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

Khidhir Alhameedi,*[ab] Björn Bohman,[a] Amir Karton,[a] and Dylan Jayatilaka*[a]

[a] School of Molecular Sciences, University of Western Australia, 35 Stirling Highway Nedlands 6009, Australia.

[b] Department of Chemistry, College of Education for Pure Science, University of Karbala, Karbala, Iraq.
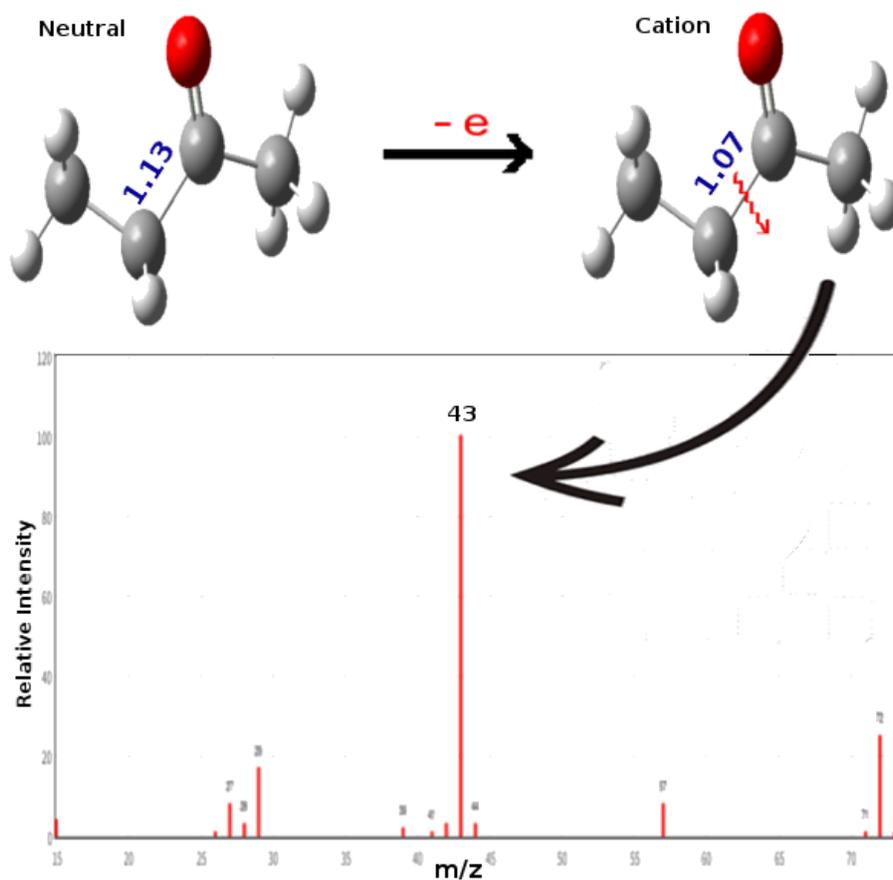
E-mail:
khidhir.abdalhussein@gmail.com
dylan.jayatilaka@uwa.edu.au,

Graphical Abstract Text and Image

The highest intensity peak (base peak) in electron impact mass spectrometry (EI-MS) is here predicted correctly for 65 out of 75 molecules using the Roby-Gould bond indices applied to ground state and cation *ab initio* wavefunctions.

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 (EI-MS). 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.

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[1] and X-ray crystallography[2]. However, these methods are not suitable when the targeted compound is present in low amounts or cannot be crystallized. Instead, mass spectrometry methods, *e.g.* electron impact mass spectrometry (EI-MS), coupled to chromatographic separation methods, *e.g.* gas chromatography (GC-MS), are routinely used for such applications[3].

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[4]. In order 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.

a) (*E*)-*β*-farnesene   **b)** 2,3-octanedione

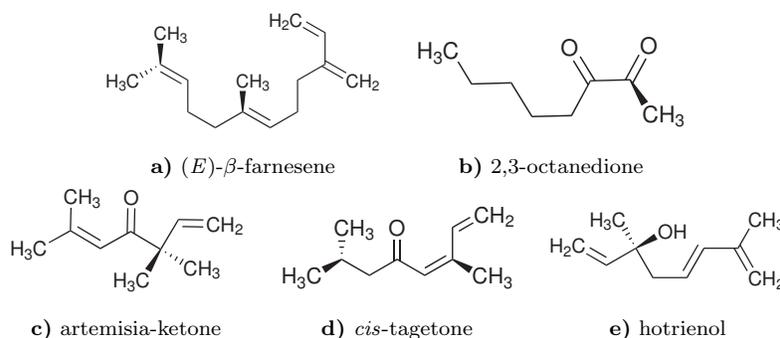**c)** artemisia-ketone   **d)** *cis*-tagetone   **e)** hotrienol

Figure 1: The molecular structure of specifically selected natural products. Note that compounds **c**, **d** and **e** are isomers with identical formula $C_{10}H_{16}O$.

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[5–9]. 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[10]. 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 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[11]. Since then a variety of methods that adopt artificial intelligence and machine-learning approaches to elucidate substructures from mass spectra have been reported[12–14]. One particular example is competitive fragment modeling for metabolite identification (CFM-ID) for predicting EI-MS[15,16].

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[17]. Recently, Grimme[18] 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[19]. 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[20]. 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 non-dynamical *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 the first to use *ab initio* methods to predict bond orders and the primary fragment in the mass spectrum[21], 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[22]. 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 paper 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[23] which is known to be stable with basis set extension. Furthermore, we re-optimize the geometry after removing one electron *i.e.* 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[24]. 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 literature[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.

$$c_{AB} = \left\langle \frac{R_{AB}}{2 \mid R_{AB} \mid} \right\rangle, \text{ and} \tag{1}$$

$$i_{AB} = \left\langle \frac{I_{AB}}{2 \mid I_{AB} \mid} \right\rangle \tag{2}$$

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,

$$R_{AB} = P_A + P_B - P_{AB}, \tag{3}$$

$$I_{AB} = P_A - P_B. \tag{4}$$

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 $V_A \bigoplus V_B$.[1] 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

$$R_{AB}^2 + I_{AB}^2 = P_{AB}^2. \tag{5}$$

---

[1]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[25], this would result in no shared electrons; in the word's of Parr and Yang[26] 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[27]. 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.

The total Roby-Gould bond index τ is defined as following;

$$\tau_{AB} = \sqrt{c_{AB}^2 + i_{AB}^2}. \tag{6}$$

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[23]. In this study, bond indices were calculated using the free Tonto program package[28] 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[29]. 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[30] in the Avogadro program[31]. 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   Data set A

In order to benchmark the bond index method, we formulate data set 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 data set 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[32].

2.4 Data set B

There have only been a few publications concerning the Roby-Gould bond indices[23,33,34]. 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 base peak predictions, we extend data set A with data set B, which comprised another 28 molecules; these are listed in Table S1 in the supplementary material. The molecules in data set B include semiochemicals such as chiloglottones[35,36] and hydroxymethylpyrazines[37,38]. The combination of the two data sets comprise 103 molecules in total.

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.

MBI0. Minimum bond index in the neutral species of data set A over all pairs of atoms,

$$\tau_{\min}^{\text{neutral}} = \min \tau^{\text{neutral}}, \text{ where} \tag{7}$$

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

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

MBI+. Minimum bond index in the cations of data set A with all types of bonds,

$$\tau_{\min}^{\text{cation}} = \min \tau^{\text{cation}}, \text{ where} \tag{9}$$

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

M0. Weakest carbon-carbon bond in the neutral,

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

M+. Weakest carbon-carbon bond in the cation,

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

MG. The biggest change in bond index between two carbon atoms in the neutral and cation species[21],

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

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

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

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.

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

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).

MBI0

Weakest bond index in neutral species

MBI+

Weakest bond index in cations

M0

Weakest C-C bond index in neutral species

M+

Weakest C-C bond index in cations

MG

The biggest change in bond index between two carbon atoms in the neutral and cation species

MC

"Consensus"
The two or three of M0, M+ and MG protocols agree with the same C-C bond
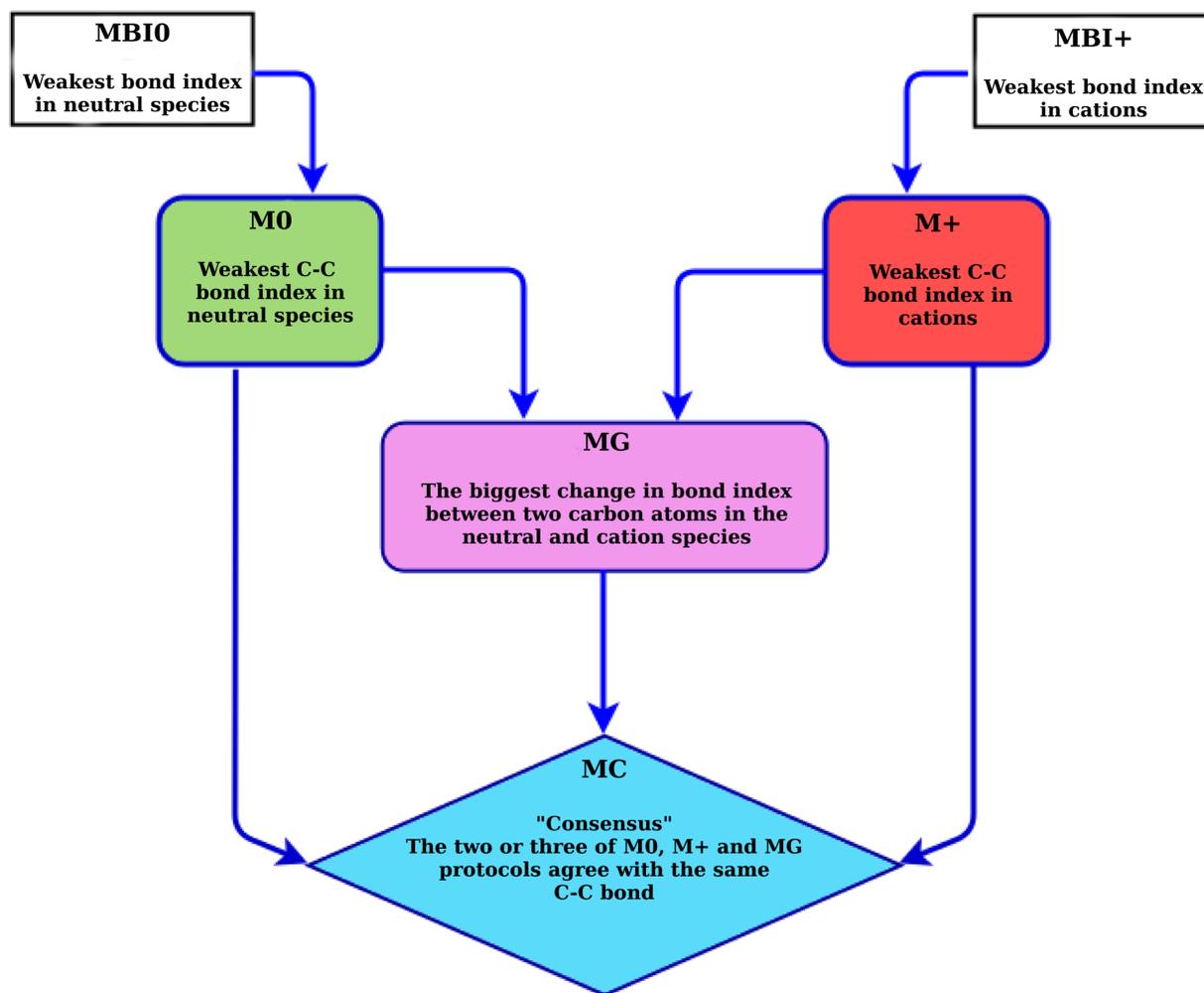
Scheme 1: The dependency of the six protocols used to predict the base peak in EI-MS. The colours used for each protocol correspond to those in Figure 6. The central panel in pink is the modified protocol of Mayer and Gomory (protocol MG).

Alcohols

Amines

Alkenes

Aldehydes & Ketones

Thiols & Sulfides

Methyldecanes

Phenols
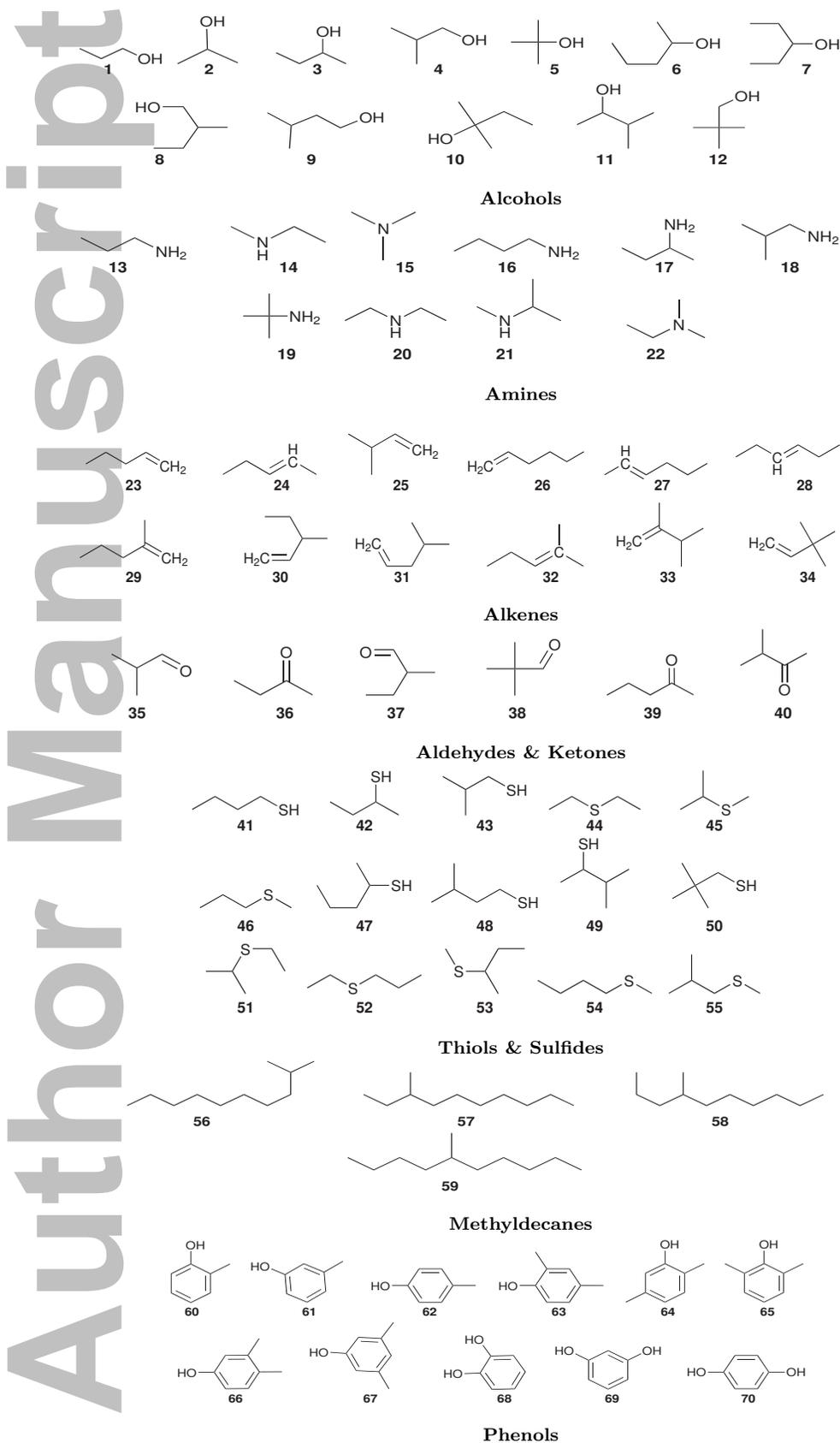
Figure 2: Structures of compounds **1**-**70** as part of data set A.

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 data sets 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.
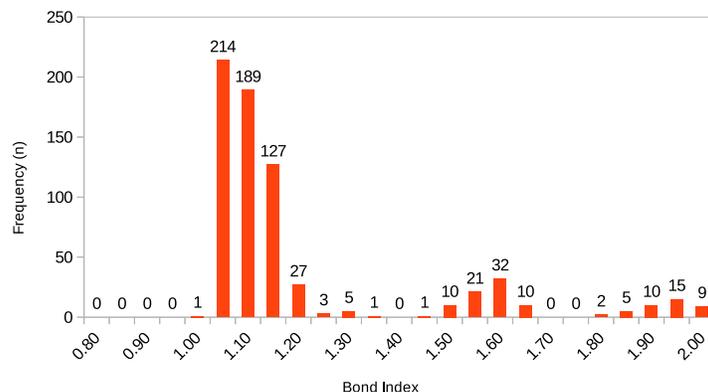
(a) Neutral species

(b) Cations

Figure 3: Distributions of C-C Roby-Gould bond index values for (a) neutral species and (b) cations for combined data sets A and B. Distributions for other bond types are given in SI.

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.

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.

14

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 0.01Å[39]; hence our results show that changes in the bond lengths of this order will hardly change the bond indices at this level of theory.
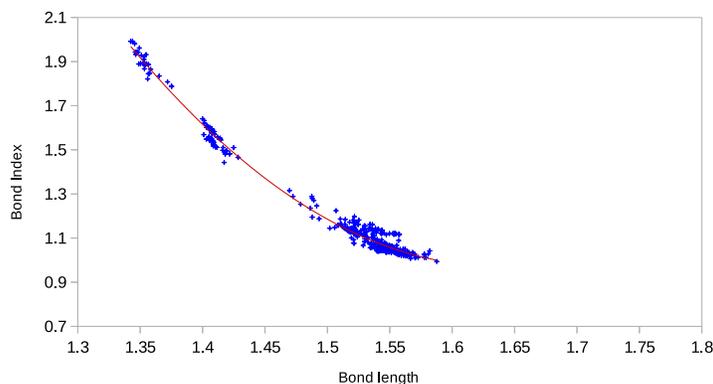
(a) Neutral species
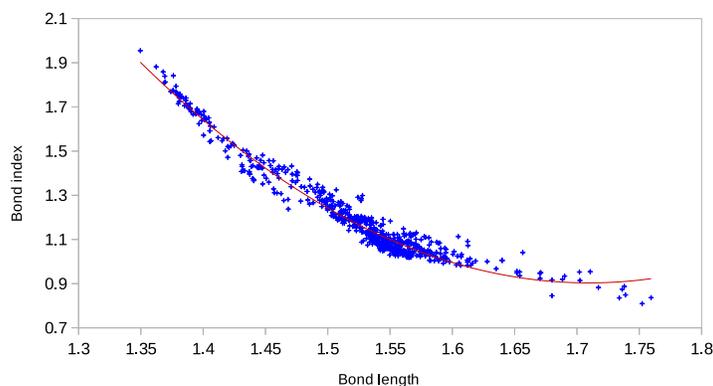
(b) Cations

Figure 4: Bond index versus bond length (in Ångstrom) for C-C bonds in data sets A and B for (a) the neutral species and (b) the cations. Correlations for other bond types are given in SI.

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 data set A are distinguishable from their base peak value. Indeed for the phenols with formula C_8H_9OH all five isomers in our data set have the same base peak, the molecular ion (which as we will see in Table 1, is correctly predicted by using protocol MG; the molecular ion is the base peak because the protocol predicts that the ring is opened). Thus even in this small data set 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 data set A. Other protocols (MBI0, MBI+, and MCMG) are discussed in the SI.

It is quite remarkable how successful protocol MG is in predicting the base peak, as shown in Figure 5(a). 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 fragments corresponded to the base peak in the experimental spectrum; this occured 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 Figures 5(b) and 5(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,

Table 1: Chemical formulae and the number of isomers for data set A. The number of isomers with the same mass-to-charge ratio for the base peak is shown as a list, specifically $(n_1, n_2, n_3, \ldots)$ 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").

Chemical Formula	No. of isomers	Isomers with same base peak	No. correct/ Near miss
Alcohols:	12		11/1
C_3H_7OH	2	(2)	
C_4H_9OH	3	(3)	
$C_5H_{11}OH$	7	(0,2,1)	
Amines:	10		10/0
C_3H_9N	3	(3)	
$C_4H_{11}N$	7	(2,1,1)	
Alkenes:	12		10/1
C_5H_{10}	3	(1,1)	
C_6H_{12}	9	(2,0,1,1)	
Aldehydes & Ketones:	6		6/0
C_4H_8O	2	(0,1)	
$C_5H_{10}O$	4	(2,1)	
Thiols & Sulfides:	15		12/2
$C_4H_{10}S$	6	(4,1)	
$C_5H_{12}S$	9	(4,1,1)	
Methyldecanes:	4		2/0
$C_{11}H_{24}$	4	(1,0,1)	
Phenols:	11		11/0
C_7H_7OH	3	(0,0,1)	
C_8H_9OH	5	(0,0,0,0,1)	
$C_6H_6O_2$	3	(0,0,1)	
Selected natural products:	5		4/0
$C_{15}H_{24}$	1	(1)	
$C_8H_{14}O_2$	1	(1)	
$C_{10}H_{16}O$	3	(3)	

Figure 5(d), 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

In order 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)-β-Farnesene (**71**) is an alarm pheromone for aphids[40]. 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[41]. Artemisia ketone (**73**) is the major constituent of essential oil of many Artemisia plants[42]. Similarily, cis-tagetone (**74**)[43] and hotrienol (**75**)[44] are known as essential oils in several plants. Their experimental spectra are shown in Figure 6.

(a) Protocol MG
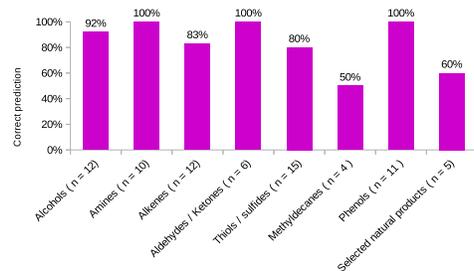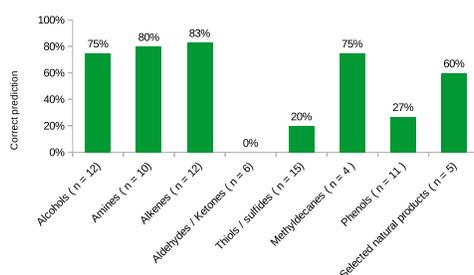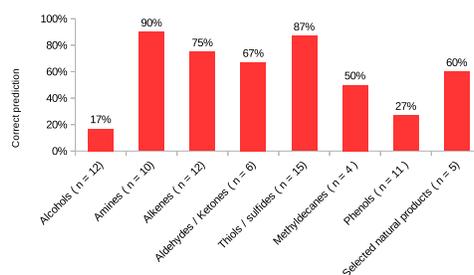
(b) Protocol M0
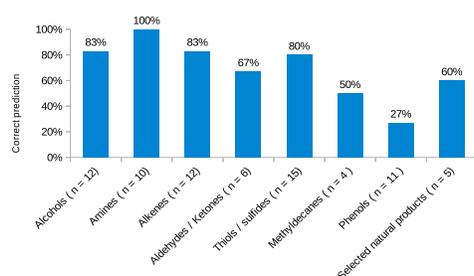
(c) Protocol M+

(d) Protocol MC

Figure 5: The percentage of correct predictions for the base peak of EI-MS in data set A, using (a) protocol MG, (b) protocol M0, (c) protocol M+, and (d) protocol MC. Number of compounds are given in brackets.
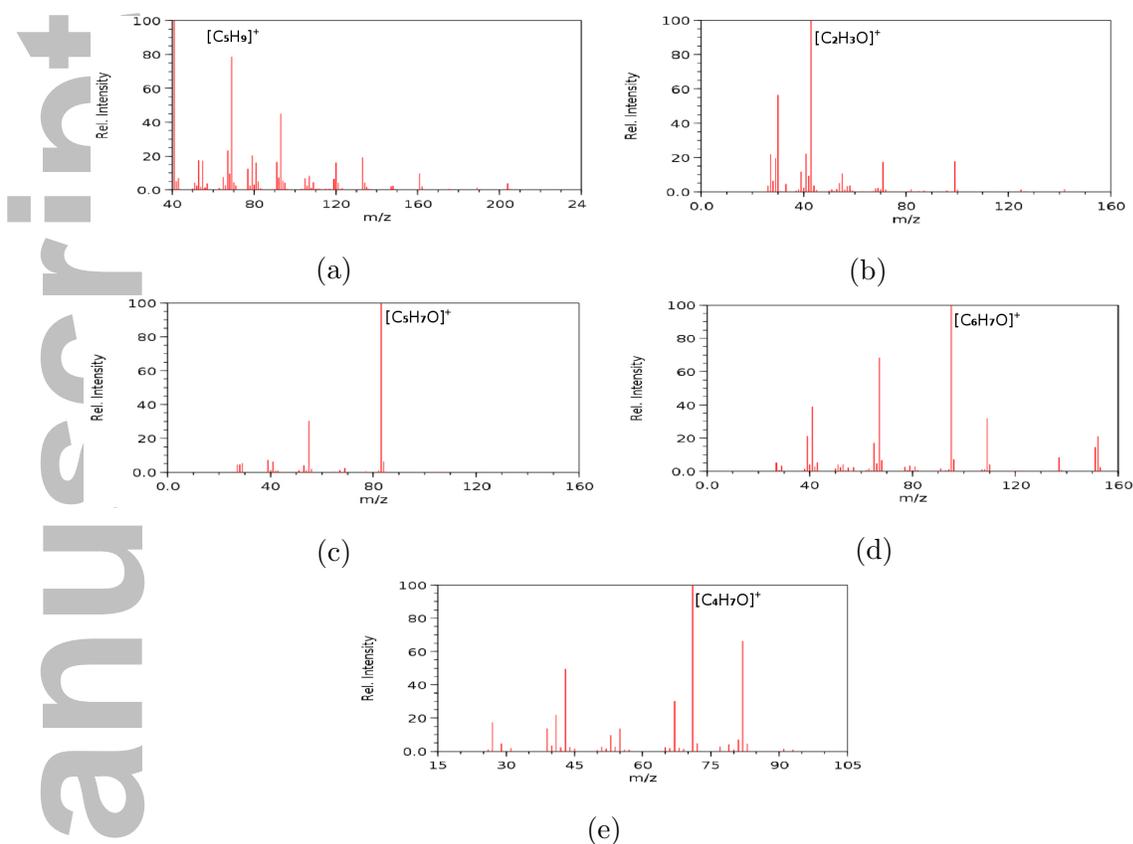
Figure 6: Electron impact mass spectra: (a) *(E)-β*-farnesene (**71**), (b) 2,3-octanedione (**72**), (c) artemisia-ketone (**73**), (d) *cis*-tagetone (**74**), and (e) hotrienol (**75**).

The last bar in Figure 5(a) 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)-β*-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 Δ_{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* physico-chemical 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 $(MS)^n$ 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.

Acknowledgements

A 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 [23] 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 [28] available on github at `github.com/tonto-chem`.

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

$$\rho = \sum_{\alpha=1}^{n} \sum_{\beta=1}^{n} |\chi_\alpha\rangle D_{\alpha\beta} \langle\chi_\beta| \tag{16}$$

   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),

$$|A_i\rangle = \sum_{\alpha=1}^{n_A} A_{\alpha i} |\chi_\alpha^A\rangle, \ i = 1, \ldots, N_A^{\text{occ}}. \tag{17}$$

   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^{occ} occupied ANOs on this atom (it is important to realise N_A^{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 ρ^A in the usual way [45], with ρ^A being defined like ρ 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 [46,47] using the octahedral group O_h. A particular spherically average ANO is deemed to be occupied if it's occupation number (the corresponding eigenvalue of ρ^A) is greater than 0.05. The spherically averaged ANOs are orthonormal i.e. $\langle A_i|A_j\rangle = \delta_{ij}$ where δ_{ij} is the Kronecker delta.

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

$$P_A = \sum_{i=1}^{N_A^{\text{occ}}} |A_i\rangle\langle A_i| \tag{18}$$

$$= \sum_{\alpha=1}^{n_A}\sum_{\beta=1}^{n_A} |\chi_\alpha^A\rangle D_{\alpha\beta}^A \langle\chi_\beta^A|, \text{ where} \tag{19}$$

$$D_{\alpha\beta}^A = \sum_{i=1}^{N_A^{\text{occ}}} A_{\alpha i} A_{\beta i}. \tag{20}$$

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

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

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

$$|\chi_\alpha^{AB}\rangle = \begin{cases} |\chi_\alpha^A\rangle & \text{if } \alpha \leq n_A \\ |\chi_{\alpha-n_A}^B\rangle & \text{if } n_A < \alpha \leq n_{AB} \end{cases} \tag{21}$$

is defined by

$$I_{AB} = \sum_{\alpha=1}^{n_{AB}}\sum_{\beta=1}^{n_{AB}} |\chi_\alpha^{AB}\rangle I_{\alpha\beta}^{AB} \langle\chi_\beta^{AB}|, \text{ where} \tag{22}$$

$$\boldsymbol{I}^{AB} = \begin{pmatrix} \boldsymbol{D}^A & \boldsymbol{0} \\ \boldsymbol{0} & -\boldsymbol{D}^B \end{pmatrix}. \tag{23}$$

Likewise, Roby's shared operator is given by

$$R_{AB} = \sum_{\alpha=1}^{n_{AB}}\sum_{\beta=1}^{n_{AB}} |\chi_\alpha^{AB}\rangle R_{\alpha\beta}^{AB} \langle\chi_\beta^{AB}|, \text{ where} \tag{24}$$

$$\boldsymbol{R}^{AB} = \begin{pmatrix} \boldsymbol{D}^A & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D}^B \end{pmatrix} - (\boldsymbol{S}^{AB})^{-1}. \tag{25}$$

5. According to the original paper[23] an ionic eigenstate characterised by Araki angle θ,

$$|\sin\theta\rangle = \sum_{\alpha=1}^{n_{AB}} I_\alpha |\chi_\alpha^{AB}\rangle, \tag{26}$$

is an eigenstates of Gould's ionic operator I_{AB} i.e.

$$I_{AB}|\sin\theta\rangle = \sin\theta|\sin\theta\rangle. \tag{27}$$

This leads to the following matrix eigenvalue equations

$$\boldsymbol{I}^{AB}\boldsymbol{S}^{AB}\boldsymbol{I} = \sin\theta_i\boldsymbol{I}, \tag{28}$$

where $\boldsymbol{S}^{AB}$ 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 non-bonding non-overlapping orbitals on the two atoms) and are likewise removed from the following analysis. The remaining non-zero eigenvalues were shown to come in pairs of opposite sign, representing bonding and antibonding pairs[23] and the degenerate paired subspaces are labeled by their angle $V_{|\theta|}$.

6. In order 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 it's 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 (equations (56), (57), (63) and (64) in the previous paper[23]) and the construction of the negative eigenvalue state from these two components (equation (64) in the previous paper). This leads to the following expression for the negative eigenvalue coefficients I_α^- in terms of the corresponding positive eigenvalue coefficients I_α^+:

$$|-\sin\theta\rangle = \sum_{\alpha=1}^{n_{AB}} I_\alpha^- |\chi_\alpha^{AB}\rangle, \tag{29}$$

$$\boldsymbol{I}^- = \boldsymbol{P}^-\boldsymbol{S}^{AB}\boldsymbol{I}^+, \tag{30}$$

$$\boldsymbol{P}^- = \begin{pmatrix} g^+\boldsymbol{D}^A & \boldsymbol{0} \\ \boldsymbol{0} & g^-\boldsymbol{D}^B \end{pmatrix} \tag{31}$$

$$g^\pm = \left[(f^- \pm f^+) + \cos\theta(f^- \mp f^+)\right](f^- \mp f^+), \tag{32}$$

$$f^\pm = \frac{\sqrt{1 \pm \cos\theta}}{\sin\theta}, \tag{33}$$

and where $\boldsymbol{I}^+$ are, by analogy with equation (29), the expansion coefficients for the eigenstate $|+\sin\theta\rangle$.

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

$$\boldsymbol{R}^+ = \frac{1}{2}\left[\boldsymbol{I}^+ + \boldsymbol{I}^-\right], \text{ and} \tag{34}$$

$$\boldsymbol{R}^- = \frac{1}{2}\left[\boldsymbol{I}^+ - \boldsymbol{I}^-\right]. \tag{35}$$

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

$$n_c^\pm = (\boldsymbol{I}^\pm)^T \boldsymbol{S}^{AB} \boldsymbol{R}^{AB} \boldsymbol{S}^{AB} \boldsymbol{I}^\pm, \tag{36}$$

$$n_i^\pm = (\boldsymbol{R}^\pm)^T \boldsymbol{S}^{AB} \boldsymbol{I}^{AB} \boldsymbol{S}^{AB} \boldsymbol{R}^\pm, \tag{37}$$

and the covalent and ionic Roby-Gould bond indices in each subspace V_θ are, respectively,

$$c_\theta = (n_c^+ - n_c^-)/2, \text{ and} \tag{38}$$

$$i_\theta = (n_i^+ - n_i^-)/2. \tag{39}$$

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

$$c_{AB} = \sum_{0<\theta<\pi/2} c_\theta, \text{ and} \tag{40}$$

$$i_{AB} = \sum_{0<\theta<\pi/2} i_\theta. \tag{41}$$

In this work, an angle is regarded equal to zero or $\pi/2$ if it's difference is less than $(0.01°)\pi/180$. In previous work we used the value $(77°)\pi/180$ 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).

References

[1] Breitmaier, E. Structure Elucidation by NMR in Organic Chemistry; John Wiley & Sons, Ltd, 2002.

[2] Giacovazzo, C. Fundamentals of crystallography; vol. 7 Oxford university press, USA, 2002.

[3] Krone, N.; Hughes, B. A.; Lavery, G. G.; Stewart, P. M.; Arlt, W.; Shackleton, C. H. L. The Journal of Steroid Biochemistry and Molecular Biology 2010, 121, 496.

[4] Stein, S. E. Journal of the American Society for Mass Spectrometry 1995, 6, 644.

[5] Bohman, B.; Phillips, R. D.; Menz, M. H. M.; Berntsson, B. W.; Flematti, G. R.; Barrow, R. A.; Dixon, K. W.; Peakall, R. New Phytologist 2014, 203, 939.

[6] Bohman, B.; Flematti, G. R.; Barrow, R. A.; Pichersky, E.; Peakall, R. Current Opinion in Plant Biology 2016, 32, 37.

[7] Cardé, R. T.; Millar, J. G. Advances in insect chemical ecology; Cambridge University Press, 2004.

[8] Reddy, G. V. P.; Guerrero, A. Trends in Plant Science 2004, 9, 253.

[9] Khan, Z. R.; James, D. G.; Midega, C. A. O.; Pickett, J. A. Biological Control 2008, 45, 210.

[10] Jacobson, M.; Hansson, H.-C.; Noone, K.; Charlson, R. Reviews of Geophysics 2000, 38, 267.

[11] Lindsay, R. K.; Buchanan, B. G.; Feigenbaum, E. A.; Lederberg, J. New York 1980.

[12] Beavis, R. C.; Colby, S. M.; Goodacre, R.; Harrington, P. d. B.; Reilly, J. P.; Sokolow, S.; Wilkerson, C. W. Encyclopedia of analytical chemistry 2000.

[13] Cross, K.; Palmer, P.; Beckner, C.; Giordani, A.; Gregg, H.; Hoffman, P.; Enke, C. ACS Publications, 1986.

John Wiley & Sons, Inc.

[14] Varmuza, K.; Werther, W. Journal of Chemical Information and Modeling 1996, 86, 323.

[15] Allen, F.; Greiner, R.; Wishart, D. Metabolomics 2015, 11, 98.

[16] Allen, F.; Pon, A.; Greiner, R.; Wishart, D. Analytical chemistry 2016, 88, 7689.

[17] Aguirre, N. F.; Díaz-Tendero, S.; Hervieux, P.-A.; Alcamí, M.; Martín, F. Journal of Chemical Theory and Computation 2017, 13, 992.

[18] Grimme, S. Angewandte Chemie International Edition 2013, 52, 6306.

[19] Scheubert, K.; Hufsky, F.; Böcker, S. Journal of Cheminformatics 2013, 5, 1.

[20] Spackman, P. R.; Bohman, B.; Karton, A.; Jayatilaka, D. International Journal of Quantum Chemistry 2017.

[21] Mayer, I.; Gomory, A. Chemical Physics Letters 2001, 344, 553.

[22] George, H. Journal of the American Chemical Society 1955, 77, 334.

[23] Gould, M. D.; Taylor, C.; Wolff, S. K.; Chandler, G. S.; Jayatilaka, D. Theoretical Chemistry Accounts 2008, 119, 275.

[24] McLafferty, F. W. Analytical Chemistry 1959, 31, 82.

[25] Bader, R. F. Atoms in Molecules; Wiley Online Library, 1990.

[26] Parr, R. G.; Yang, W. Density-functional theory of atoms and molecules; Oxford university press, 1989.

[27] Breuer, H.-P.; Petruccione, F. The theory of open quantum systems; Oxford University Press on Demand, 2002.

[28] Jayatilaka, D.; Grimwood, D. TONTO: A Fortran Based Object-Oriented System for Quantum Chemistry and Crystallography; The University of Western Australia, Perth, Australia, 2003.

[29] Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, .; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian09 Revision E.01. Gaussian Inc. Wallingford CT 2009.

[30] Rappé, A. K.; Casewit, C. J.; Colwell, K.; Goddard Iii, W.; Skiff, W. Journal of the American chemical society 1992, 114, 10024.

[31] Hanwell, M. D.; Curtis, D. E.; Lonie, D. C.; Vandermeersch, T.; Zurek, E.; Hutchison, G. R. Journal of Cheminformatics 2012, 4, 1.

[32] NIST Mass Spec Data Center, Stein, S. E. director, mass spectra in nist chemistry webbook, NIST standard reference database Number 69, ed. Linstrom, P. J. and Mallard, W. G., National Institute of Standards and Technology, Gaithersburg, M. D., 2015.

[33] Grabowsky, S.; Luger, P.; Buschmann, J.; Schneider, T.; Schirmeister, T.; Sobolev, A. N.; Jayatilaka, D. Angewandte Chemie International Edition 2012, 51, 6776.

[34] Thomas, S. P.; Jayatilaka, D.; Guru Row, T. N. Physical Chemistry Chemical Physics 2015, 17, 25411.

[35] Franke, S.; Ibarra, F.; Schulz, C. M.; Twele, R.; Poldy, J.; Barrow, R. A.; Peakall, R.; Schiestl, F. P.; Francke, W. Proceedings of the National Academy of Sciences of the United States of America 2009, 106, 8877.

[36] Peakall, R.; Ebert, D.; Poldy, J.; Barrow, R. A.; Francke, W.; Bower, C. C.; Schiestl, F. P. New Phytologist 2010, 188, 437.

[37] Bohman, B.; Jeffares, L.; Flematti, G.; Byrne, L. T.; Skelton, B. W.; Phillips, R. D.; Dixon, Kingsley, W.; Peakall, R.; Barrow, R. A. Journal of Natural Products 2012, 75, 1589.

[38] Bohman, B.; Berntsson, B.; Dixon, R. C. M.; Stewart, C. D.; Barrow, R. A. Organic Letters 2014, 16, 2787.

[39] Martin, J. M.; El-Yazal, J.; François, J.-P. Molecular Physics 1995, 86, 1437.

[40] Cui, L. L.; Francis, F.; Heuskin, S.; Lognay, G.; Liu, Y. J.; Dong, J.; Chen, J. L.; Song, X. M.; Liu, Y. Biological Control 2012, 60, 108.

[41] Serrano, E.; Cornu, A.; Kondjoyan, N.; Agabriel, J.; Micol, D. ANIMAL 2011, 5, 641.

[42] Goswamia, P.; Chauhan, A.; Verma, R. S.; Padalia, R. C.; Verma, S. K.; Darokar, M. P.; Chanotiya, C. S. Journal of Essential Oil Research 2016, 28, 71.

[43] Singh, G.; Singh, O. P.; De Lampasona, M.; Catalan, C. A. Flavour and Fragrance Journal 2003, 18, 62.

[44] Radulović, N.; Denić, M.; Stojanović-Radić, Z.; Skropeta, D. Journal of the American Oil Chemists' Society 2012, 89, 2165.

[45] Davidson, E. Reduced Density Matrices in Quantum Chemistry; Academic Press, New York, 1976.

[46] Bauschlicher, C. W.; Taylor, P. R. Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta) 1988, 74, 63.

[47] Jayatilaka, D.; Graham, S. C. Molecular Physics 1997, 92, 471.
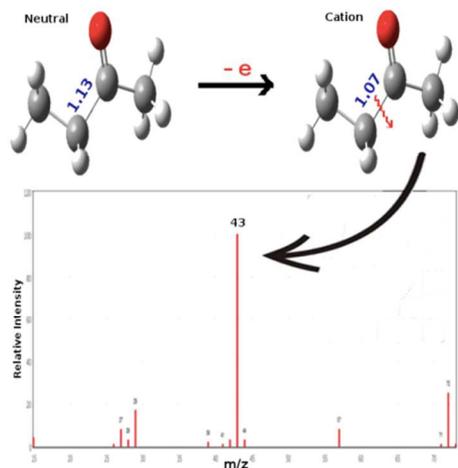
There is currently a scarcity of computational methods supporting the elucidation of unknown compounds by mass spectrometry. The highest intensity peak (base peak) in electron impact mass spectrometry (EI-MS) can be predicted correctly for 65 out of 75 test molecules using the a newly developed protocol in which Roby-Gould bond indices are applied to ground state and cation ab-initio wavefunctions.

Supporting Information

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

Khidhir Alhameedi,[*ab] Björn Bohman,[a] Amir Karton,[a] and Dylan Jayatilaka[*a]

[a] School of Molecular Sciences, University of Western Australia, 35 Stirling Highway Nedlands 6009, Australia.

[b] Department of Chemistry, College of Education for Pure Science, University of Karbala, Karbala, Iraq.

E-mail:

dylan.jayatilaka@uwa.edu.au,

khidhir.alhameedi@research.uwa.edu.au

Contents

ii

Table S1: Chemical formulae for data set B.

Chemical Formula	No. of isomers
Ethers and Carbonyls:	
$C_8H_{14}O_2$	3
$C_{10}H_{16}O$	5
Chiloglottones:	
$C_{11}H_{18}O_2$	6
$C_{13}H_{22}O_2$	3
Pyrazines:	
$C_9H_{14}N_2$	1
$C_{11}H_{18}N_2$	1
$C_{10}H_{16}N_2O$	1
$C_{11}H_{18}N_2O$	1
Heptanenitrile derivative:	
$C_7H_{13}NS$	4
Unusual ring systems:	
$C_{15}H_{24}$	3

S1 Detailed analysis of predicting base peak for each class

Here we discuss in more detail observations concerning the application of protocol M for each chemical class.

S1.1 Alcohols 1-12

The alcohols here have molecular formulae C_3H_7OH (2 isomers), C_4H_9OH (3 isomers) and $C_5H_{11}OH$ (7 isomers) and comprise primary, secondary and tertiary alcohols. One example is shown in Figure S1.
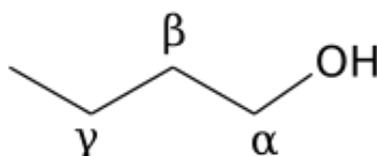
Figure S1: 1-butanol molecule as example of showing the position of the C_α–C_β bond in all molecules.

- Protocol M gives the correct result for all molecules except 3-methyl-1-butanol, **9**.

- In the case of molecule **9**, we predicted the base peak to be 57 compared to the experimental value of 55; this may correspond to a double hydrogen loss.

- In all cases of alcohols, the C_α–C_β bond was broken during optimization of the cations. This may be rationalized by noting that the Roby-Gould bond index values of C$-$O increased in the cationic species after removing an electron from neutral species (and the corresponding bond length values decreased) and so presumably the C_α–C_β bond was weakened enough to rupture.

- For secondary and tertiary alcohols, the C_α–C_β bond in the long (rather than short) part of the alkyl chain was predicted to be cleaved.

- In the cases where the cation dissociated during optimization which occurred with all alcohols except 3-pentanol and 2-methylbutanol, we assumed that those two fragments corresponded to the base peak in the experimental spectrum.

- The positive charge is associated with the fragment that had the OH group, except when the other fragment has a longer carbon chain.

2

S1.2 Amines 13-22

The amines here have molecular formulae C_3H_9N (3 isomers) and $C_4H_{11}N$ (7 isomers) and comprise primary, secondary and tertiary amines.

- Protocol M gives the correct result for all molecules.

- In all cases of amines, except trimethylamine (**15**), the C_α–C_β bond was predicted to break. This may be rationalized by noting that the Roby-Gould bond index values of C−N increased in the cationic species after removing an electron from neutral species (and the corresponding bond length values decreased) and so presumably the C_α–C_β bond was weakened enough to rupture.

- For trimethylamine, **15**, the C−H bond is predicted to break, as there is no C−C bond.

- For secondary and tertiary amines, the C_α–C_β bond in the long (rather than short) part of the alkyl chain was predicted to be cleaved.

S1.3 Alkenes 23-34

The alkenes here have molecular formulae C_5H_{10} (3 isomers) and C_6H_{12} (9 isomers).

- Protocol M gives the correct result for all molecules except 3-hexene (**28**) and 2-methyl-1-pentene (**29**) .

- In the case of molecule **28**, we predicted the base peak to be 55 compared to the experimental value of 56; this may correspond to a hydrogen gain.

- In the case of molecule **29**, we predicted the 4th experimental peak (m/z=69) in EI-MS.

- In all cases of alkenes, the C_β–C_γ bond was was predicted to break. This may be rationalized by noting that the Roby-Gould bond index values of C=C decreased in the cationic species after removing an electron from neutral species (and the corresponding bond length values increased) and so presumably the C_β–C_γ bond was weakened enough to rupture.

- In all cases, the C=C double bond have the biggest value of (Δ_{CC}^{max}) in protocol M.

S1.4   Aldehydes and Ketones 35-40

The aldehydes and ketones here have molecular formulae C_4H_8O formula (2 isomers) and $C_5H_{10}O$ formula (4 isomers).

- Protocol M gives the correct result for all molecules.

- In all cases of aldehyde and ketone compounds, the C_α–C_β bond was predicted to break. This may be rationalized by noting that the Roby-Gould bond index values of C=O increased in the cationic species after removing an electron from neutral species (and the corresponding bond length values decreased) and so presumably the C_α–C_β bond was weakened enough to rupture.

- For ketones, the predicted cleavages took place in the C_α–C_β bond in the long part of the alkyl chain.

S1.5   Thiols and Sulfides 41-55

The thiols and sulfides here have molecular formulae $C_4H_{10}S$ (9 isomers) and $C_5H_{12}S$ (6 isomers) and comprise primary and secondary thiols.

- Protocol M gives the correct result for 12 out of 15 molecules. The incorrect predictions were with 1-butanethiol **41**, 2-butanethiol **42** and 3-methyl-1-butanethiol **48**.

- In the case of molecules **41** and **48**, we predicted the base peak to be 43 and 57 compared to the experimental value of 41 and 55 respectively; this may correspond to a double hydrogen loss.

- In the case of molecule **42**, we predicted the other peaks which are the second base peak to be 61 and 29 in EI-MS.

- In all cases of thiols, the C_α–C_β bond was broken during optimization of the cations. This may be rationalized by noting that the Roby-Gould bond index values of C−S increased in the cationic species after removing an electron from neutral species (and the corresponding bond length values decreased) and so presumably the C_α–C_β bond was weakened enough to rupture.

- For secondary thiols, the C_α–C_β bond in the long (rather than short) part of the alkyl chain was predicted to be cleaved.

- In the cases where the cation dissociated during optimization which occurred with all alcohols except 3-pentanol and 2-methylbutanol, we assumed that those two fragments corresponded to the base peak in the experimental spectrum.

- The positive charge is associated with the fragment that had the SH group, except when the other fragment has a longer carbon chain.

S1.6 Methyldecanes 56-59

The methyldecanes here have molecular formulae $C_{11}H_{24}$ (4 isomers), as a long hydrocarbon chains.

- Protocol M gives the correct result for 2 out of 4 molecules. The incorrect prediction were with 3-methyldecane **57** and 5-methyldecane **59** .

- In the case of molecules **57** and **59**, we predicted the peak at experimental m/z ratio of 29 (around 35 relative intensity in EI-MS).

S1.7 Phenols 60-70

The phenols here have molecular formulae C_7H_7OH (3 isomers), C_8H_9OH (5 isomers) and $C_6H_6O_2$ (3 isomers).

- Protocol M gives the correct result for all molecules.

- In the case of molecules **57** and **59**, we predicted the peak at experimental m/z ratio of 29 (around 35 relative intensity in EI-MS).

- In all cases of phenols, the C-C ring bond was predicted to break. This gives molecular ions as base peak peak in all molecules.

S1.8 Selected natural products 71-75

These compounds have molecular formulae $C_{15}H_{24}$ (1 molecule), $C_8H_{14}O_2$ (1 molecule) and $C_{10}H_{16}O$ (3 isomers).

- Protocol M gives the correct result for three molecules out of five. These two failure were *(E)-β*-farnesene **71** and *cis*-tagetone **74**.

- In the case of molecule **71**, we predicted the second highest peak in EI-MS by adopting protocol M.

- In the case of molecule **74**, we predicted the base peak with the second value of $\Delta_{\rm CC}^{\rm max}$ in protocol M.

S2   Analysis of the protocols MBI0, MBI+ and MCMG

Figure S2 shows the results of predicting base peak peaks for selected molecules in the data set A with protocol MBI0. Although it achieved some success for alcohols, amines and alkenes, it is rather unsuccessful for the other five chemical classes.

The same Figure S2 shows the results with protocol MBI+. The success rate here is even lower than that with MBI0 protocol.

Consequently we reject these two as a viable protocols for predicting base peak.

Because the success rate of these protocols is rather low, and these protocols differ from those in the main text only in the fact that bond indices between all atoms are considered (rather than only C–C bonds) we conclude that if the bond order differences are valid at all for the non-C–C bonds, then they must be on a different scale than those between carbon atoms (otherwise we would have obtained correct predictions). This scale change is not surprising since the bond dissociation energies for different bonds are also group roughly on different scales according to bond type. It is therefore rather fortunate that we were able to restrict our attention to only C–C bonds and obtain correct results; it may have happened that some of the C–X bonds were more likely to fragment that the C–C bonds but our data set did not include any such examples, although there undoubtedly must be such cases. Future protocols focused on predicting the base peak of inorganic compounds will therefore need to place bond index differences on a common scale, perhaps an energy scale.

(a) MBI0 protocol.
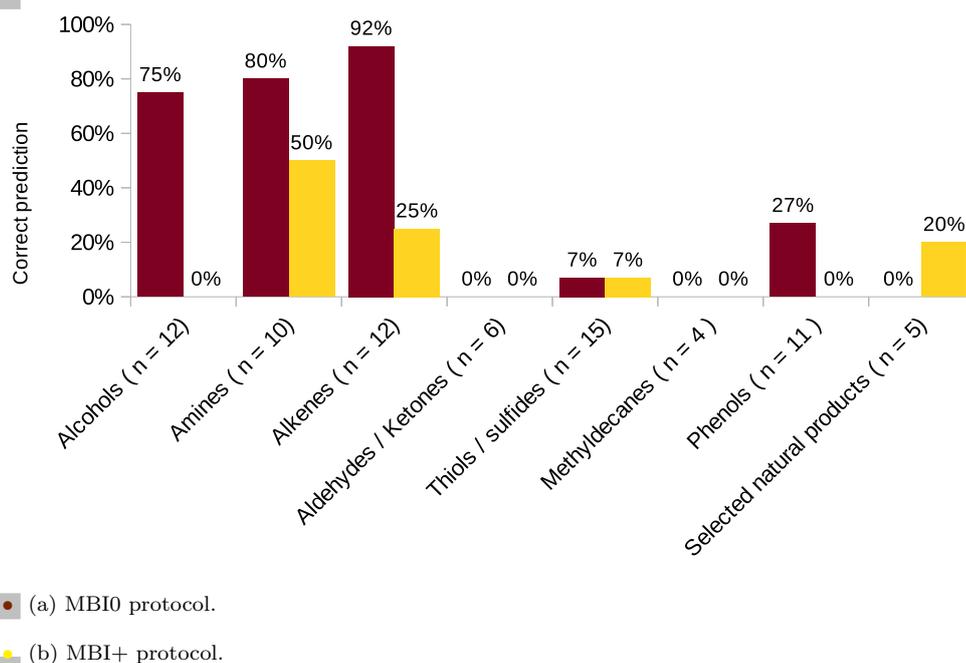
(b) MBI+ protocol.

Figure S2: The percentage of correct predictions for the highest intensity peak in the mass spectrum (base peak) in data set A, using (a) MBI0 protocol and (b) MBI+ protocol Percentages as fractions are given on top of the histogram. Number of molecules are given beside each class name.

Figure S3 illustrates the number of correct predictions for base peak for the MCMG protocol, evaluated on data set A. Since this protocol is a combination of MC and M it is perhaps not too surprising that the results are slightly better than for the MC method alone. In fact the success rate of MCMG protocol is exactly the same M protocol, discussed in the main text. This means MCMG protocol has no effect on the M protocol in predicting experimental base peak peaks in mass spectrum. Consequently we reject this protocol in favor of protocol M alone.
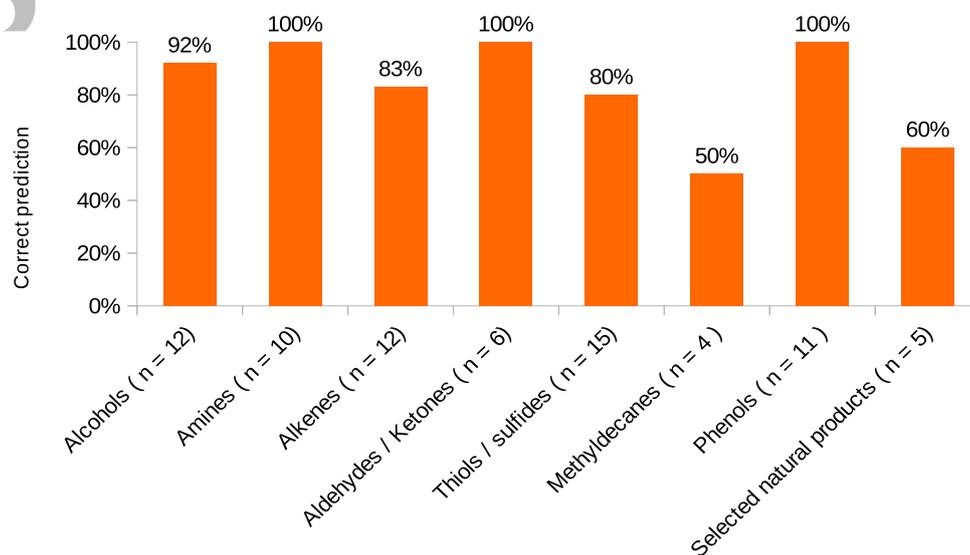
Figure S3: The number and percentage of correct predictions for the highest intensity peak in the mass spectrum (base peak) in data set A, using MCMG protocol. Percentages as fractions are given on top of the histogram. Number of molecules are given beside each class name.

S3   Molecules in data set A

In this section we present the molecular structures of neutral species and cations for data set A (Table 1 in the main text) as well as the Roby-Gould bond indices, percent covalency, and the experimental mass spectra taken from the NIST Web book database.

For the alcohols , **1-12**, all formed cations except from 3-pentanol **7** and 2-methylbutanol **9** dissociated during optimization; consequently, Roby-Gould bond indices do not appear above the broken bond for those cases.

It is important to realise that these figures were generated automatically from the gaussian FChk files using a bash script which called the `openbabel` program, the `mol2chemfig` package, and LaTeX program. Therefore some of the layout and drawing of the structures reflect idiosyncrasies of this automatic scheme.

S3.1   Alcohols

S3.1.1   Isomers of C_3H_7OH formula 1-2
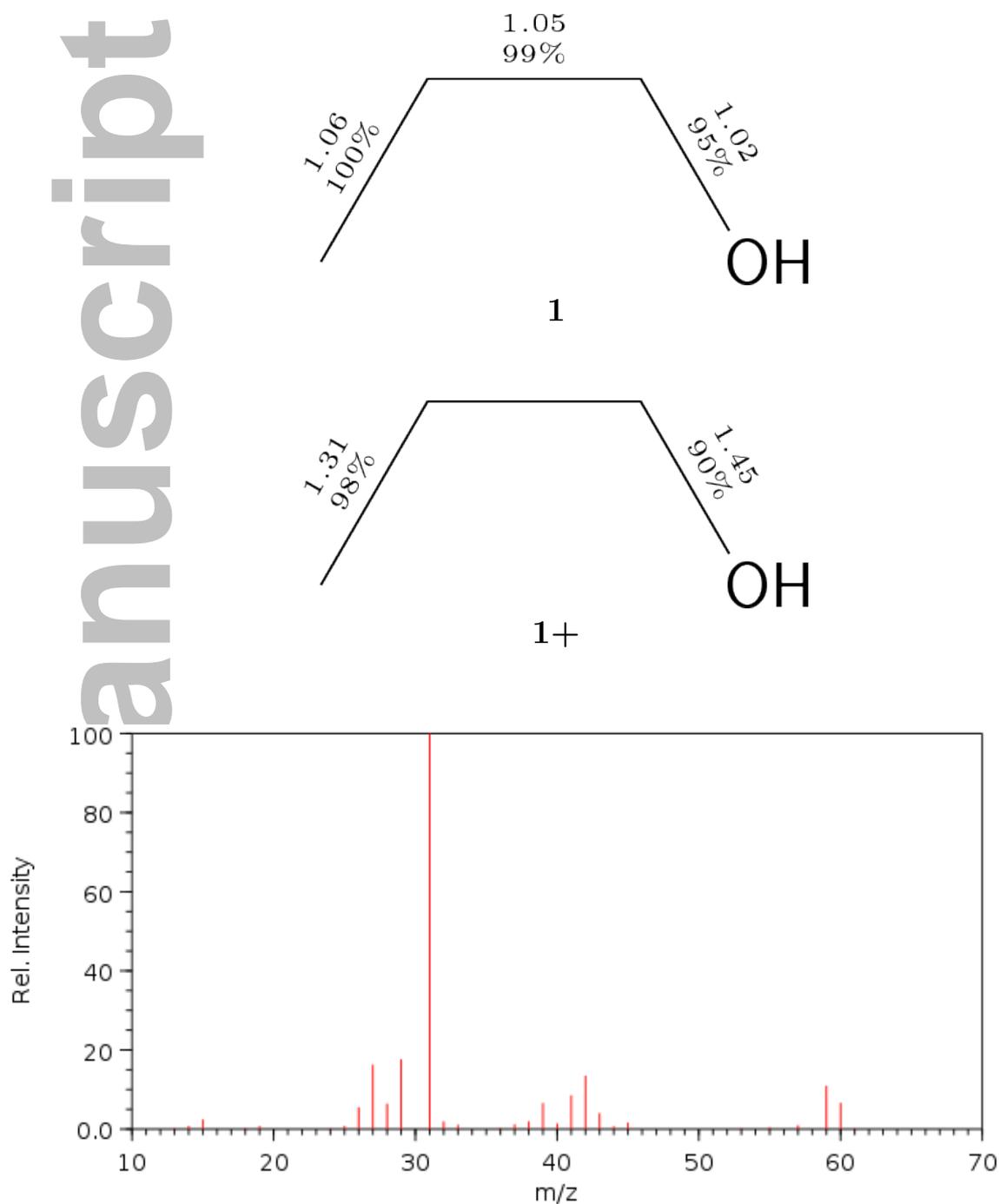
Figure S1: Structure of molecule **1** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**1**) and cation (**1+**).
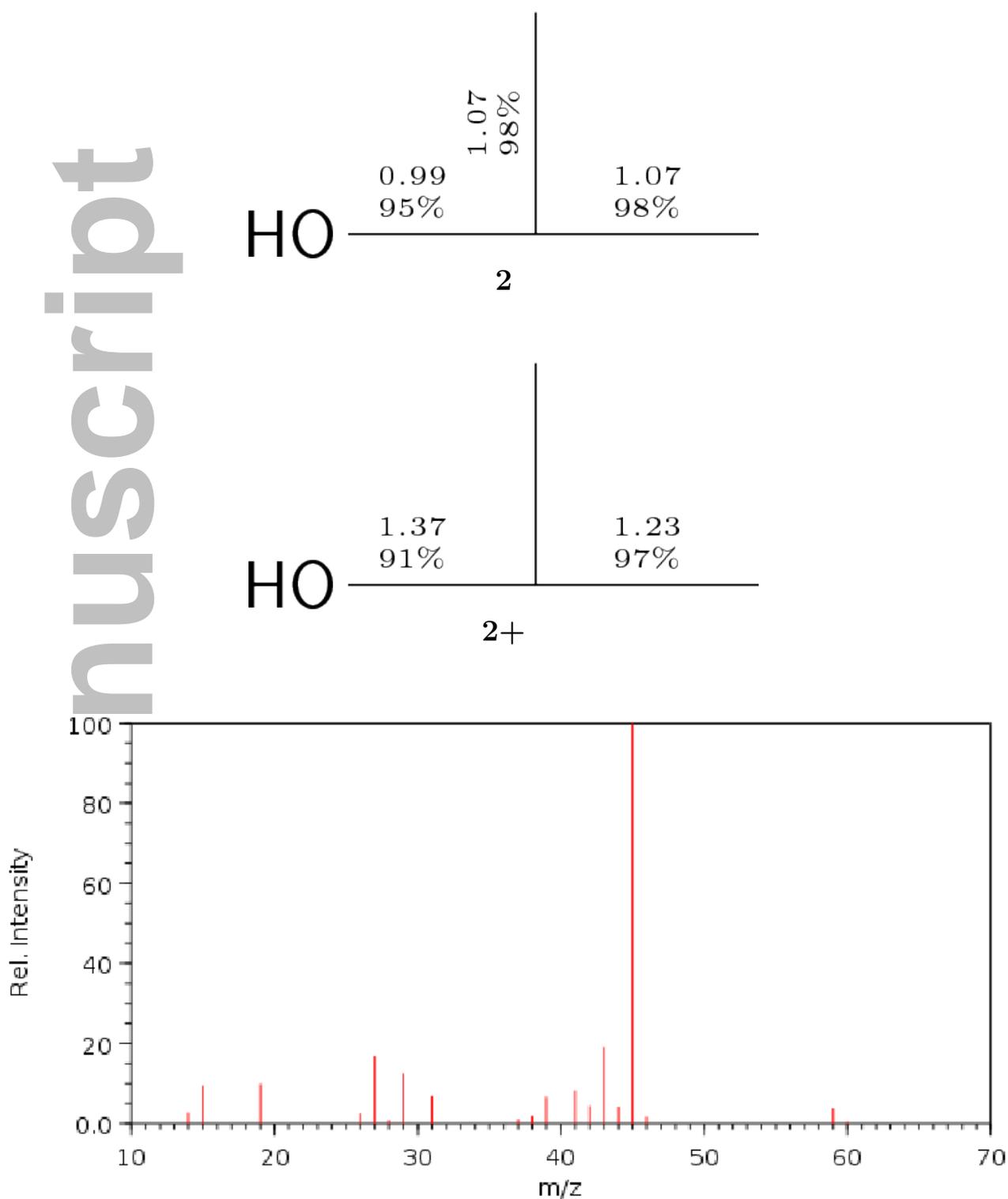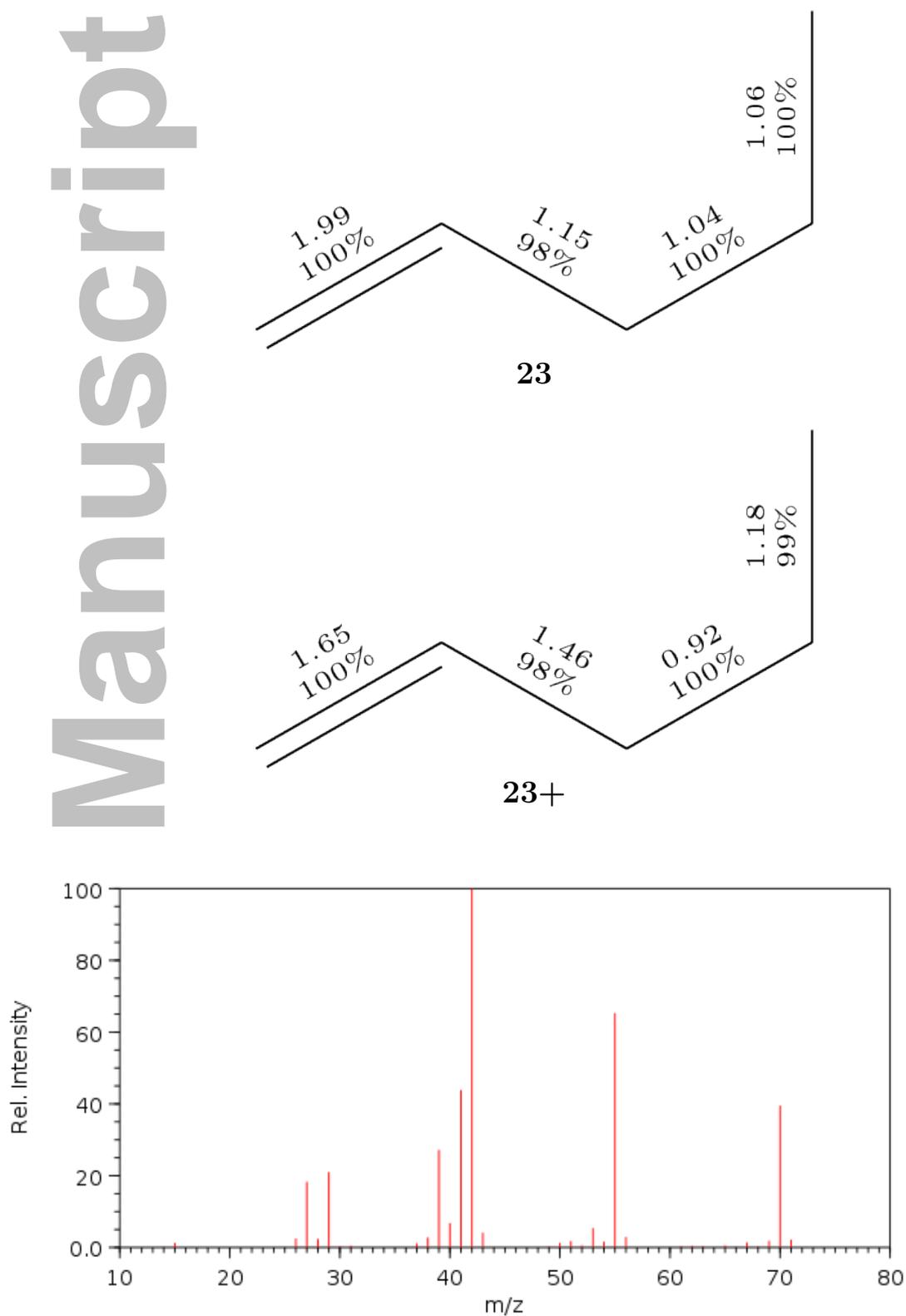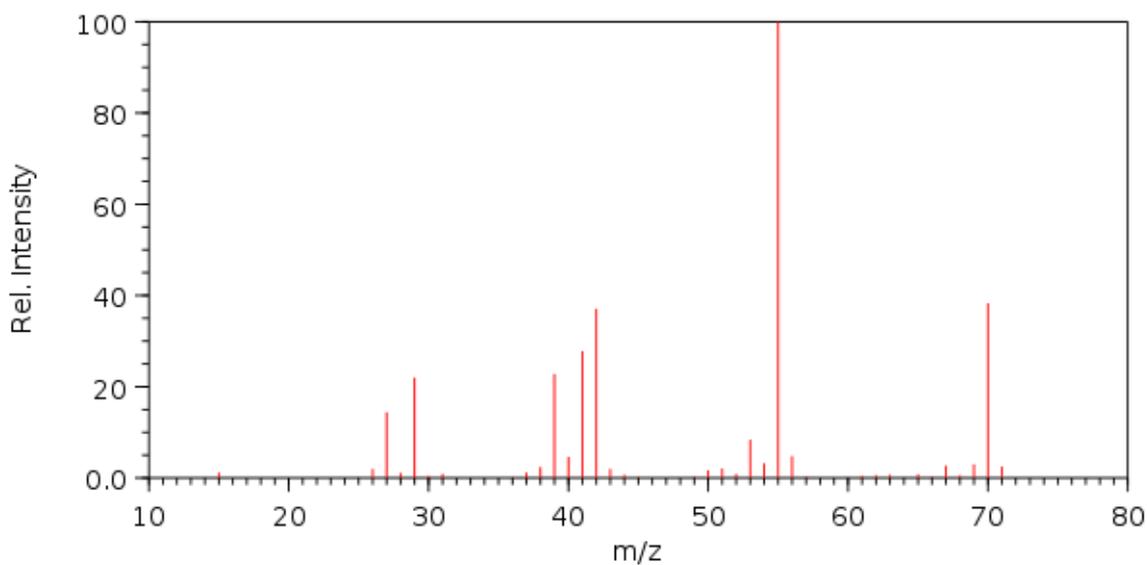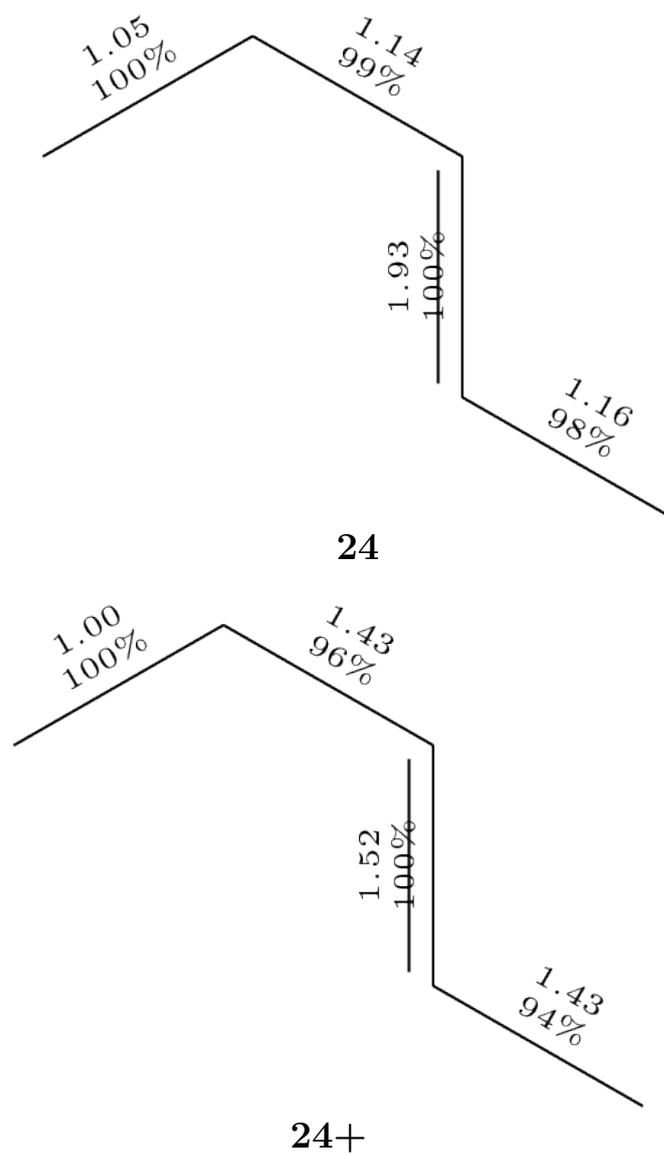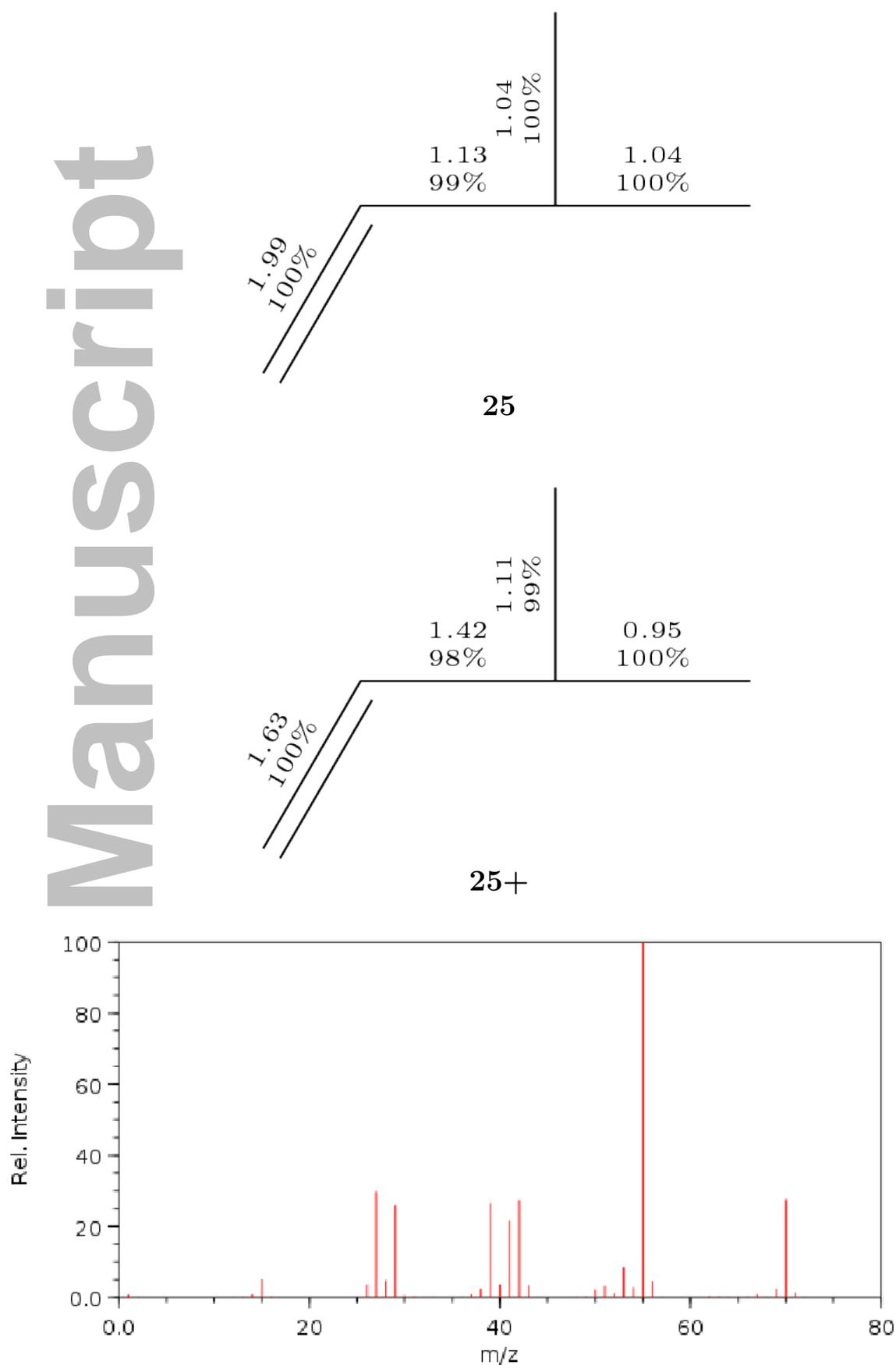
Figure S2: Structure of molecule **2** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**2**) and cation (**2+**).
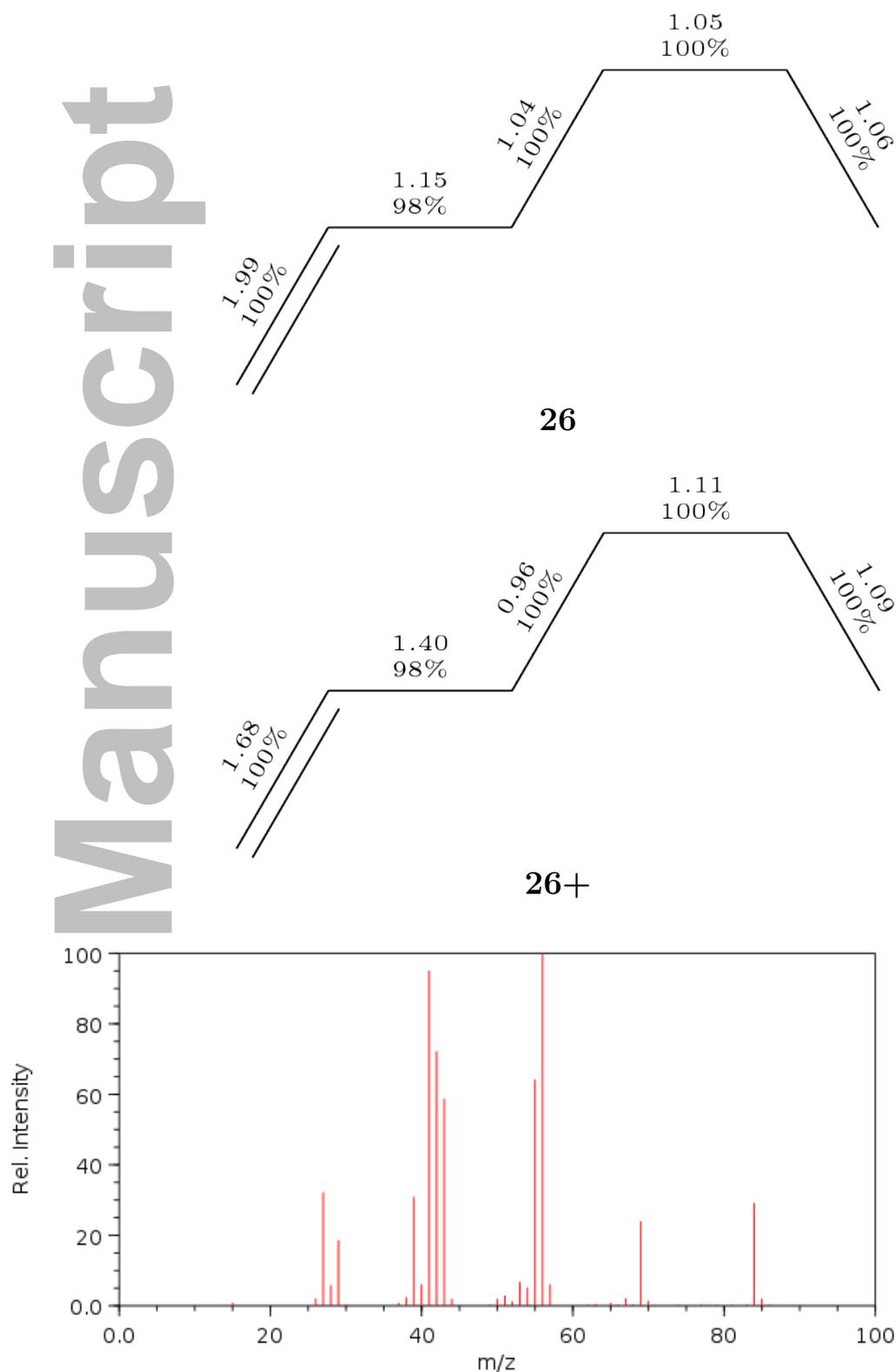
S3.1.2   Isomers of C_4H_9OH formula 3-5
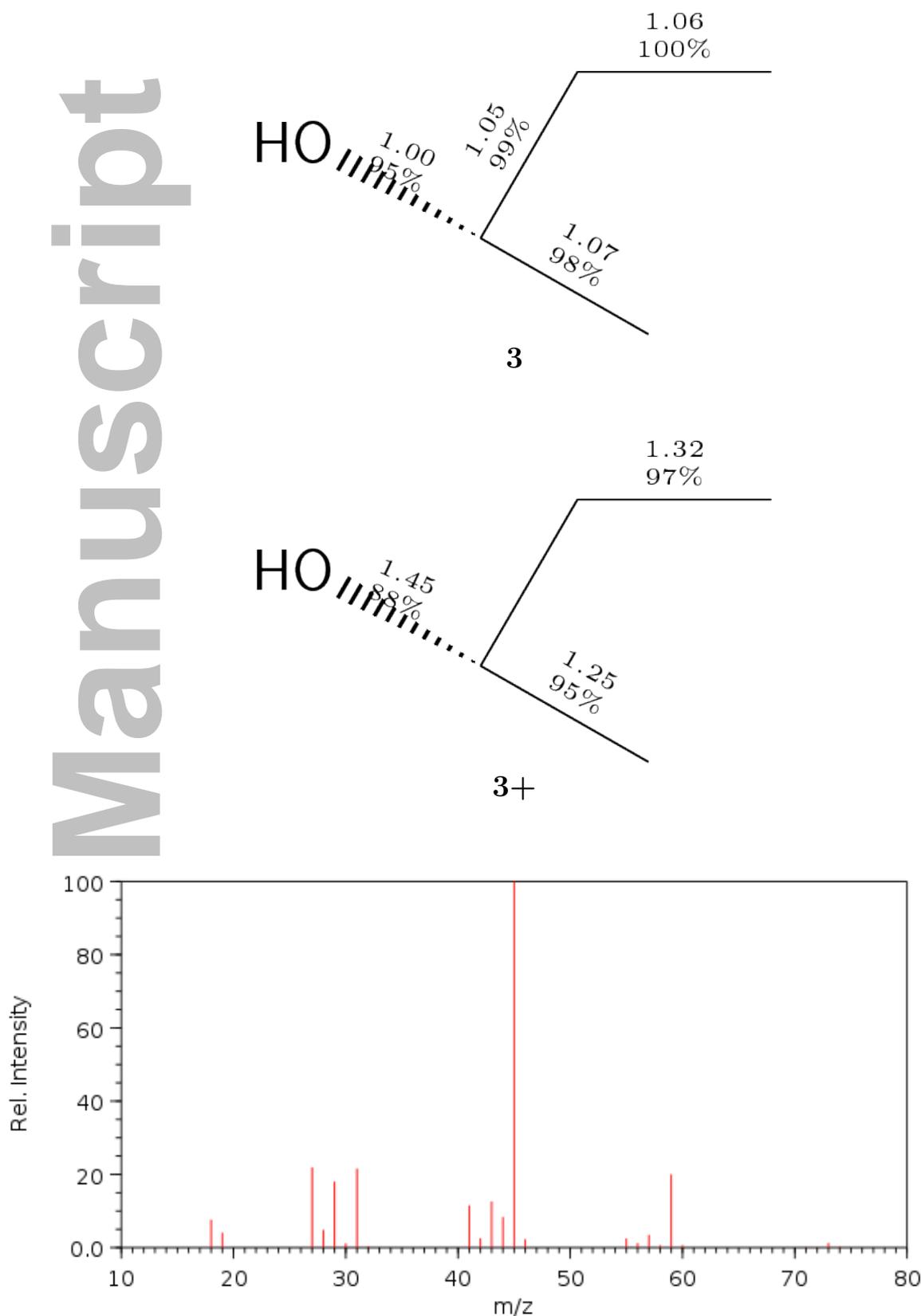
Figure S3: Structure of molecule **3** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**3**) and cation (**3+**).
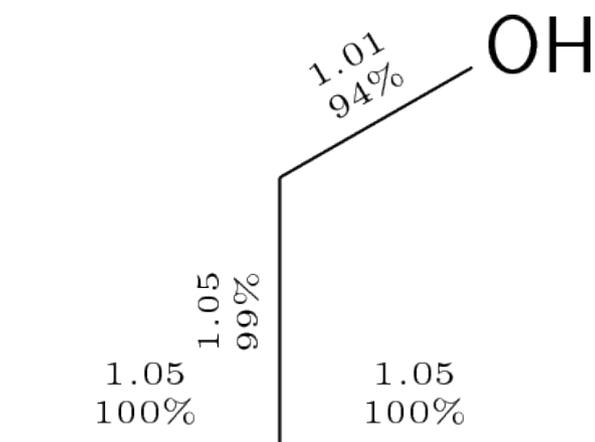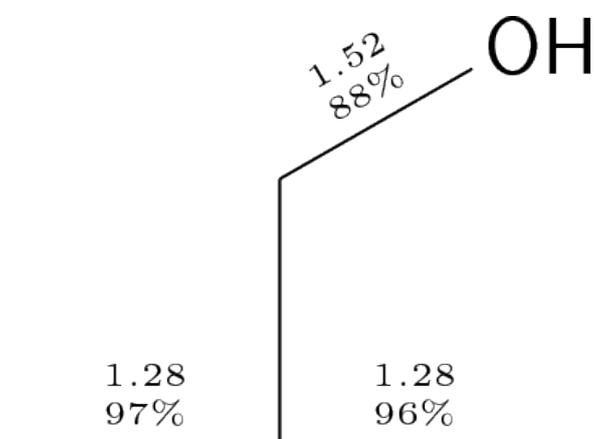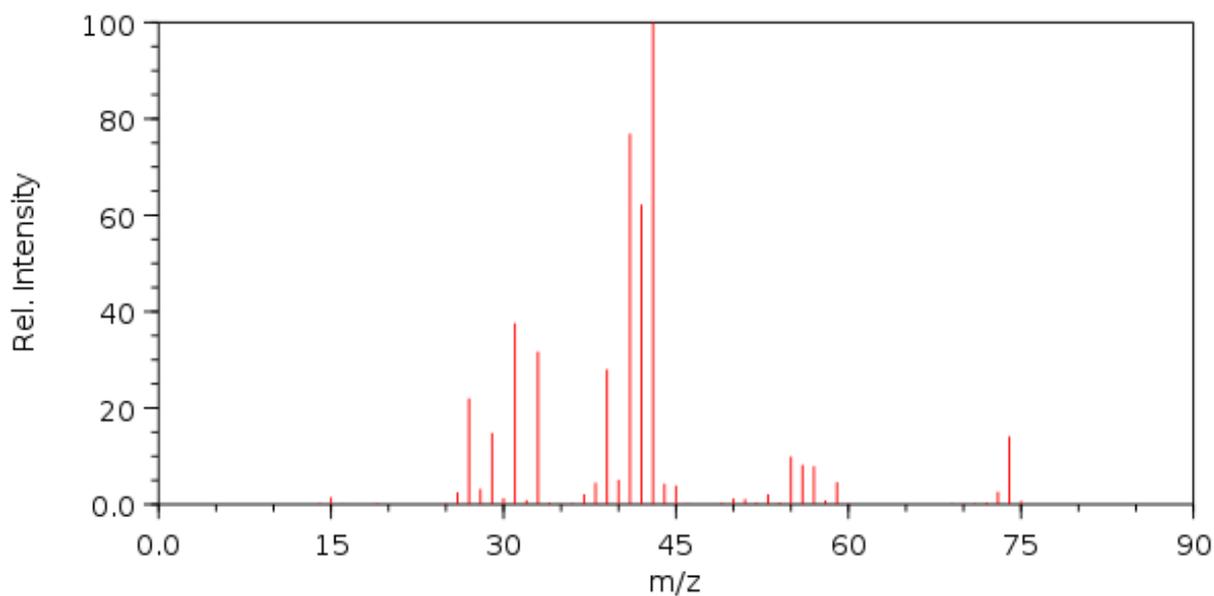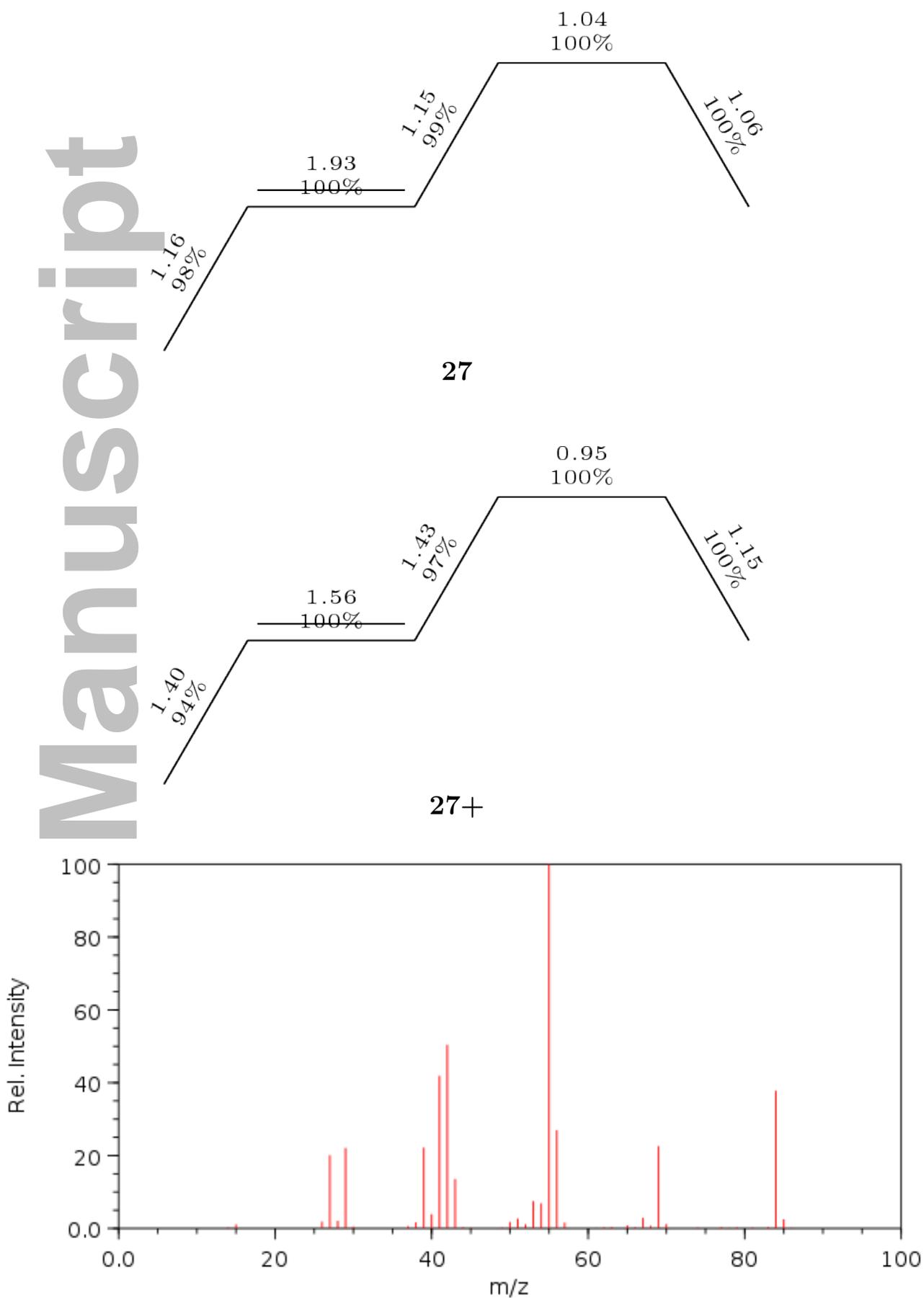
Figure S4: Structure of molecule **4** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**4**) and cation (**4+**).
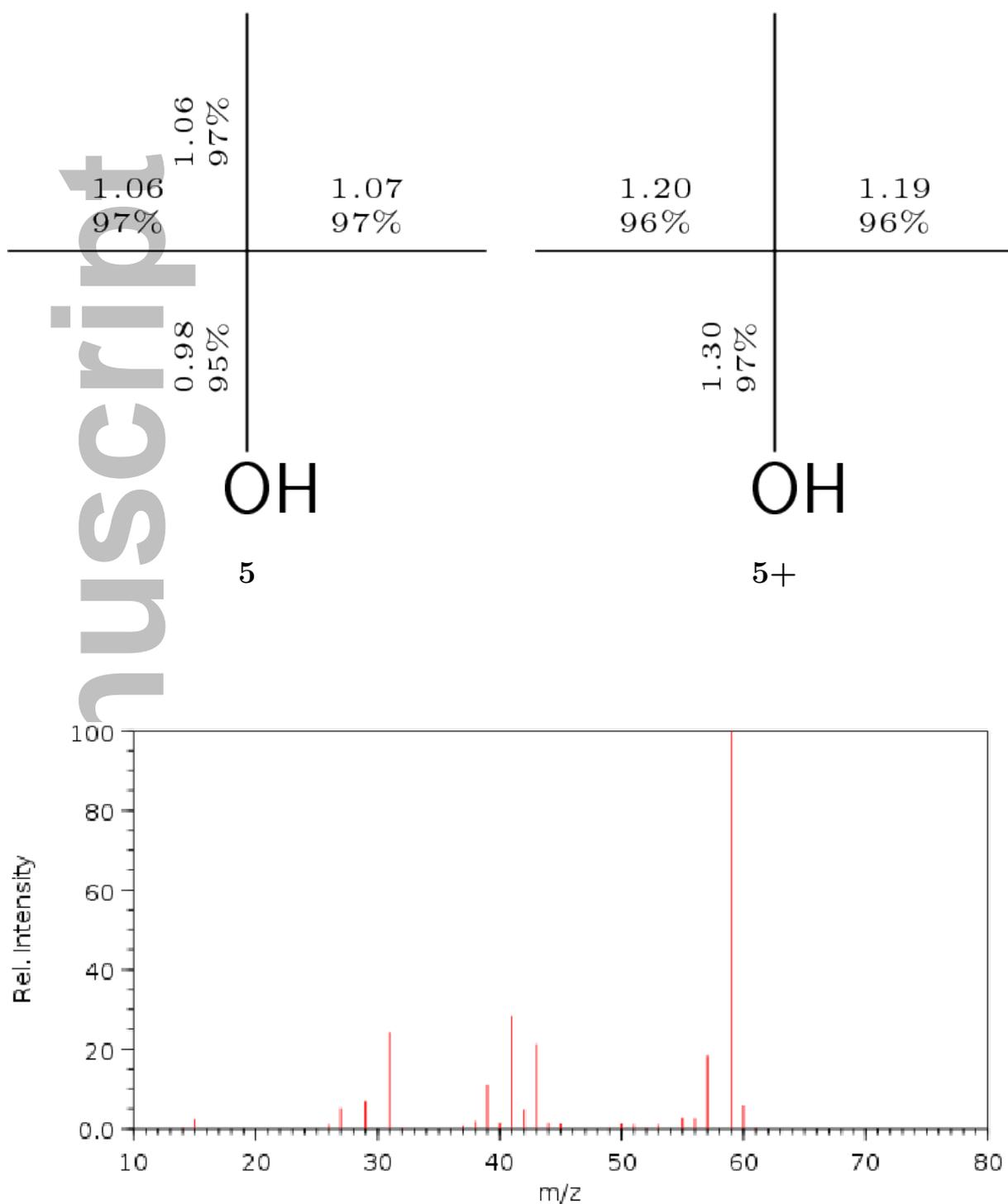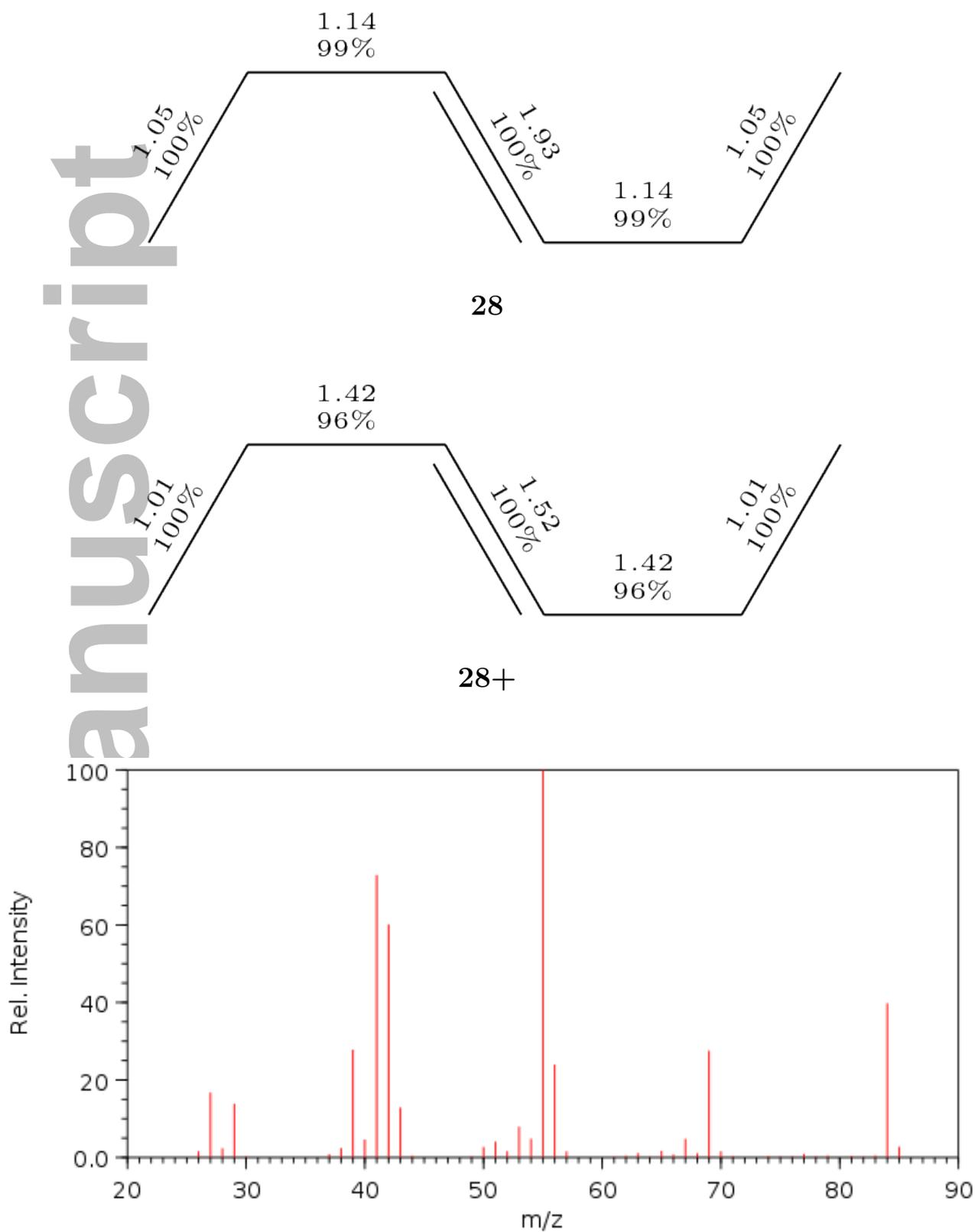
Figure S5: Structure of molecule **5** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**5**) and cation (**5+**).

S3.1.3   Isomers of $C_5H_{11}OH$ formula 6-12
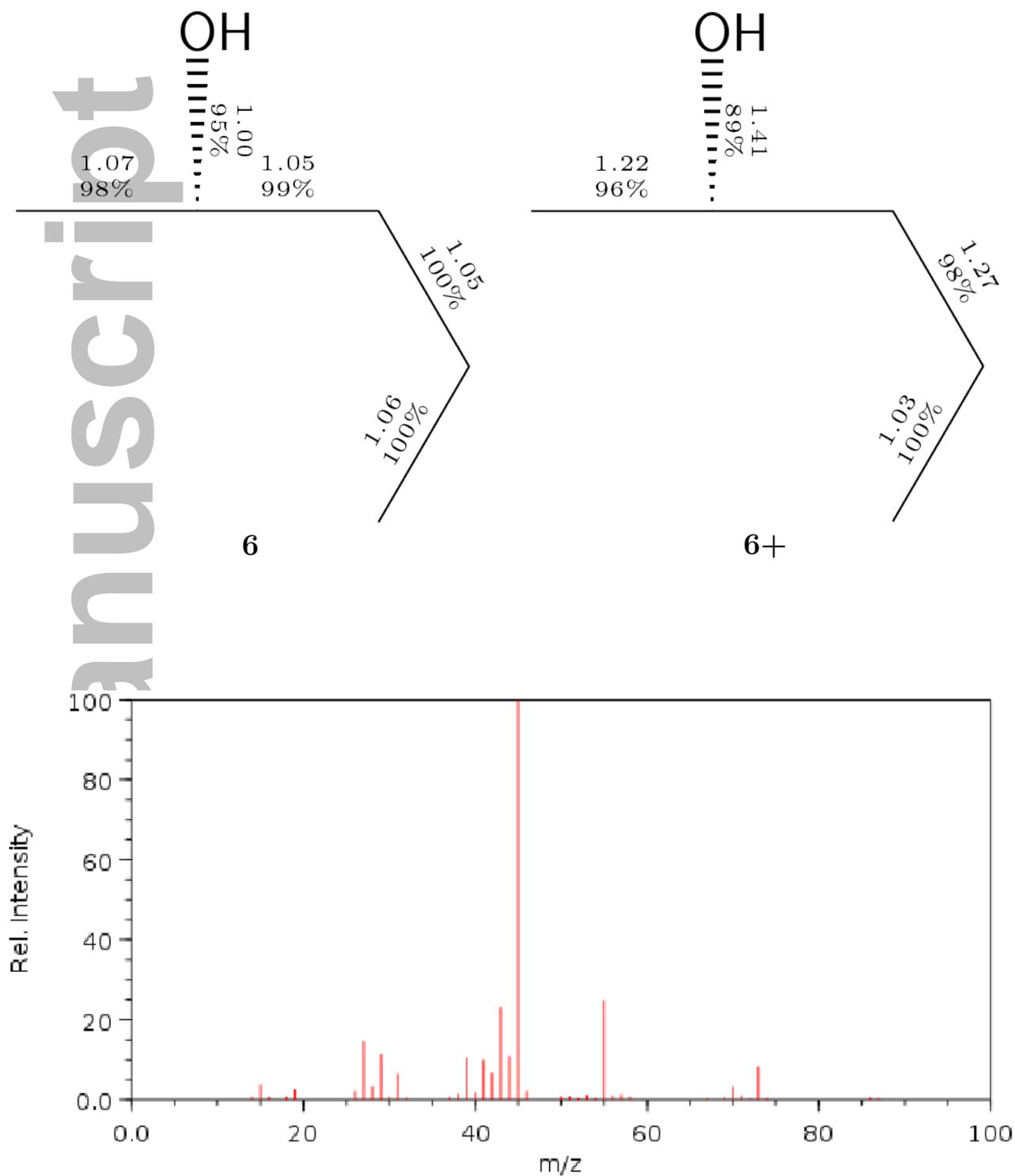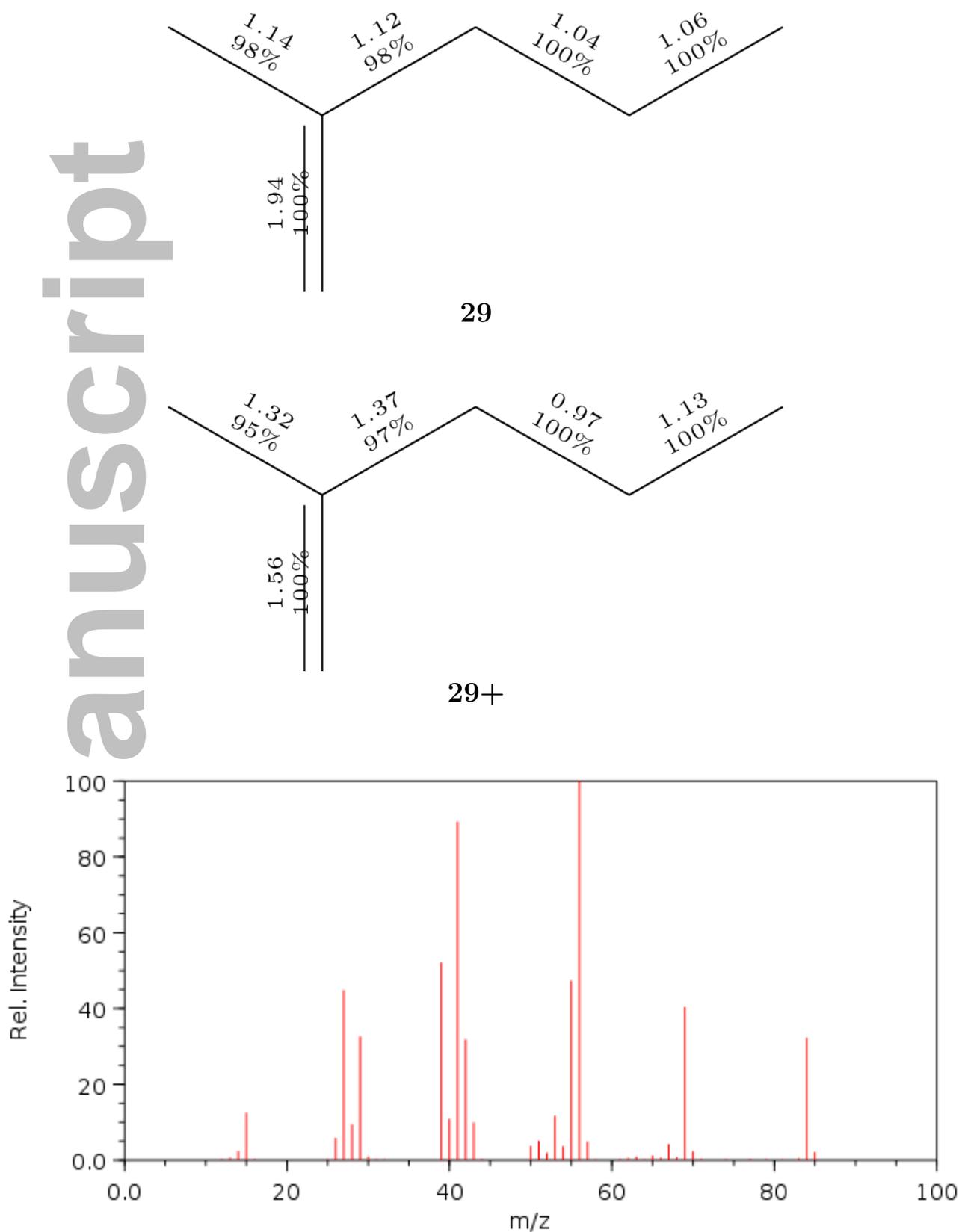

Figure S6: Structure of molecule **6** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**6**) and cation (**6+**).
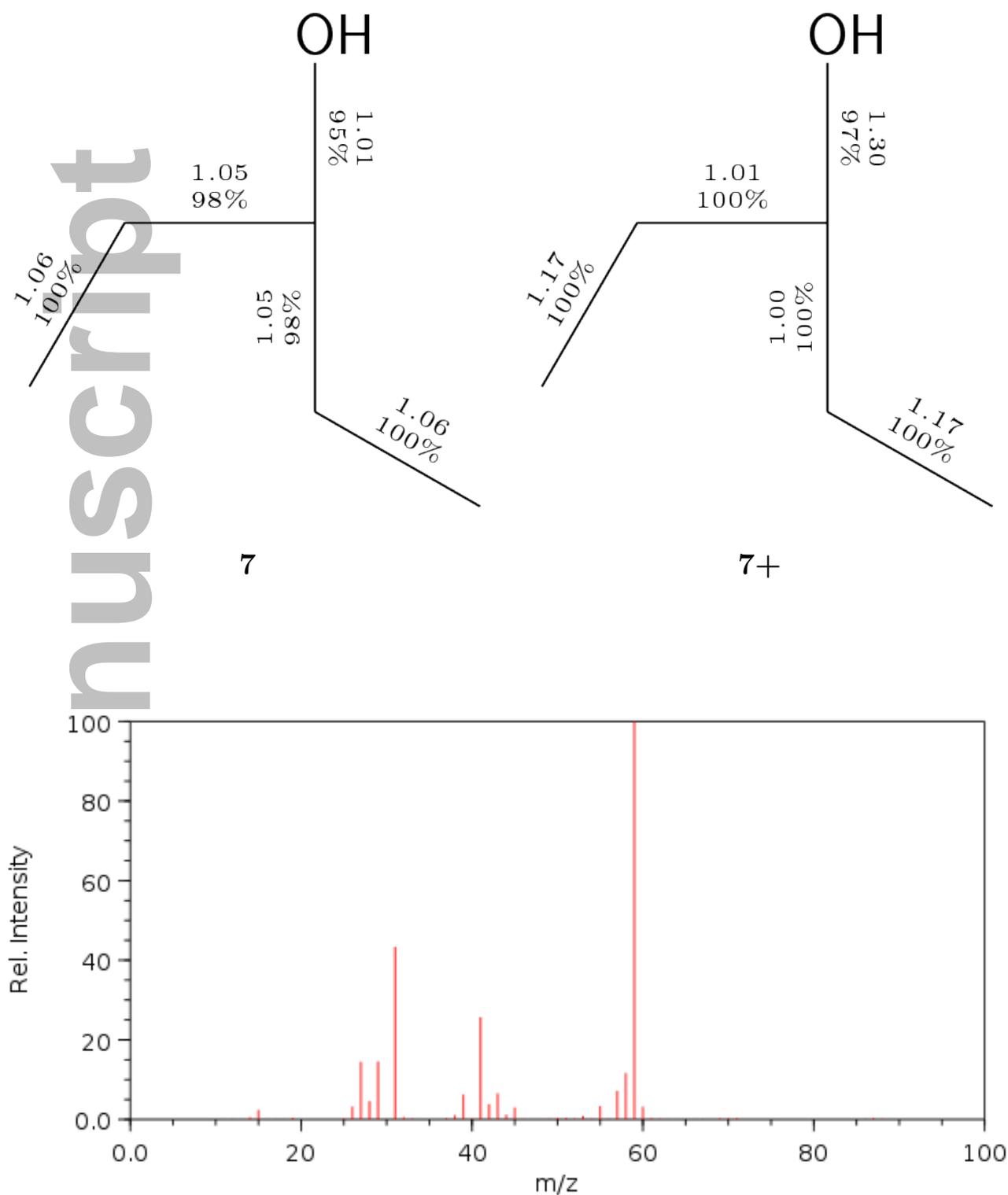
Figure S7: Structure of molecule **7** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**7**) and cation (**7+**).
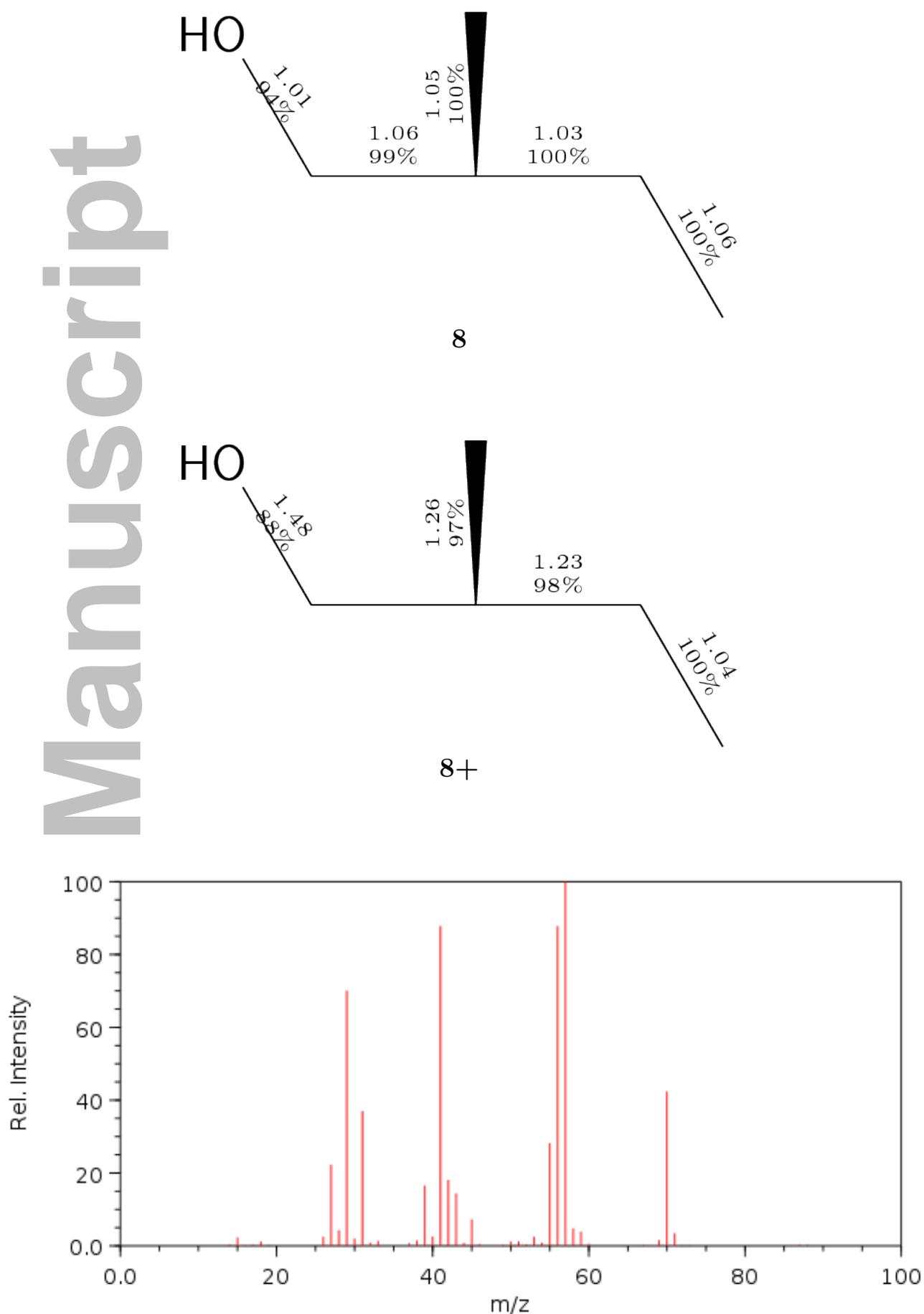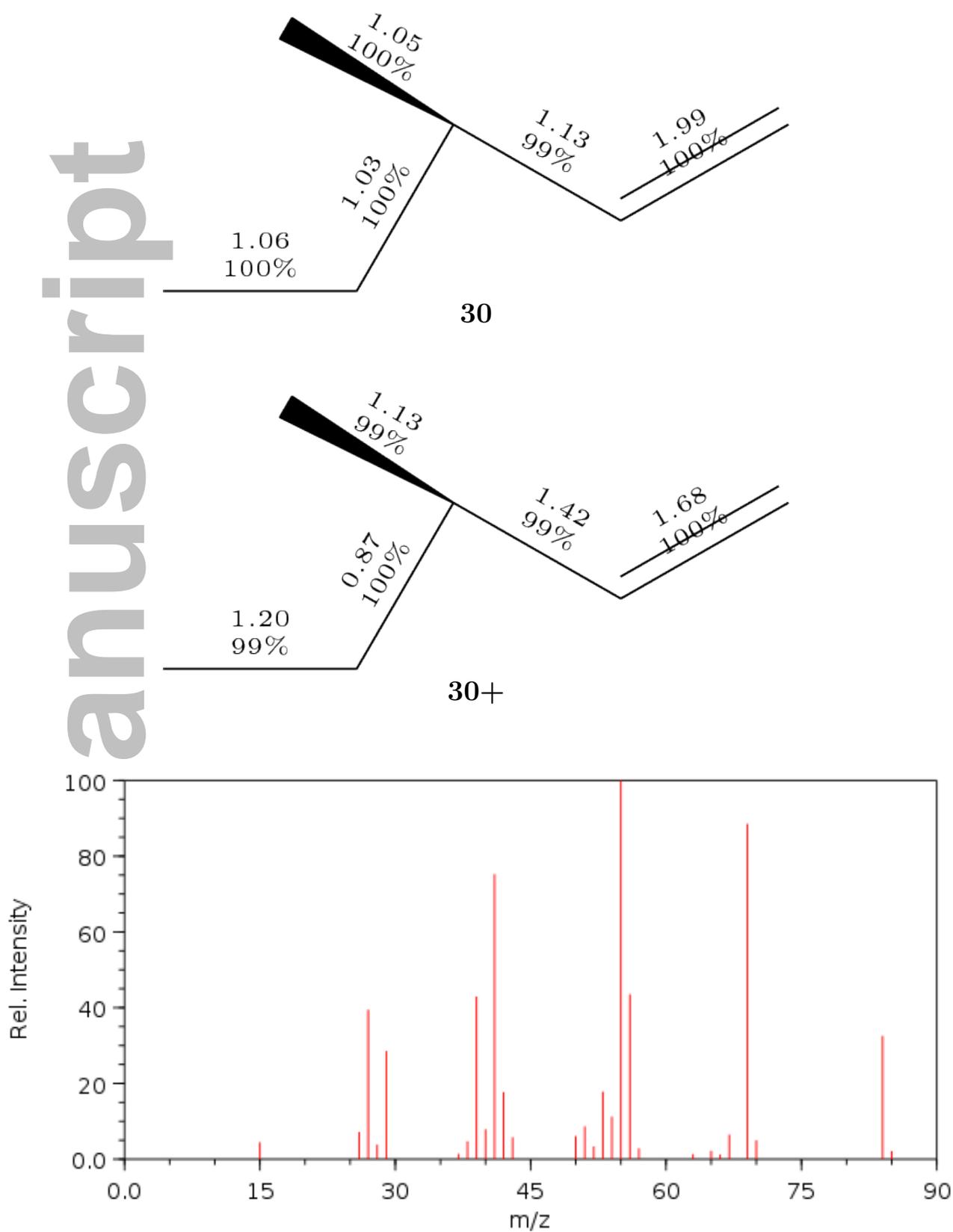
Figure S8: Structure of molecule **8** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**8**) and cation (**8+**).
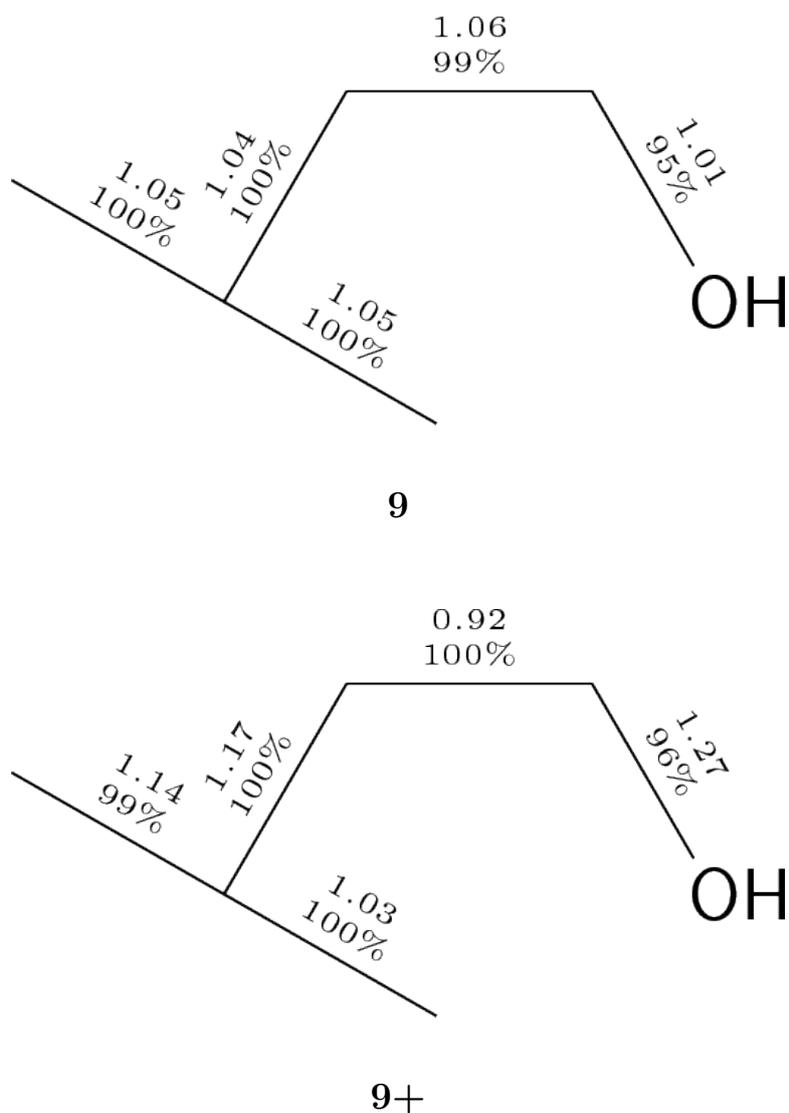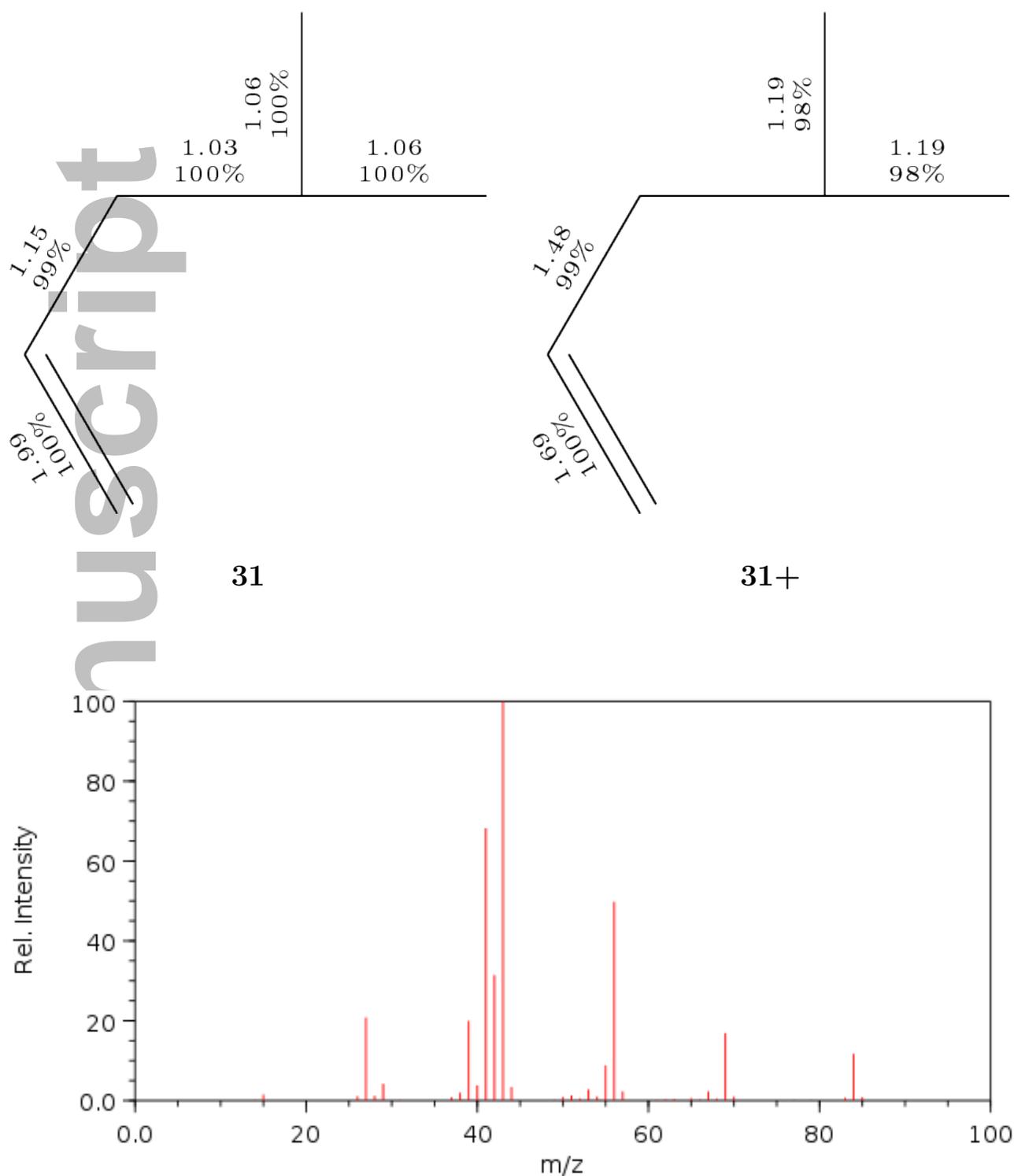
Figure S9: Structure of molecule **9** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**9**) and cation (**9+**).
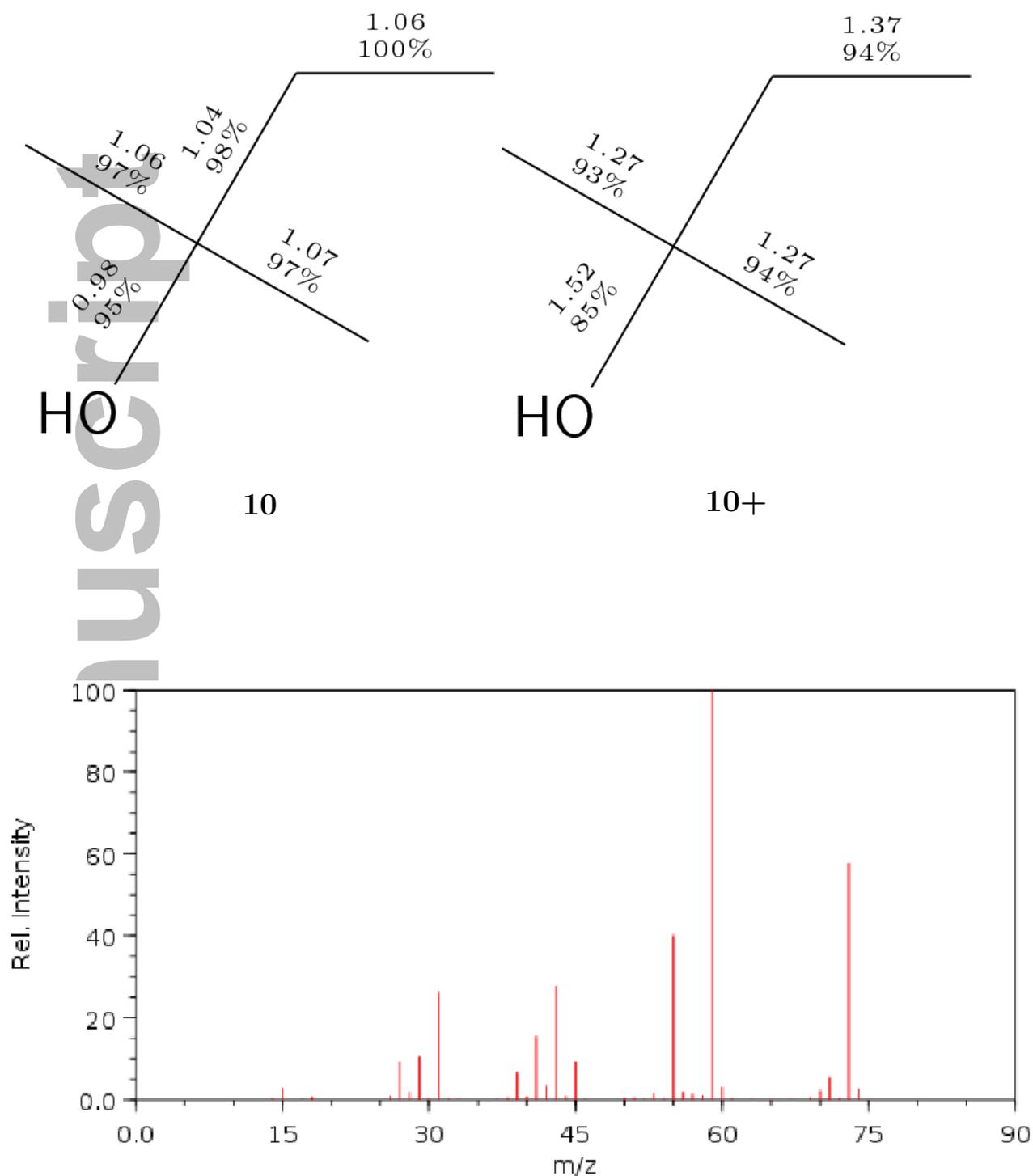
Figure S10: Structure of molecule **10** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**10**) and cation (**10+**).
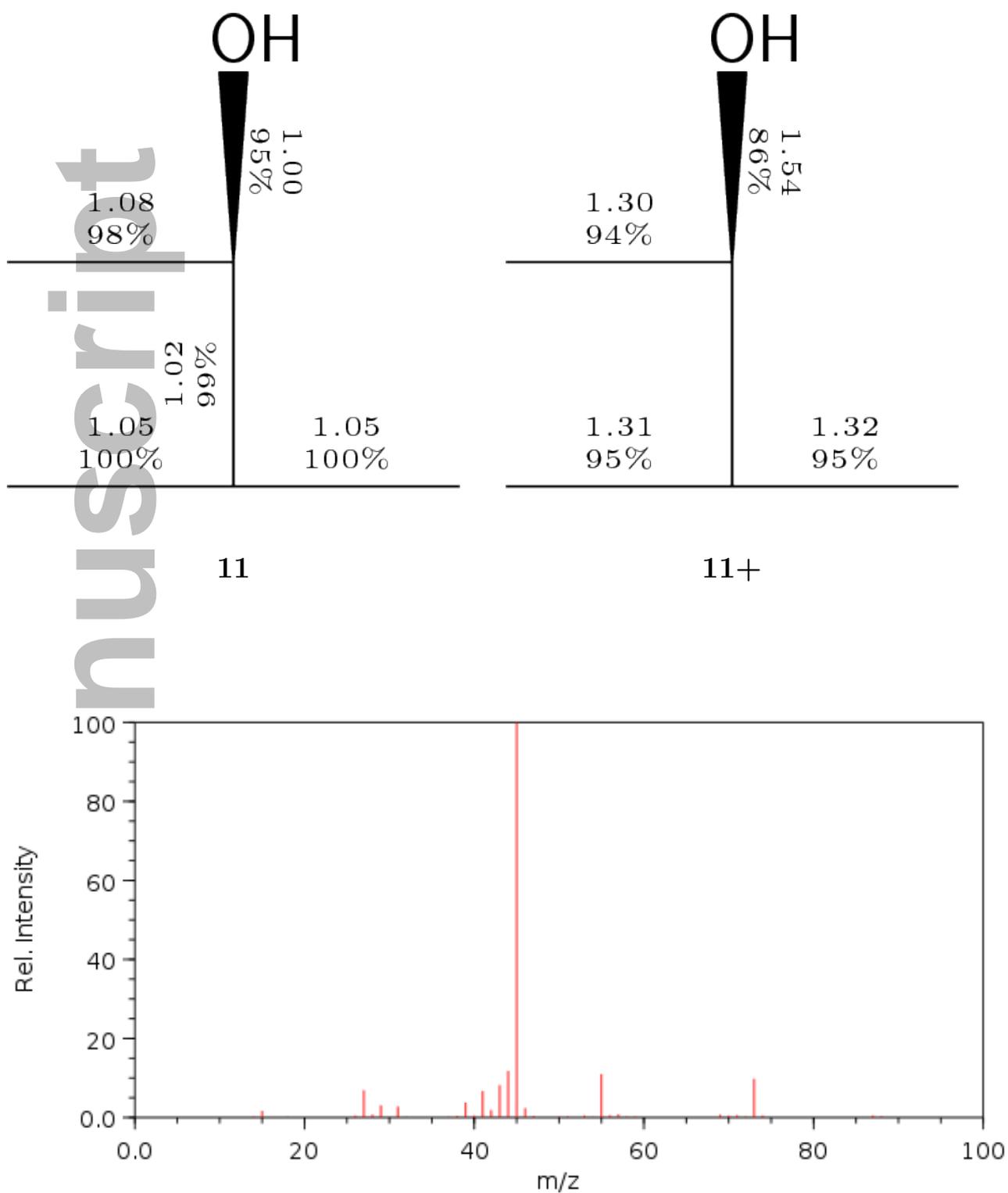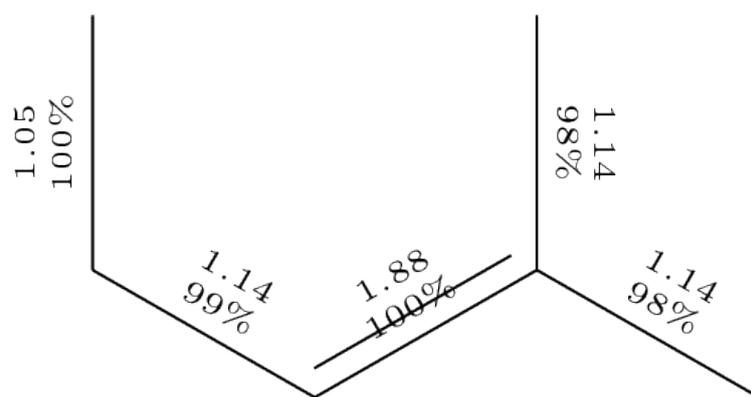
Figure S11: Structure of molecule **11** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**11**) and cation (**11+**).
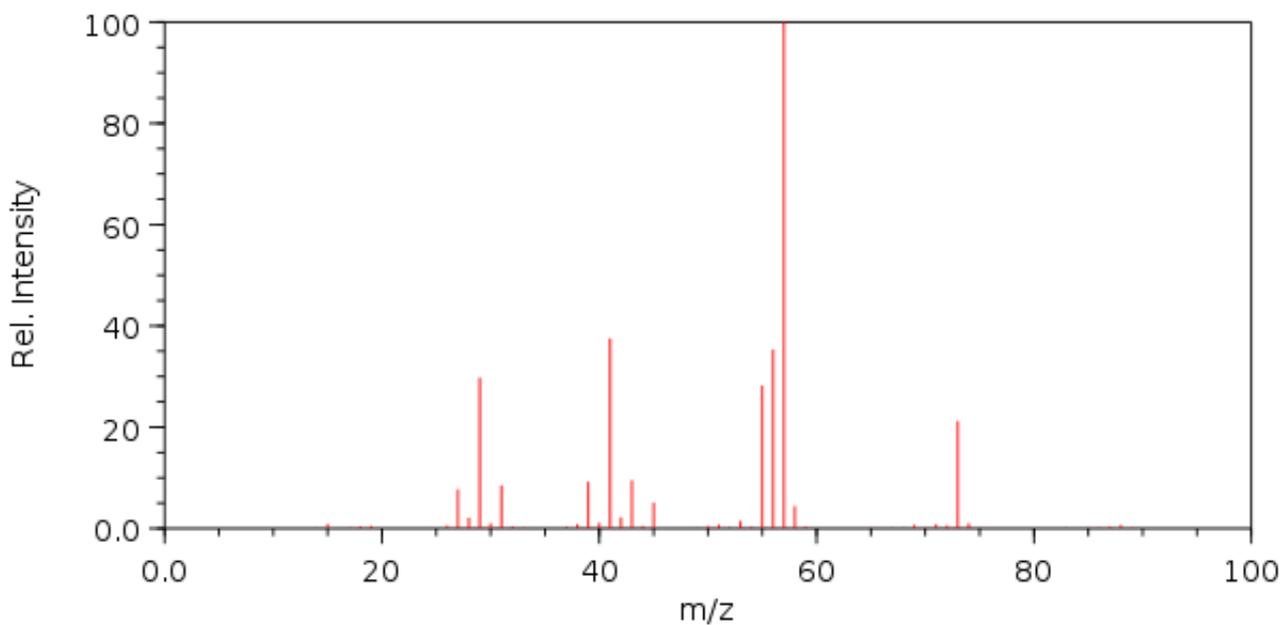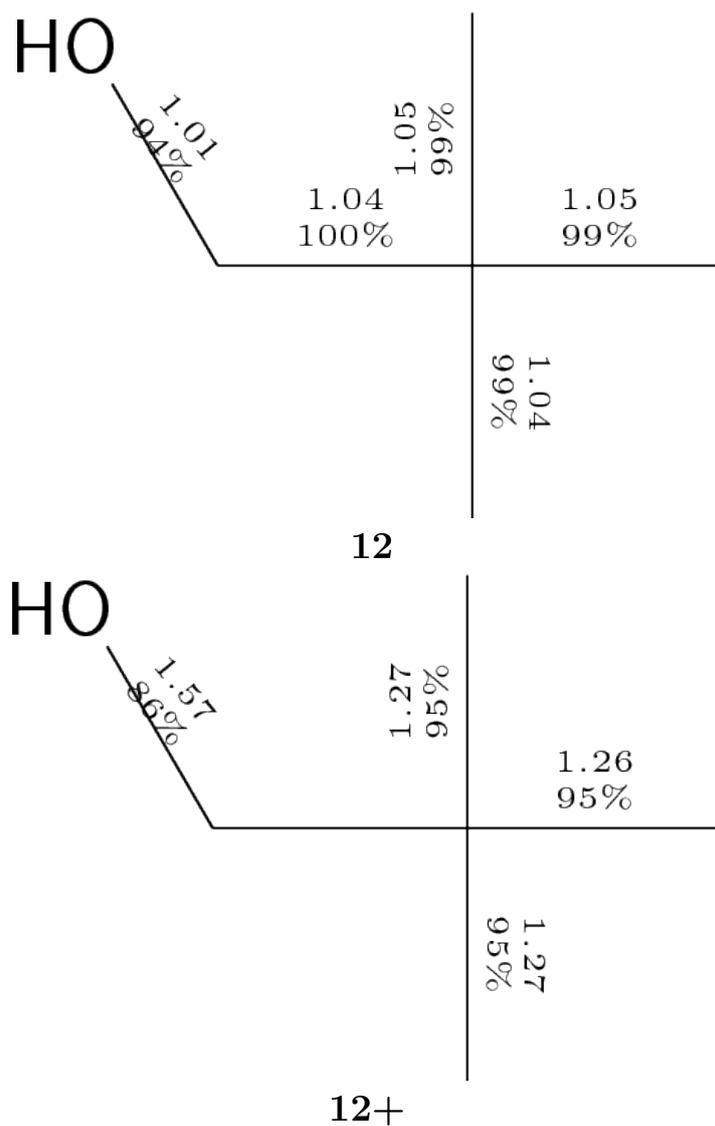
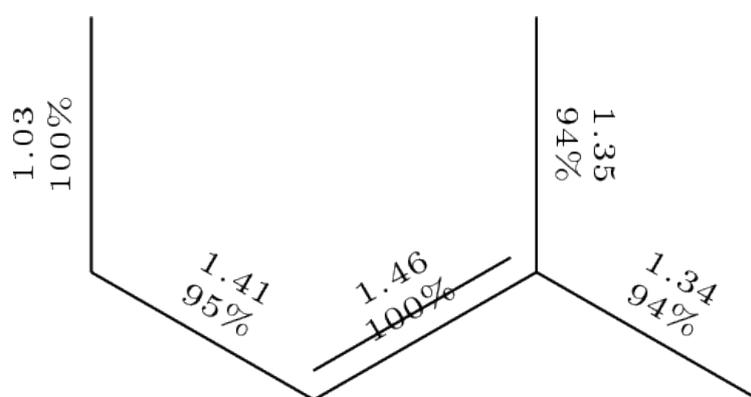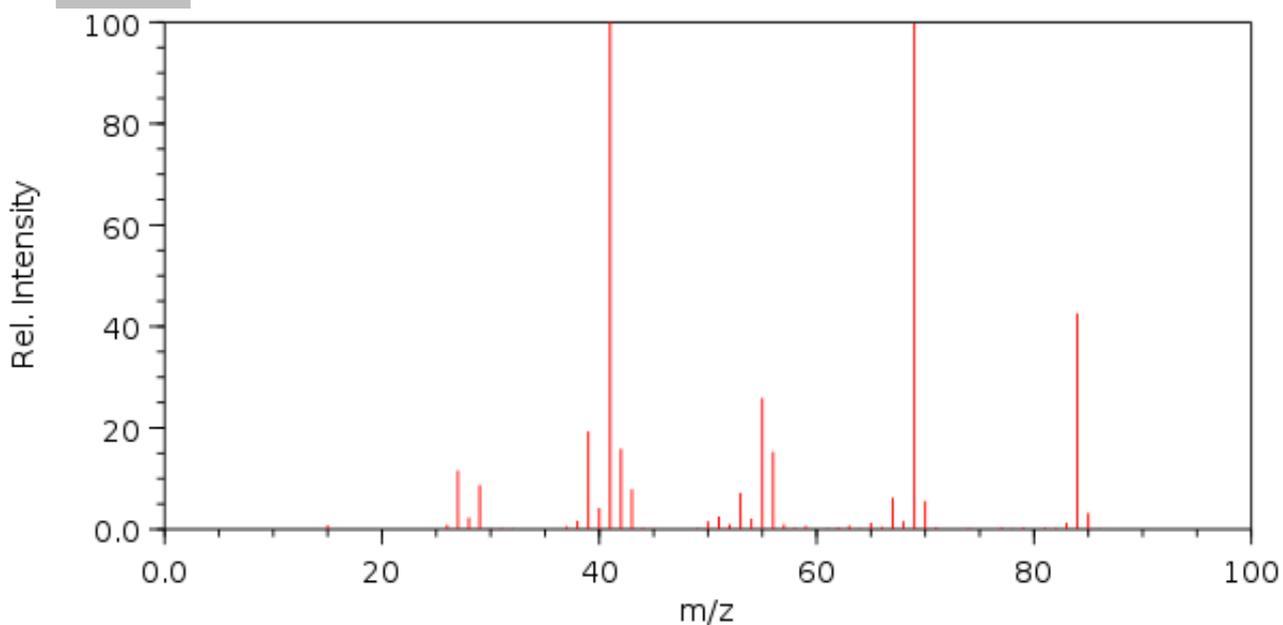Figure S12: Structure of molecule **12** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**12**) and cation (**12**+).

S3.2   Amines

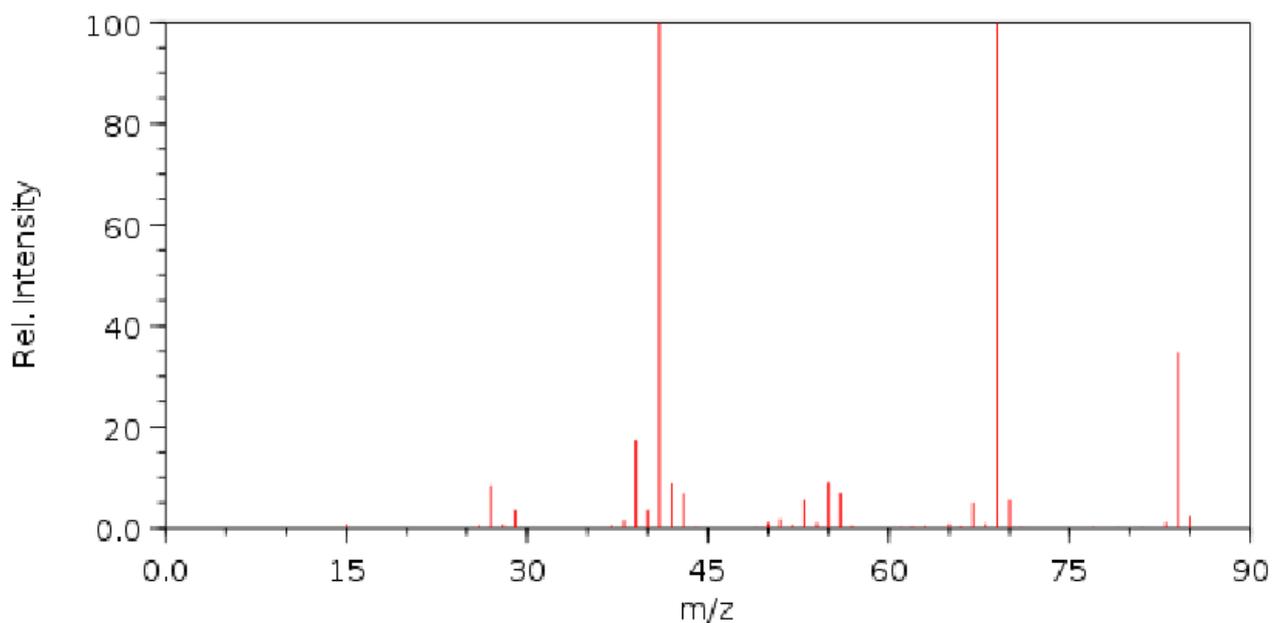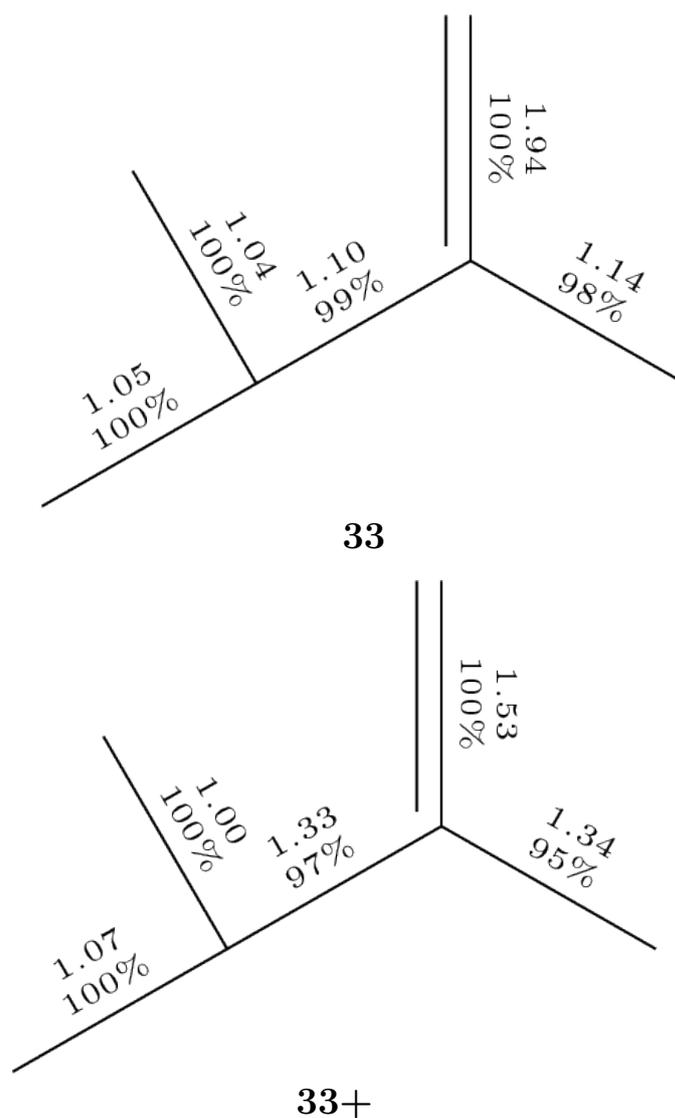### S3.2.1   Isomers of C_3H_9N formula 13-15
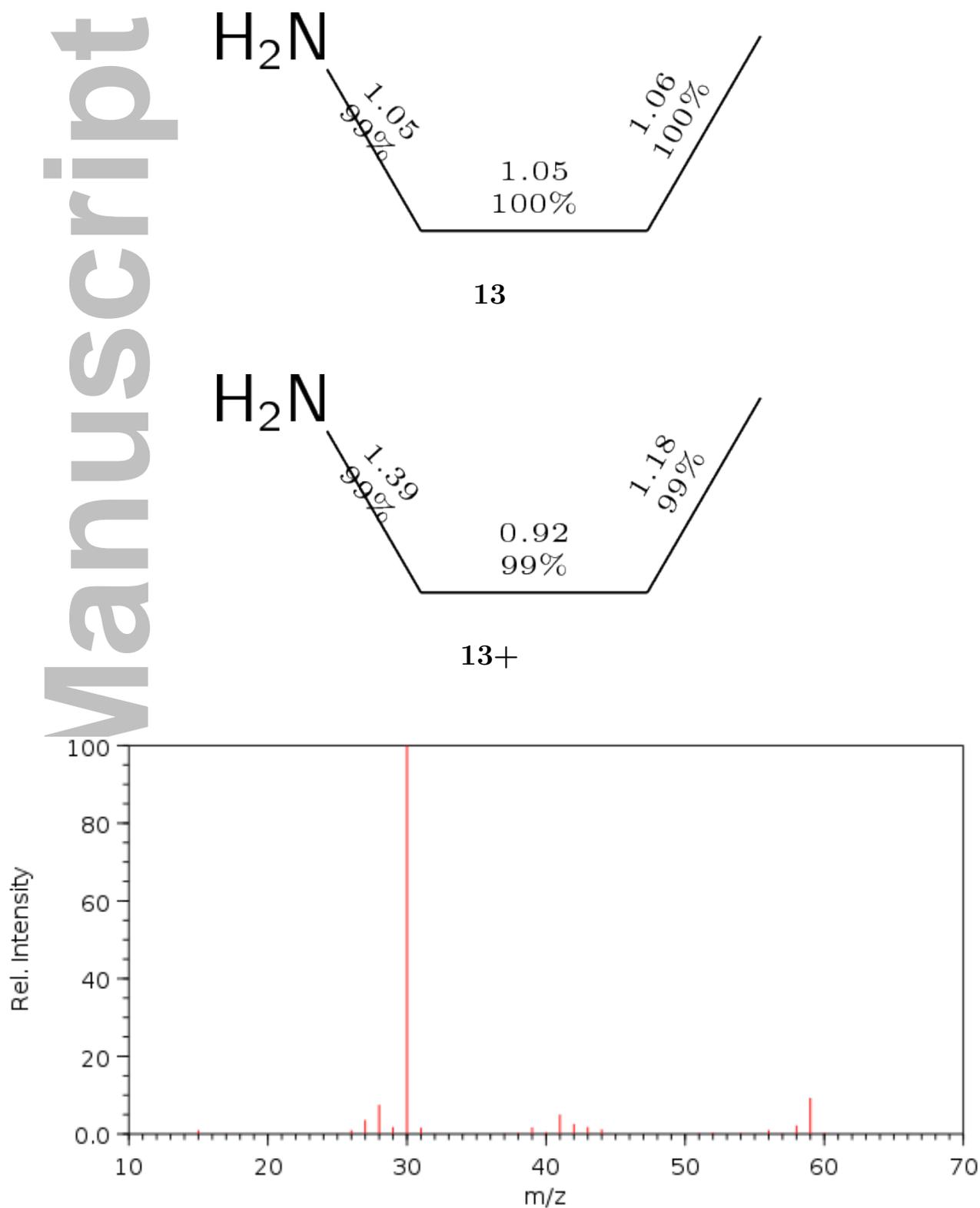
Figure S13: Structure of molecule **13** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**13**) and cation (**13+**).
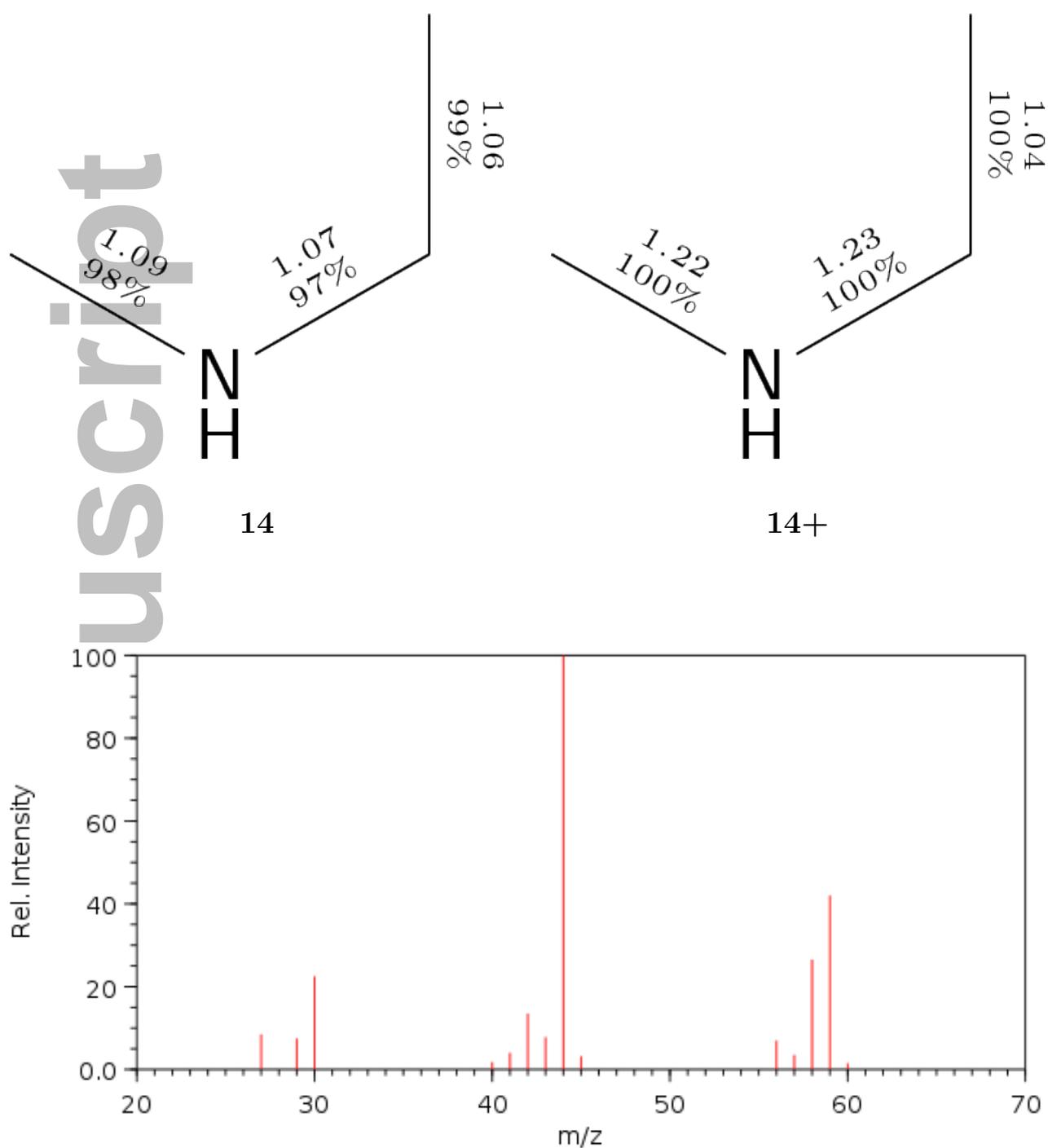
Figure S14: Structure of molecule **14** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**14**) and cation (**14+**).
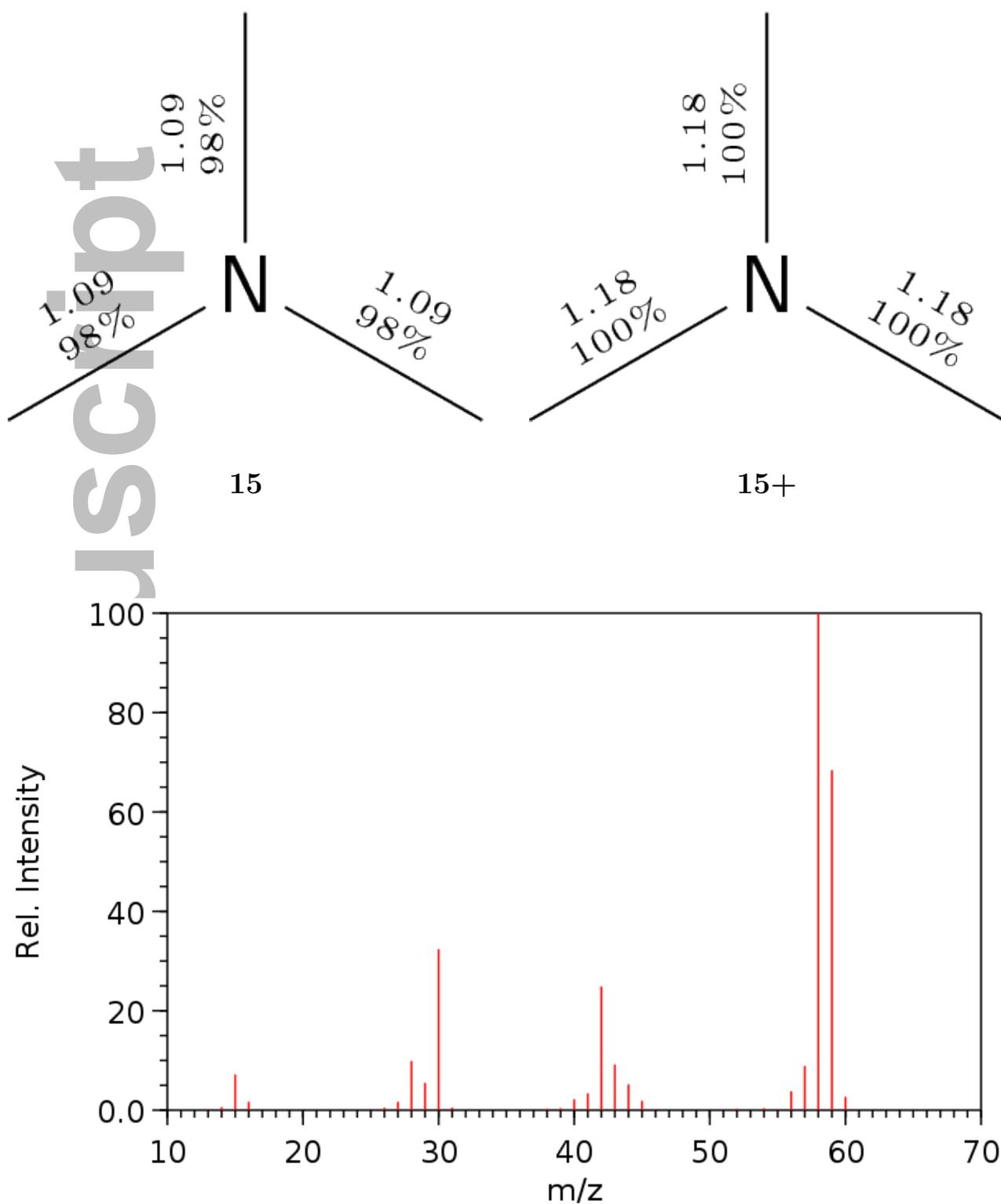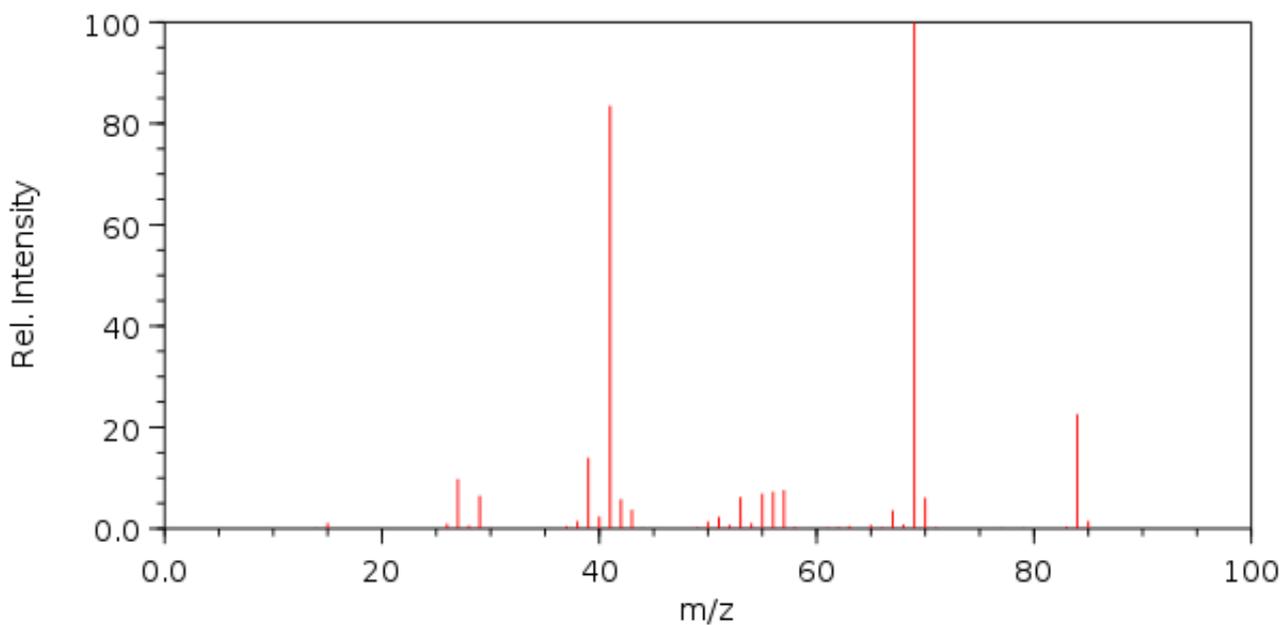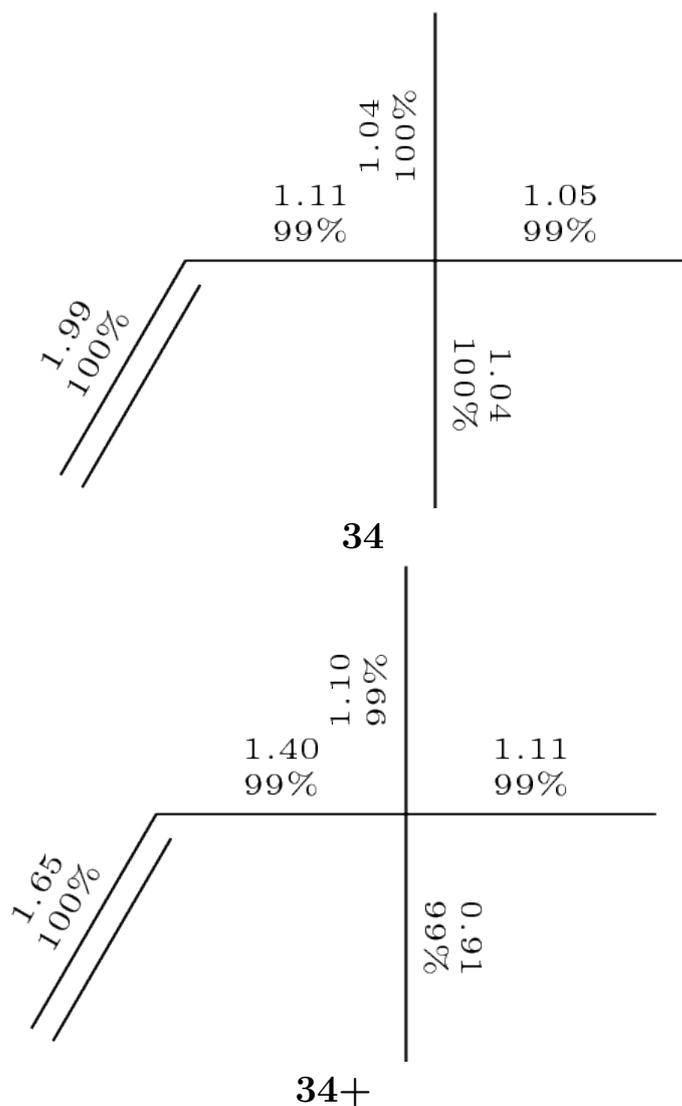
Figure S15: Structure of molecule **15** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**15**) and cation (**15+**).
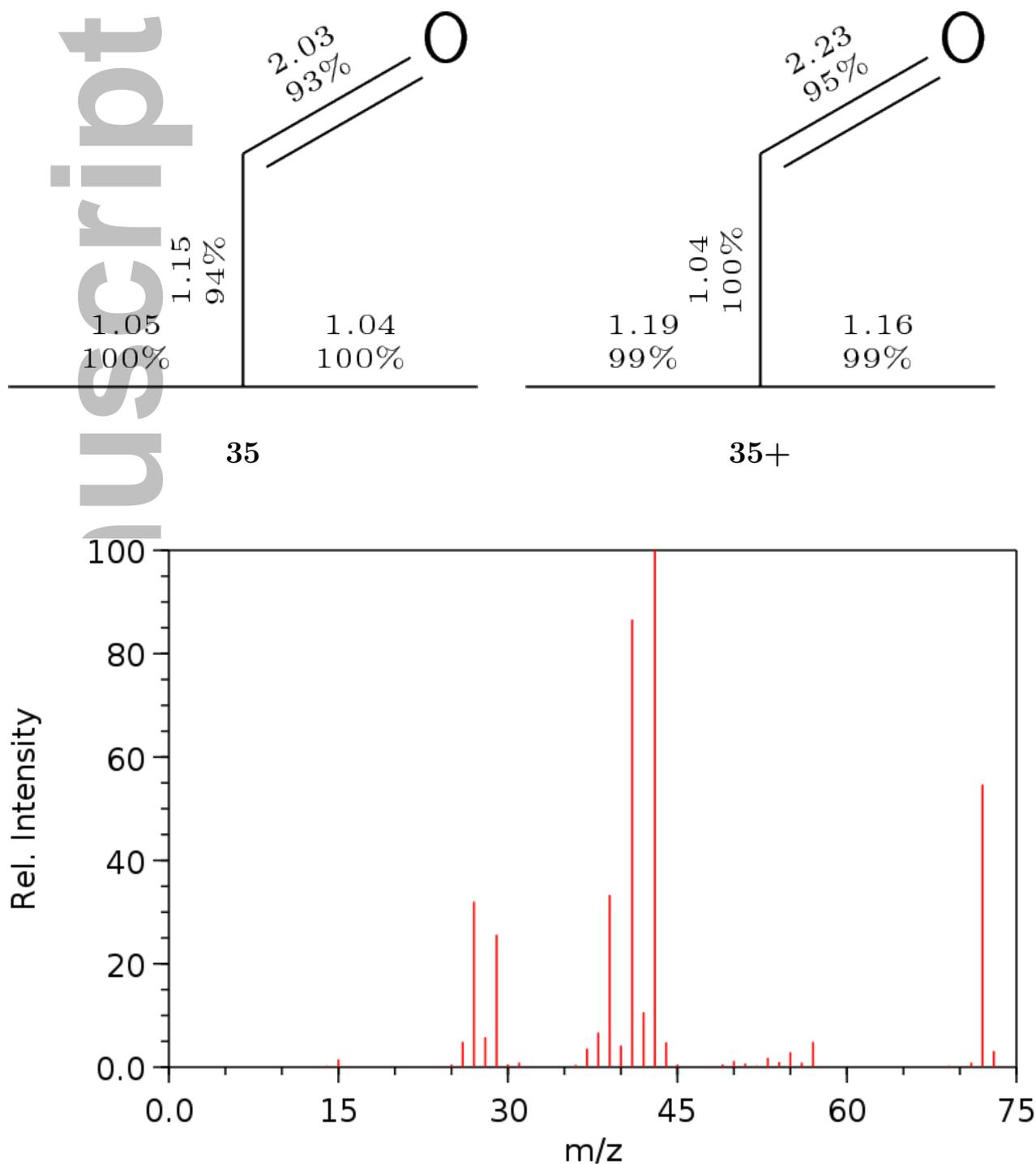
S3.2.2 Isomers of $C_4H_{11}N$ formula 16-22
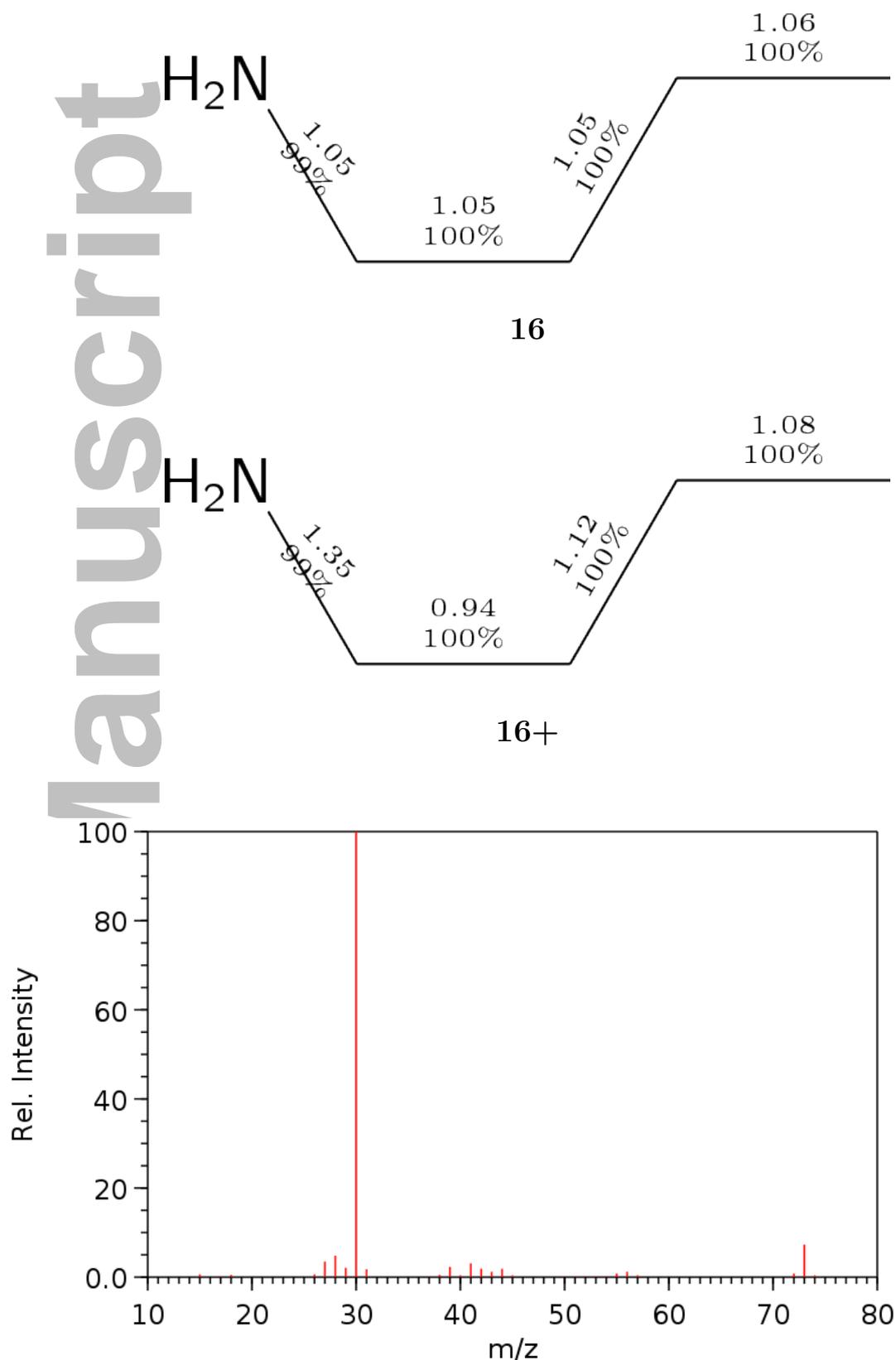
Figure S16: Structure of molecule **16** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**16**) and cation (**16+**).

H₂N

1.04
99%

1.04
100%

1.06
100%

1.06
99%

17

H₂N

1.38
99%

0.88
99%

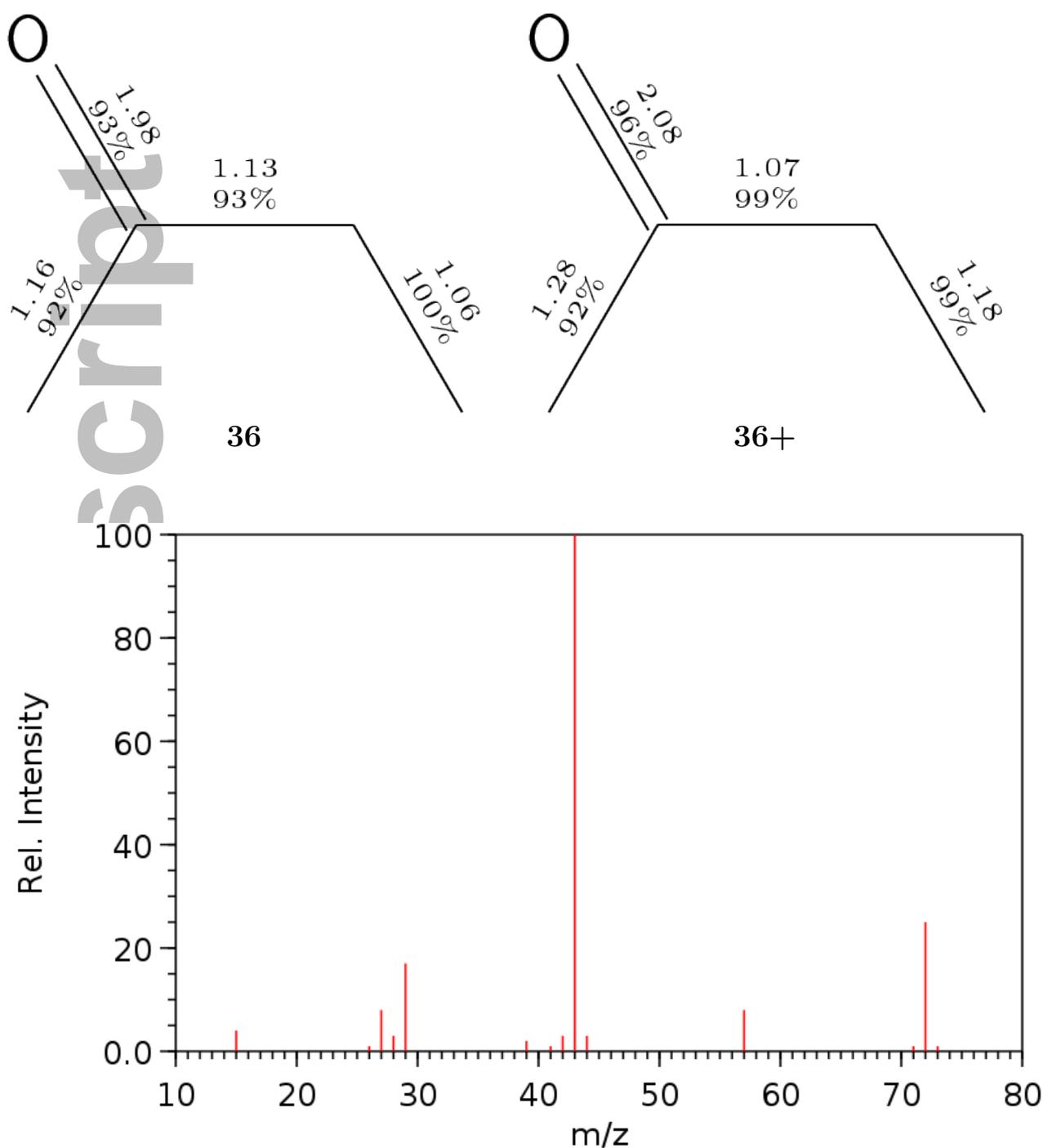1.19
99%

1.12
98%

17+
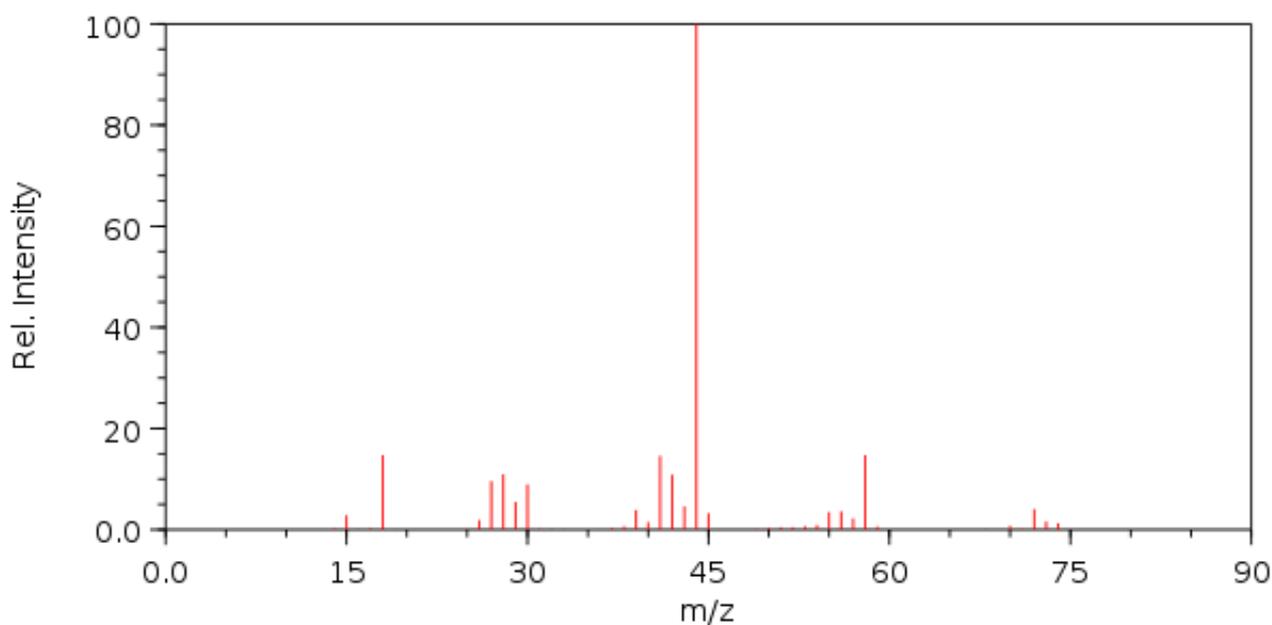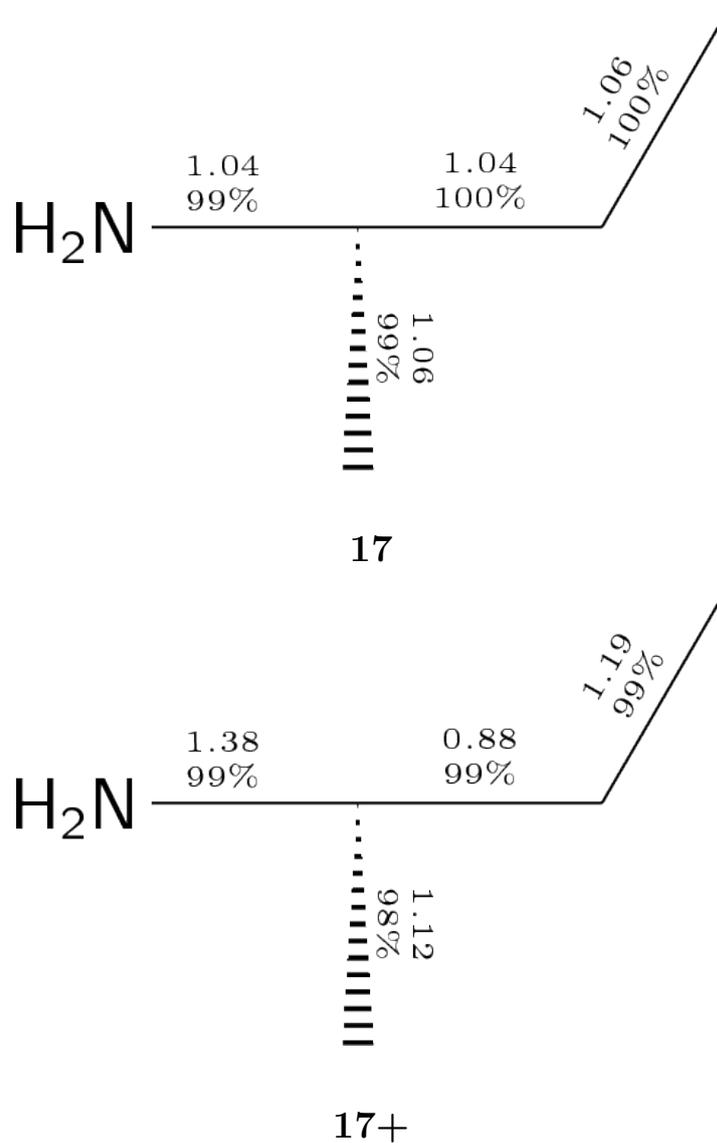
Figure S17: Structure of molecule **17** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**17**) and cation (**17+**).
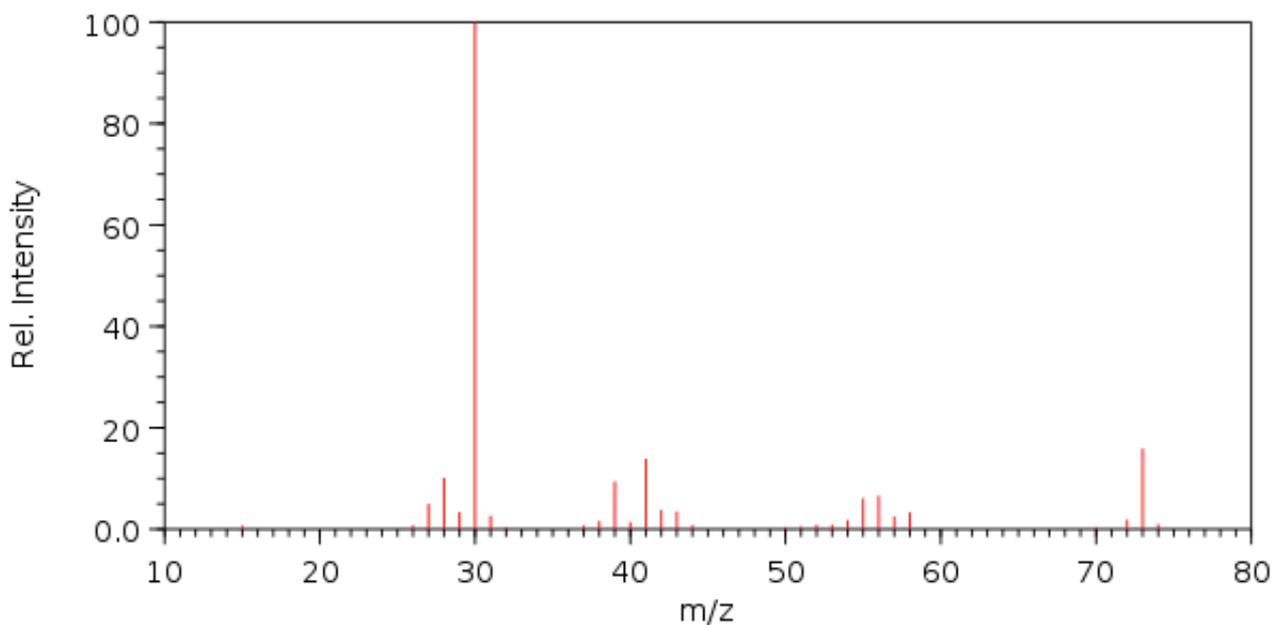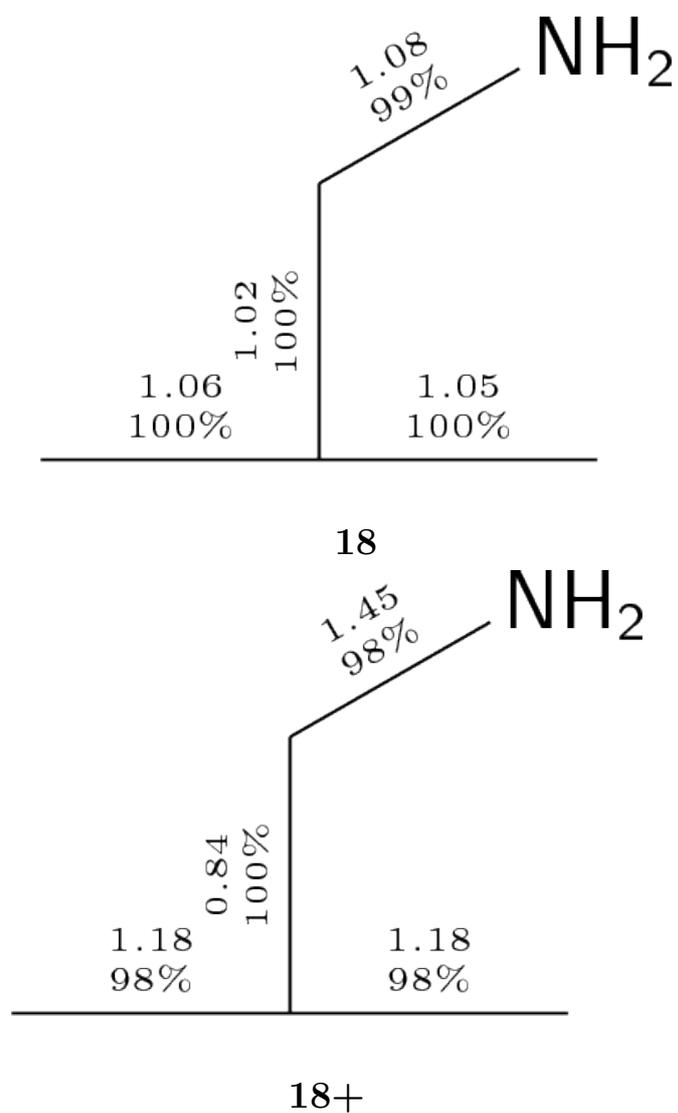
Figure S18: Structure of molecule **18** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**18**) and cation (**18+**).
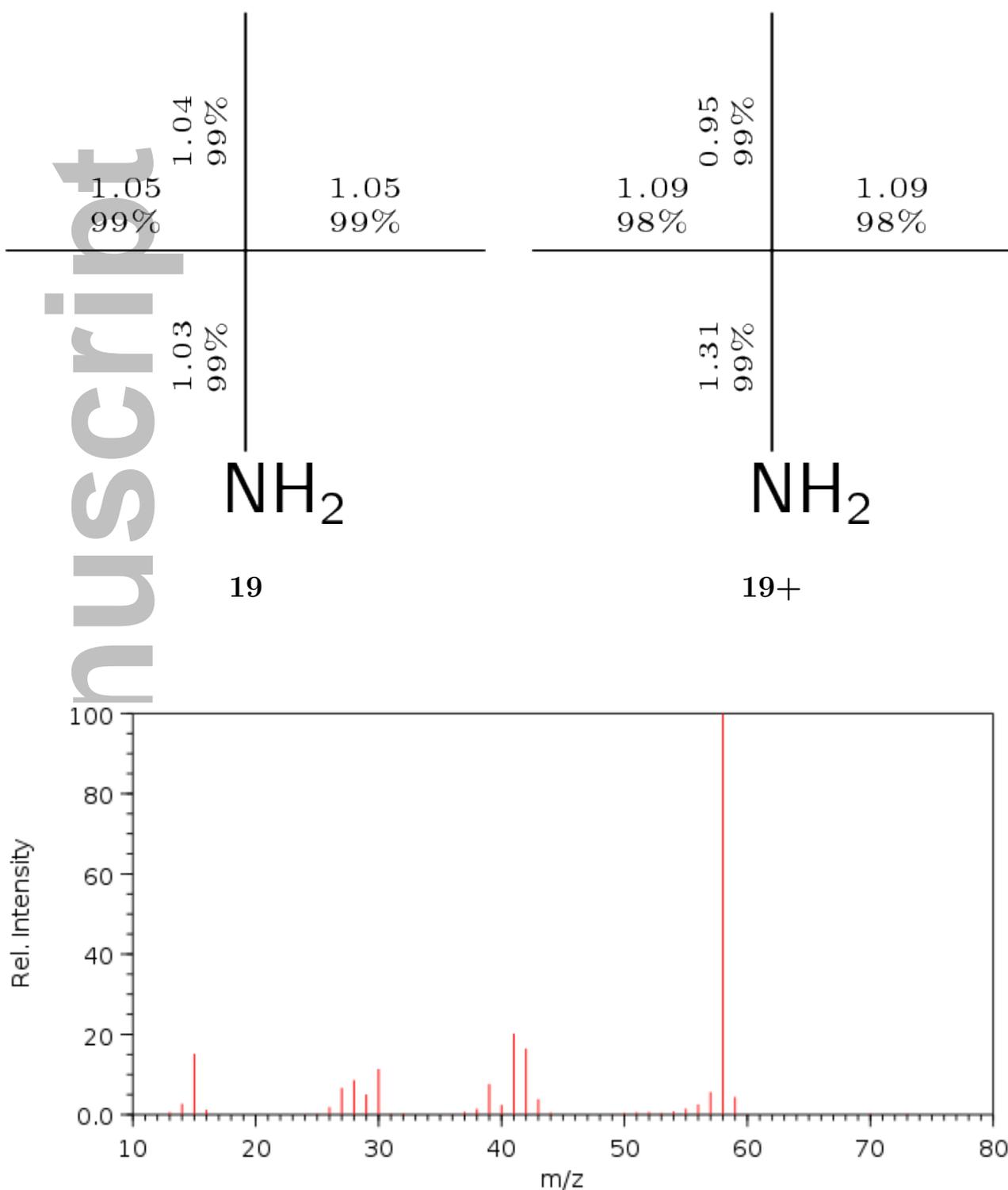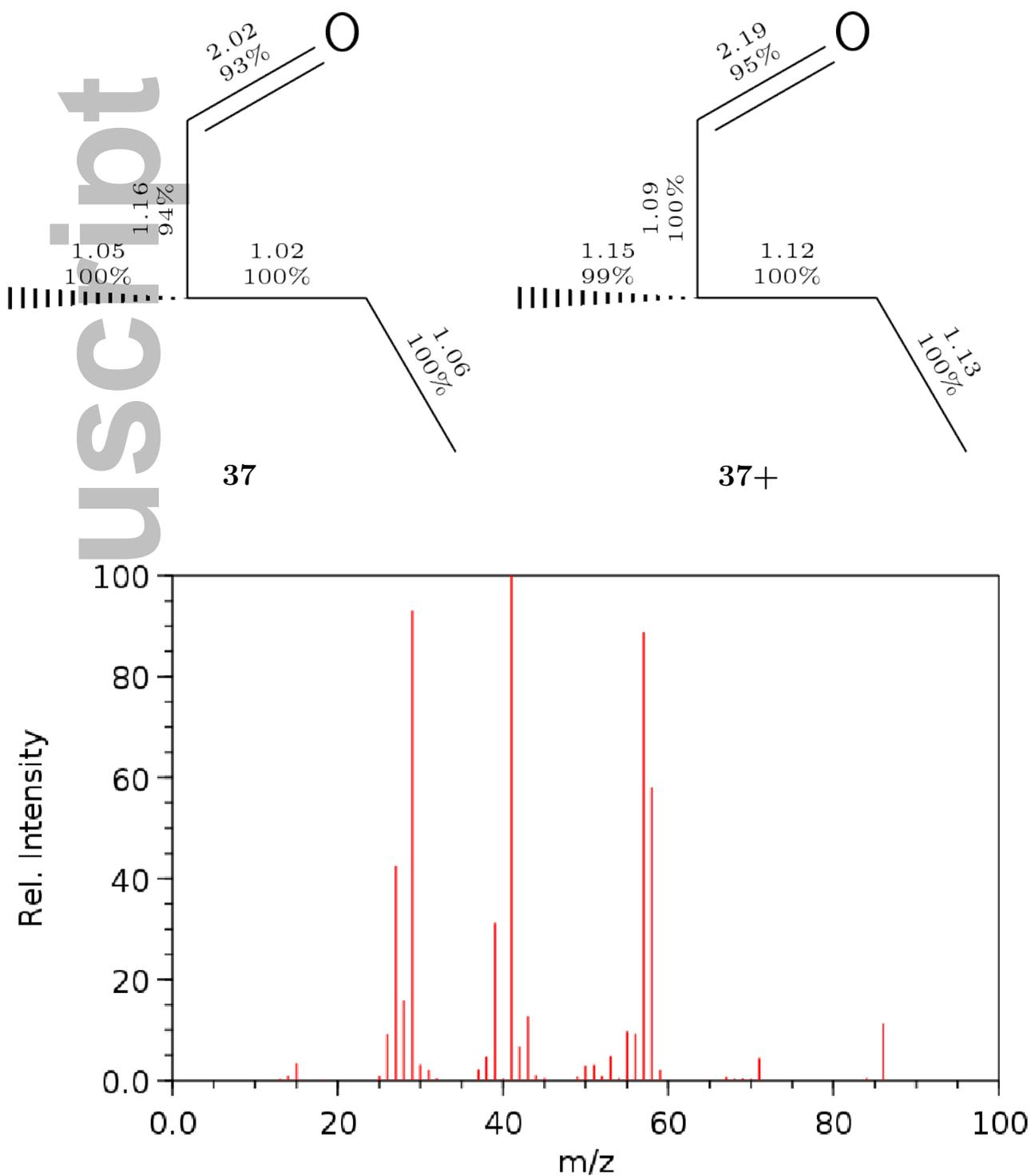
Figure S19: Structure of molecule **19** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**19**) and cation (**19+**).
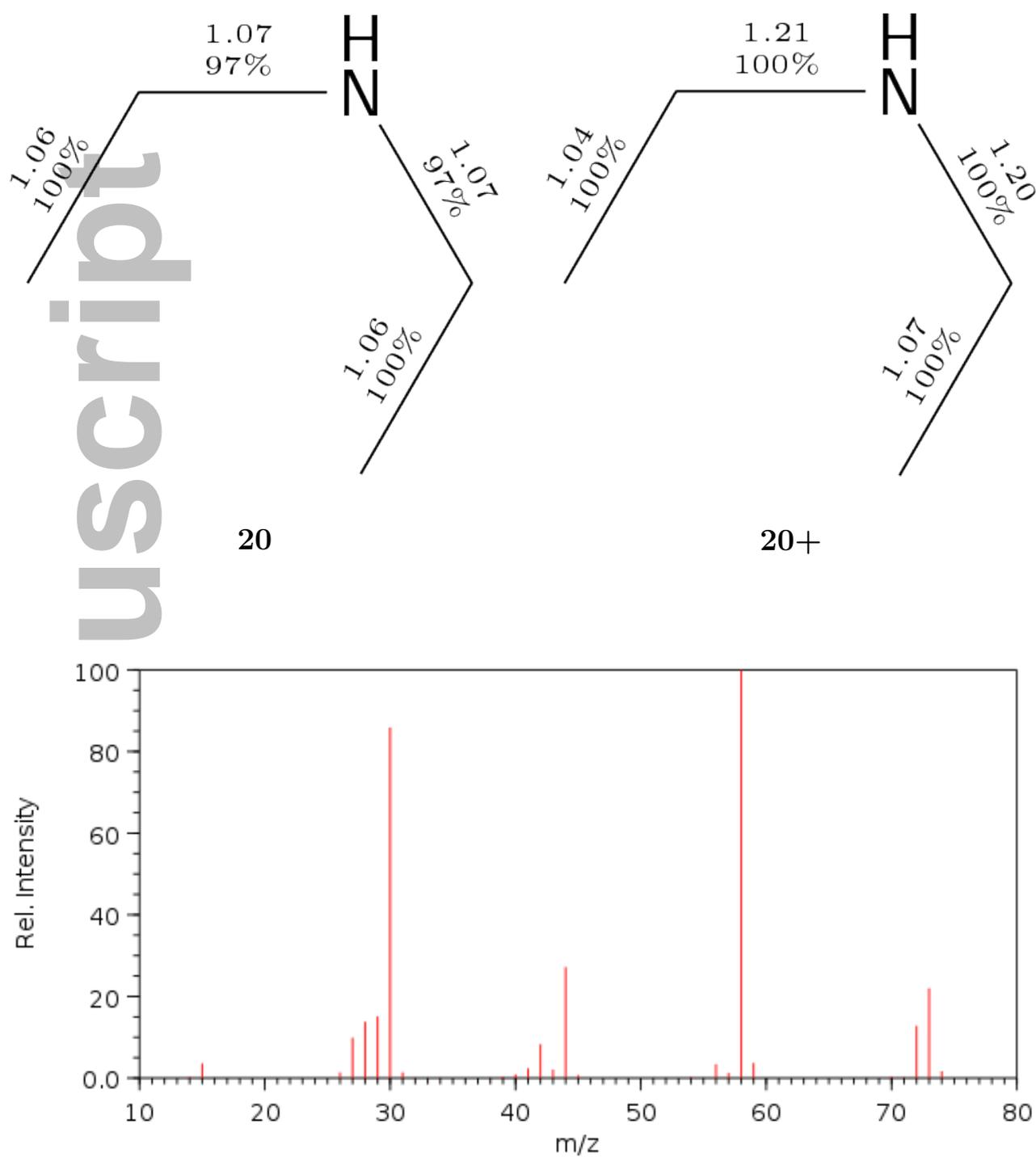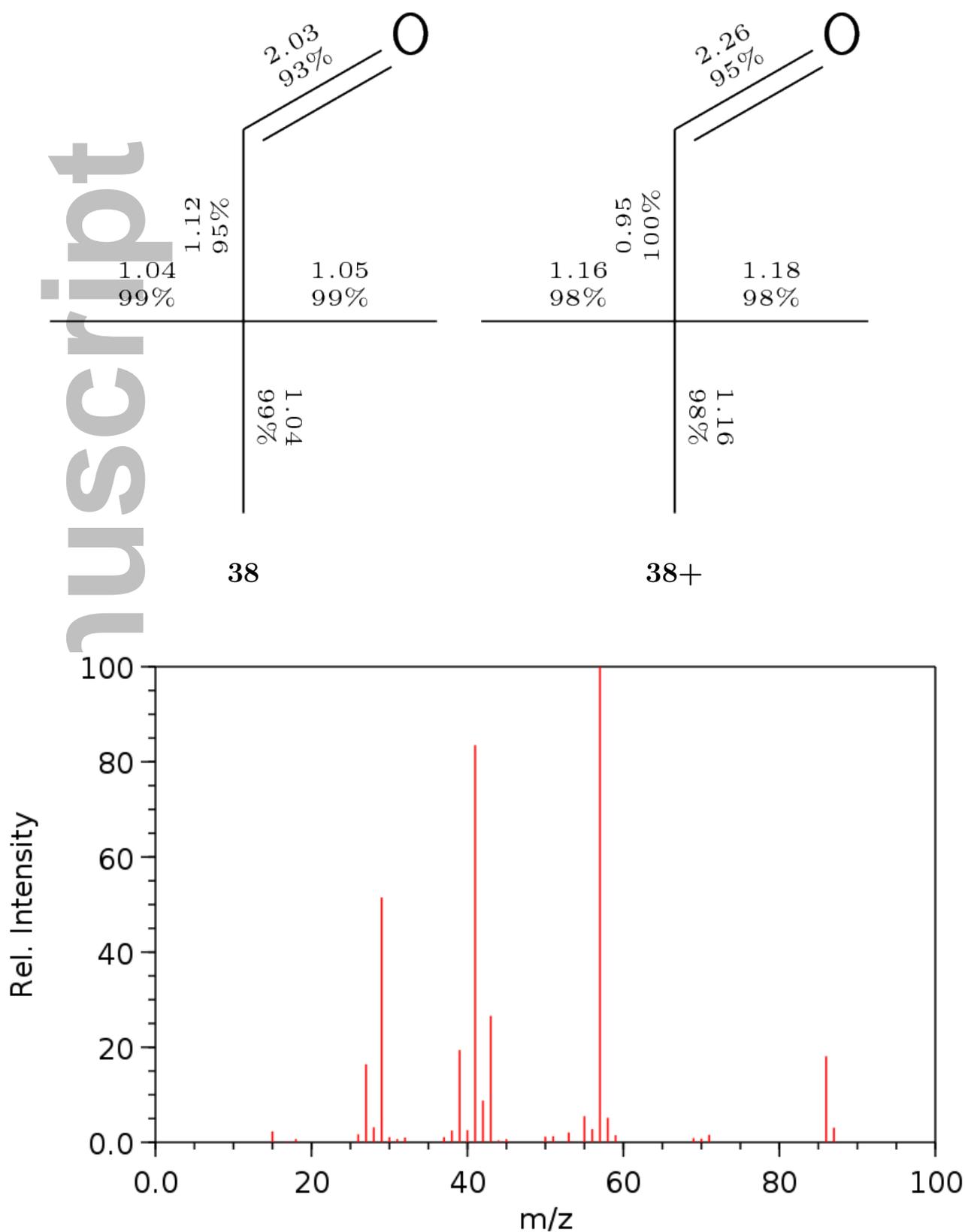
Figure S20: Structure of molecule **20** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**20**) and cation (**20+**).
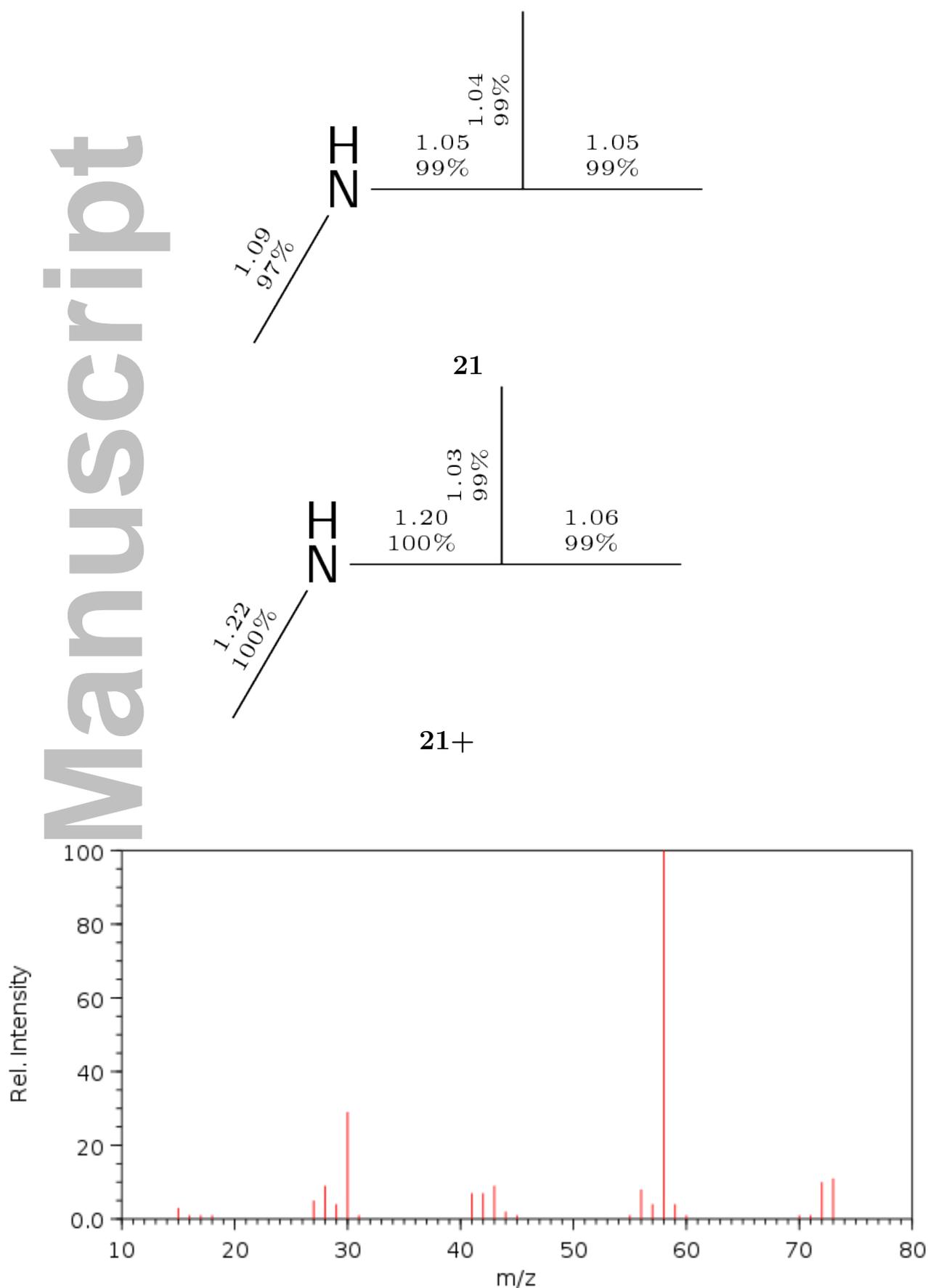
Figure S21: Structure of molecule **21** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**21**) and cation (**21+**).
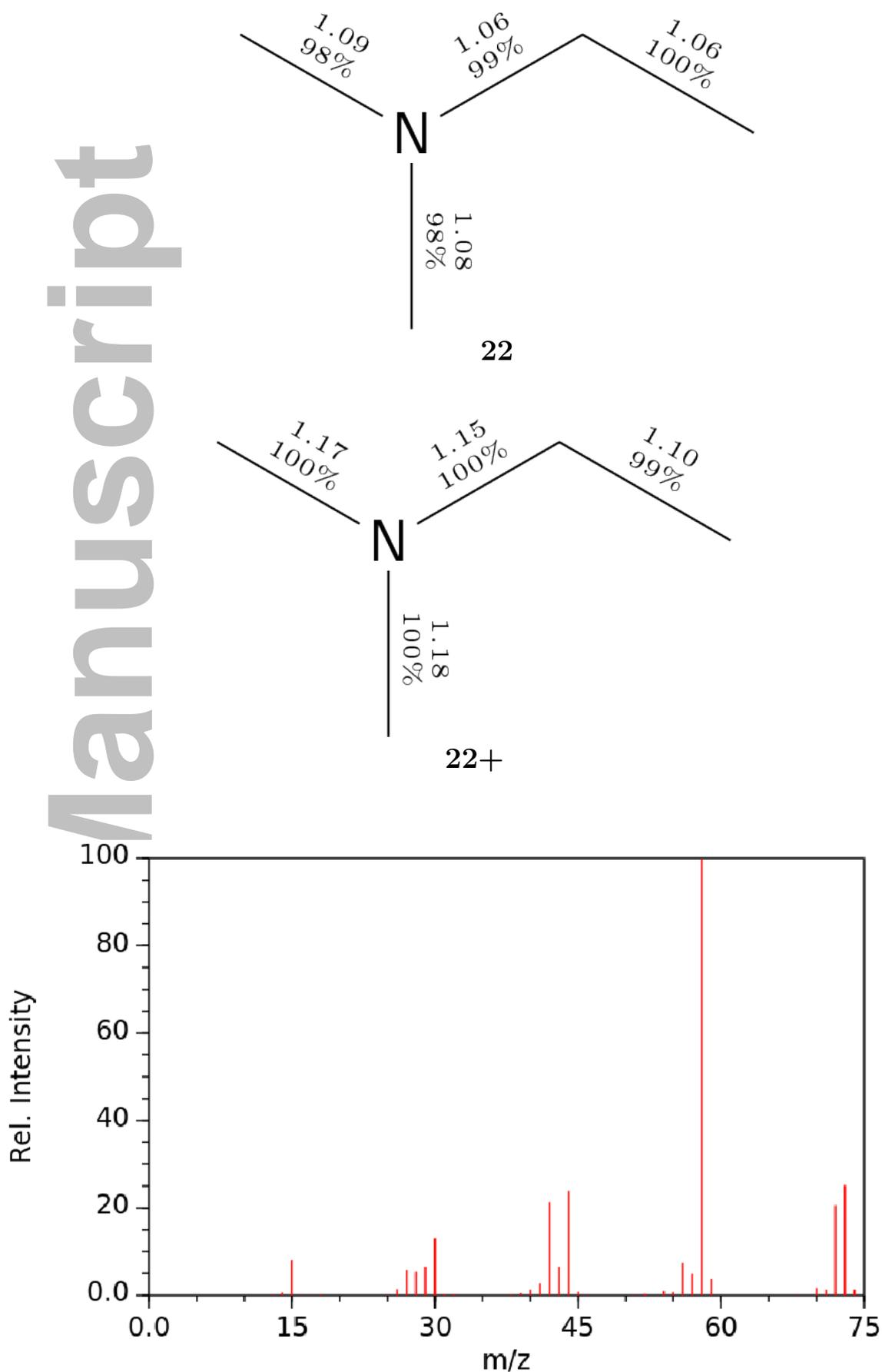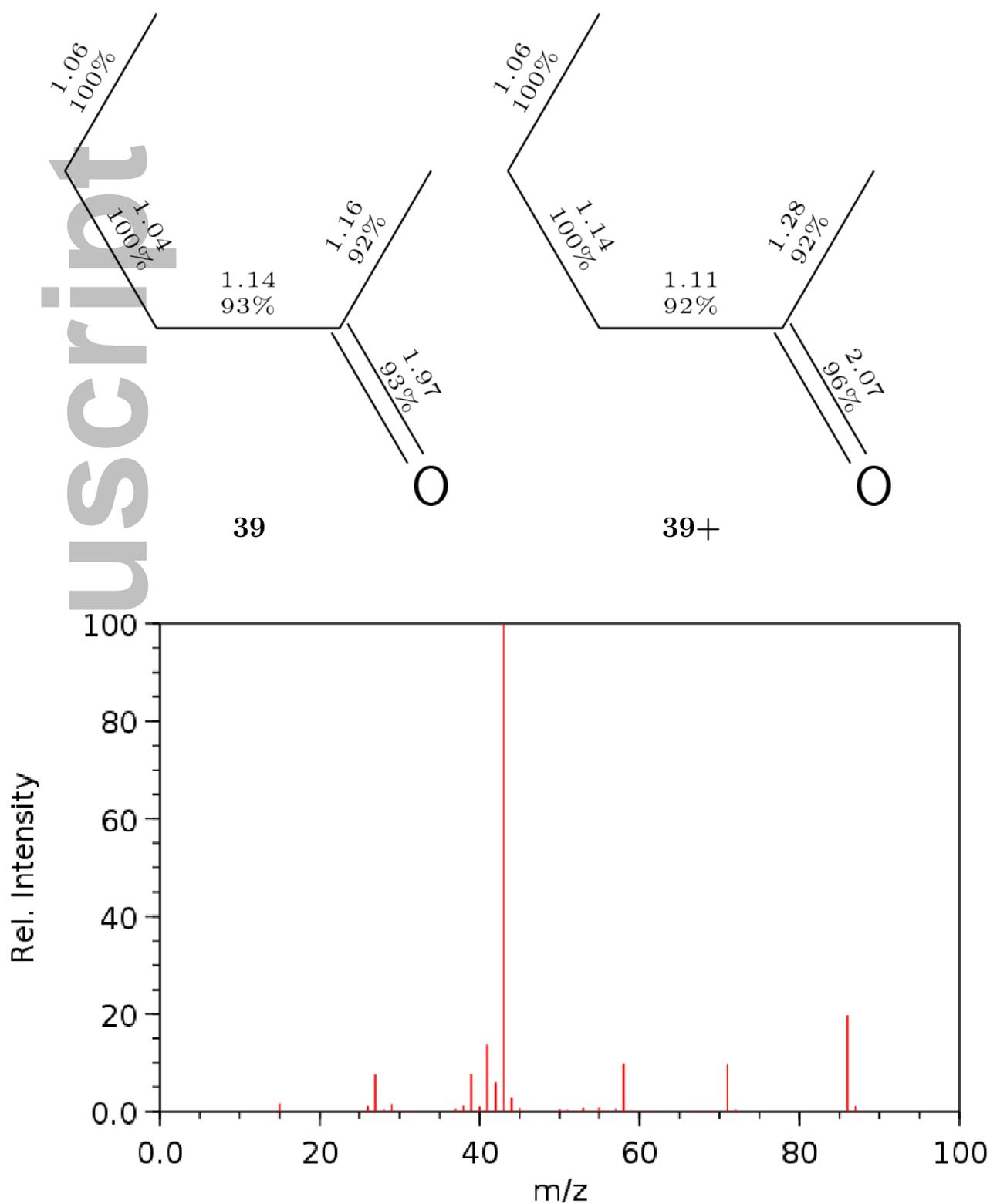
Figure S22: Structure of molecule **22** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**22**) and cation (**22+**).

S3.3   Alkenes

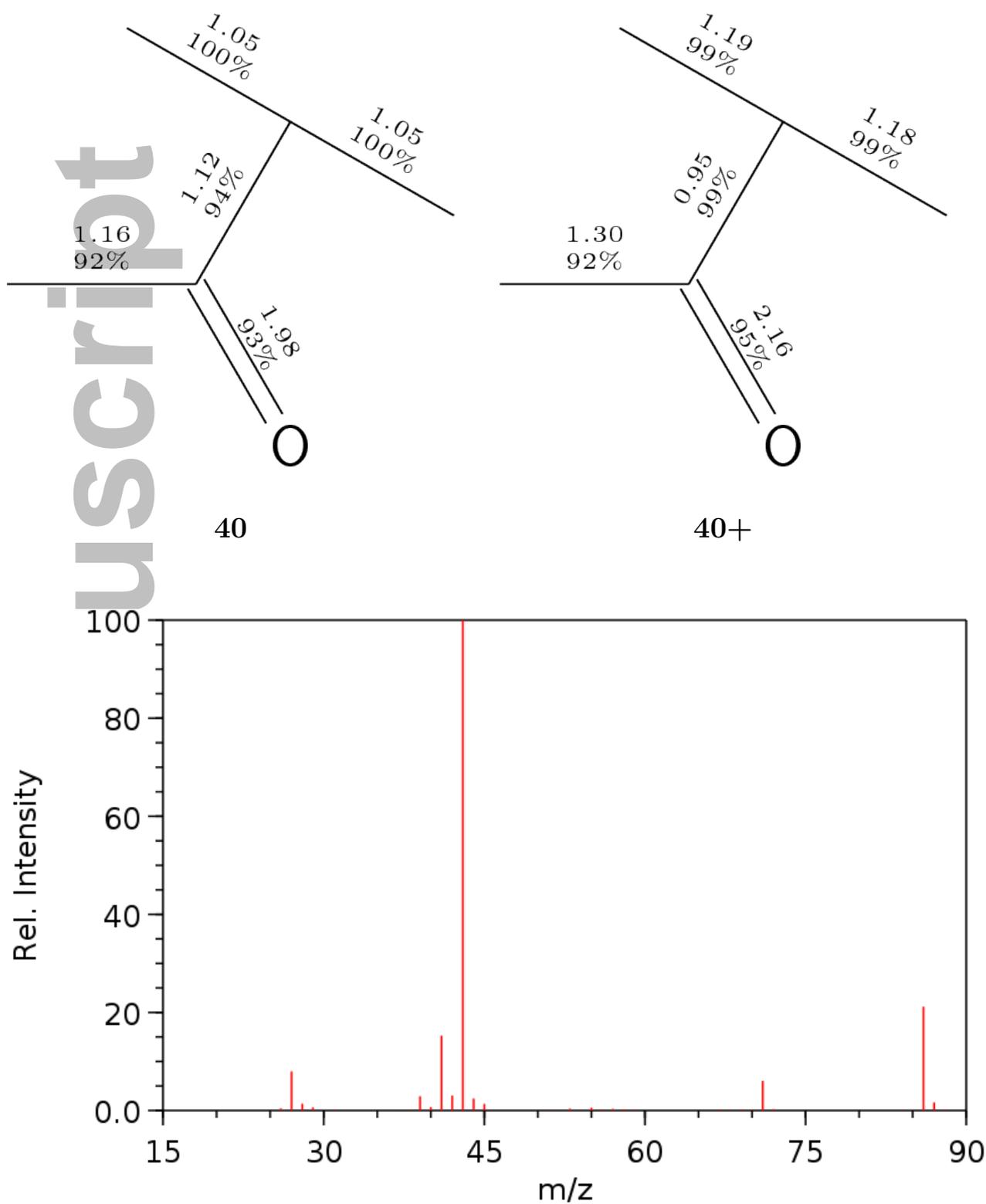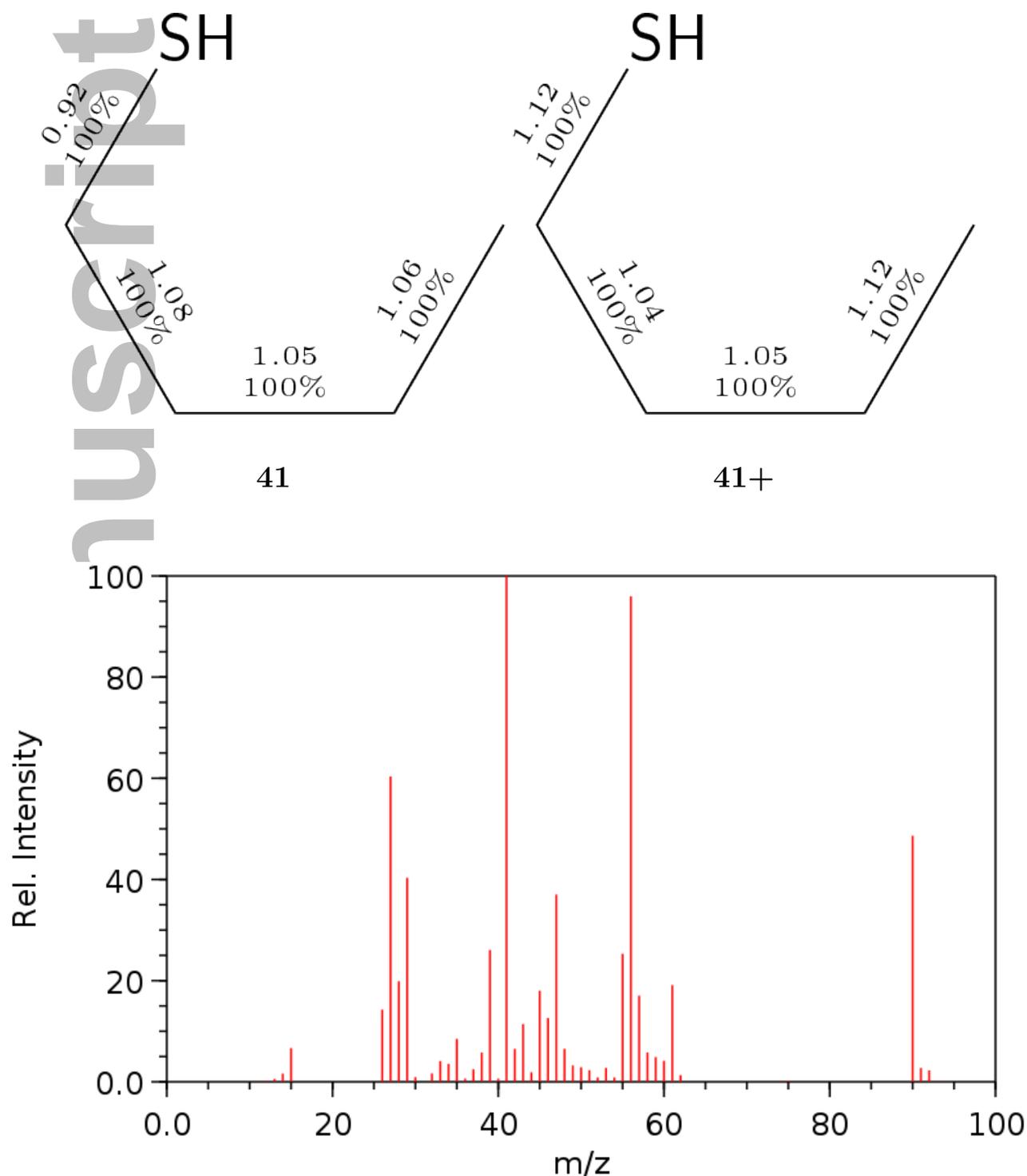### S3.3.1   Isomers of C_5H_{10} formula 23-25

Figure S23: Structure of molecule **23** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**23**) and cation (**23+**).

Figure S24: Structure of molecule **24** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**24**) and cation (**24+**).
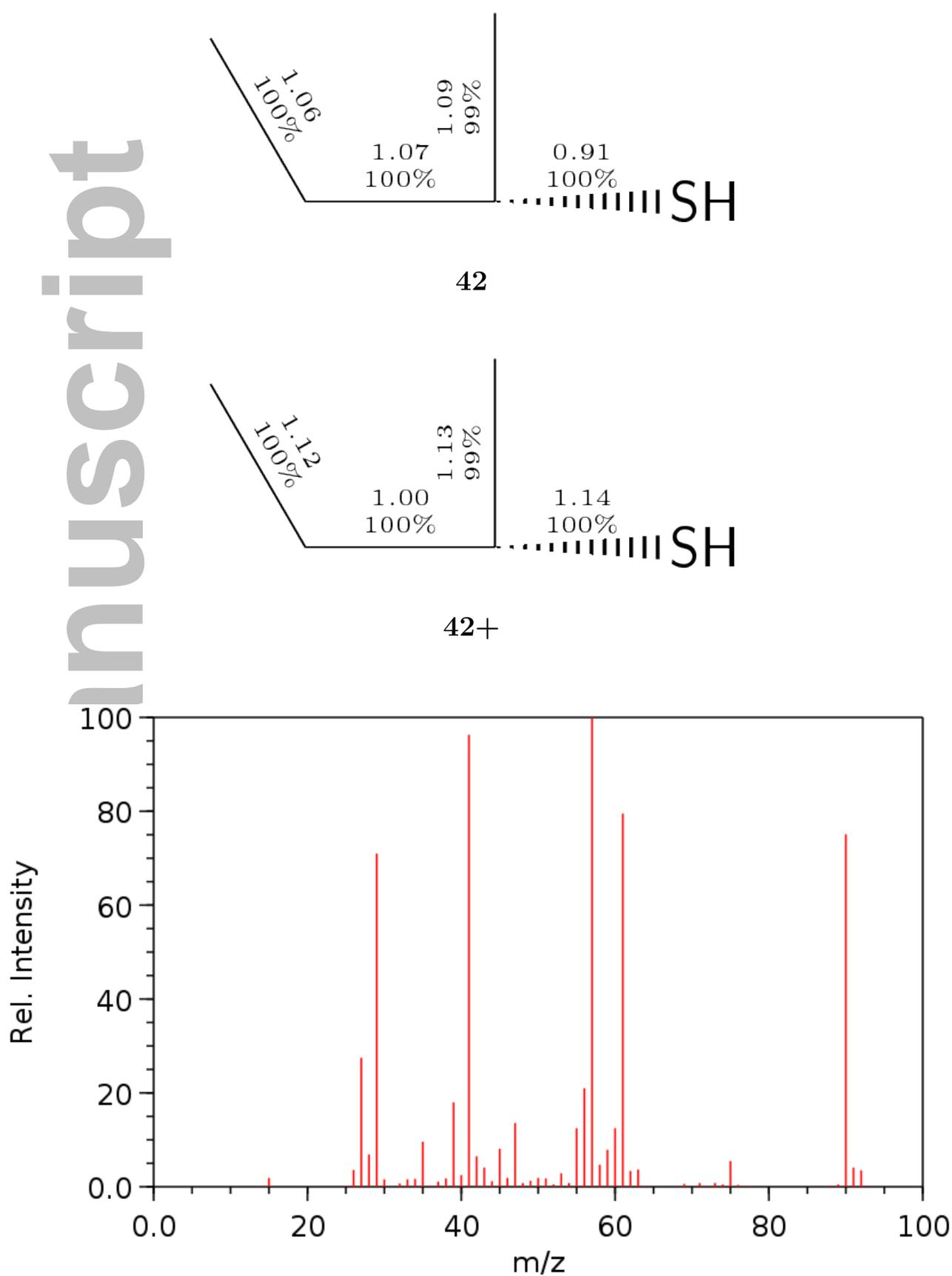
Figure S25: Structure of molecule **25** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**25**) and cation (**25+**).

S3.3.2   Isomers of C_6H_{12} formula 26-34

Figure S26: Structure of molecule **26** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**26**) and cation (**26**+).
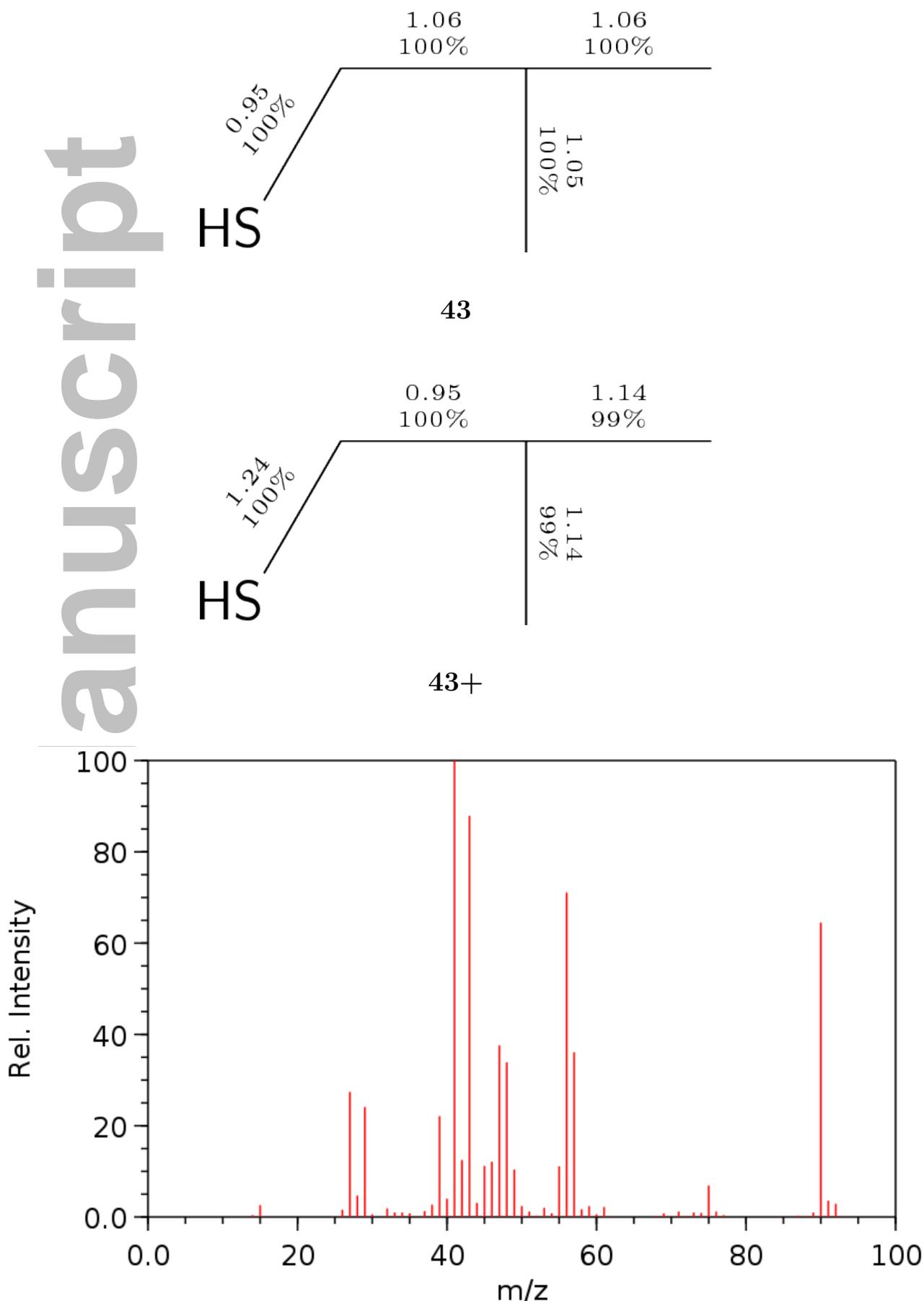
Figure S27: Structure of molecule **27** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**27**) and cation (**27+**).
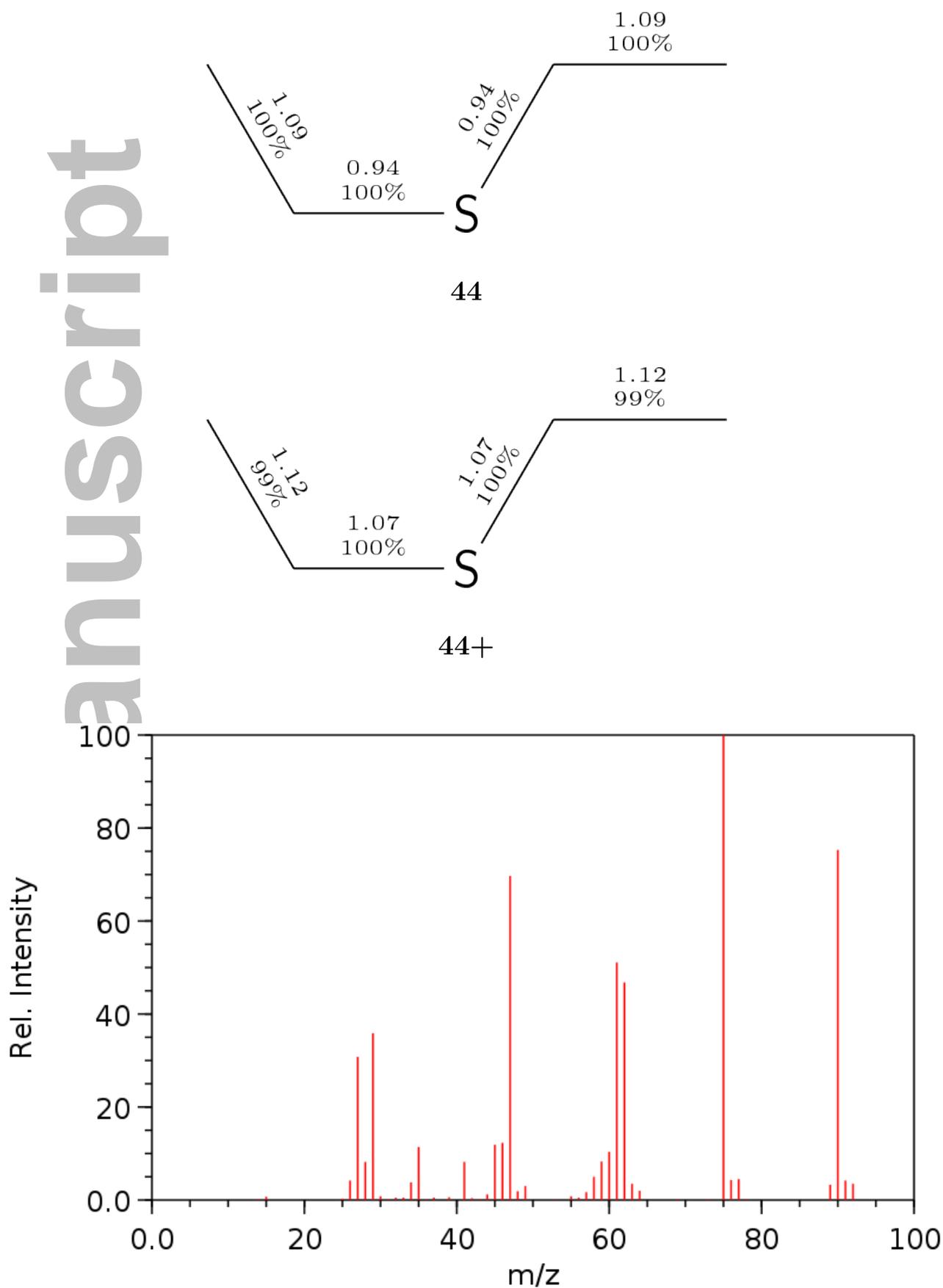
Figure S28: Structure of molecule **28** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**28**) and cation (**28+**).
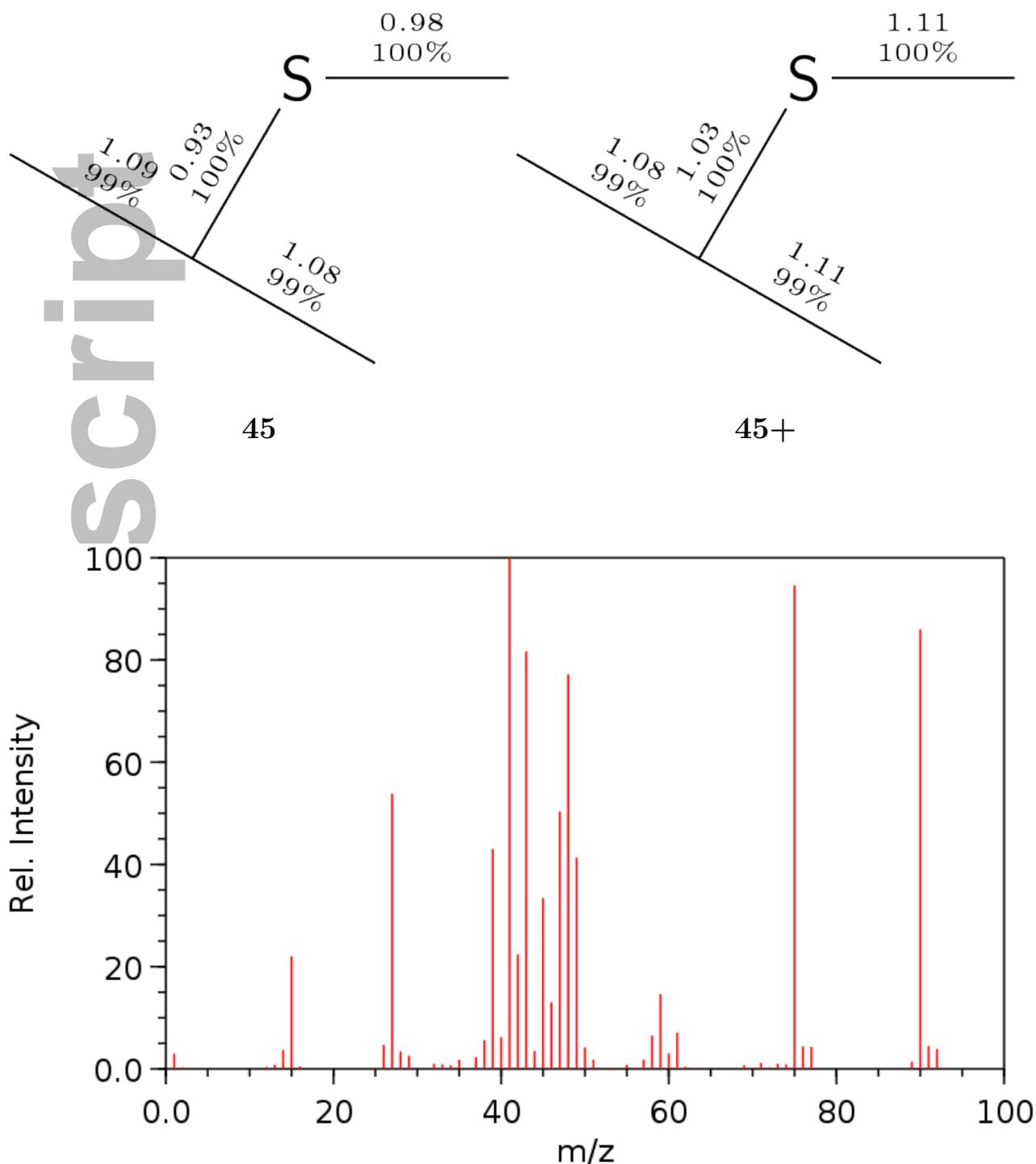
29

29+

Figure S29: Structure of molecule **29** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**29**) and cation (**29+**).

Figure S30: Structure of molecule **30** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**30**) and cation (**30+**).
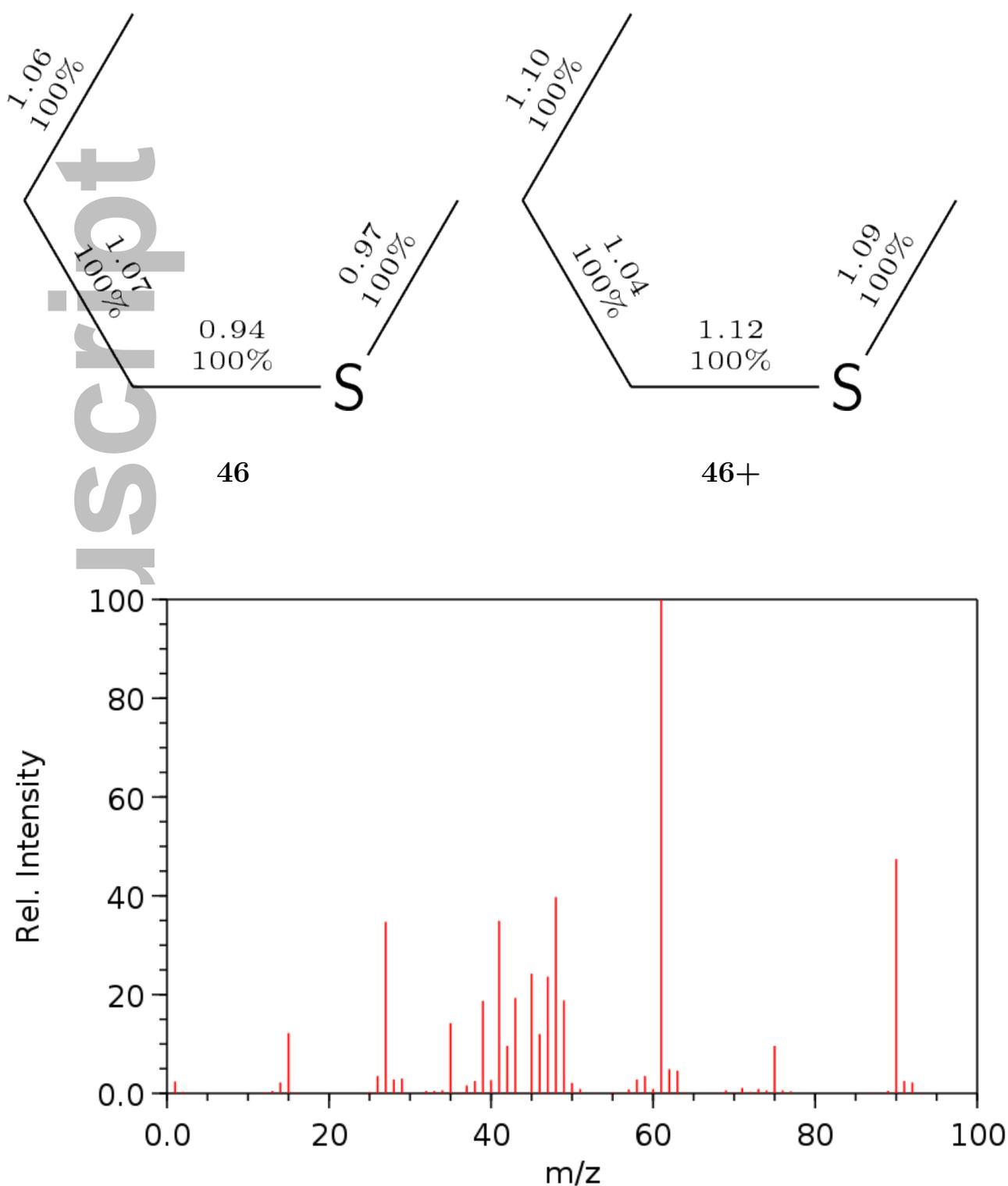
31

31+

Figure S31: Structure of molecule **31** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**31**) and cation (**31+**).
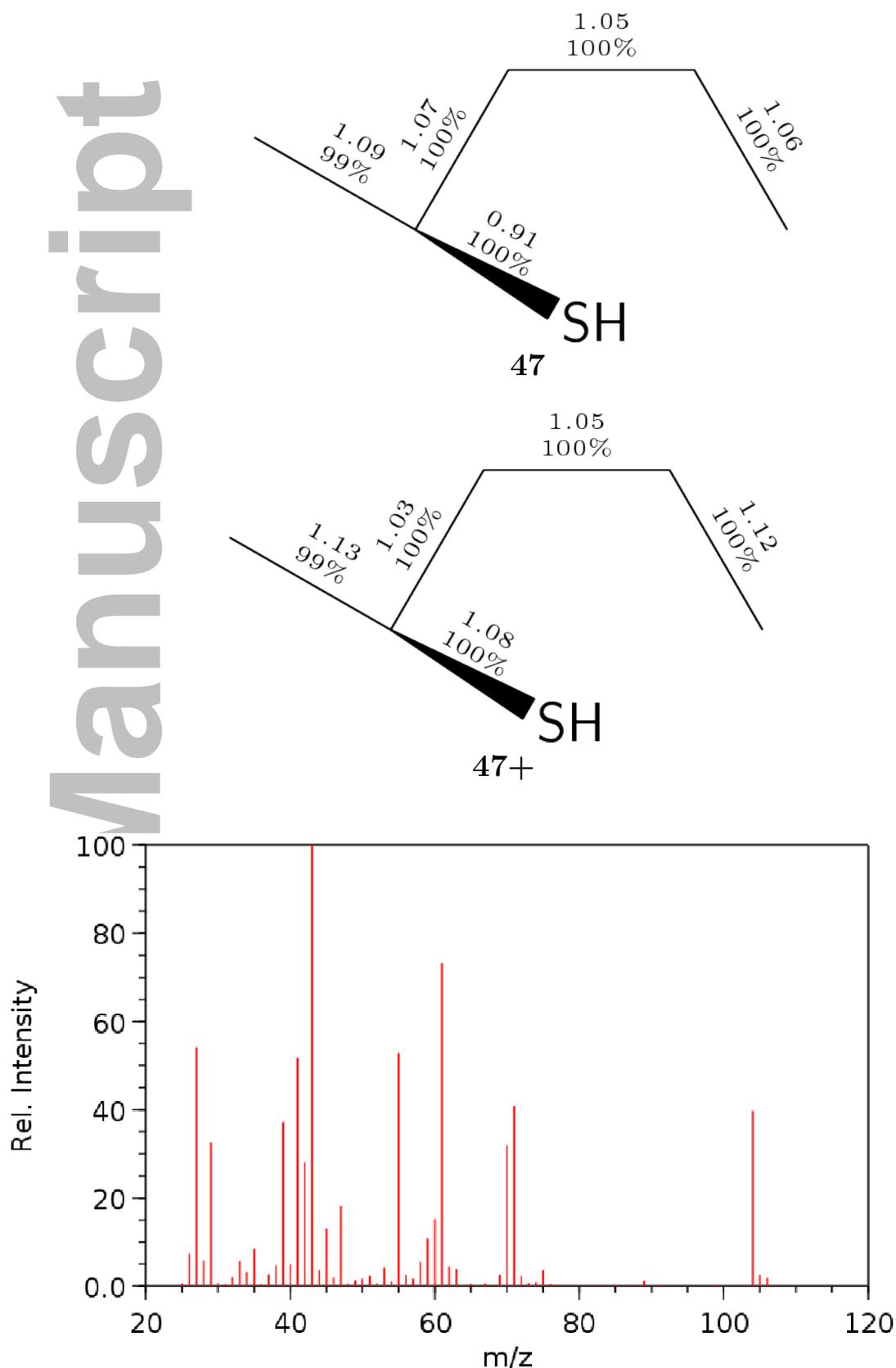
Figure S32: Structure of molecule **32** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**32**) and cation (**32+**).
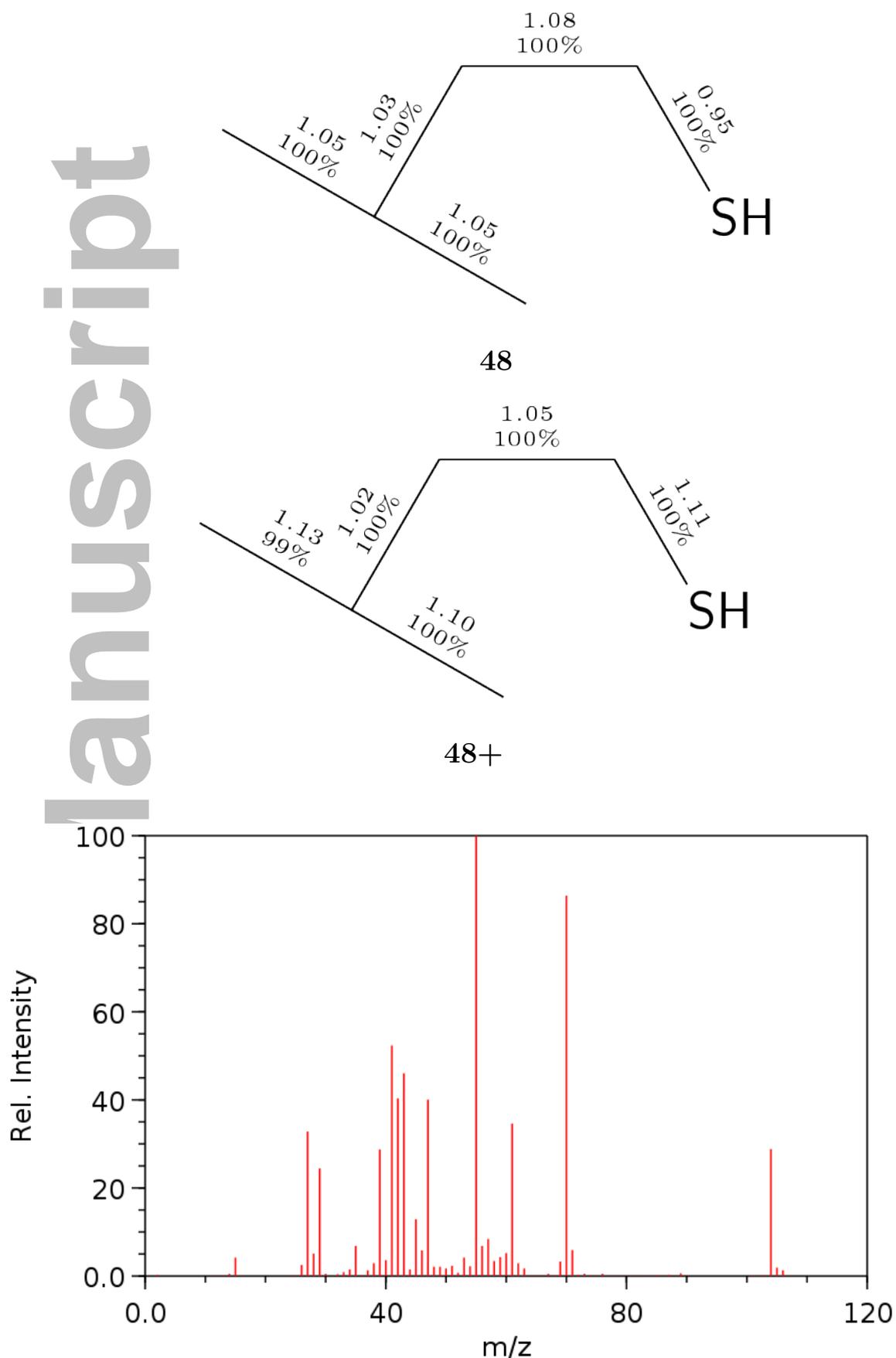
Figure S33: Structure of molecule **33** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**33**) and cation (**33+**).
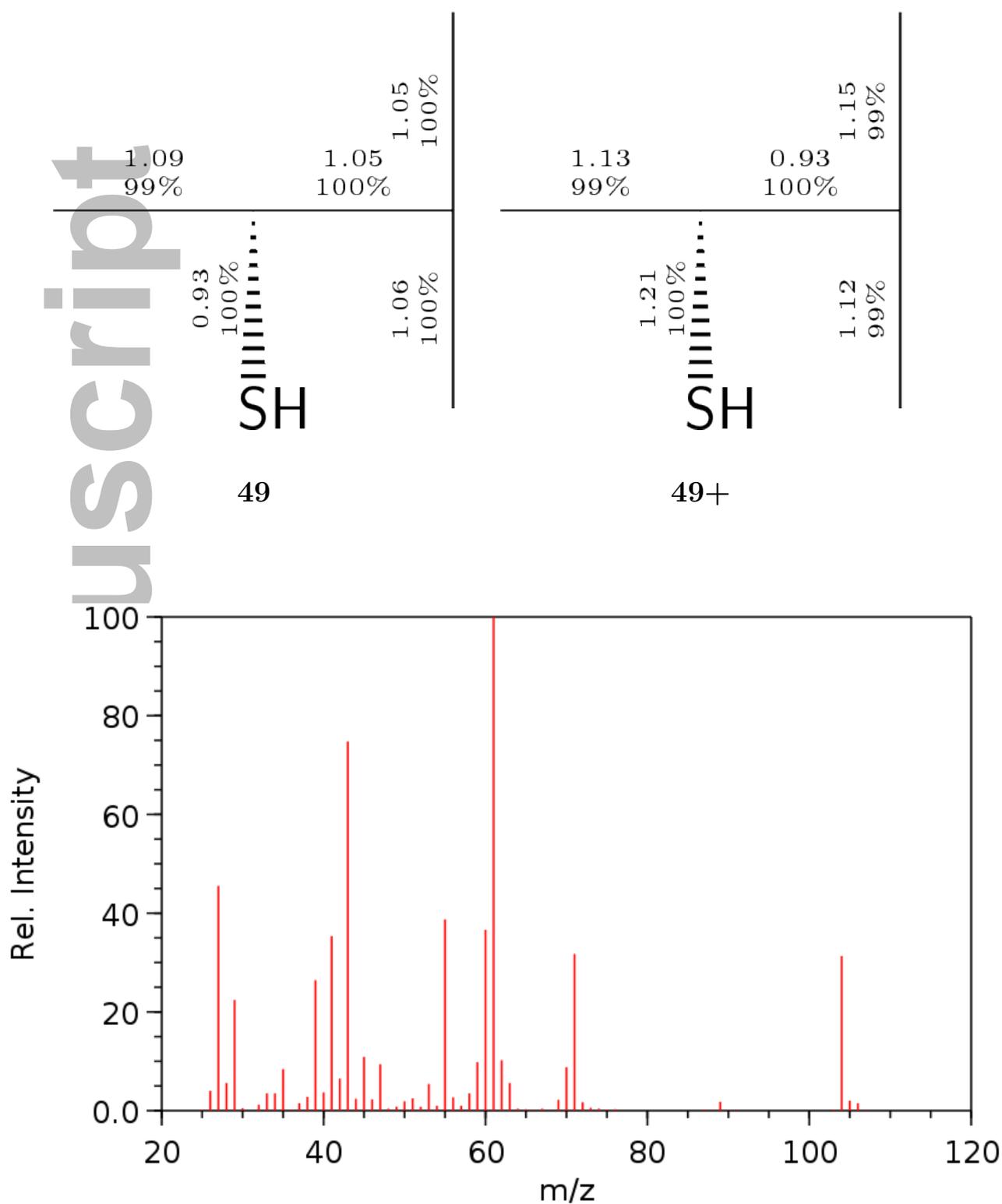
Figure S34: Structure of molecule **34** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**34**) and cation (**34+**).

S3.4   Aldehydes and Ketones

S3.4.1   Isomers of C_4H_8O formula 35-36

Figure S35: Structure of molecule **35** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**35**) and cation (**35+**).
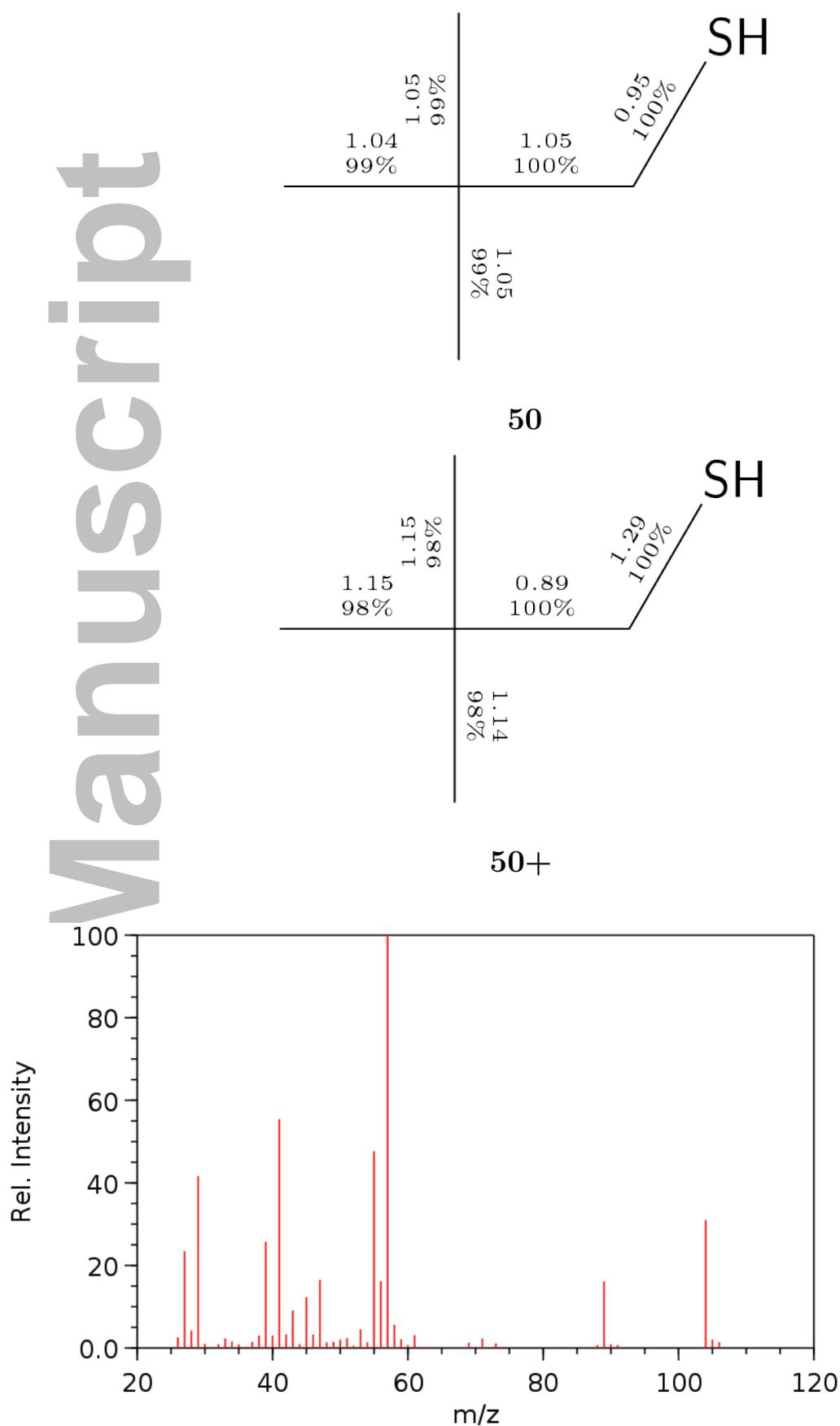
Figure S36: Structure of molecule **36** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**36**) and cation (**36+**).

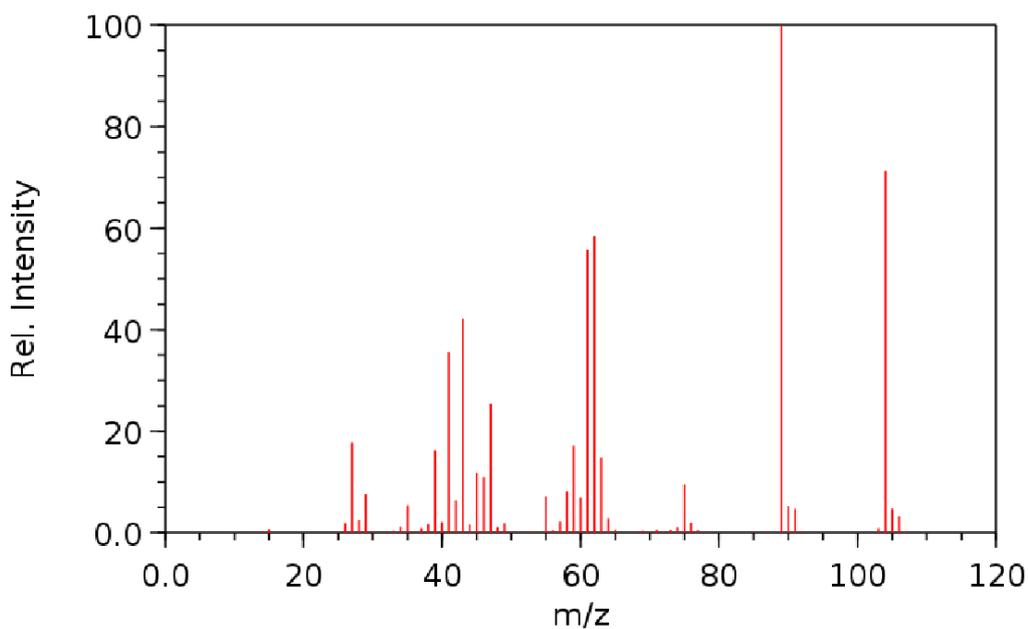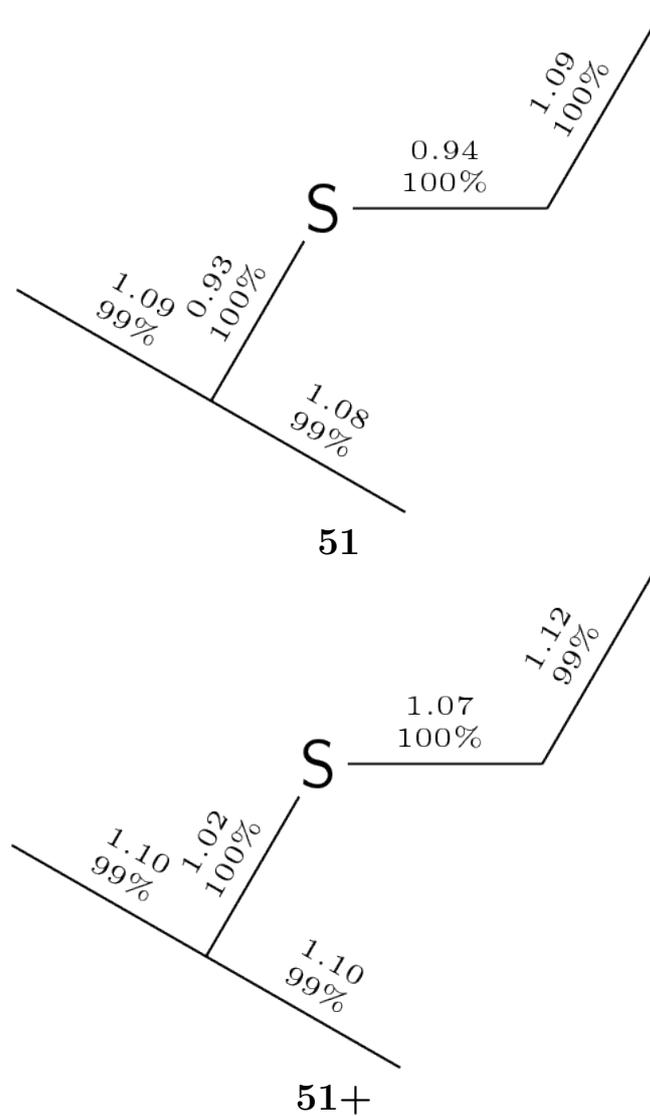## S3.4.2 Isomers of $C_5H_{10}O$ formula 37-40

Figure S37: Structure of molecule **37** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**37**) and cation (**37+**).

Figure S38: Structure of molecule **38** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**38**) and cation (**38+**).
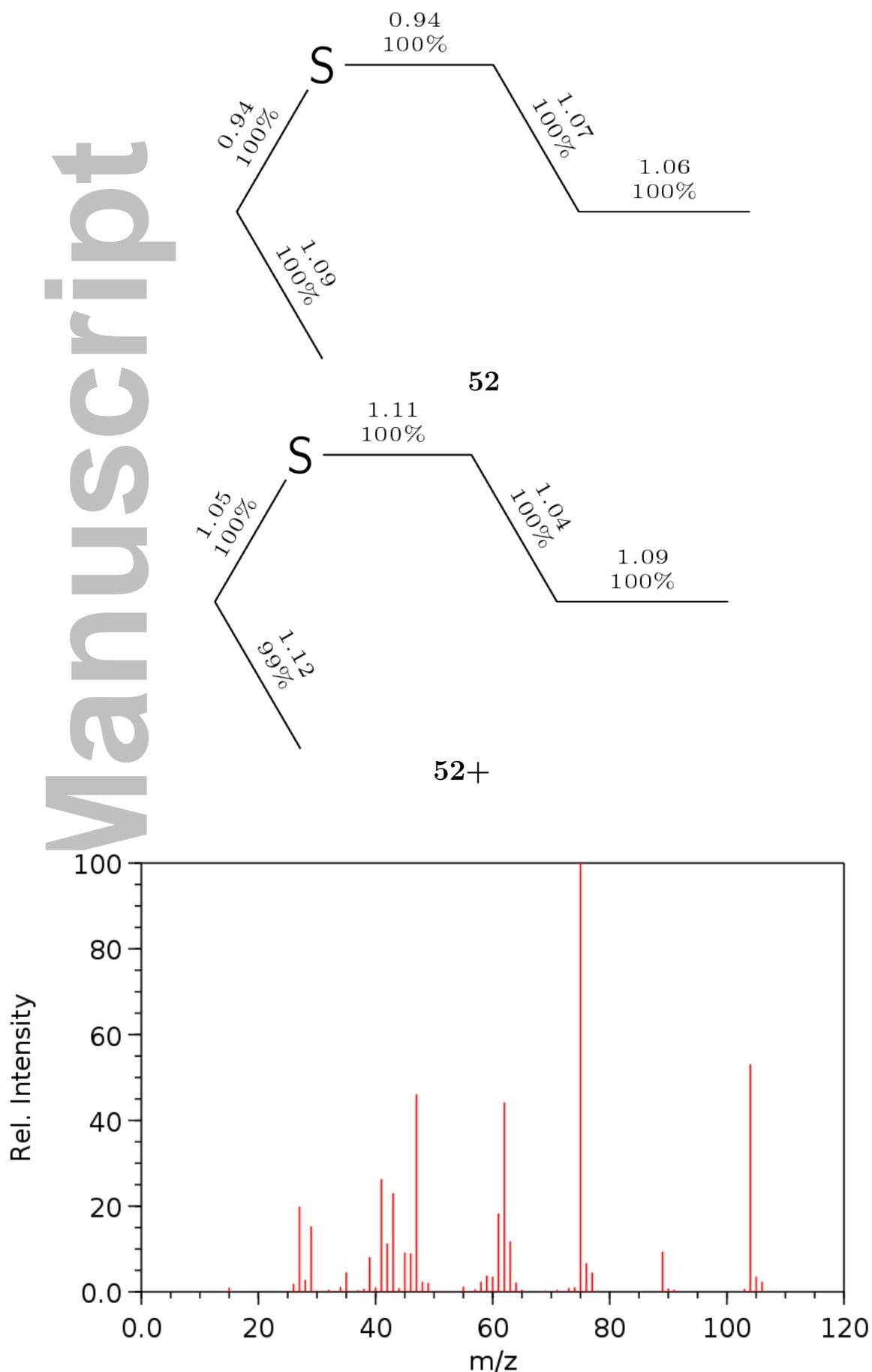
Figure S39: Structure of molecule **39** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**39**) and cation (**39+**).
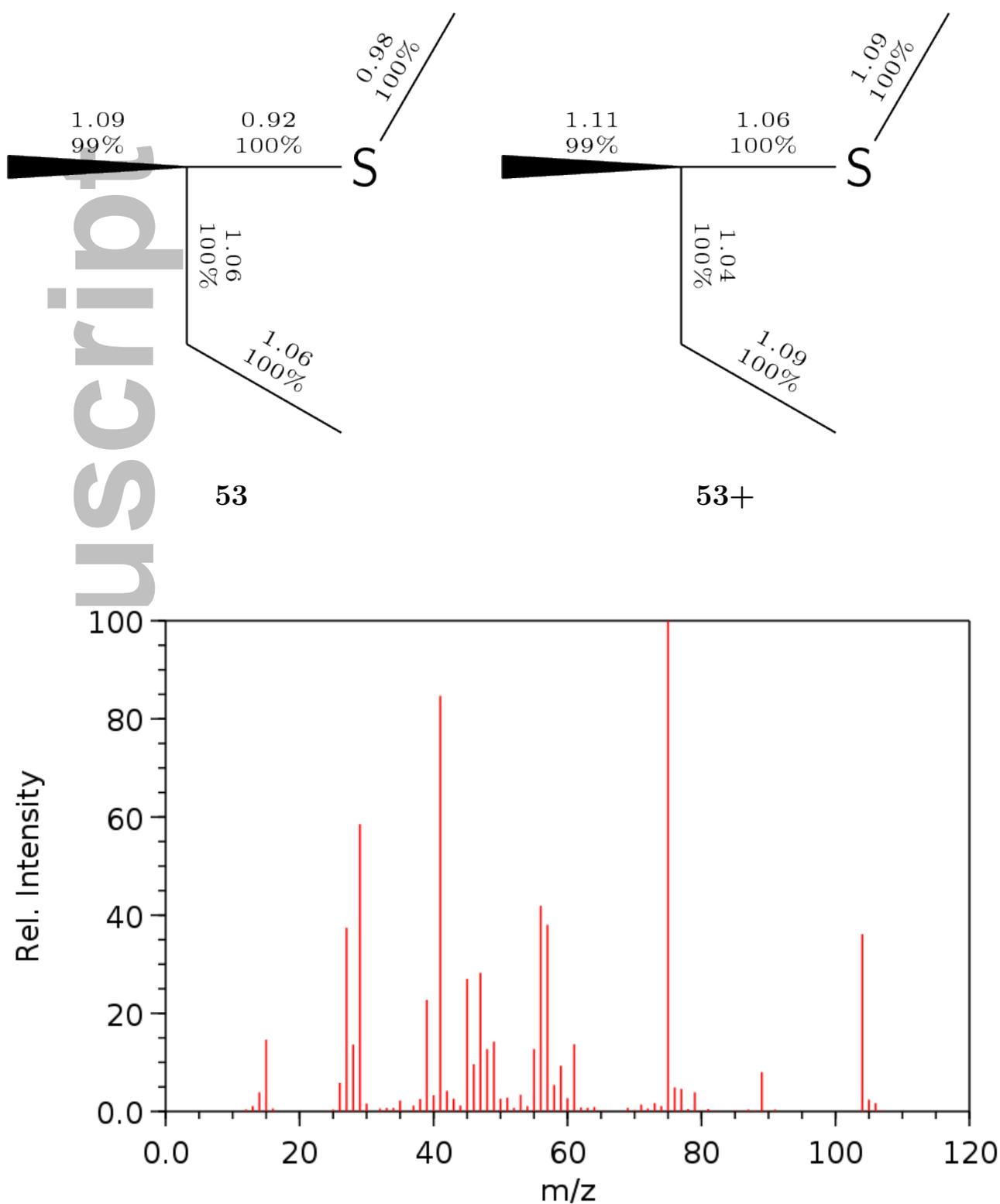
Figure S40: Structure of molecule **40** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**40**) and cation (**40+**).

S3.5   Thiols and Sulfides

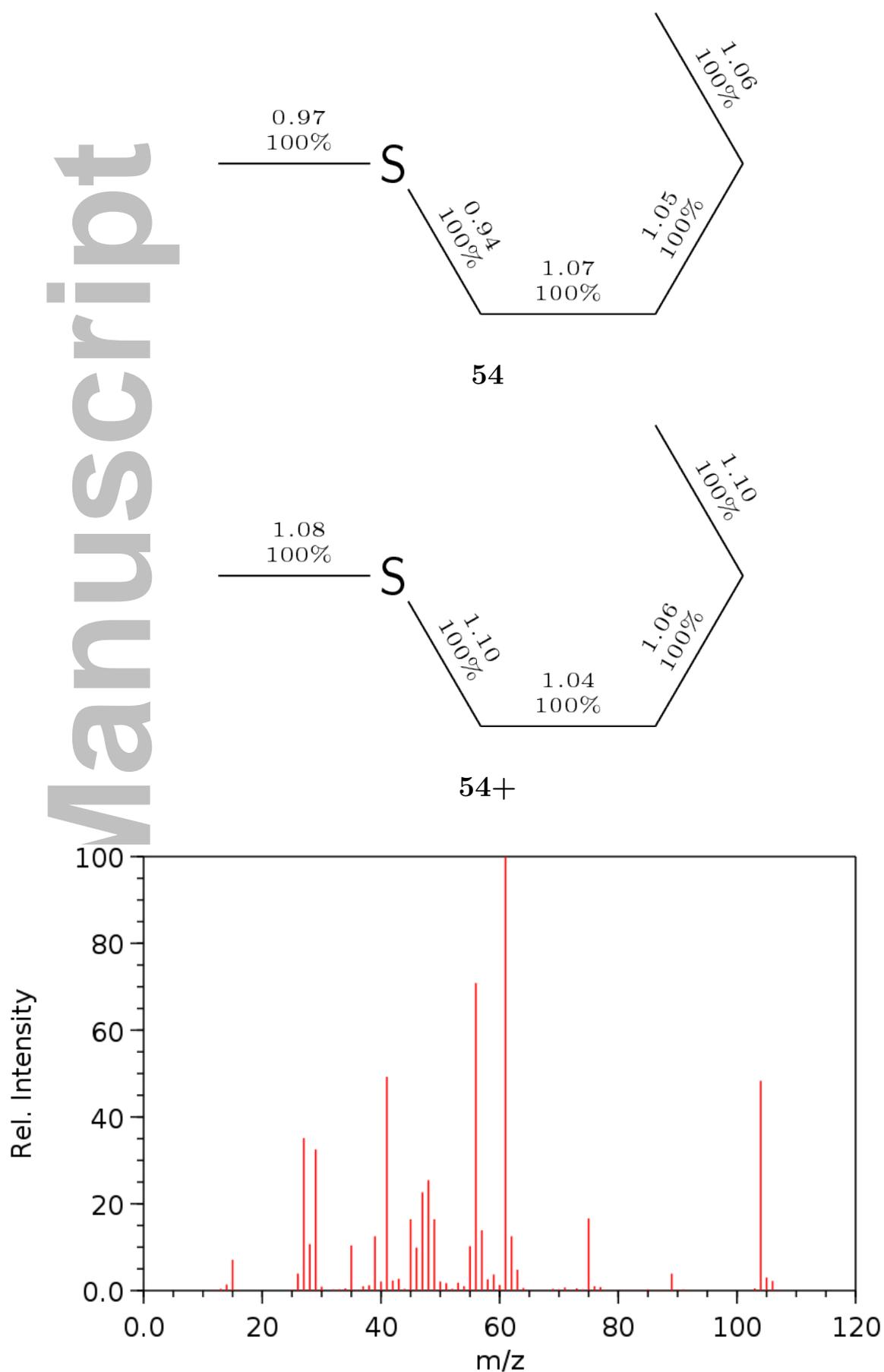### S3.5.1   Isomers of $C_4H_{10}S$ formula 41-46

Figure S41: Structure of molecule **41** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**41**) and cation (**41+**).

Figure S42: Structure of molecule **42** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**42**) and cation (**42+**).
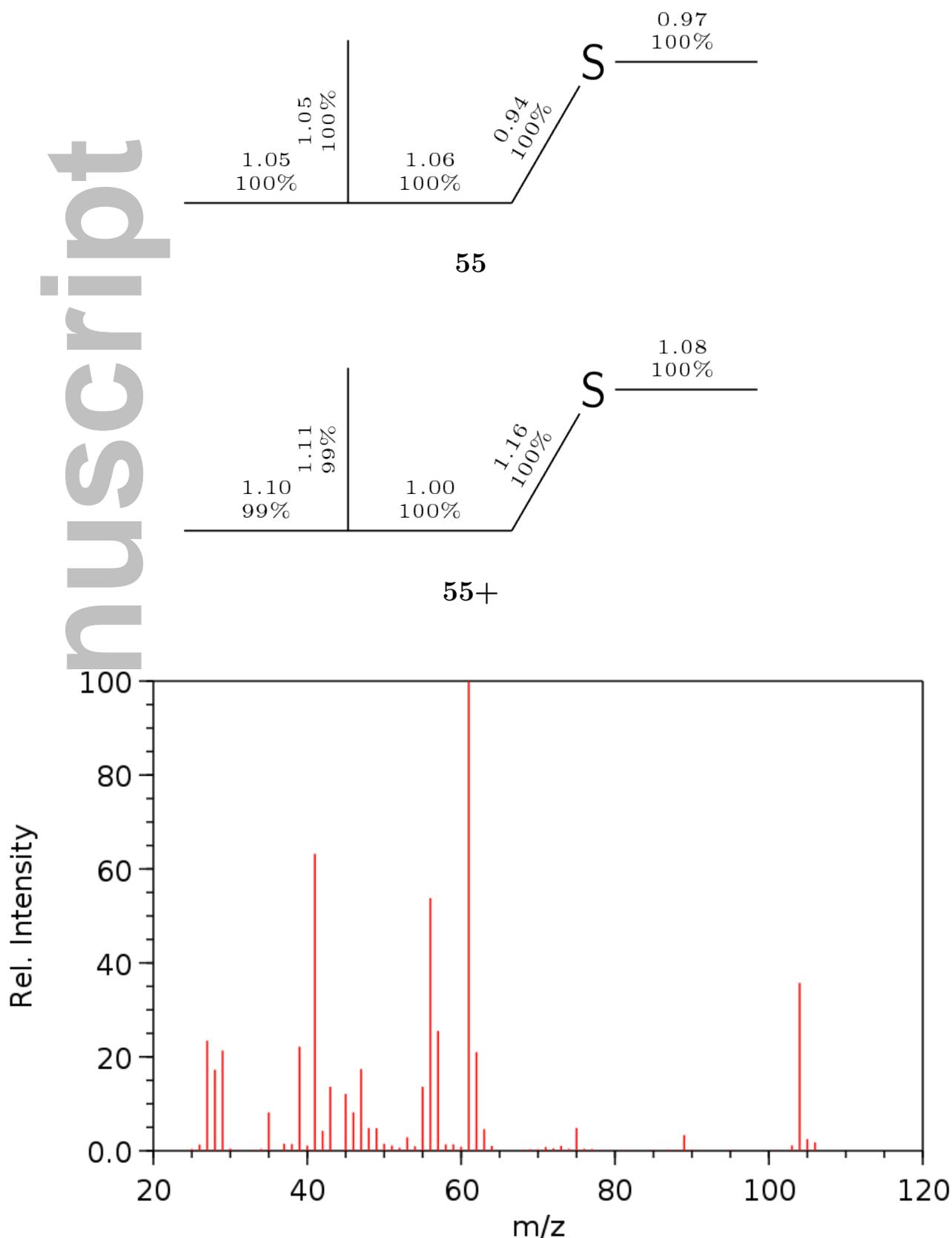
Figure S43: Structure of molecule **43** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**43**) and cation (**43+**).
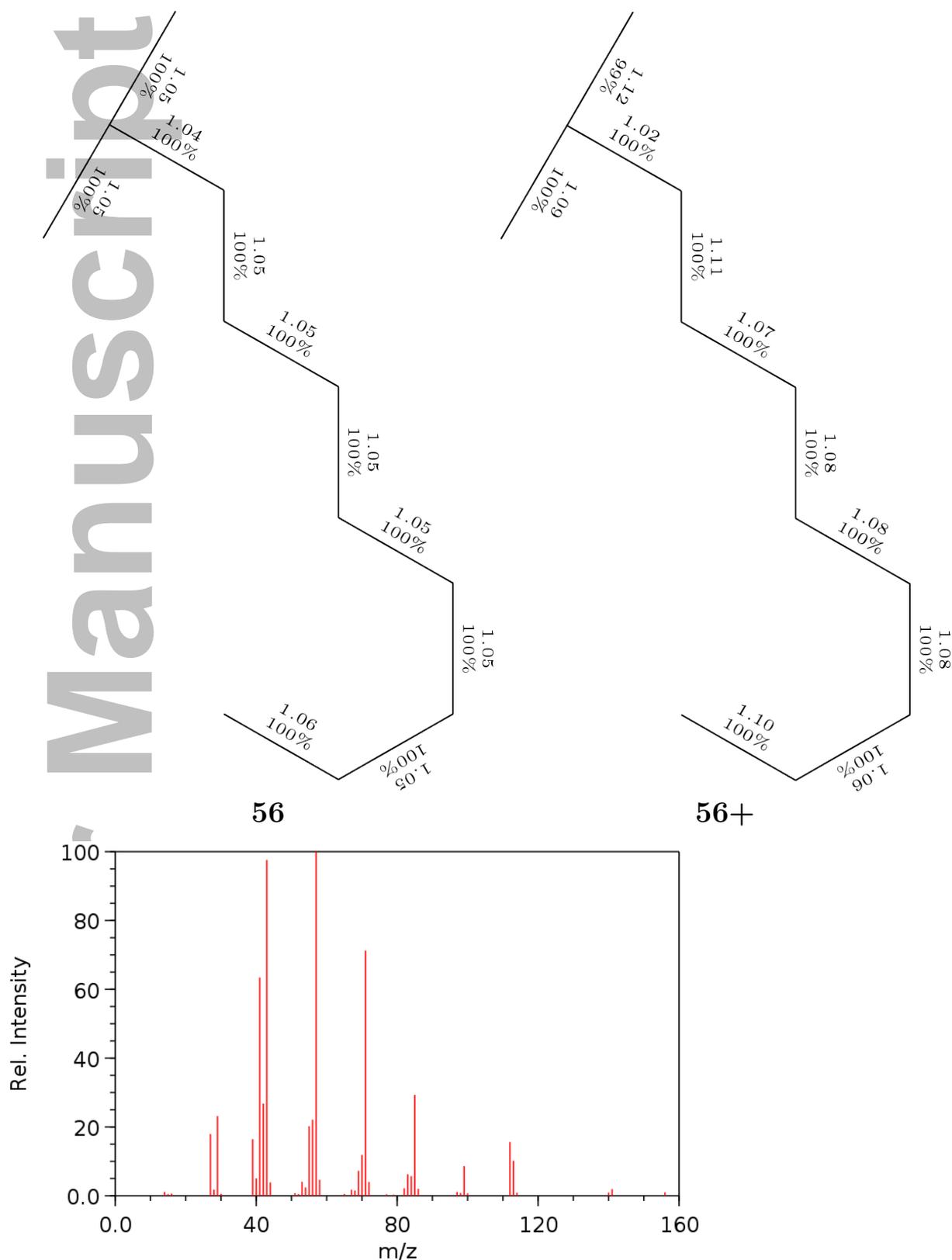
Figure S44: Structure of molecule **44** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**44**) and cation (**44+**).
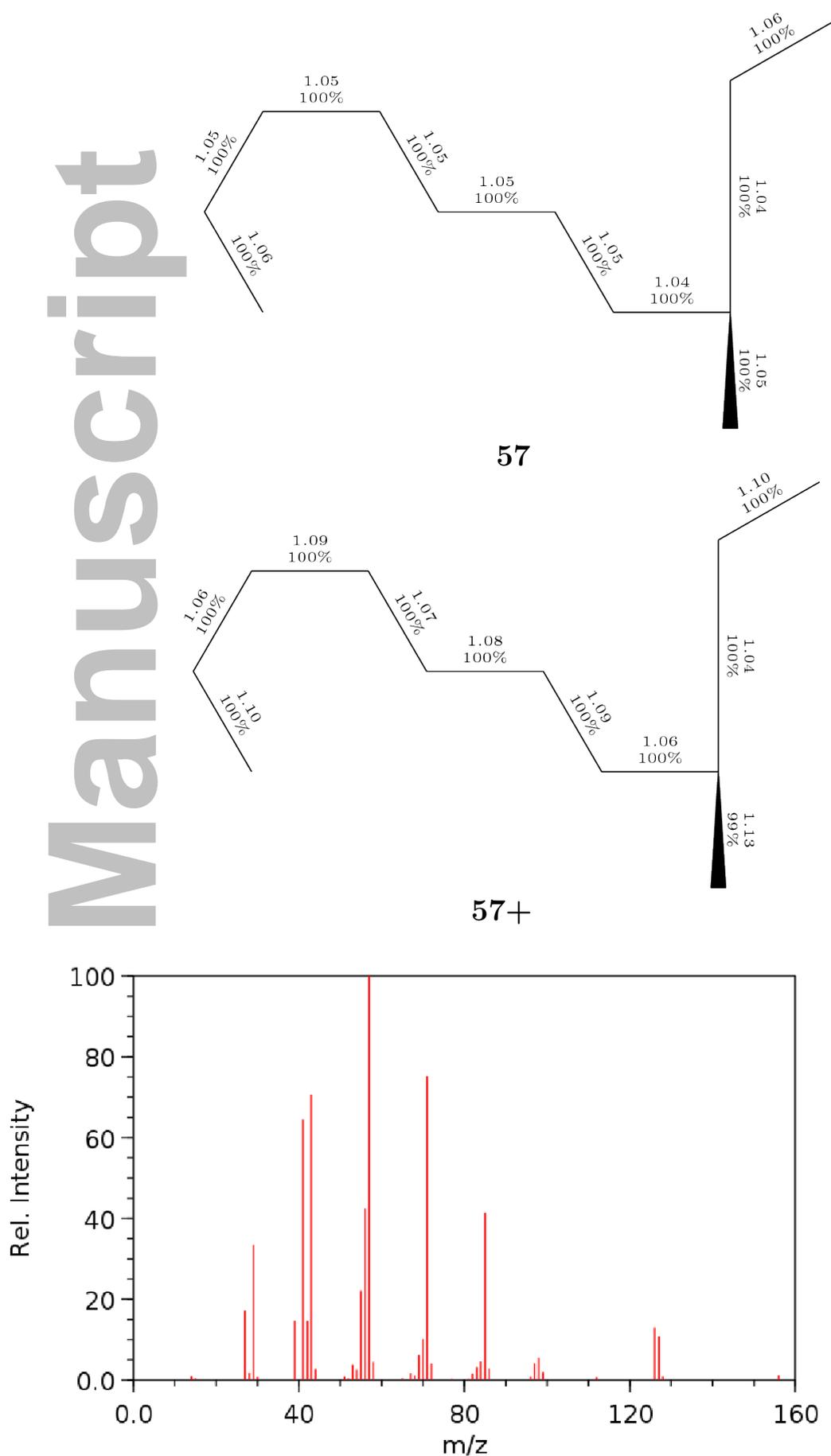
Figure S45: Structure of molecule **45** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**45**) and cation (**45+**).

Figure S46: Structure of molecule **46** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**46**) and cation (**46+**).

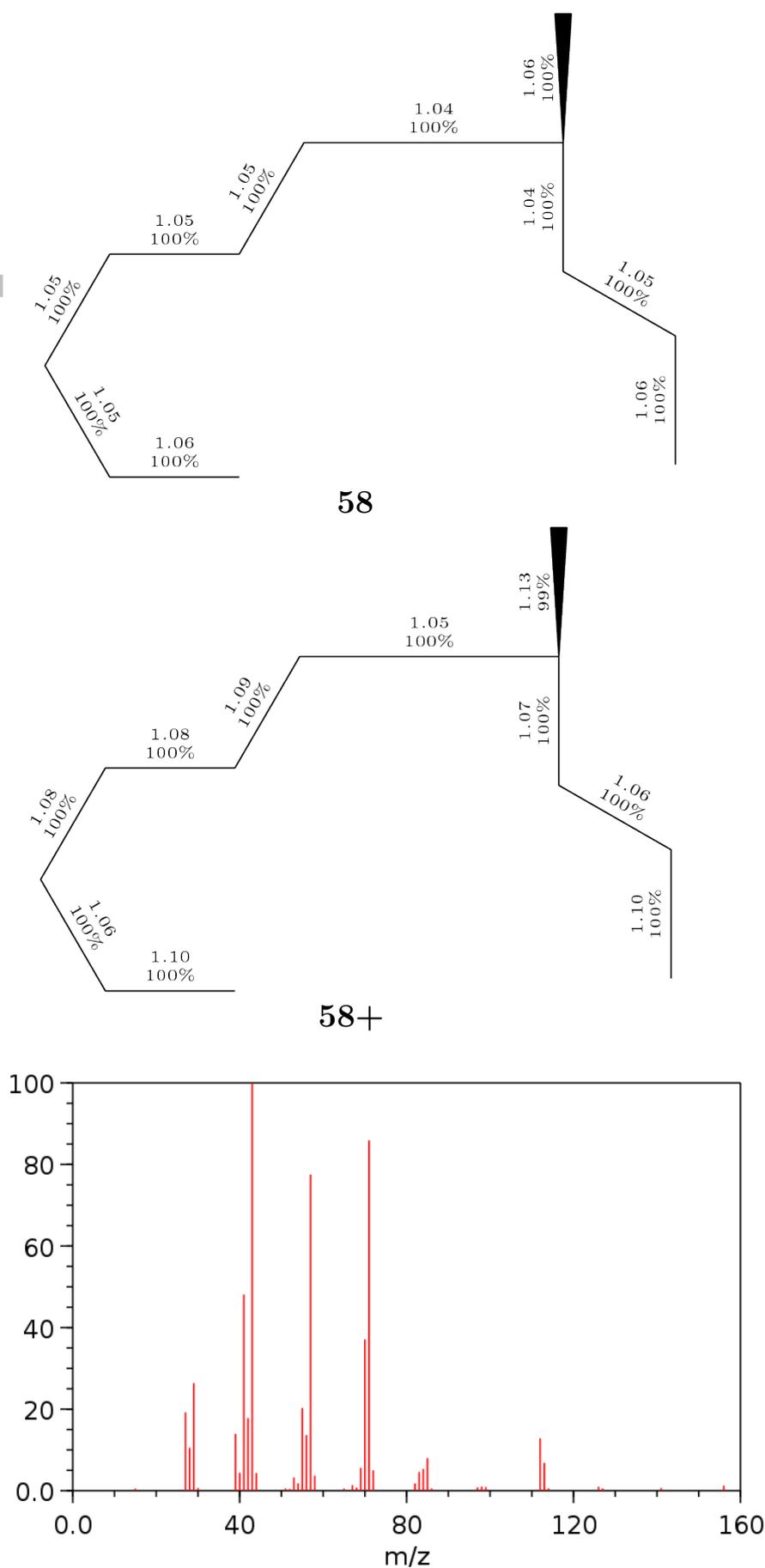## S3.5.2   Isomers of $C_5H_{12}S$ formula 47-55

Figure S47: Structure of molecule **47** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**47**) and cation (**47+**).
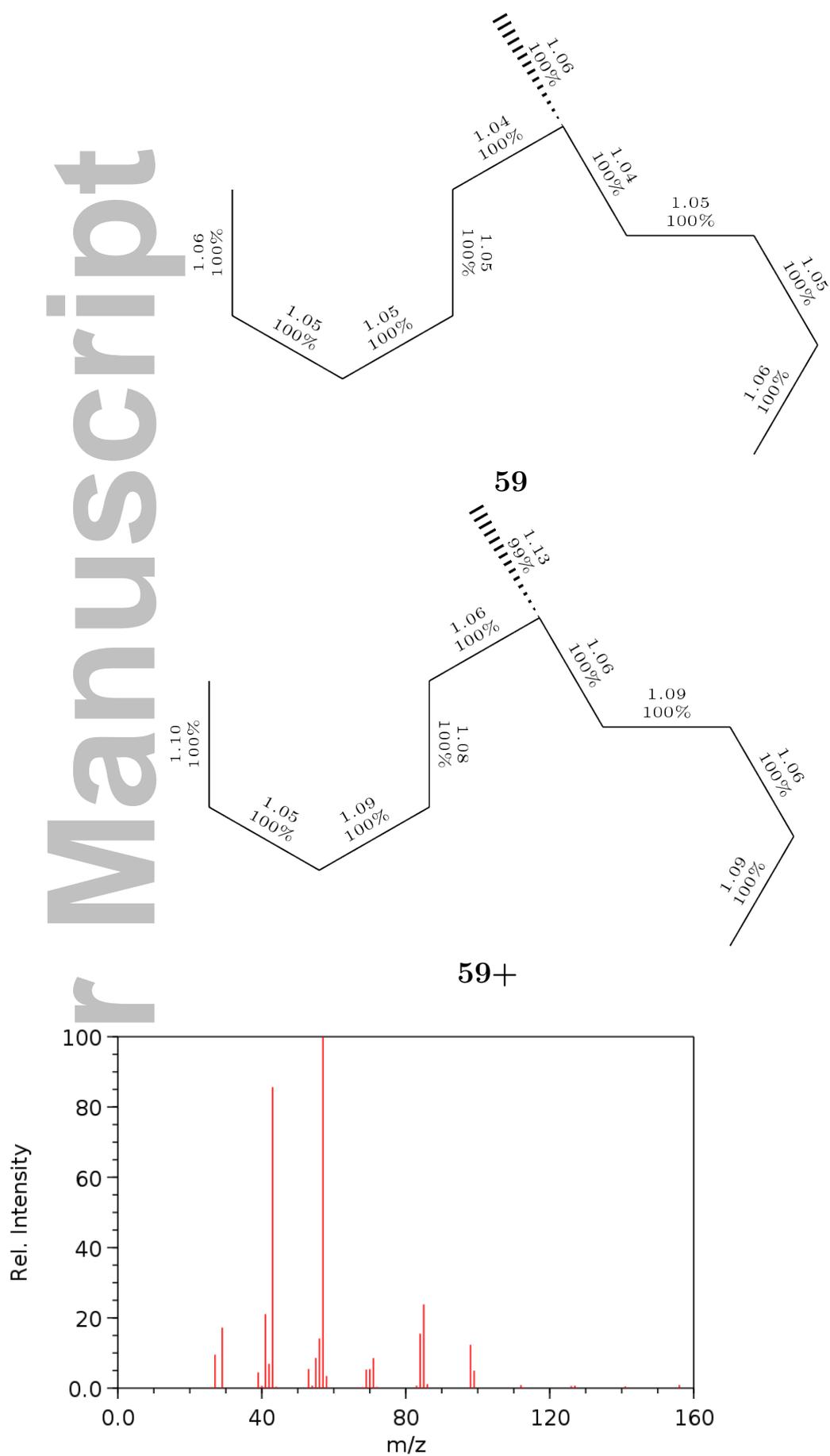
Figure S48: Structure of molecule **48** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**48**) and cation (**48+**).

Figure S49: Structure of molecule **49** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**49**) and cation (**49+**).

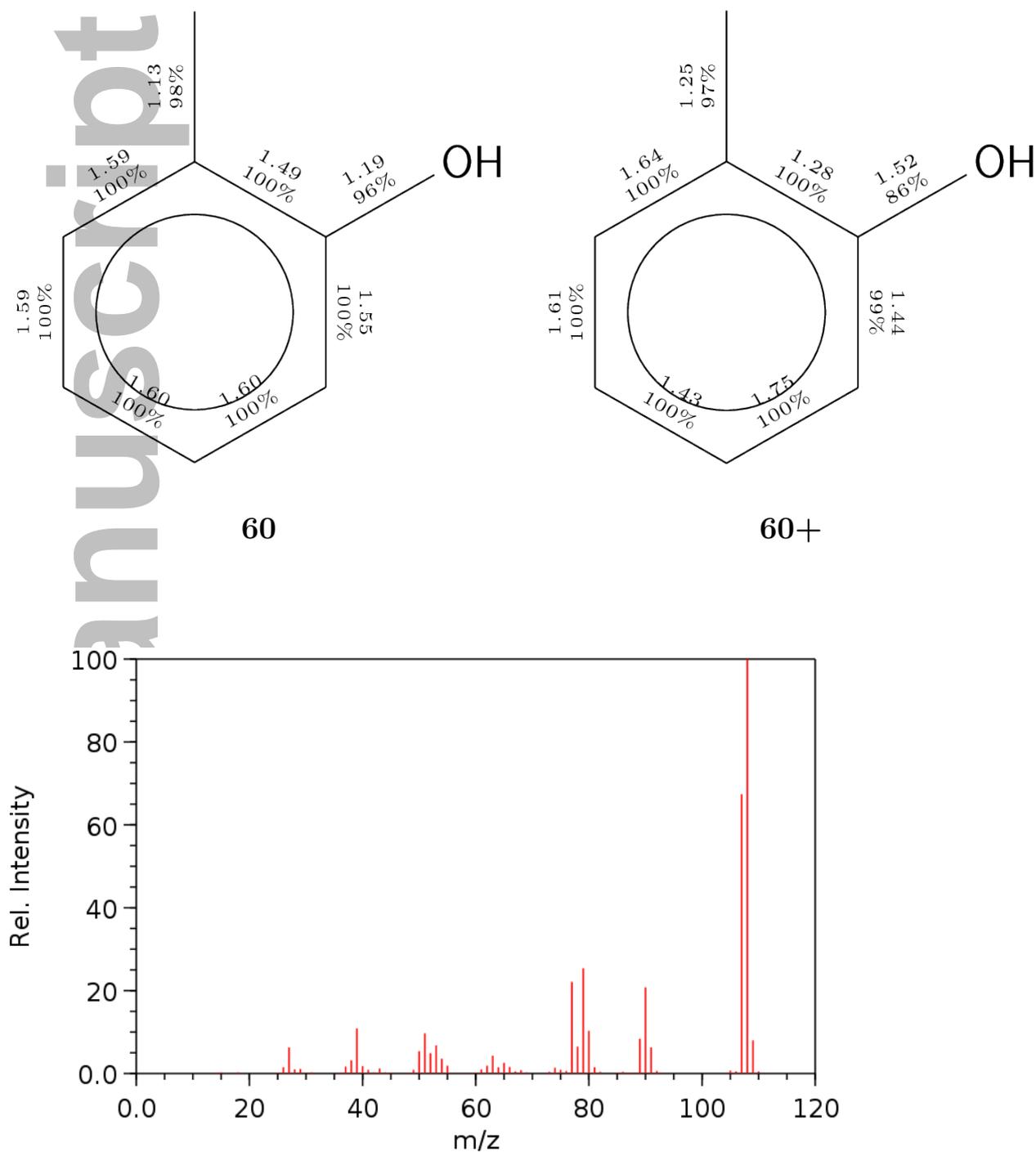Figure S50: Structure of molecule **50** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**50**) and cation (**50+**).
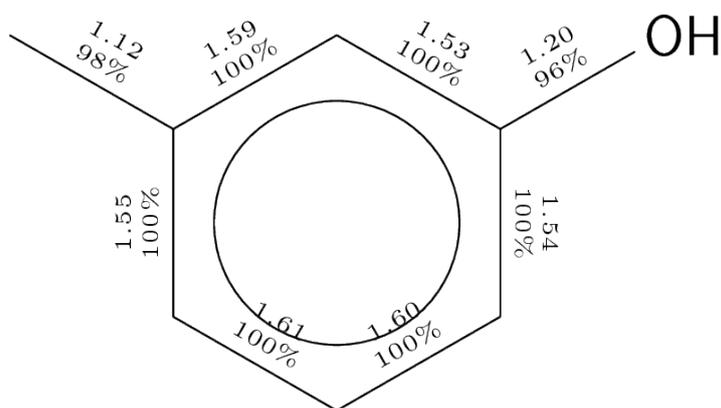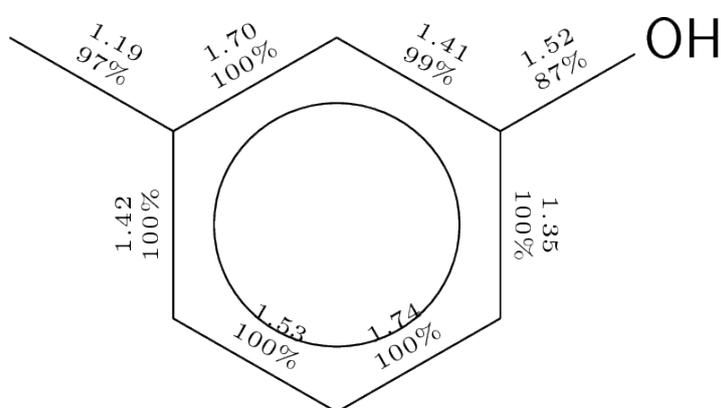
Figure S51: Structure of molecule **51** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**51**) and cation (**51+**).
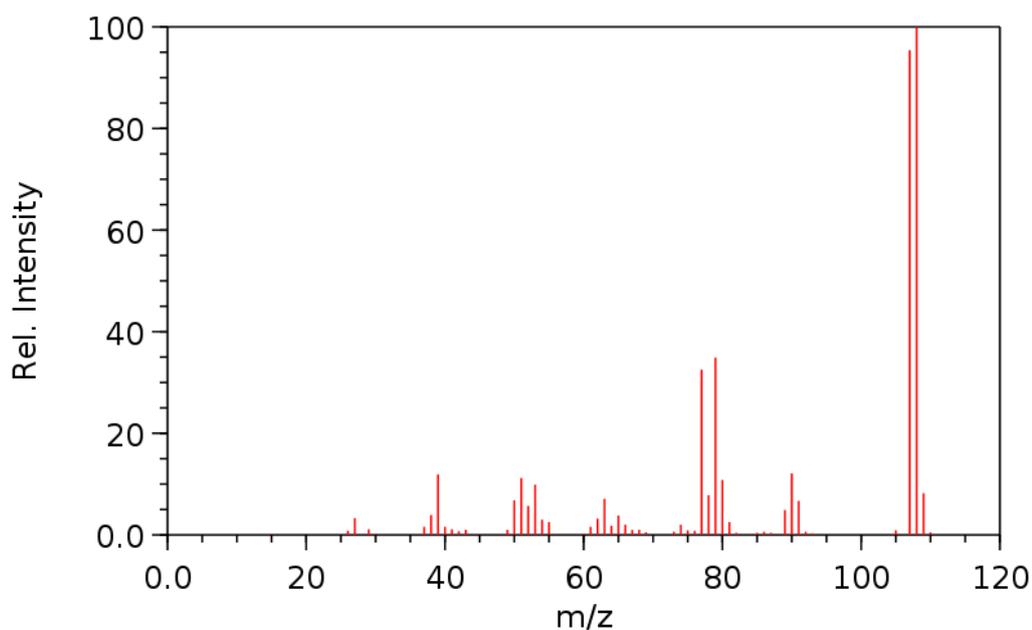
Figure S52: Structure of molecule **52** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**52**) and cation (**52**+).

Figure S53: Structure of molecule **53** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**53**) and cation (**53+**).

Manuscript

0.97
100%

S

0.94
100%

1.07
100%

1.06
100%

1.05
100%

54

1.08
100%

S

1.10
100%

1.04
100%

1.10
100%

1.06
100%

54+

Figure S54: Structure of molecule **54** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**54**) and cation (**54+**).

Figure S55: Structure of molecule **55** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**55**) and cation (**55+**).
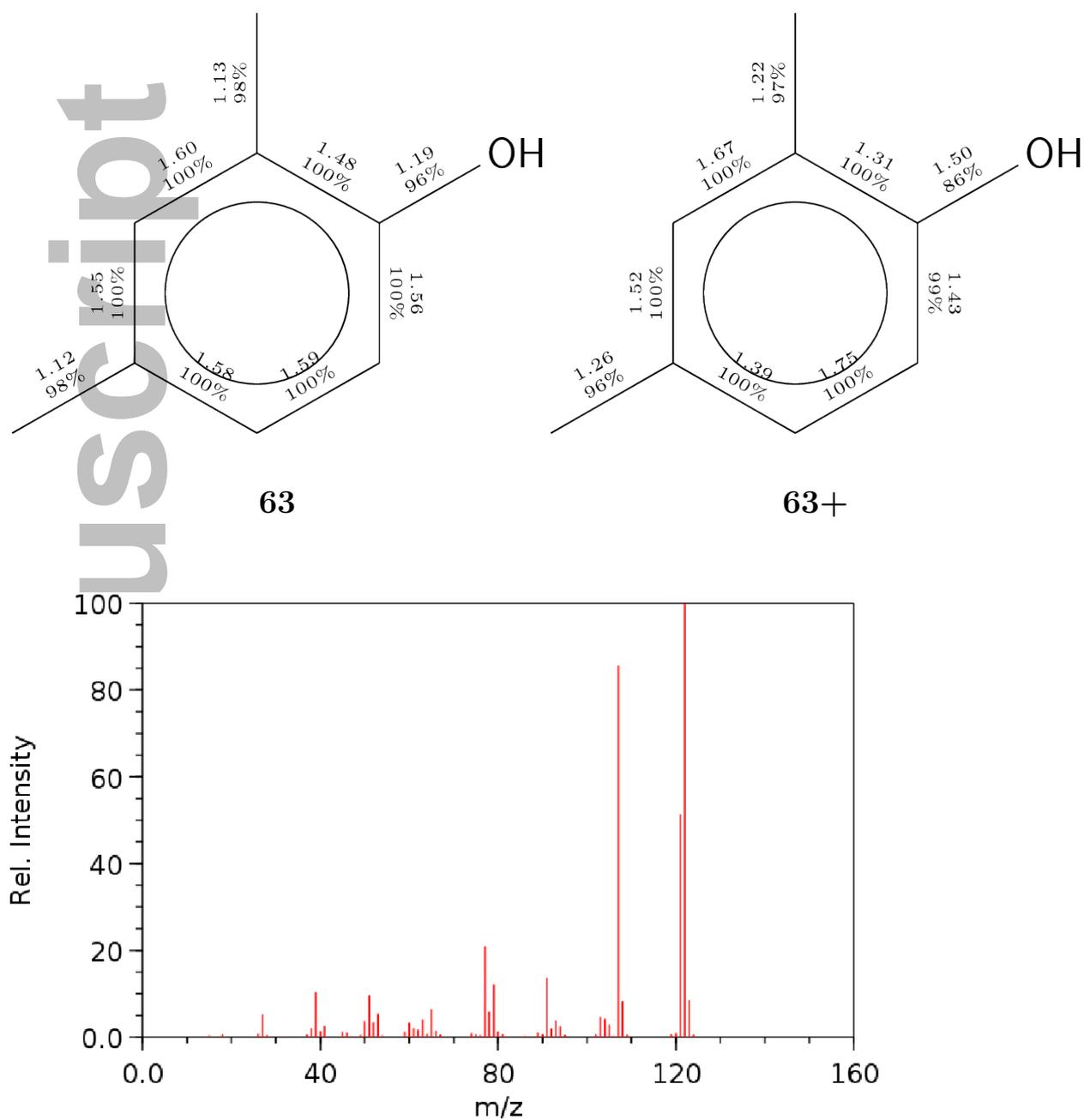
S3.6 Methyldecanes
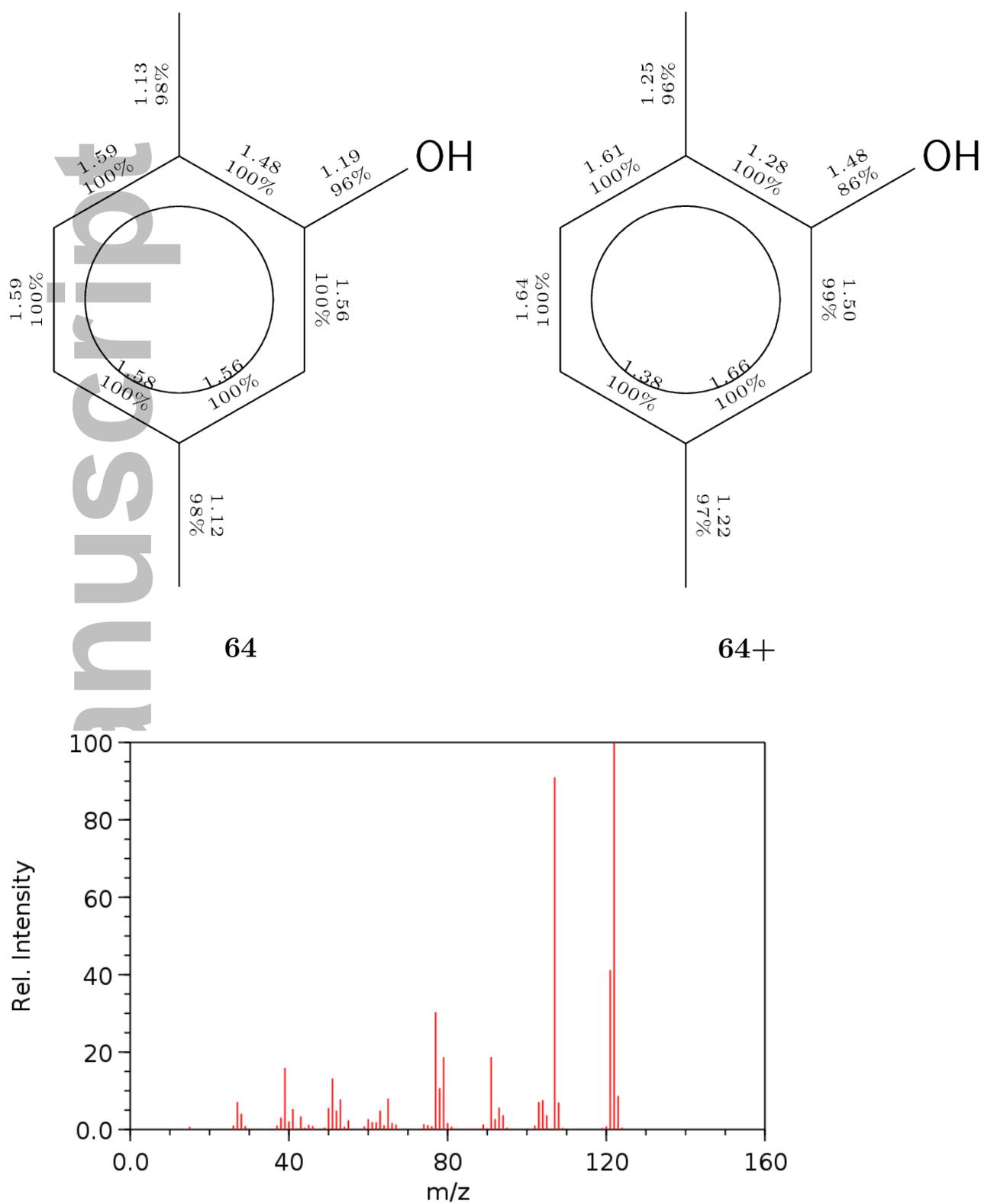
S3.6.1 Isomers of $C_{11}H_{24}$ formula 56-59

Figure S56: Structure of molecule **56** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**56**) and cation (**56+**).
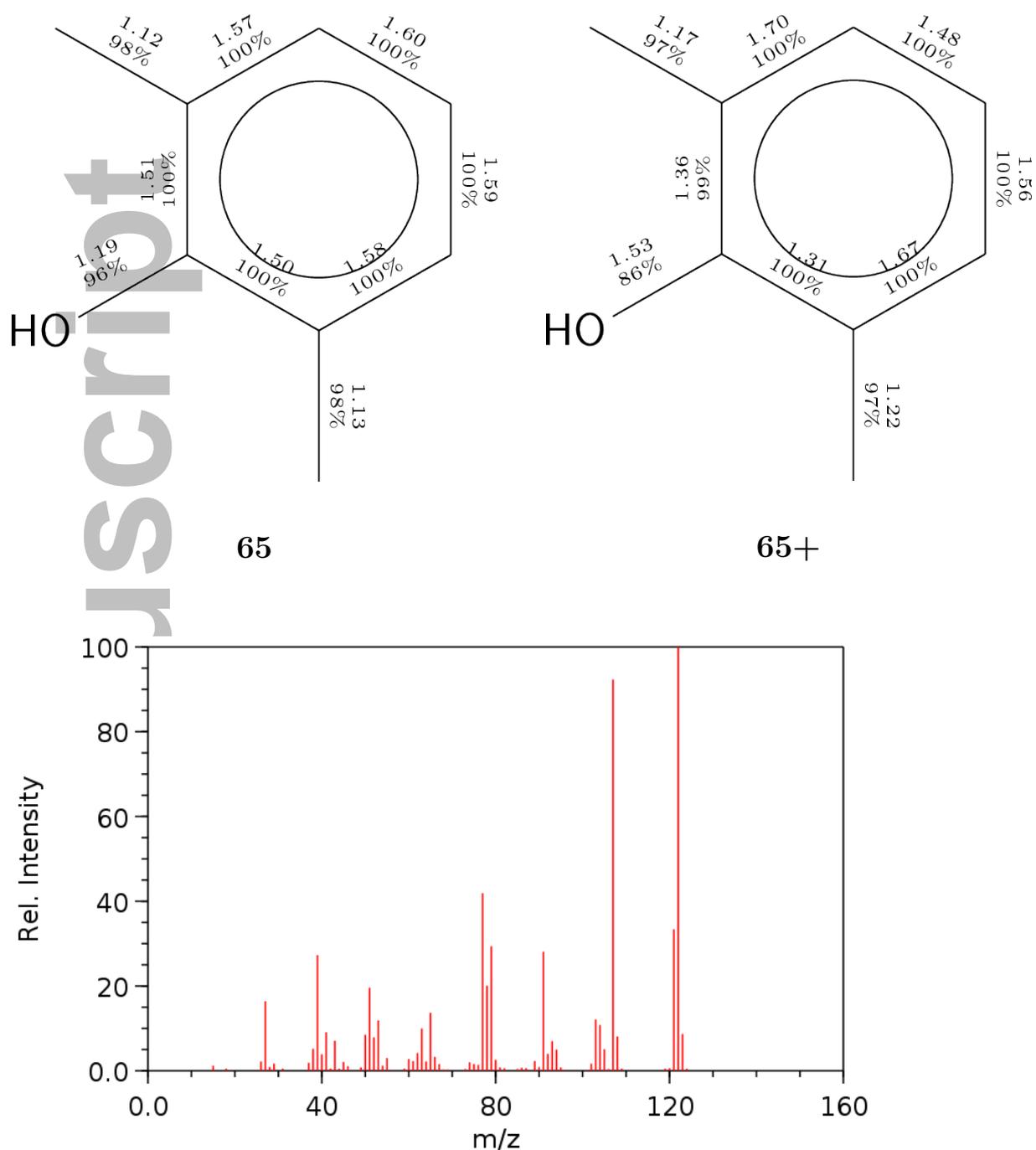
Figure S57: Structure of molecule **57** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**57**) and cation (**57+**).

Figure S58: Structure of molecule **58** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**58**) and cation (**58+**).
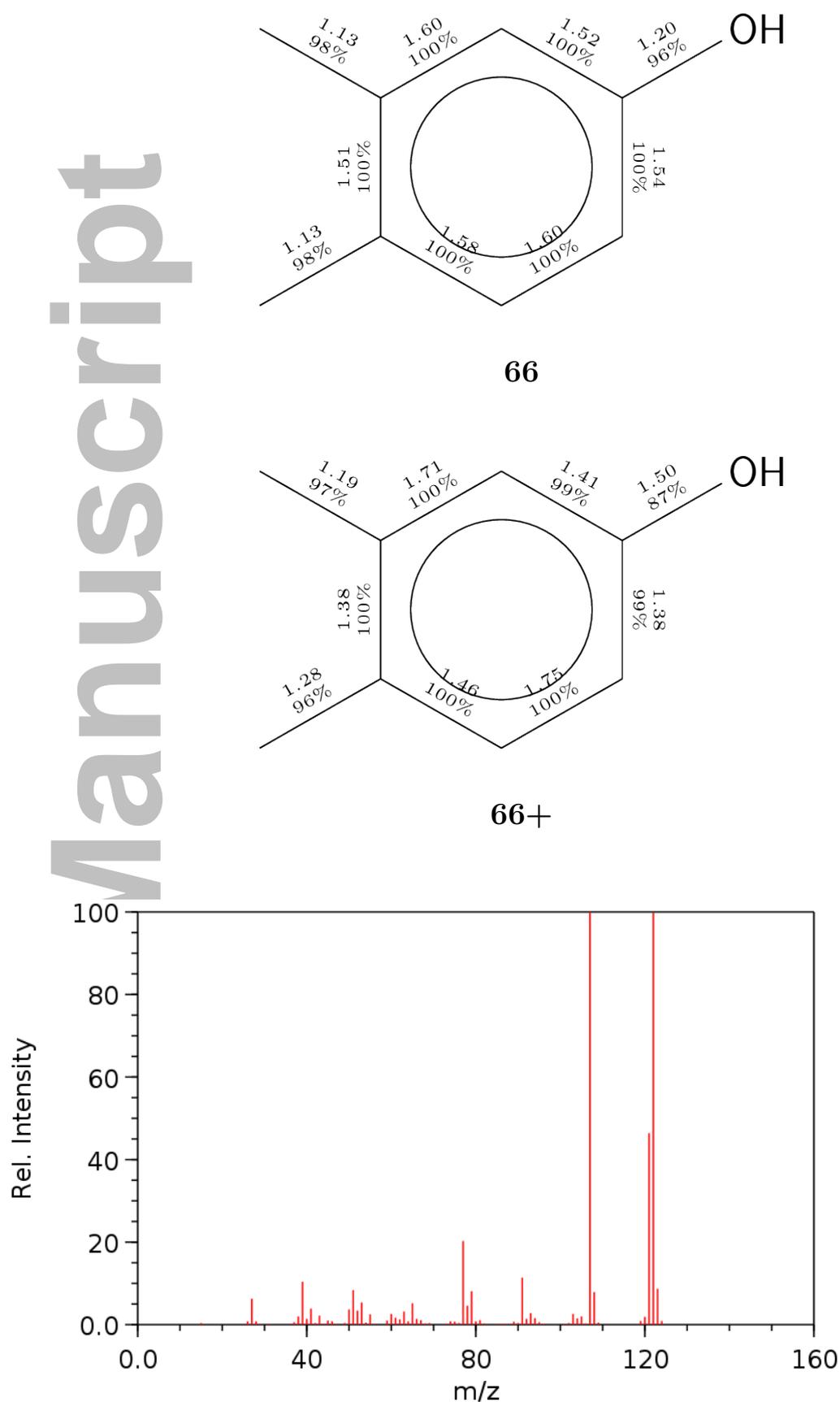
Figure S59: Structure of molecule **59** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**59**) and cation (**59+**).

S3.7   Phenoles

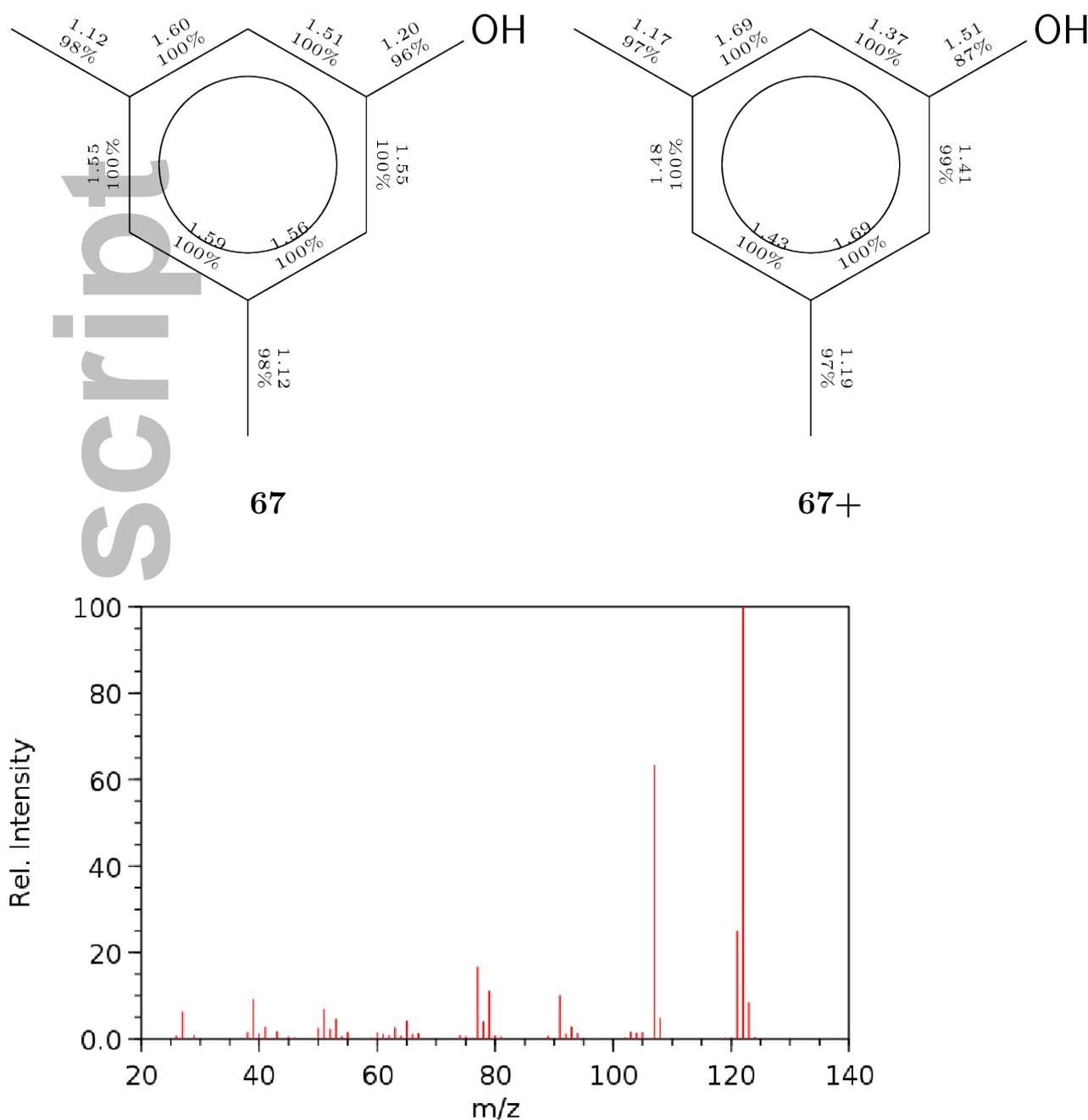### S3.7.1   Isomers of C_7H_7OH formula 60-62

Figure S60: Structure of molecule **60** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**60**) and cation (**60+**).

Figure S61: Structure of molecule **61** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**61**) and cation (**61+**).
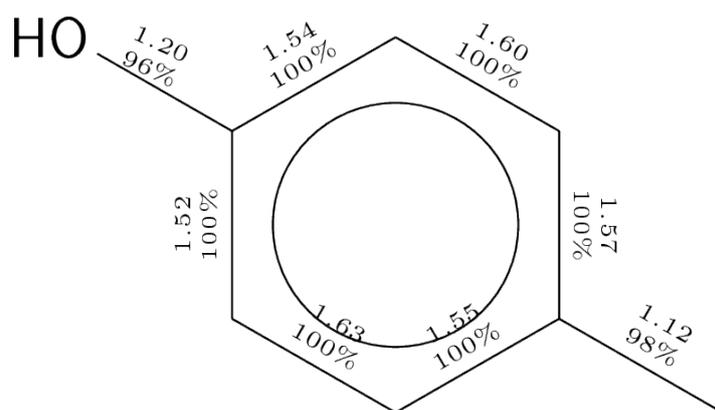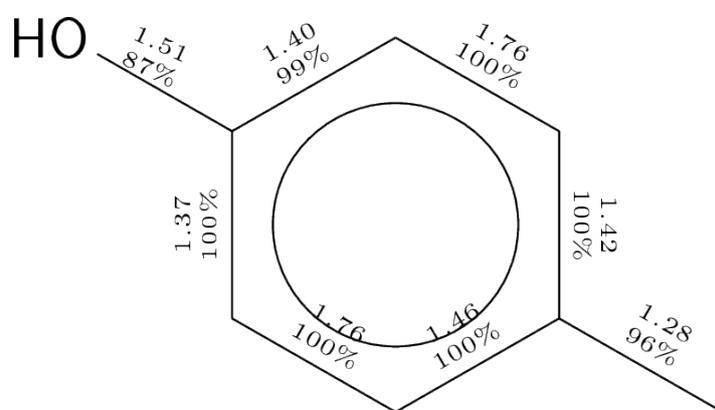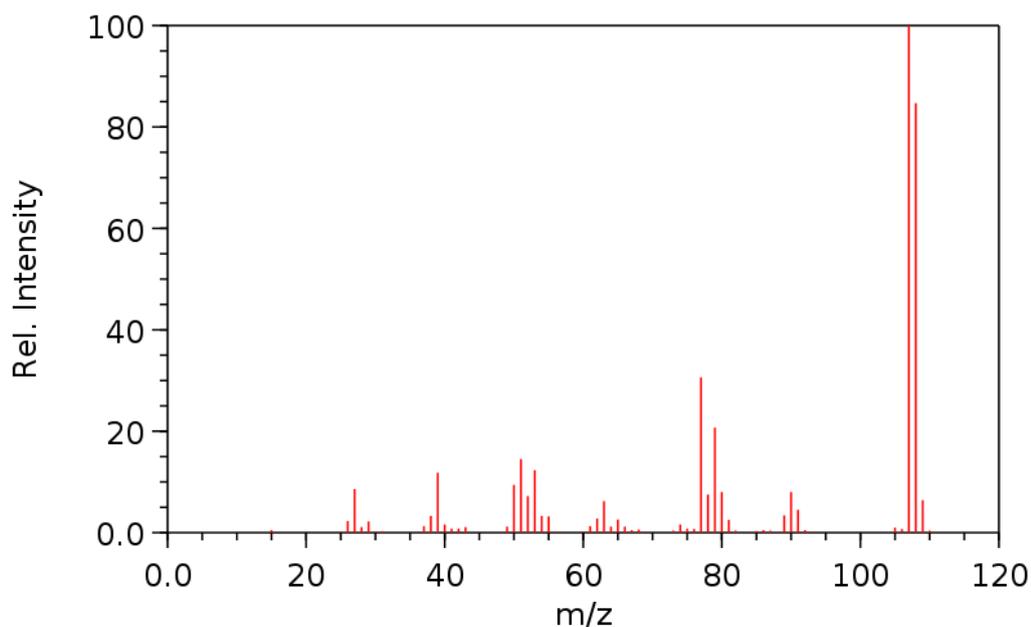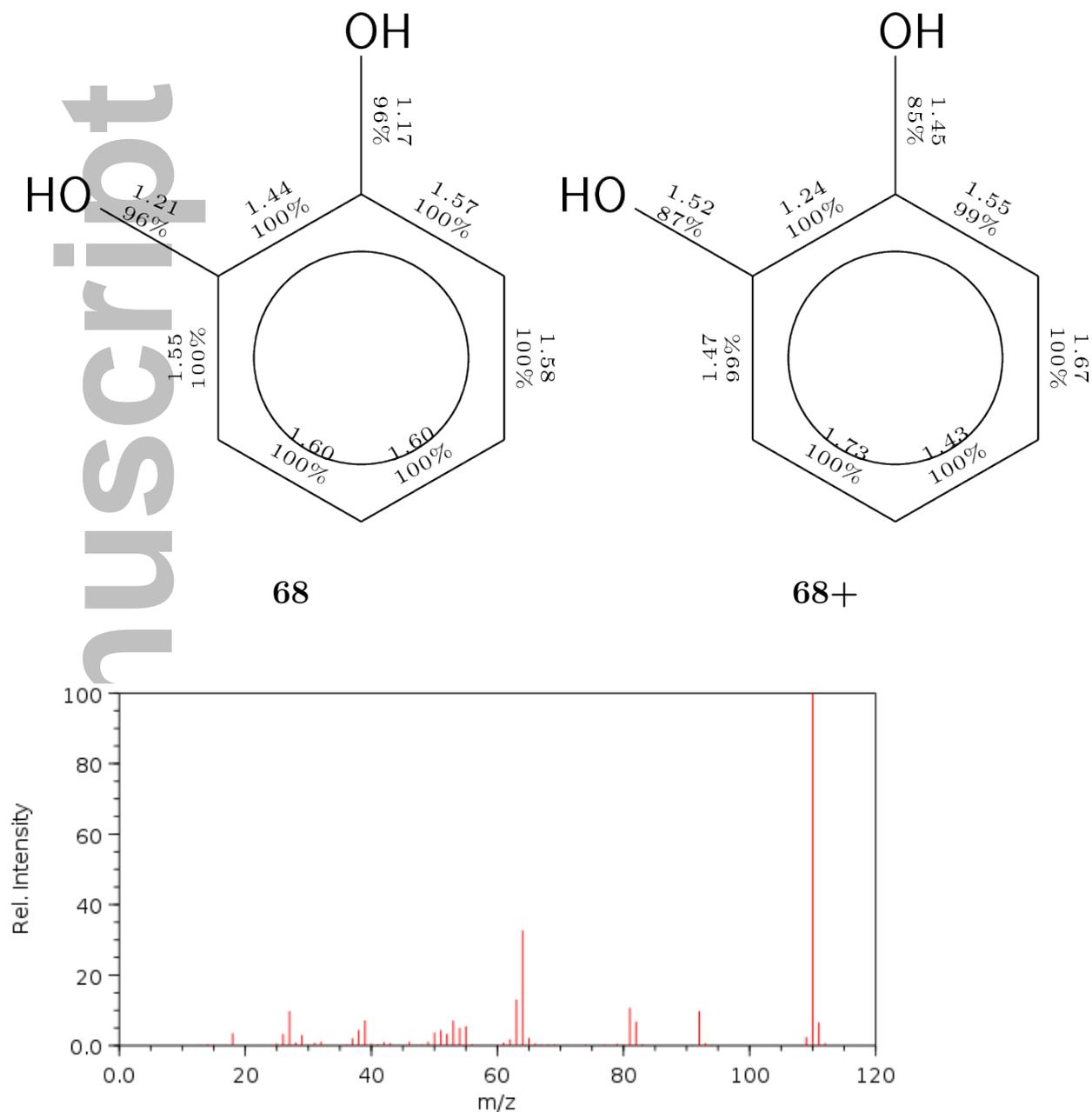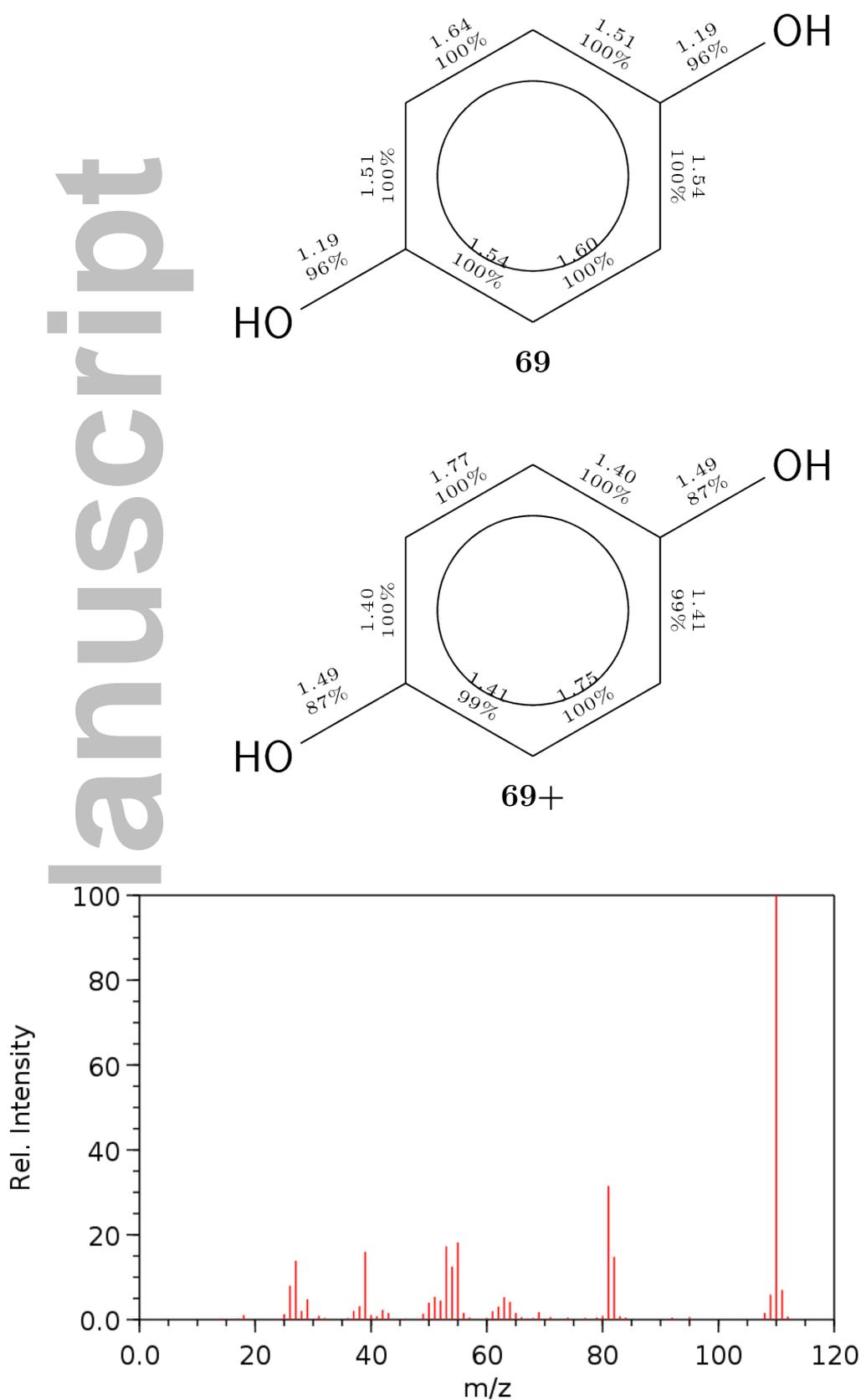
Figure S62: Structure of molecule **62** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**62**) and cation (**62**+).

S3.7.2 Isomers of C_8H_9OH formula 63-67

63

63+

Figure S63: Structure of molecule **63** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**63**) and cation (**63+**).
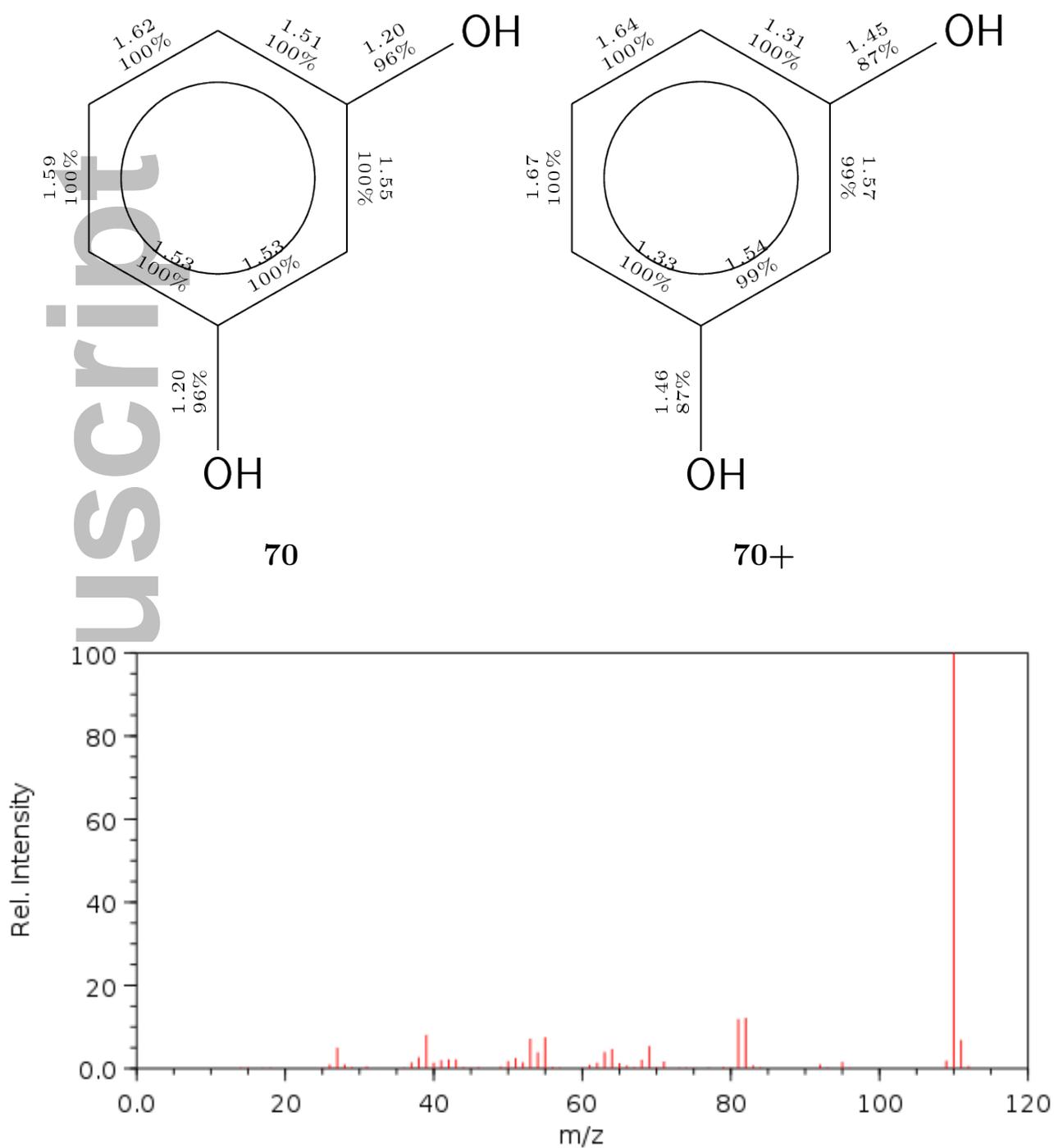
Figure S64: Structure of molecule **64** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**64**) and cation (**64+**).

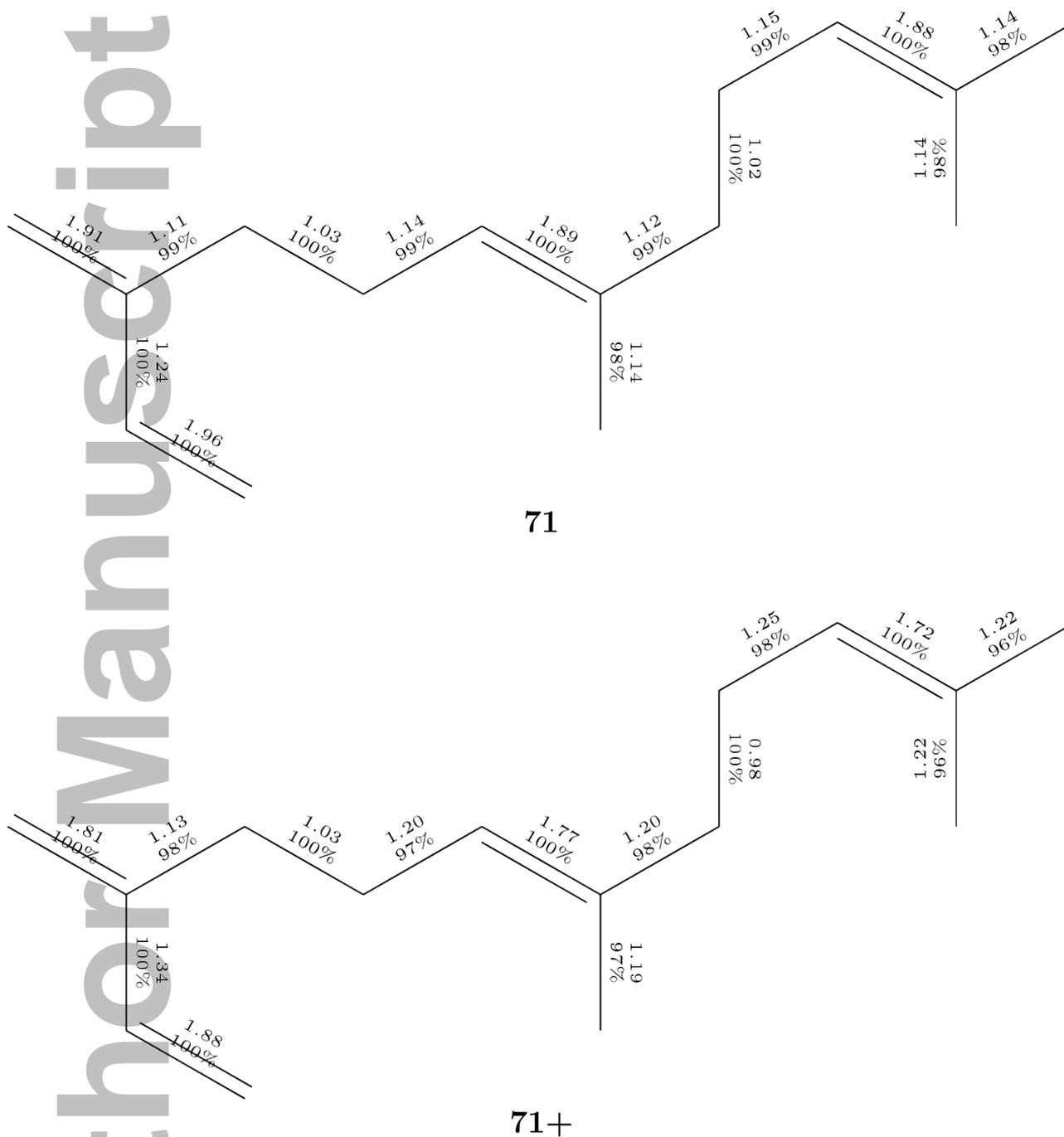Figure S65: Structure of molecule **65** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**65**) and cation (**65+**).

Figure S66: Structure of molecule **66** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**66**) and cation (**66+**).
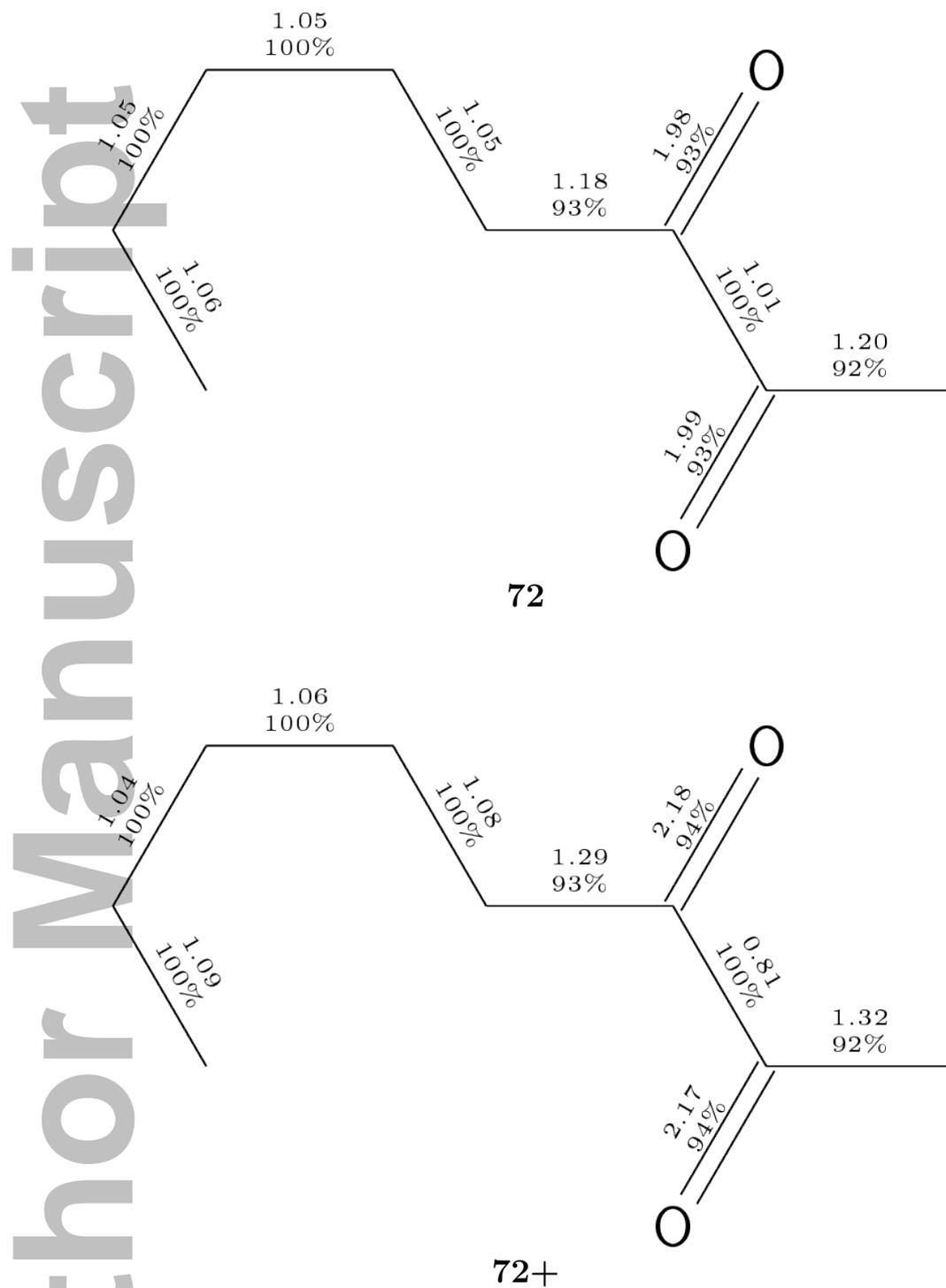
Figure S67: Structure of molecule **67** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**67**) and cation (**67+**).

S3.7.3   Isomers of $C_6H_6O_2$ formula 68-70

Figure S68: Structure of molecule **68** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**68**) and cation (**68+**).
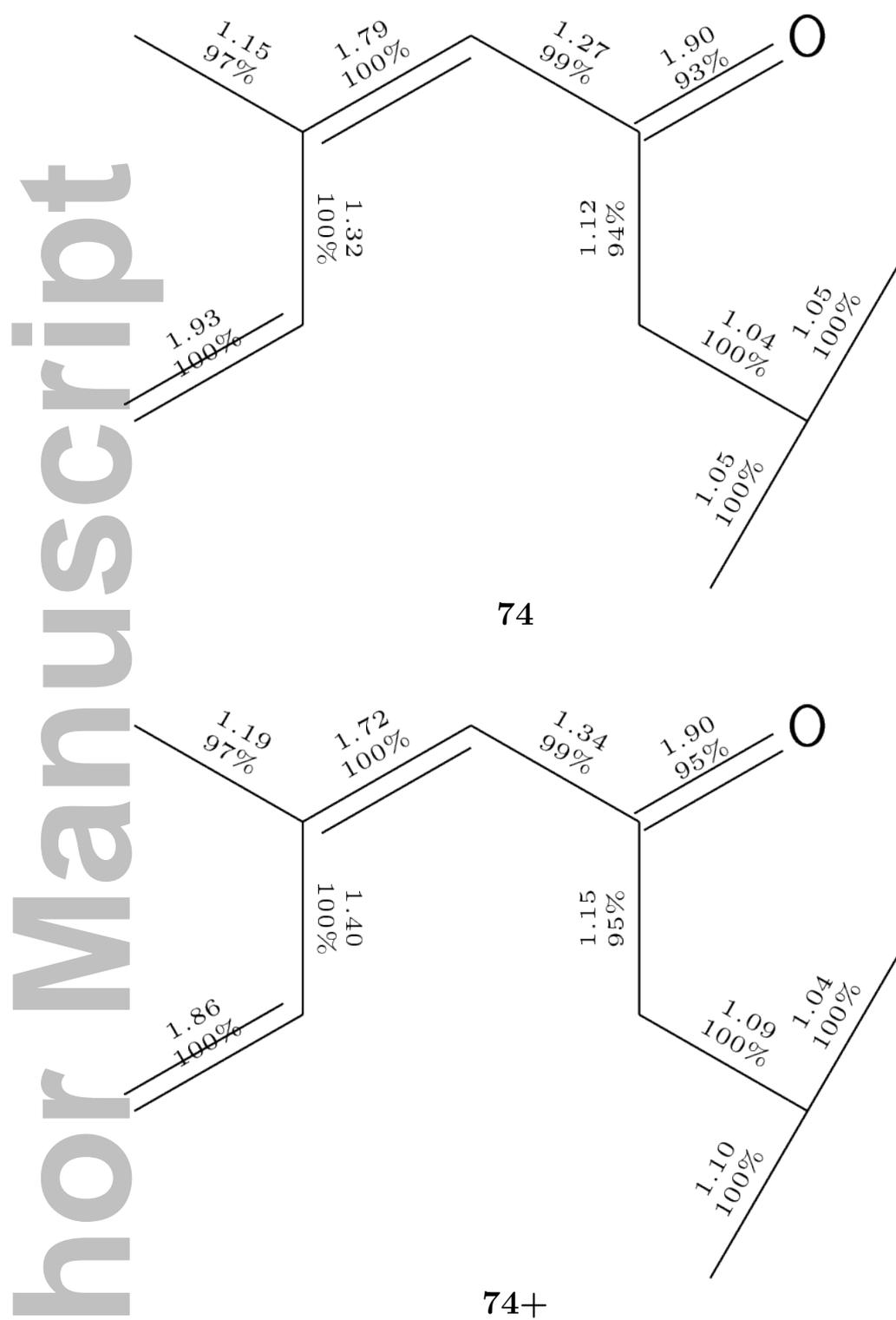
Figure S69: Structure of molecule **69** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**69**) and cation (**69+**).

Figure S70: Structure of molecule **70** and the experimental mass spectrum from NIST. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**70**) and cation (**70+**).
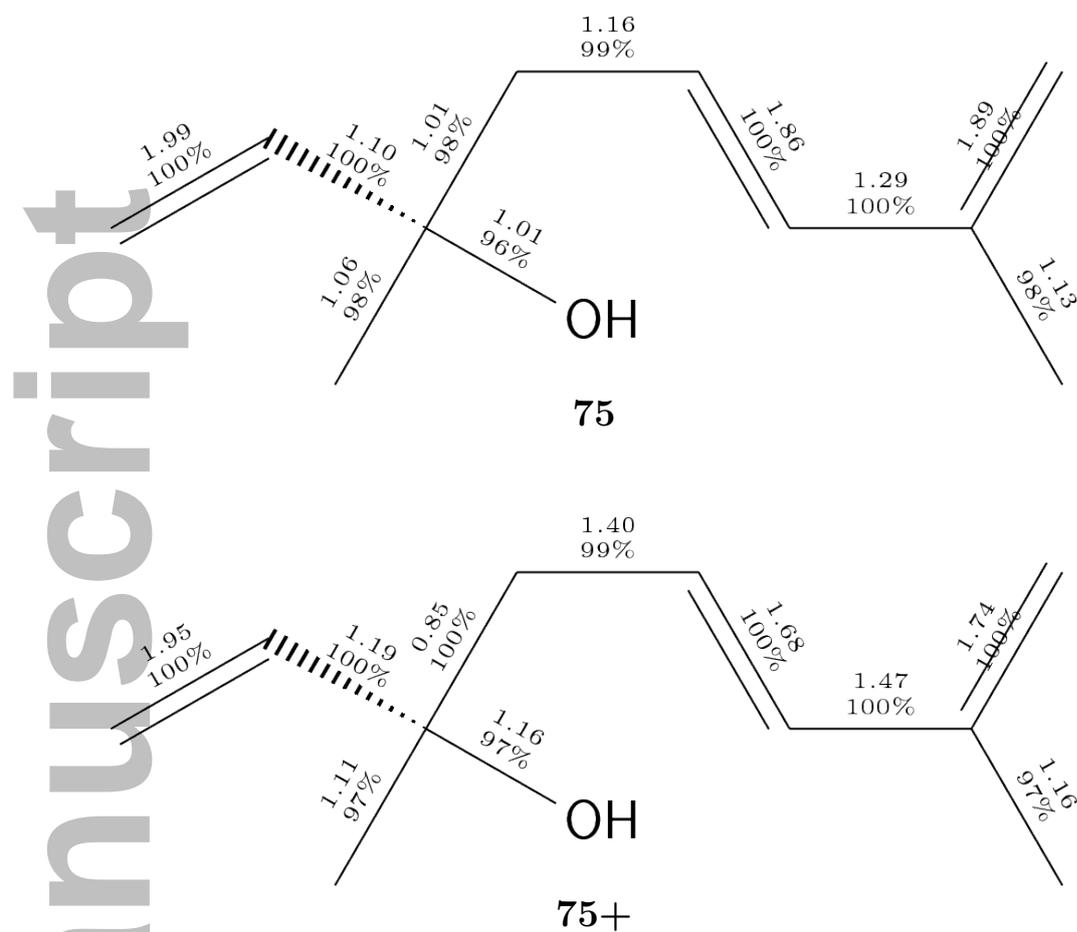
S3.8   Selected natural products

S3.8.1 $C_{15}H_{24}$ formula 71

71

71+

Figure S71: Structure of molecule **71**. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**71**) and cation (**71+**).

S3.8.2 $C_8H_{14}O_2$ formula 72

Figure S72: Structure of molecule **72**. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**72**) and cation (**72+**).
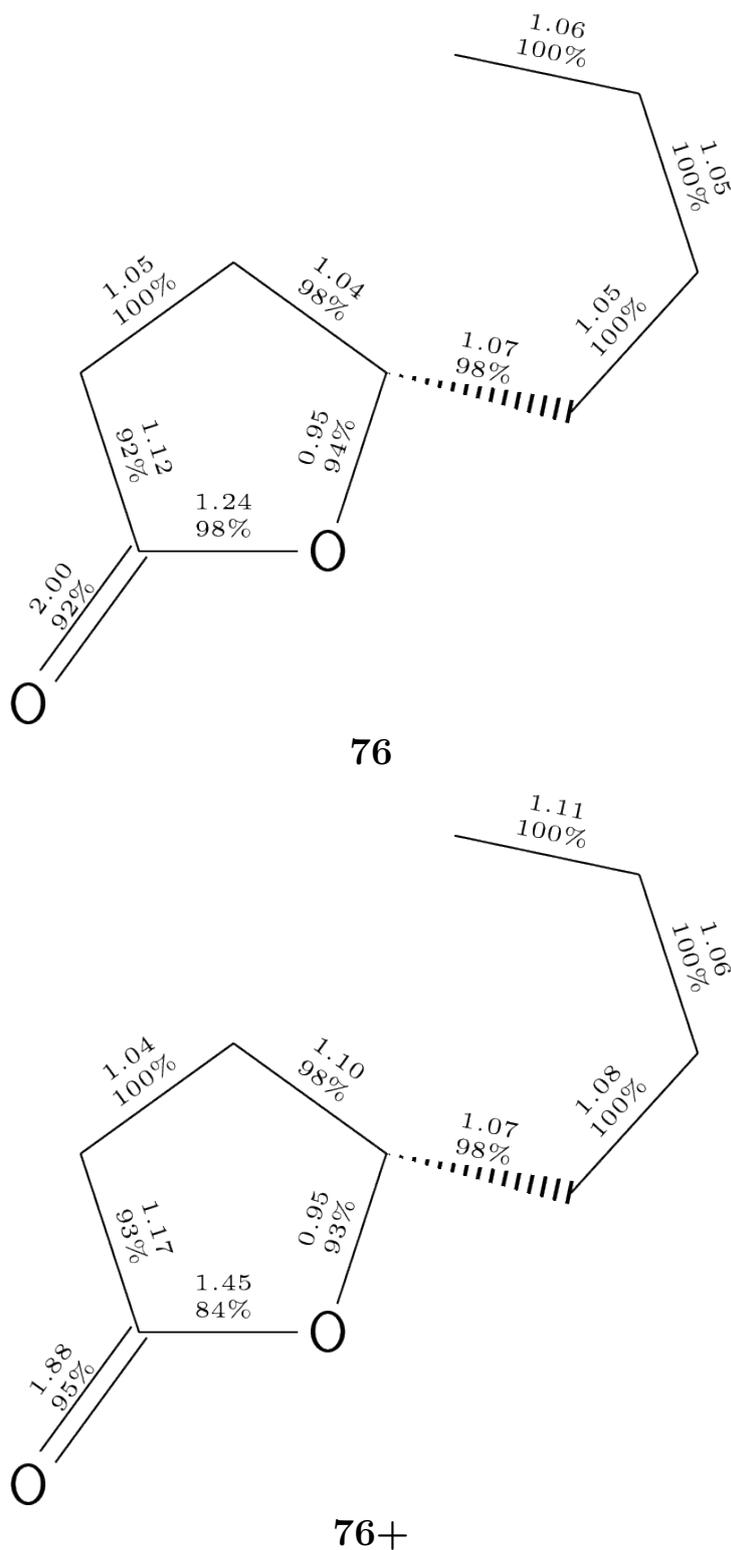
S3.8.3 Isomers of $C_{10}H_{16}O$ formula 73-75
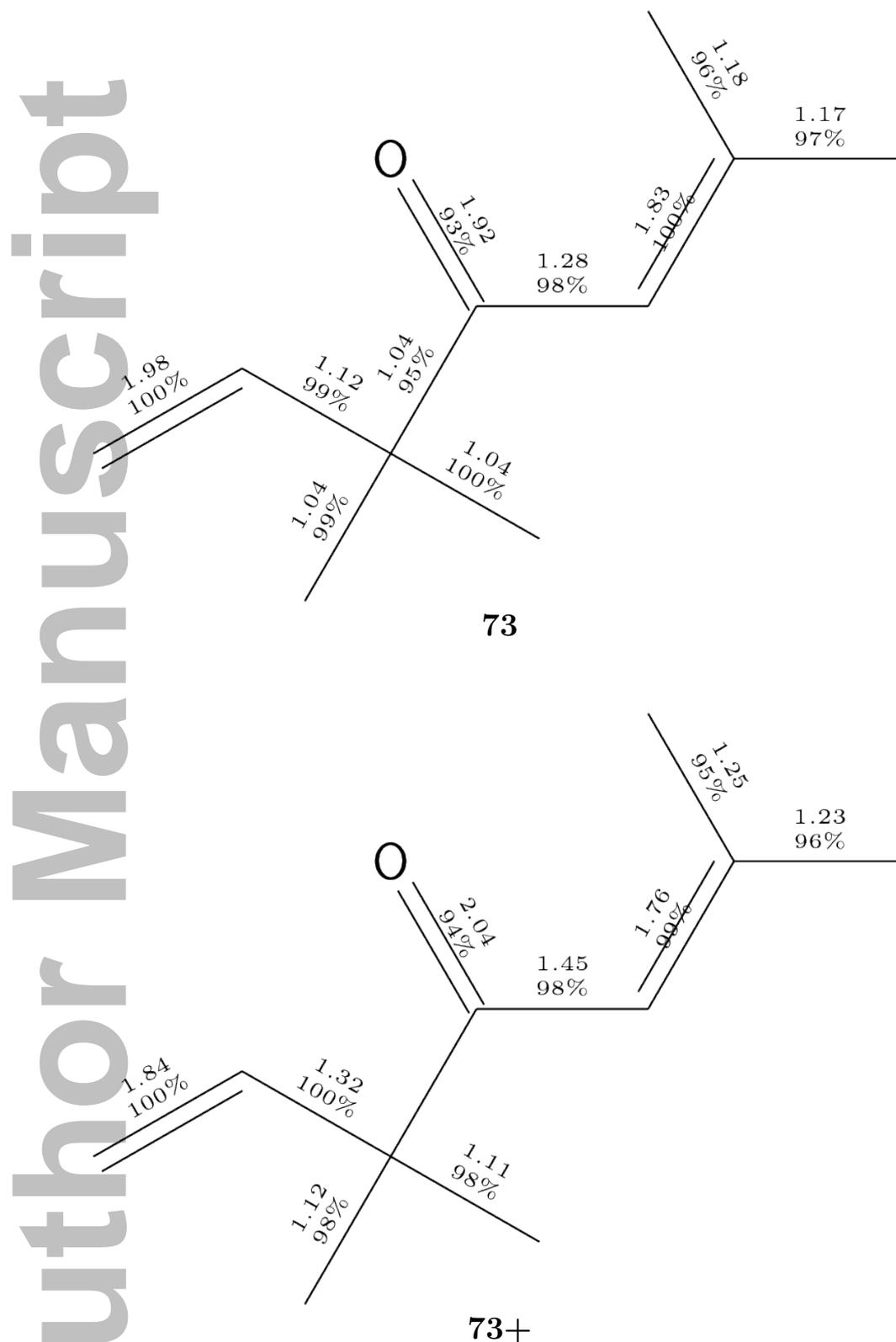
Figure S73: Structure of molecule **73**. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**73**) and cation (**73+**).

Figure S74: Structure of molecule **74**. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**74**) and cation (**74+**).

Figure S75: Structure of molecule **75**. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**75**) and cation (**75+**).

S4   Molecules in data set B

In Table S1 we present the molecular structures of neutral and cations for the data set B which were used to provide Roby-Gould bond indices for a variety of different bond types. They are all isomers of known semiochemical classes, some of them with special names, as detailed below.

- Ethers and carbonyls, **76**-**83**, 8 molecules. These include more examples of C=O bonds and well as C-O-C ether type bonds in pent- and hexane type rings.

- Chiloglottones, **84**-**92**, 9 molecules. These are alkyl double quinones.

- Pyrazines, **93**-**96**, 4 molecules. Alkyl and/or hydroxide substituted pyrazine rings.

- Heptanenitrile derivative, **97**-**100**, 4 molecules. These contain examples of C=N bonds, C=S bonds, with one example of a five-membered ring.

- Unusual ring systems, **101**-**103**, 3 molecules. These carbohydrates contain examples tricyclo ring systems.

It is important to realise that, as for data set A, these figures were generated automatically from the gaussian FChk files using a bash script which called the **openbabel** *program, the* **mol2chemfig** *package, and* LaTeX *program. Therefore some of the layout and drawing of the structures reflect idiosyncracies of this automatic scheme.*
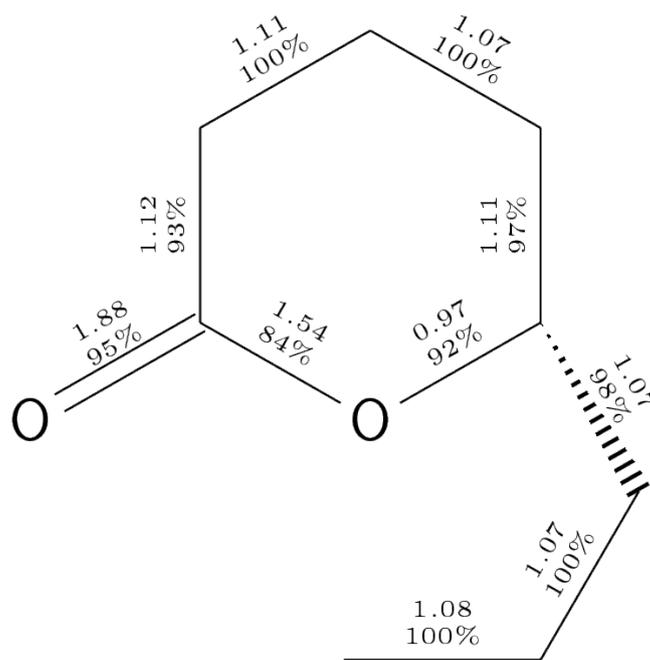
S4.1   Ethers and Carbonyls

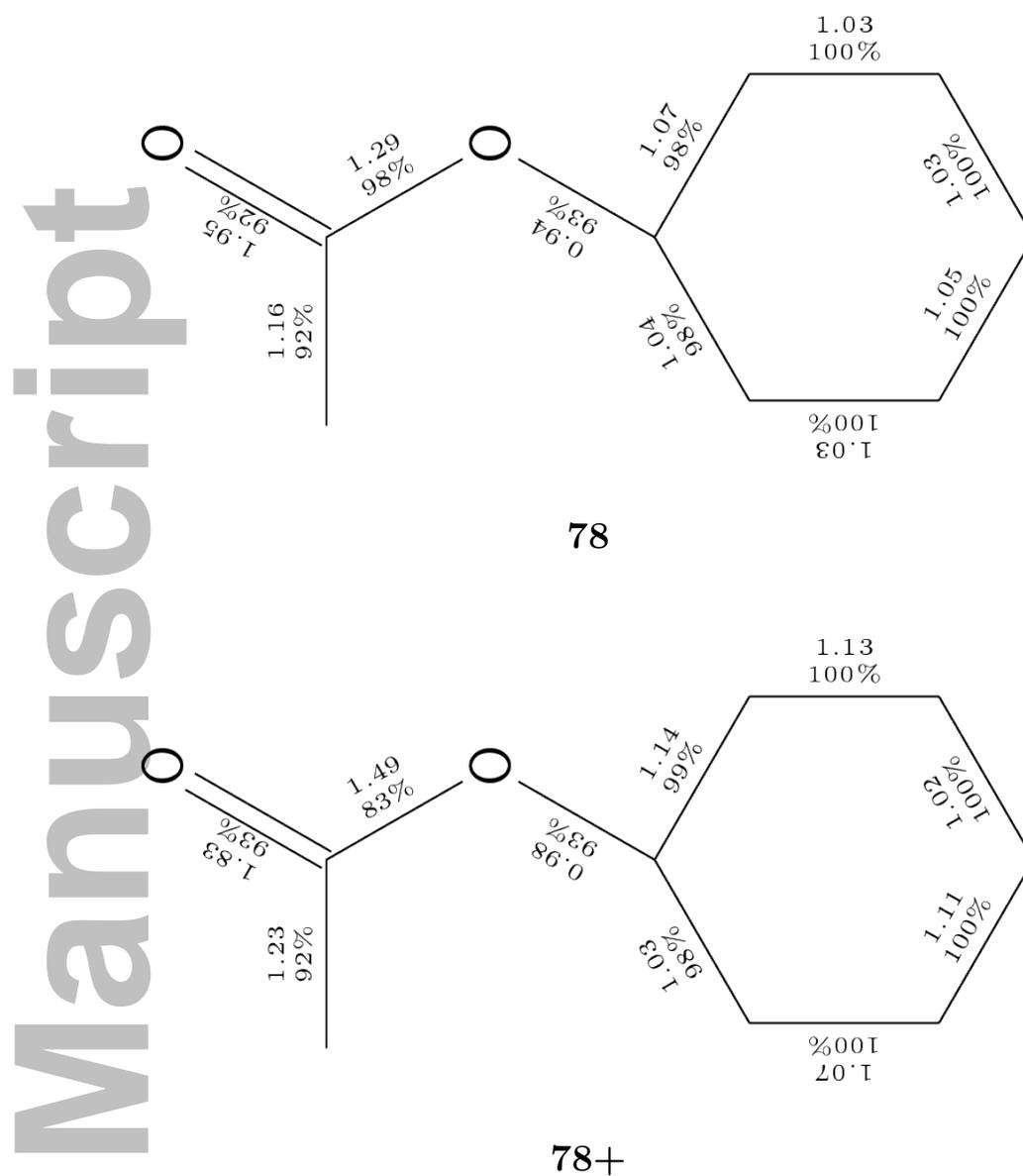### S4.1.1   Isomers of $C_8H_{14}O_2$ formula 76-78

Figure S76: Structure of molecule **76**. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**76**) and cation (**76+**).
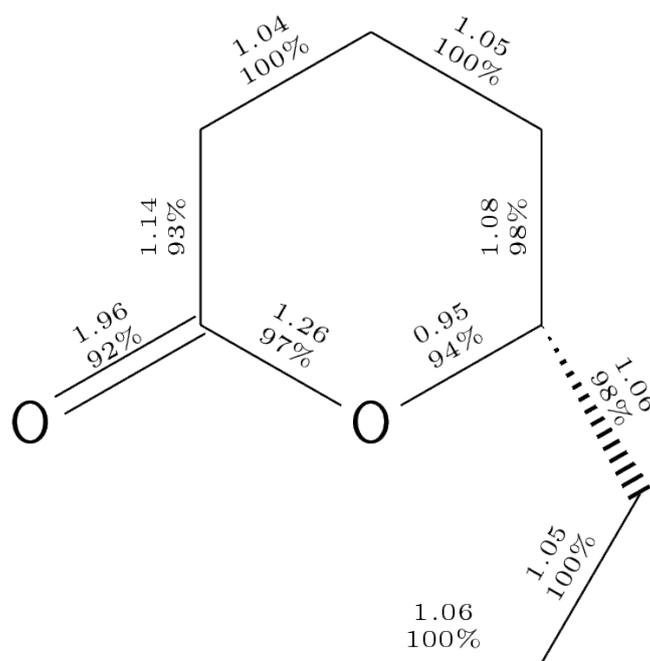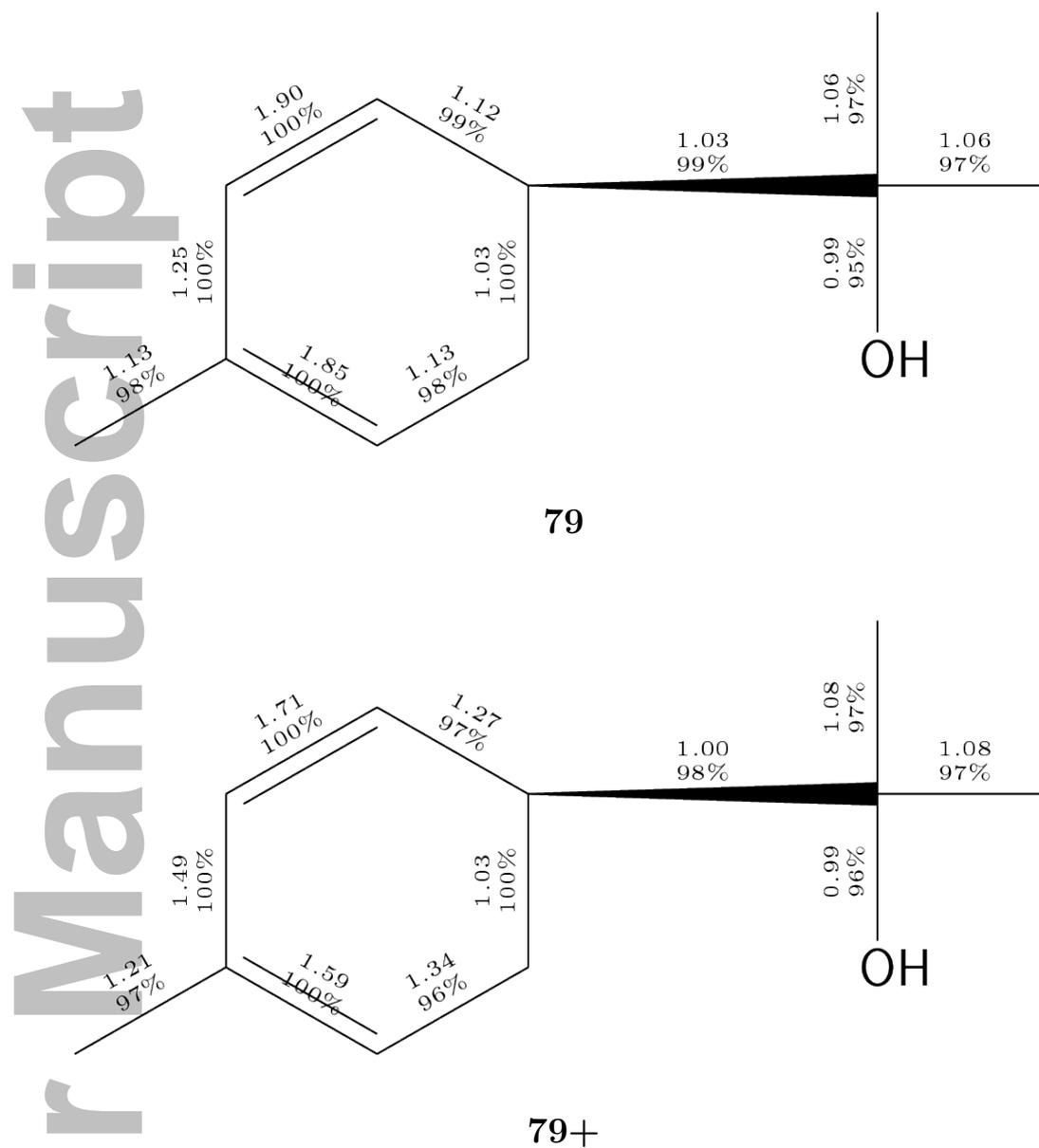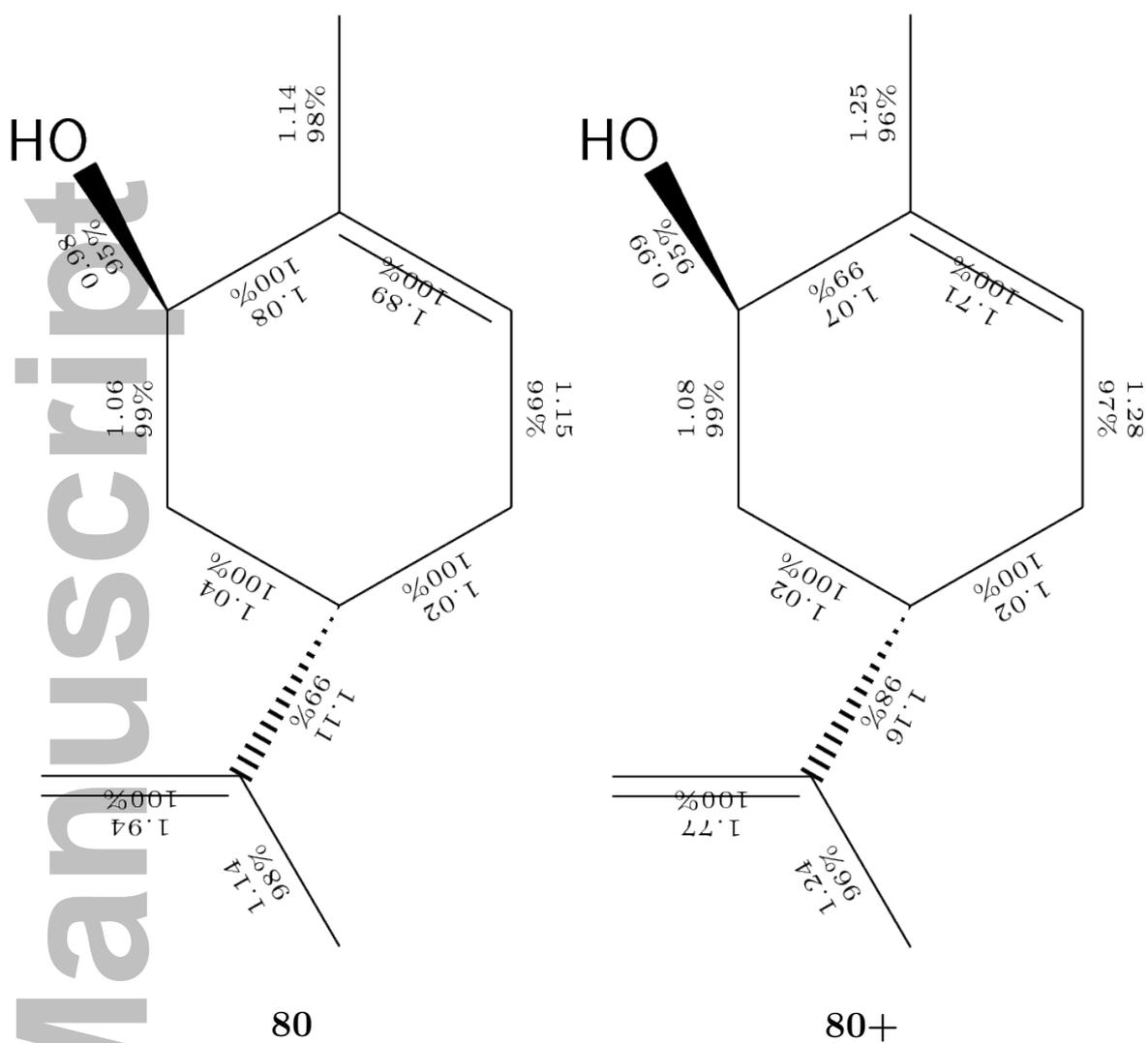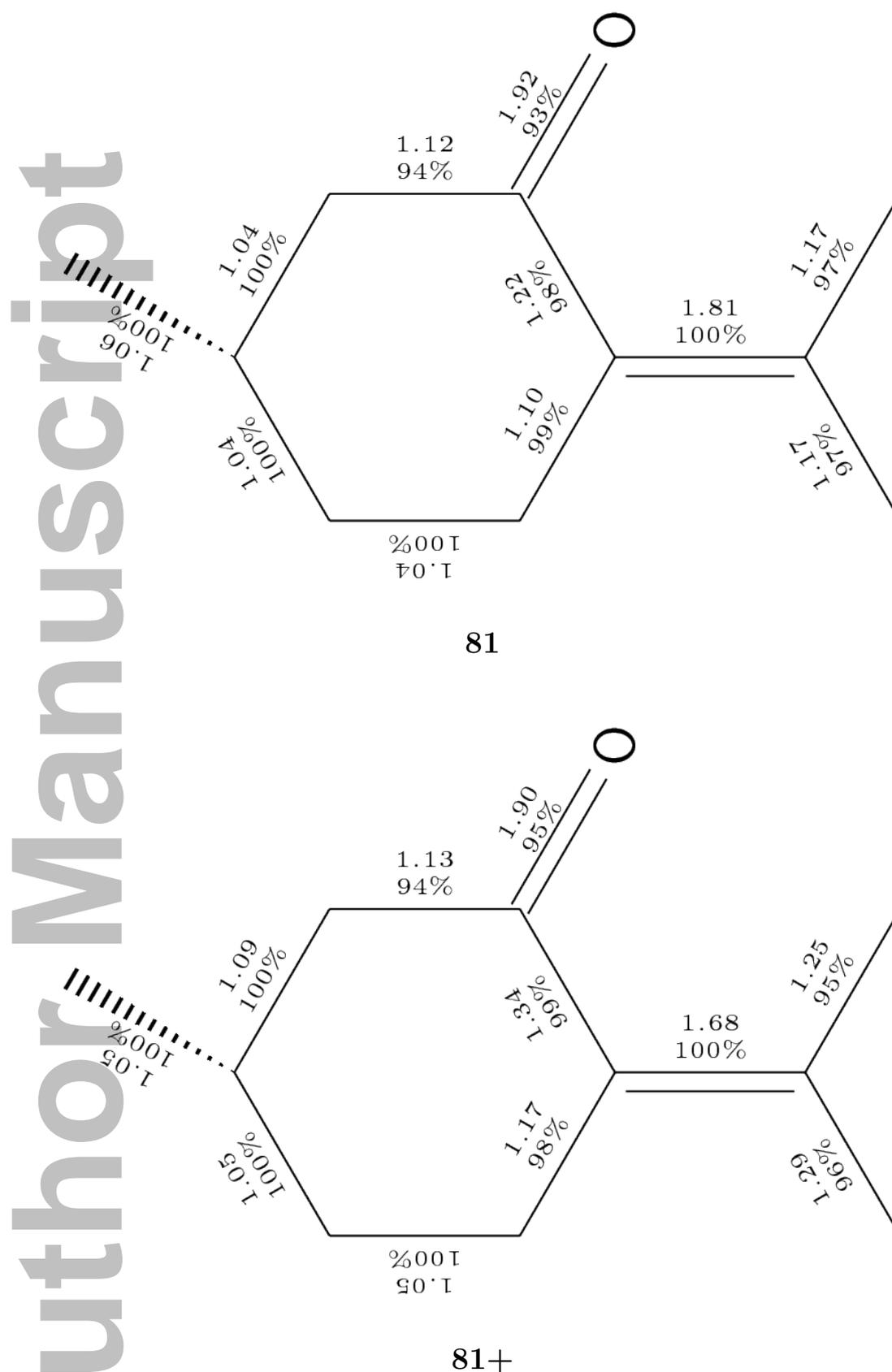
Figure S77: Structure of molecule **77**. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**77**) and cation (**77+**).
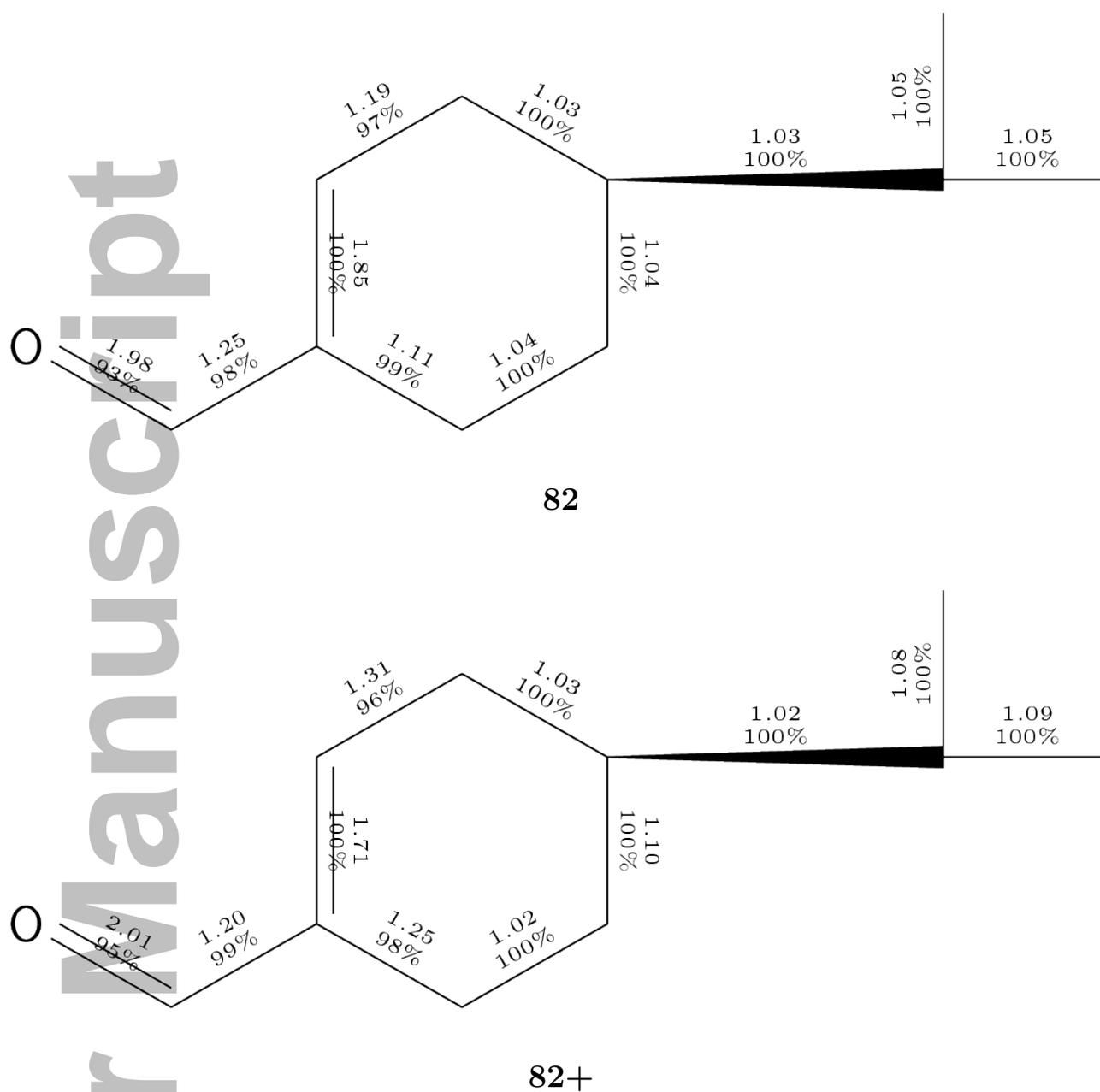
78

78+

Figure S78: Structure of molecule **78**. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**78**) and cation (**78+**).

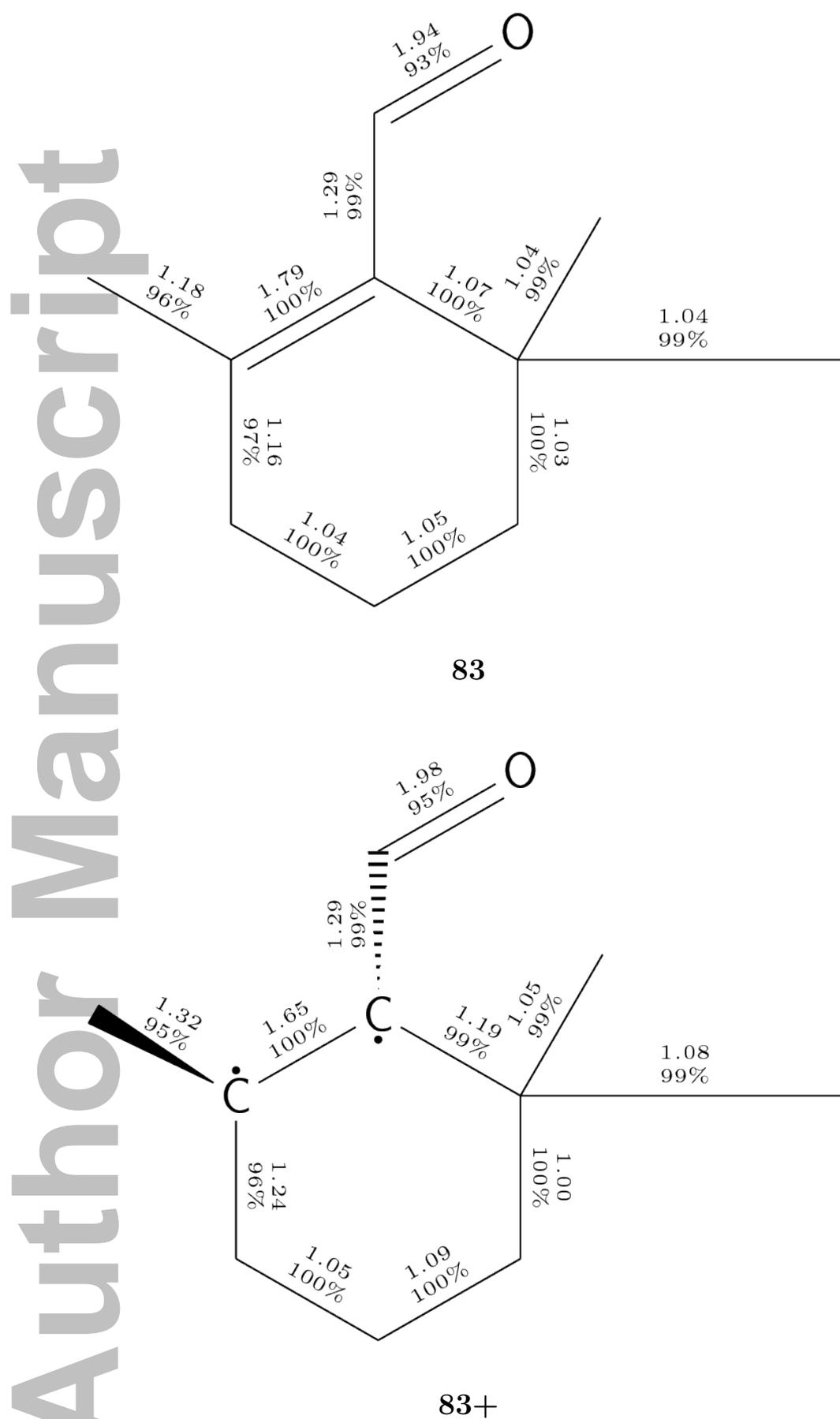### S4.1.2 Isomers of $C_{10}H_{16}O$ formula 79-83

79

79+

Figure S79: Structure of molecule **79**. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**79**) and cation (**79+**).

HO

1.14 98%

0.96 98%
0.98

1.08 100%

1.89 100%

1.06 99%

1.15 99%

1.04 100%

1.02 100%

1.11 99%

1.94 100%

1.14 98%

80

HO

1.25 96%

0.99 95%
0.99

1.07 99%

1.71 100%

1.08 99%

1.28 97%

1.02 100%

1.02 100%

1.16 98%

1.77 100%

1.24 96%

80+

Figure S80: Structure of molecule **80**. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**80**) and cation (**80+**).

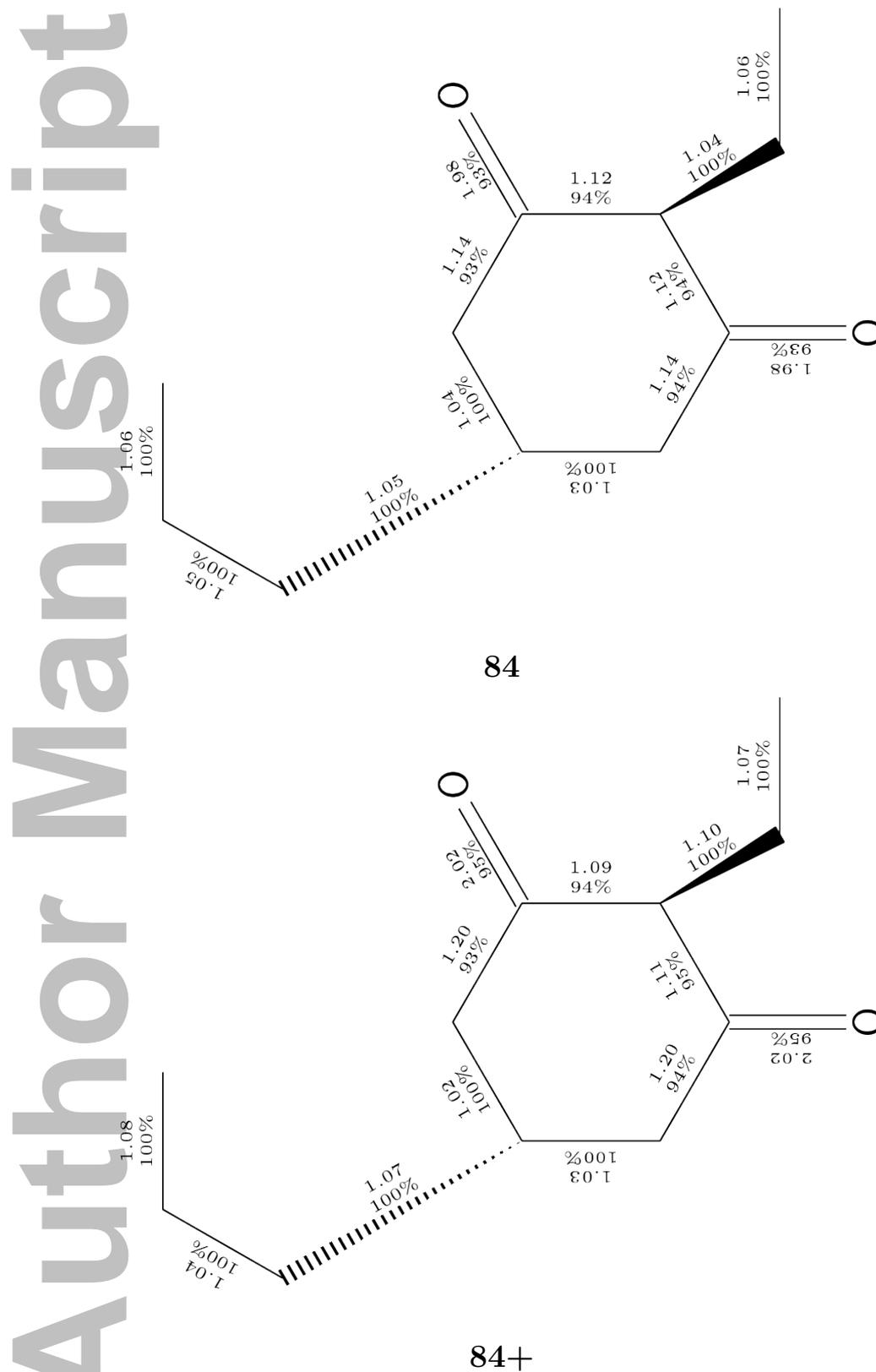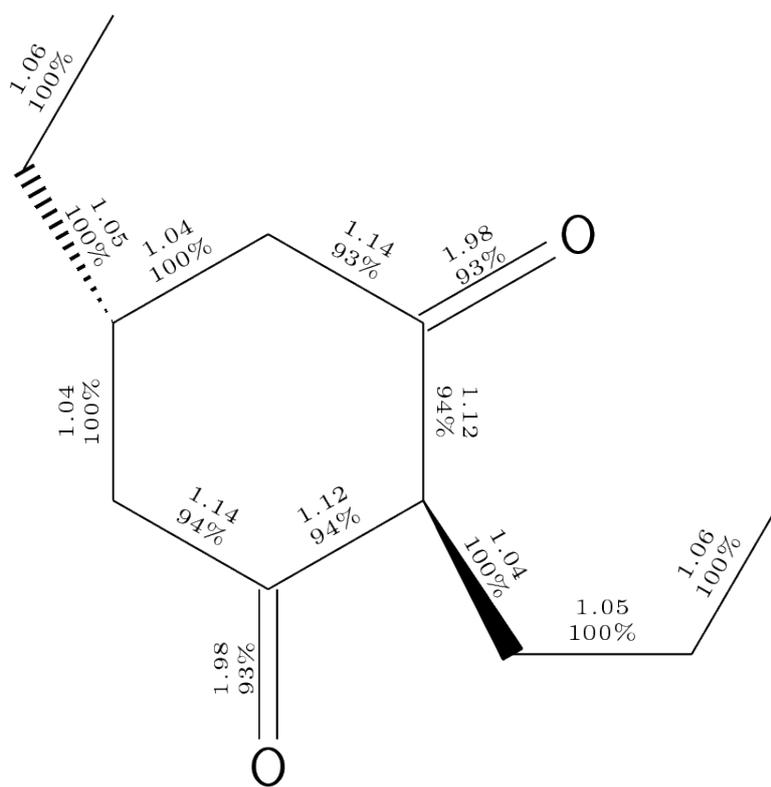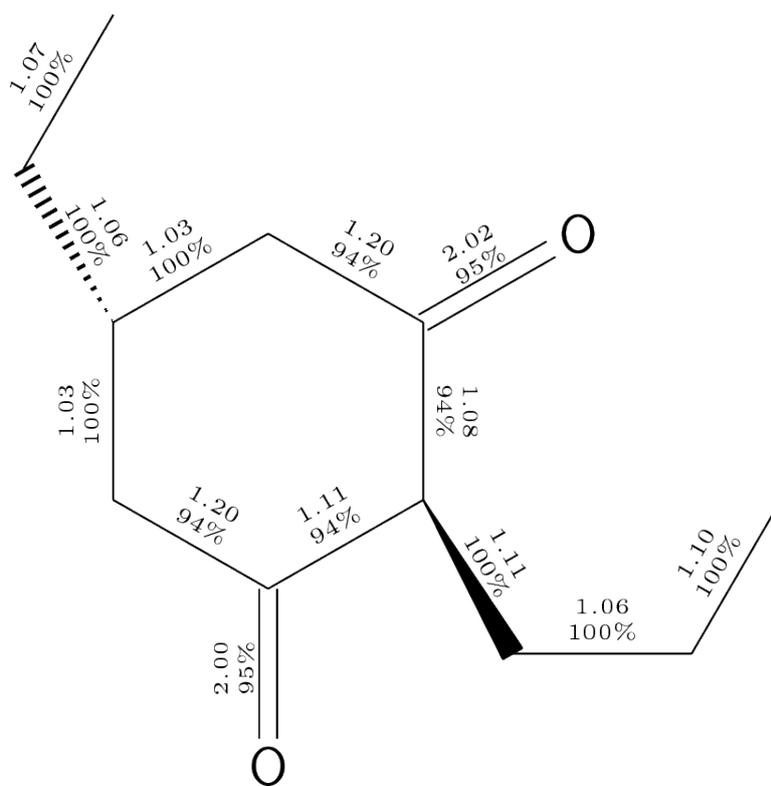Figure S81: Structure of molecule **81**. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**81**) and cation (**81+**).

Figure S82: Structure of molecule **82**. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**82**) and cation (**82+**).
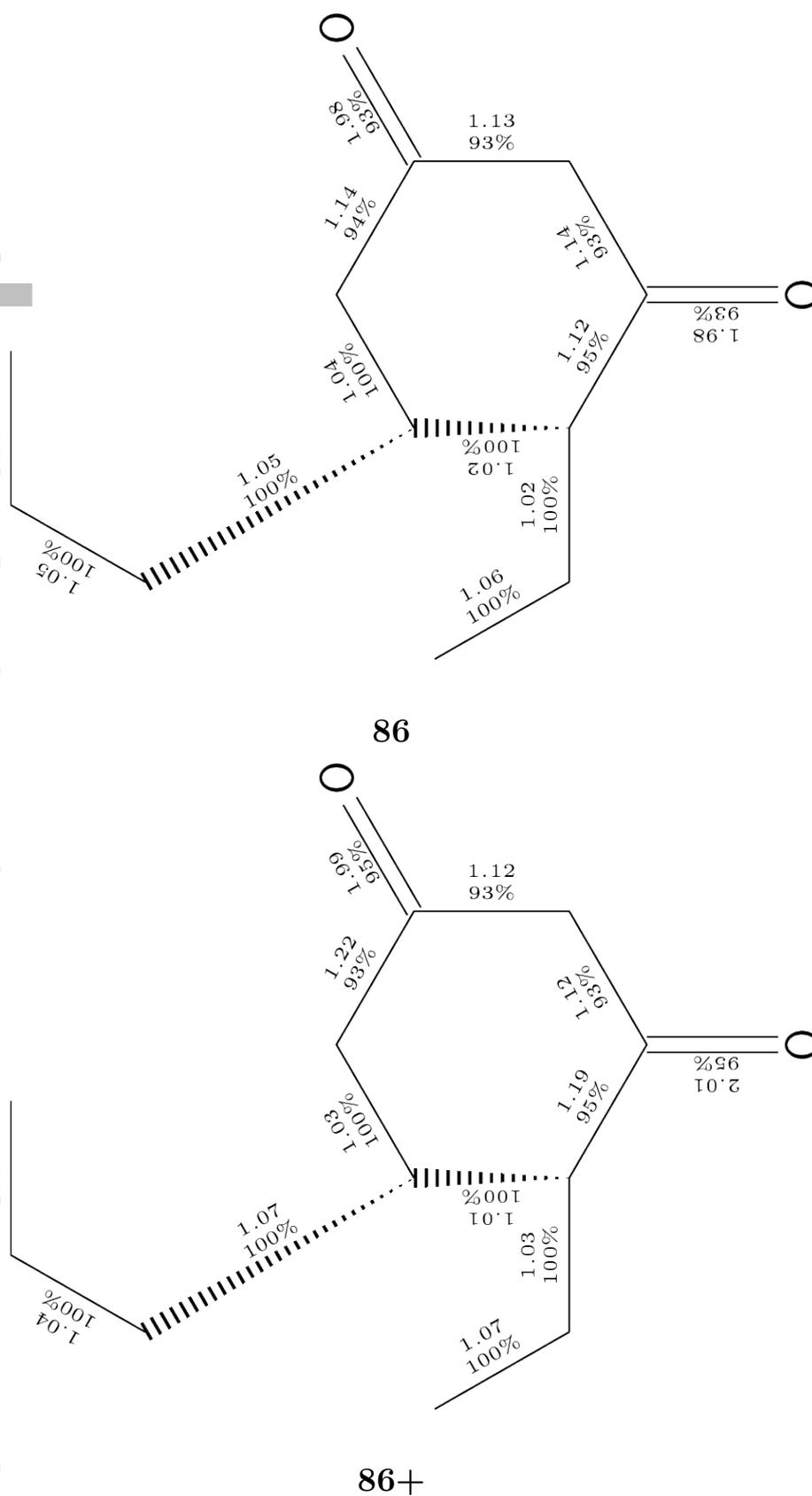
Figure S83: Structure of molecule **83**. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**83**) and cation (**83+**).

S4.2 Chiloglottones

S4.2.1 Isomers of $C_{11}H_{18}O_2$ formula 84-89

84

84+

Figure S84: Structure of molecule **84**. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**84**) and cation (**84+**).
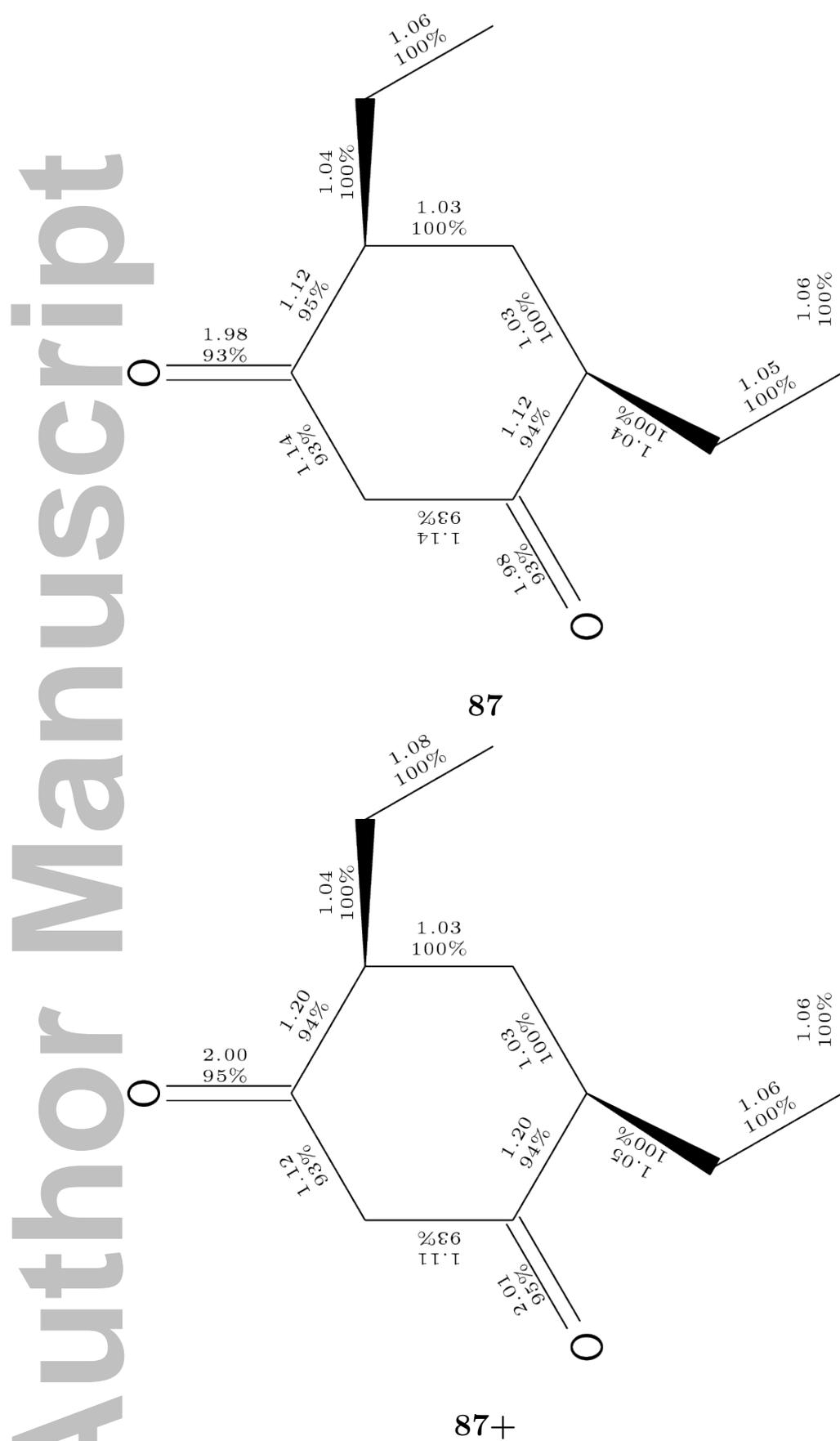
85

85+

Figure S85: Structure of molecule **85**. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**85**) and cation (**85+**).
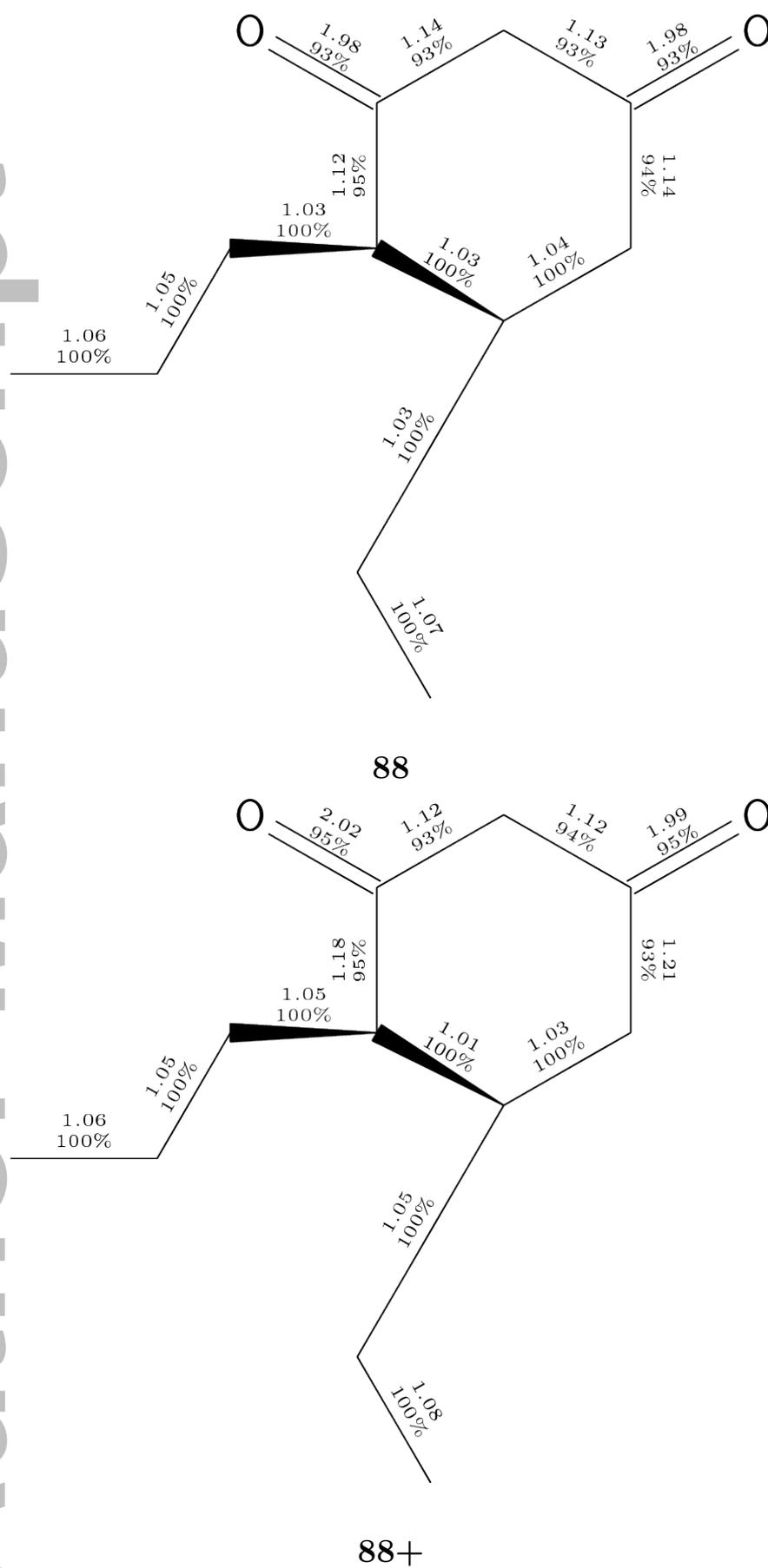
Figure S86: Structure of molecule **86**. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**86**) and cation (**86+**).
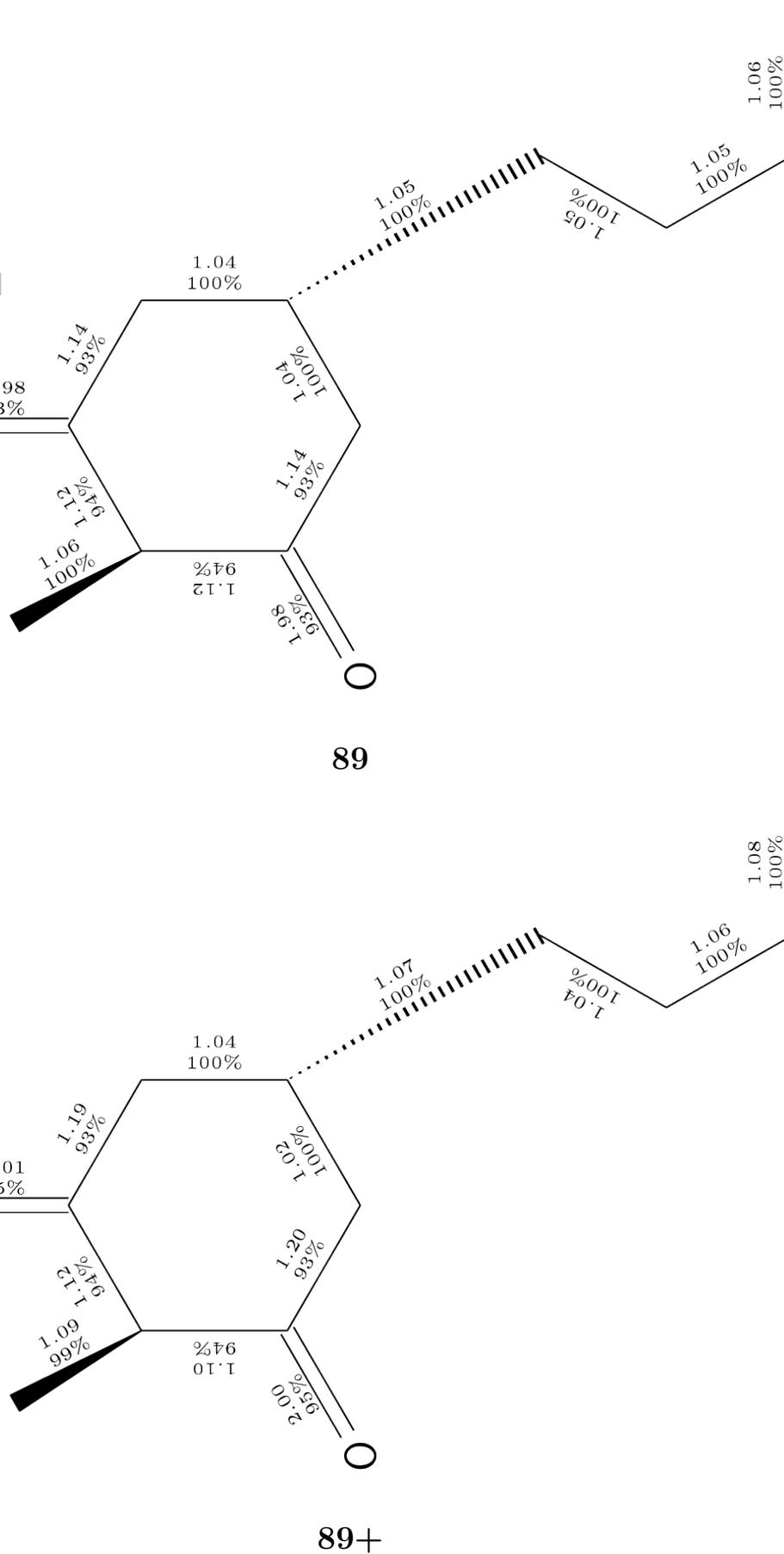
Figure S87: Structure of molecule **87**. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**87**) and cation (**87+**).

88

88+

Figure S88: Structure of molecule **88**. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**88**) and cation (**88+**).

Figure S89: Structure of molecule **89**. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**89**) and cation (**89+**).
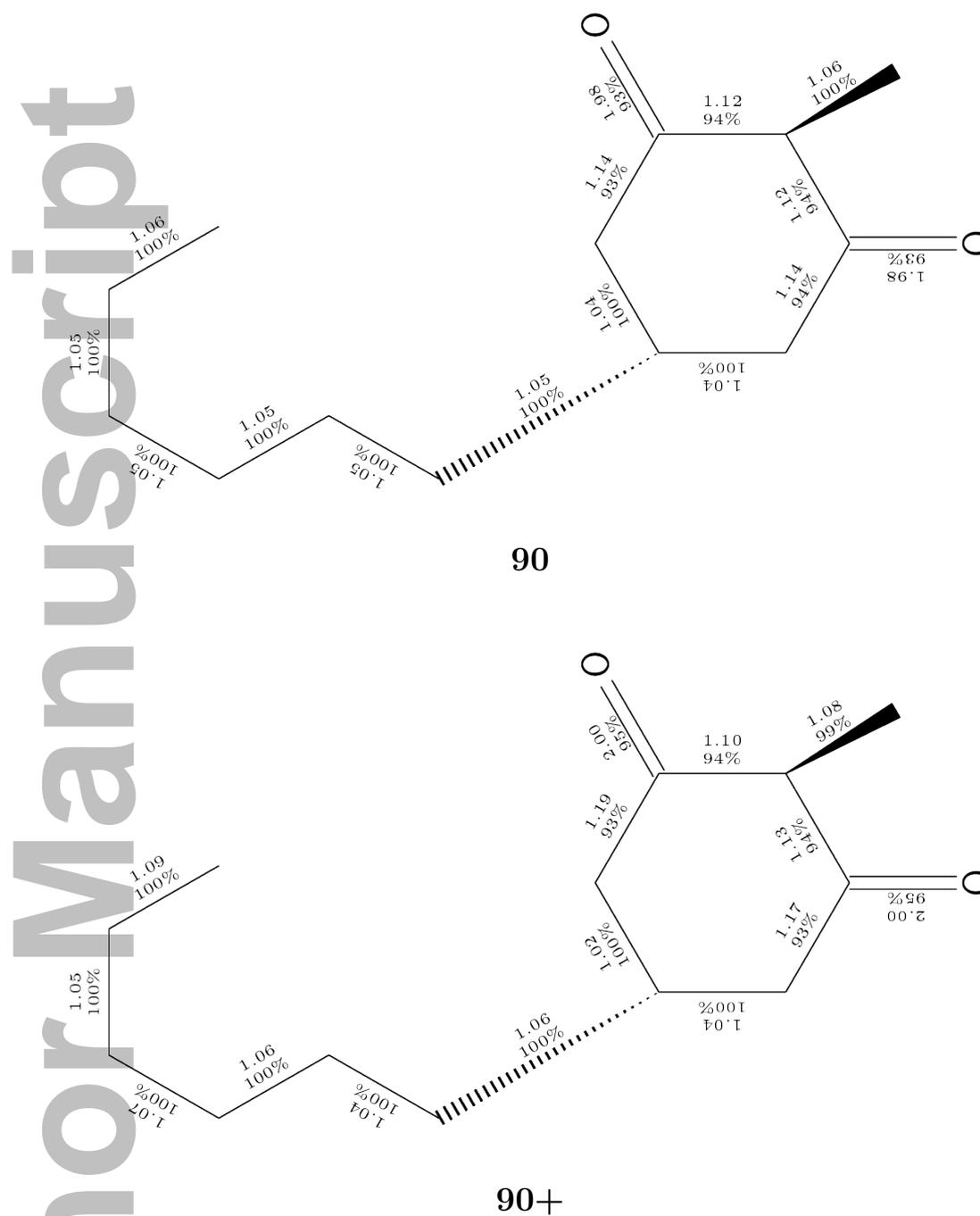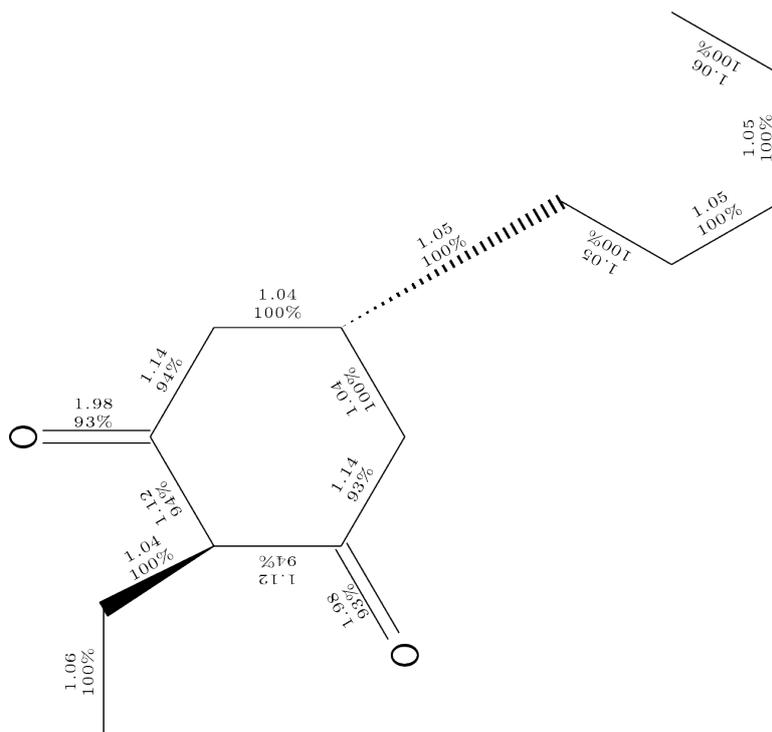
S4.2.2 Isomers of $C_{13}H_{22}O_2$ formula 90-92

90

90+

Figure S90: Structure of molecule **90**. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**90**) and cation (**90+**).
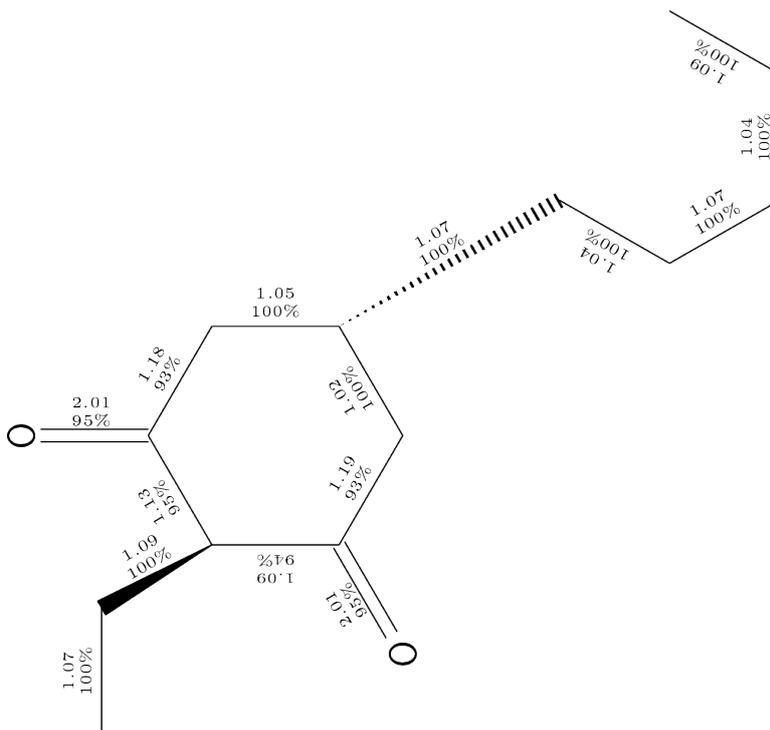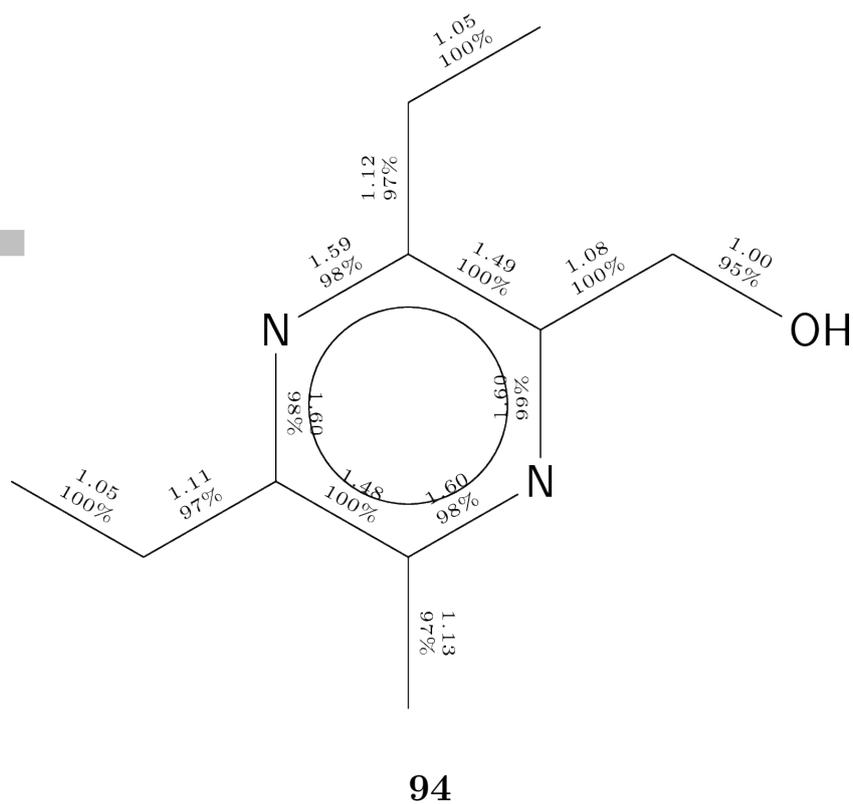
91

91+

Figure S91: Structure of molecule **91**. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**91**) and cation (**91+**).

92

92+

Figure S92: Structure of molecule **92**. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**92**) and cation (**92+**).

S4.3   Pyrazines

S4.3.1   Isomers of $C_9H_{14}N_2$ formula 93

93

93+

Figure S93: Structure of molecule **93**. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**93**) and cation (**93+**).
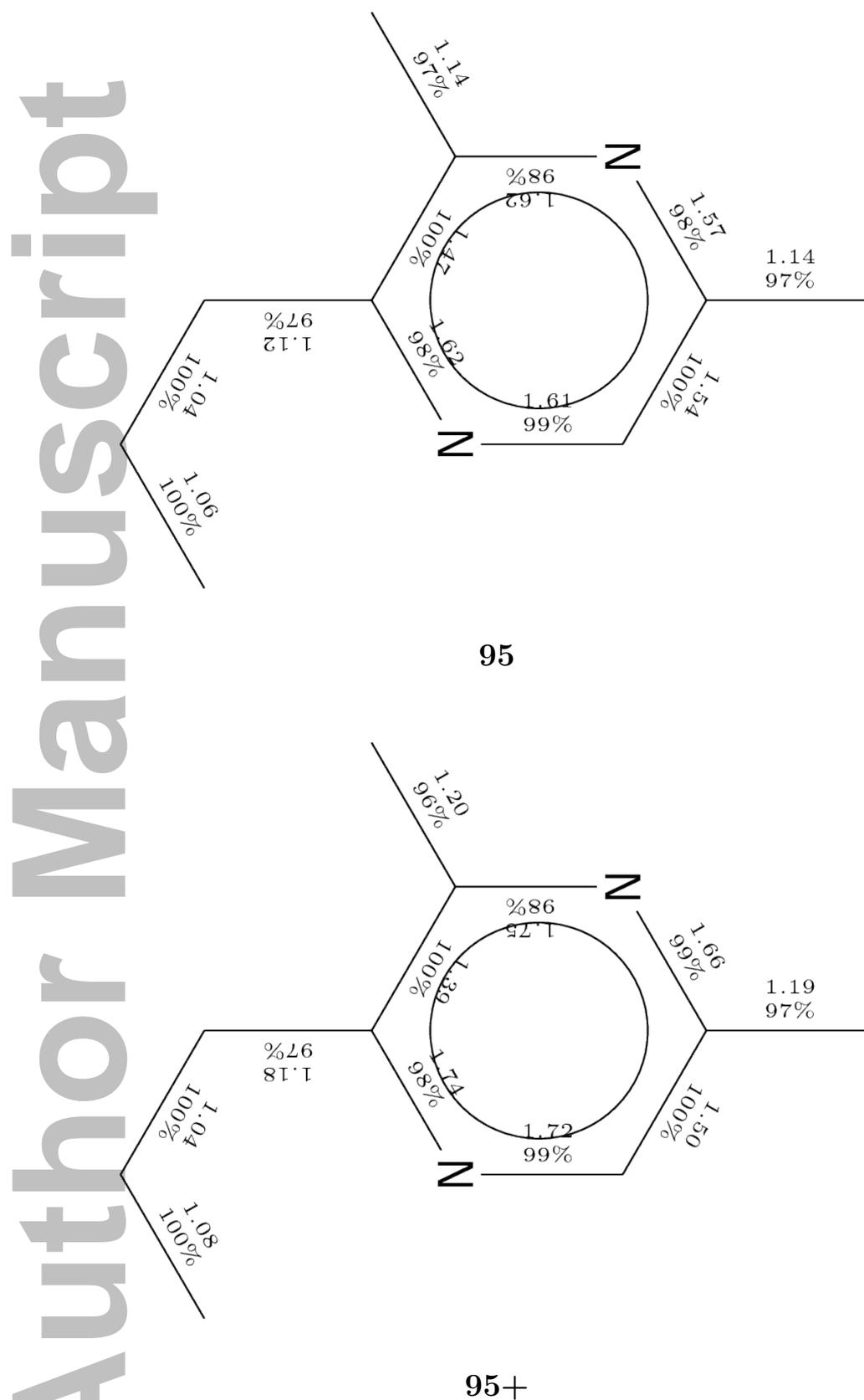
S4.3.2   Isomers of $C_{11}H_{18}N_2$ formula 94

94

94+

Figure S94: Structure of molecule **94**. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**94**) and cation (**94+**).
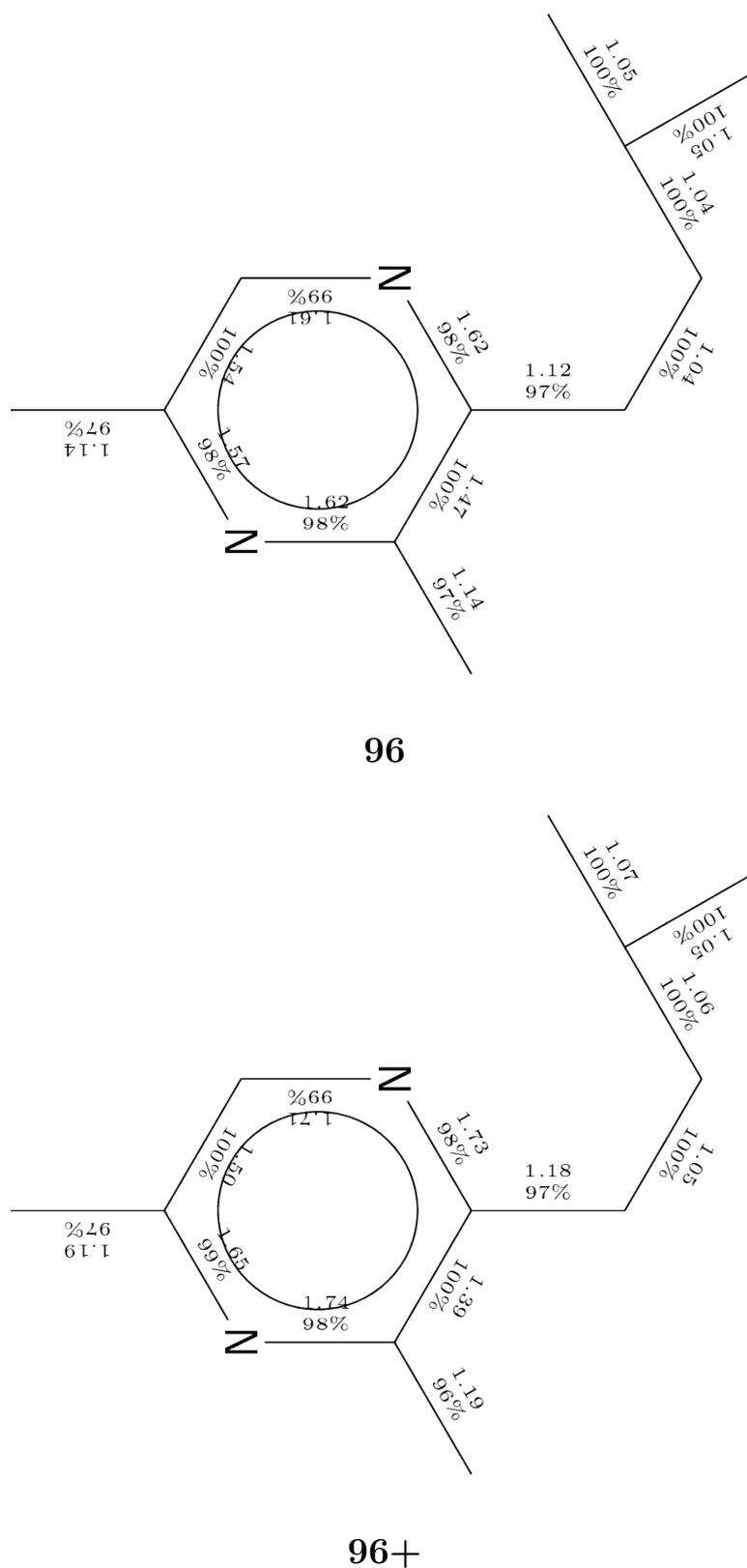
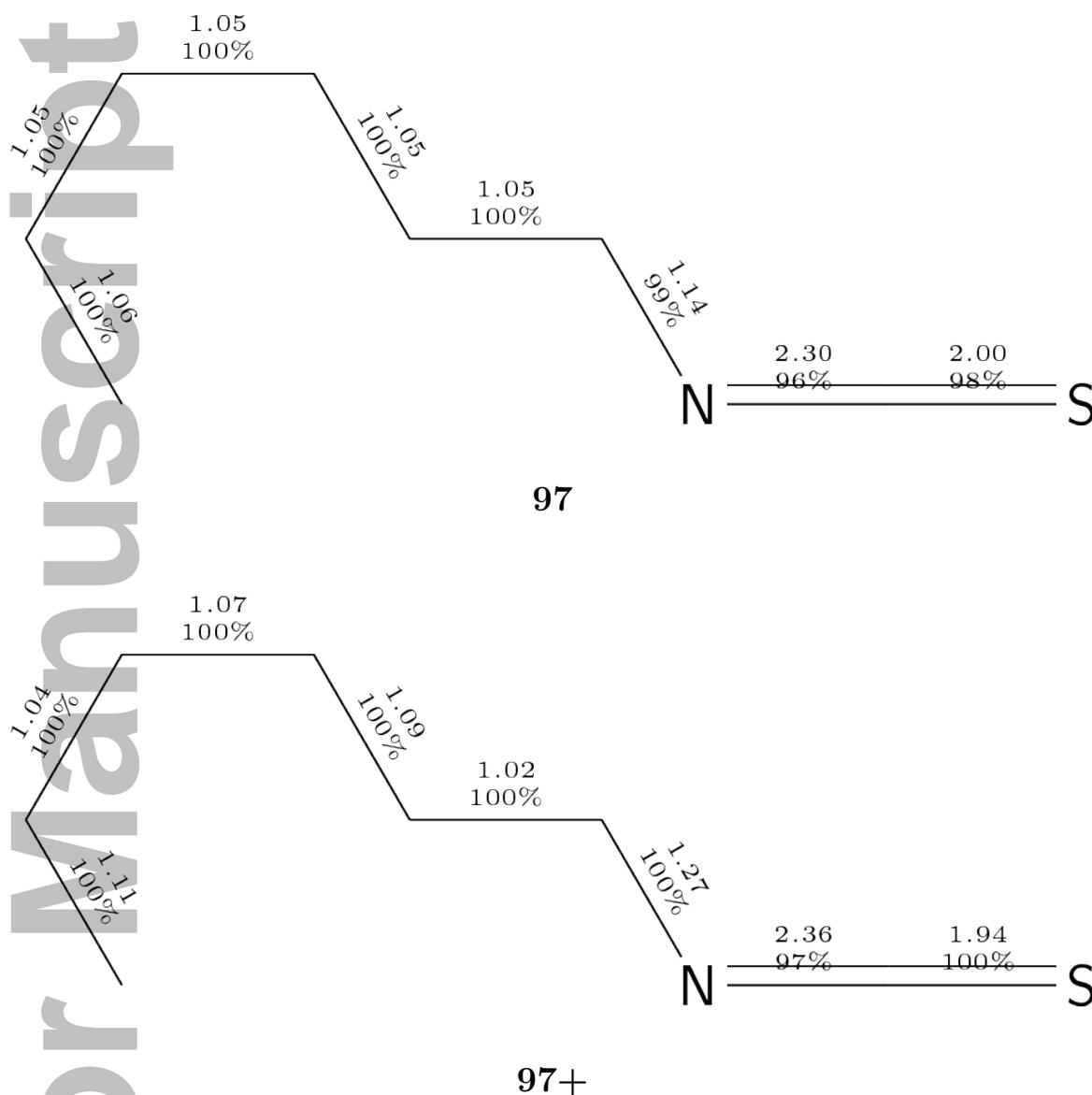### S4.3.3　Isomers of $C_{10}H_{16}N_2O$ formula 95

95

95+

Figure S95: Structure of molecule **95**. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**95**) and cation (**95+**).

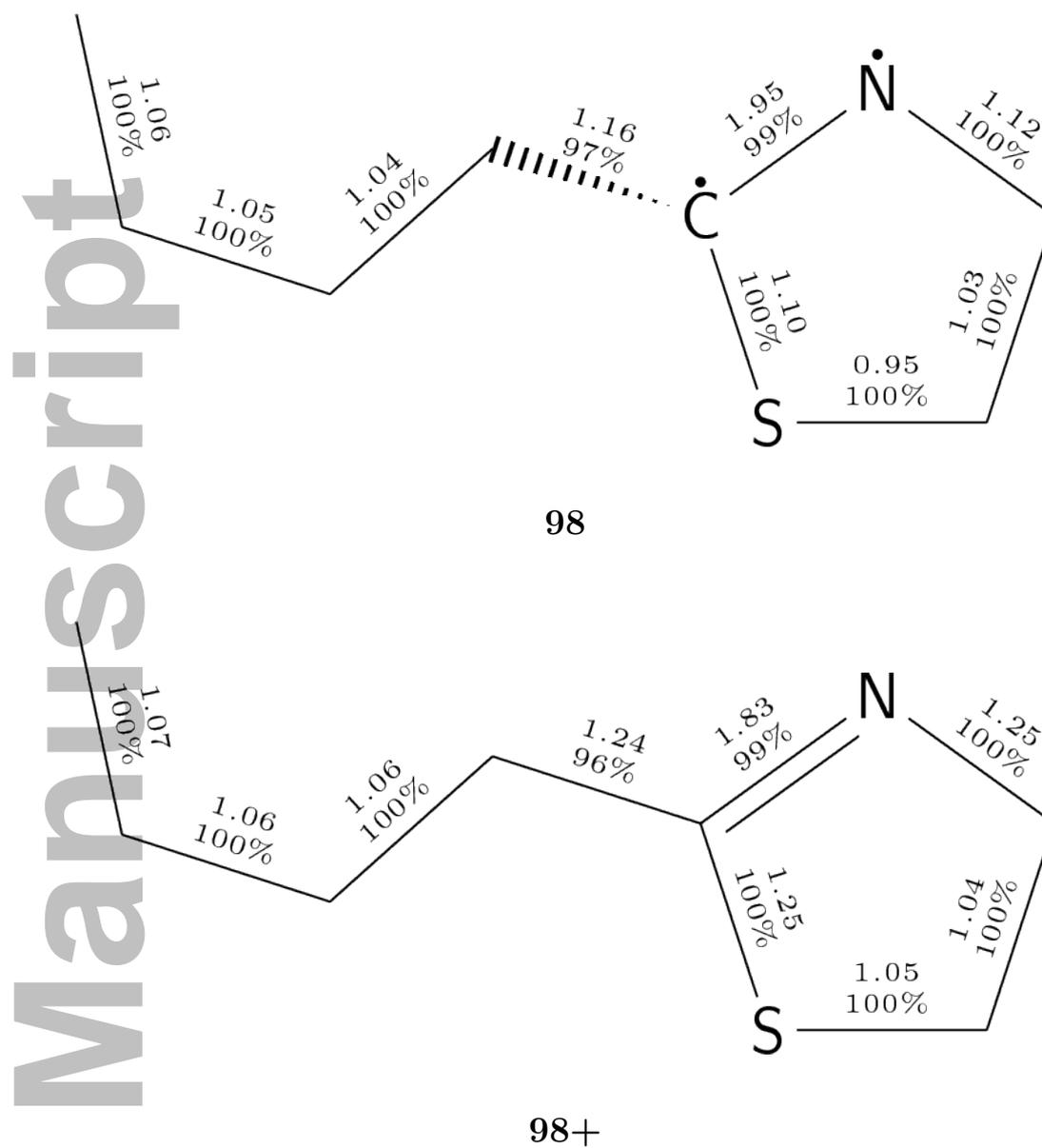### S4.3.4   Isomers of $C_{11}H_{18}N_2O$ formula 96

96

96+

Figure S96: Structure of molecule **96**. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**96**) and cation (**96+**).

S4.4 Heptanenitrile derivatives

S4.4.1 Isomers of $C_7H_{13}NS$ formula 97-100

Figure S97: Structure of molecule **97**. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**97**) and cation (**97+**).
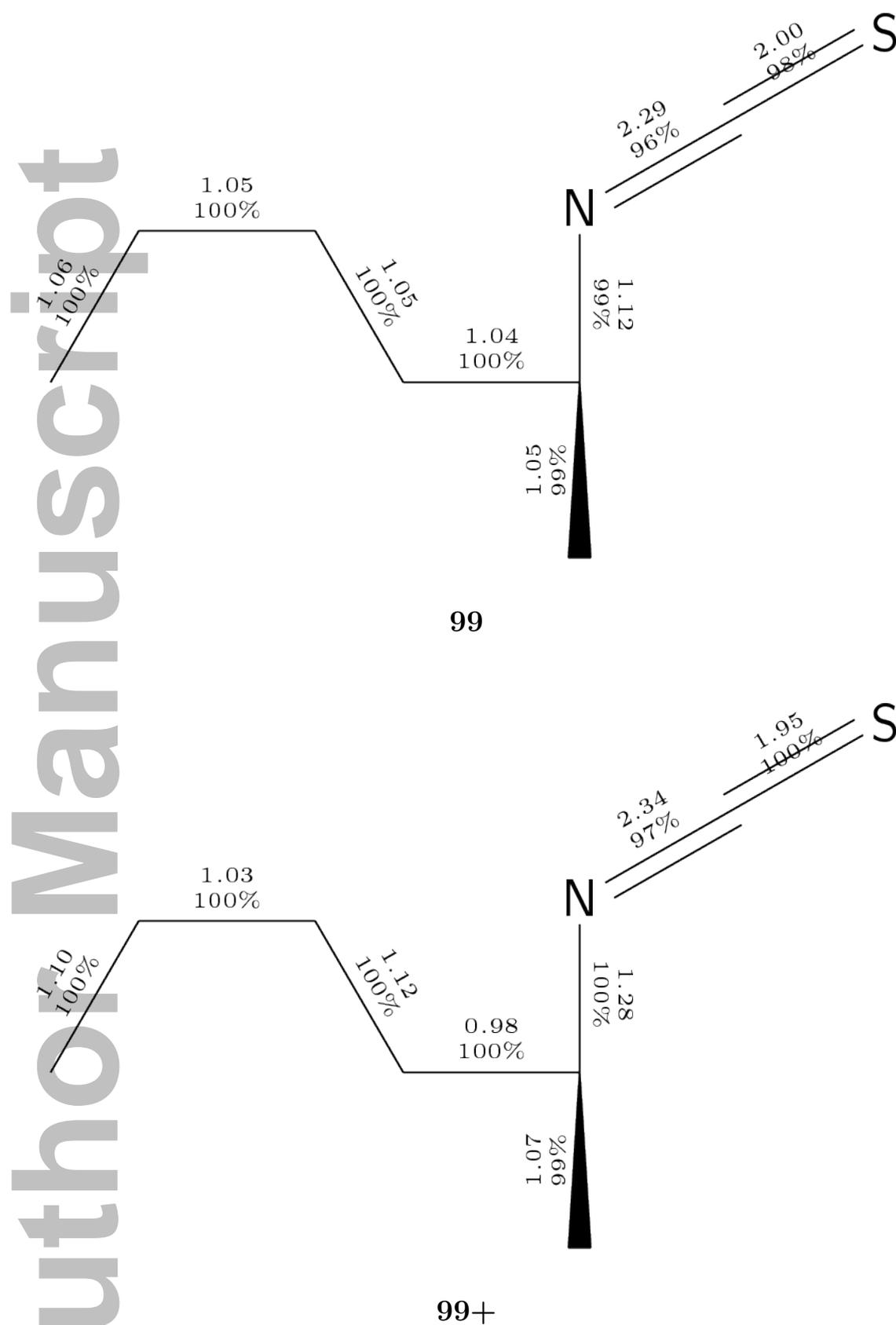
98

98+

Figure S98: Structure of molecule **98**. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**98**) and cation (**98+**).

Figure S99: Structure of molecule **99**. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**99**) and cation (**99+**).
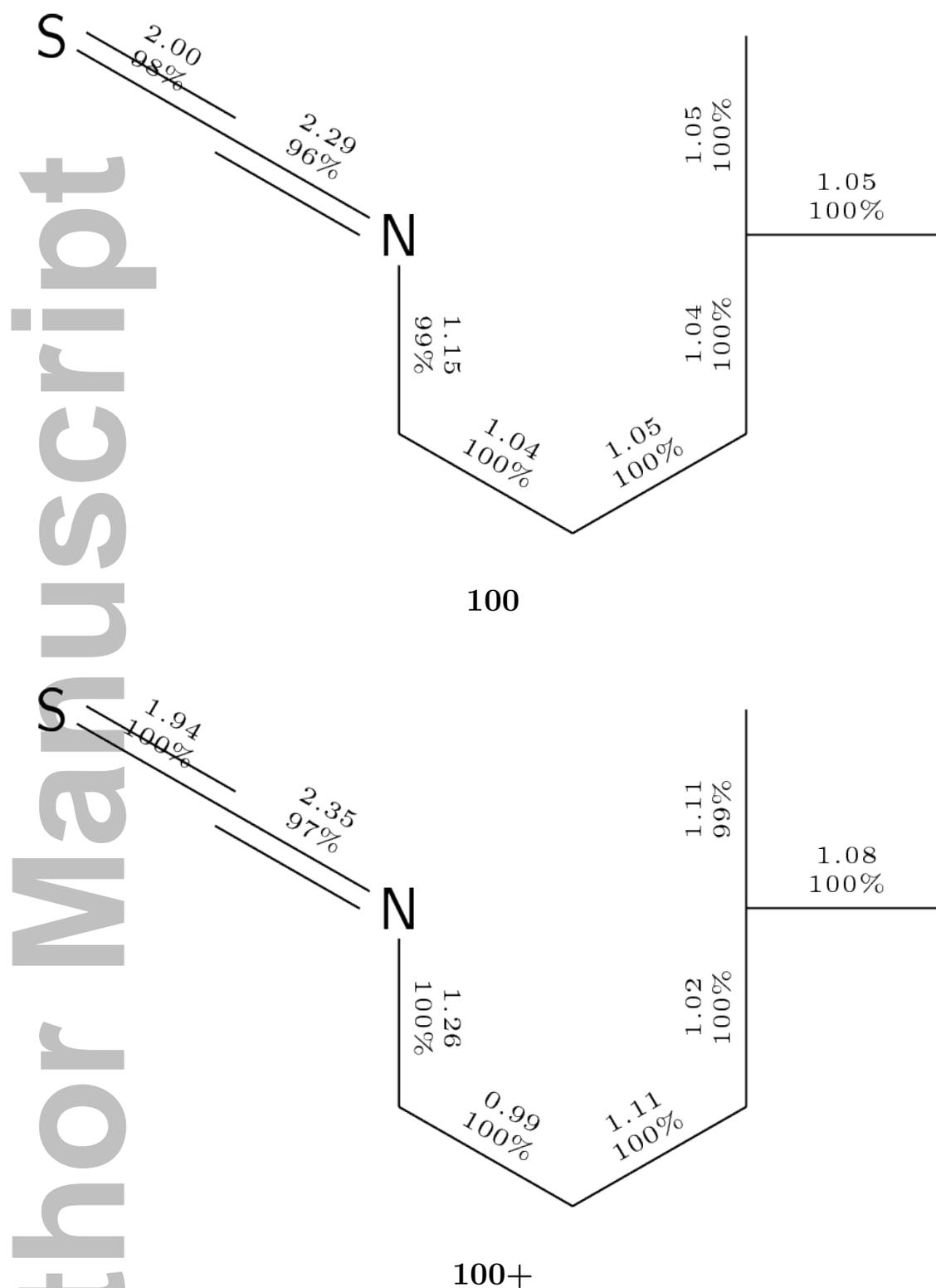
S
2.00
98%
2.29
96%
N
1.05
100%
1.05
100%
1.04
100%
1.15
99%
1.04
100%
1.05
100%

100

S
1.94
100%
2.35
97%
N
1.11
99%
1.08
100%
1.02
100%
1.26
100%
0.99
100%
1.11
100%

100+

Figure S100: Structure of molecule **100**. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**100**) and cation (**100+**).
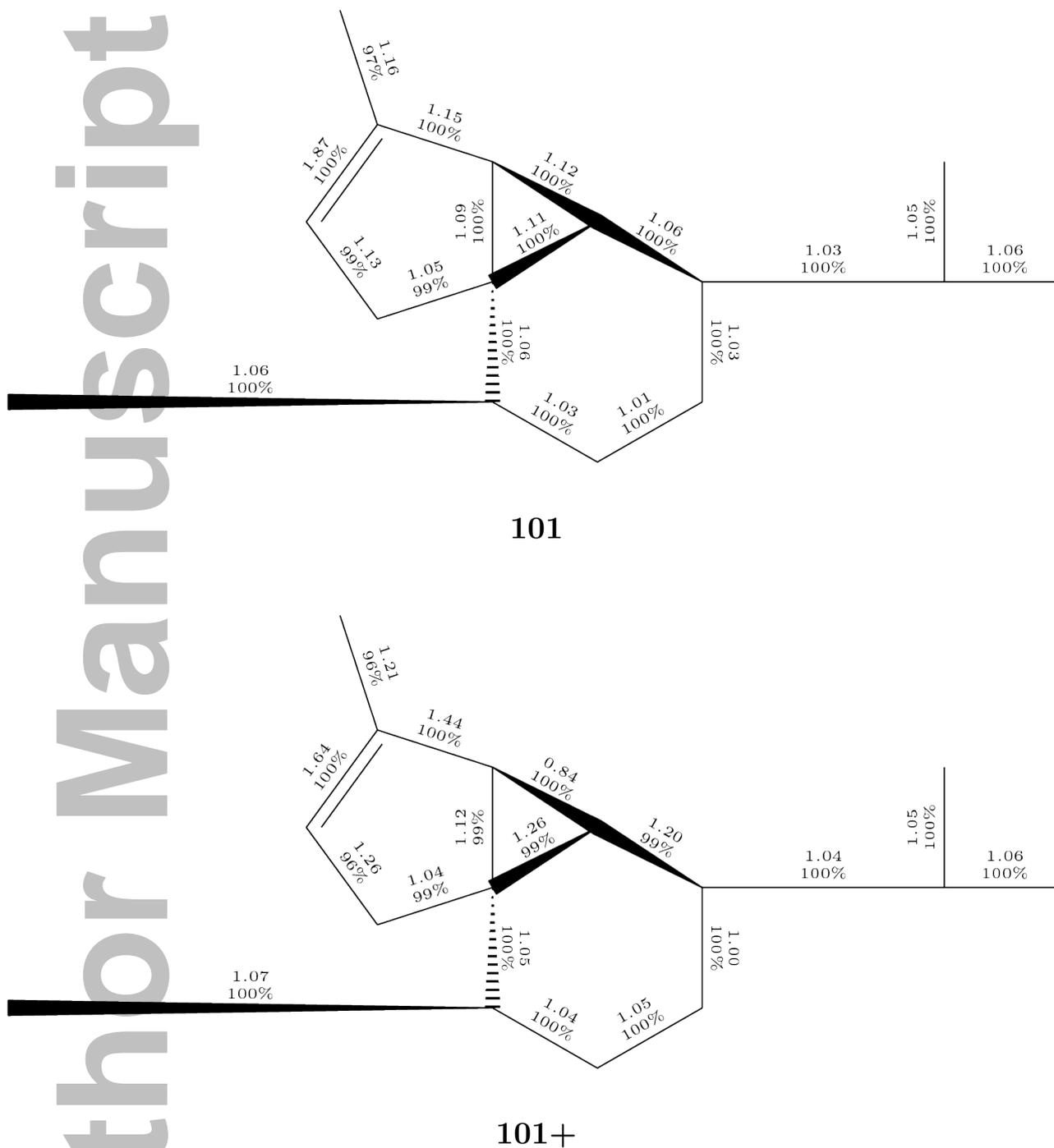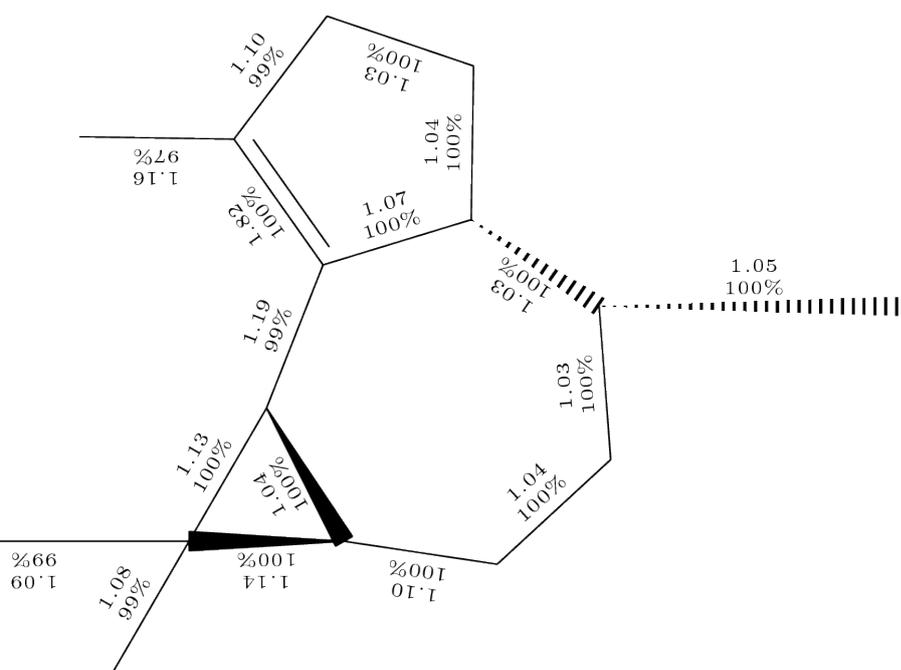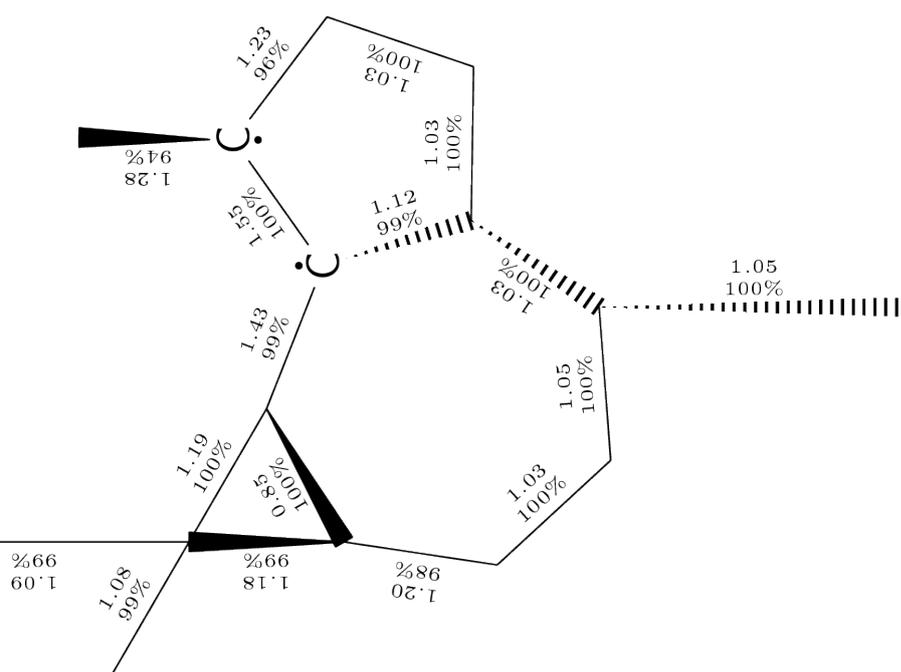
S4.5   Unusual ring systems

S4.5.1   Isomers of $C_{15}H_{24}$ formula 101-103

101

101+

Figure S101: Structure of molecule **101**. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**101**) and cation (**101+**).

102

102+

Figure S102: Structure of molecule **102**. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**102**) and cation (**102+**).
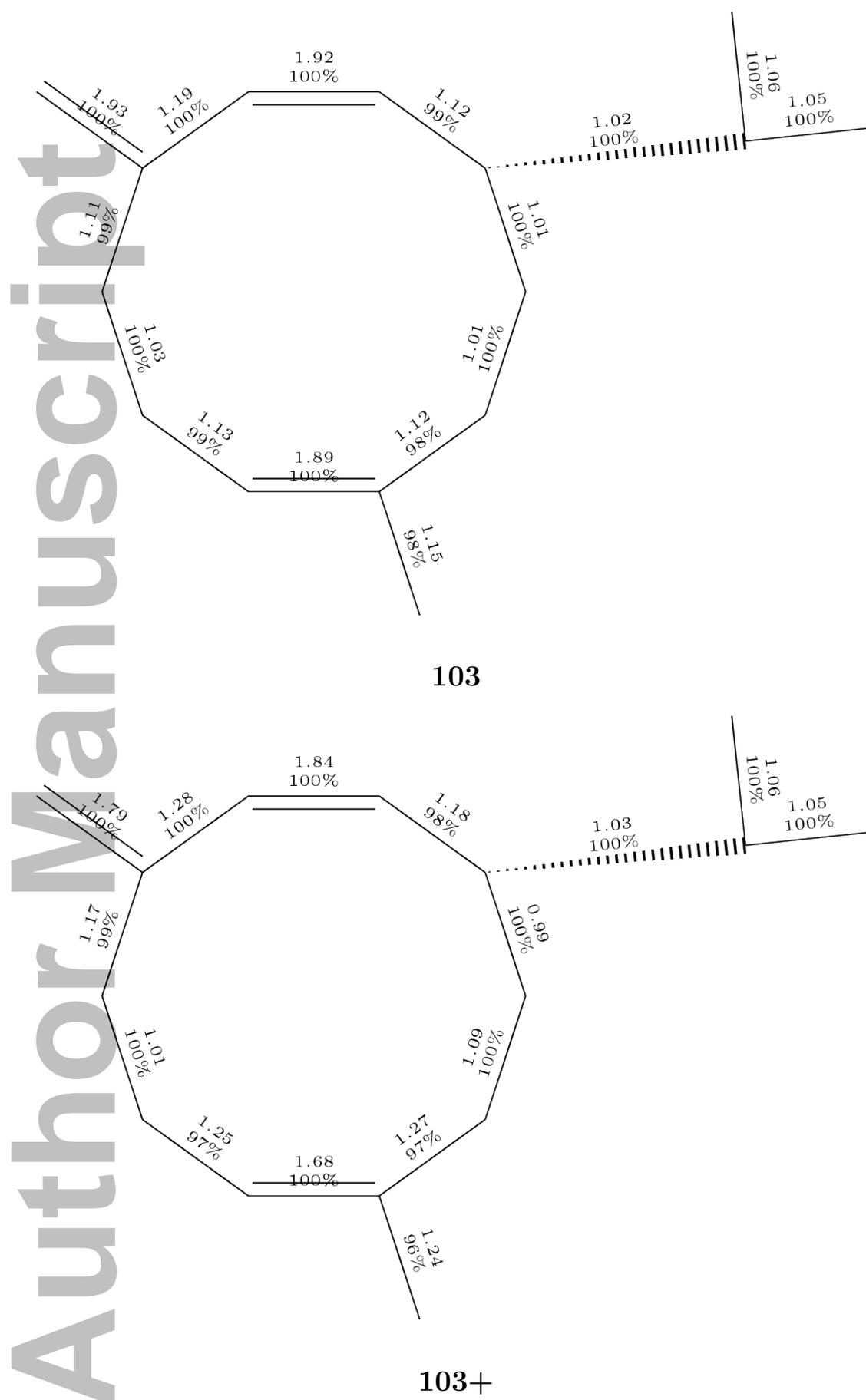
103

103+

Figure S103: Structure of molecule **103**. The Roby-Gould bond indices appear in the chemical structures (upper number) and the percentage covalency (lower number) for neutral species (**103**) and cation (**103+**).