



# Predicting the primary fragments in mass spectrometry using *ab initio* Roby–Gould bond indices

Khidhir Alhameedi<sup>1,2</sup>  | Björn Bohman<sup>1</sup> | Amir Karton<sup>1</sup>  | Dylan Jayatilaka<sup>1</sup>

<sup>1</sup>School of Molecular Sciences, The University of Western Australia, 35 Stirling Hwy, Crawley WA 6009, Australia

<sup>2</sup>Department of Chemistry, College of Education for Pure Science, University of Karbala, Karbala, Iraq

## Correspondence

Khidhir Alhameedi, School of Molecular Sciences, University of Western Australia, 35 Stirling Highway Nedlands 6009, Australia.

Email: khidhir.abdaloussein@gmail.com

and

Dylan Jayatilaka, School of Molecular Sciences, University of Western Australia, 35 Stirling Highway Nedlands 6009, Australia.

Email: dylan.jayatilaka@uwa.edu.au

## Funding information

Higher Committee for Education Development in Iraq (HCED) (PhD scholarship); Australian Research Council, Grant/Award Number: DE160101313

## Abstract

There is currently a lack of computational methods supporting the elucidation of unknown compounds by mass spectrometry. In this study, we develop and evaluate seven different protocols, based on the *ab initio* Roby–Gould bond indices [Gould *et al.*, *Theor. Chem. Acc.*, 2008, 119, 275] for predicting the mass-to-charge ratio of the highest intensity peak (base peak) in electron impact mass spectra. The protocols are applied to a dataset of 75 molecules, including five directly targeted semiochemicals. The Roby–Gould bond indices are also surveyed exhaustively, for the first time, for a dataset of 103 molecules with 682 C–C bonds. For neutral species we find that the bond indices are, as may be expected, highly correlated with the bond length; for cations, although there is a correlation, the bond indices are more variable. One of our protocols, protocol MG, correctly predicts the base peak in the mass spectra for 65 out of 75 cases. The correct base peak was calculated for three out of five targeted natural products.

## KEYWORDS

mass spectrometry, predicting base peak, Roby–Gould bond indices

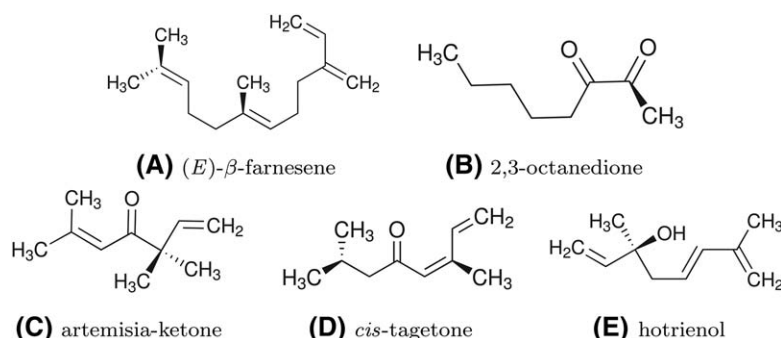
## 1 | INTRODUCTION

The determination of chemical structure, the relative arrangement of atomic nuclei in three dimensional space, is fundamental to chemistry. The two most widely used experimental techniques to elucidate chemical structure are nuclear magnetic resonance (NMR) spectroscopy<sup>[1]</sup> and X-ray crystallography.<sup>[2]</sup> However, these methods are not suitable when the targeted compound is present in low amounts or cannot be crystallized. Instead, mass spectrometry methods, for example, electron impact mass spectrometry (EI-MS), coupled to chromatographic separation methods, for example, gas chromatography (GC-MS), are routinely used for such applications.<sup>[3]</sup>

EI-MS is most commonly used to identify compounds with known mass spectra by comparing experimental data with commercial databases. Identification of new compounds, where reference spectra are unavailable, is much more challenging.<sup>[4]</sup> To confirm the structure of an unknown compound, one must examine the mass fragmentations; the mechanisms involved, and compare with literature data. It is extremely difficult to reduce the candidates to a single compound, and often it is required to laboriously synthesize multiple candidate compounds, subsequently compare mass spectra, and in an interactive manner predict new candidates until the spectra match.

One particular application of interest to us, where GC-MS is the key method for natural product identification, is insect semiochemicals critical in pest control and evolutionary studies.<sup>[5–9]</sup> Some examples of compounds targeted in this work are shown in Figure 1. Yet another example of great interest would be the identification of unknown volatile organic compounds in atmospheric pollution chemistry applications with GC-MS.<sup>[10]</sup> Furthermore, with efficient tools facilitating correlations of chemical structure and mass spectrum, mass spectrometric protocols could become affordable alternatives to NMR spectroscopy in many routine applications.

With the introduction of high resolution mass spectrometers coupled to GC in many laboratories, the chemical formulae for molecular ions and fragments can easily be obtained. Consequently, the main obstacle is to predict, *de novo*, the structure of the unknown molecule from its molecular



**FIGURE 1** The molecular structure of specifically selected natural products. Note that compounds C, D and E are isomers with identical formula  $C_{10}H_{16}O$

formula. Unfortunately the number of possible candidate isomers grows combinatorically with the size of the molecule. Therefore, some form of modeling would be invaluable to identify or narrow down the possible candidate structures.

Machine-learning or artificial intelligence methods were the first to be used in conjunction with mass spectrometry for predicting the structure of a molecule from its mass spectrum. In fact, the Dendral project for just this purpose was also one of the first significant machine-learning projects ever attempted.<sup>[11]</sup> Since then a variety of methods that adopt artificial intelligence and machine-learning approaches to elucidate substructures from mass spectra have been reported.<sup>[12–14]</sup> One particular example is competitive fragment modeling for metabolite identification (CFM-ID) for predicting EI-MS.<sup>[15,16]</sup>

A different modeling approach is possible using *ab initio* quantum chemistry. In *ab initio* quantum chemistry, it is possible to predict the probability of obtaining certain mass fragments unbiased by experimental inputs used in training or learning algorithms.<sup>[17]</sup> Recently, Grimme<sup>[18]</sup> reported such a predictive method that accounts for both the thermodynamics and kinetics of the molecular fragmentation processes. This groundbreaking method is unfortunately computationally demanding and time-consuming, limiting its use in identifying molecular isomers from mass spectra.<sup>[19]</sup> In this context, an important realization which has not yet been emphasized in the literature is that it may not be necessary to predict the full mass spectra to identify the structure of the unknown compound: rather, it may be sufficient to reliably predict just a few critical and unique parts of the spectrum, so-called “fingerprints” or “signatures.” More recent, a comparison between the method of Grimme and coworkers (*ab initio* Quantum Chemical Electron Impact Mass Spectrum [QCEIMS]) with the CFM-ID model has been made.<sup>[20]</sup> This study showed that the performance of the QCEIMS for predicting EI-MS is more efficient than CFM-ID model.

In this work, we continue investigations based on *ab initio* wavefunction methods for mass spectra predictions. However, rather than trying to model the fragmentation process we focus on using nondynamical *ab initio* bond indices for predicting only the mass-to-charge ratio of the base peak in the spectrum. Importantly, the density functional theory (DFT) methods we use here are already sufficiently fast to use for small to medium sized molecules.

Mayer and Gomory were the first to use *ab initio* methods to predict bond orders and the primary fragment in the mass spectrum,<sup>[21]</sup> and they showed that large differences between the bond orders of the neutral and cation wavefunction were useful for predicting the primary fragmentation product. The argument supporting this hypothesis is not clear, but it seems to be that the bond undergoing the “most weakening” (as measured by bond index change) is the one most likely to fragment. We consider this to be essentially a kinetic argument—the weakening of the bond being associated with a reduction in the barrier for bond breaking—coupled to a Hammond-like postulate that information on the transition state is available from the nearby equilibrium structure.<sup>[22]</sup> In calculating the bond index changes, Mayer and Gomory used a cation wavefunction obtained by simply removing the HOMO or second HOMO of the neutral molecule (kept at the same geometry)—the so-called vertical “quasi-Koopmans approximation.”

In this article, we extend Mayer and Gomory’s idea and test a range of molecules with various chemical structures. We consider using not only the difference in the bond orders between the neutral and cation, but the bond orders of the neutral and cation species themselves, as well as a combination of all three quantities. Unlike Mayer and Gomory we use the Roby–Gould bond index<sup>[23]</sup> which is known to be stable with basis set extension. Furthermore, we re-optimize the geometry after removing one electron, that is, we consider an adiabatic rather than vertical transition. We restrict our attention to C–C bond cleavages, and we do not attempt to predict hydrogen rearrangements.<sup>[24]</sup> As a necessary preliminary to this work we also characterize and describe the behavior of the values of the Roby–Gould bond indices for a series of chemical structures.

## 2 | METHODS AND MATERIALS

### 2.1 | Roby–Gould bond indices

Roby–Gould bond indices have been described in the Ref. [23] hence only a brief summary of their properties is given here.

In short there are two Roby–Gould bond indices: one covalent  $c_{AB}$  and one ionic  $i_{AB}$ . In fact, the Roby–Gould bond indices are calculated as expectation values from a wavefunction, in exactly the same way that all quantum mechanical properties;  $c_{AB}$  and  $i_{AB}$  accord with the usual definition of a bond order.

$$c_{AB} = \left\langle \frac{R_{AB}}{2|R_{AB}|} \right\rangle, \text{ and} \quad (1)$$

$$i_{AB} = \left\langle \frac{I_{AB}}{2|I_{AB}|} \right\rangle \quad (2)$$

These are defined between two regions  $A$  and  $B$  which are usually (but not necessarily) associated with two atoms. The operator  $R_{AB}$  is the Roby shared population operator, while  $I_{AB}$  is Gould's population difference operator,

$$R_{AB} = P_A + P_B - P_{AB}, \quad (3)$$

$$I_{AB} = P_A - P_B. \quad (4)$$

Here,  $A$  and  $B$  label subspaces  $V_A$  and  $V_B$ , respectively, which are supposed to represent a pair of "atoms."  $P_A$  and  $P_B$  are the idempotent projection operators associated with the spaces  $V_A$  and  $V_B$ . In this work  $V_A$  (resp.  $V_B$ ) is equal to the span (i.e., the subspace obtained by all linear combinations) of the "occupied" atomic natural spin orbitals (ANOs) obtained from a spherically averaged isolated-atom unrestricted BLYP calculation for atom  $A$  (resp. atom  $B$ ) located at its position in the actual molecule, and using the atomic basis set for that atom. By "occupied," we mean a spherically averaged ANOs with a population larger than 0.05 electrons.  $P_{AB}$  is the projection operator onto  $V_A \oplus V_B$  (Although different choices for the subspaces  $V_A$  and  $V_B$  are possible, and we do not claim that our choice is optimal, the atomic subspaces used should have some overlap. For example, although one could define  $P_A$  to be associated with the subspace spanned by a set of Dirac delta functions on a Bader atomic basin,<sup>[25]</sup> this would result in no shared electrons; in the words of Parr and Yang<sup>[26]</sup> p. 222 this kind of exclusive and disjoint partitioning causes the chemical bond to "vanish into thin air." Alternatively, from a quantum subsystem point of view, the trace distance between such basins would be equal to 1, i.e., such basis are orthogonal and separate.<sup>[27]</sup> This is not to say that such basis are not useful for chemistry, only that they are not useful for performing a Roby–Gould bonding analysis.). Finally, the notation  $(1/|X|)$  refers to the pseudoinverse of the operator  $X = \sqrt{X^\dagger X}$ .

The expectation value of  $R_{AB}$  is Roby's shared electron population, while the expectation value of  $I_{AB}$  is the difference in electron population between the two atoms. The eigenvalues of  $R_{AB}$  and  $I_{AB}$  are known to occur in pairs with opposite value: the positive-eigenvalue eigenstates represent "bonding" states, while the negative-eigenvalue eigenstates represent "antibonding" states. Further, the pairs of eigenstates of  $I_{AB}$  are related to those of  $R_{AB}$  by a 45 degree rotation, as per the Pythagorean relationship

$$R_{AB}^2 + I_{AB}^2 = P_{AB}^2. \quad (5)$$

The total Roby–Gould bond index  $\tau$  is defined as following:

$$\tau_{AB} = \sqrt{c_{AB}^2 + i_{AB}^2}. \quad (6)$$

A more detailed description of the Roby–Gould bond index method in the Appendix to facilitate reproduction of our results.

It has been established that the bond indices from the Roby–Gould method are generally chemically acceptable and are stable to basis set extension.<sup>[23]</sup> In this study, bond indices were calculated using the free Tonto program package<sup>[28]</sup> using the text version of the Gaussian checkpoint file, which containing the geometry and Kohn–Sham orbitals.

## 2.2 | Wavefunctions

Wavefunctions were calculated at the BLYP/6-31G(d) level with Cartesian Gaussian basis sets, using the Gaussian 09 program.<sup>[32]</sup> Closed-shell species used the restricted formalism, while cation wavefunctions employed the unrestricted formalism. Initial geometries were generated by using universal force field (UFF) method<sup>[33]</sup> in the Avogadro program.<sup>[34]</sup> Optimized geometries were used throughout, corresponding to the adiabatic rather than vertical electronic transition. The geometries and output files are deposited for open access figshare.com under <https://figshare.com/s/42857cc7421faf14cbc2>.

## 2.3 | Dataset A

To benchmark the bond index method, we formulate dataset A comprising 75 molecules; **1–70** are given in Figure 2. The chemical formulae and number of isomers are detailed in Table 1. This set also includes five specifically targeted semiochemicals, **71–75**, which are given as **a–e**, respectively, in Figure 1. The molecules in this dataset were chosen to have a variety of the most common functional groups and bond types, hence containing alcohols, amines, alkanes, alkenes, aldehydes, ketones, thiols, and phenols. All of the selected compounds and corresponding structural analogues have associated experimental mass spectra available online in NIST web book.<sup>[35]</sup>

## 2.4 | Dataset B

There have only been a few publications concerning the Roby–Gould bond indices.<sup>[23,36,37]</sup> Therefore, it is necessary to examine and characterize the distribution of typical values of this bond index to see if the values obtained are chemically sensible. For this purpose, which is distinct from the

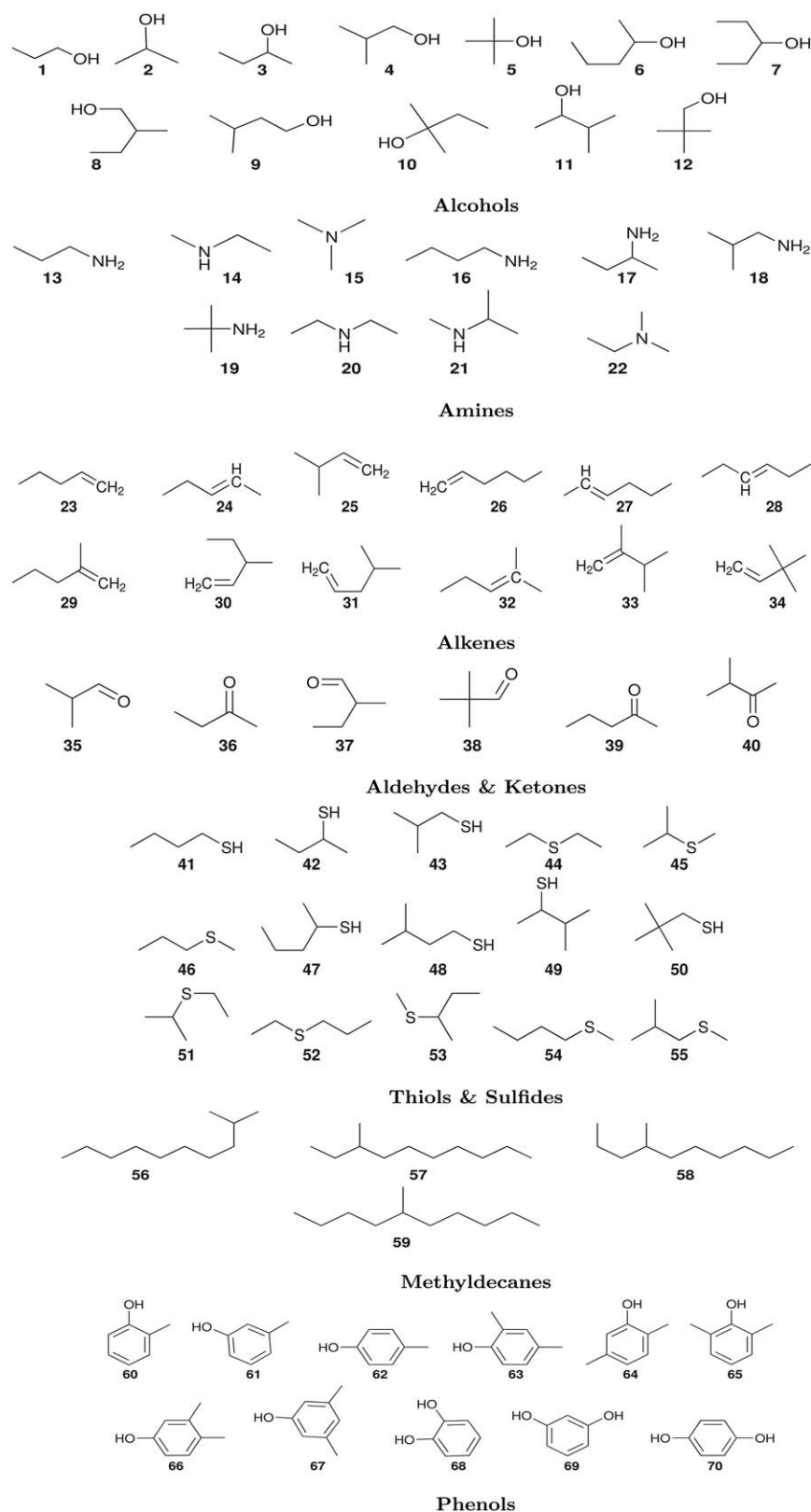


FIGURE 2 Structures of compounds 1–70 as part of dataset A

base peak predictions, we extend dataset A with dataset B, which comprised another 28 molecules; these are listed in Table S1 in the Supporting Information. The molecules in dataset B include semiochemicals such as chiloglottones<sup>[38,39]</sup> and hydroxymethylpyrazines.<sup>[40,41]</sup> The combination of the two datasets comprise 103 molecules in total.

TABLE 1 Chemical formulae and the number of isomers for dataset A

Chemical Formula	No. of isomers	Isomers with same base peak	No. correct/near miss
<b>Alcohols:</b>	12		11/1
C <sub>3</sub> H <sub>7</sub> OH	2	(2)	
C <sub>4</sub> H <sub>9</sub> OH	3	(3)	
C <sub>5</sub> H <sub>11</sub> OH	7	(0,2,1)	
<b>Amines:</b>	10		10/0
C <sub>3</sub> H <sub>9</sub> N	3	(3)	
C <sub>4</sub> H <sub>11</sub> N	7	(2,1,1)	
<b>Alkenes:</b>	12		10/1
C <sub>5</sub> H <sub>10</sub>	3	(1,1)	
C <sub>6</sub> H <sub>12</sub>	9	(2,0,1,1)	
<b>Aldehydes &amp; Ketones:</b>	6		6/0
C <sub>4</sub> H <sub>8</sub> O	2	(0,1)	
C <sub>5</sub> H <sub>10</sub> O	4	(2,1)	
<b>Thiols &amp; Sulfides:</b>	15		12/2
C <sub>4</sub> H <sub>10</sub> S	6	(4,1)	
C <sub>5</sub> H <sub>12</sub> S	9	(4,1,1)	
<b>Methyldecane:</b>	4		2/0
C <sub>11</sub> H <sub>24</sub>	4	(1,0,1)	
<b>Phenols:</b>	11		11/0
C <sub>7</sub> H <sub>7</sub> OH	3	(0,0,1)	
C <sub>8</sub> H <sub>9</sub> OH	5	(0,0,0,0,1)	
C <sub>6</sub> H <sub>6</sub> O <sub>2</sub>	3	(0,0,1)	
<b>Selected natural products:</b>	5		4/0
C <sub>15</sub> H <sub>24</sub>	1	(1)	
C <sub>8</sub> H <sub>14</sub> O <sub>2</sub>	1	(1)	
C <sub>10</sub> H <sub>16</sub> O	3	(3)	

The number of isomers with the same mass-to-charge ratio for the base peak is shown as a list, specifically  $(n_1, n_2, n_3, \dots)$  indicates there are  $n_1$  spectra which have a distinct base peak,  $n_2$  pairs of spectra which have the same base peak,  $n_3$  triples of spectra which have the same base peak, etc. The total number of correct base peak predictions is given ("No. correct"), as well as the number predicted within one or two mass units of the base peak ("Near miss").

## 2.5 | Protocols for predicting base peaks

Here, we present seven different protocols for predicting the base peak in EI-MS. We call these protocols: MBI0, the smallest bond index in the neutral species; MBI+, the smallest bond index in the cations; M0, the smallest CC bond index in the neutral species; M+, the smallest CC bond index in the cation species; MG, the biggest change in bond index between two carbon atoms in the neutral and cation species; MC, the two or three of M0, M+, and MG protocols agree with the same C—C bond; and MCMG, a combination of protocols MC and MG.

For each protocol, we test whether the calculated bond index corresponds to the base peak in experimental mass spectra. Scheme 1 summarizes how the protocols MG and MC are built from the M0 and M+ protocols.

### 2.5.1 | MBI0

Minimum bond index in the neutral species of dataset A overall pairs of atoms,

$$\tau_{\min}^{\text{neutral}} = \min \tau^{\text{neutral}}, \text{ where} \quad (7)$$

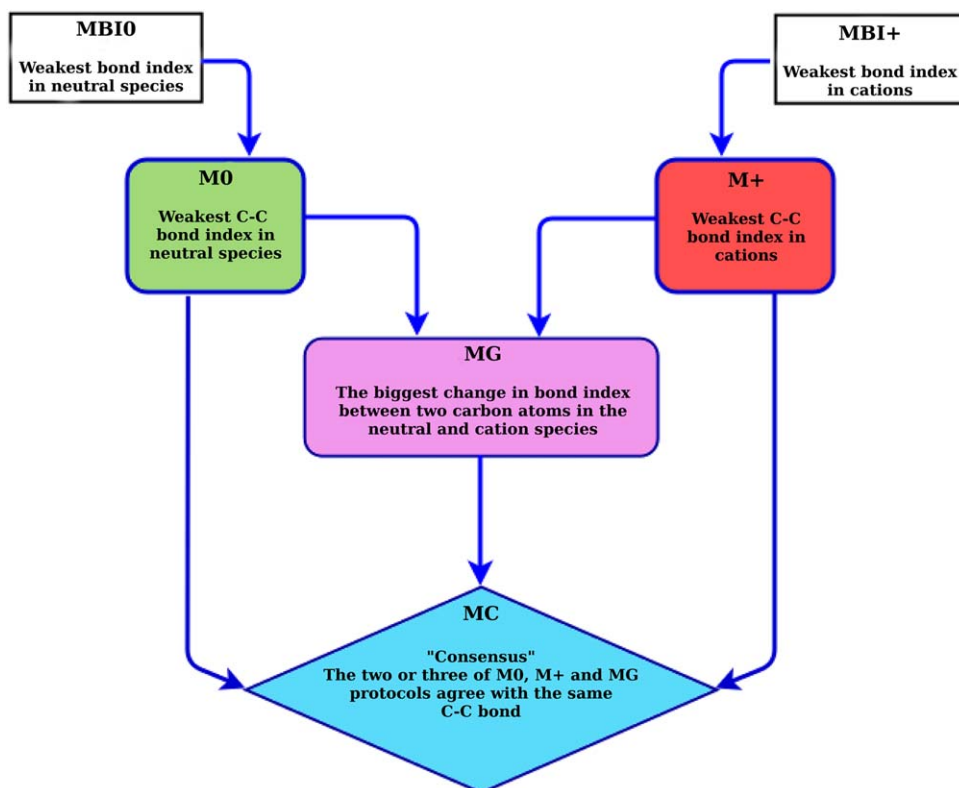
$$\tau^{\text{neutral}} = \{ \tau_{A_i A_j}^{\text{neutral}} | i, j = 1, n_A \}, \quad (8)$$

and where  $n_A$  is the number of atoms in the molecule.

### 2.5.2 | MBI+

Minimum bond index in the cations of dataset A with all types of bonds,

$$\tau_{\min}^{\text{cation}} = \min \tau^{\text{cation}}, \text{ where} \quad (9)$$



**SCHEME 1** The dependency of the six protocols used to predict the base peak in EI-MS. The colors used for each protocol correspond to those in Figure 5. The central panel in pink is the modified protocol of Mayer and Gomory (protocol MG)

$$\tau^{\text{cation}} = \{\tau_{A_i A_j}^{\text{cation}} | i, j = 1, n_A\}. \quad (10)$$

### 2.5.3 | M0

Weakest carbon–carbon bond in the neutral,

$$\tau_{\min(\text{CC})}^{\text{neutral}} = \min \{\tau_{C_i C_j}^{\text{neutral}} | i, j = 1, n_C\}. \quad (11)$$

### 2.5.4 | M+

Weakest carbon–carbon bond in the cation,

$$\tau_{\min(\text{CC})}^{\text{cation}} = \min \{\tau_{C_i C_j}^{\text{cation}} | i, j = 1, n_C\}. \quad (12)$$

### 2.5.5 | MG

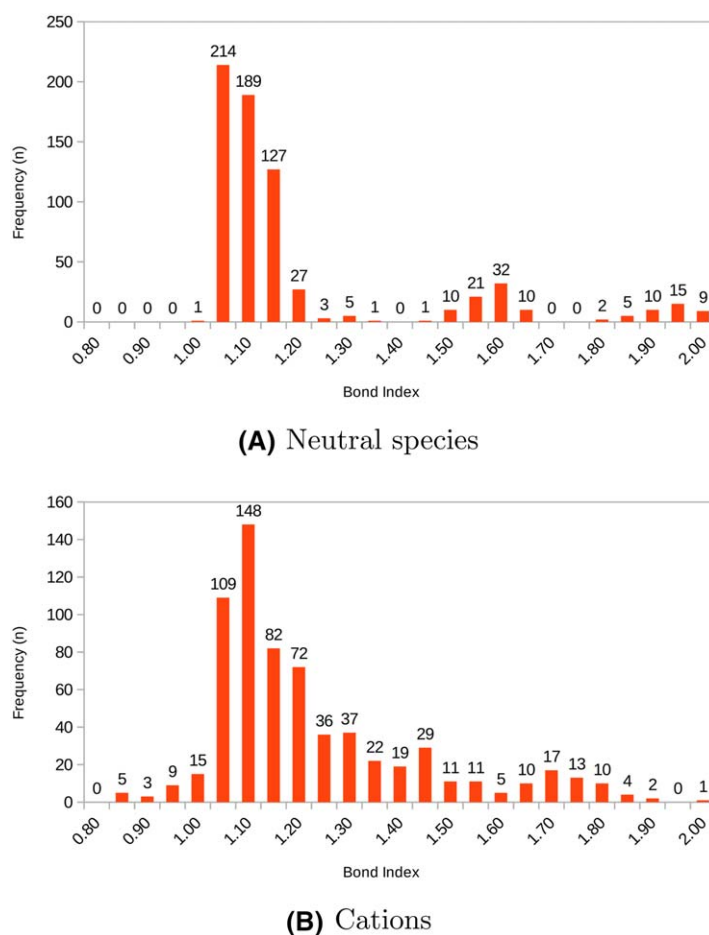
The biggest change in bond index between two carbon atoms in the neutral and cation species,<sup>[21]</sup>

$$\Delta_{\text{CC}}^{\max} = \max \Delta_{\text{CC}}, \text{ where} \quad (13)$$

$$\Delta_{\text{CC}} = \{\Delta(C_i C_j) | i, j = 1, n_C\}, \text{ and} \quad (14)$$

$$\Delta(C_i C_j) = \tau_{C_i C_j}^{\text{neutral}} - \tau_{C_i C_j}^{\text{cation}}. \quad (15)$$

Here,  $n_C$  is the number of carbon atoms in the molecule, and  $\tau_{C_i C_j}^{\text{neutral}}$  and  $\tau_{C_i C_j}^{\text{cation}}$  are the bond index values in the neutral and cation species, respectively. Unlike the work of Mayer and Gomory, the geometries of both species are optimized and we use fully relaxed wavefunctions. We then test whether the largest peak in the mass spectrum corresponds to the fragment(s) obtained when bond  $C_k C_l$  breaks, where  $C_k C_l$  corresponds to  $\Delta_{\text{CC}}^{\max}$ ; and we consider the result correct if the mass-to-charge ratio value ( $m/z$ ) of either fragment matches. Peaks whose magnitude are within 10% are regarded as equal in the mass spectra.



**FIGURE 3** Distributions of C—C Roby-Gould bond index values for (A) neutral species and (B) cations for combined datasets A and B

### 2.5.6 | MC

A “consensus” method: if two or more of the above three protocols (M0, M+, and MG) agree in predicting that the same bond  $C_kC_l$  breaks, then that bond is regarded as the one that actually breaks; otherwise no prediction is made.

### 2.5.7 | MCMG

We test a combination of protocols MC and MG whereby if MC does not give a prediction, we use the result from protocol MG (the only difference to protocol MC is that a prediction is always made in this method).

## 3 | RESULTS AND DISCUSSION

### 3.1 | Characterization of the C—C bond index

We have analysed the Roby-Gould bond indices for C—C bonds in the combined datasets A and B. These bonds are the most important for our protocol to predict the base peak. The distribution of values is presented in Figure 3. We observe for these closed shell neutral species that the bond indices are peaked around values of 1.00, 1.50, and 2.00 in accord with expectations of C—C single, aromatic, and double bonds. In contrast, for the associated cation species, the bond indices are more evenly distributed: indeed, the peaks around the value of 2.00, corresponding to C=C double bonds in the neutral species, are virtually absent in the cation case (Figure 3B). Consequently, whereas the bond indices in the neutral species are easily estimated using simple Lewis structure diagrams, those for cations require quantum mechanical wavefunction calculations.

#### 3.1.1 | Correlation between bond index and bond length

Figure 4 shows plots of the 1362 C—C Roby-Gould bond indices as a function of the bond length, for both neutral and cationic species.

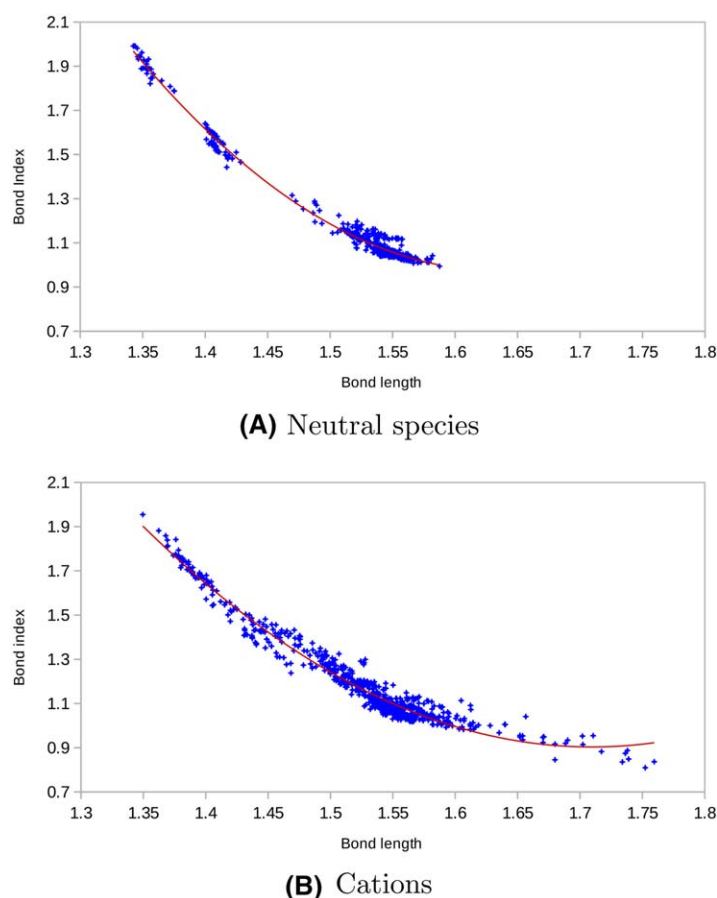


FIGURE 4 Bond index versus bond length (in Å) for C—C bonds in datasets A and B for (A) the neutral species and (B) the cations

We observe that the data points lie on a quadratic curve with a correlation coefficient  $R^2$  of 0.99 and 0.96 for neutral species and cations, respectively. Again, it can be seen that the cationic species bond indices are widely distributed across the range of bond lengths whereas the neutral species bond lengths are grouped into three distinct regions.

### 3.1.2 | Adequacy of the BLYP wavefunctions

An important conclusion to note from Figure 4 is that any small change in molecular geometry will only slightly affect the Roby–Gould bond indices that are obtained. Specifically, at the level of BLYP/6–31G(d) that we have used, the bond length changes relative to high level CCSD(T) calculations are only  $0.01\text{Å}$ ,<sup>[42]</sup> hence our results show that changes in the bond lengths of this order will hardly change the bond indices at this level of theory.

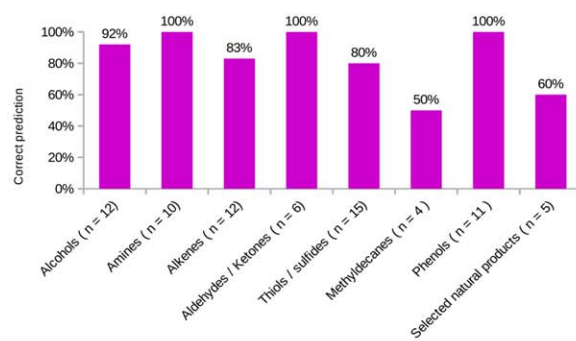
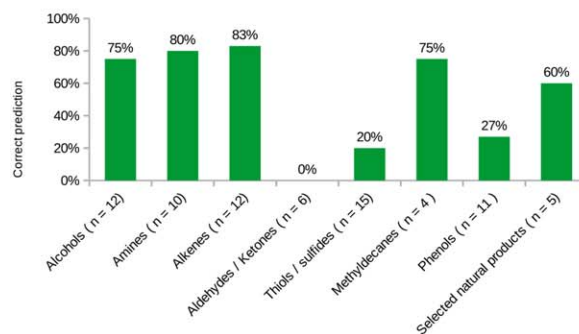
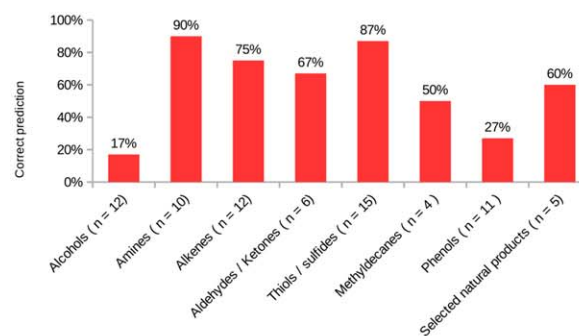
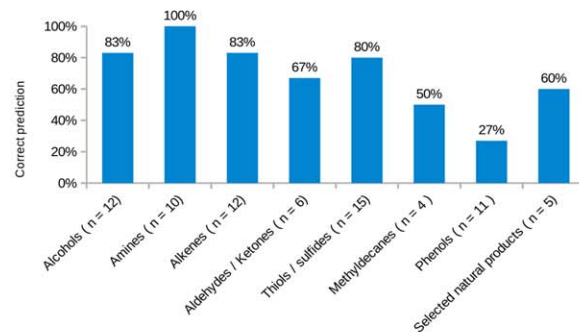
## 3.2 | Base peak as a tool for identifying structure

Table 1 lists the number of isomers corresponding to each chemical formula, and it further lists in the third column how many of these isomers have different base peaks. We observe that only 22 of the 70 molecules in dataset A are distinguishable from their base peak value. Indeed for the phenols with formula  $\text{C}_8\text{H}_9\text{OH}$  all five isomers in our dataset have the same base peak, the molecular ion (which as we will see in Table 1, is correctly predicted using protocol MG; the molecular ion is the base peak because the protocol predicts that the ring is opened). Thus, even in this small dataset the use of base peak is not sufficient to determine the structure from a set of mass spectra for isomers. The base peak is nevertheless useful, because if we consider how many of these are distinct up to a pair of compounds then the table shows that 35 out of the 75 molecules are distinguishable; the ability to narrow down the structure of the compound to a pair of structures is still very useful considering how labor intensive the standard procedure of identifying unknown compounds from EI-MS is.

## 3.3 | Performance of the protocols for predicted base peak

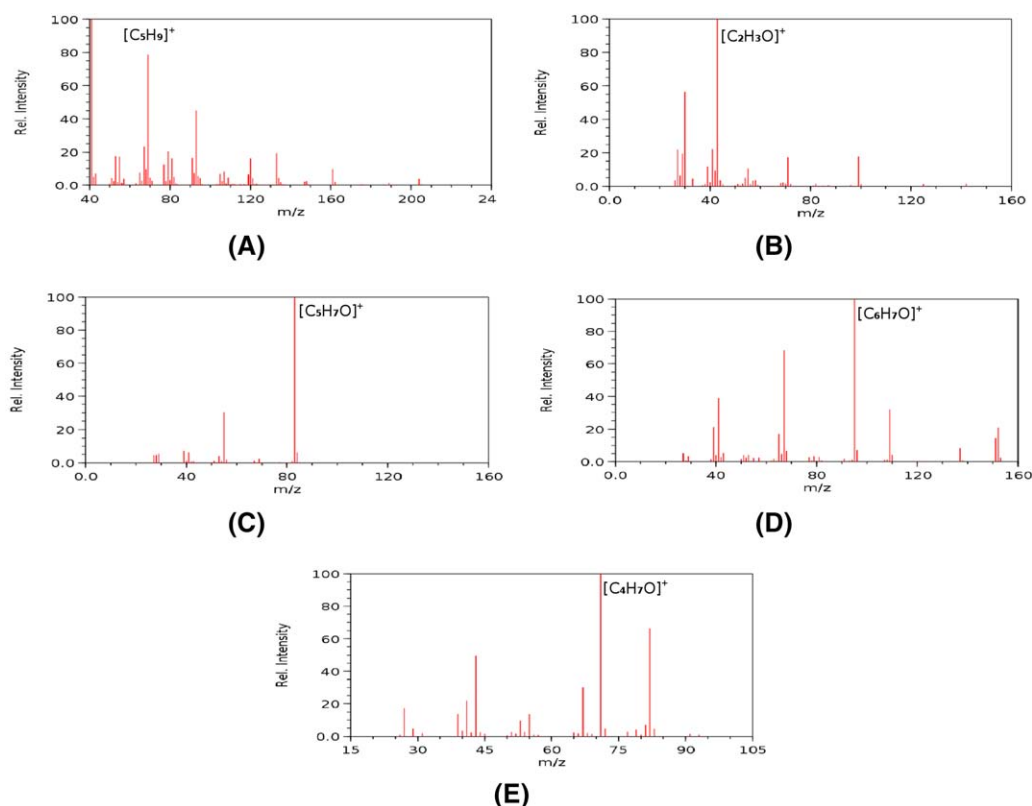
Figure 5 shows the number of correct predictions for the base peak for our four protocols (MG, M0, M+, and MC), evaluated on dataset A. Other protocols (MBIO, MBI+, and MCMG) are discussed in the Supporting Information.



**(A) Protocol MG****(B) Protocol M0****(C) Protocol M+****(D) Protocol MC**

**FIGURE 5** The percentage of correct predictions for the base peak of EI-MS in dataset A, using (A) protocol MG, (B) protocol M0, (C) protocol M+, and (D) protocol MC. Number of compounds are given in brackets

It is quite remarkable how successful protocol MG is in predicting the base peak, as shown in Figure 5A. The lowest success rate is for methyldecanes with 50% correctly predicted, albeit out of a total number of four molecules. The base peak was predicted correctly for 80% or more of the molecules. In the cases, where the cation dissociated during optimization, we assumed that one of the two resulting



**FIGURE 6** Electron impact mass spectra: (A) (*E*)- $\beta$ -farnesene (**71**), (B) 2,3-octanedione (**72**), (C) artemisia-ketone (**73**), (D) *cis*-tagetone (**74**), and (E) hotrienol (**75**)

fragments corresponded to the base peak in the experimental spectrum; this occurred for all the cations from alcohols except 3-pentanol **7** and 3-methyl-1-butanol **9**.

If we further analyze whether the success of the method comes from the weakening of the neutral or cation bond indices, we find that it is neither on its own, as shown in Figure 5B,C. For example, for the eight phenols, it is *only* the difference that correctly predicts the base peak. Therefore, it is not very surprising that while the consensus protocol MC, Figure 5D, produces better results than both M0 and M+, it is not better than protocol MG.

We emphasize that our protocols are not capable of dealing with rearrangement processes since they only apply to simple bond breaking, not bond breaking coupled to bond formation. This can be seen in Table 1 with four molecules where the predicted base peak differs by one or two units, presumably due to hydrogen rearrangement.

We also emphasize that our protocols do not predict which of the two fragments is charged. However, if needed, standard thermochemical methods in quantum chemistry are certainly able to solve this problem by calculating the relative energies of the charged fragments.

### 3.4 | Predicted base peak for selected natural products

To demonstrate the efficacy of protocol MG, we have selected five diverse natural products with the molecular formulae  $C_{15}H_{24}$  (**71**),  $C_8H_{14}O_2$  (**72**), and  $C_{10}H_{16}O$  (three isomers **73–75**) shown in Figure 1. (*E*)- $\beta$ -Farnesene (**71**) is an alarm pheromone for aphids.<sup>[43]</sup> 2,3-Octanedione (**72**) has been shown to be an important indicator of botanical diversity of use in assessing the quality of feed for ruminants.<sup>[44]</sup> Artemisia ketone (**73**) is the major constituent of essential oil of many Artemisia plants.<sup>[45]</sup> Similarly, *cis*-tagetone (**74**)<sup>[46]</sup> and hotrienol (**75**)<sup>[47]</sup> are known as essential oils in several plants. Their experimental spectra are shown in Figure 6.

The last bar in Figure 5A shows the percentage of base peaks predicted correctly for these selected natural products using protocol MG. We obtained three correct results out of five compounds. For (*E*)- $\beta$ -farnesene (**71**), the second largest peak ( $m/z = 69$ ) was correctly predicted. For *cis*-tagetone (**74**), the base peak was correctly predicted by the second largest value in the set  $\Delta_{CC}$ , Equation 14.

## 4 | CONCLUSION

In this study, we have demonstrated that *ab initio* Roby–Gould bond indices obtained from the optimized molecular geometries of the neutral and cationic molecular species can be used to reliably predict, using our protocol MG, the mass-to-charge ratio of the base peak in the EI-MS for a wide

range of systems where breaking of C—C bonds is responsible for the fragmentation processes. As far as we are aware, this is one of the few times that a large number of bond indices have been benchmarked statistically so as to demonstrate reliable predictions of a *bona fide* physicochemical property, the EI-MS base peak.

In contrast to neutral species, the Roby–Gould bond indices for C—C bonds in cationic species are much more unpredictable. Since we have demonstrated that predicting the main fragment of the mass spectrum via protocol MG required properties of the cation species, it follows that methods which attempt to predict fragmentation patterns based solely on the chemical structure of the neutral species are in essence trying to predict the quantum mechanics of cationic species—a very difficult task indeed for simple fitting or artificial intelligence methods.

Clearly, one advantage of using our protocol for predicting base peak in EI-MS is that it is not computationally demanding or time-consuming, unlike thermodynamic methods which may involve an unfeasibly large number of computations on molecular fragments as the molecule size increases.

Finally, we allow ourselves a brief speculation. Although the use of the base peak is bound to be less useful for larger molecular species, our work here on small compounds is still relevant to the analysis of (MS)<sup>n</sup> experiments, where the fragments of a mass spectrum are themselves subject to further electron impact fragmentation. After only a few of these steps, relatively small fragments are obtained. The use of the base peak protocol MG, or some other quantum mechanically based protocol, coupled with analysis of the spectra of such fragments offers the possibility to obtain the correct molecular structure *via* an *aufbau* process (even with a success rate of 88%, as found in this work) because one may expect that the likelihood of obtaining successively correct base peaks for fragments of a putative molecule which is not the actual molecule will become more and more unlikely, relative to the actual molecule, the more fragments that are analyzed. This is simply the product rule for probabilities, used to great effect in, for example, DNA profiling.

## ACKNOWLEDGMENTS

KA is grateful to the higher committee for education development in Iraq (HCED) for the award of a PhD scholarship. BB acknowledges funding from the Australian Research Council for grant DE160101313.

## ORCID

Khidhir Alhameedi  <http://orcid.org/0000-0003-3155-2716>

Amir Karton  <http://orcid.org/0000-0002-7981-508X>

## REFERENCES

- [1] E. Breitmaier, *Structure Elucidation by NMR in Organic Chemistry*, Wiley, Hoboken, New Jersey **2002**.
- [2] C. Giacovazzo, *Fundamentals of Crystallography*, Vol. 7, Oxford University Press, USA **2002**.
- [3] N. Krone, B. A. Hughes, G. G. Lavery, P. M. Stewart, W. Arlt, C. H. L. Shackleton, *J. Steroid Biochem. Mol. Biol.* **2010**, 121, 496.
- [4] S. E. Stein, *J. Am. Soc. Mass Spectrom.* **1995**, 6, 644.
- [5] B. Bohman, R. D. Phillips, M. H. M. Menz, B. W. Berndtsson, G. R. Flematti, R. A. Barrow, K. W. Dixon, R. Peakall, *New Phytol.* **2014**, 203, 939.
- [6] B. Bohman, G. R. Flematti, R. A. Barrow, E. Pichersky, R. Peakall, *Curr. Opin. Plant Biol.* **2016**, 32, 37.
- [7] R. T. Cardé, J. G. Millar, *Advances in Insect Chemical Ecology*, Cambridge University Press, Cambridge, UK **2004**.
- [8] G. V. P. Reddy, A. Guerrero, *Trends Plant Sci.* **2004**, 9, 253.
- [9] Z. R. Khan, D. G. James, C. A. O. Midega, J. A. Pickett, *Biol. Control* **2008**, 45, 210.
- [10] M. Jacobson, H.-C. Hansson, K. Noone, R. Charlson, *Rev. Geophys.* **2000**, 38, 267.
- [11] R. K. Lindsay, B. G. Buchanan, E. A. Feigenbaum, J. Lederberg, New York **1980**.
- [12] R. C. Beavis, S. M. Colby, R. Goodacre, P. d. B. Harrington, J. P. Reilly, S. Sokolow, C. W. Wilkerson, *Encyclopedia of Analytical Chemistry* Wiley Online Library **2000**.
- [13] K. Cross, P. Palmer, C. Beckner, A. Giordani, H. Gregg, P. Hoffman, C. Enke, ACS Publications **1986**.
- [14] K. Varnuza, W. Werther, *J. Chem. Inf. Model.* **1996**, 86, 323.
- [15] F. Allen, R. Greiner, D. Wishart, *Metabolomics* **2015**, 11, 98.
- [16] F. Allen, A. Pon, R. Greiner, D. Wishart, *Anal. Chem.* **2016**, 88, 7689.
- [17] N. F. Aguirre, S. Díaz-Tendero, P.-A. Hervieux, M. Alcamí, F. Martín, *J. Chem. Theory Comput.* **2017**, 13, 992.
- [18] S. Grimme, *Angew. Chem. Int. Ed.* **2013**, 52, 6306.
- [19] K. Scheubert, F. Hufsky, S. Böcker, *J. Cheminf.* **2013**, 5, 12.
- [20] P. R. Spackman, B. Bohman, A. Karton, D. Jayatilaka, *Int. J. Quantum Chem.* Wiley Online Library **2017**, 18, 2.
- [21] I. Mayer, A. Gomory, *Chem. Phys. Lett.* **2001**, 344, 553.
- [22] H. George, *J. Am. Chem. Soc.* **1955**, 77, 334.
- [23] M. D. Gould, C. Taylor, S. K. Wolff, G. S. Chandler, D. Jayatilaka, *Theor. Chem. Acc.* **2008**, 119, 275.

- [24] F. W. McLafferty, *Anal. Chem.* **1959**, 31, 82.
- [25] R. F. Bader, *Atoms in Molecules*; Wiley, *Acc Chem Res* **1990**, 9.
- [26] R. G. Parr, W. Yang, *Density-Functional Theory of Atoms and Molecules*, Oxford University Press, Oxford **1989**.
- [27] H.-P. Breuer, F. Petruccione, *The Theory of Open Quantum Systems*, Oxford University Press on Demand, Oxford **2002**.
- [28] D. Jayatilaka, D. Grimwood, *TONTO: A Fortran Based Object-Oriented System for Quantum Chemistry and Crystallography*, The University of Western Australia, Perth, Australia **2003**.
- [29] E. Davidson, *Reduced Density Matrices in Quantum Chemistry*, Academic Press, New York **1976**.
- [30] C. W. Bauschlicher, P. R. Taylor, *Theor. Chem. Acc.* **1988**, 74, 63.
- [31] D. Jayatilaka, S. C. Graham, *Mol. Phys.* **1997**, 92, 471.
- [32] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, D. J. Fox, *Gaussian09 Revision E.01*, Gaussian Inc., Wallingford, CT **2009**.
- [33] A. K. Rappé, C. J. Casewit, K. Colwell, W. Goddard III, W. Skiff, *J. Am. Chem. Soc.* **1992**, 114, 10024.
- [34] M. D. Hanwell, D. E. Curtis, D. C. Lonie, T. Vandermeersch, E. Zurek, G. R. Hutchison, *J. Cheminf.* **2012**, 4, 17.
- [35] NIST Mass Spec Data Center, S. E. Stein, in *NIST Chemistry Webbook, NIST Standard Reference Database Number 69* (Eds: P. J. Linstrom, W. G. Mallard), National Institute of Standards and Technology, Gaithersburg, MD **2015**.
- [36] S. Grabowsky, P. Luger, J. Buschmann, T. Schneider, T. Schirmeister, A. N. Sobolev, D. Jayatilaka, *Angew. Chem. Int. Ed.* **2012**, 51, 6776.
- [37] S. P. Thomas, D. Jayatilaka, T. N. Guru Row, *Phys. Chem. Chem. Phys.* **2015**, 17, 25411.
- [38] S. Franke, F. Ibarra, C. M. Schulz, R. Twele, J. Poldy, R. A. Barrow, R. Peakall, F. P. Schiestl, W. Francke, *Proc. Natl. Acad. Sci. USA* **2009**, 106, 8877.
- [39] R. Peakall, D. Ebert, J. Poldy, R. A. Barrow, W. Francke, C. C. Bower, F. P. Schiestl, *New Phytol.* **2010**, 188, 437.
- [40] B. Bohman, L. Jeffares, G. Flematti, L. T. Byrne, B. W. Skelton, R. D. Phillips, K. W. Dixon, R. Peakall, R. A. Barrow, *J. Nat. Prod.* **2012**, 75, 1589.
- [41] B. Bohman, B. Berntsson, R. C. M. Dixon, C. D. Stewart, R. A. Barrow, *Org. Lett.* **2014**, 16, 2787.
- [42] J. M. Martin, J. El-Yazal, J.-P. François, *Mol. Phys.* **1995**, 86, 1437.
- [43] L. L. Cui, F. Francis, S. Heuskin, G. Lognay, Y. J. Liu, J. Dong, J. L. Chen, X. M. Song, Y. Liu, *Biol. Control* **2012**, 60, 108.
- [44] E. Serrano, A. Cornu, N. Kondjoyan, J. Agabriel, D. Micol, *Animal* **2011**, 5, 641.
- [45] P. Goswamia, A. Chauhan, R. S. Verma, R. C. Padalia, S. K. Verma, M. P. Darokar, C. S. Chanotiya, *J. Essent. Oil Res.* **2016**, 28, 71.
- [46] G. Singh, O. P. Singh, M. De Lampasona, C. A. Catalan, *Flavour Fragr. J.* **2003**, 18, 62.
- [47] N. Radulović, M. Denić, Z. Stojanović-Radić, D. Skropeta, *J. Am. Oil Chem. Soc.* **2012**, 89, 2165.

## SUPPORTING INFORMATION

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

**How to cite this article:** Alhameedi K, Bohman B, Kartan A, Jayatilaka D. Predicting the primary fragments in mass spectrometry using *ab initio* Roby–Gould bond indices. *Int J Quantum Chem.* 2018;e25603. <https://doi.org/10.1002/qua.25603>

## APPENDIX : ROBY–GOULD BOND INDICES

In this appendix, we describe and present formula for the Roby–Gould bond indices in a finite molecular space-spin basis set. The original paper<sup>[23]</sup> did not give such explicit expressions and it is hoped that this presentation will be less abstract for those interested in reproducing our results and checking the required algebraic manipulations. Should an actual implementation of these equations be needed, this is available in the open source program Tonto<sup>[28]</sup> available on github at [github.com/tonto-chem](https://github.com/tonto-chem).

1. First, the density operator for the molecule is defined by

$$\rho = \sum_{\alpha=1}^n \sum_{\beta=1}^n |\chi_{\alpha}\rangle D_{\alpha\beta} \langle \chi_{\beta}| \quad (\text{A1})$$

The molecular density matrix  $D$  in the equation above is obtained from the quantum chemical method for the desired molecular state (charge and multiplicity).  $n$  is the number of basis spin-orbitals, normally twice the number of spatial basis functions.

2. Next, one must obtain the coefficients  $A_{\alpha i}$  for the spherically averaged atomic natural spinorbitals (ANOs),

$$|A_i\rangle = \sum_{\alpha=1}^{n_A} A_{\alpha i} |\chi_{\alpha}^A\rangle, \quad i=1, \dots, N_A^{\text{occ}}. \quad (\text{A2})$$

Here,  $\{|\chi_{\alpha}^A\rangle\}_{\alpha=1}^{n_A}$  is the set of basis spin-orbitals on atom  $A$  of which there are  $n_A$ , and  $N_A^{\text{occ}}$  occupied ANOs on this atom (it is important to realise  $N_A^{\text{occ}}$  is not necessarily equal to the number of electrons on the atom  $A$ ). The ANOs are obtained by finding the eigenstates of the spherically average density operator  $\rho^A$  in the usual way,<sup>[29]</sup> with  $\rho^A$  being defined like  $\rho$  above but being the density operator for the spherically averaged isolated atom calculated using the unrestricted version of the quantum mechanical method used in the first step. Spherical averaging is performed as described in Refs. [30,31] using the octahedral group  $O_h$ . A particular spherically average ANO is deemed to be occupied if its occupation number (the corresponding eigenvalue of  $\rho^A$ ) is greater than 0.05. The spherically averaged ANOs are orthonormal, that is,  $\langle A_i | A_j \rangle = \delta_{ij}$  where  $\delta_{ij}$  is the Kronecker delta.

3. The the projector  $P_A$  for the atom  $A$  in the space  $V_A = \text{span}\{|\chi_{\alpha}^A\rangle\}_{\alpha=1}^{n_A}$  is defined by

$$P_A = \sum_{i=1}^{N_A^{\text{occ}}} |A_i\rangle \langle A_i| \quad (\text{A3})$$

$$= \sum_{\alpha=1}^{n_A} \sum_{\beta=1}^{n_A} |\chi_{\alpha}^A\rangle D_{\alpha\beta}^A \langle \chi_{\beta}^A|, \quad \text{where} \quad (\text{A4})$$

$$D_{\alpha\beta}^A = \sum_{i=1}^{N_A^{\text{occ}}} A_{\alpha i} A_{\beta i}. \quad (\text{A5})$$

4. Gould's operator  $I_{AB} = P_A - P_B$  for atoms  $A$  and  $B$  in the space

$$V_{AB} = V_A \oplus V_B = \text{span}\{|\chi_{\alpha}^{AB}\rangle\}_{\alpha=1}^{n_{AB}=n_A+n_B},$$

where  $\chi_{\alpha}^{AB}$  is the list of concatenated atomic spin-basis functions on atoms  $A$  and  $B$ ,

$$|\chi_{\alpha}^{AB}\rangle = \begin{cases} |\chi_{\alpha}^A\rangle & \text{if } \alpha \leq n_A \\ |\chi_{\alpha-n_A}^B\rangle & \text{if } n_A < \alpha \leq n_{AB} \end{cases} \quad (\text{A6})$$

is defined by

$$I_{AB} = \sum_{\alpha=1}^{n_{AB}} \sum_{\beta=1}^{n_{AB}} |\chi_{\alpha}^{AB}\rangle I_{\alpha\beta}^{AB} \langle \chi_{\beta}^{AB}|, \quad \text{where} \quad (\text{A7})$$

$$I^{AB} = \begin{pmatrix} D^A & 0 \\ 0 & -D^B \end{pmatrix}. \quad (\text{A8})$$

Likewise, Roby's shared operator is given by

$$R_{AB} = \sum_{\alpha=1}^{n_{AB}} \sum_{\beta=1}^{n_{AB}} |\chi_{\alpha}^{AB}\rangle R_{\alpha\beta}^{AB} \langle \chi_{\beta}^{AB}|, \quad \text{where} \quad (\text{A9})$$

$$R^{AB} = \begin{pmatrix} D^A & 0 \\ 0 & D^B \end{pmatrix} - (S^{AB})^{-1}. \quad (\text{A10})$$

5. According to the original paper<sup>[23]</sup> an ionic eigenstate characterised by Araki angle  $\theta$ ,

$$|\sin \theta\rangle = \sum_{\alpha=1}^{n_{AB}} I_{\alpha} |\chi_{\alpha}^{AB}\rangle, \quad (\text{A11})$$

is an eigenstates of Gould's ionic operator  $I_{AB}$  that is,

$$I_{AB}|\sin\theta\rangle = \sin\theta|\sin\theta\rangle. \quad (\text{A12})$$

This leads to the following matrix eigenvalue equations

$$I^{AB}S^{AB}I = \sin\theta I, \quad (\text{A13})$$

where  $S^{AB}$  is the overlap matrix between the basis spin-orbitals in  $V_{AB}$ . The eigenstates corresponding to a zero eigenvalue (i.e., a zero Araki angle,  $\theta_i=0$ ) correspond to linear dependencies and are a mathematical artefact, therefore, not of interest in bond analysis. Likewise, angles corresponding to a unit eigenvalue (i.e., an Araki angle of  $\theta_i=\pi/2$  correspond to nonbonding nonoverlapping orbitals on the two atoms) and are likewise removed from the following analysis. The remaining nonzero eigenvalues were shown to come in pairs of opposite sign, representing bonding and antibonding pairs<sup>[23]</sup> and the degenerate paired subspaces are labeled by their angle  $V_{|\theta|}$ .

6. To construct the eigenstates of the Roby operator  $R_{AB}$  we need to ensure that any degeneracies are handled properly. To this end, we explicitly construct the negative ionic eigenstates  $|\sin\theta\rangle$  from its paired positive eigenstates counterpart  $|+\sin\theta\rangle$  by projecting that state with  $P_A$  and  $P_B$  to, respectively, obtained the two linearly independent components in the space  $V_A$  and  $V_B$  (eqs. 56, 57, 63, and 64) in the previous paper<sup>[23]</sup> and the construction of the negative eigenvalue state from these two components (eq. 64) in the previous paper). This leads to the following expression for the negative eigenvalue coefficients  $I_\alpha^-$  in terms of the corresponding positive eigenvalue coefficients  $I_\alpha^+$ :

$$|\sin\theta\rangle = \sum_{\alpha=1}^{n_{AB}} I_\alpha^- |\chi_\alpha^{AB}\rangle, \quad (\text{A14})$$

$$I^- = P^- S^{AB} I^+, \quad (\text{A15})$$

$$P^- = \begin{pmatrix} g^+ D^A & 0 \\ 0 & g^- D^B \end{pmatrix} \quad (\text{A16})$$

$$g^\pm = [(f^- \pm f^+) + \cos\theta(f^- \mp f^+)](f^- \mp f^+), \quad (\text{A17})$$

$$f^\pm = \frac{\sqrt{1 \pm \cos\theta}}{\sin\theta}, \quad (\text{A18})$$

and where  $I^+$  are, by analogy with Equation A14, the expansion coefficients for the eigenstate  $|+\sin\theta\rangle$ .

7. The covalent eigenstates are equal to a  $\pi/2$  rotation of the ionic eigenstates,

$$R^+ = \frac{1}{2} [I^+ + I^-], \text{ and} \quad (\text{A19})$$

$$R^- = \frac{1}{2} [I^+ - I^-]. \quad (\text{A20})$$

8. The covalent and ionic populations of the paired states are given by, respectively,

$$n_c^\pm = (I^\pm)^T S^{AB} R^{AB} S^{AB} I^\pm, \quad (\text{A21})$$

$$n_i^\pm = (R^\pm)^T S^{AB} I^{AB} S^{AB} R^\pm, \quad (\text{A22})$$

and the covalent and ionic Roby–Gould bond indices in each subspace  $V_\theta$  are, respectively,

$$c_\theta = (n_c^+ - n_c^-)/2, \text{ and} \quad (\text{A23})$$

$$i_\theta = (n_i^+ - n_i^-)/2. \quad (\text{A24})$$

Finally, the two components of the Roby–Gould bond indices are

$$c_{AB} = \sum_{0 < \theta < \pi/2} c_\theta, \text{ and} \quad (\text{A25})$$

$$i_{AB} = \sum_{0 < \theta < \pi/2} i_\theta. \quad (\text{A26})$$

In this work, an angle is regarded equal to zero or  $\pi/2$  if its difference is less than  $(0.01^\circ)\pi/180$ . In previous work, we used the value  $(77^\circ)\pi/180$  but we have found that, for analyses between atoms (as opposed to groups of atoms) any numerically small value is adequate. This essentially removes one of the *ad hoc* constants that had been introduced in the previous paper (the only remaining arbitrary constant is the value of 0.05 used to decide an occupied ANO, in step 2).