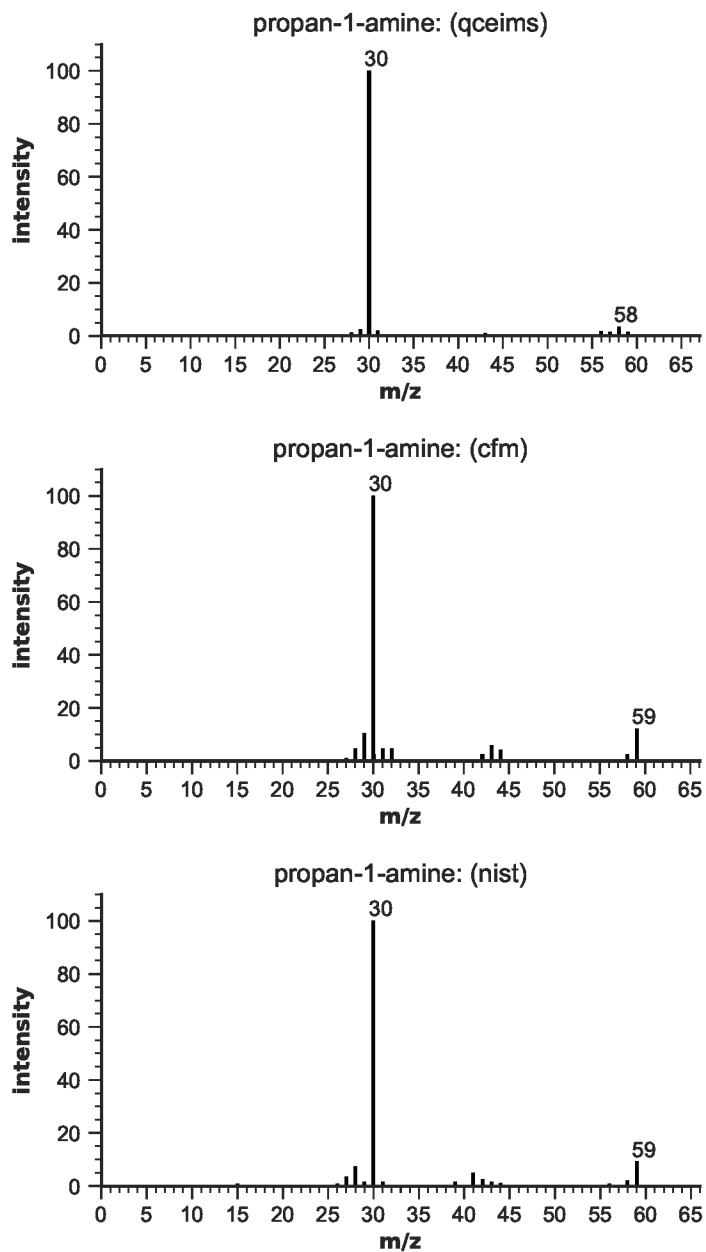


Figure S33: EIMS for propan-1-amine

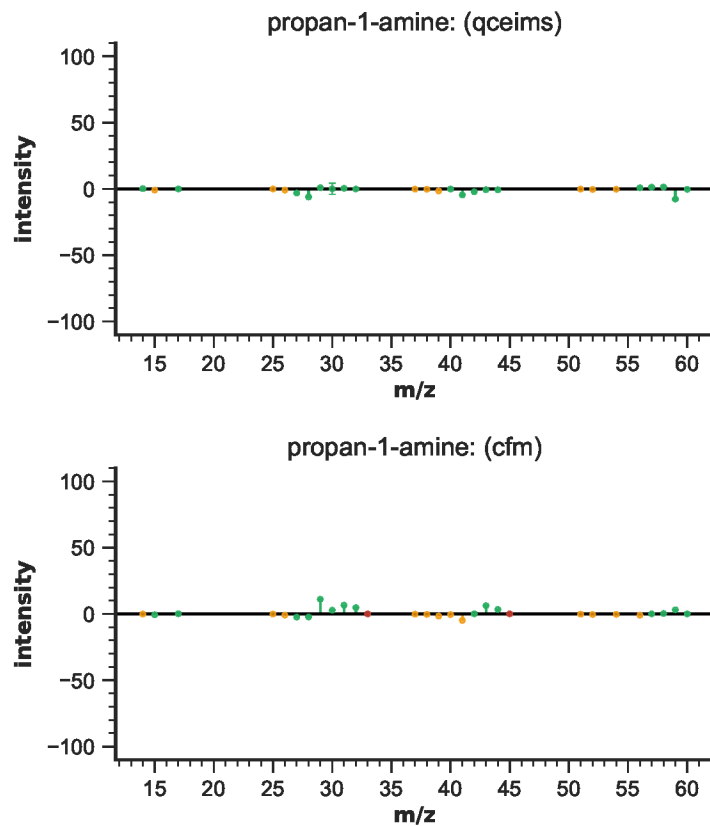


Figure S34: EIMS differences for propan-1-amine

2.4.2 18: propan-2-amine

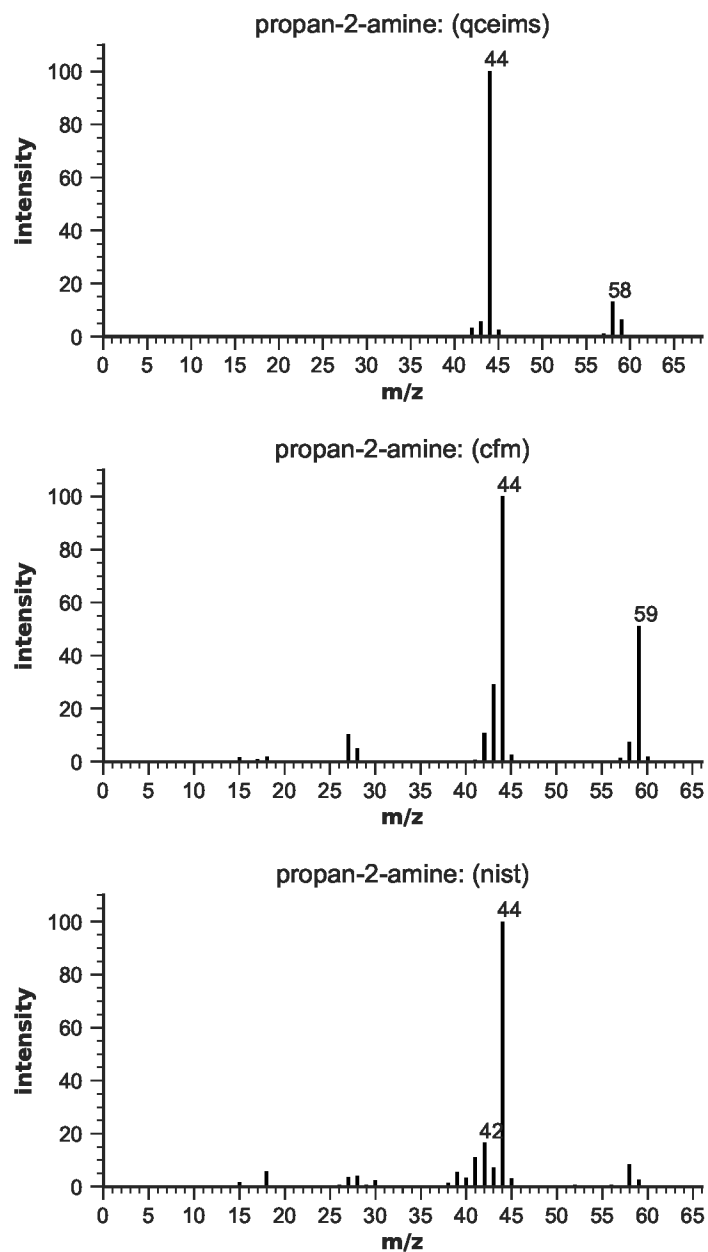


Figure S35: EIMS for propan-2-amine

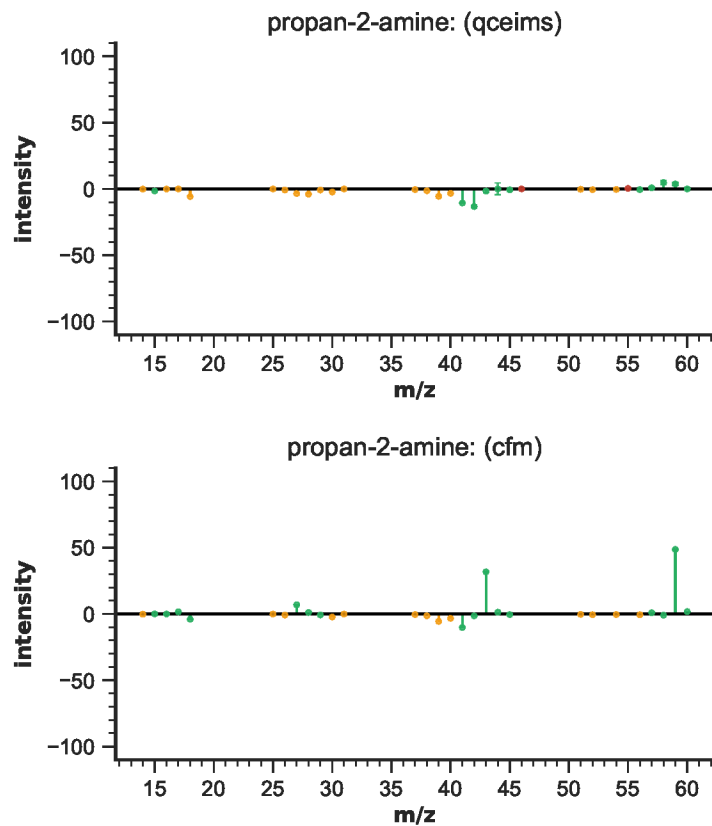


Figure S36: EIMS differences for propan-2-amine

2.4.3 19: butan-1-amine

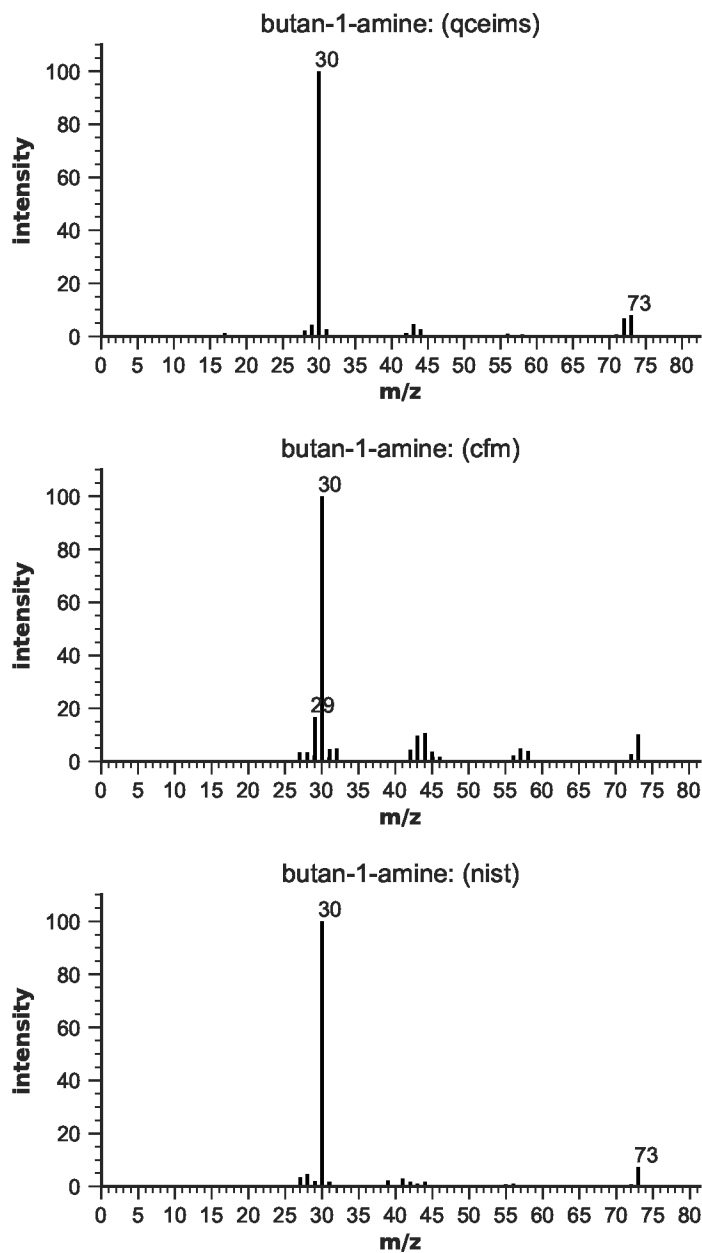


Figure S37: EIMS for butan-1-amine

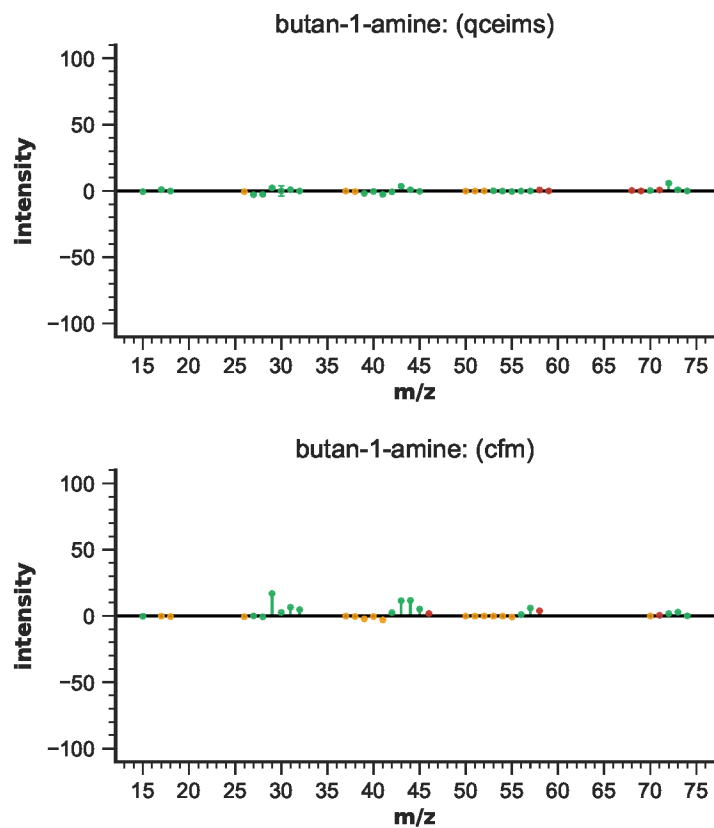


Figure S38: EIMS differences for butan-1-amine

2.4.4 20: butan-2-amine

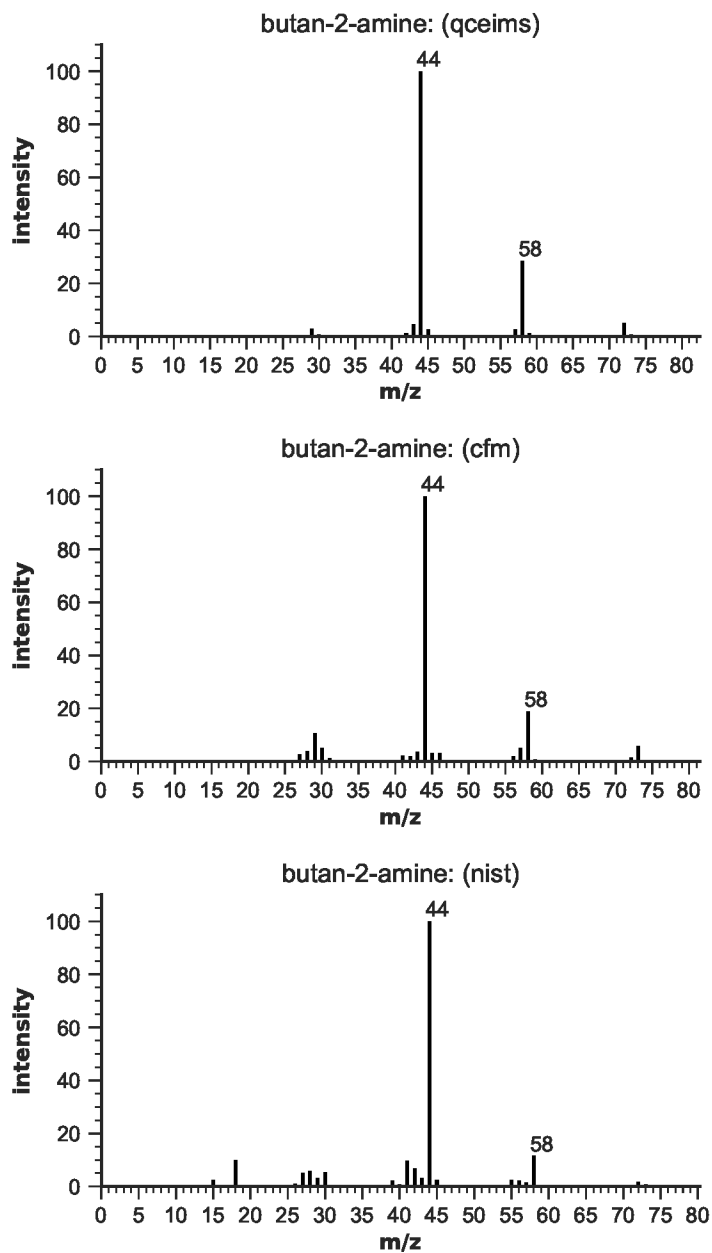


Figure S39: EIMS for butan-2-amine

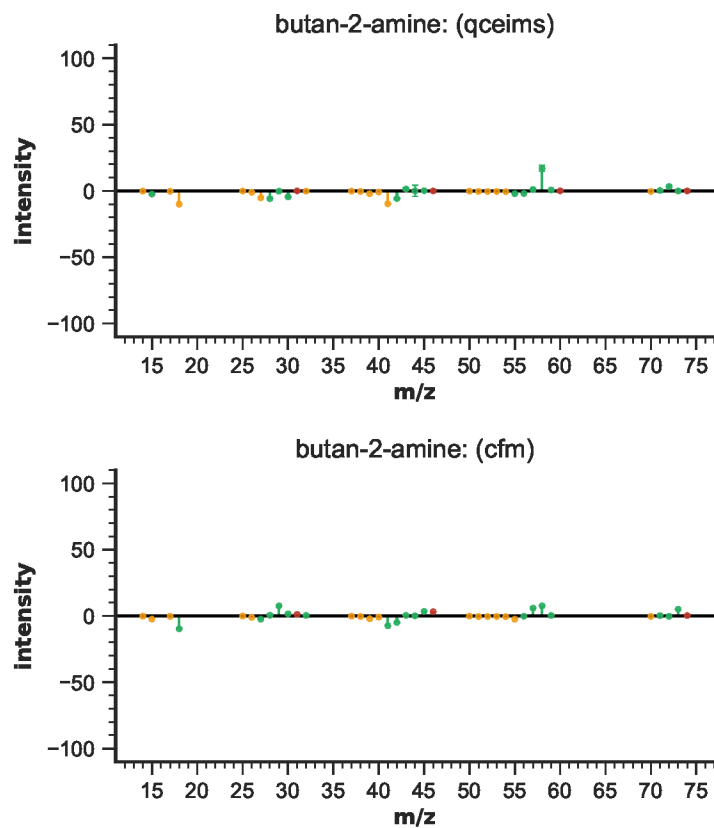


Figure S40: EIMS differences for butan-2-amine

2.4.5 21: 2-methylpropan-1-amine

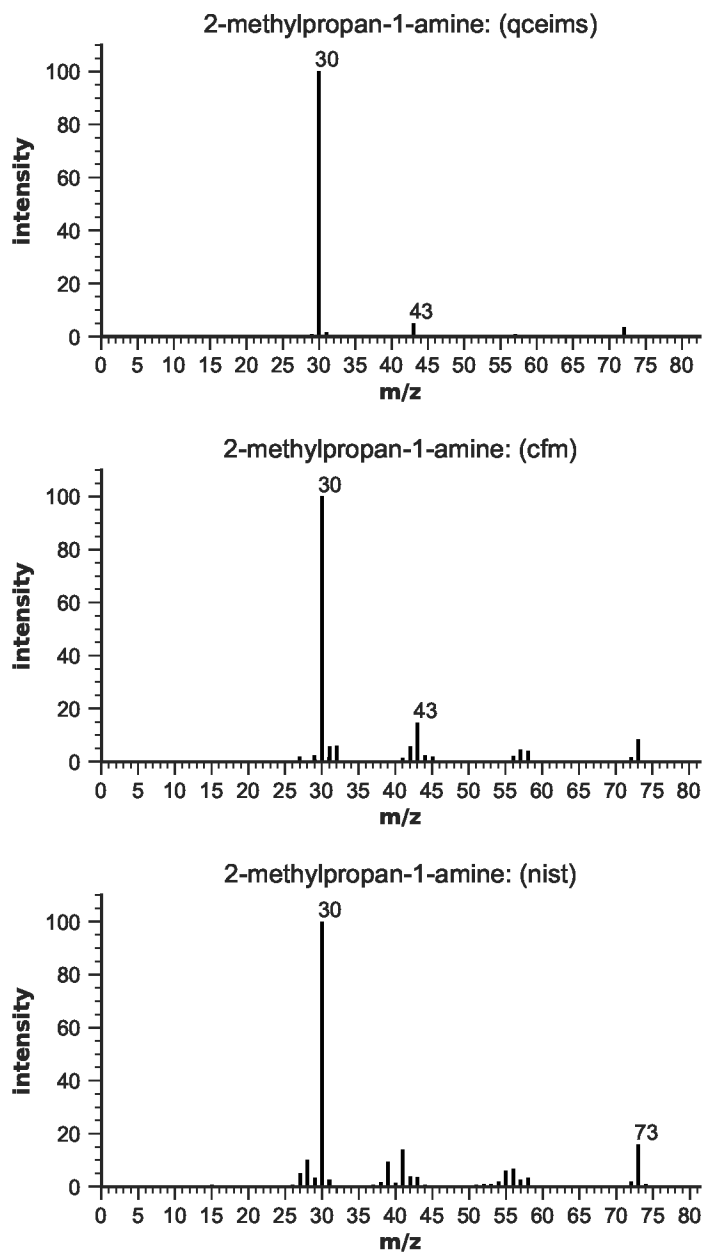


Figure S41: EIMS for 2-methylpropan-1-amine

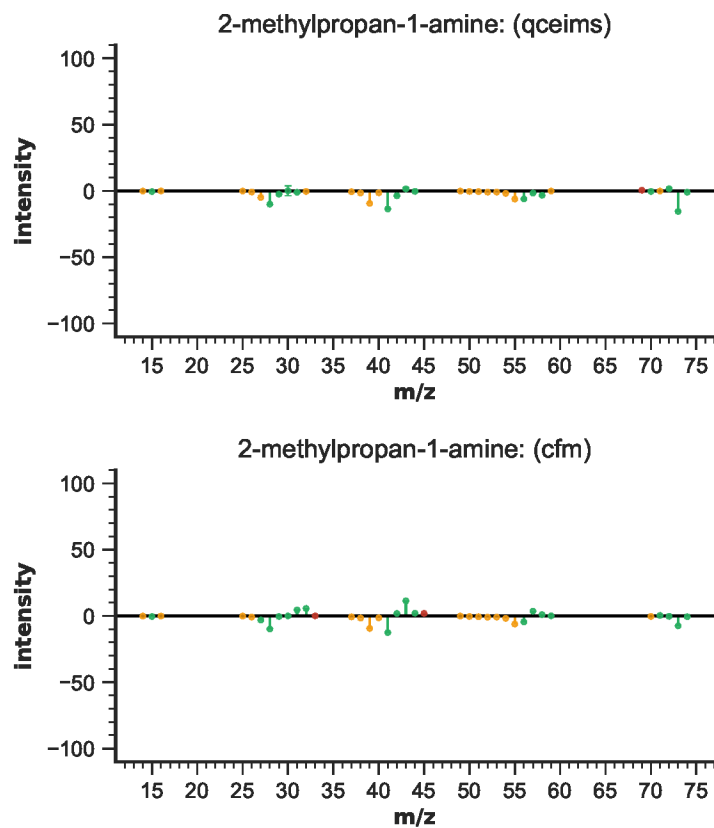


Figure S42: EIMS differences for 2-methylpropan-1-amine

2.4.6 22: 2-methylpropan-2-amine

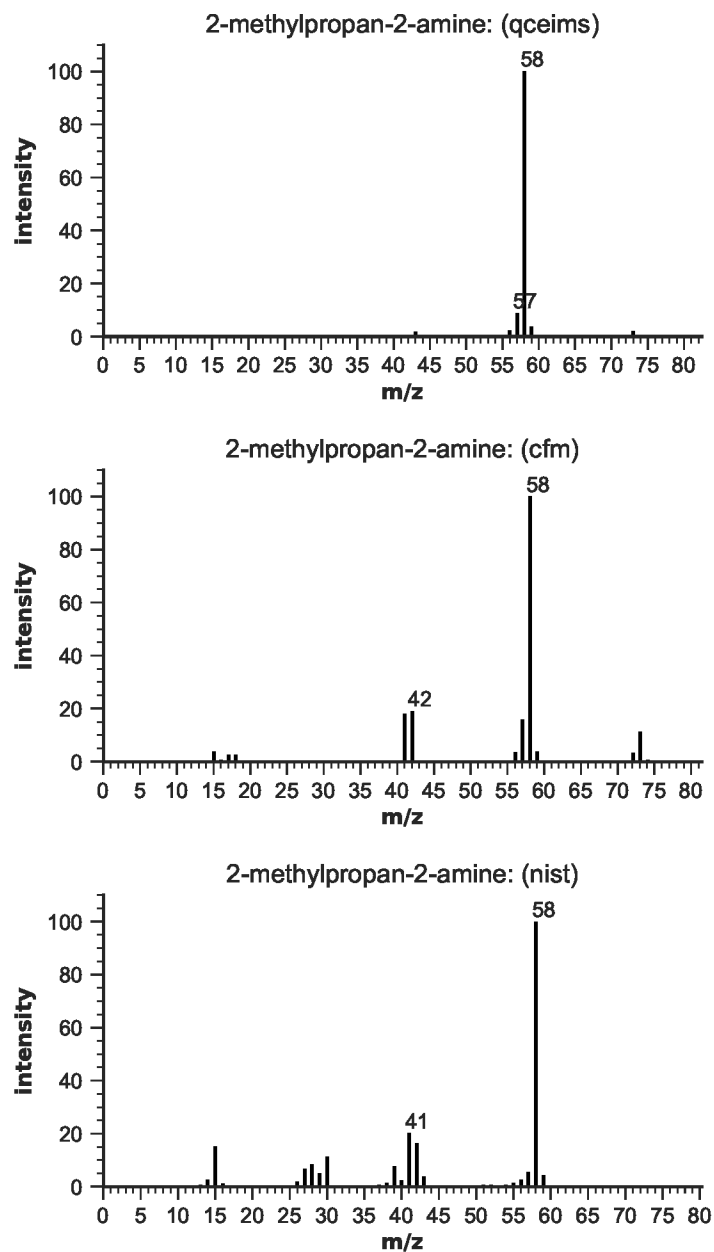


Figure S43: EIMS for 2-methylpropan-2-amine

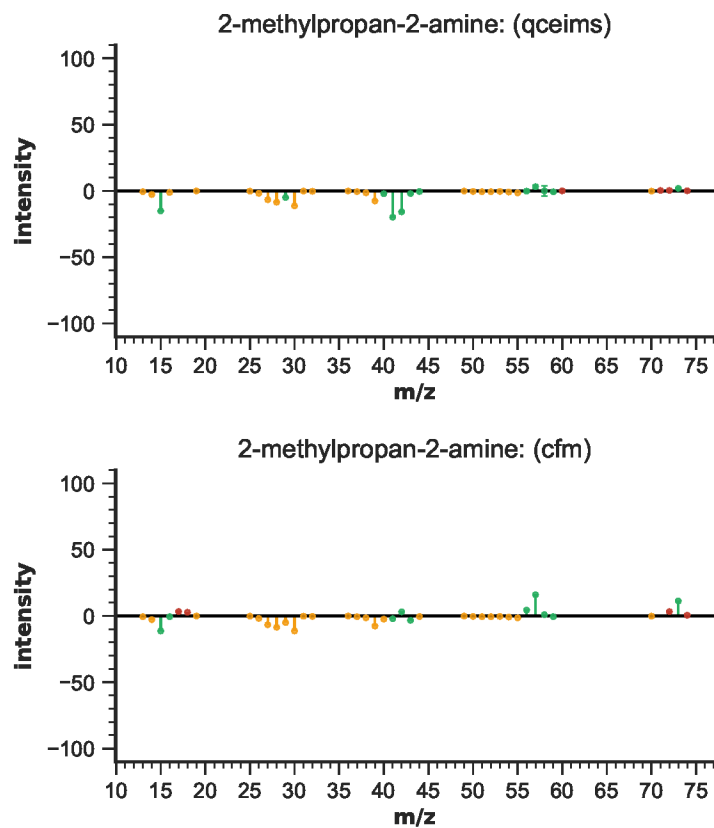


Figure S44: EIMS differences for 2-methylpropan-2-amine

2.5 Benzenes

2.5.1 23: benzene

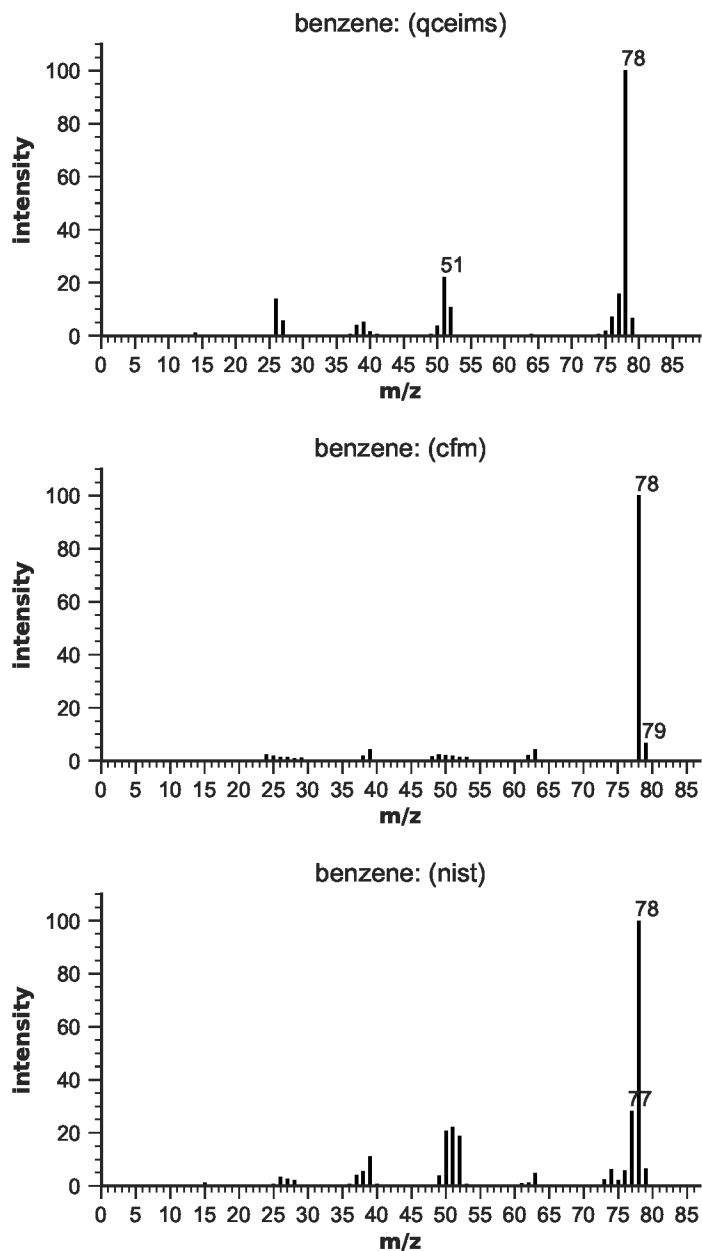


Figure S45: EIMS for benzene

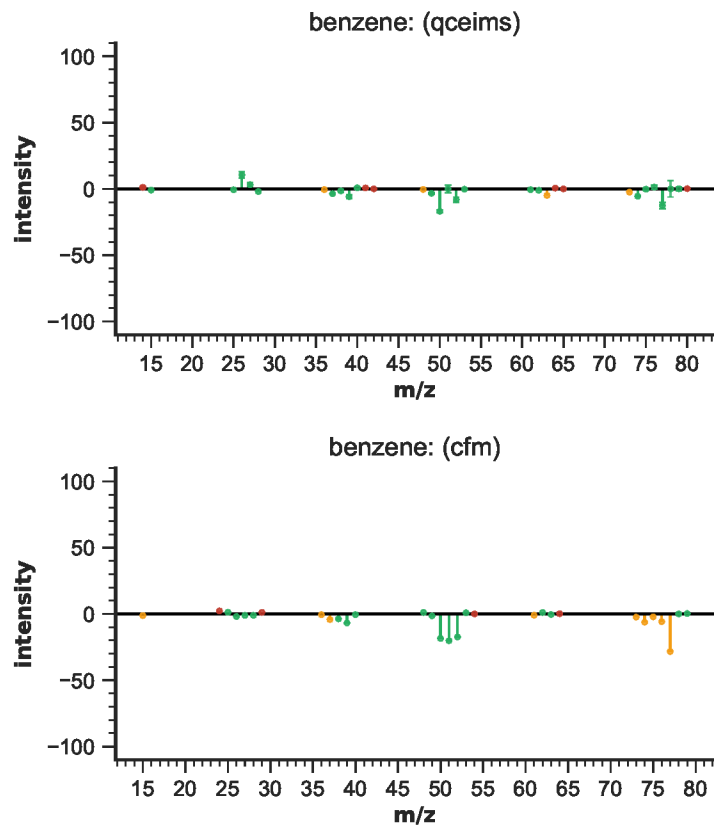


Figure S46: EIMS differences for benzene

2.5.2 24: toluene

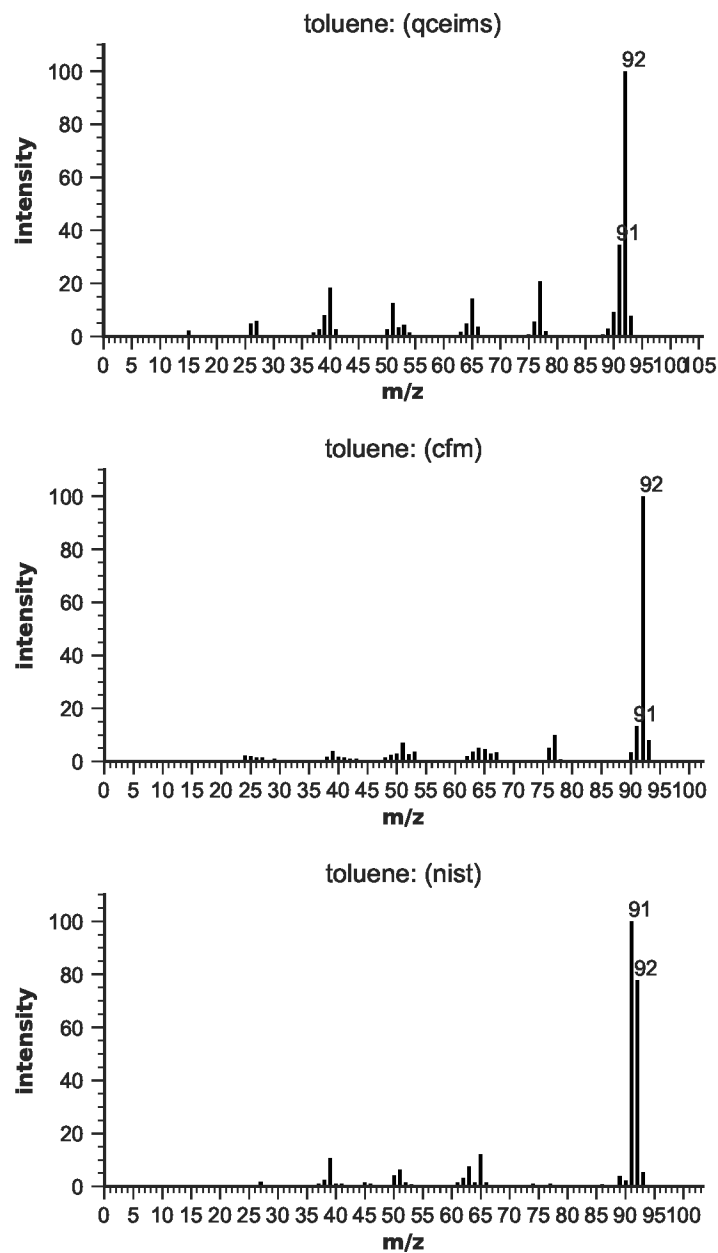


Figure S47: EIMS for toluene

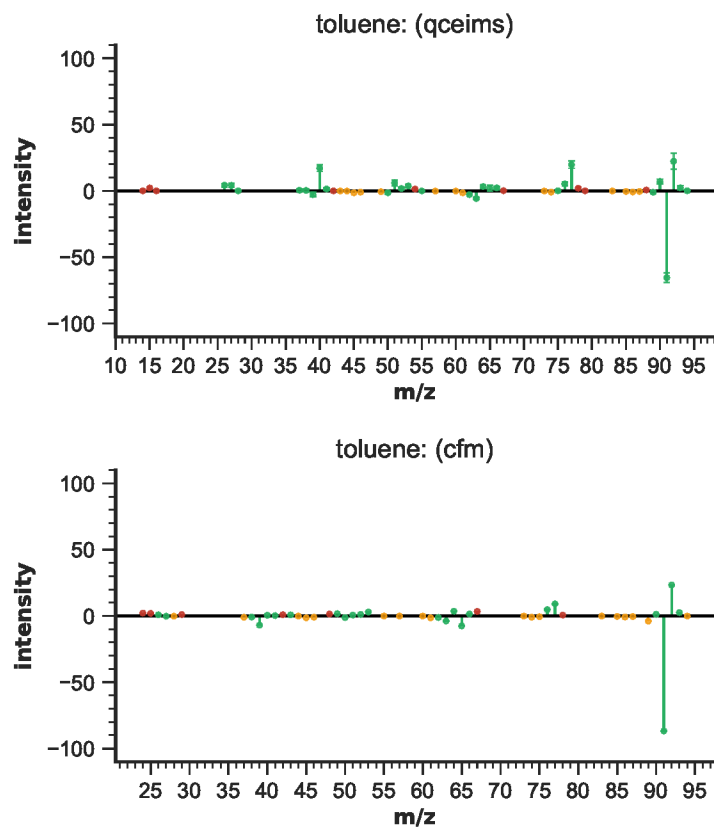


Figure S48: EIMS differences for toluene

2.5.3 25: ethylbenzene

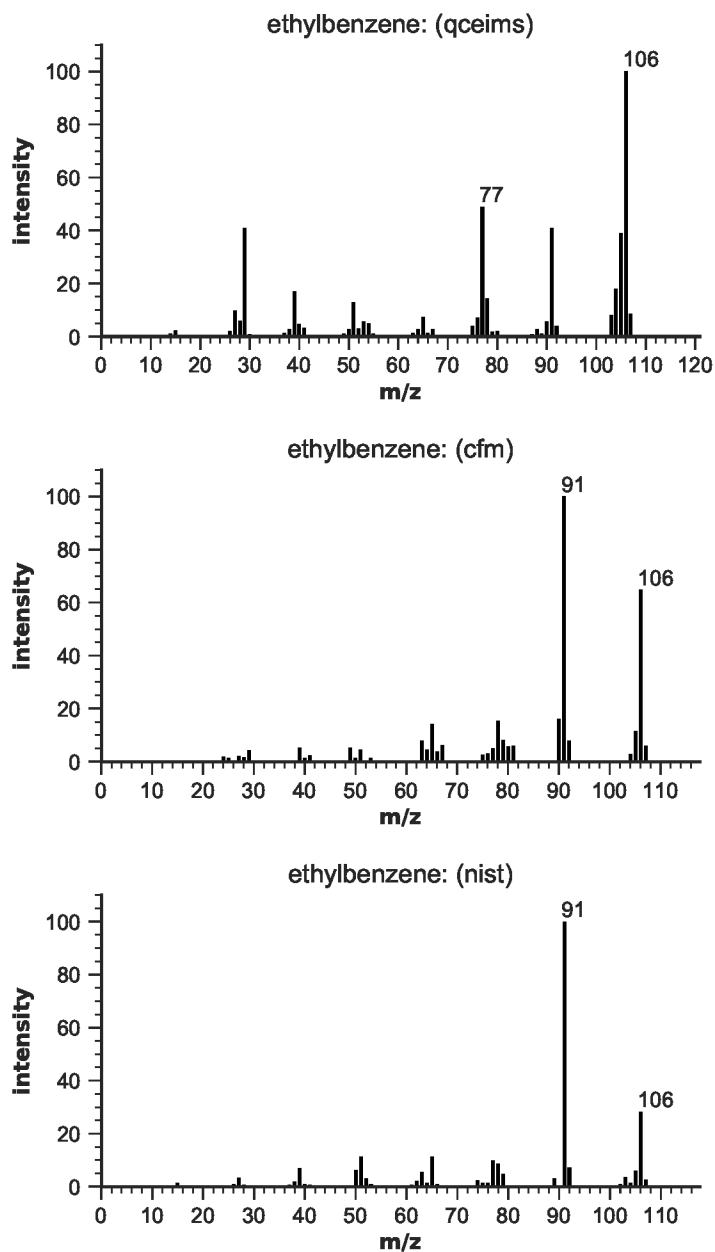


Figure S49: EIMS for ethylbenzene

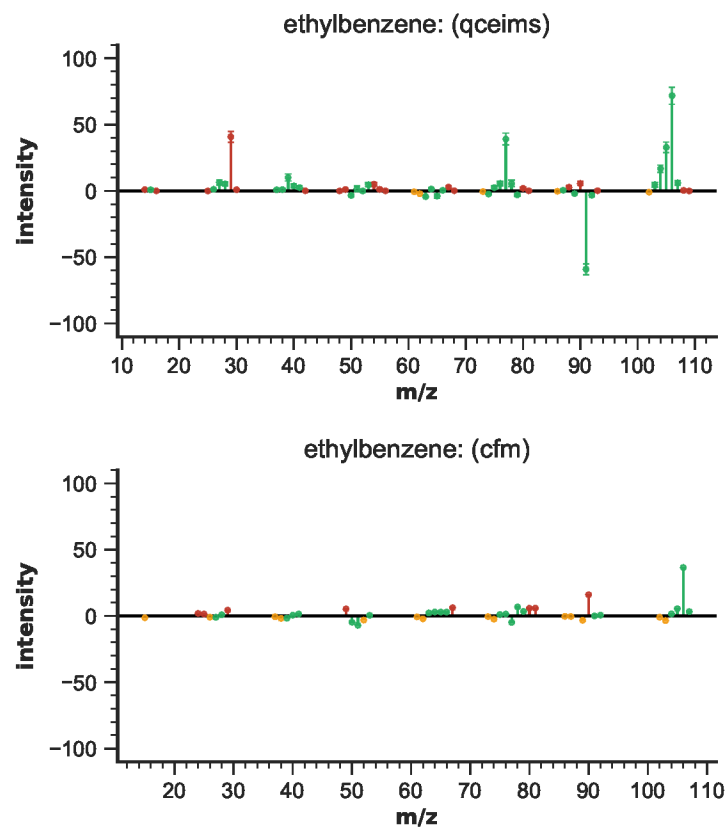


Figure S50: EIMS differences for ethylbenzene

2.5.4 26: propylbenzene

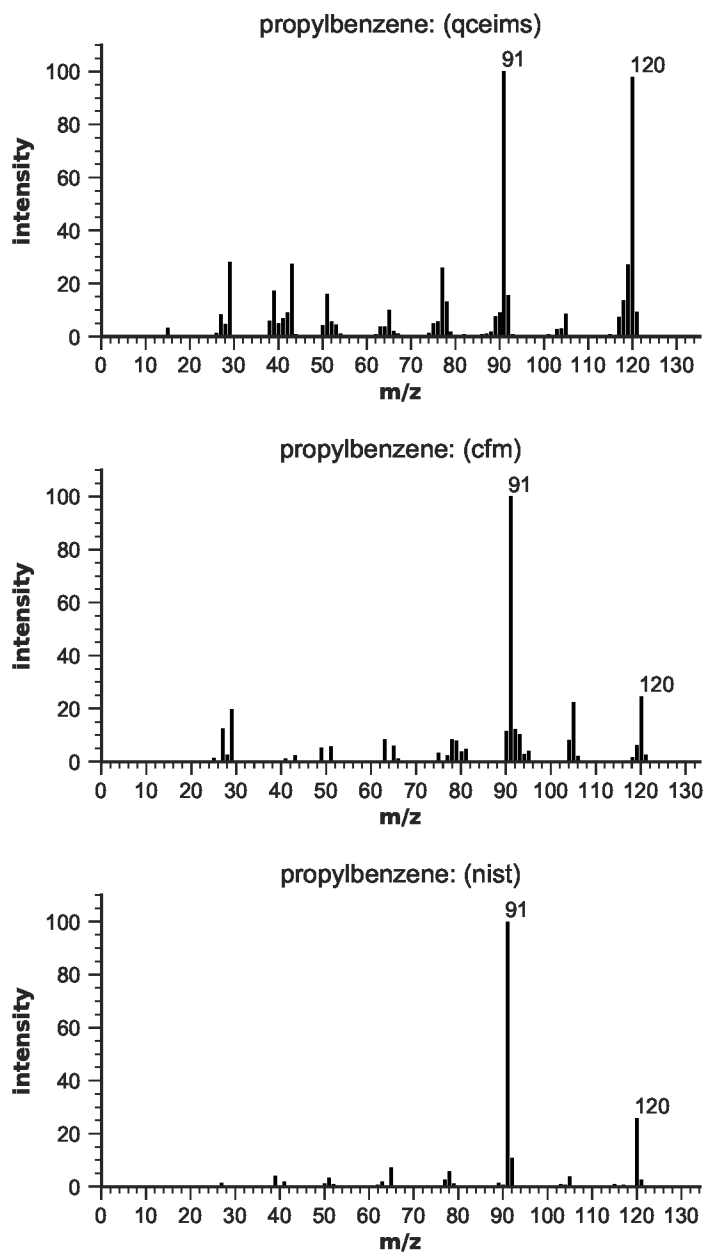


Figure S51: EIMS for propylbenzene

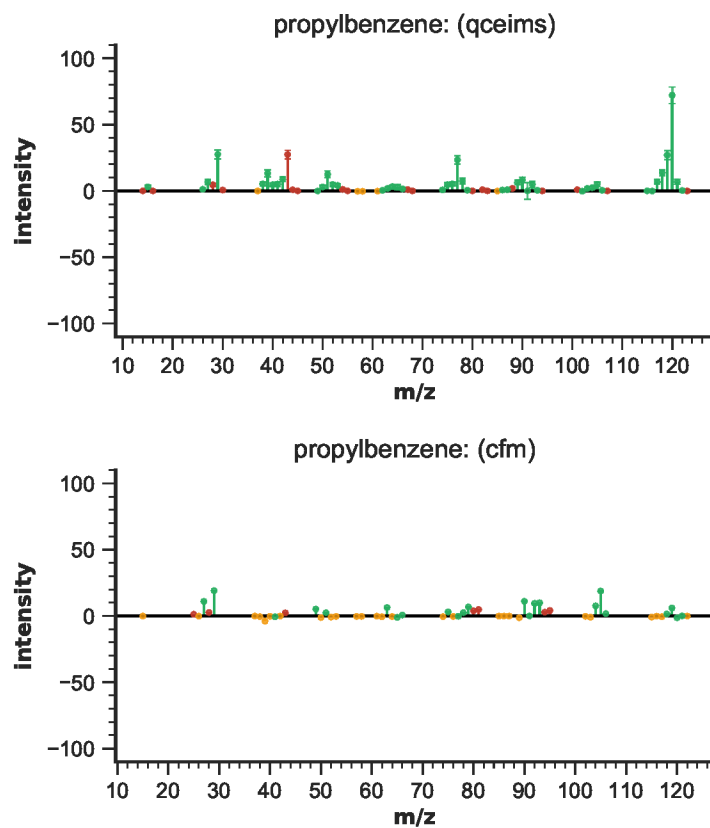


Figure S52: EIMS differences for propylbenzene

2.5.5 27: cumene

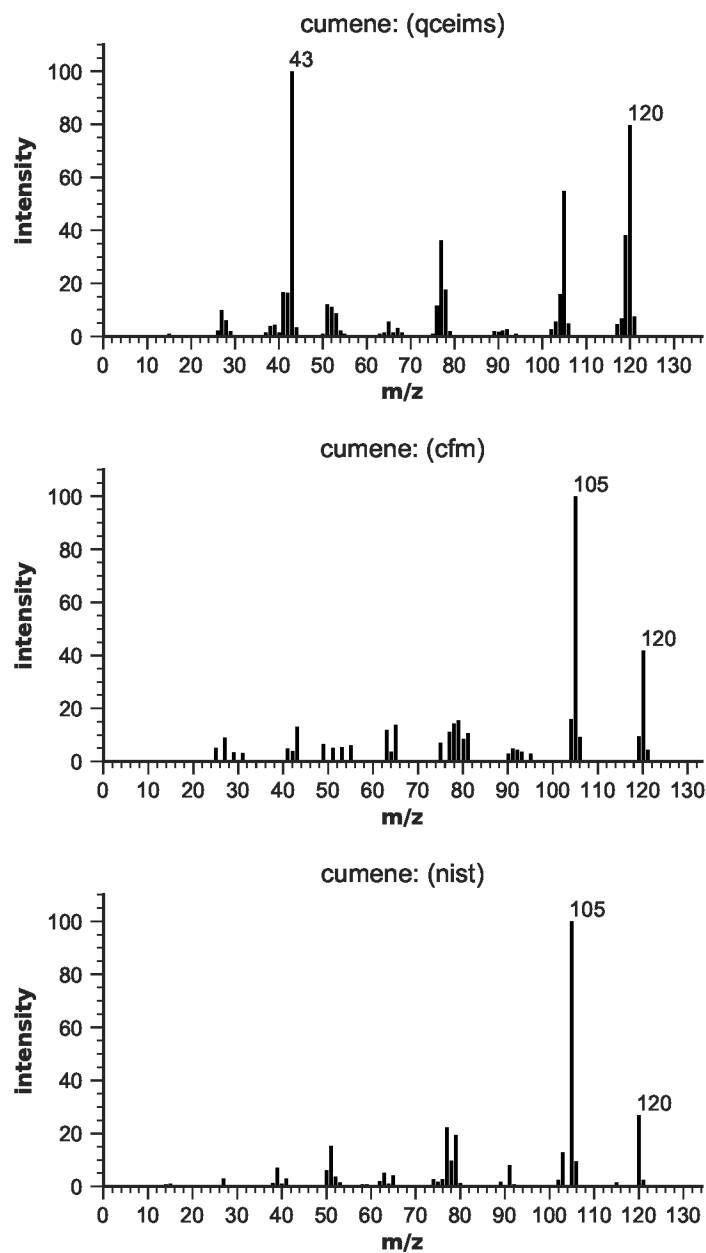


Figure S53: EIMS for cumene

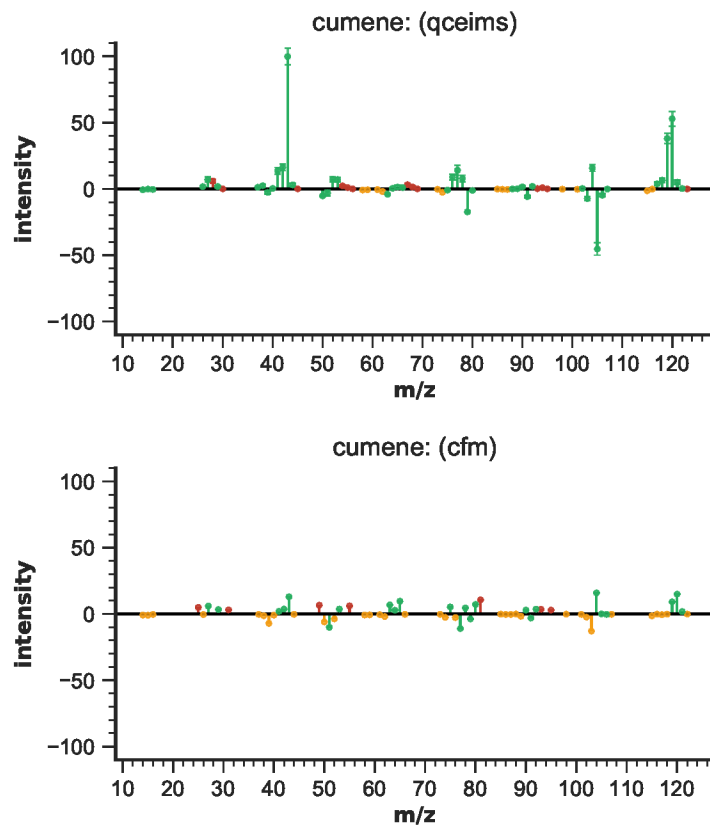


Figure S54: EIMS differences for cumene

2.5.6 28: butylbenzene

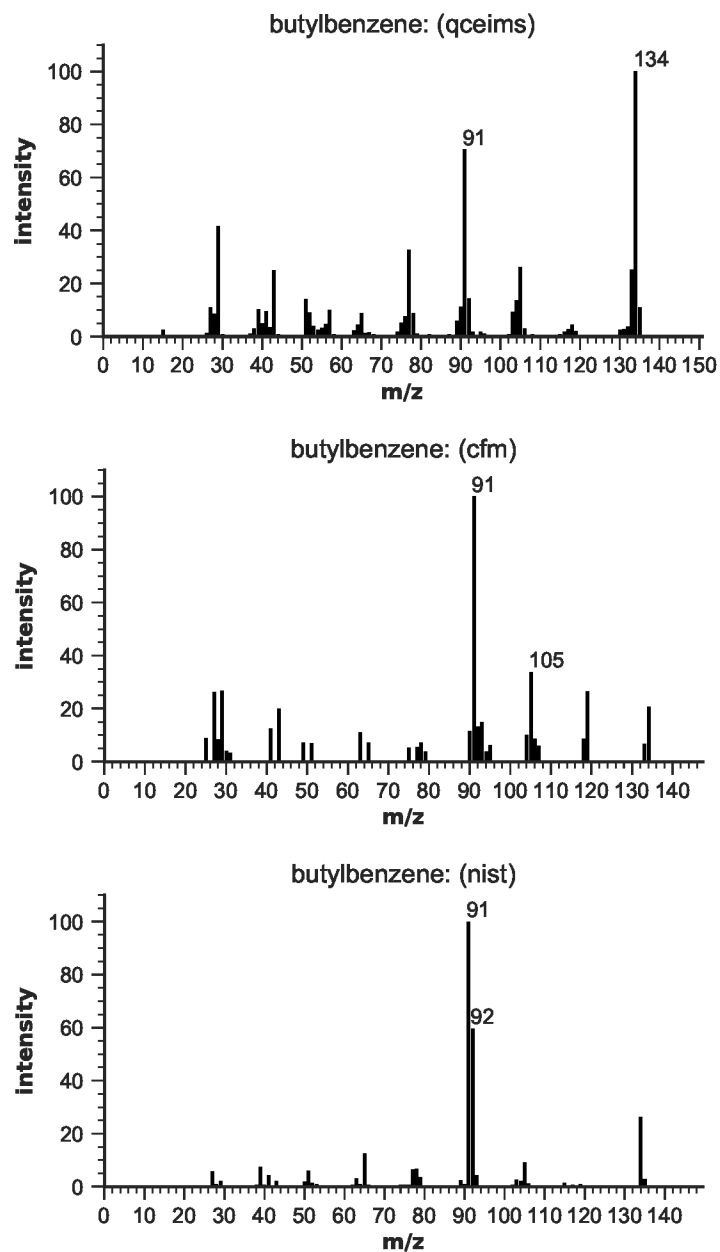


Figure S55: EIMS for butylbenzene

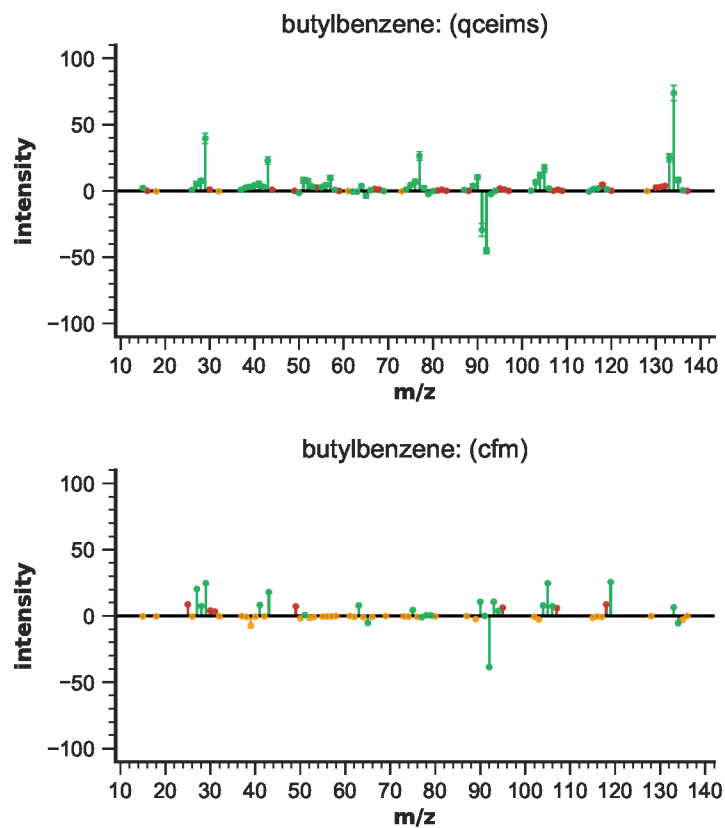


Figure S56: EIMS differences for butylbenzene

2.5.7 29: butan-2-ylbenzene

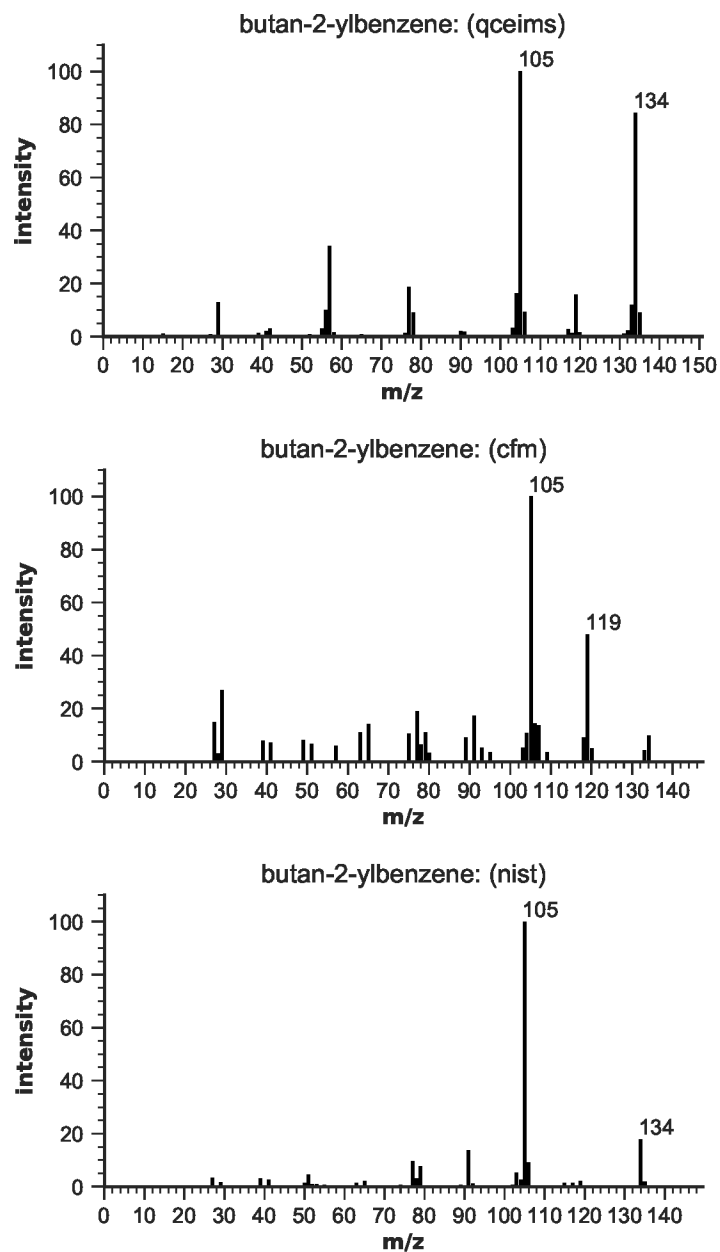


Figure S57: EIMS for butan-2-ylbenzene

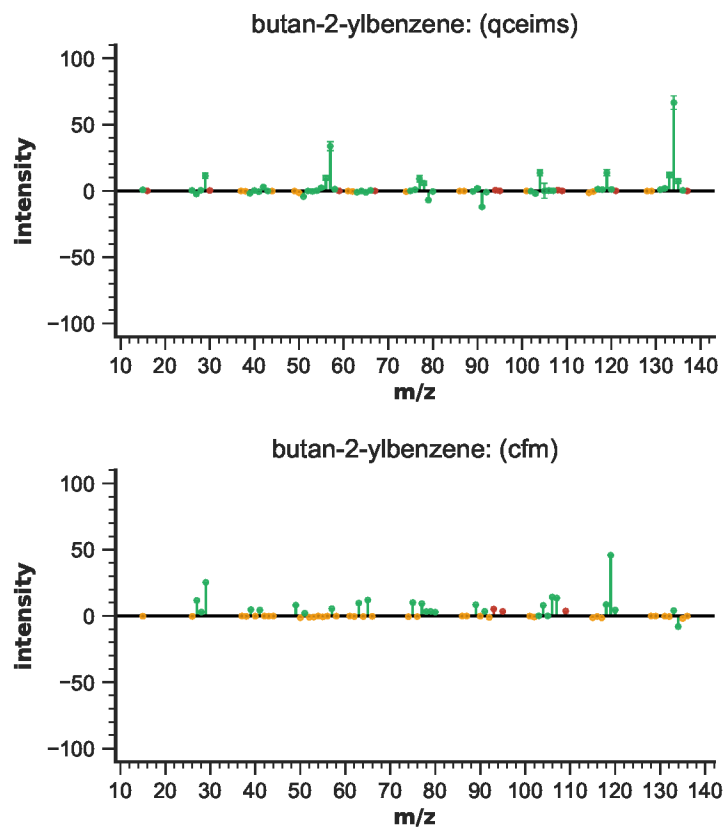


Figure S58: EIMS differences for butan-2-ylbenzene

2.5.8 30: 2-methylpropylbenzene

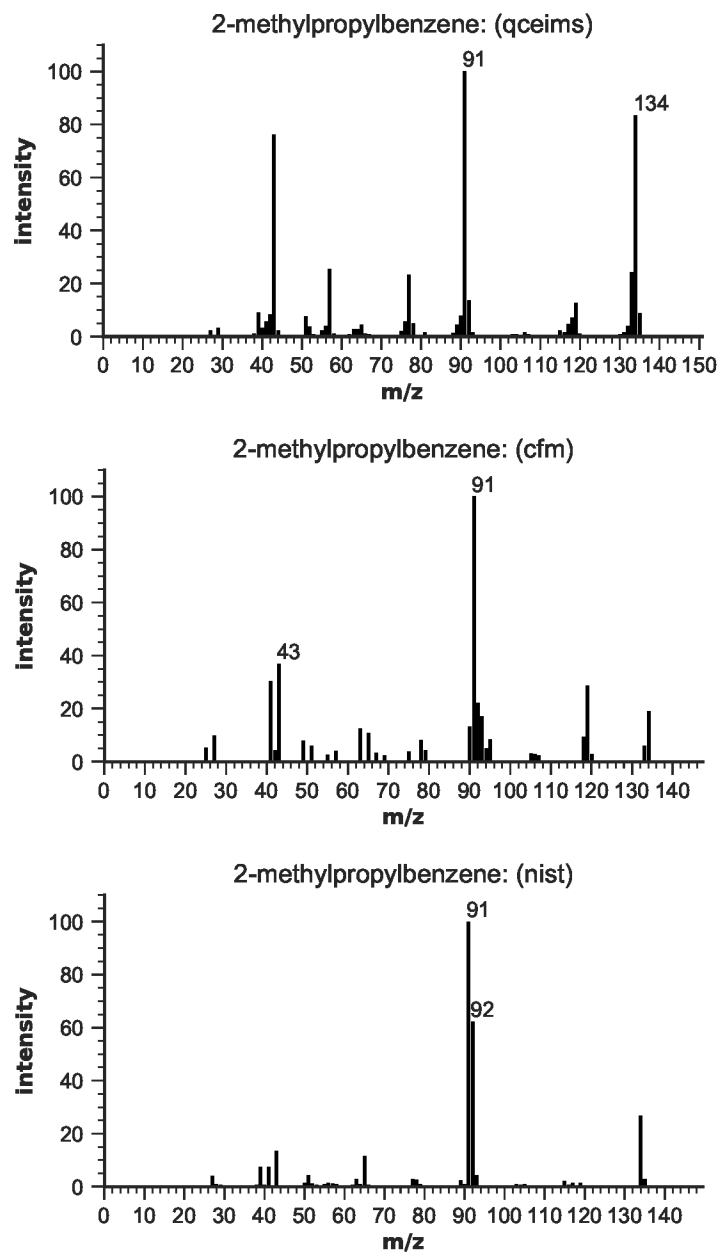


Figure S59: EIMS for 2-methylpropylbenzene

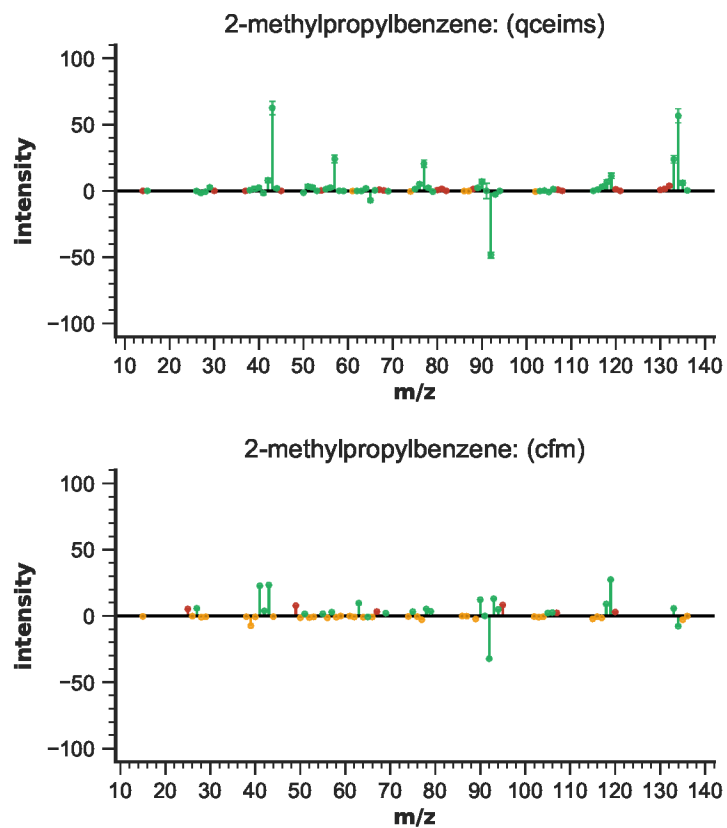


Figure S60: EIMS differences for 2-methylpropylbenzene

2.6 Carboxylic acids

2.6.1 31: propanoic acid

Accepted Article

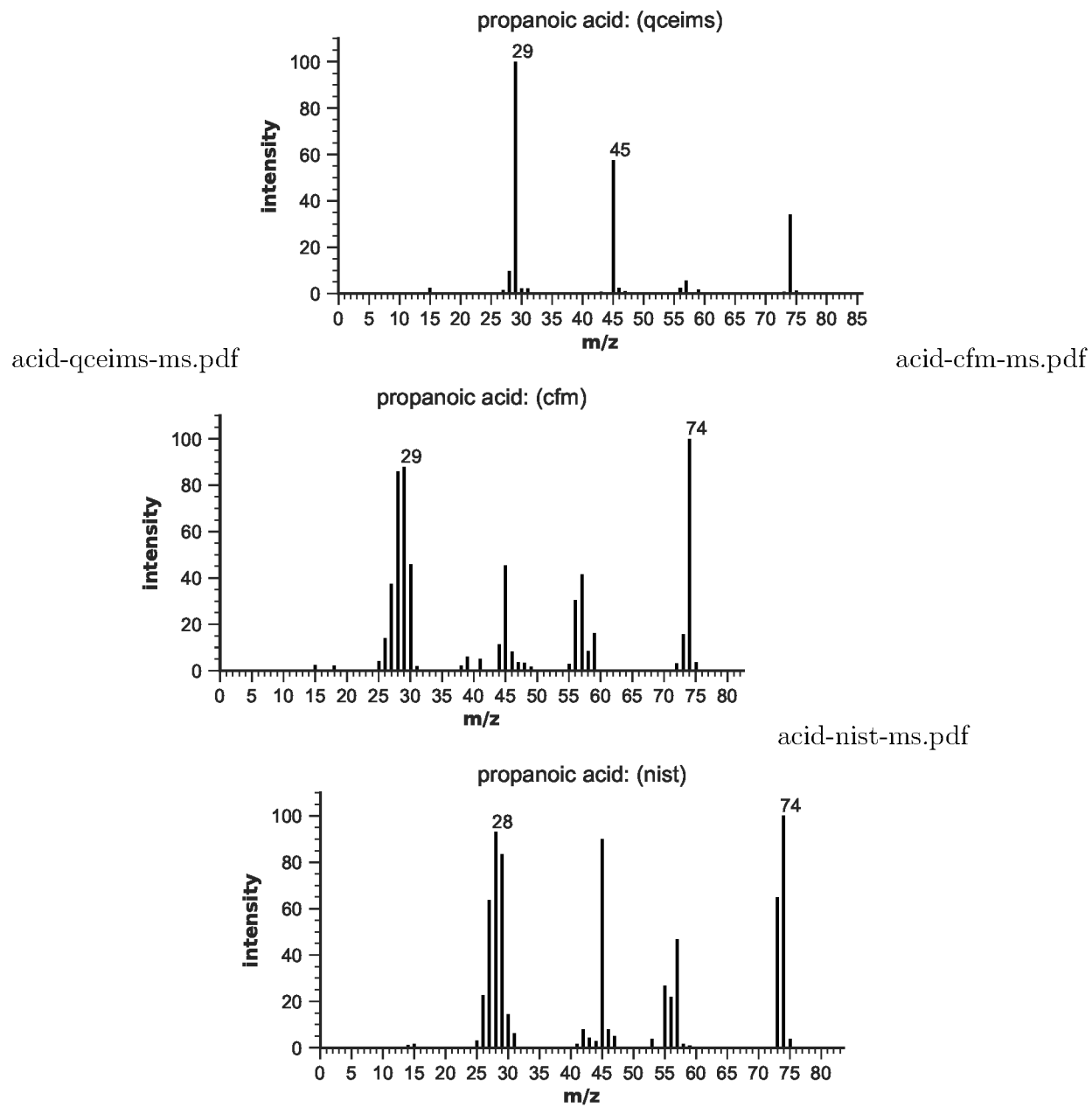


Figure S61: EIMS for propanoic acid

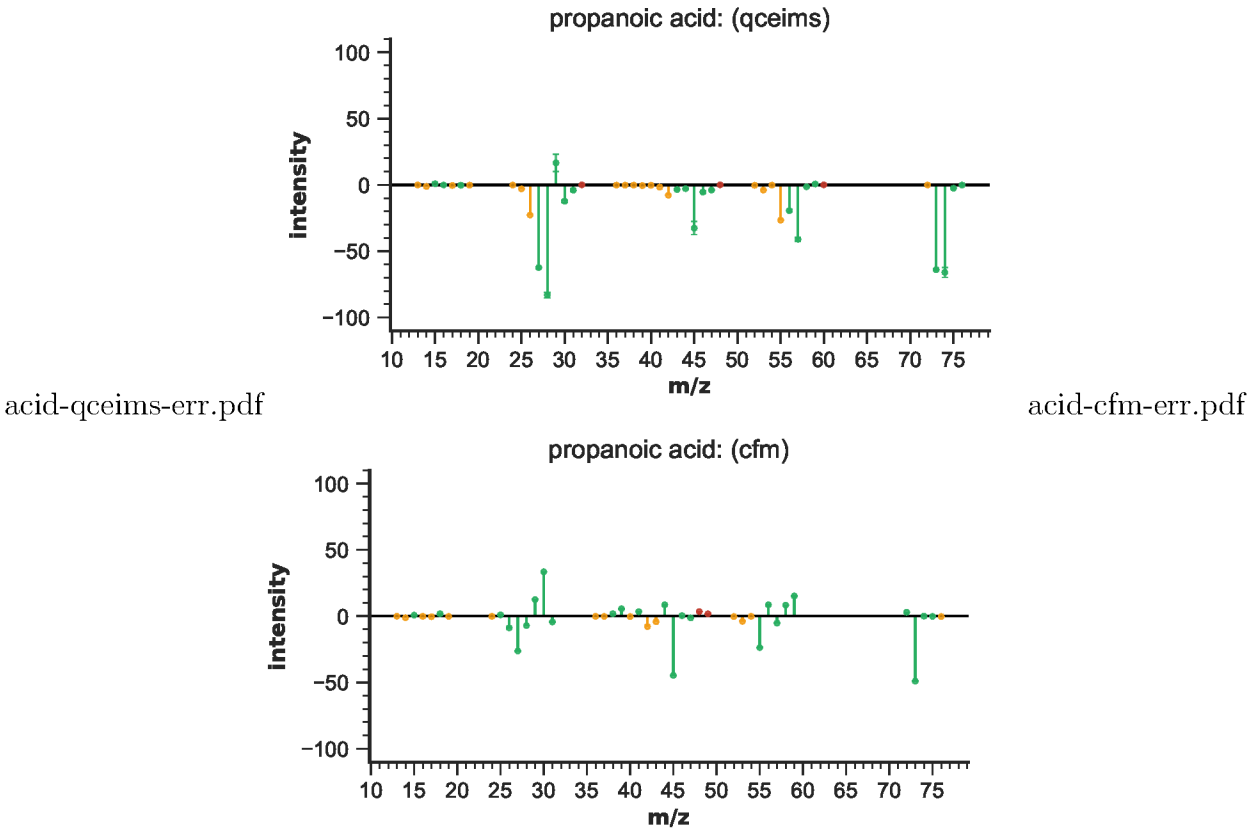


Figure S62: EIMS differences for propanoic acid

2.6.2 32: butanoic acid

Accepted Article

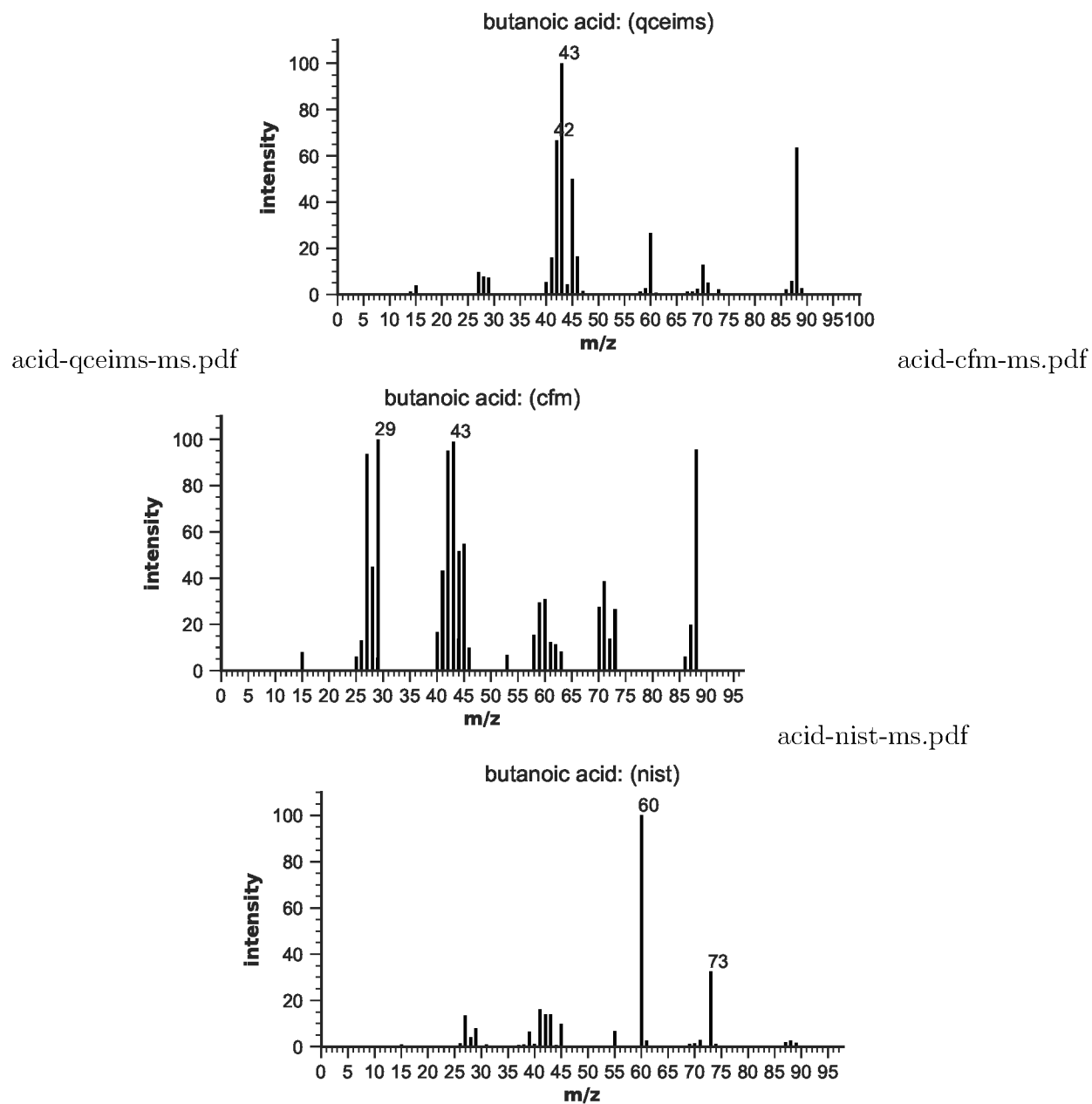
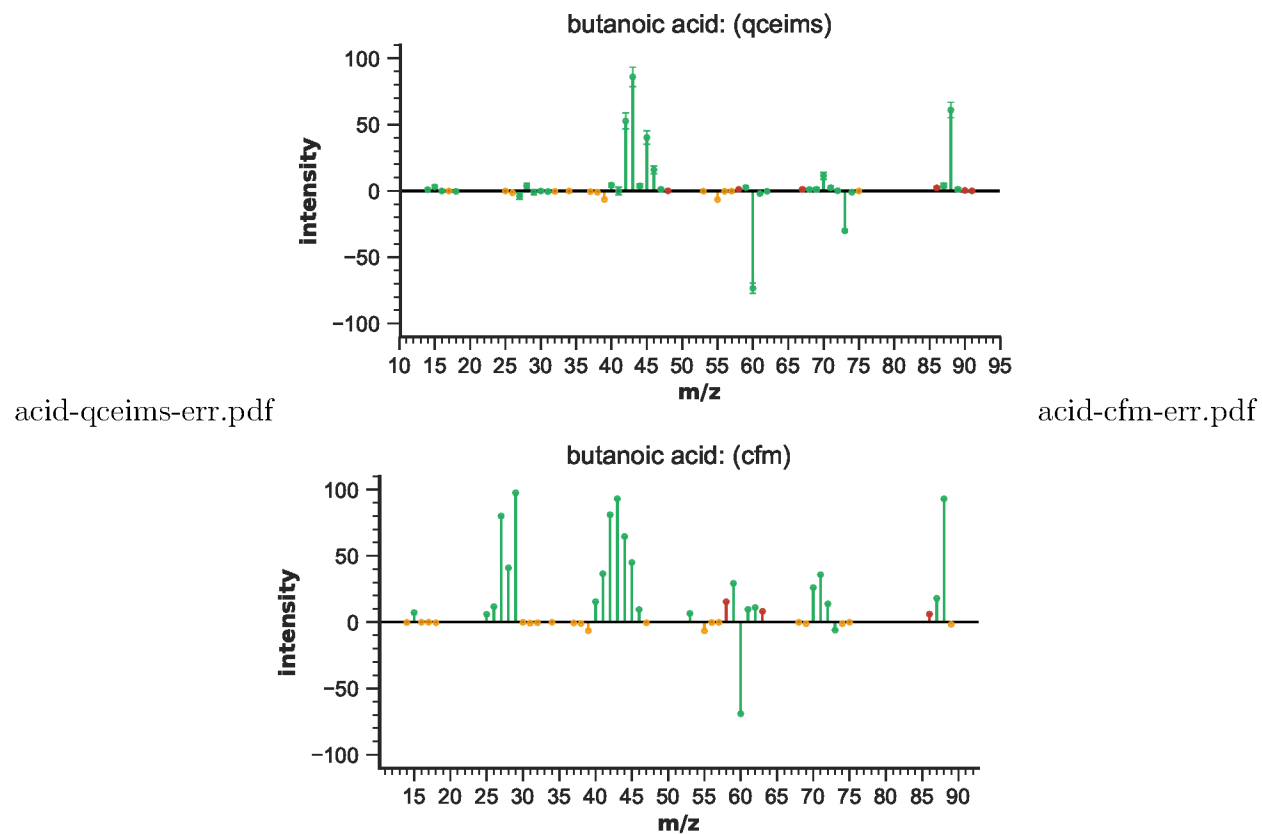


Figure S63: EIMS for butanoic acid



2.6.3 33: 2-methylpropanoic acid

Accepted Article

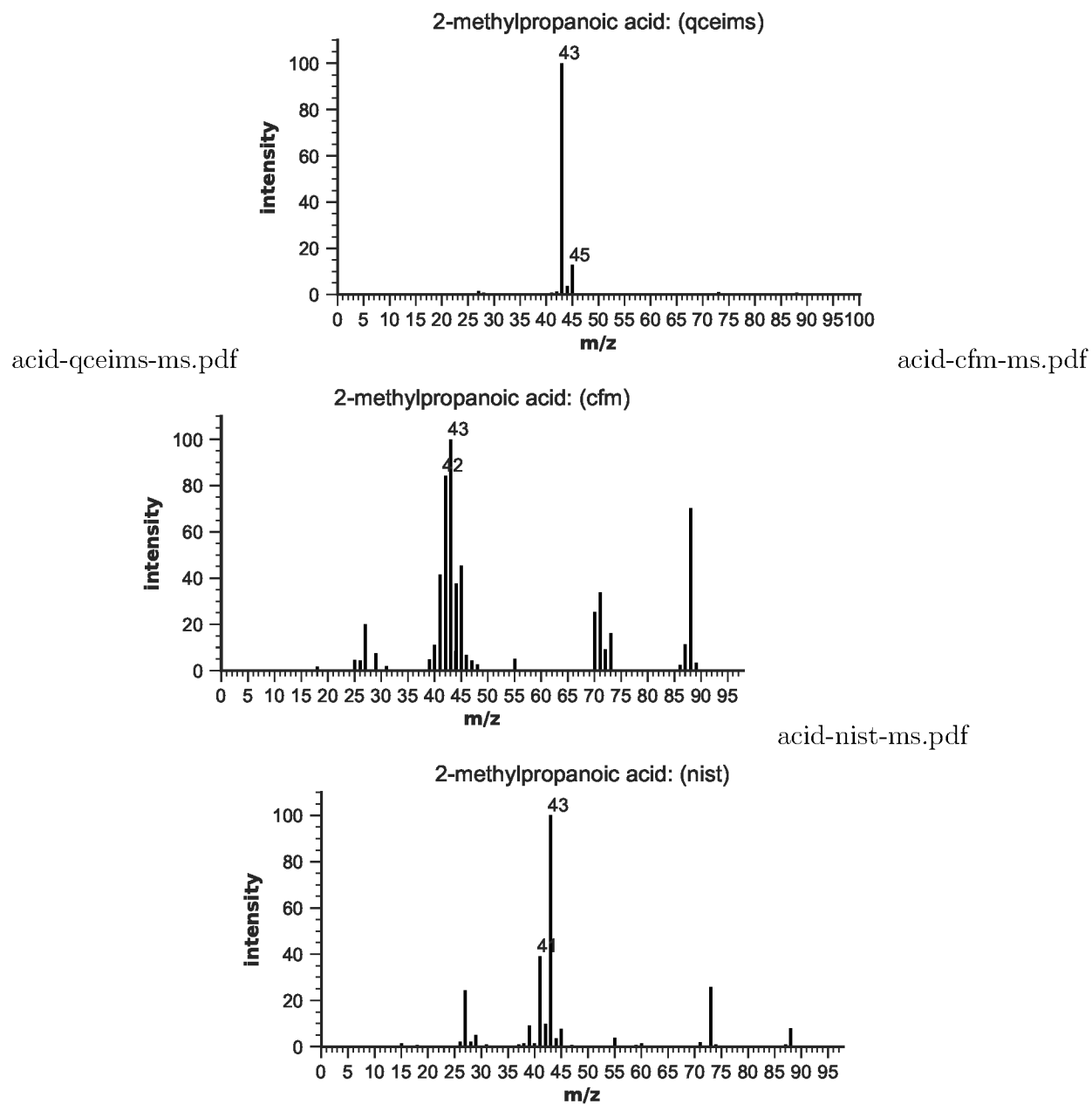


Figure S65: EIMS for 2-methylpropanoic acid

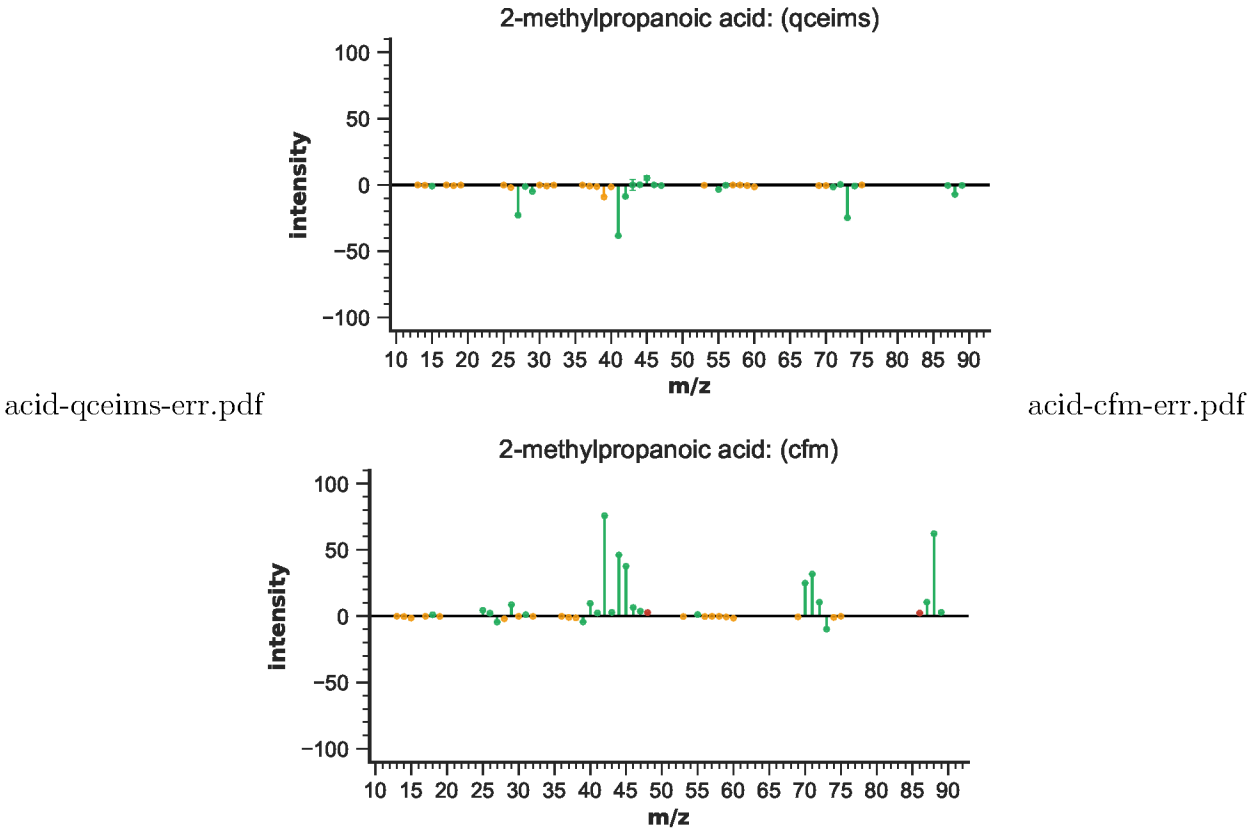


Figure S66: EIMS differences for 2-methylpropanoic acid

2.6.4 34: 2-methylbutanoic acid

Accepted Article

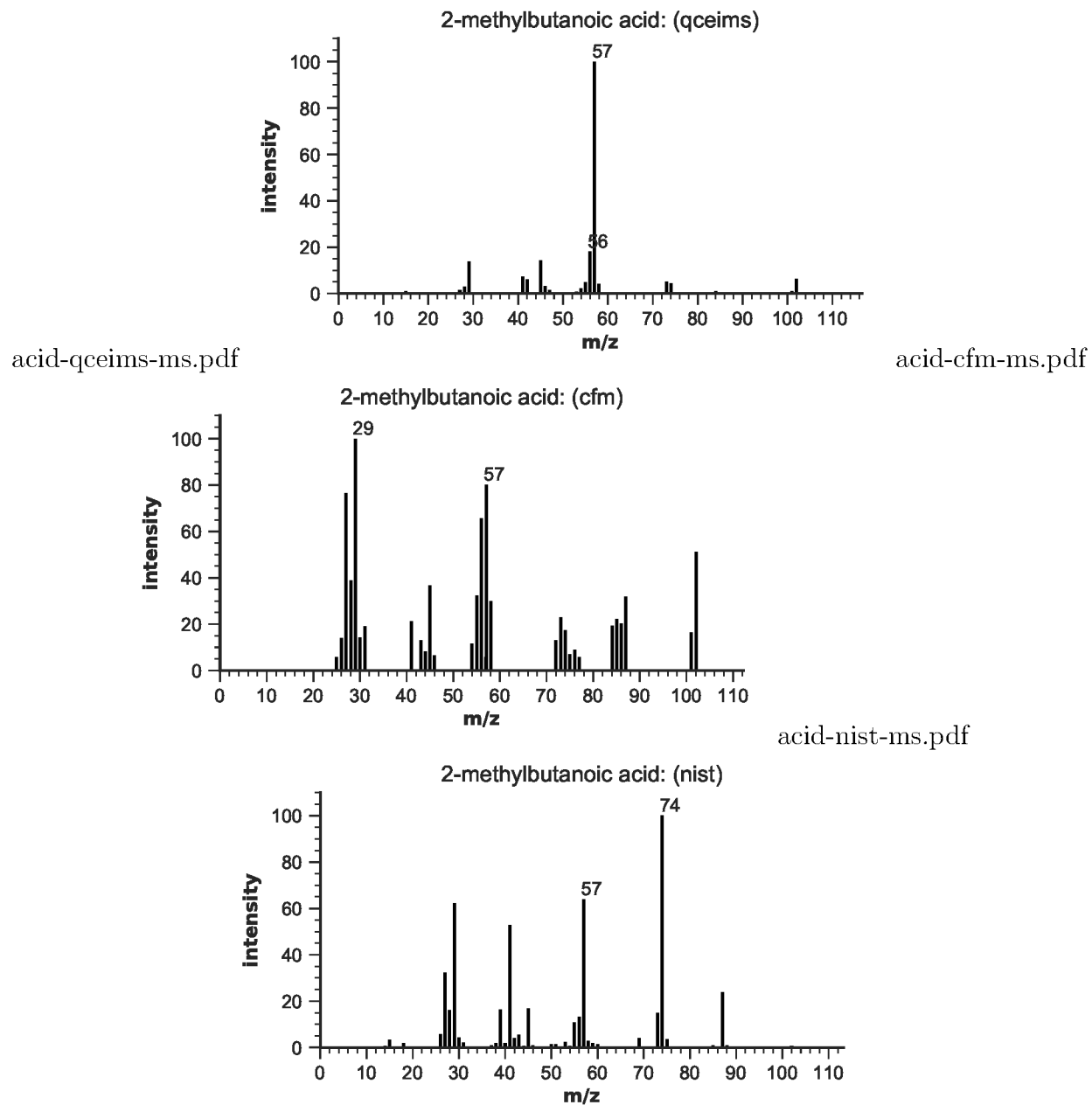


Figure S67: EIMS for 2-methylbutanoic acid

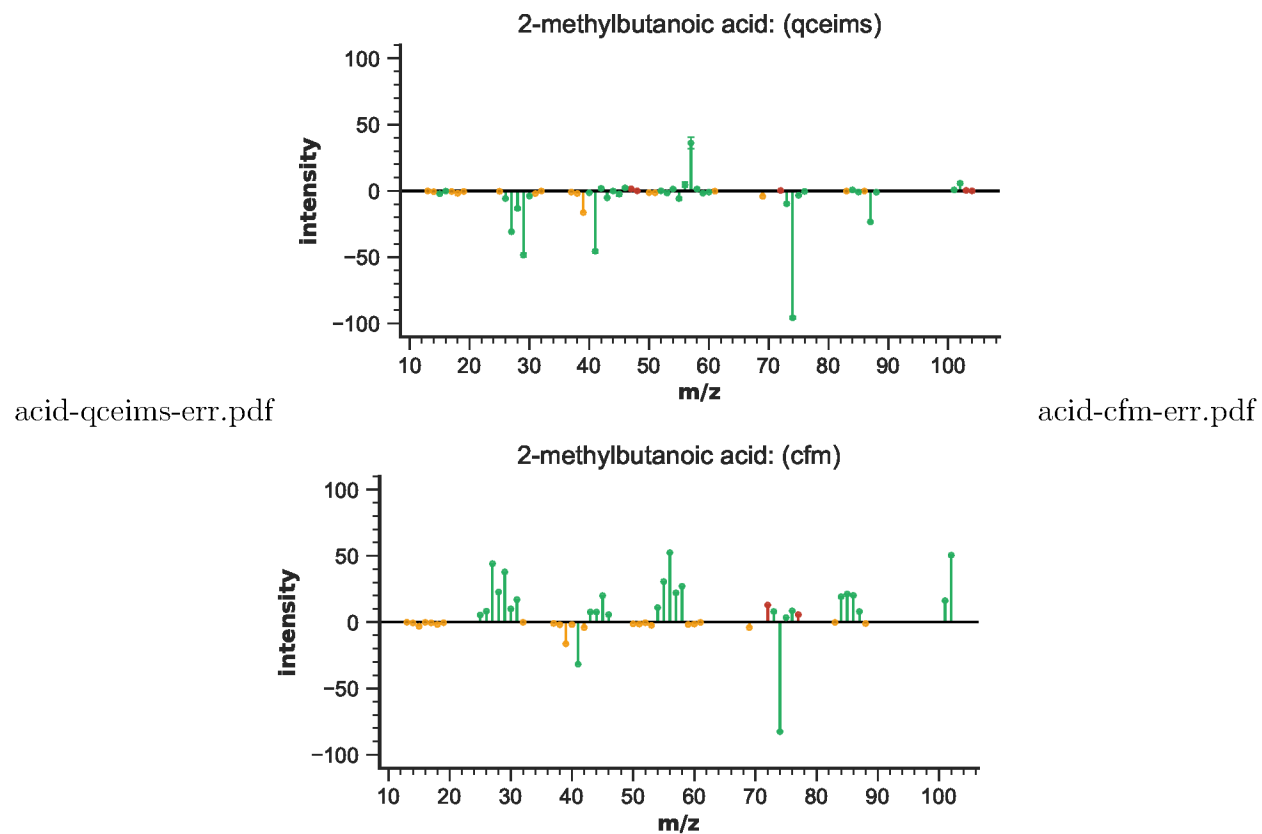


Figure S68: EIMS differences for 2-methylbutanoic acid

2.6.5 35: 3-methylbutanoic acid

Accepted Article

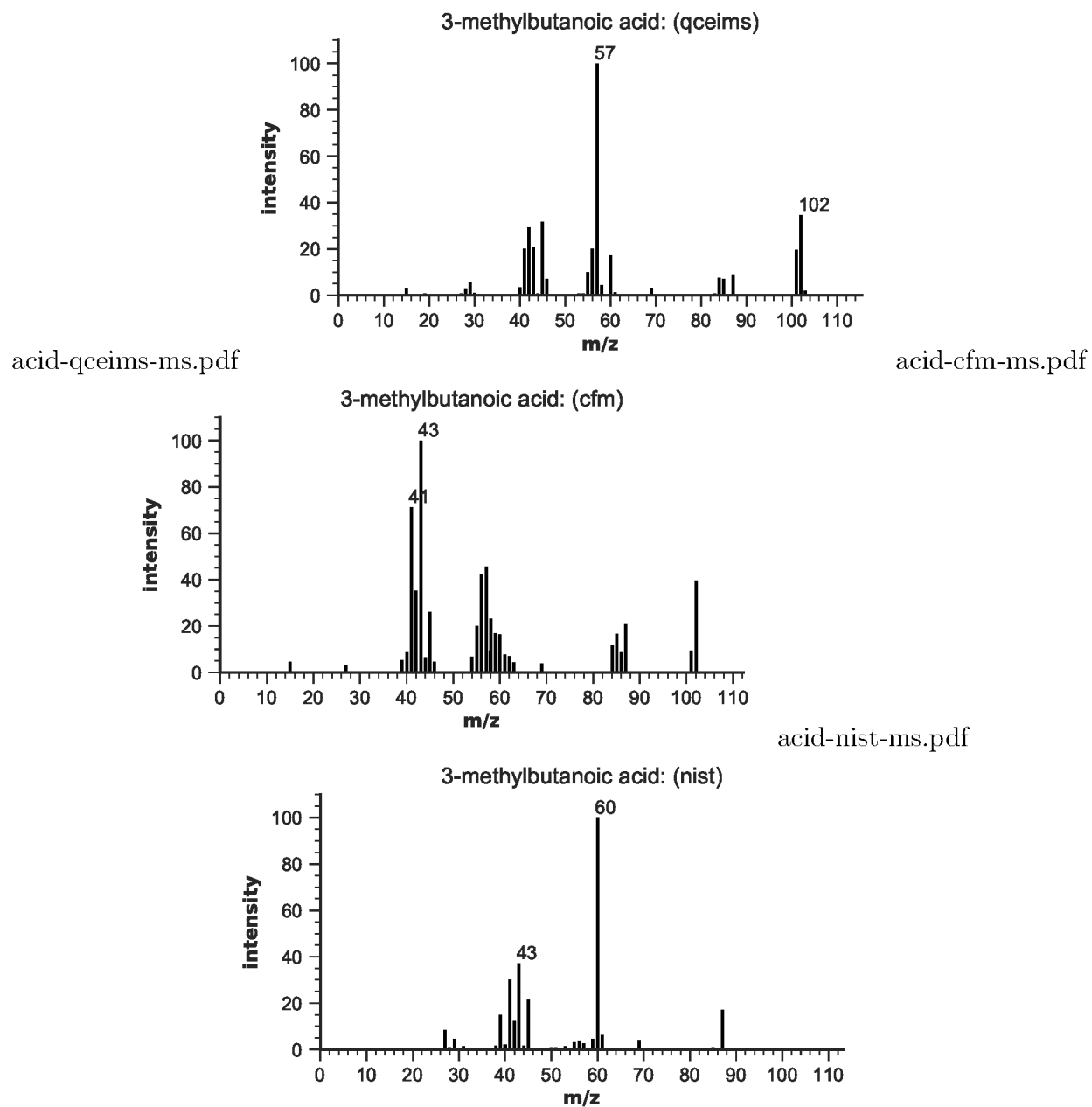


Figure S69: EIMS for 3-methylbutanoic acid

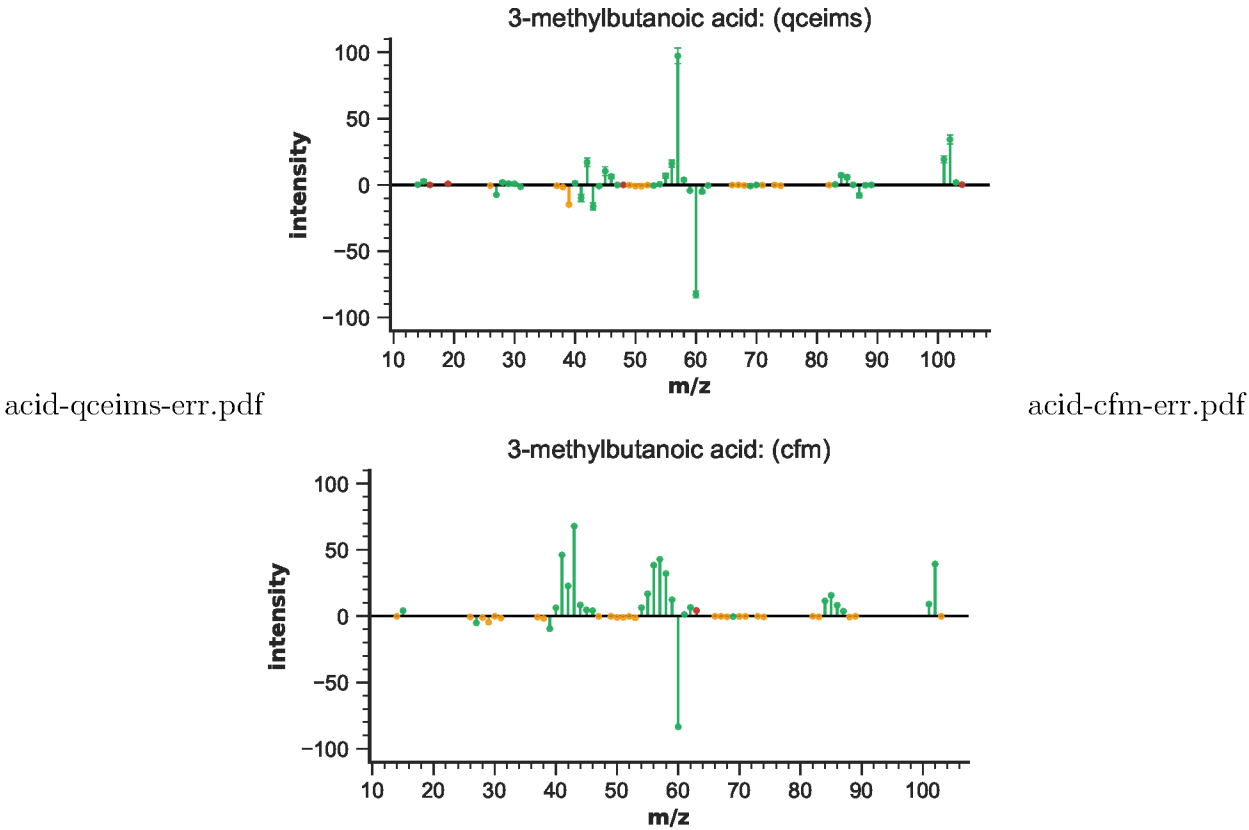


Figure S70: EIMS differences for 3-methylbutanoic acid

2.7 Esters

2.7.1 36: methyl formate

Accepted Article

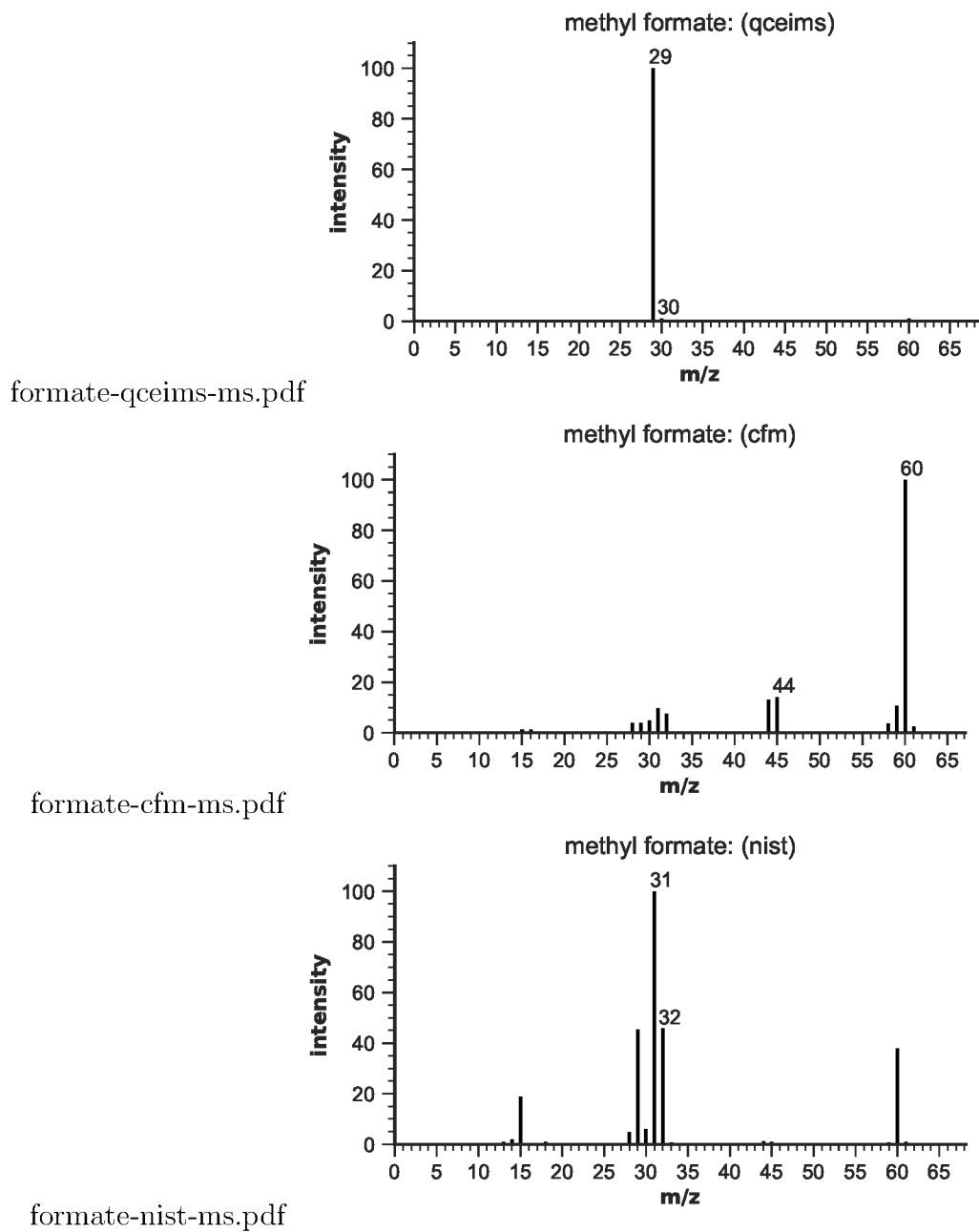


Figure S71: EIMS for methyl formate

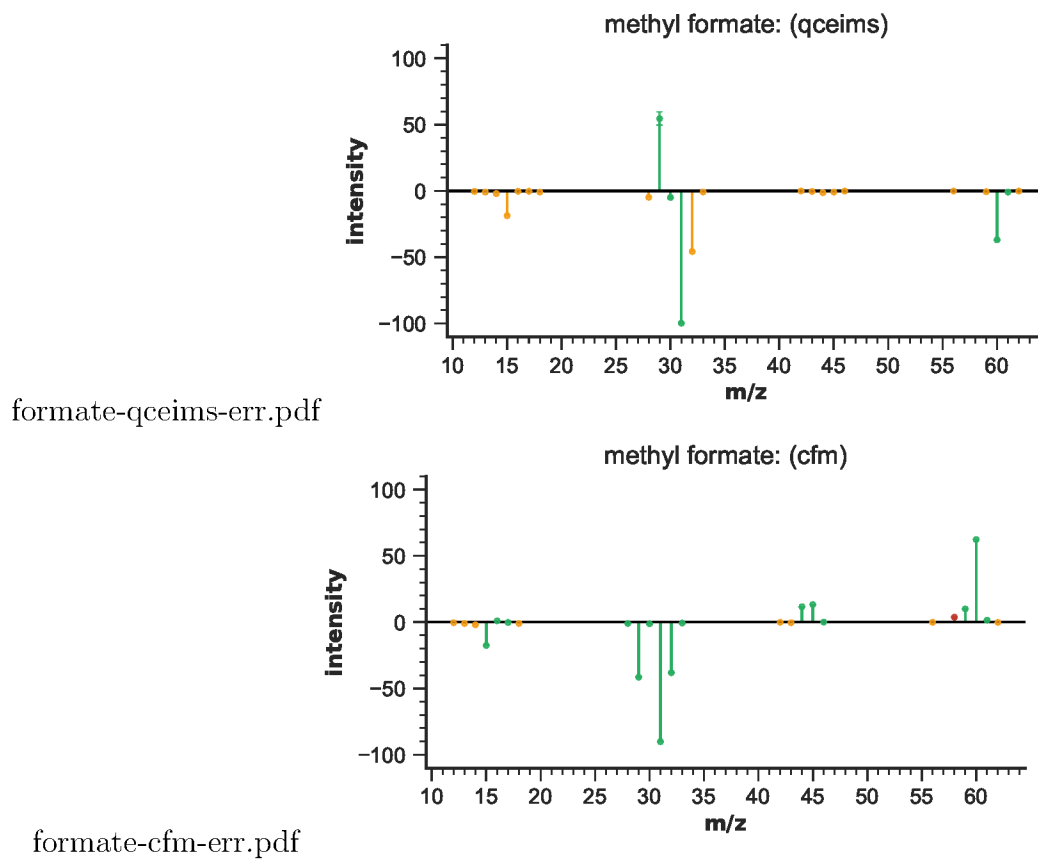
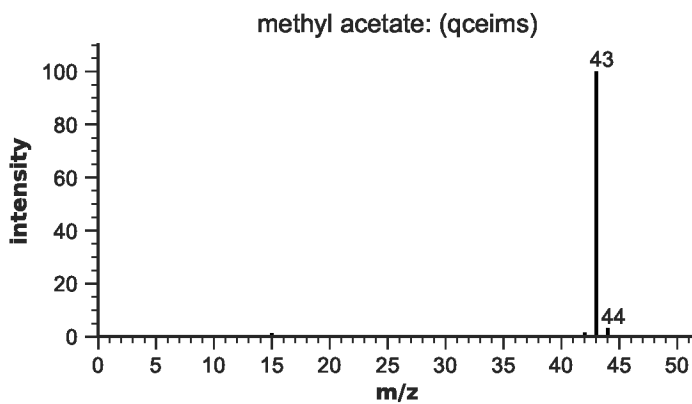


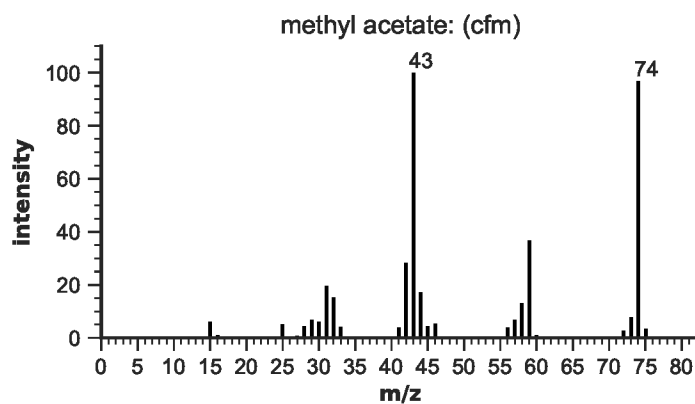
Figure S72: EIMS differences for methyl formate

2.7.2 37: methyl acetate

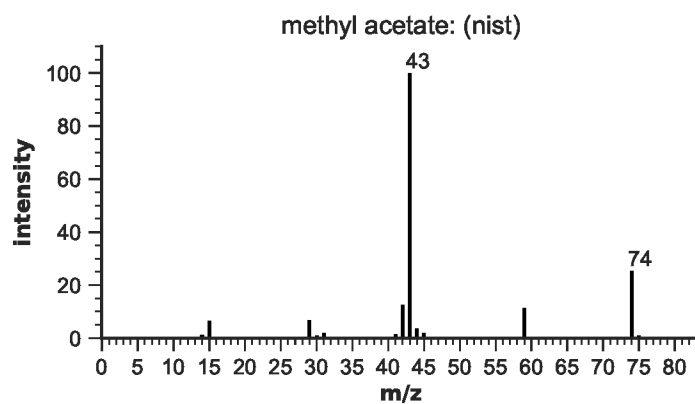
Accepted Article



acetate-qceims-ms.pdf



acetate-cfm-ms.pdf



acetate-nist-ms.pdf

Figure S73: EIMS for methyl acetate

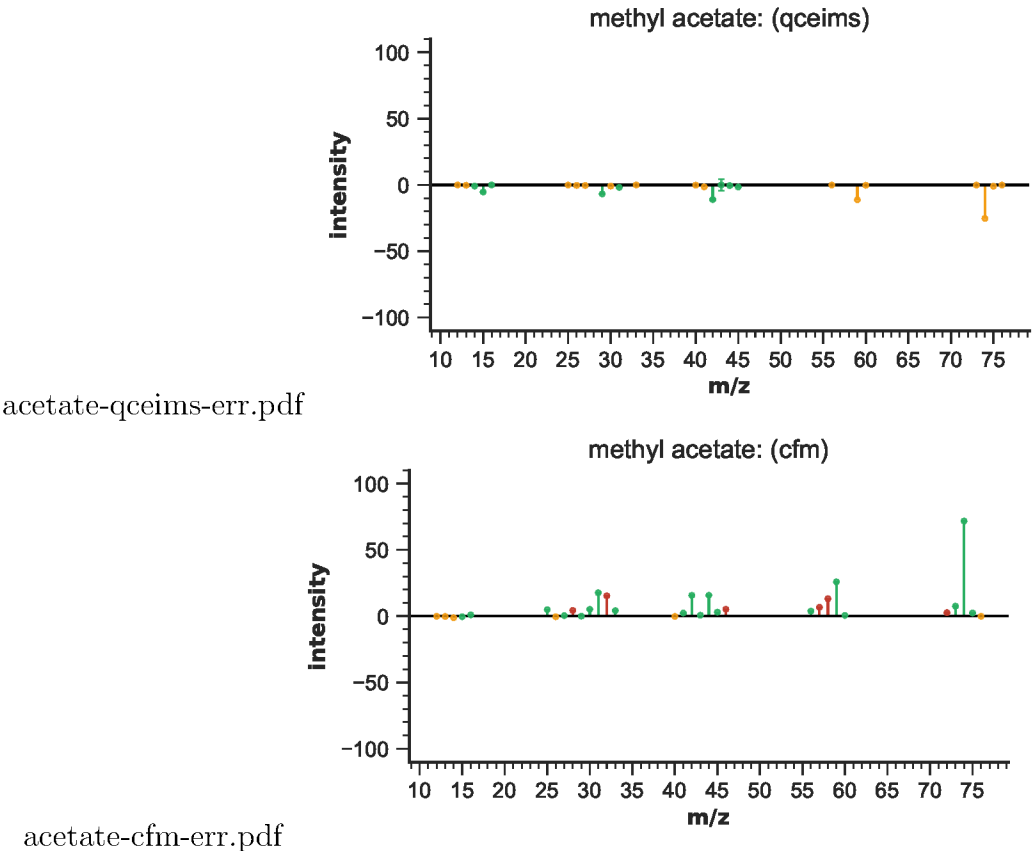
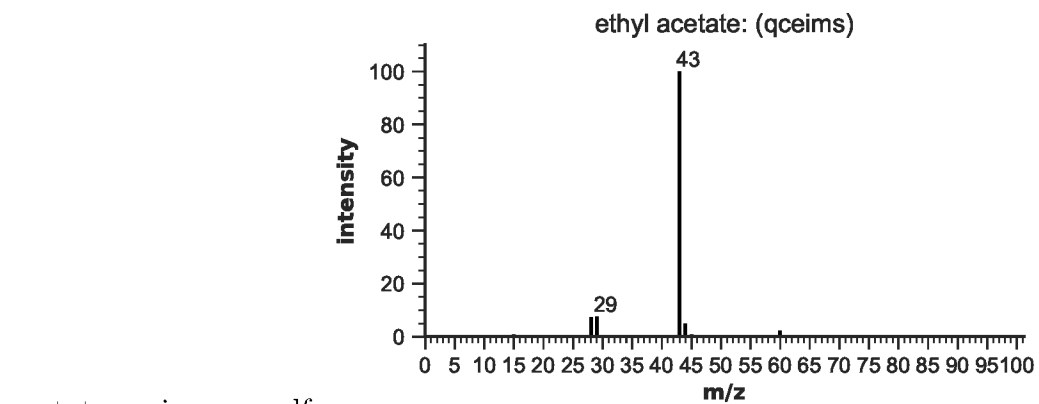


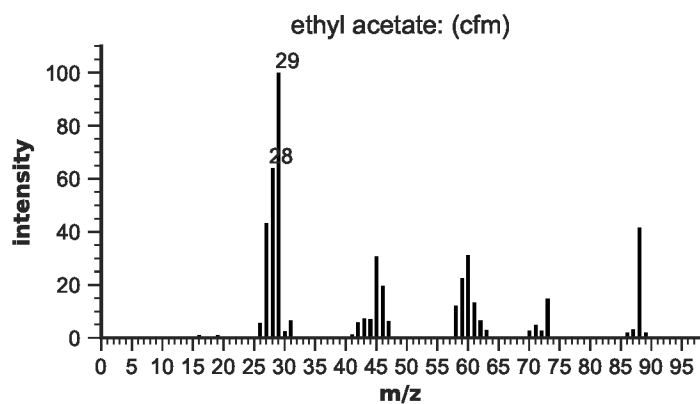
Figure S74: EIMS differences for methyl acetate

2.7.3 38: ethyl acetate

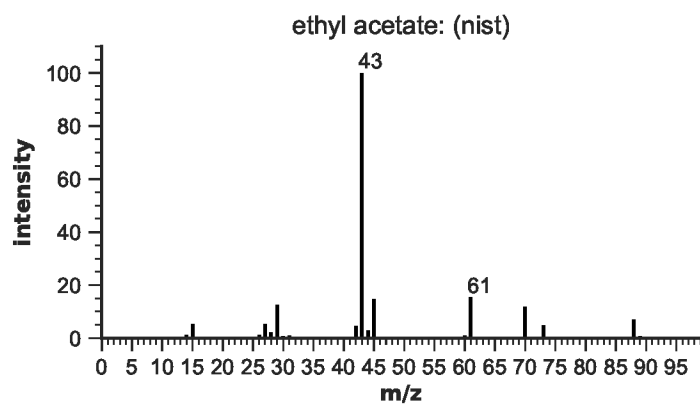
Accepted Article



acetate-qceims-ms.pdf

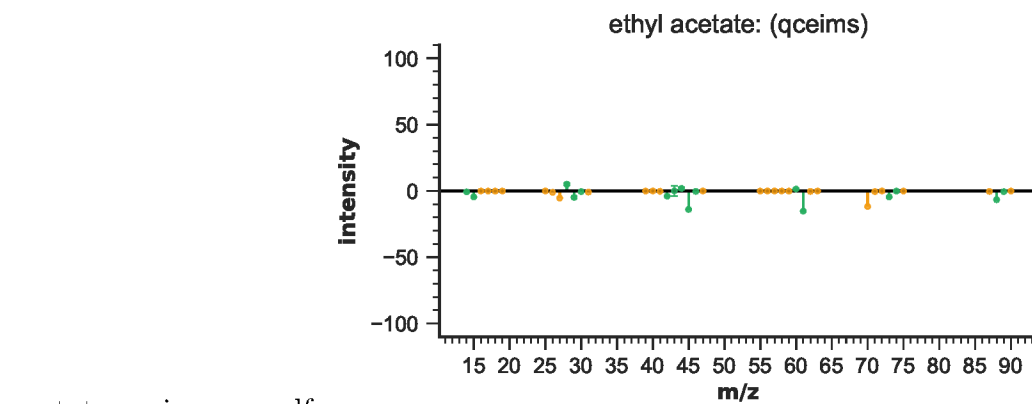


acetate-cfm-ms.pdf

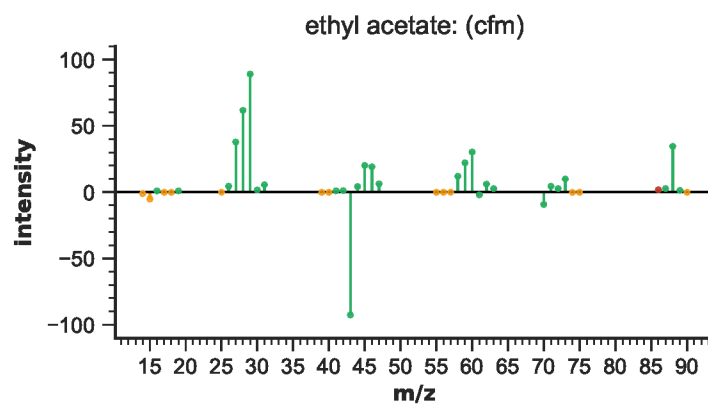


acetate-nist-ms.pdf

Figure S75: EIMS for ethyl acetate



acetate-qceims-err.pdf

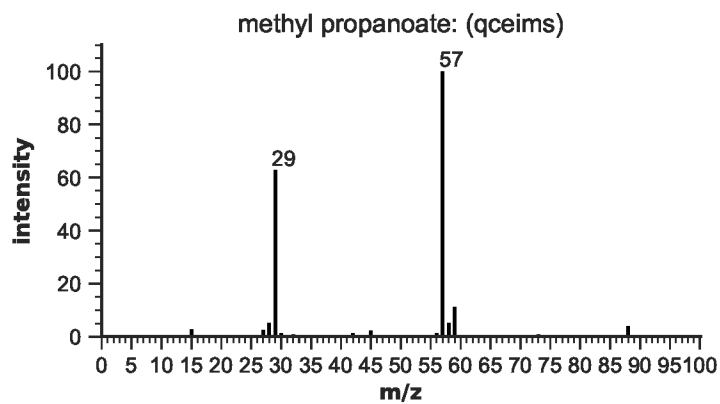


acetate-cfm-err.pdf

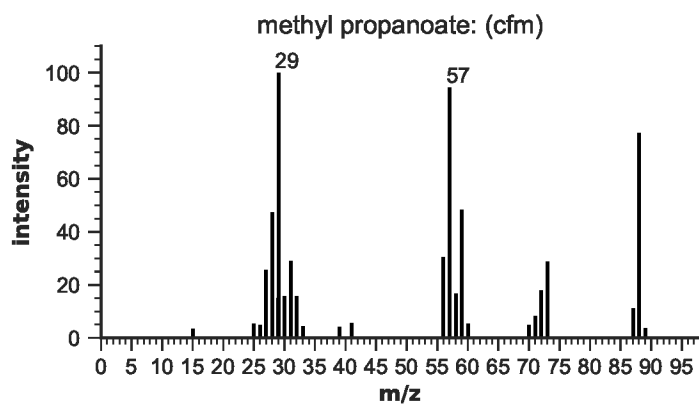
Figure S76: EIMS differences for ethyl acetate

2.7.4 39: methyl propanoate

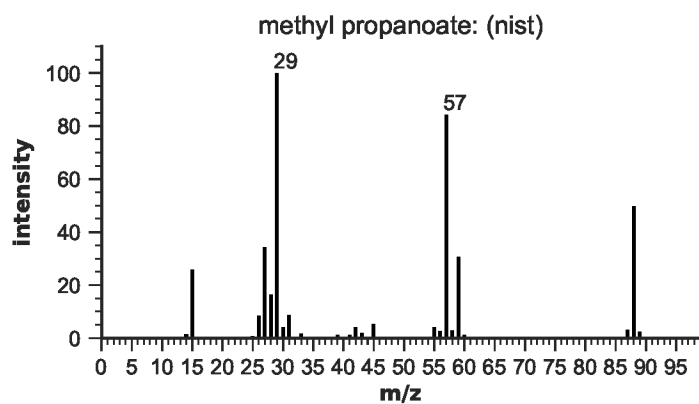
Accepted Article



propanoate-qceims-ms.pdf



propanoate-cfm-ms.pdf



propanoate-nist-ms.pdf

Figure S77: EIMS for methyl propanoate

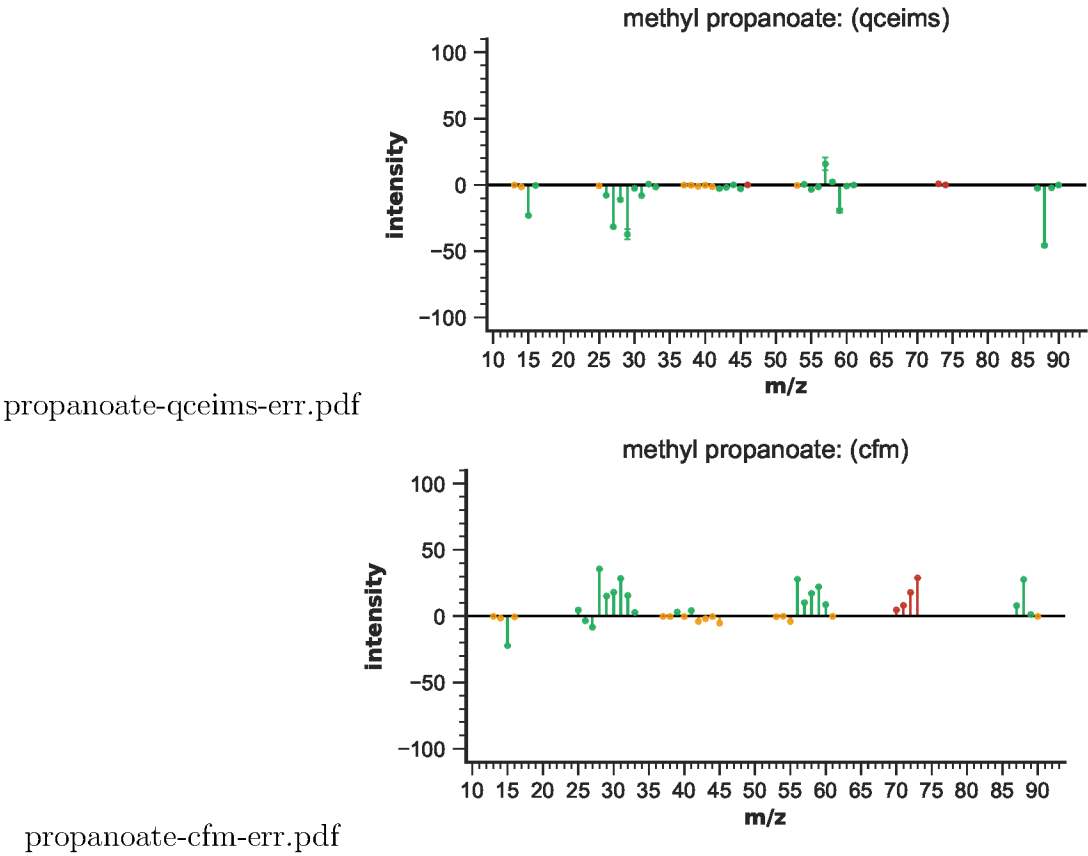
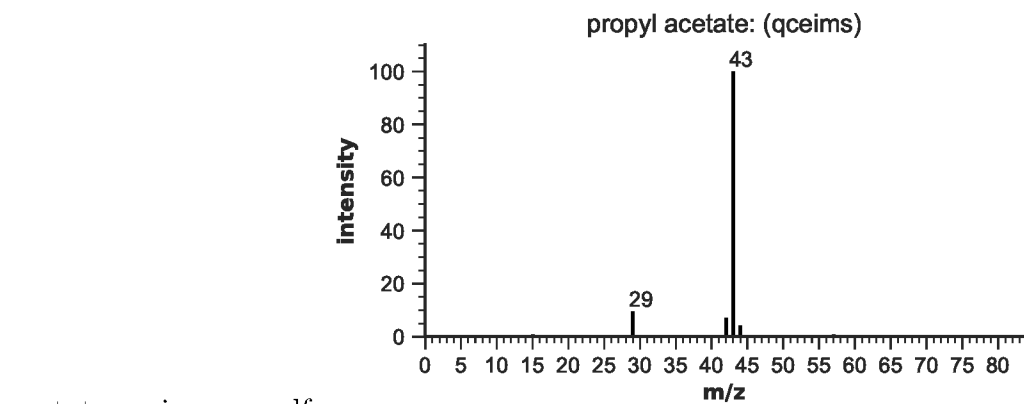


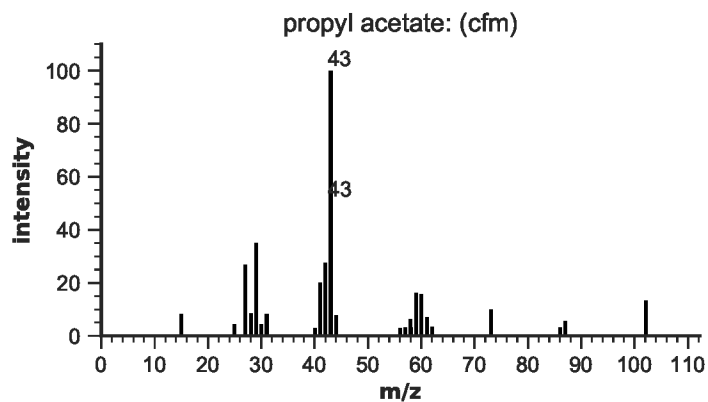
Figure S78: EIMS differences for methyl propanoate

2.7.5 40: propyl acetate

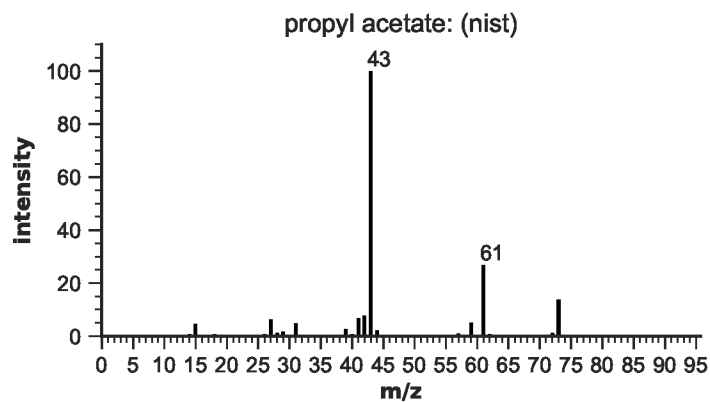
Accepted Article



acetate-qceims-ms.pdf

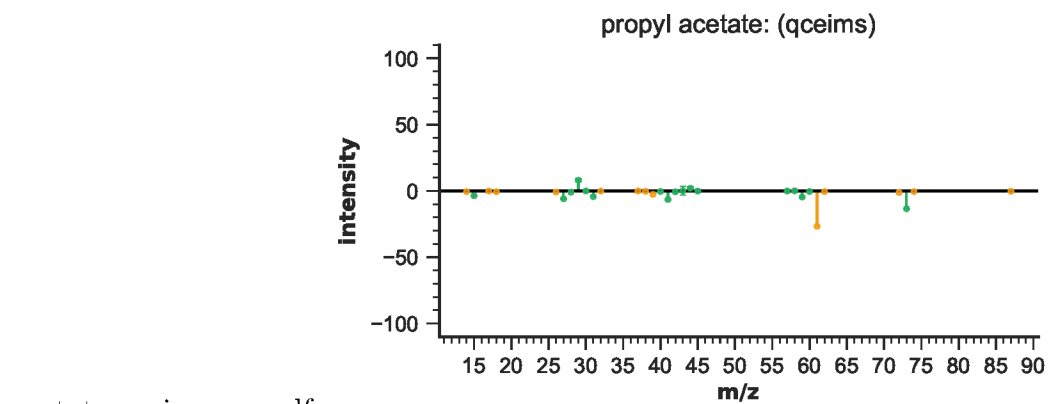


acetate-cfm-ms.pdf

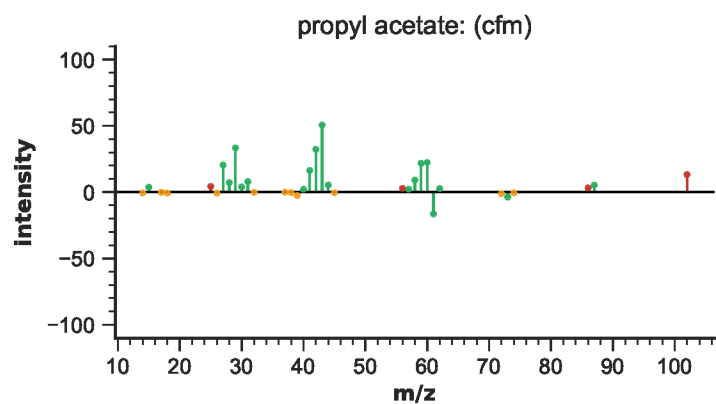


acetate-nist-ms.pdf

Figure S79: EIMS for propyl acetate



acetate-qceims-err.pdf

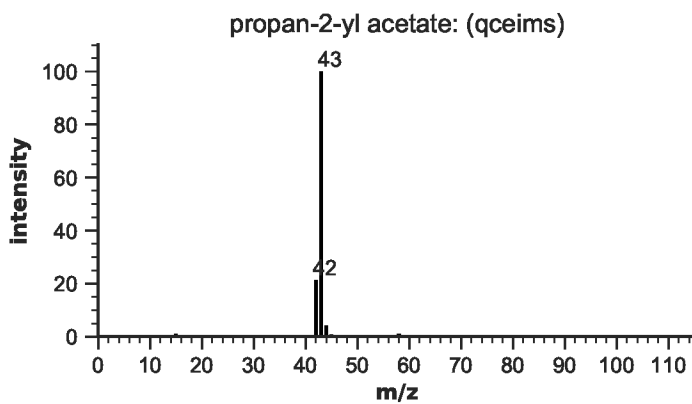


acetate-cfm-err.pdf

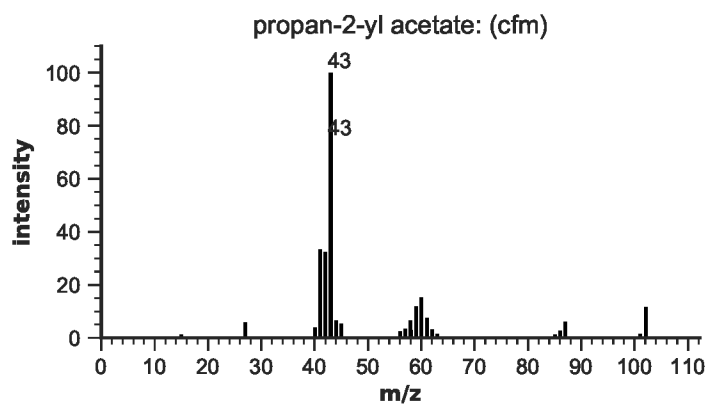
Figure S80: EIMS differences for propyl acetate

2.7.6 41: propan-2-yl acetate

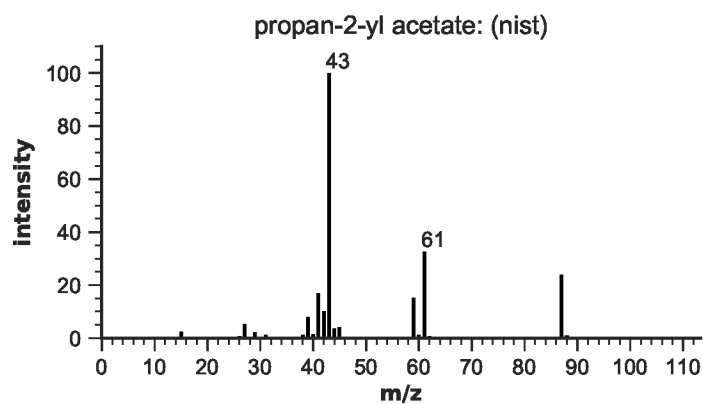
Accepted Article



acetate-qceims-ms.pdf



acetate-cfm-ms.pdf



acetate-nist-ms.pdf

Figure S81: EIMS for propan-2-yl acetate

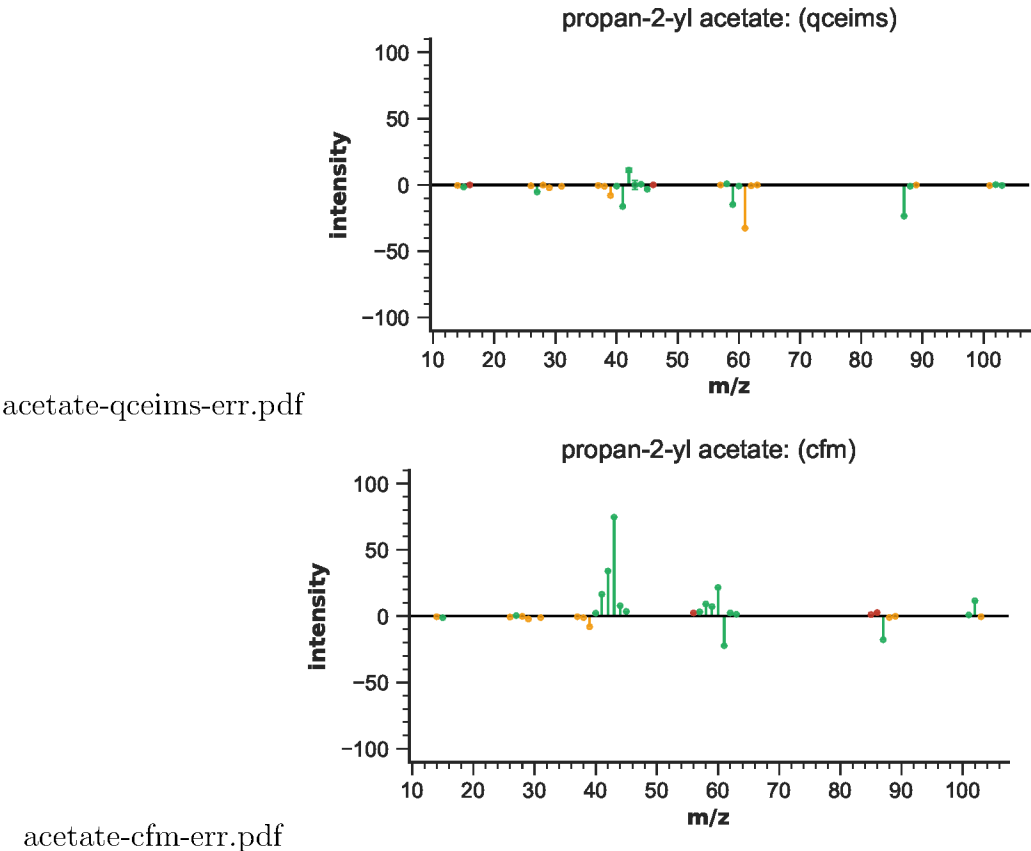


Figure S82: EIMS differences for propan-2-yl acetate

2.7.7 42: methyl 2-methylpropanoate

Accepted Article

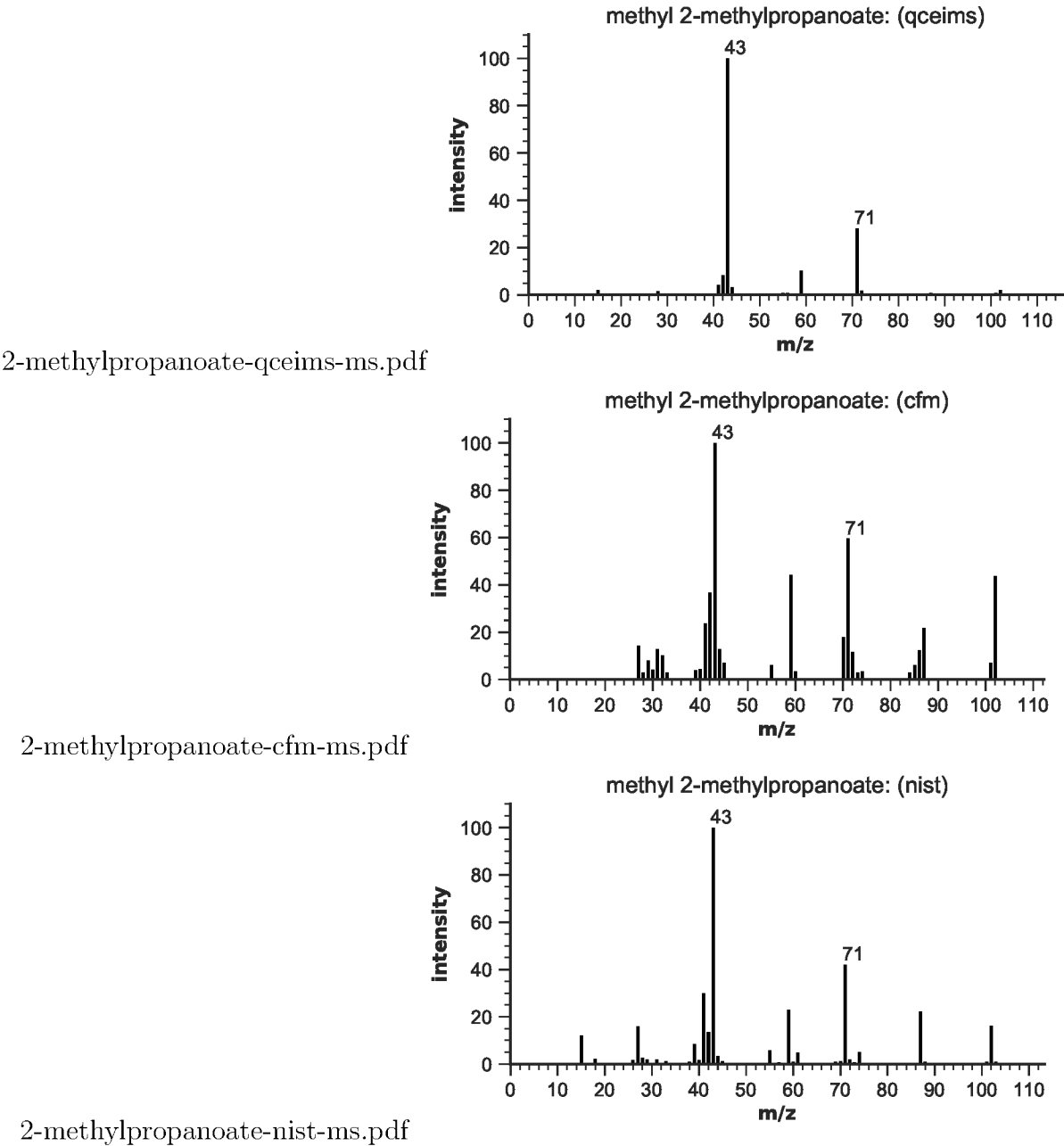
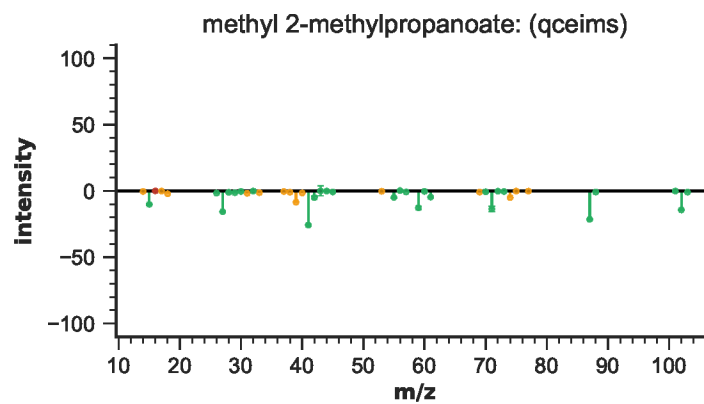
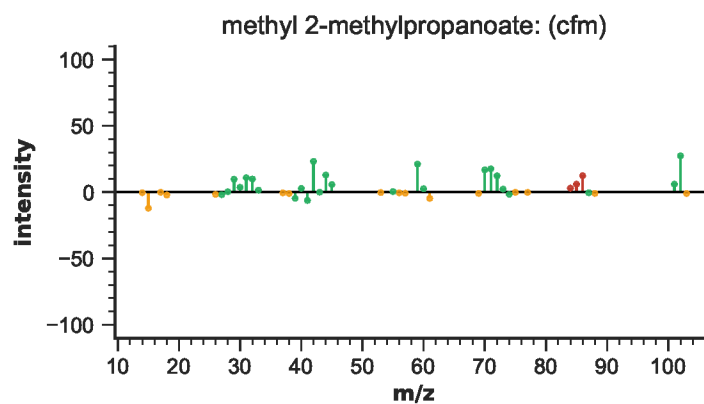


Figure S83: EIMS for methyl 2-methylpropanoate



2-methylpropanoate-qceims-err.pdf

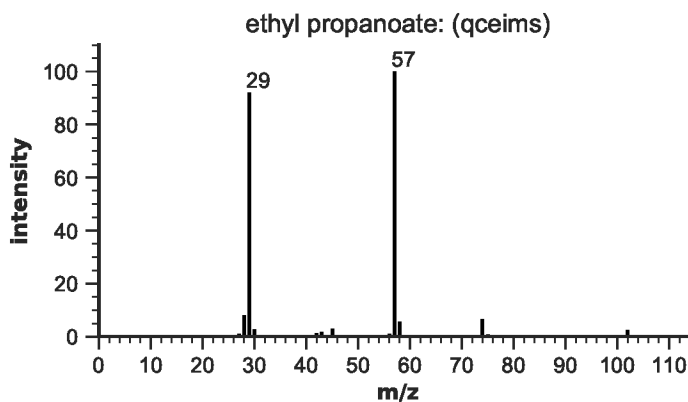


2-methylpropanoate-cfm-err.pdf

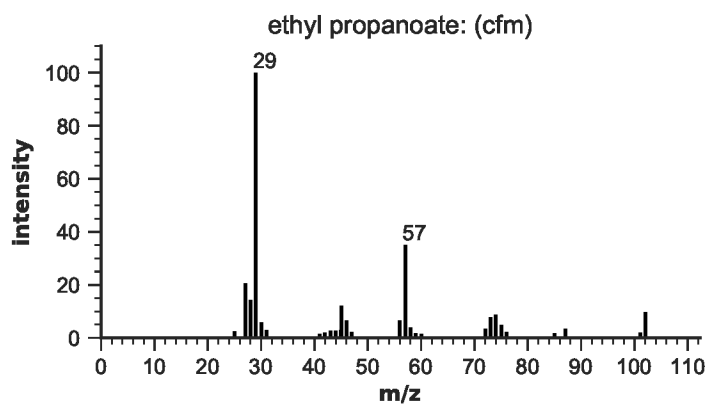
Figure S84: EIMS differences for methyl 2-methylpropanoate

2.7.8 43: ethyl propanoate

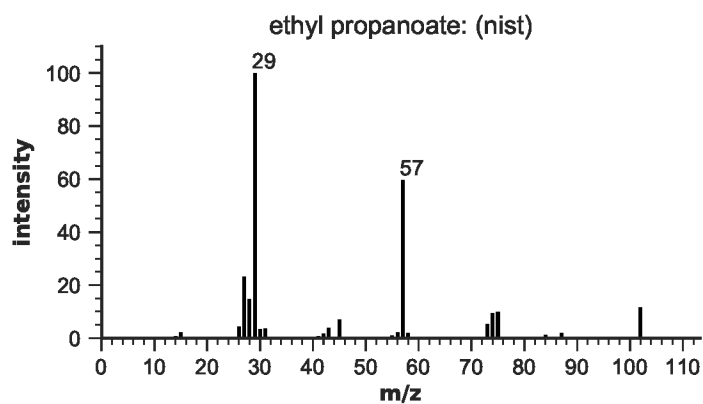
Accepted Article



propanoate-qceims-ms.pdf



propanoate-cfm-ms.pdf



propanoate-nist-ms.pdf

Figure S85: EIMS for ethyl propanoate

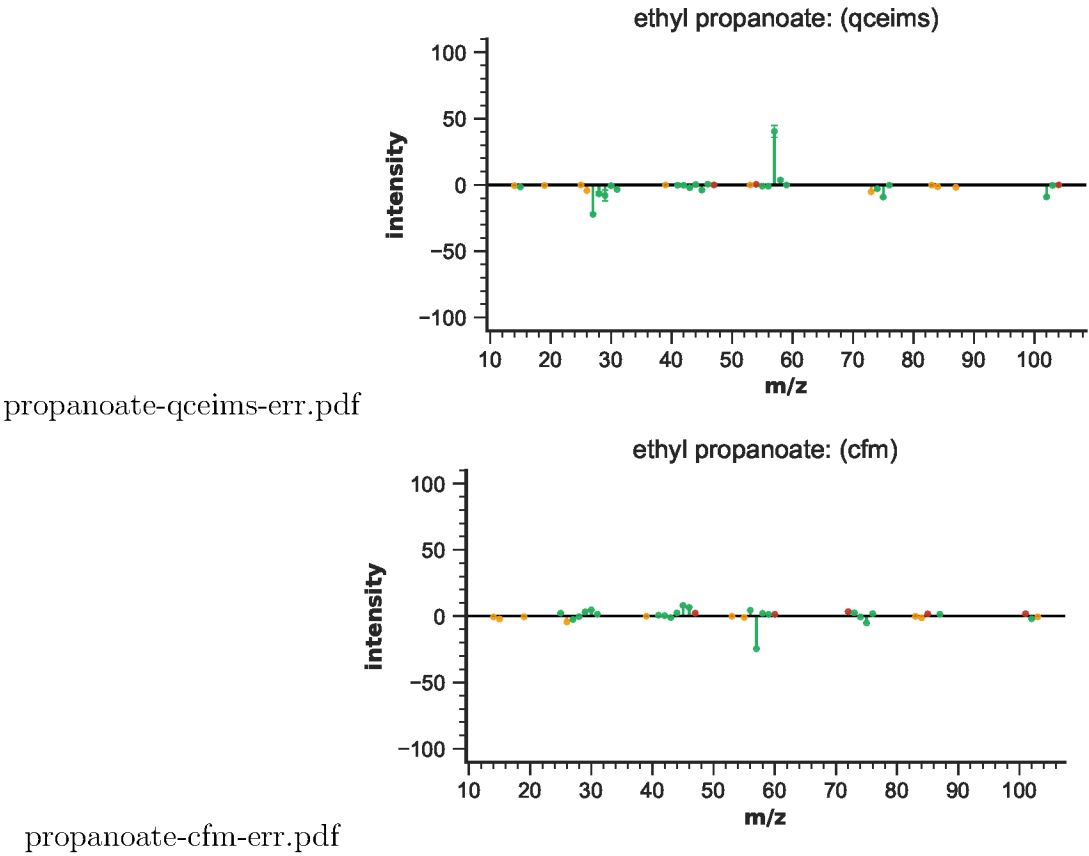
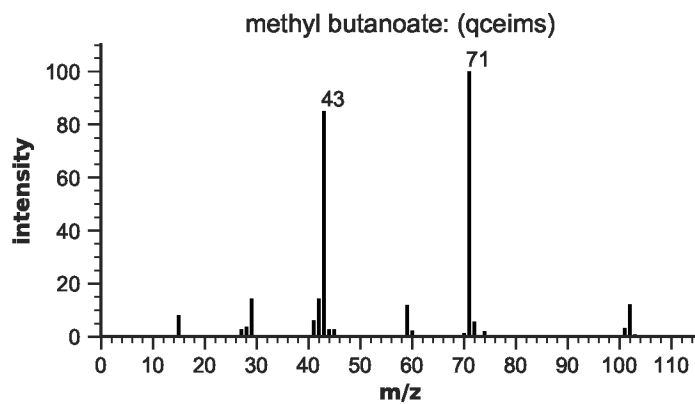


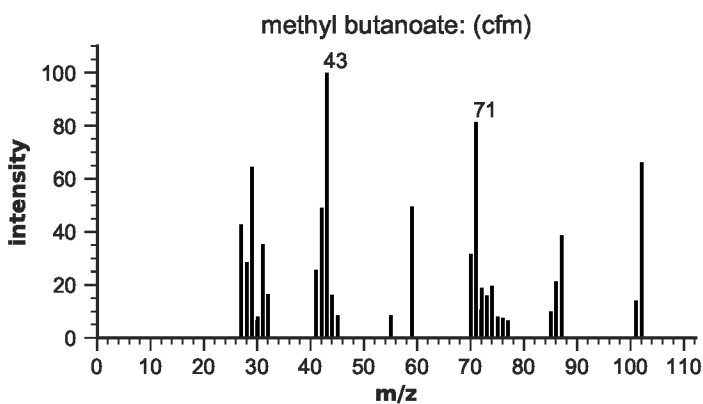
Figure S86: EIMS differences for ethyl propanoate

2.7.9 44: methyl butanoate

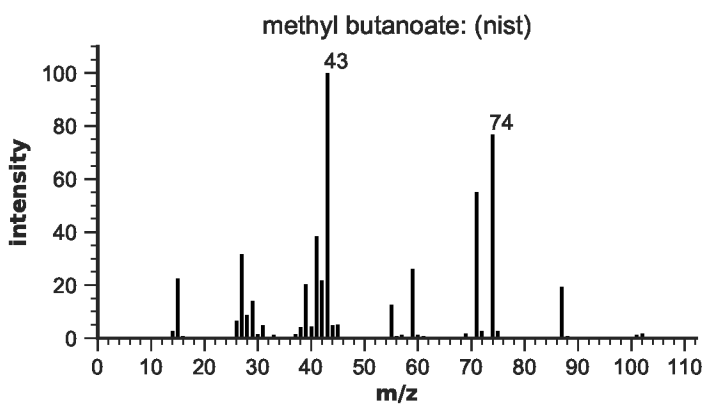
Accepted Article



butanoate-qceims-ms.pdf



butanoate-cfm-ms.pdf



butanoate-nist-ms.pdf

Figure S87: EIMS for methyl butanoate

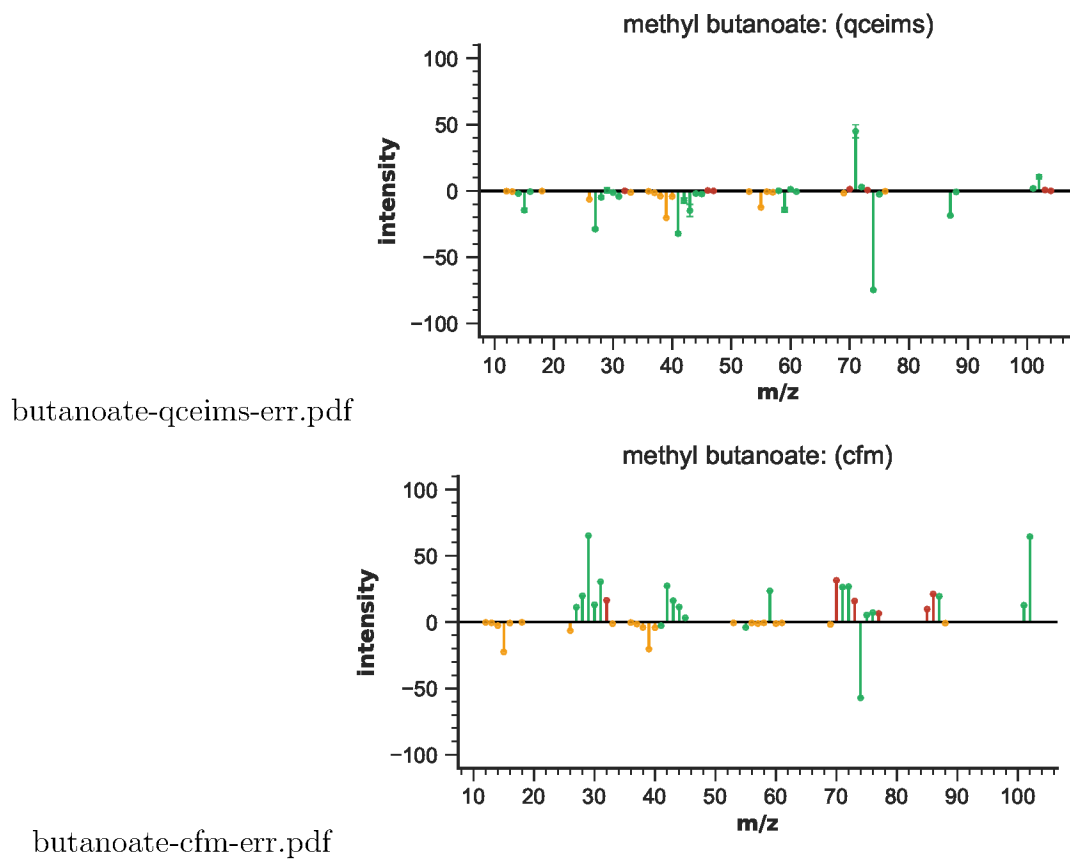
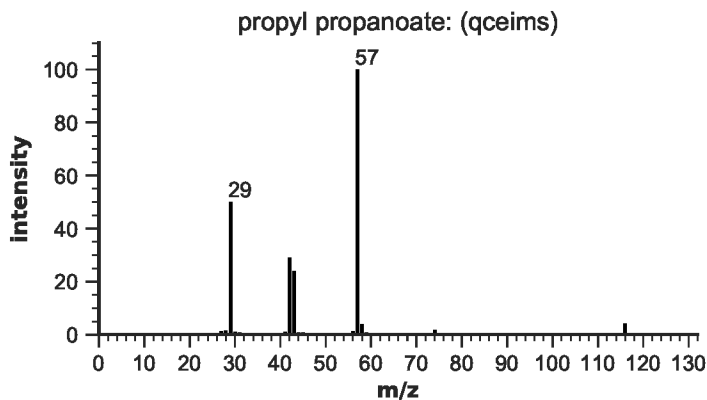


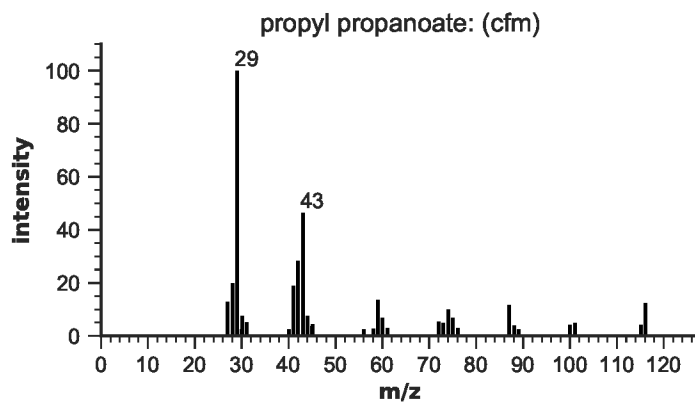
Figure S88: EIMS differences for methyl butanoate

2.7.10 45: propyl propanoate

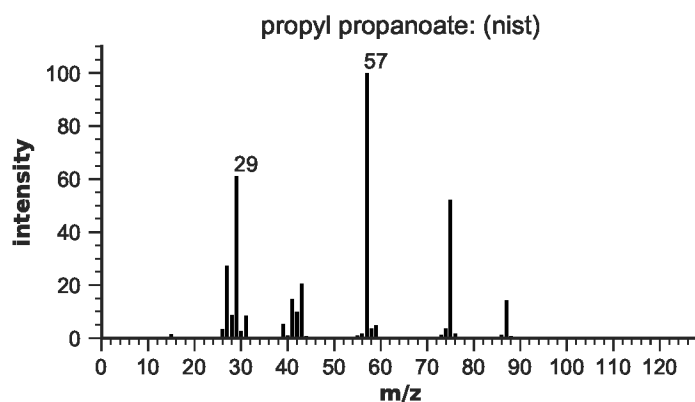
Accepted Article



propanoate-qceims-ms.pdf



propanoate-cfm-ms.pdf



propanoate-nist-ms.pdf

Figure S89: EIMS for propyl propanoate

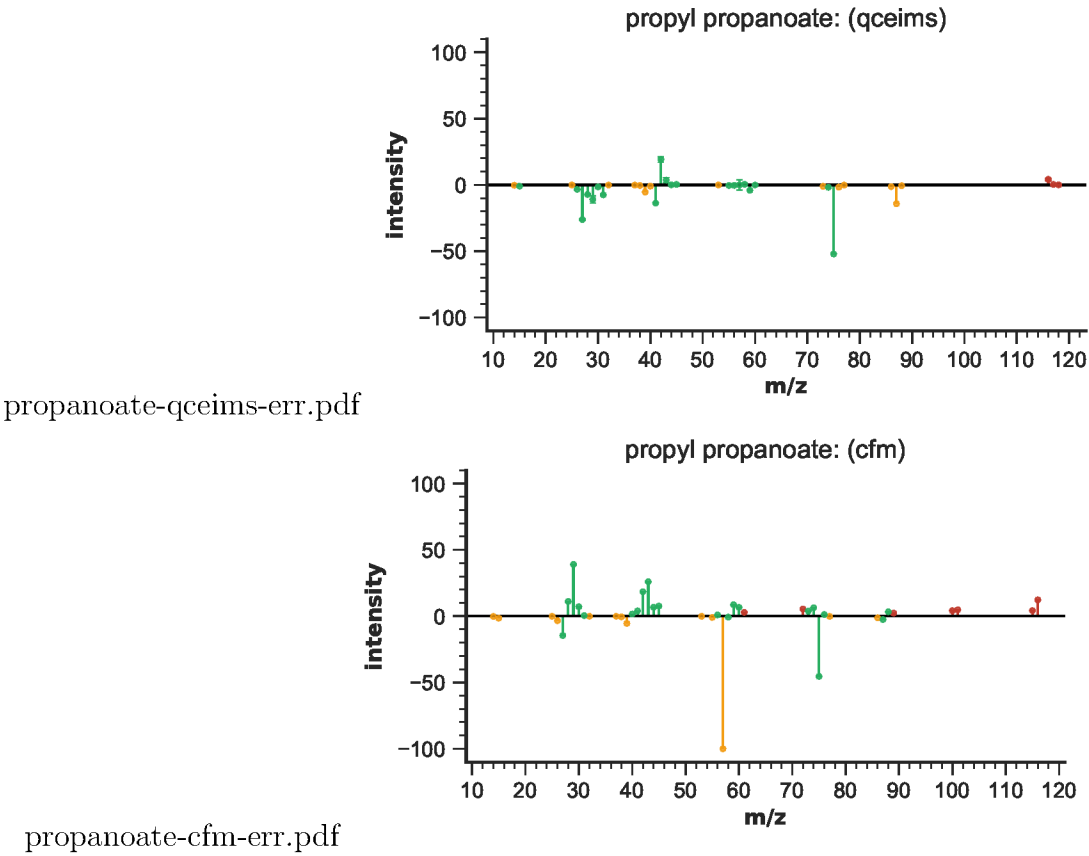


Figure S90: EIMS differences for propyl propanoate

2.7.11 46: propan-2-yl propanoate

Accepted Article

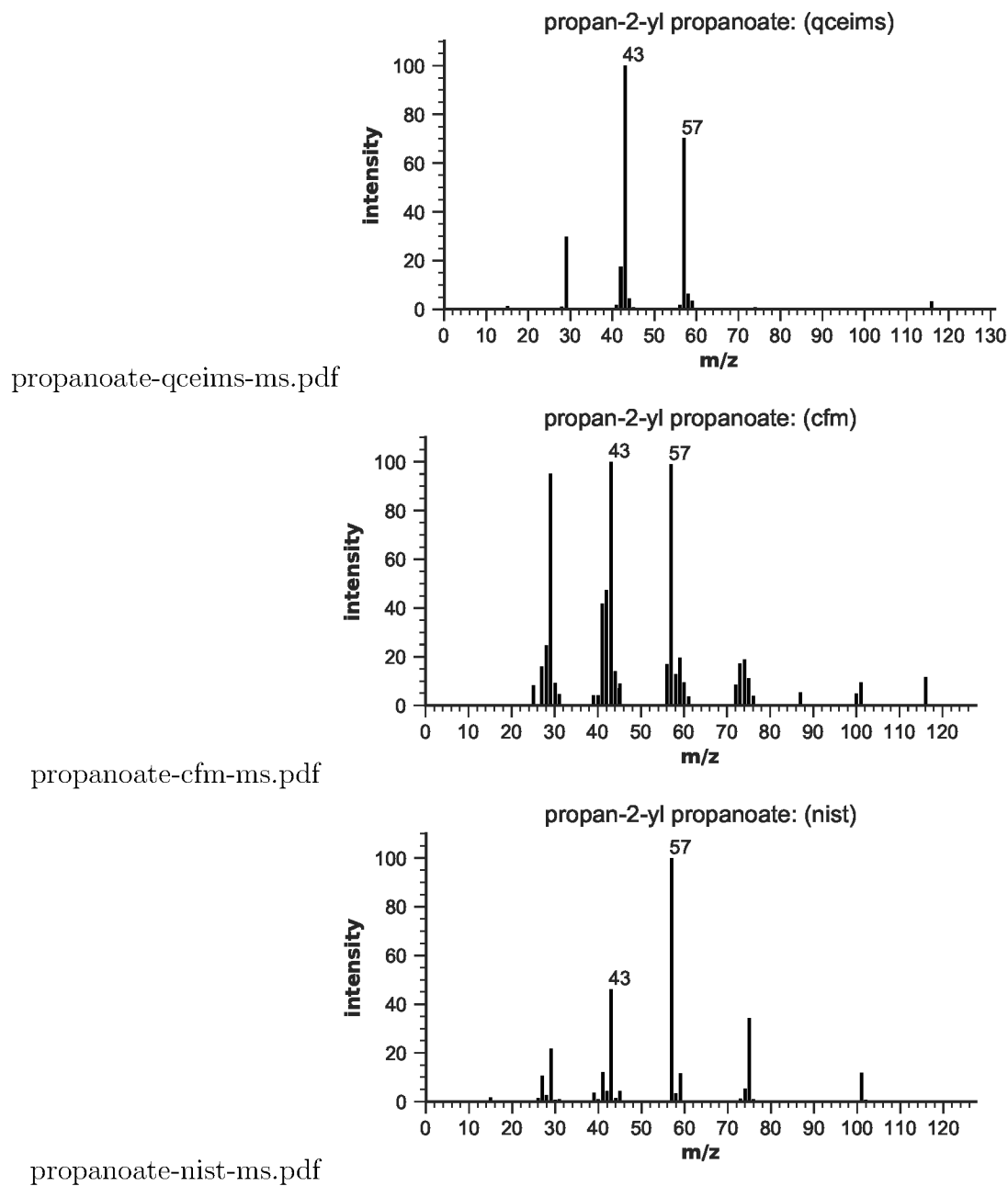
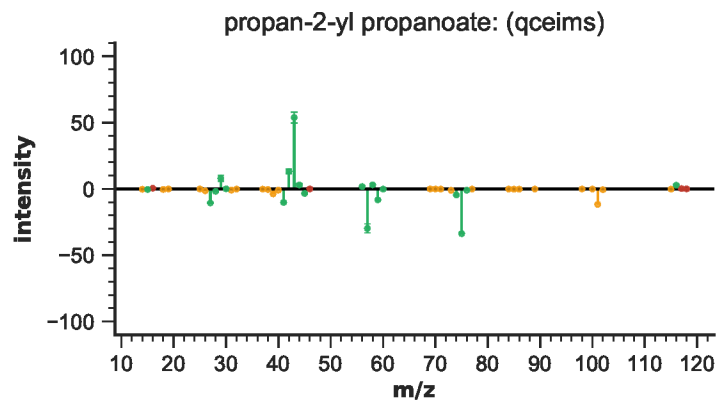
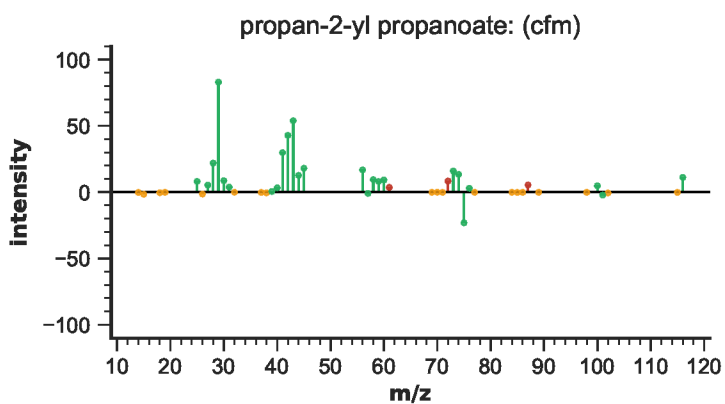


Figure S91: EIMS for propan-2-yl propanoate



propanoate-qceims-err.pdf

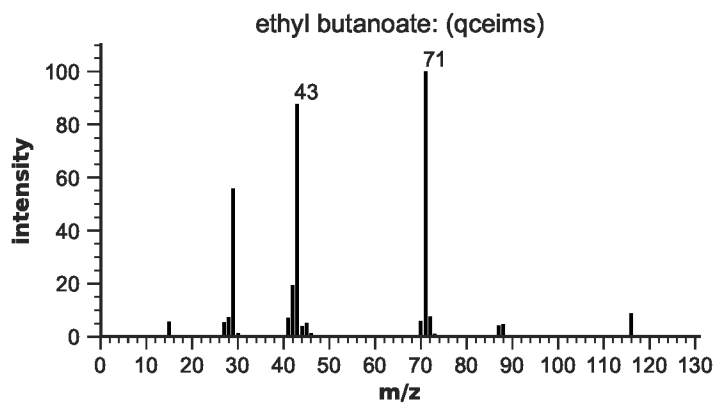


propanoate-cfm-err.pdf

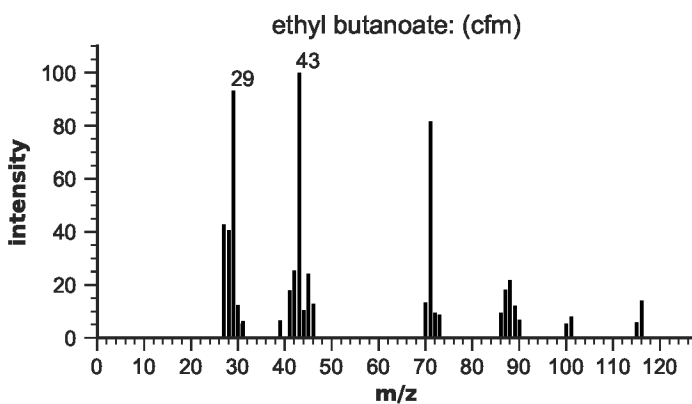
Figure S92: EIMS differences for propan-2-yl propanoate

2.7.12 47: ethyl butanoate

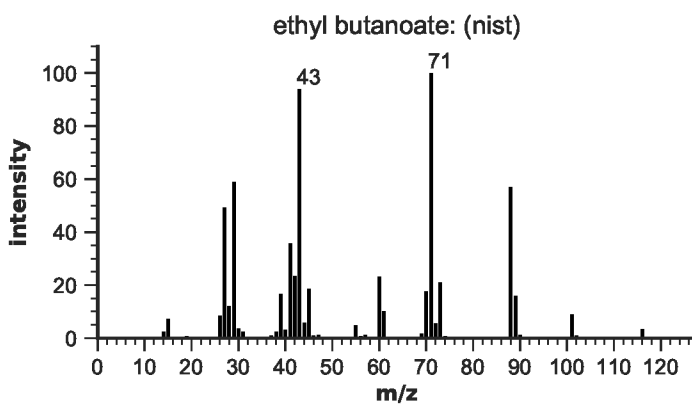
Accepted Article



butanoate-qceims-ms.pdf



butanoate-cfm-ms.pdf



butanoate-nist-ms.pdf

Figure S93: EIMS for ethyl butanoate

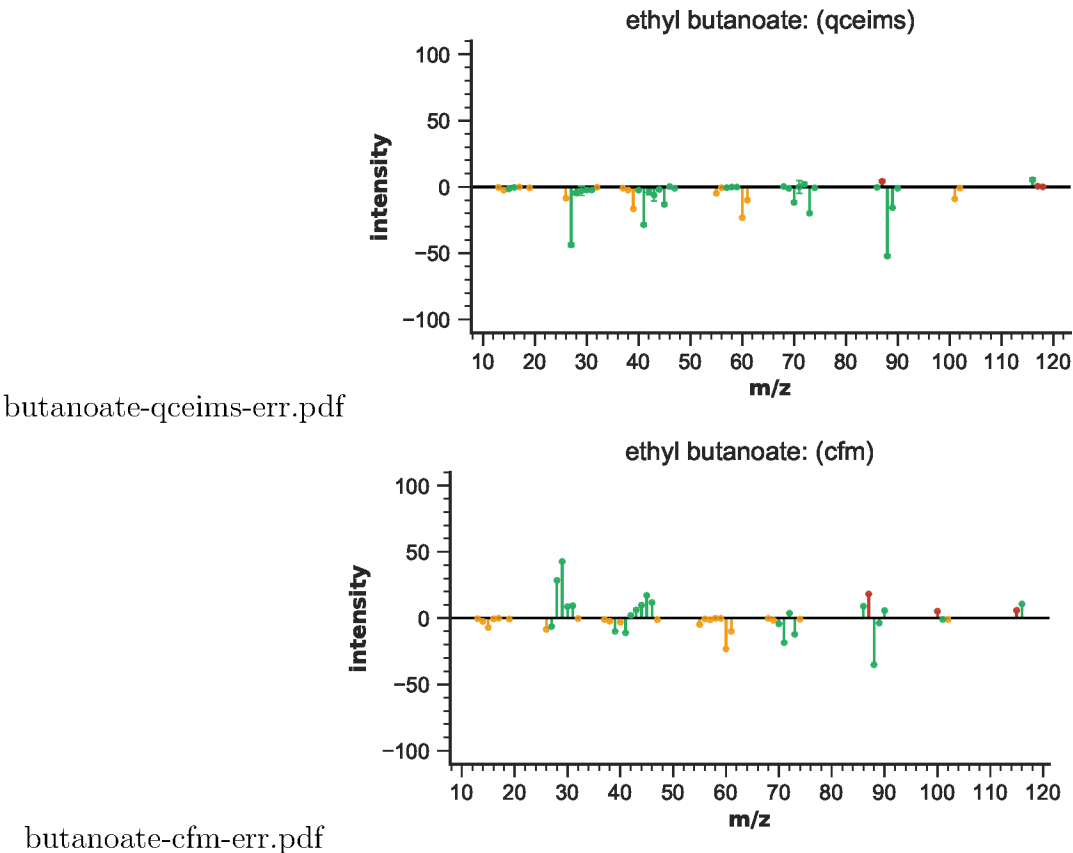
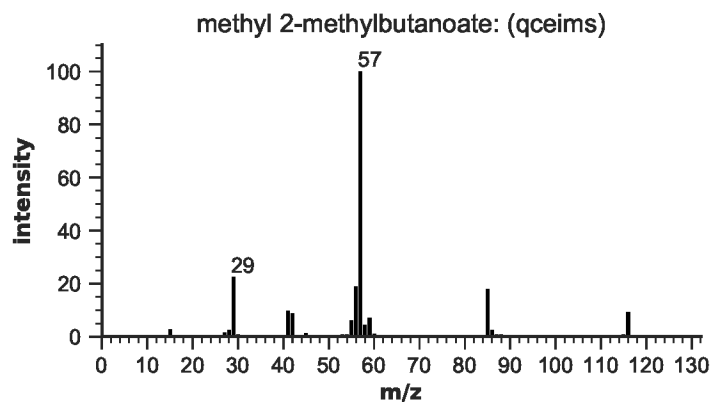


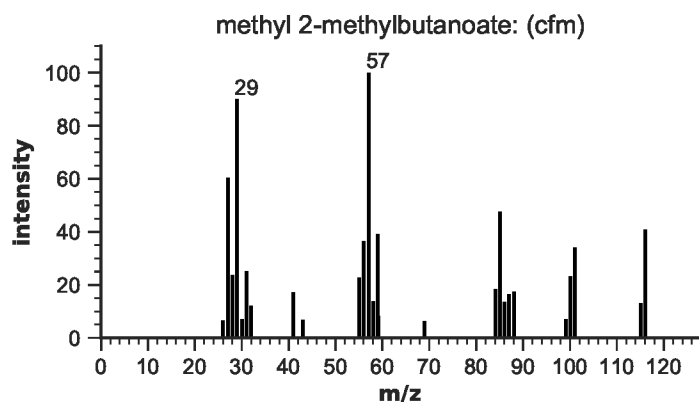
Figure S94: EIMS differences for ethyl butanoate

2.7.13 48: methyl 2-methylbutanoate

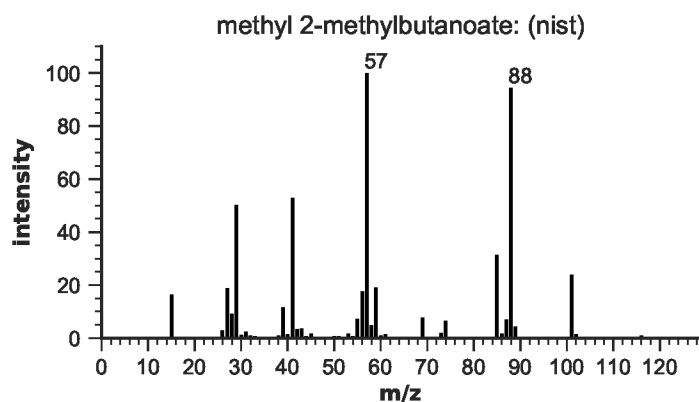
Accepted Article



2-methylbutanoate-qceims-ms.pdf

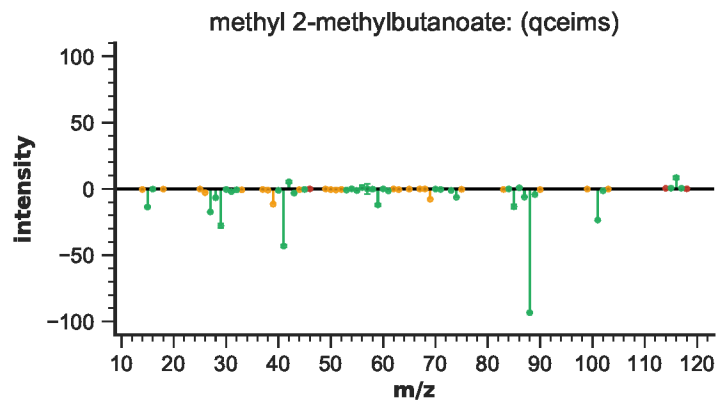


2-methylbutanoate-cfm-ms.pdf

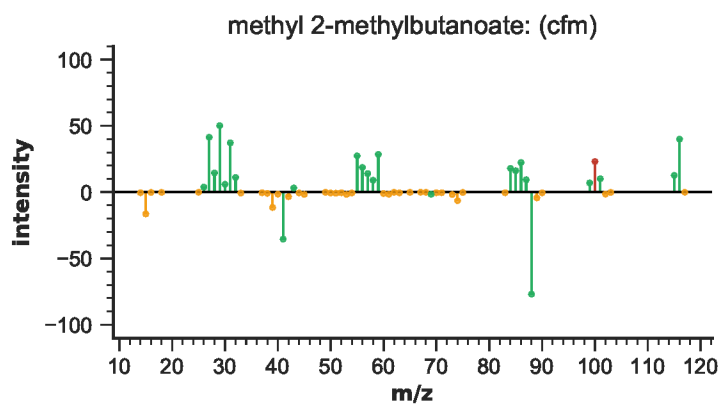


2-methylbutanoate-nist-ms.pdf

Figure S95: EIMS for methyl 2-methylbutanoate



2-methylbutanoate-qceims-err.pdf

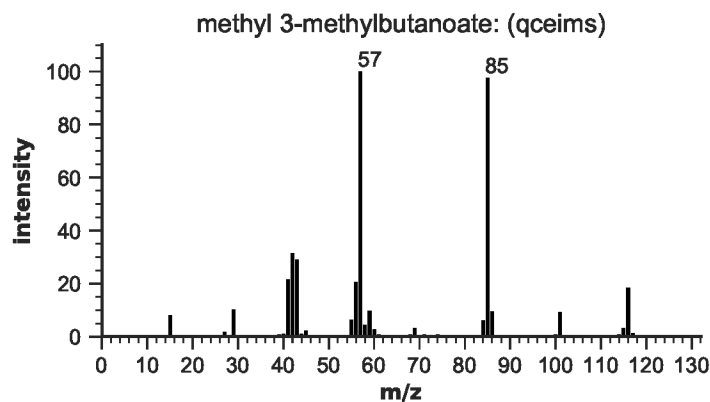


2-methylbutanoate-cfm-err.pdf

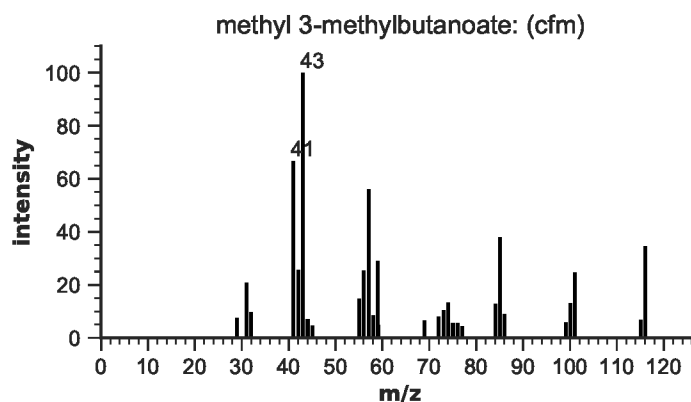
Figure S96: EIMS differences for methyl 2-methylbutanoate

2.7.14 49: methyl 3-methylbutanoate

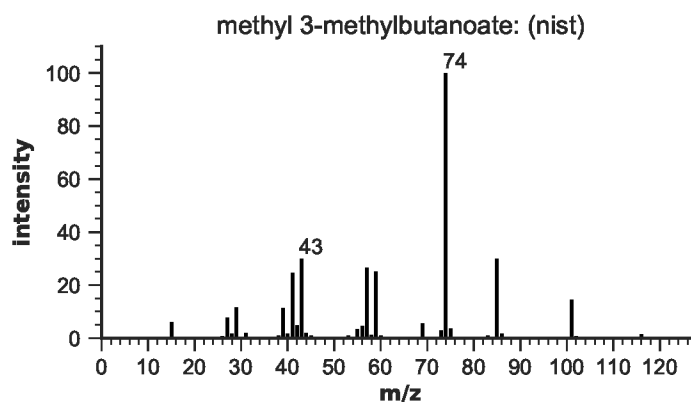
Accepted Article



3-methylbutanoate-qceims-ms.pdf



3-methylbutanoate-cfm-ms.pdf



3-methylbutanoate-nist-ms.pdf

Figure S97: EIMS for methyl 3-methylbutanoate

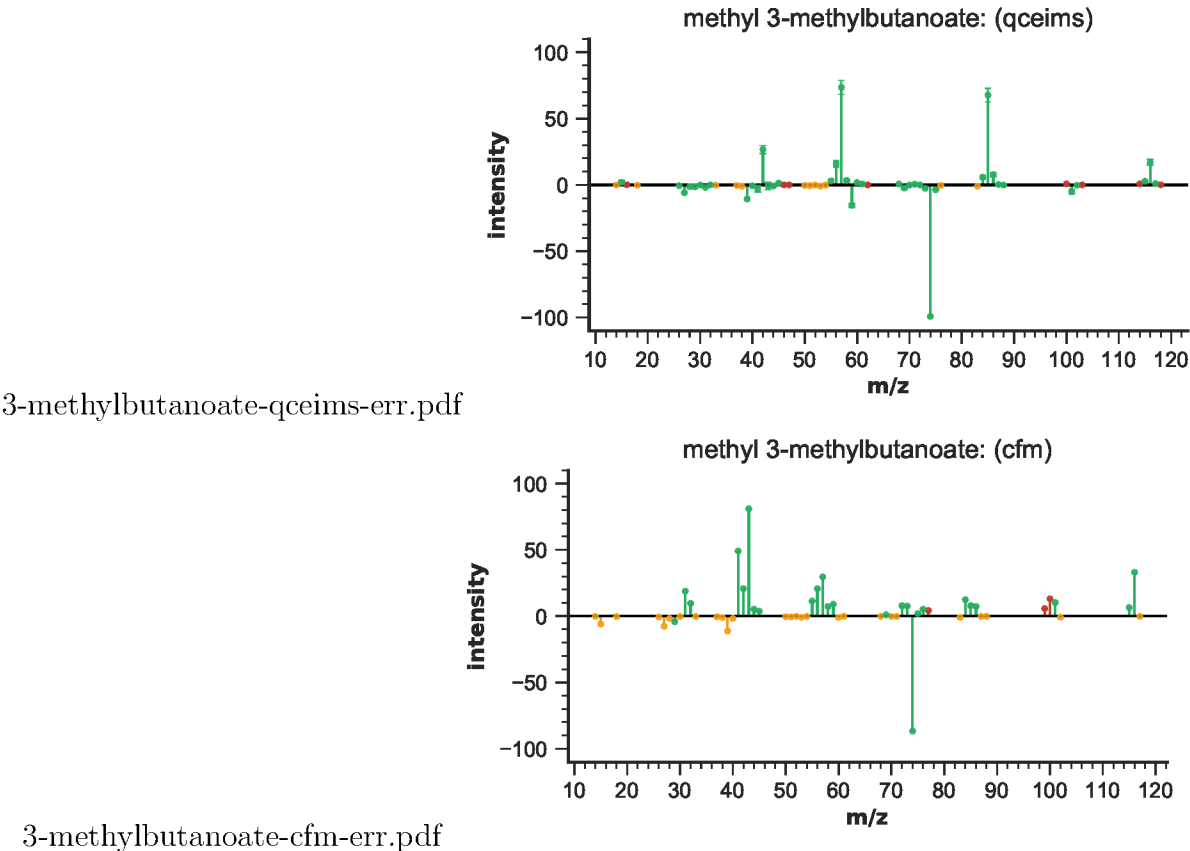
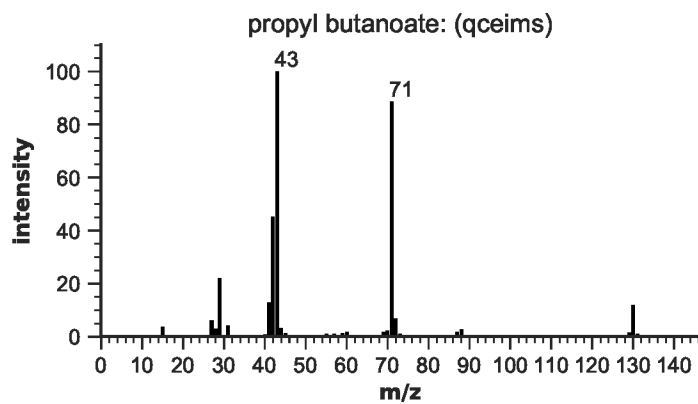


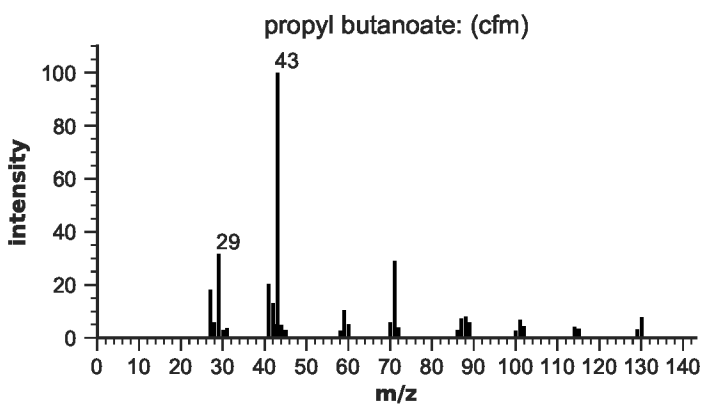
Figure S98: EIMS differences for methyl 3-methylbutanoate

2.7.15 50: propyl butanoate

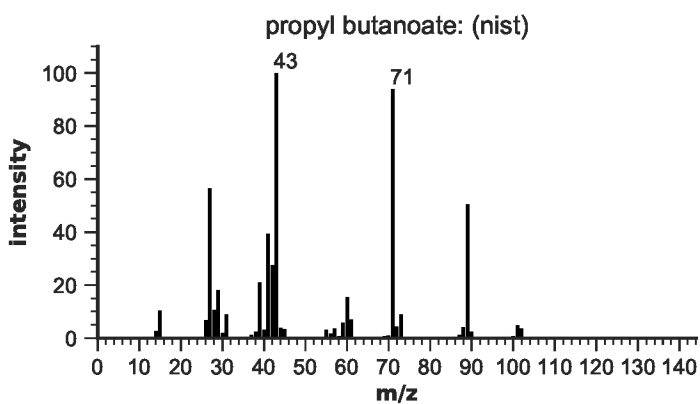
Accepted Article



butanoate-qceims-ms.pdf



butanoate-cfm-ms.pdf



butanoate-nist-ms.pdf

Figure S99: EIMS for propyl butanoate

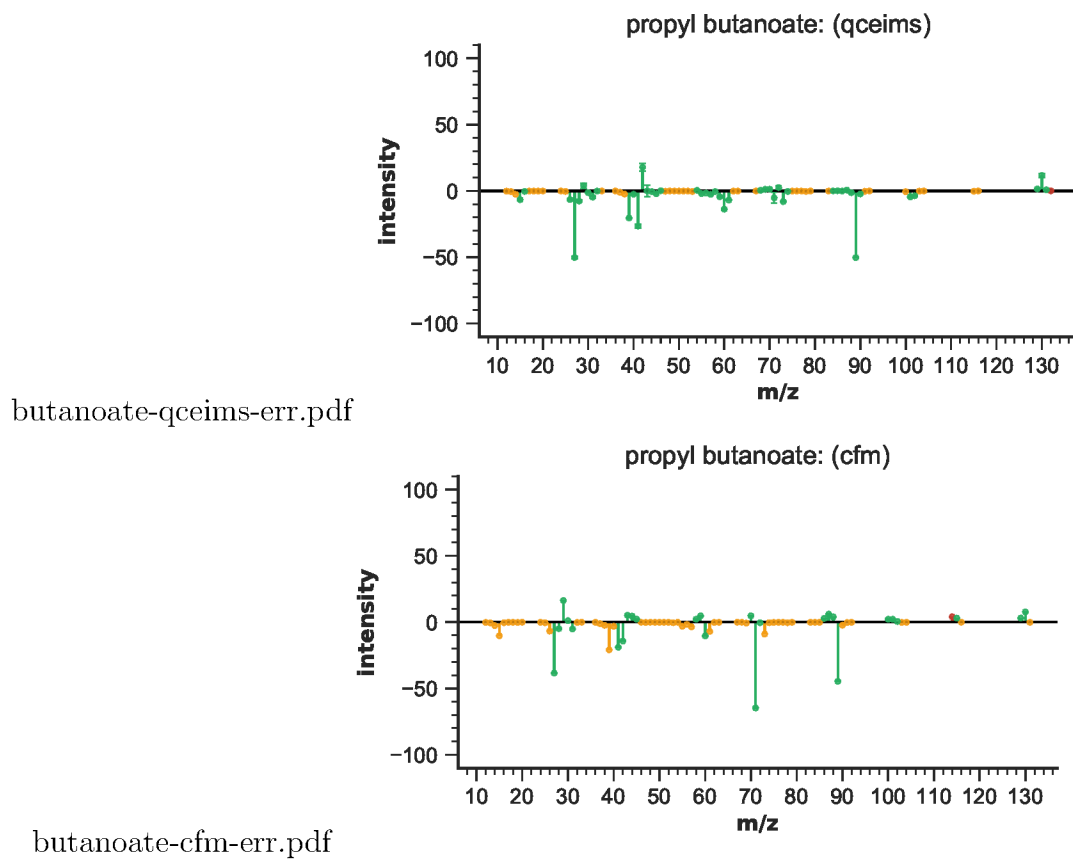
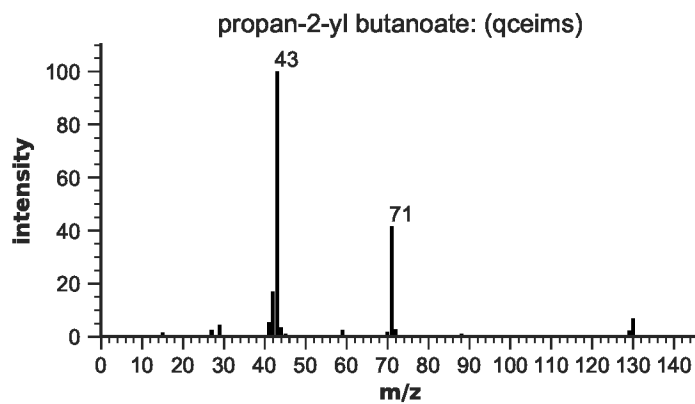


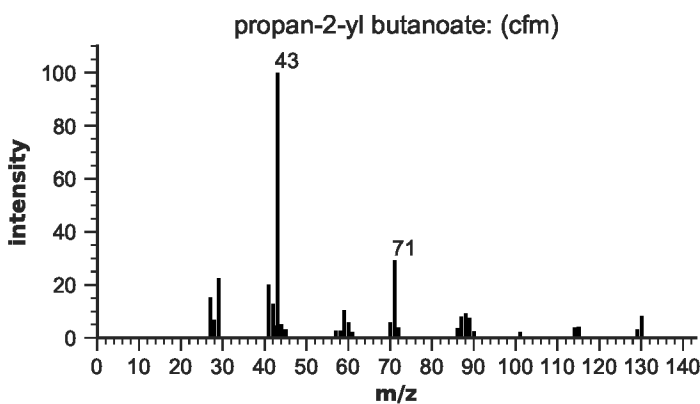
Figure S100: EIMS differences for propyl butanoate

2.7.16 51: propan-2-yl butanoate

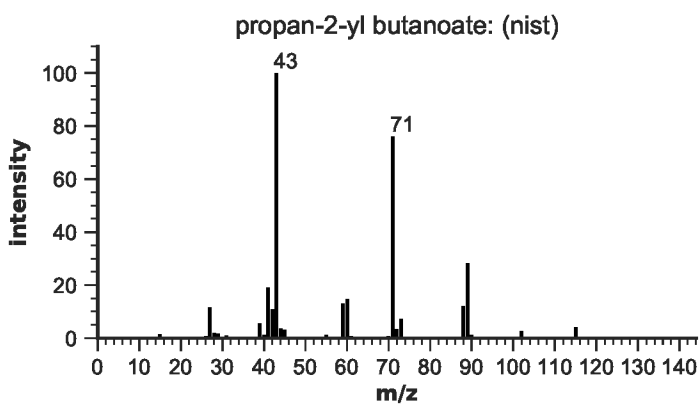
Accepted Article



butanoate-qceims-ms.pdf



butanoate-cfm-ms.pdf



butanoate-nist-ms.pdf

Figure S101: EIMS for propan-2-yl butanoate

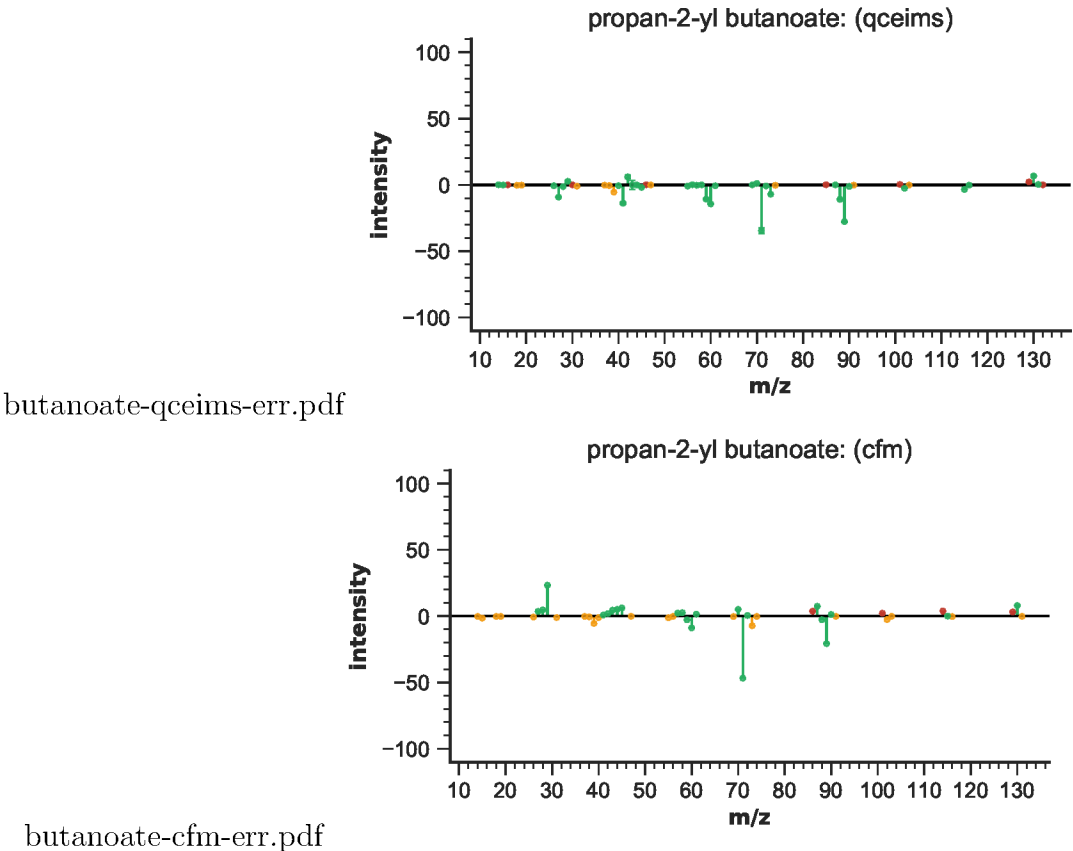


Figure S102: EIMS differences for propan-2-yl butanoate

2.8 Ketones

2.8.1 52: acetone

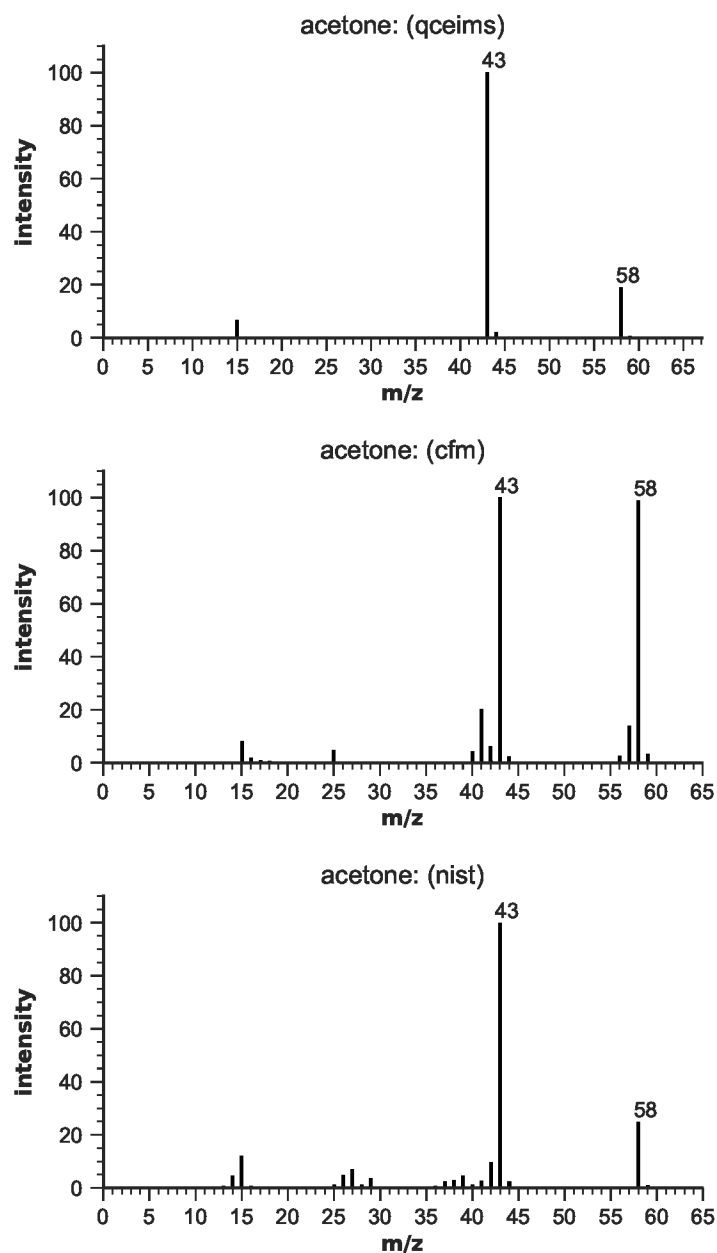


Figure S103: EIMS for acetone

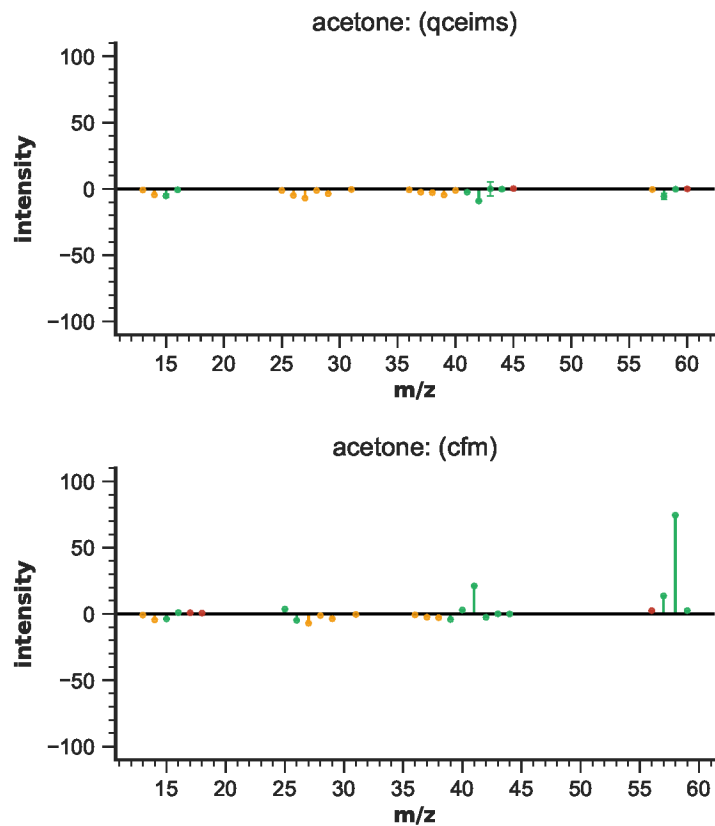


Figure S104: EIMS differences for acetone

2.8.2 53: butan-2-one

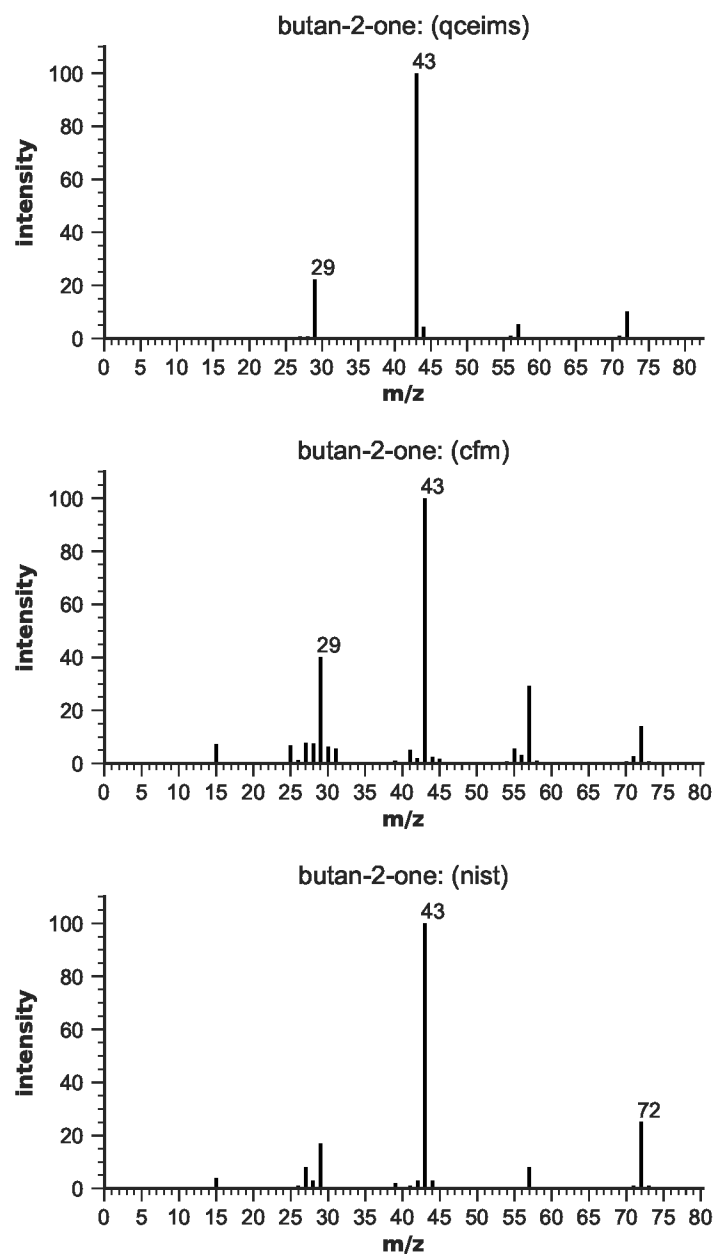


Figure S105: EIMS for butan-2-one

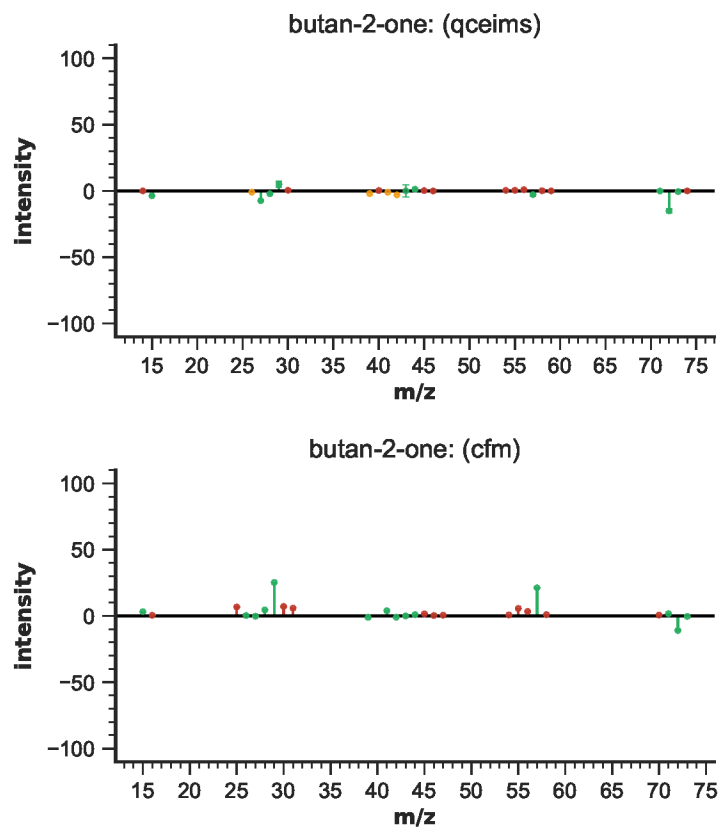


Figure S106: EIMS differences for butan-2-one

2.9 Phenols

2.9.1 54: phenol

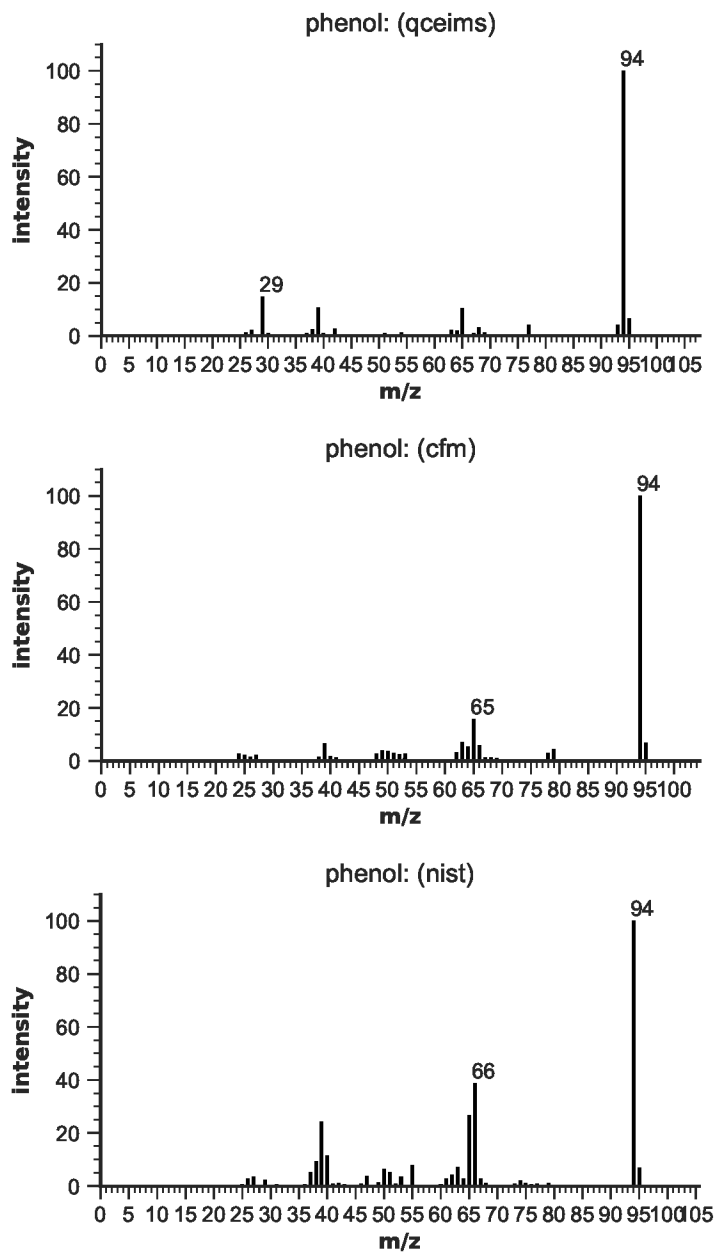


Figure S107: EIMS for phenol

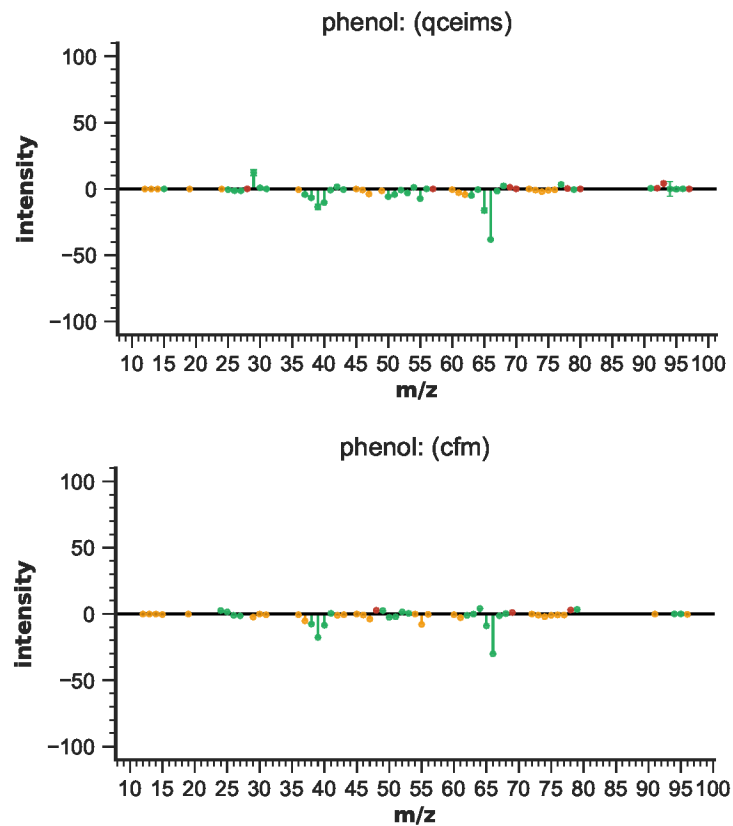


Figure S108: EIMS differences for phenol

2.9.2 55: 4-methylphenol

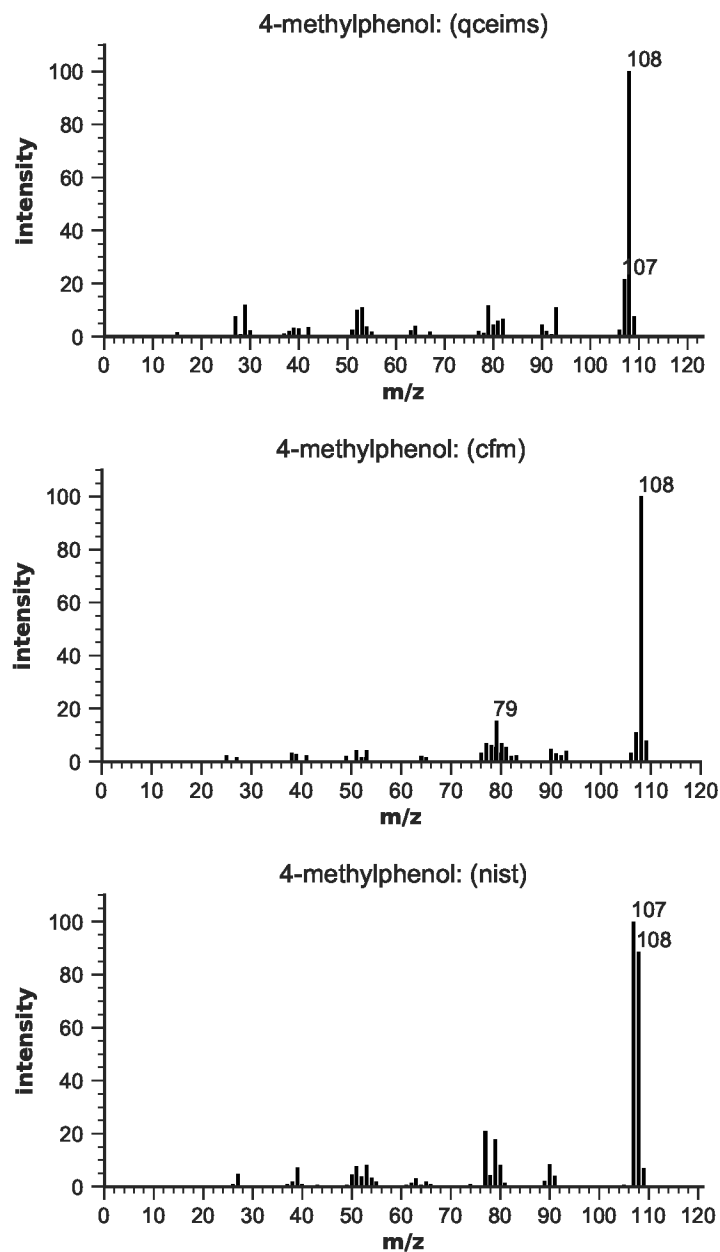


Figure S109: EIMS for 4-methylphenol

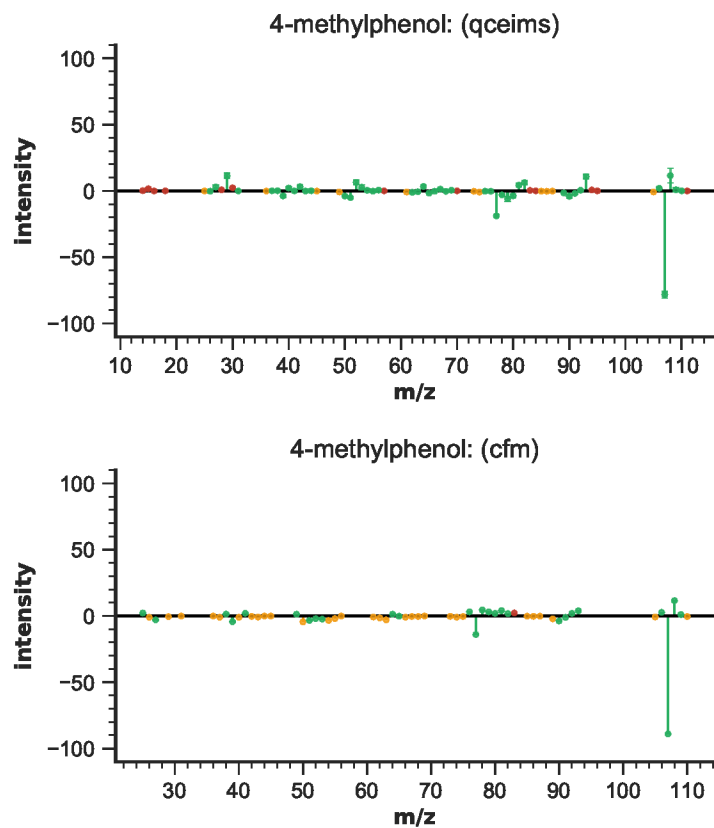


Figure S110: EIMS differences for 4-methylphenol

2.9.3 56: 4-ethylphenol

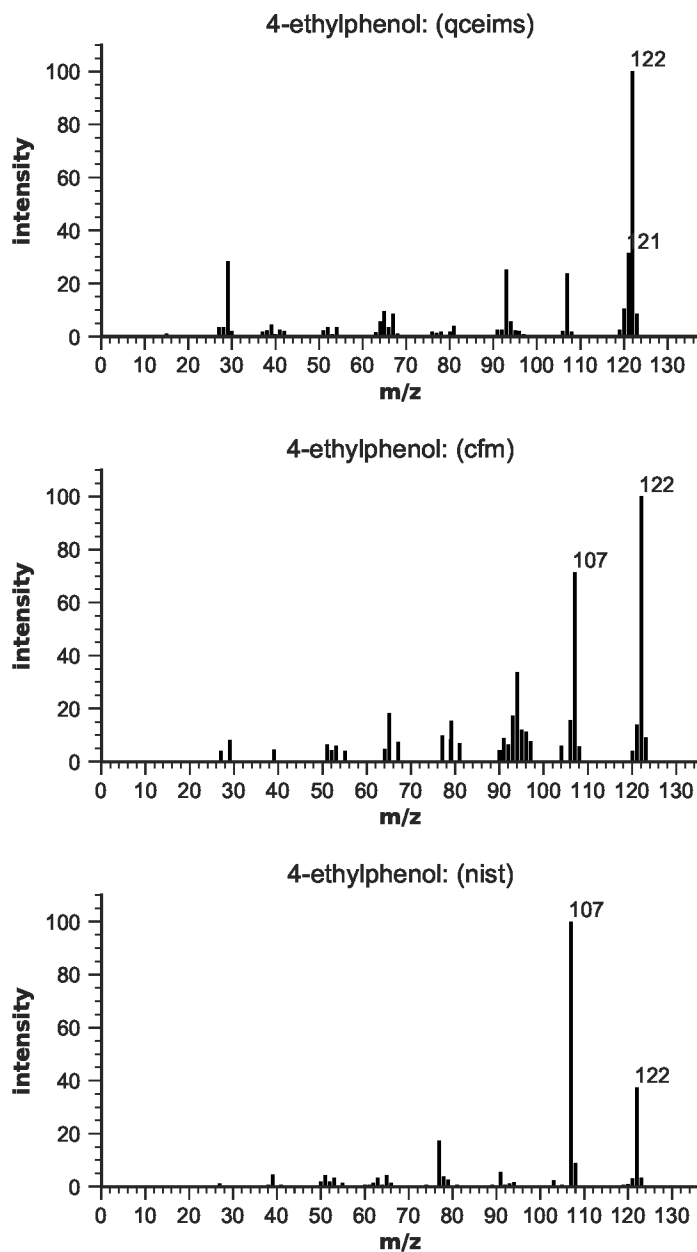


Figure S111: EIMS for 4-ethylphenol

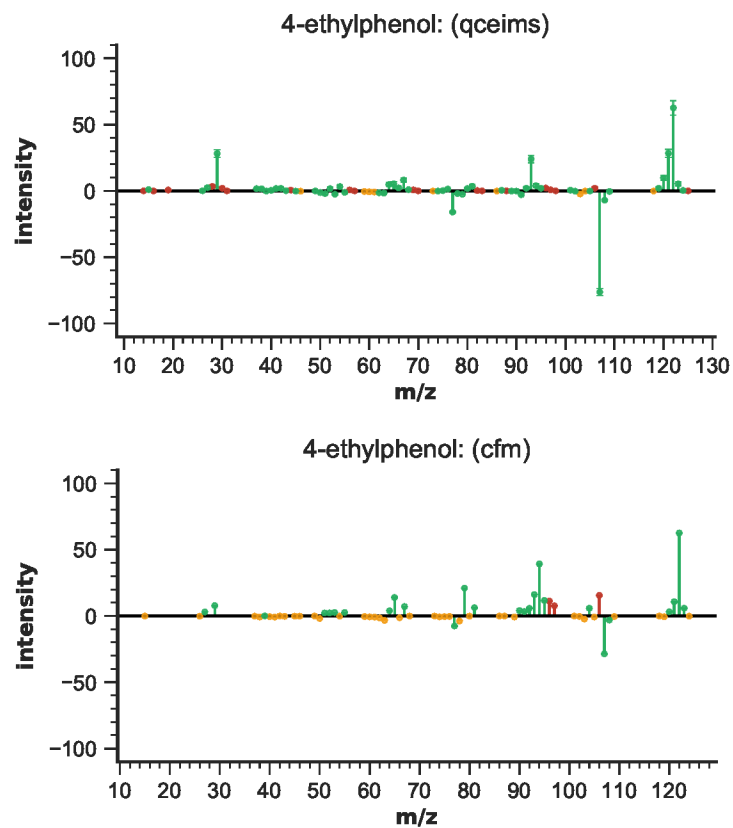


Figure S112: EIMS differences for 4-ethylphenol

2.9.4 57: 4-propylphenol

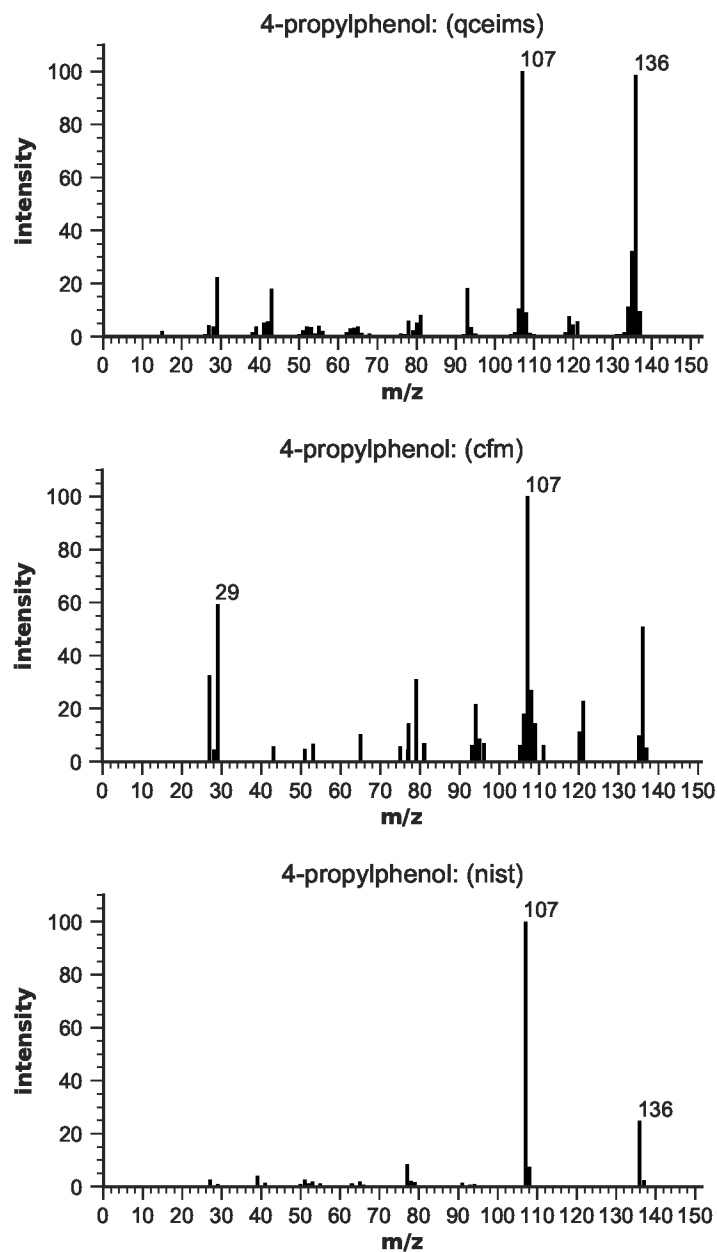


Figure S113: EIMS for 4-propylphenol

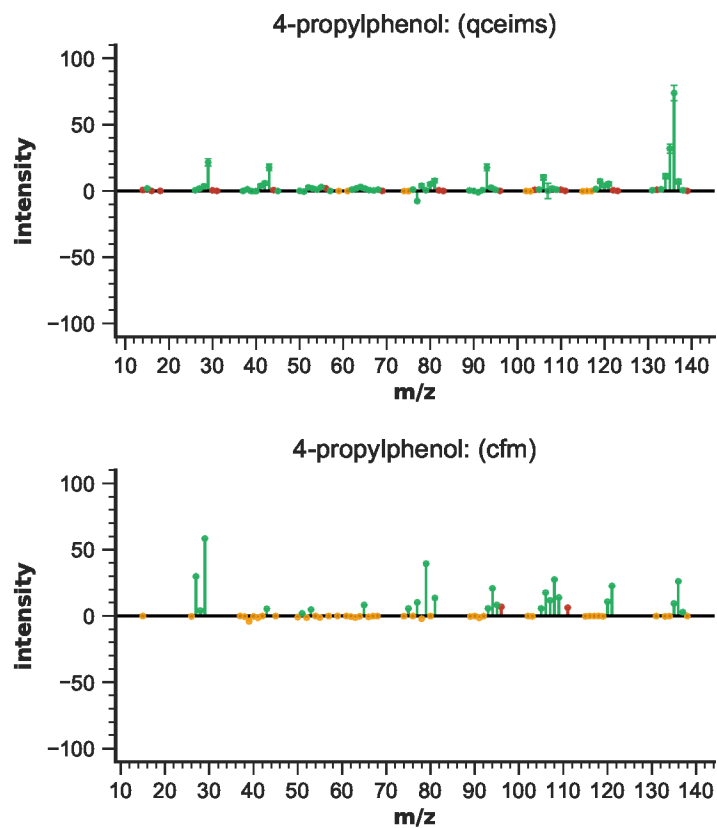


Figure S114: EIMS differences for 4-propylphenol

2.9.5 58: 4-propan-2-ylphenol

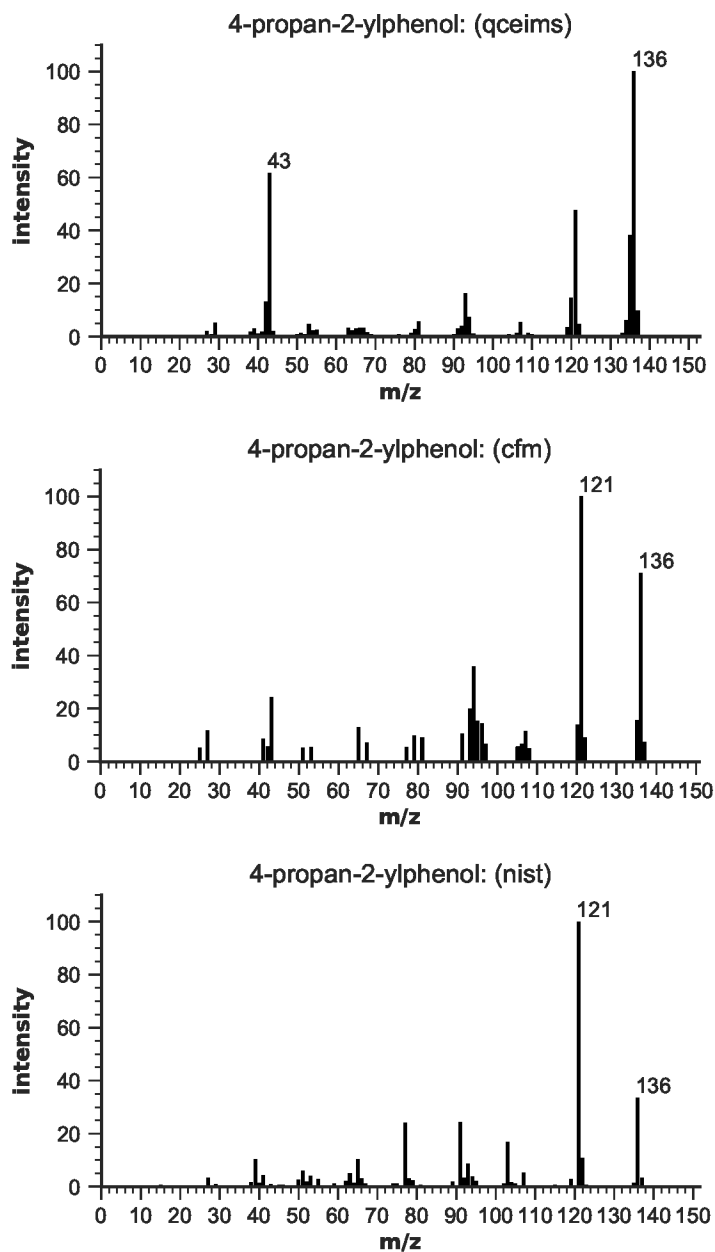


Figure S115: EIMS for 4-propan-2-ylphenol

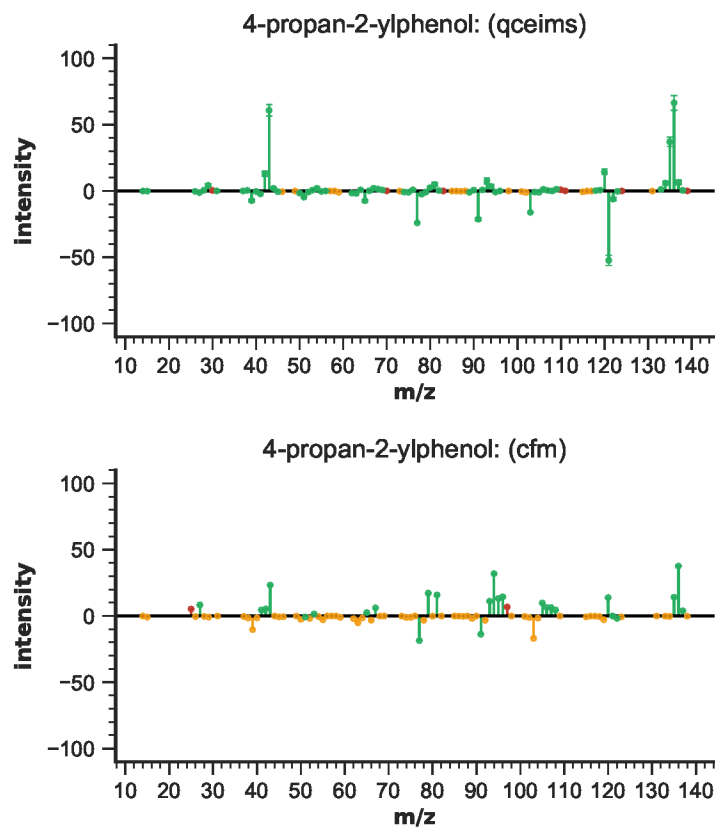


Figure S116: EIMS differences for 4-propan-2-ylphenol

2.9.6 59: 4-butylphenol

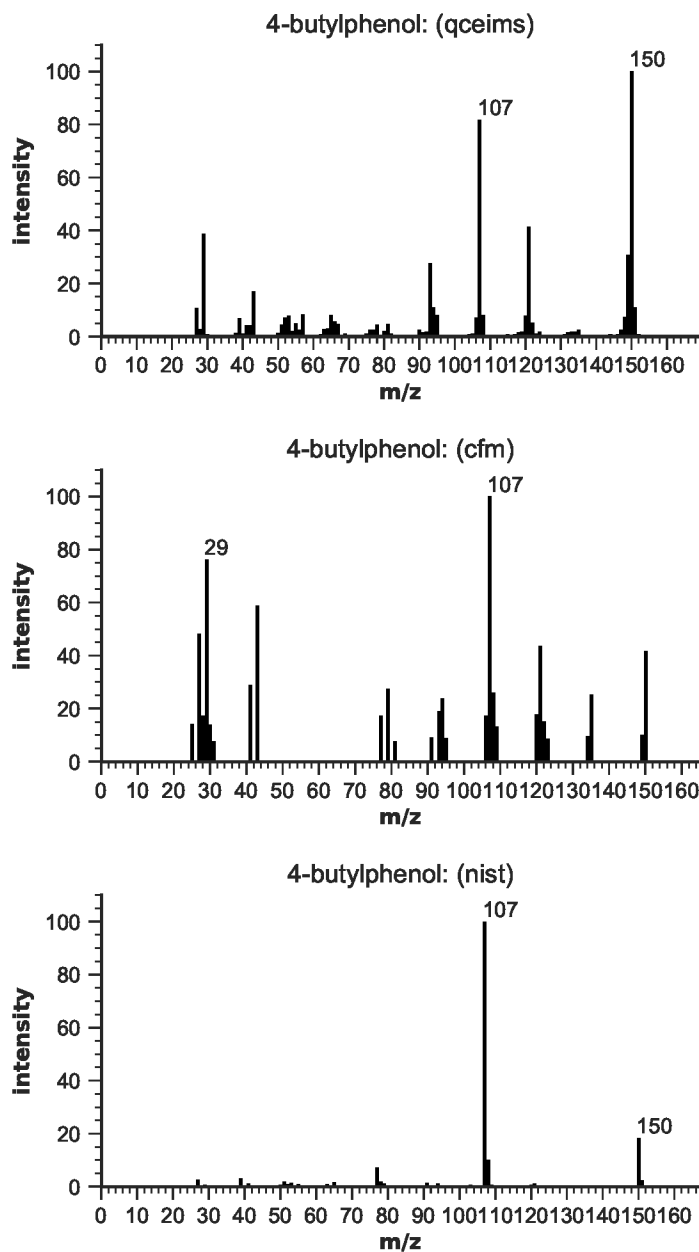


Figure S117: EIMS for 4-butylphenol

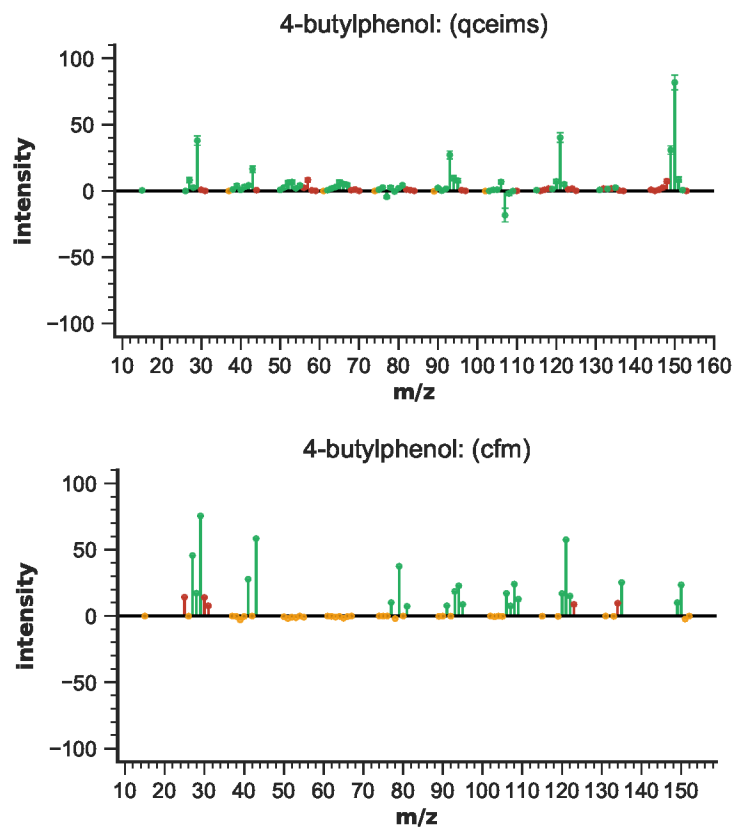


Figure S118: EIMS differences for 4-butylphenol

2.9.7 60: 4-butan-2-ylphenol

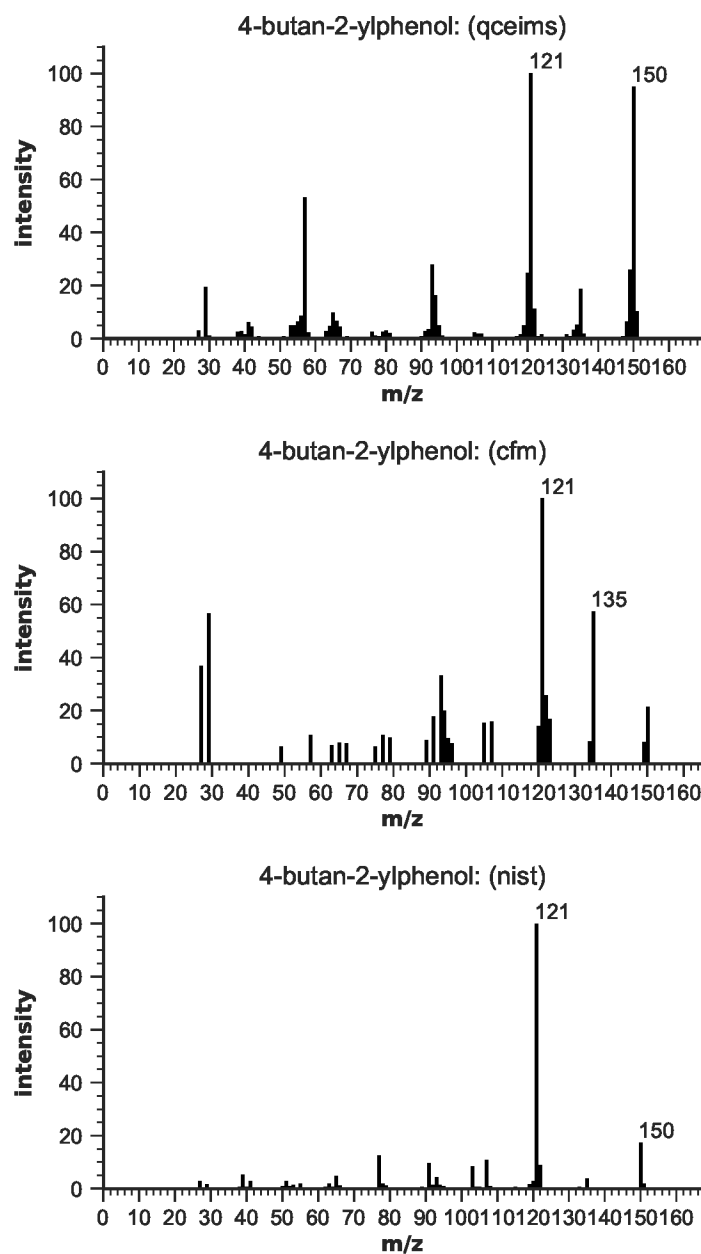


Figure S119: EIMS for 4-butan-2-ylphenol

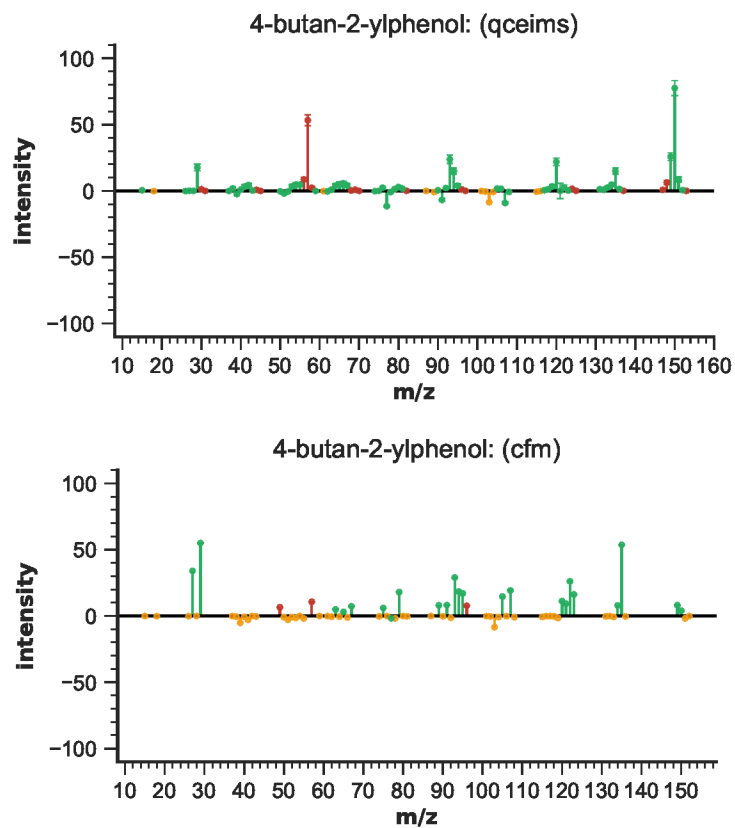


Figure S120: EIMS differences for 4-butan-2-ylphenol

2.9.8 61: 4-(2-methylpropyl)phenol

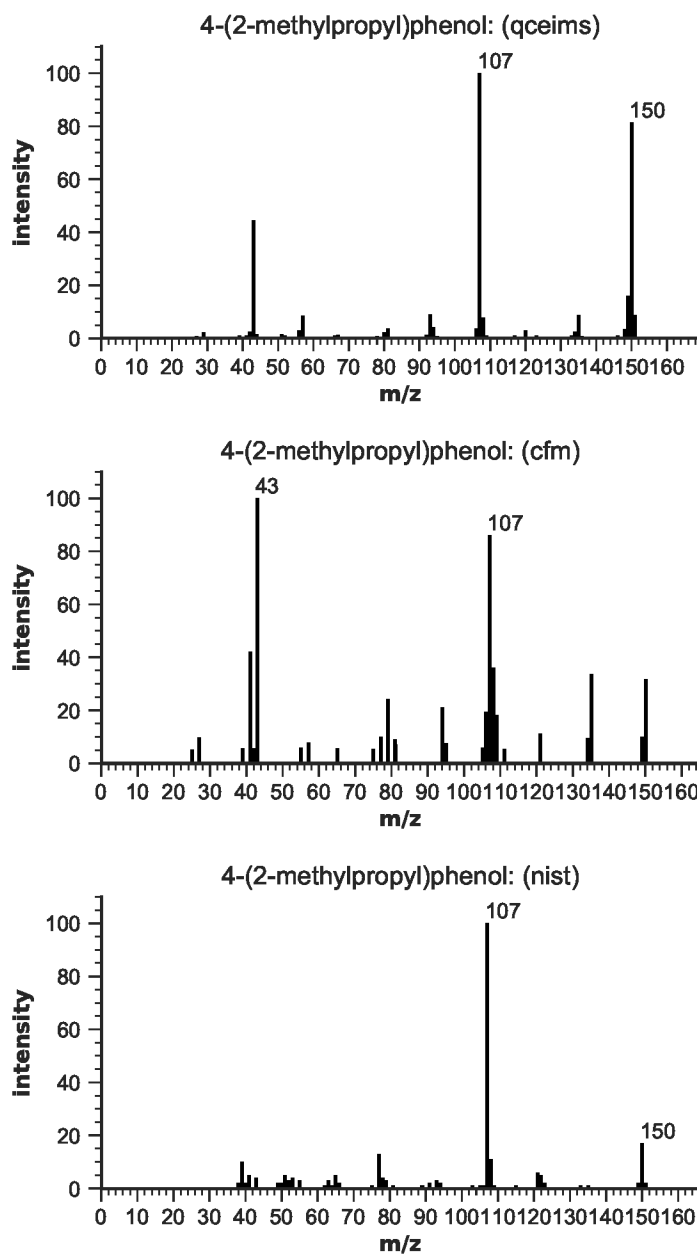


Figure S121: EIMS for 4-(2-methylpropyl)phenol

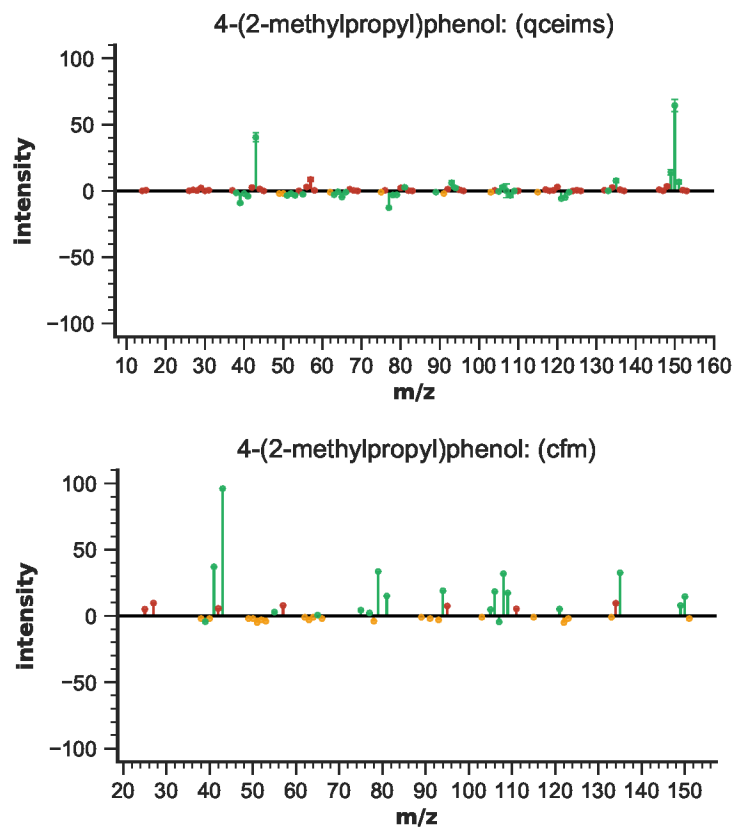


Figure S122: EIMS differences for 4-(2-methylpropyl)phenol

Differences in EIMS for an increased number of trajectories in the case of methylbutanoate

Peter R. Spackman,* Björn Bohman, Amir Karton, and Dylan Jayatilaka

*School of Chemistry and Biochemistry, University of Western Australia, Crawley, 6009,
WA, Australia*

E-mail: spackp01@student.uwa.edu.au

Accepted Article

The following page consists of three mass spectra predicted by QCEIMS for methylbutanoate, with 425, 510 and 595 trajectories used respectively. Of particular note is the lack of any real change in the $m/z = 74$ peak associated with McLafferty rearrangement.

Accepted Article

