

# Identification of complex molecules at surfaces: G-SIMS and SMILES fragmentation pathways

F.M. Green<sup>a,\*</sup>, E.J. Dell<sup>b</sup>, I.S. Gilmore<sup>a</sup>, M.P. Seah<sup>a</sup>

<sup>a</sup> *Quality of Life Division, National Physical Laboratory, Teddington, Middlesex, TW 11 0LW, UK*

<sup>b</sup> *Department of Physical and Theoretical Chemistry, University of Oxford, Oxford, UK*

Received 27 September 2007; received in revised form 19 December 2007; accepted 19 December 2007

Available online 20 January 2008

## Abstract

In this study, we develop a simple method using the SMILES molecular structure format to simulate fragmentation pathways in secondary ion mass spectrometry (SIMS). These pathways are found to have good agreement with fragmentation pathways identified using G-SIMS-FPM (Fragmentation Pathway Mapping) using the two examples of folic acid and Irganox 1010. G-SIMS is an easy-to-use method that considerably simplifies complex static SIMS spectra. G-SIMS-FPM allows the molecular structure to be re-assembled by following fragmentation pathways as the G-SIMS surface plasma temperature is varied. The simulated pathways help reduce the wide choice of possible structures faced by analysts as the molecular structure is reassembled, leading to more reliable molecular identification. A rapid method to establish a foundation database of simulated pathways using the community and a web-based system is proposed.

Crown Copyright © 2008 Published by Elsevier B.V. All rights reserved.

**Keywords:** Static SIMS; G-SIMS; Fragmentation pathway; Simulated spectra; SMILES

## 1. Introduction

Static SIMS is a powerful analytical tool for the analysis of complex molecules at the outermost surface (approximately 1 nm) of a solid with femtomole sensitivity, molecular specificity, high spatial resolution of better than 200 nm and with excellent reliability of better than 1%. It is essential to analysis in a wide range of technologies from the interactions of proteins at surfaces [1] and polymer interfacial behaviour [2] to drug delivery systems [3] and organic electronics [4]. A major use of static SIMS is the identification of molecules and organics at surfaces. However, the complexity of static SIMS spectra and the difficulties in interpretation and identification have been significant barriers to the wider uptake of the technique. G-SIMS has been demonstrated to be a powerful method for simplifying static SIMS spectra and leads to a direct identification for polymers and molecules [5–7] for biodegradable polymers [8,9] and for adhesives [10]. Details of the G-SIMS method are given in references [5–7]. Briefly, the static SIMS spectra are composed

of parent fragment ions amongst a large number of high intensity degradation products which make interpretation difficult. In the G-SIMS analysis, the fragmentation involves the partition functions of the fragments emitted from the surface plasma with effective temperature ( $T_p$ ). This theory allows one to extrapolate to a much lower value of  $T_p$  than is possible experimentally, by use of the ratio of two static SIMS spectra acquired under lower fragmentation and higher fragmentation conditions.

The significant peaks in the G-SIMS spectra are those peaks which would be emitted from a surface with a plasma of very low temperature and thus have little fragmentation and post-emission rearrangement. This allows a more direct interpretation of the spectra and identification of the molecule. This is adequate for small molecules and polymers. However, for larger molecules (>100 u) the mass alone may be insufficient to identify the molecule unambiguously. With a careful calibration of the mass scale [11] and instrument optimisation, an accuracy of the mass scale of 10 ppm can be achieved [11]. Unfortunately, this is insufficient to separate the many chemical permutations accommodated by this level of uncertainty. Traditionally in mass spectrometry an MS/MS experiment would be performed, fragmenting the molecular ion of interest. This is usually achieved via a low energy collision with target atoms, typically of an inert

\* Corresponding author. Tel.: +44 20 8943 6153; fax: +44 20 8943 6453.  
E-mail address: [Felicia.Green@npl.co.uk](mailto:Felicia.Green@npl.co.uk) (F.M. Green).

gas, followed by a mass analysis of the fragmentation products. In ion trap systems, this process may be repeated many times depending on the number of initial ions available and so is called MS<sup>n</sup>. This is a very powerful method and is routine for molecular identification. Unfortunately, at the present time none of the commercial ToF-SIMS instruments, widely used by industry and academia, have MS/MS capability. This is because of the need for high repetition rate (100  $\mu$ s) for fast imaging and also the small volumes of material that are available from the surface. For example in a monolayer, a 200 nm  $\times$  200 nm square pixel with only 1% ionisation, only 10 molecules are available assuming a transmission of 50%. The first ToF-SIMS with MS/MS capability is now being developed by Vickerman's group which takes advantage of C<sub>60</sub> primary ions to work beyond the "static limit" [12]. This has powerful potential, especially for biological applications.

Recently [7,13], we have shown that G-SIMS may be used to elucidate the molecular structure by varying  $T_p$ , to map out the variation of G-SIMS intensities by changing the G-SIMS index,  $g$ -index, from 0 (high fragmentation) to 40 (low fragmentation). This has two important advantages; firstly, it may be used on all commercially available instruments as demonstrated in an interlaboratory study [14] and, secondly, it only requires two mass spectra rather than the one for static SIMS and so does not require large amounts of analyte material at the surface.

Here, we demonstrate the G-SIMS molecular structure approach for two different molecules at surfaces, folic acid and Irganox 1010. The examples of folic acid and Irganox illustrate the approach but, of course, the structure of these molecules is known a priori. For analysts trying to identify an unknown, this is certainly not the case. In the second part of this study, we develop a novel approach based on Simplified Molecular Input Line Entry Specification (SMILES) [15], to identify the reassembly process through evaluation of fragmentation pathways of molecular structures. The SMILES description format allows the molecular structure of a molecule or fragment to be expressed in a logical computer readable way. Here, we show how this technique can be combined with G-SIMS-FPM (Fragmentation Pathway Mapping) to build up fragment libraries that will enable structural identification of unknown molecules from SIMS spectra.

The prediction of mass spectra from structural information and vice versa has been a long standing challenge since the mid 1960s when computers became sufficiently powerful to study small molecules. The DENDRAL project [16] 1965–1990 was the first. There, molecules are represented topologically using graph theory which, combined with a chemical knowledge base, led to predicted mass spectra for comparison with experiment. That is one of the first examples of an expert system. However, as noted in reference [17] that pioneering project did not meet its ultimate objective of automatic structure elucidation from mass spectral data. Gasteiger et al. [17] have used a different approach to predict a mass spectrum from the molecular structure including details of fragmentation and rearrangements occurring in the spectrometer. They used an automatic algorithm, FRANZ, that takes as input the molecular structure and experimental mass spectra to produce rapidly a knowledge base

of fragmentation and re-arrangement that also incorporated a rule base of fundamental reaction steps. FRANZ produces a fragmentation network of all possible reactions, which is then reduced through comparison with the peaks and intensities in the experimental data. That knowledge base is then used with another algorithm, MASSIMO, to predict the mass spectrum. Excellent agreement is shown between predicted and experimental spectra for small molecules with around 10 constituents excluding hydrogen. However, it is clear that these systems are not, as yet, readily applicable to complex molecules as evidenced by the usage in the community. We show that our simple system based on SMILES, coupled with the undegraded G-SIMS spectra, forms a powerful combination for the identification of complex molecules.

## 2. Experimental

The instrument used in this study is an ION-TOF IV (ION-TOF-GmbH, Germany) of the single stage reflection design. The instrument is equipped with an electron impact primary ion source as well as a separate Cs<sup>+</sup> primary ion source, which share a single focusing column mounted at 45° to the sample surface normal. The ion beam is readily switched from Cs<sup>+</sup> to Ar<sup>+</sup> for G-SIMS analysis. G-SIMS spectra were acquired as described in detail elsewhere [5]. For each analysis, the ion beam was digitally rastered with a 128  $\times$  128 array over an area of 200  $\mu$ m  $\times$  200  $\mu$ m using a beam current of less than 1 pA. The order of analysis was Cs<sup>+</sup> followed Ar<sup>+</sup> from the same area giving a combined total primary ion dose of  $<1 \times 10^{16}$  ions/m<sup>2</sup>.

Folic acid powder (SIGMA F-8798) of 24.45 mg was dispersed in 25 ml of tetrahydrofuran. A droplet (approximately 5  $\mu$ l) was dispensed on to the surface of a freshly UV/ozone cleaned 1 cm  $\times$  1 cm piece of silicon wafer. Following evaporation of the solvent, the folic acid was homogeneously distributed over the sample surface. Thin films of Irganox 1010, were prepared from a 1 mg/ml chloroform solution by spin coating onto clean silicon wafers. The molecular structures for folic acid and Irganox 1010 are given in Fig. 1 together with labels that will be used to identify particular sub-structures later.

## 3. G-SIMS fragmentation pathway mapping

Firstly, G-SIMS spectra are calculated separately for the two secondary ion polarities from the ratio of spectral intensities,  $F_x$ , for Cs<sup>+</sup> and Ar<sup>+</sup> primary ions, as defined in Ref. [5], where  $x$  is an index referencing each secondary ion peak in the mass spectrum. We may now generate a G-SIMS spectrum,  $I_x$ , by multiplying an existing spectrum,  $N_x$ , (the static SIMS spectrum with least fragmentation) with the factor  $F_x^{13}$ . This forms the G-SIMS spectrum with intensities given by:

$$I_x = M_x N_x F_x^g \quad (1)$$

where  $g$ , the G-SIMS index, if as given above, would be 13. We have shown elsewhere [5–7] that 13 is a convenient value that removes most of the degraded peaks. The additional factor  $M_x$ , the mass of the emitted fragment, is found useful to enhance the

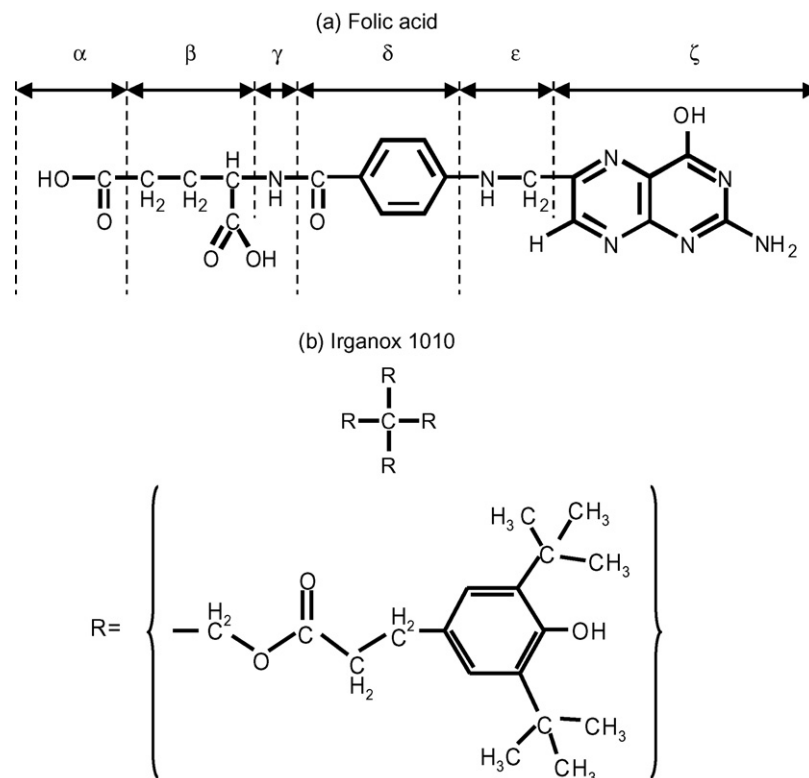


Fig. 1. The molecular structure of (a) folic acid labelled with six subunits:  $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$ ,  $\epsilon$ ,  $\zeta$  and (b) Irganox 1010—a central carbon atom surrounded by 4 equal side-chains denoted by the symbol R.

natural fall in emission with mass. We shall discuss the effect of the choice of the  $g$  index value later. In addition, the procedure described in Ref. [6] was used where a tangent gradient is selected to normalise the tops of the fragmentation cascades observed in  $F_x$  [5] to have similar values across the mass range. For negative ions, it was found that the molecular ion exhibited a significantly higher  $F_x$  value than other fragments and was, therefore, not included in the tangent gradient selection [13]. If this was not done, the G-SIMS spectra would be entirely dominated by the parent molecule and the intermediate degradation products, necessary for studying the breakdown of molecular structure, would be obscured. The positive and negative ion G-SIMS spectra for folic acid and Irganox 1010 are shown in Figs. 2 and 3, respectively. These spectra are much simpler than the static SIMS spectra which consist of a high population of small mainly fragmented ions as can be seen in Fig. 2(a) of reference [7] and Fig. 4(d) of reference [18]. The peaks that are of strong intensity in the G-SIMS spectra are those fragments that are characteristic of the parent structure with simple fragmentation and no rearrangements. The G-SIMS spectra more closely resemble Electron Impact (EI) or Electrospray mass spectra. It is immediately apparent that the positive and negative spectra are complementary and recently we have developed a methodology to incorporate both polarities in a combined G-SIMS-FPM analysis [13].

We may investigate the effect of varying the surface plasma temperature on the G-SIMS intensities by changing the G-SIMS index,  $g$ , of Eq. (1). This has some analogy to changing the centre of mass energy in the collision of ions with inert gas target

atoms in a traditional MS/MS experiment. Fig. 4 shows this for both positive and negative ions [13] with values of  $g$  from 0 to 40 for folic acid and Irganox 1010. For each value of  $g$ , the G-SIMS intensities are normalised by dividing by the total G-SIMS intensity rather than the maximum intensity used in Figs. 2 and 3. With  $g = 0$  we begin in the SSIMS regime and as  $g$  is increased we progress to the G-SIMS regime (lower surface plasma temperature). Firstly, we see the substrate and low molecular weight fragments decay rapidly. At the same time, the intensities of more intact fragments begin to grow but these are soon overtaken by even larger intact fragments which grow more strongly. Consequently, the intensities of the smaller degraded fragments begin to decay. This evolves for larger and more intact fragments until eventually the dominant parent fragment has the highest G-SIMS intensity. For each fragment, the G-SIMS intensity goes through a maximum at a characteristic value,  $g_{\max}$ , and generally, the smaller the fragment, the lower  $g_{\max}$ .

Fig. 5 shows the re-assembly plot with values of  $g_{\max}$  for all the fragments in the G-SIMS spectrum. We analyse the fragmentation process by first selecting the highest mass fragment with the highest value of  $g_{\max}$ . This is a parent fragment with a mass,  $M_p$ . We next choose a fragment at lower mass with the next lowest value of  $g_{\max}$ . This is the daughter fragment with mass,  $M_D$ . We now postulate that the parent fragmented into the daughter (which exhibits a higher surface plasma temperature) with up to two fragmentation products, co-daughters, with masses  $M_{C1}$  and  $M_{C2}$ , respectively. All possible combinations of up to two fragments are computed from those fragments present in the G-SIMS spectrum of Fig. 3. The positive fragment ions

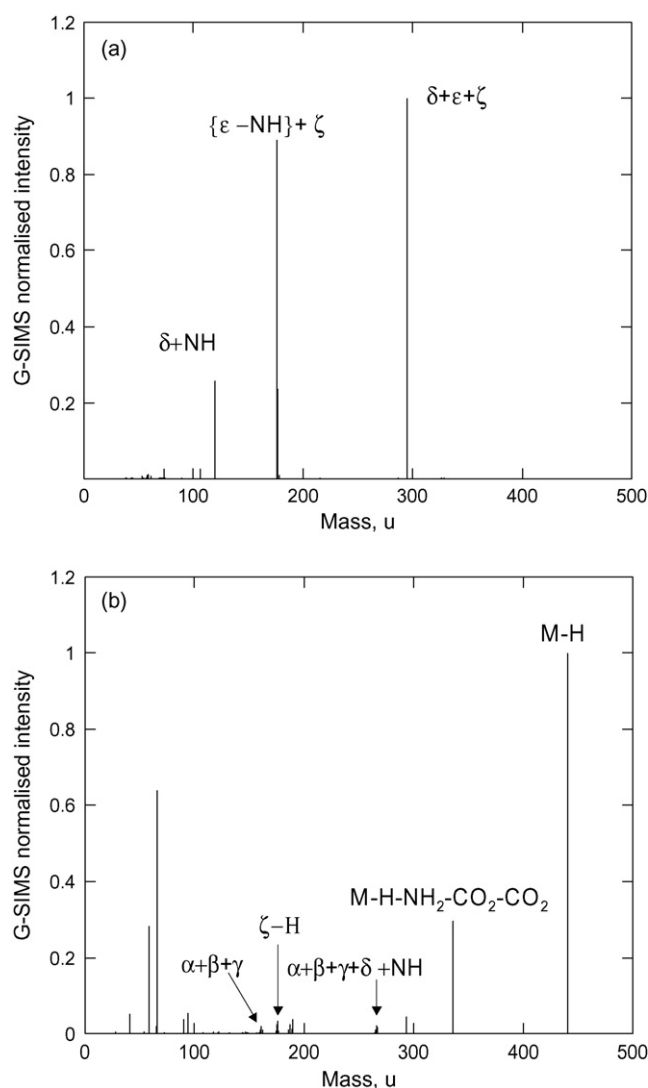


Fig. 2. G-SIMS spectra (see Eq. (1)) of folic acid (a) positive ions and (b) negative ions.

in the G-SIMS spectra are likely to be protonated and the negative ions deprotonated so the possibility of a difference 1 or 2 hydrogen atoms with mass,  $n\text{H}$  is permitted, such that

$$M_{\text{P}} = M_{\text{D}} + M_{\text{C1}} + M_{\text{C2}} + n\text{H} \quad (2)$$

where  $-2 \leq n \leq 2$ .

The possible co-daughter fragments are then validated by calculating their composition from the peak centroid mass and comparison with chemically possible structures. Typically,  $M_{\text{C}} \ll M_{\text{P}}$  and therefore the chemical composition may be assigned with higher certainty.

In Fig. 5(a), we apply this method to folic acid and use the sub-unit labels  $\alpha, \beta, \gamma, \delta, \epsilon, \zeta$  defined in Fig. 1 to identify fragments. Here we highlight four key fragmentation pathways, from the molecular ion, and these are indicated in the figure. With careful mass measurement [11], it may be seen that fragmentation pathway 1 leads to three products which are identified as  $\delta + 2\text{H}$ ,  $\epsilon - \text{H}$  and  $\zeta - \text{H}$  (with nominal masses 106, 28 and 161 u) so that the parent ion for that pathway is  $\delta + \epsilon + \zeta$  (with mass 295 u),

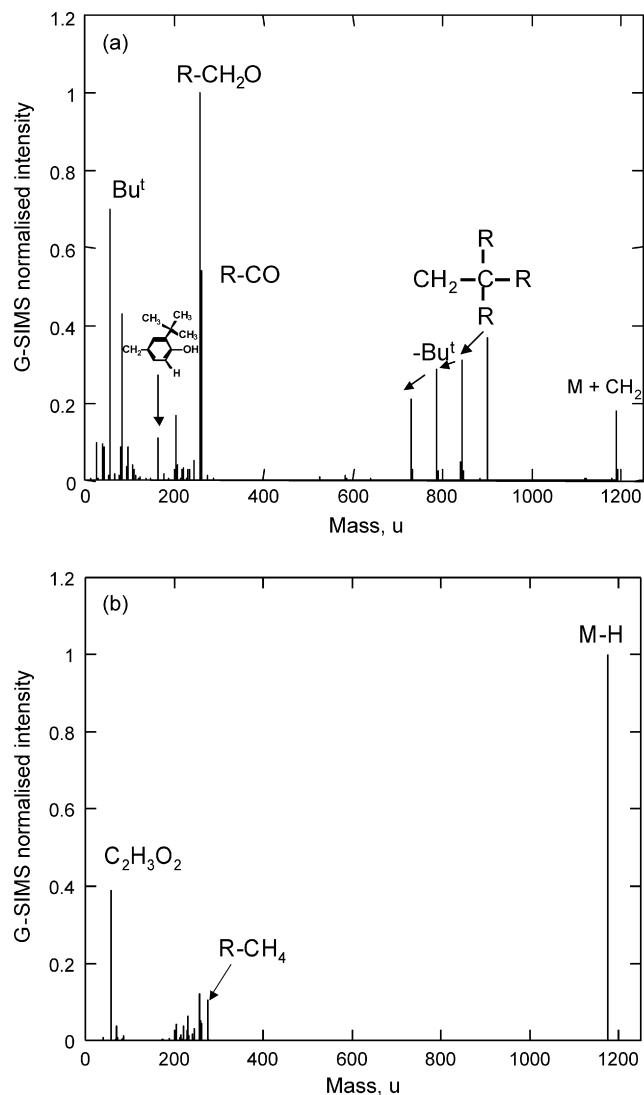


Fig. 3. G-SIMS spectra of Irganox 1010 (a) positive ions and (b) negative ions.

where the symbol order does not necessarily define the order in the molecule. This represents one end of the molecule and provides the identity of the daughter (with mass 295 u) for fragmentation pathway 3. The co-daughter is identified as  $\alpha + \beta + \gamma$  (with mass 145 u), which then leads to the identification of the entire molecular structure  $\alpha + \beta + \gamma + \delta + \epsilon + \zeta$  (with mass 440 u). Fragmentation pathways 2 and 4 support these assignments and are summarised in Table 1. Many more fragmentation pathways may be identified in Fig. 5 and later we develop a method using simulated pathways to help analysts do this.

In Fig. 5(b), three fragmentation pathways from the Irganox parent ion (mass 1176 u) are identified each leading to daughter and co-daughters. We call these fragmentation pathways 4, 3 and 2. Pathways 2 and 3 each have a common co-daughter (at mass 260 u). Similarly, a fragmentation pathway for that co-daughter to its own next generation daughter and two co-daughters may also be identified (masses 163, 41, 58 u), called pathway 1. These daughter and co-daughter ions are now below 200 u so the chemical composition and structure of these smaller entities may be identified from the peak positions with an accu-

Table 1  
Summary of fragmentation pathways for folic acid identified in Fig. 5(a) with identified molecular structures

Fragment pathway	Parent	Daughter	Co-daughter 1	Co-daughter 2
1	$\delta + \varepsilon + \zeta$	$\zeta - \text{H}$	$\delta + 2\text{H}$	$\varepsilon - \text{H}$
2	$\delta + \varepsilon + \zeta$	$\{\varepsilon - \text{NH}\} + \zeta$	$\delta + \text{NH}$	
3	$\alpha + \beta + \gamma + \delta + \varepsilon + \zeta - \text{H}$	$\delta + \varepsilon + \zeta$	$\alpha + \beta + \gamma$	
4	$\alpha + \beta + \gamma + \delta + \varepsilon + \zeta - \text{H}$	$\alpha + \beta + \gamma + \delta + \text{NH}$	$\{\varepsilon - \text{NH}\} + \zeta$	

Note: The symbols  $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $\delta$ ,  $\varepsilon$  and  $\zeta$  denote sub-structures of folic acid identified in Fig 1(a). The molecular ion is denoted by the symbol M is equivalent to  $\alpha + \beta + \gamma + \delta + \varepsilon + \zeta$ .

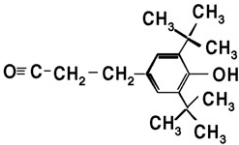
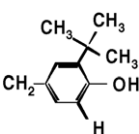
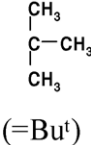
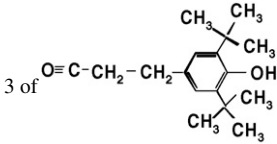
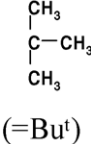
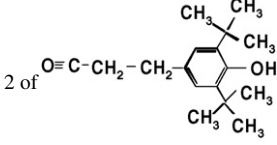

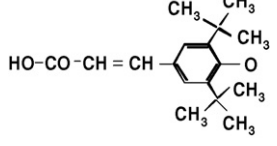
rately calibrated mass scale [11]. The structures are identified in Table 2 and are consistent with the structures in the SSIMS library [19]. These may be assembled to form a complete end of the side chain R (see Fig. 1), leaving  $\text{CH}_2\text{-OH}$  attached to the Irganox 1010 central carbon. We have now identified the co-daughter for fragmentation pathways 2 and 3 (mass 260 u). It is then clear that the daughters for fragment pathways 2 and 3 (masses 322 and 640 u) are the molecular ion with 3 and 2 side chains broken off, respectively. Similarly, the fragment pathway 4 shows the loss of 1 side chain (to mass 900 u). By following the fragmentation pathways from a large molecule ion to smaller entities, those entities may be identified. The original parent molecule structure is then reassembled in a manner similar to traditional MS/MS experiments. In this analysis, we have treated positive and negative ions equally and have not forced charge balance. This is permitted because we are not following individual molecular events but an average ensemble

where the main populations are neutrals which are not observed.

### 3.1. SMILES fragmentation pathways

So far, we see how the molecular structure may be reassembled using G-SIMS fragmentation pathway mapping for folic acid and Irganox 1010. Of course, in this example, the molecular structure of the parent molecule is known a priori. Typically, in industrial analysis it is the molecular structure that must be identified for an unknown and this makes the interpretation of the fragmentation pathways significantly more challenging. What analysts need is an efficient system to guide them through the reassembly process providing suggested options. This may provide a set of possible molecular structures and not ultimately a unique molecular structure. However, this would be a significant advance on the reliance of the mass spectrum as a chemical

Table 2  
Summary of fragmentation pathways for Irganox 1010 identified in Fig. 5(b) with identified molecular structures

Fragment pathway	Parent	Daughter	Co-daughter 1	Co-daughter 2
1			$\text{C}_2\text{HO}$	
2	$\text{R} - \text{C}(\text{R}) - \text{R}$ [R-H]	$\text{CH}_2 - \text{C}(\text{R}) - \text{CH}_2 - \text{OH}$ [R-Bu <sup>t</sup> ] OH	3 of 	
3	$\text{R} - \text{C}(\text{R}) - \text{R}$ [R-H]	$\text{CH}_2 - \text{C}(\text{R}) - \text{CH}_2 - \text{OH}$ OH	2 of 	
4	$\text{R} - \text{C}(\text{R}) - \text{R}$ [R-H]	$\text{CH}_2 - \text{C}(\text{R}) - \text{R}$ R		

Note: The side chain, R, having lost one hydrogen or a tertiary butyl group is denoted by [R-H] and [R-Bu<sup>t</sup>] respectively. The molecular side chain is denoted by the symbol R, as identified in Fig. 1(b).



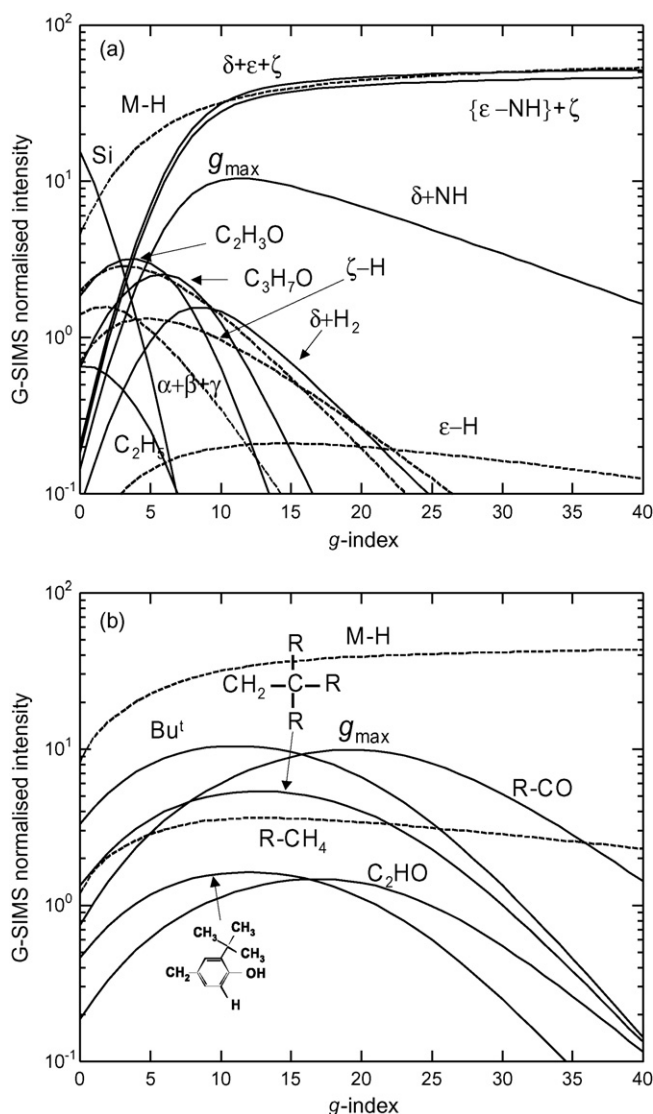


Fig. 4. Effect of increasing  $g$ -index from 0 to 40 on normalised G-SIMS intensities for positive ions (solid lines) and negative ions (dashed lines) (a) folic acid and (b) Irganox 1010.

fingerprint where the fingerprint databases are, and are likely to remain, relatively small.

On the left in Fig. 6, we illustrate the fragmentation pathway identified using G-SIMS-FPM, starting with the parent molecule at the top and fragmenting to smaller sub-structures as we move down to either positive or negative ions. Fragments that are ionised account for around only 1% of the total emitted fragments. Fragmentation to neutrals are not observed and would leave gaps in the fragmentation pathways. On the right in Fig. 6, we illustrate a complementary simulated fragmentation pathway constructed by again starting at the top with the parent molecule and then conceptually fragmenting the molecule into ever smaller sub-structures until the constituent elements. In the following, we show how these pathways may be simulated in a simple way and demonstrate the complementarity of these pathways with the G-SIMS-FPM analysis of folic acid and Irganox 1010 given earlier. What Fig. 6 does not do is

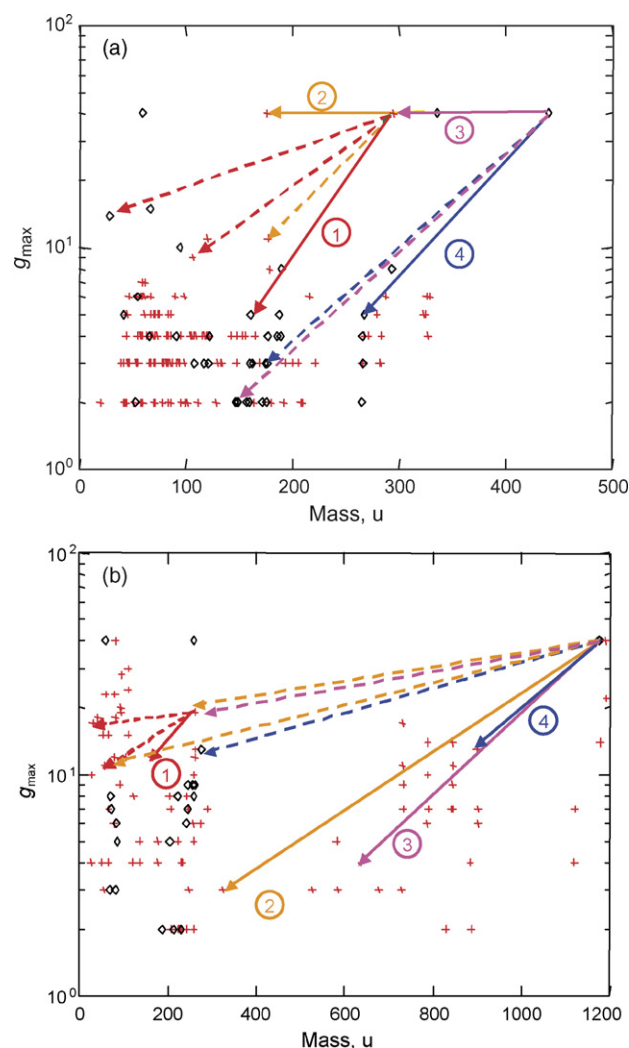
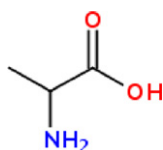


Fig. 5. Reassembly plot of  $g_{\max}$  from all mass peaks in the G-SIMS spectra of (a) folic acid and (b) Irganox 1010. Four key fragmentation pathways are shown, in each case as a solid line for the parent to the main daughter and as dashed lines for parent to co-daughters (see Eq. (2)) with the set of pathways for each ion shown in a given colour. The fragmentation pathway sets are labelled with a circumscribed number for reference in the text and in Tables 1 and 2. After Gilmore and Seah [7].

to consider a rearrangement of the molecule that subsequently fragments.

Simplified Molecular Input Line Entry Specification allows the structure of a molecule to be unambiguously expressed in a logical computer readable way using an ASCII text string. SMILES was originally developed by Weininger [15] and more recently developed by Daylight Chemical Information Systems [20]. The SMILES format is a very popular format used in informatics and is integrated into a wide range of software freely available on the internet to give, for example, molecular structure [21,22], chemical information such as  $pK_A$  [23], solubility [23], molecular volume [22] and  $\log P$  values [22] as well as physical parameters such as density [23], refractive index [23], molar mass [24] and mass spectral isotope patterns [24]. The syntax and grammar of SMILES is described in detail elsewhere [15,20]. In brief, each element is denoted by

its chemical symbol with the first letter given in the upper case. Adjacent atoms in the SMILES string (read from left to right) are considered to have a single bond unless otherwise denoted by an inserted “=” or “#” character representing double and triple bonds, respectively. Hydrogen is implicit. Branches are described using parentheses and it is possible to nest parentheses for sub-branches. Cyclic structures are defined by a number identifying which atoms open and close the ring structure so that cyclohexane is represented by C1CCCCC1. Aromatic atoms are shown in lower case so that benzene is represented by c1ccccc1. Ionic bonds may also be described as well as chirality. Canonical form ensures that the SMILES structure is unique for a given molecular structure. For example, CC(C(=O)O)N represents the following structure (without chirality):



A computer programme developed at NPL using MATLAB (The MathWorks, Natick, USA) simulates full fragmentation pathways for any molecule given in the SMILES format. This starts with the parent molecule of unlimited complexity and cleaves it at a bond into two parts. This is repeated for each bond in the parent molecule and the fragment products are listed in the first level of a tree structure, as illustrated in Fig. 6. For example, an illustrative molecule ABCD would be fragmented by one fracture into six sub-structures A, BCD, AB, CD, ABC and D. Clearly, the fragmentation terminates for A and D but, for example, BCD may be fragmented into 4 further sub-units B, CD, BC and D giving a second level of fragmentation and similarly CD and BC may be fragmented to a third level of fragmentation. In this example the third level of fragmentation is the final level so that at that stage the molecule has been fragmented to its constituent elements. At the bottom, in the right hand example in Fig. 6, we would usually finish with fewer possibilities in the final level and possibly, for example, just C, H and O. It is also easy to see in the above simple example that the sub-unit BC will be found from both the fragmentation of BCD and of ABC. There is no need to fragment the same structure twice and, to save computer time for each fragment, the computer checks in a look-up table to see if a fragment has been done before. If so, a link is stored and the fragmentation termi-

nated. Later, we shall see how this can be extended to include fragments in a larger database for many parent molecules. This method significantly reduces computational time. The computer programme uses a structured array with each element containing details of the fragment including SMILES text string, exact mass and links to other parts in the fragmentation pathway with the same fragment. The exact mass is calculated by a separate programme that, for a given SMILES text string, sums the main isotope masses of all the constituent elements as well as the number of implicit hydrogens through determination of atom valency and bonding. We use these values later to compare with experimental G-SIMS data.

As discussed earlier, branching within a molecule is described by enclosing that part of the SMILES text within parentheses. Fragmentation of branched structures is accurately executed by the programme. For example, in the molecule AB(EFG)CD, cleavage of the branch EFG from the atom B should result in the two fragments ABCD and EFG and is very unlikely to rearrange into AB and (EFG)CD, which could result from incorrect treatment of the branching. A further complication for the computer programme is branching within branches. This may continue to many levels of branching, yielding a SMILES text with many nested parentheses. The computer programme was carefully developed to deal efficiently with such systems which do often occur in complex molecules.

### 3.2. Simulated SMILES fragmentation pathways for folic acid and Irganox 1010

The SMILES text for folic acid is readily available from the web, as follows:

```
OC(=O)CCC(C(=O)O)NC(=O)c1ccc
(NCc2nc3c(O)nc(N)nc3nc2)cc1
```

This structure is provided as input to the fragmentation pathway simulator software and the fragmentation pathways are calculated. On a Pentium PC with a 2.66 GHz processor this takes approximately 20 min. To fragment the molecule to the constituent atoms C, N and O requires 34 levels of fragmentation. However it is assumed that stable ring structures, such as benzene, are unlikely to fragment, such that folic acid only requires 17 levels of fragmentation to break into constituent parts. The output of the program contains a structured array described earlier containing each fragment structure in SMILES

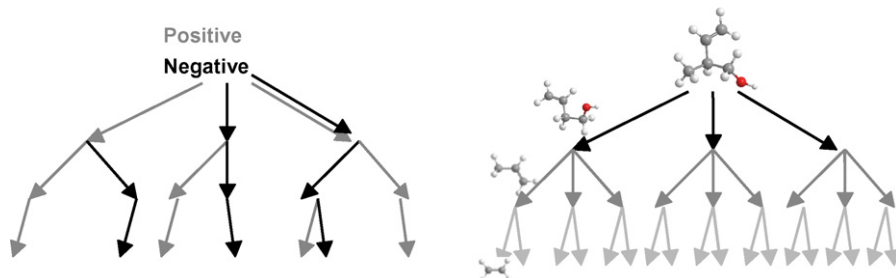


Fig. 6. Illustrative example of the fragmentation pathways mapped out in G-SIMS for positive and negative ions (left) and the complementary simulated pathways using the SMILES fragmentation software (right).

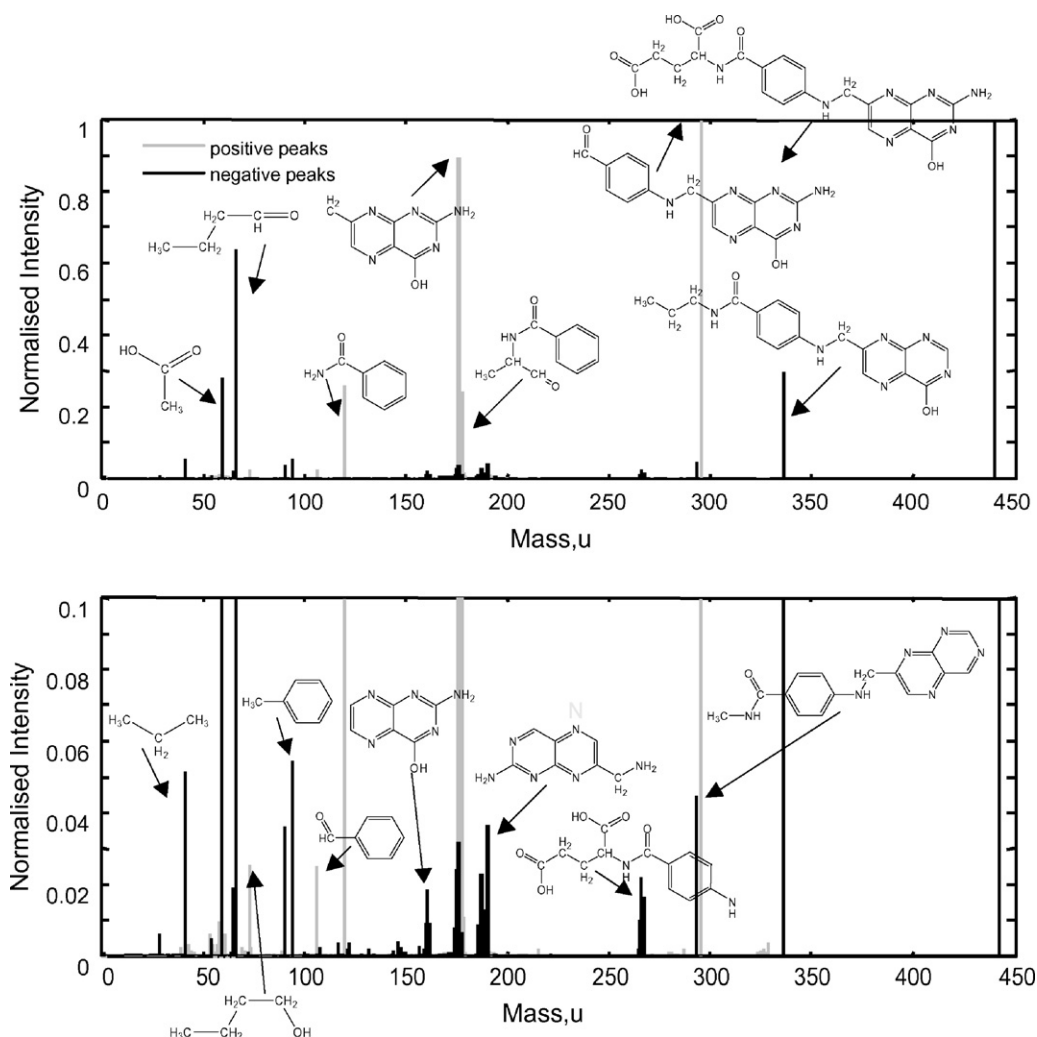


Fig. 7. Combined G-SIMS positive ion (grey peaks) and negative ion (black peaks) spectra of folic acid (from Fig. 2), lower frame is the upper spectrum expanded 10-fold on the ordinate scale. The neutral molecular structures are shown with the related molecular ion peaks, for those with significant intensity. These peaks match the SMILES fragments within a mass tolerance of 0.1 u ( $\pm 1$  Hydrogen). These structures may all be found in Fig. 1(a) and exhibit only one or two bond scissions in the backbone.

text and the main isotope accurate mass position. For this system to be useful it must be able to describe say at least 75% of the fragments observed in the key G-SIMS spectrum. In Fig. 7, we populate the G-SIMS spectrum of folic acid of Fig. 2(a) with the SMILES fragments matching within a mass tolerance of 100 ppm ( $\pm 1$  Hydrogen) from the simulated fragmentation pathway. As can be seen the simulated fragments account for 90% of peaks with an intensity of  $>0.02$ . This excellent agreement between experimental G-SIMS data and our simple system for simulating fragments is because the G-SIMS spectrum consists of simple fragments that are undegraded and have not undergone post-emission rearrangements [5].

Similarly, the SMILES text for Irganox 1010 is given by:

```
C(COC(=O)CCc1cc(C(C)(C)C)c(O)c(C(C)(C)C)c1)
(COC(=O)CCc1cc(C(C)(C)C)c(O)c(C(C)(C)C)c1)
(COC(=O)CCc1cc(C(C)(C)C)c(O)c(C(C)(C)C)c1)
COC(=O)CCc1cc(C(C)(C)C)c(O)c(C(C)(C)C)c1
```

Again, an excellent agreement between experimental G-SIMS and simulated data is found. Earlier, we discussed the endeavours of over 40 years of research to simulate mass spectra from gas phase ionisation and the limited success that had been achieved for complex molecules. To do this for SIMS would indeed be ambitious. Fortunately, G-SIMS significantly simplifies the problem so that our simple approach is quite successful.

The simulated fragmentation pathways stored in the structural array are essentially organised in a tree structure. However, for analytical purposes this is not so convenient since fragments of a similar mass are not necessarily juxtaposed. Instead, in Fig. 8(a) we show the simulated fragmentation pathways for folic acid as a mass based tree structure with the fragmentation level on the ordinate and the fragment mass on the abscissa. So in the top right we start with the parent molecule going through successive fragmentations until eventually ending up with the constituent structures (exact aromatic ring structures which are kept intact) and elements C, N, O in the bottom left. This plot is then similar to the G-SIMS reassembly plot shown earlier in



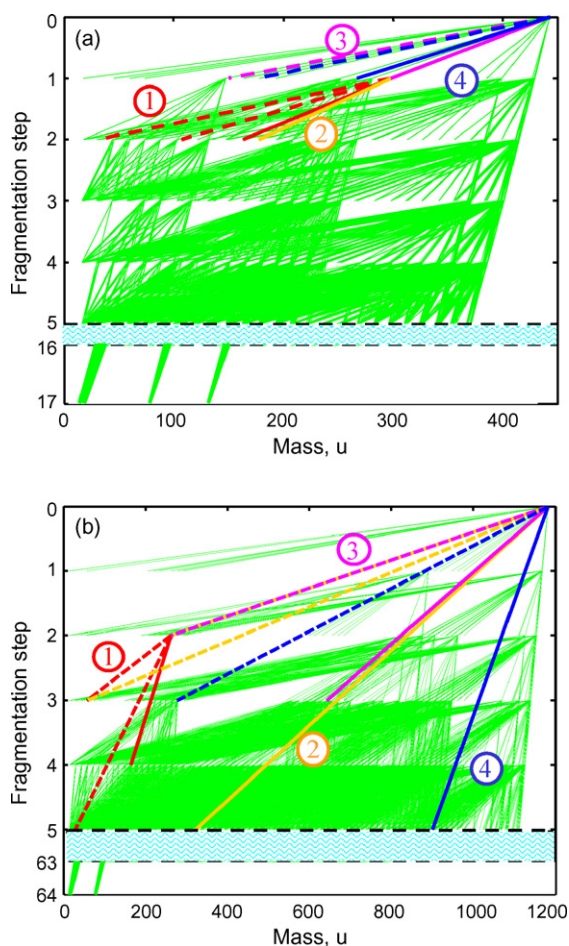


Fig. 8. Simulated fragmentation pathways displayed on a plot of fragment mass and fragmentation level for (a) folic acid and (b) Irganox 1010. For clarity, not all fragmentation levels are shown. The fragmentation pathways identified in Fig. 5 are also highlighted for comparison.

Fig. 5. Each of the 4 fragmentation pathways identified earlier may also be found on the simulated fragmentation pathway and these are highlighted. At the 16 and 17th fragmentation levels all of the structural information has been lost and are shown here for completeness. Practically, the 10th level of fragmentation is typically required for structurally significant peaks.

In Fig. 8(b), we show the mass based tree structure of fragmentation for Irganox 1010 for comparison with the G-SIMS FPM reassembly plot of Fig. 5. The 4 pathways identified in Table 2 and Fig. 5 are highlighted.

### 3.3. Interpretation of G-SIMS fragmentation pathways using simulated fragments

Earlier, we showed how the molecular structure could be reassembled from small fragments by working in reverse up the fragmentation pathways. For the illustrative examples of folic acid and Irganox 1010, this was helped by knowing the molecular structure. What analysts require is a simple system to guide them through the possible choices as the molecule is reassembled but its structure is unknown.

An analyst may select any fragment in the G-SIMS fragmentation pathway and the mass of this fragment may then be automatically compared with a simulated library of fragments containing many types of parent molecule. A suite of possible fragmentation routes may then be identified to build the fragment to one of the larger mass that is contained in the experimental data. This guided process may then be followed until reaching the largest fragment in the experimental data. This may not identify a unique solution but may produce a reduced set of options. This accelerates and improves the reliability of establishing the molecular structure and molecular identity.

A great advantage of a simulated fragmentation library is that it does not rely on the contribution of experimental data from the community, which is always the rate-limiting factor. It is, of course, essential to validate the simulated pathways with experimental data, as started in the example here. An analysis of this will also allow the identification of the most effective fragmentation rules from the mass spectrometry community that are relevant to the processes occurring in G-SIMS. Further development of this approach will include this rule base which will simplify the number of possible fragmentation pathways. One mechanism to grow the library very rapidly is to provide the SMILES simulation fragmentation pathway software freely available on the web for the community to use on molecules of interest to them and for the generated simulated fragmentation pathways to then be added to an open-access central web library, which is freely accessible to all users. This approach is now being developed and will be tested.

## 4. Conclusion

We have developed a novel method for simulating fragmentation pathways using the SMILES format for molecular structure. A computer program rapidly generates the fragmentation pathways. The simulated fragments are shown to be in very good agreement with G-SIMS-FPM experimental data for folic acid and Irganox 1010 with 90% of the fragment ions explained. A further development allows the simulated fragmentation pathways to be shown in a similar way to the G-SIMS reassembly plot. To help analysts reassemble unknown molecules, a library of simulated fragment pathways is in the process of being developed and validated against experimental data for a wide selection of molecules including amino acids and simple peptides [25]. As the G-SIMS and SMILES method develops, we hope that it will guide analysts through appropriate choices as pathways are selected in the reassembly. This will either provide a unique solution to the identity of a molecule or will help significantly to identify a short list of strong possibilities. A method has been proposed to grow such a library rapidly, using freely available web access.

## Acknowledgements

This work forms part of the Chemical and Biological Programme of the National Measurement System of the UK Department of Innovation, Universities and Skills (DIUS). The

authors are grateful to Dr G. O'Connor and Mr P. Stokes of LGC for helpful discussions regarding MS/MS fragmentation.

## References

- [1] M.S. Wagner, D.G. Castner, *Appl. Surf. Sci.* 231 (2004) 366.
- [2] I.S. Gilmore, M.P. Seah, J.E. Johnstone, *Surf. Interf. Anal.* 35 (2003) 888.
- [3] A.M. Belu, M.C. Davies, J.M. Newton, N. Patel, *Anal. Chem.* 72 (2000) 5625.
- [4] R. Pinna, F. Jamme, F.J.M. Rutten, E.F. Smith, M.R. Willis, D. Briggs, M.R.S. McCoustra, *Appl. Surf. Sci.* 252 (2006) 6672.
- [5] I.S. Gilmore, M.P. Seah, *Appl. Surf. Sci.* 161 (2000) 465.
- [6] I.S. Gilmore, M.P. Seah, *Appl. Surf. Sci.* 203/204 (2003) 551.
- [7] I.S. Gilmore, M.P. Seah, *Appl. Surf. Sci.* 231/232 (2004) 224.
- [8] R. Ogaki, F.M. Green, S. Li, M. Vert, M.R. Alexander, I.S. Gilmore, M.C. Davies, *Appl. Surf. Sci.* 252 (2006) 6797.
- [9] R. Ogaki, F.M. Green, I.S. Gilmore, A.G. Shard, S. Luk, M.R. Alexander, M.C. Davies, *Surf. Interf. Anal.* 39 (2007) 852.
- [10] P.N. Hawtin, M.-L. Abel, J.F. Watts, J. Powell, *Appl. Surf. Sci.* 252 (2006) 6676.
- [11] F.M. Green, I.S. Gilmore, M.P. Seah, *J. Am. Soc. Mass Spectrom.* 17 (2006) 514.
- [12] J.C. Vickerman, private communication, 2006.
- [13] I.S. Gilmore, F.M. Green, M.P. Seah, *Appl. Surf. Sci.* 252 (2006) 6601.
- [14] I.S. Gilmore, F.M. Green, M.P. Seah, *Surf. Interf. Anal.* 39 (2007) 817.
- [15] D. Weininger, *J. Chem. Inf. Comput. Sci.* 28 (1988) 31.
- [16] R.K. Lindsay, B.G. Buchanan, E.A. Feigenbaum, J. Lederberg, *Applications of Artificial Intelligence for Organic Chemistry: The Dendral Project*, McGraw-Hill, New York, 1980.
- [17] J. Gasteiger, W. Hanebeck, K.-P. Schulz, *J. Chem. Inf. Comput. Sci.* 32 (1992) 264.
- [18] F.M. Green, J.L.S. Lee, I.S. Gilmore, M.P. Seah, NPL Report DQL-AS 029. Available at <http://www.npl.co.uk/server.php?show=ConWebDoc.2324>, 2006.
- [19] J.C. Vickerman, D. Briggs, A. Henderson, *The Static SIMS Library*, SurfaceSpectra Ltd, Manchester, 1998.
- [20] SMILES, Daylight Chemical Information Systems, <http://www.daylight.com/smiles/>.
- [21] SMILES, Daylight Chemical Information Systems, Depict <http://www.daylight.com/daycgi-tutorials/depict.cgi>.
- [22] Molinspiration, <http://www.molinspiration.com/cgi-bin/properties>.
- [23] S.W. Karickhoff, L.A. Carreira, S.H. Hilal, SPARC on-line calculator, <http://ibmlc2.chem.uga.edu/sparc/>.
- [24] Molar mass and mass spectral isotope pattern <http://www.colby.edu/chemistry/NMR/scripts/smileMM.html>.
- [25] F.M. Green, I.S. Gilmore, M.P. Seah, *Appl. Surf. Sci.*, accepted.