Received: 27 June 2017 | Revised: 4 August 2017 | Accepted: 9 August 2017

DOI: 10.1002/qua.25460

FULL PAPER

WILEY International Journal of QUANTUM CHEMISTRY

Quantum chemical electron impact mass spectrum prediction for de novo structure elucidation: Assessment against experimental reference data and comparison to competitive fragmentation modeling

Peter R. Spackman | Björn Bohman | Amir Karton | Dylan Jayatilaka

School of Molecular Sciences, University of Western Australia, Crawley

Correspondence
Peter Robert Spackman, School of Molecular Sciences, University of Western Australia, Crawley, WA 6008.
Email: peterspackman@fastmail.com

Funding information
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.

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 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, 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.

KEYWORDS

machine learning, mass spectrometry, quantum chemistry, simulation

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 methods.[1] Indeed, MS is often the only feasible identification method[2–4] in contexts like the identification of semiochemicals,[5–7] where abundances of critical compounds may be in the ng range.

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.[8–13] 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}.[14] 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 (although the structure of the close matches may still provide insight).

© 2017 Wiley Periodicals, Inc.

In this article, 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 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 Grimme.[15] 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.[16,17] 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 rules[18] concerned the identification of molecular structure from MSp: the DENDRAL[19] project. Regardless of the successes (or failures[1,20]) of the original DENDRAL project, recently learning-based methods such as competitive fragment modeling (CFM)[21,22] for predicting EIMS spectra (CFM-EI) have gained more traction, and proved more successful. It should be noted, although, 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 matrix[23]). Unfortunately such calculations were, in general, time consuming to the point of being impractical. However, recently the QCEIMS program from Grimme et al.[24] 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,[15,24,25] coupled with knowledge of isotopic distributions,[26] QCEIMS has proven remarkably accurate. Other quantum chemical methods based on bond indices have also been proposed,[27] and are examined by us in related work.

The purpose of this article is to investigate the performance of, QCEIMS, which, although having been applied to many molecules,[15,24,25,28,29] 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 article was performed using the Python programming language. Notably, the numpy,[30] pandas[31] packages, and matplotlib[32] 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, that is, made up of a variety of functional groups, they must be relatively small molecules ($n_{\text{atom}} < 25$). 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

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)

No.	Name	Chemical formula	RRP /n_{isomers}
	Alcohols		
1	propan-1-ol	C_3H_8O	1/3
2	propan-2-ol	C_3H_8O	1/3
3	butan-1-ol	$C_4H_{10}O$	1/8
4	butan-2-ol	$C_4H_{10}O$	1/8
5	2-methylpropan-1-ol	$C_4H_{10}O$	1/8
6	2-methylpropan-2-ol	$C_4H_{10}O$	5/8
	Aldehydes		
7	propanal	C_3H_6O	1/7
8	2-methylpropanal	C_4H_8O	3/21
9	butanal	C_4H_8O	3/21
	Alkenes		
10	but-1-ene	C_4H_8	4/6
11	but-2-ene	C_4H_8	1/6
12	pent-1-ene	C_5H_{10}	3/13
13	pent-2-ene	C_5H_{10}	3/13
14	hex-1-ene	C_6H_{12}	6/30
15	hex-2-ene	C_6H_{12}	1/30
16	hex-3-ene	C_6H_{12}	9/30
	Amines		
17	propan-1-amine	C_3H_9N	1/4
18	propan-2-amine	C_3H_9N	1/4
19	butan-1-amine	$C_4H_{11}N$	1/8
20	butan-2-amine	$C_4H_{11}N$	1/8
21	2-methylpropan-1-amine	$C_4H_{11}N$	2/8
22	2-methylpropan-2-amine	$C_4H_{11}N$	1/8
	Benzenes		
23	benzene	C_6H_6	1/5
24	toluene	C_7H_8	3/13
25	ethylbenzene	C_8H_{10}	8/23
26	propylbenzene	C_9H_{12}	2/42
27	cumene	C_9H_{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
	Carboxylic acids		
31	propanoic acid	$C_3H_6O_2$	1/9
32	butanoic acid	$C_4H_8O_2$	7/27
33	2-methylpropanoic acid	$C_4H_8O_2$	3/27

(Continues)

TABLE 1 (Continued)

No.	Name	Chemical formula	RRP /n_{isomers}
34	2-methylbutanoic acid	$C_5H_{10}O_2$	3/47
35	3-methylbutanoic acid	$C_5H_{10}O_2$	5/47
	Esters		
36	methyl formate	$C_2H_4O_2$	1/3
37	methyl acetate	$C_3H_6O_2$	2/9
38	ethyl acetate	$C_4H_8O_2$	1/27
39	methyl propanoate	$C_4H_8O_2$	1/27
40	propyl acetate	$C_5H_{10}O_2$	NM/47
41	propan-2-yl acetate	$C_5H_{10}O_2$	4/47
42	methyl 2-methylpropanoate	$C_5H_{10}O_2$	1/47
43	ethyl propanoate	$C_5H_{10}O_2$	1/47
44	methyl butanoate	$C_5H_{10}O_2$	5/47
45	propyl propanoate	$C_6H_{12}O_2$	1/80
46	propan-2-yl propanoate	$C_6H_{12}O_2$	3/80
47	ethyl butanoate	$C_6H_{12}O_2$	4/80
48	methyl 2-methylbutanoate	$C_6H_{12}O_2$	2/80
49	methyl 3-methylbutanoate	$C_6H_{12}O_2$	7/80
50	propyl butanoate	$C_7H_{14}O_2$	2/83
51	propan-2-yl butanoate	$C_7H_{14}O_2$	NM/83
	Ketones		
52	acetone	C_3H_6O	1/7
53	butan-2-one	C_4H_8O	1/21
	Phenols		
54	phenol	C_6H_6O	2/4
55	4-methylphenol	C_7H_8O	2/17
56	4-ethylphenol	$C_8H_{10}O$	16/43
57	4-propylphenol	$C_9H_{12}O$	NM/112
58	4-propan-2-ylphenol	$C_9H_{12}O$	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

Ranking is relative to the total number isomers n_{isomers}. NM indicates that no match was found in the database.

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 Avogadro2[33] and then optimized using the generalized Amber force field.[34] MSp prediction calculations were then performed using the QCEIMS program version 2.26, as described by Grimme and coworkers[15,24] using the default parameters, namely: using the OM2-D3[35] semi-empirical quantum mechanical wavefunction method from MNDO99[36] version

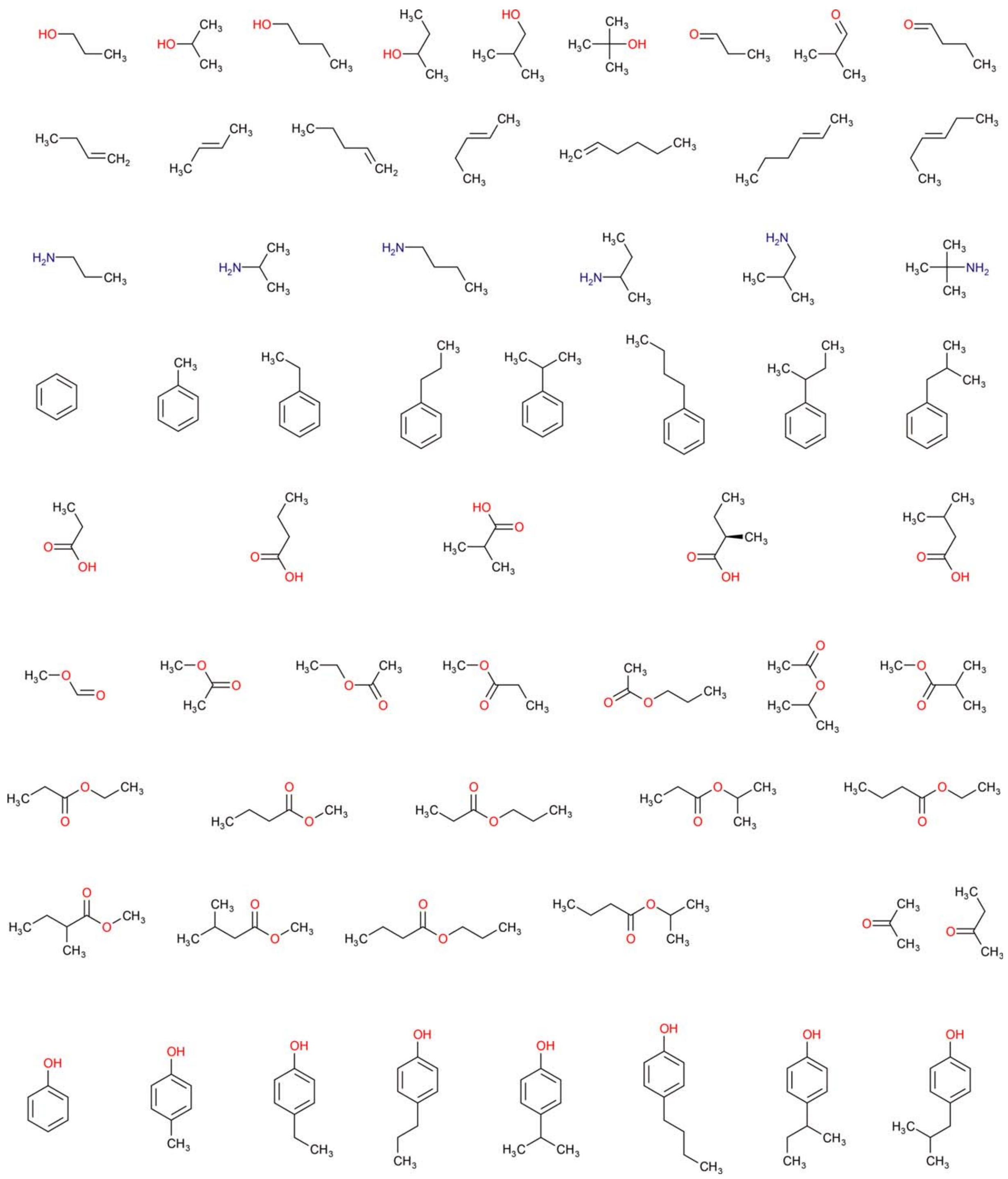

FIGURE 1 Structures of the molecules chosen for analysis grouped into chemical classes

7.0. The default QCEIMS parameters were used, with an initial temperature of 500 K, at 70 eV and the number of trajectories set to 25×$n_{\text{heavyatoms}}$. 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 Supporting Information 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.[22] 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 the mass weighting) is negligible for our purposes—see Ref. 10 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[3,22] has been used:

$$\mathrm{RRP}=\frac{r-1}{n-1} \tag{1}$$

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.[3]). Thus, a rank of one maps 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 Isenhour[37] 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 $N_{\mathrm{total}}=\sum_i N_i$ 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 $\sqrt{N_i}/N_{\mathrm{total}}$. 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%.[24]

3 | RESULTS

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

3.1 | The QCEIMS method

3.1.1 | 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 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.

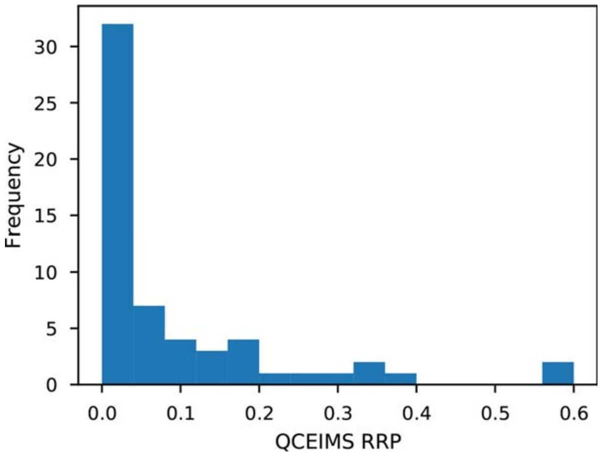

FIGURE 2 The number of molecules predicted correctly at a given rank for the QCEIMS method. There were four molecules which were not matched to any molecule in the NIST database

3.1.2 | 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.

3.1.3 | 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 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.

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

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

Overall mean RRPs are given in the final line. Only molecules in which both methods matched are compared.

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

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

Overall mean RMSDs and MADs are given in the final line.

3.1.4 | Mean RRPs for each chemical class

Table 2 summarizes some of the discussion above in a more quantitative way, presenting the mean RRPs for each chemical class. Although the results for ketones and aldehydes seem very good, there are too few systems to consider the mean RRPs meaningful for these two classes of compounds. However, we see that the esters and amines are very well predicted, with the means RRPs 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, for example, for **30** we have 1/74 and for **61** we have 2/163.

3.1.5 | 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 alchohols, alkenes, and ketones. In the Supporting Information, we present the mass spectra produced for all the compounds. Most informative are the differences between the predicted and observed spectra: peaks are colored 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 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 statics. Also, the difference 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 in presented in Figure 4, ethylbenzene, discussed in more detail later.

3.2 | The CFM-EI method

3.2.1 | Overall success rate

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

3.2.2 | 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 Supporting Information. The following observations can be made.

- The alcohols are well ranked except for one outlier (**3**) which is placed in the fifth position (this is not the same compound ranked fifth 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.

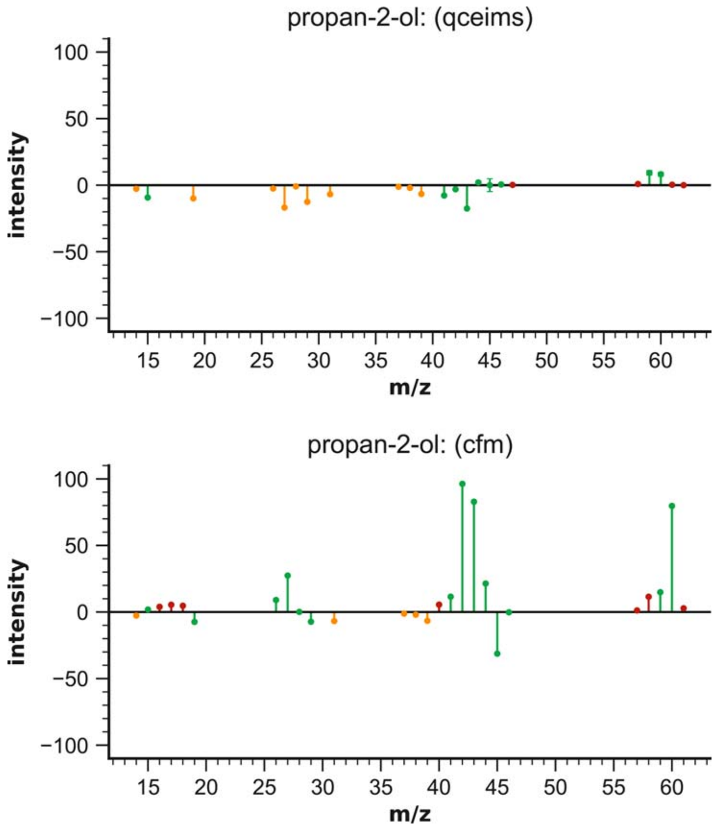

FIGURE 3 Difference between predicted (QCEIMS/CFM-EI) and experimental (NIST) spectra. Green bars indicate the presence of a correct peak, yellow indicates a missing peak, and red indicates a peak which is incorrect, that is, should not have been predicted

- 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.

3.2.3 | Mean RRPs 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 vs. 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 RRPs respectively 0.12 vs. 0.09).

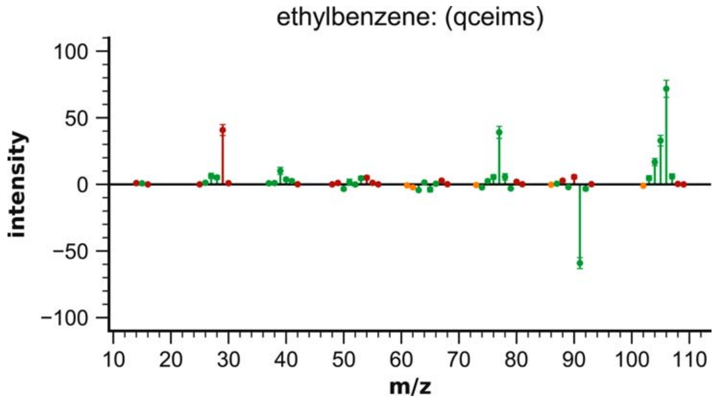

FIGURE 4 Difference the between predicted (QCEIMS) and experimental (NIST) mass spectrum for ethylbenzene

3.2.4 | 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.

3.2.5 | 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, that is, the spectrum shows fewer outlying peaks relative to experiment. Nevertheless, a qualitatively better MSp does not translate into a a superior ranking of the MSp (see discussion on the RRPs).

4 | DISCUSSION

4.1 | Detailed analysis of the predicted spectra

4.1.1 | 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.

Conversely, 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 Figure 3). CFM-EI particularly in the cases of 2-methylpropan-1-ol and, to a lesser extent, butan-1-ol predicts significant peaks nonexistent in the experimental MSp.

4.1.2 | 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 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.

4.1.3 | 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 prediciton techniques, although 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 ($\mathbf{M}-1$), 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.

4.1.4 | 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).[38] 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 pentenes and butenes. The regions $25<m/z<30$ and $40<m/z<45$, and the $m/z=55$ peak for pentene 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 predications are not accurate enough for the purposes of matching within sets of structural isomers.

4.1.5 | 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 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.

4.1.6 | 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 ($\mathbf{M}-15$) is seen with both methods in most alkyl benzenes.

4.1.7 | Phenols

Errors were present in the cases of propylphenol 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 article 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. Conversely, the worst ranked spectra were the alkenes and substituted phenols with mean relative ranking positions above 0.18. It should be noted, although, 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 bew 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 rearragements), 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.[1]), 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.

ACKNOWLEDGMENTS

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.

ORCID

Peter R. Spackman http://orcid.org/0000-0002-6532-8571
Amir Karton http://orcid.org/0000-0002-7981-508X

REFERENCES

[1] K. Scheubert, F. Hufsky, S. Böcker, *J. Cheminform.* **2013**, *5*, 12.
[2] I. K. Mun, F. W. Mclafferty, in *Supercomputers in Chemistry* (pp. 117–124), n.d. doi:10.1021/bk-1981-0173.ch008.
[3] A. Kerber, M. Meringer, C. Rücker, *Croat. Chem. Acta* **2006**, *79*, 449.
[4] T. Kind, O. Fiehn, *Bioanal. Rev.* **2010**, *2*, 23.
[5] B. Bohman, R. D. Phillips, M. H. M. Menz, B. W. Berntsson, G. R. Flematti, R. A. Barrow, et al. *New Phytol.* **2014**, *203*, 939.
[6] B. Bohman, G. R. Flematti, R. A. Barrow, *J. Mass Spectrom.* **2015**, *50*, 987.
[7] B. Bohman, R. D. Phillips, G. R. Flematti, R. A. Barrow, R. Peakall, *Angew. Chem. Int. Ed. Engl.* **2017**, *56*, 8455.
[8] H. S. Hertz, R. A. Hites, K. Biemann, *Anal. Chem.* **1971**, *43*, 681
[9] S. E. Stein, *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 316.
[10] S. E. Stein, D. R. Scott, *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 859.
[11] S. E. Stein, *J. Am. Soc. Mass Spectrom.* **1999**, *10*, 770.
[12] K. X. Wan, I. Vidavsky, M. L. Gross, *J. Am. Soc. Mass Spectrom.* **2002**, *13*, 85.
[13] F. W. McLafferty, D. A. Stauffer, S. Y. Loh, C. Wesdemiotis, *J. Am. Soc. Mass Spectrom.* **1999**, *10*, 1229.
[14] C. M. Dobson, *Nature* **2004**, *432*, 824.
[15] C. A. Bauer, S. Grimme, *J. Phys. Chem. A* **2016**, *120*, 3755.
[16] J. Zupan, J. Gasteiger, *Anal. Chim. Acta* **1991**, *248*, 1.
[17] M. Jalali-Heravi, M. H. Fatemi, *Anal. Chim. Acta* **2000**, *415*, 95.
[18] S. J. Russell, P. Norvig, *Artificial Intelligence. A Modern Approach*, Prentice Hall, NJ: Englewood Cliffs, **2010**.
[19] B. G. Buchanan, J. Lederberg, E. Feigenbaum, *The Heuristic DENDRAL Program for Explaining Empirical Data* **1971**.
[20] J. Gasteiger, W. Hanebeck, K. P. Schulz, *J. Chem. Inf. Model.* **1992**, *32*, 264.
[21] F. Allen, R. Greiner, D. Wishart, *Metabolomics* **2015**, *11*, 98.
[22] F. Allen, A. Pon, R. Greiner, D. Wishart, *Anal. Chem.* **2016**, *88*, 7689.
[23] N. N. Bogolubov, *Introduction to Quantum Statistical Mechanics*, Singapore: World Scientific **2010**.

[24] S. Grimme, *Angew. Chem. Int. Ed. Engl.* **2013**, *52*, 6306.

[25] S. Grimme, C. A. Bauer, *Org. Biomol. Chem.* **2014**, *12*, 8737.

[26] D. Valkenborg, I. Mertens, F. Lemière, E. Witters, T. Burzykowski, *Mass Spectrom. Rev.* **2012**, *31*, 96.

[27] I. Mayer, M. Knapp-Mohammady, S. Suhai, *Chem. Phys. Lett.* **2004**, *389*, 34.

[28] C. A. Bauer, S. Grimme, *J. Phys. Chem. A* **2014**, *118*, 11479.

[29] V. Ásgeirsson, C. A. Bauer, S. Grimme, *Chem. Sci.***2017**, *52*, 6306.

[30] S. van der Walt, S. C. Colbert, G. Varoquaux, *Comput. Sci. Eng.* **2011**, *13*, 22.

[31] W. McKinney, in *Proc. 9th Python Sci. Conf*. (Eds: v. d. W. Se, J. Millman, pp. 51–56) **2010**. https://pdfs.semanticscholar.org/f6da/c1c52d3b07c993fe52513b8964f86e8fe381.pdf

[32] J. D. Hunter, *Comput. Sci. Eng.* **2007**, *9*, 90.

[33] M. D. Hanwell, D. E. Curtis, D. C. Lonie, T. Vandermeersch, E. Zurek, G. R. Hutchison, *J. Cheminform.* **2012**, *4*, 17.

[34] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case, *J. Comput. Chem.* **2004**, *25*, 1157.

[35] W. Weber, W. Thiel, *Theor. Chem. Acc.* **2000**, *103*, 495.

[36] W. Thiel, *MNDO99*, *version 7.0* **2004**.

[37] G. T. Rasmussen, T. L. Isenhour, *J. Chem. Inf.* **1979**.

[38] F. W. McLafferty, F. Turecek, *Interpretation of Mass Spectra*, Organic chemistry series, Sausalito, CA: University Science Books, (pp. 226–230). **1993**.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Spackman PR, Bohman B, Karton A, Jayatilaka D. Quantum chemical electron impact mass spectrum prediction for de novo structure elucidation: Assessment against experimental reference data and comparison to competitive fragmentation modeling . *Int J Quantum Chem*. 2018;118:e25460. https://doi.org/10.1002/qua.25460