

## *PhD Dissertation 07/2011*

### **Integrated Analytical and Computer Tools for Toxicant Identification in Effect-Directed Analysis**

Emma Schymanski

# **Integrated Analytical and Computer Tools for Toxicant Identification in Effect-Directed Analysis**

Von der Fakultät für Chemie und Physik

der Technischen Universität Bergakademie Freiberg

genehmigte

**DISSERTATION**

zur Erlangung des akademischen Grades

Doctor rerum naturalium

(Dr. rer. nat.)

vorgelegt

von Emma Schymanski B.Sc. (Hons) B.E. (Hons)

geboren am 26.05.1980 in Adelaide, Australien

Gutachter: Prof. Dr. rer. nat. habil. Gerrit Schüürmann, Freiberg  
Prof. Dr. rer. nat. habil. Matthias Otto, Freiberg

Tag der Verleihung: 21.04.2011



## Summary

Effect-Directed Analysis (EDA) uses fractionation combined with chemical and biological analysis to isolate and identify toxicants in complex environmental samples. This thesis describes methods to identify toxicants based on gas chromatography electron impact mass spectrometry (GC-EI-MS), where the spectrum may not necessarily be present in a database. Structure generation and mass spectral substructure classifiers are used to generate all possible structures matching the spectral data. Several filtering criteria are then used to select or eliminate candidates based on the analytical information, including spectral match values, partitioning behaviour, retention behaviour and steric energy. The use of these criteria was investigated and combined into an automated sequence to streamline the identification of unknown GC-EI-MS spectra. Theoretical examples and actual EDA samples were used to develop and test the method, which was subsequently used to tentatively identify compounds in active EDA fractions.

## Extended Summary

Effect-Directed Analysis (EDA) combines biological and chemical analysis to isolate and identify compounds in complex samples that are toxicologically significant to specific organisms. Without the identification of significant unknowns, comprehensive confirmation of the observed effects cannot take place in many cases. Here, identification refers to the assignment of a structure to a given spectrum, while confirmation involves gathering additional evidence to support the identification. In EDA, the amount and purity of the sample is usually sufficient for identification using chromatographic techniques coupled with mass spectrometry, but not often for other analytical techniques such as nuclear magnetic resonance or infra red spectroscopy that may yield additional structural information. As a result, chemical analysis in EDA studies typically start with gas chromatography coupled to electron impact mass spectrometry (GC-EI-MS), where identification of spectra usually involves a database search followed by confirmation based on the retention time of an analytical standard. This thesis presents an alternative method to tentatively identify unknown compounds measured using GC-EI-MS without necessarily relying on a good database match.

A groundwater EDA study performed by C. Meinert, which included fractionation according to lipophilicity, was used as a source of unknown spectra to develop the initial method. The concept of structure generation was used to generate all possible candidates matching the analytical information, as an alternative to database searching. The program MOLGEN-MS and the NIST mass spectral database were used to determine the presence

and absence of substructures based on the unknown GC-EI-MS spectra. This information was then used within MOLGEN-MS to calculate the matching molecular formula(e) and then generate all possible structures matching the formula and substructure restrictions. A predicted spectrum was then calculated for each candidate based on fragmentation rules, which was then compared with the experimental spectrum to determine a match value to rank each candidate. The predicted logarithm of the octanol-water partitioning coefficient ( $\log K_{ow}$ ) of the structures was used in combination with the mass spectral match value to select or eliminate candidates for further consideration. While this initial method reduced the number of candidates matching the spectrum by several orders of magnitude compared with the number of structures possible given the molecular formula alone, there were still many cases with too many candidates remaining for identification purposes.

The results from the initial method development indicated that the predicted spectra and the resulting match values calculated by MOLGEN-MS were not always very useful for structure elimination. Thus, alternative programs to predict the mass spectral fragmentation were assessed to determine if more advanced prediction techniques and additional reactions could improve the use of predicted mass spectra to select or eliminate candidate structures. Two additional programs, Mass Frontier and ACD MS Fragmenter, were used with different settings and compared with the results of MOLGEN-MS using spectra taken from the NIST MS database. The outcome indicated that in fact the simplest spectral prediction settings were the best in terms of selecting the correct candidates. Although higher match values resulted from more complicated program settings, this applied to all candidates, not just the correct ones. As this investigation showed that the prediction selectivity in MOLGEN-MS was amongst the best of all programs and settings assessed, the incorporation of other criteria into the overall method described above is necessary to improve the selection of the correct structural candidate.

A theoretical example using 29  $C_{12}H_{10}O_2$  isomers taken from the NIST database was used to assess possible method extensions to improve the elimination of incorrect candidate structures. This included the use of modified substructure classifier selection prior to structure generation as well as a Lee Retention Index - boiling point correlation and steric energy calculation post structure generation. 19 of the 29 compounds were purchased and measured within the Department to supplement the theoretical example with experimental data, especially on the use of retention behaviour in candidate selection. The incorporation of these additional criteria in the overall method improved the elimination of structures by several orders of magnitude, in many cases to the correct group of substitution isomers.

Finally, the method developed as part of this work was used to tentatively identify unknown spectra from real EDA studies conducted within the Department of Effect-Directed Analysis at UFZ. Of 71 unknown spectra isolated during the Bitterfeld groundwater EDA performed by C. Meinert, 52 could be tentatively identified using structure generation, substructure classifiers and log  $K_{ow}$  ranges. 20 of these spectra did not have any suitable database match and would otherwise not have been identified. An EDA of diclofenac exposed to sunlight, performed by S. Weiss, resulted in the isolation of one transformation product responsible for the sample toxicity. This product was tentatively identified using the methods here as 2-[2-(chlorophenyl)amino]benzaldehyde (CPAB) and confirmed analytically and toxicologically by T. Schulze et al. An EDA of river water collected using blue rayon as a passive sampler, conducted by C. Gallampois, was used to source unknown spectra associated with toxic fractions. Two peaks of interest were detected in an active fraction and tentatively identified as phthalimide and phthalic anhydride using the methods developed as part of this work. These were confirmed analytically using GC-MS and LC-MS/MS by C. Gallampois. These three examples show the feasibility of the methods developed as part of this work in identifying unknown spectra measured using GC-EI-MS without relying on database matches.

## List of Publications Arising from this Work

Brack, W., Schmitt-Jansen, M., Machala, M., Brix, R., Barcelo, D., Schymanski, E., Streck, G., Schulze, T. (2008). How to confirm identified toxicants in effect-directed analysis, *Analytical and Bioanalytical Chemistry*, 390 (8), 1959-1973.

Schymanski, E. L., Meinert, C., Meringer, M., Brack, W. (2008). The use of MS classifiers and structure generation to assist in the identification of unknowns in effect-directed analysis. *Analytica Chimica Acta*, 615 (2), 136-147.

Schymanski, E. L., Bataineh, M., Goss, K.-U., Brack, W. (2009). Integrated analytical and computer tools for structure elucidation in effect-directed analysis, *Trends in Analytical Chemistry*, 28 (5) 550-561.

Schymanski, E. L., Meringer, M., Brack, W. (2009). Matching Structures to Mass Spectra Using Fragmentation Patterns: Are the Results As Good As They Look?, *Analytical Chemistry*, 81 (9), 3608-3617.

Meinert, C., Schymanski, E., Küster, E., Kühne, R., Schüürmann, G., Brack, W. (2010). Application of preparative capillary gas chromatography (pcGC), automated structure generation and mutagenicity prediction to improve effect-directed analysis of genotoxicants in a contaminated groundwater, *Environmental Science and Pollution Research*, 17, 885-897.

Schulze, T., Weiss, S., Schymanski, E., von der Ohe, P.C., Schmitt-Jansen, M., Altenburger, R., Streck, G., Brack, W. (2010). Identification of a phytotoxic photo-transformation product of diclofenac using effect-directed analysis, *Environmental Pollution*, 158, 1461-1466.

Schymanski, E. L., Meringer, M., Brack, W. (2011). Automated Strategies to Identify Compounds on the Basis of GC/EI-MS and Calculated Properties, *Analytical Chemistry*, 83, 903-912.

Schymanski, E., Schulze, T., Hermans, J., Brack, W. (2011). Chapter 8: Computer Tools for Structure Elucidation in EDA, *Handbook of Environmental Chemistry: Effect-Directed Analysis of Complex Environmental Contamination*, Vol. 15, Ed. W. Brack, Springer-Verlag Berlin Heidelberg, Germany.

Schymanski, E. L., Gallampois, C., Brack, W. (2011). Identification of Unknown GC/EI-MS Spectra in Elbe EDA Study, *in preparation*.

## Table of Contents

|                                                                                 |           |
|---------------------------------------------------------------------------------|-----------|
| <b>SUMMARY .....</b>                                                            | <b>I</b>  |
| <b>EXTENDED SUMMARY .....</b>                                                   | <b>I</b>  |
| <b>LIST OF PUBLICATIONS ARISING FROM THIS WORK .....</b>                        | <b>IV</b> |
| <b>1 INTRODUCTION.....</b>                                                      | <b>1</b>  |
| <b>2 BACKGROUND.....</b>                                                        | <b>3</b>  |
| 2.1 EFFECT-DIRECTED ANALYSIS .....                                              | 3         |
| 2.2 ANALYTICAL TECHNIQUES USED IN EDA.....                                      | 4         |
| 2.3 DATABASE SEARCHING AND MASS SPECTRA .....                                   | 5         |
| 2.4 IDENTIFICATION OF THE MOLECULAR FORMULA .....                               | 8         |
| 2.5 STRUCTURE GENERATION .....                                                  | 9         |
| 2.6 SUBSTRUCTURE CLASSIFIERS .....                                              | 10        |
| 2.7 ASSESSMENT OF SPECTRAL MATCH .....                                          | 11        |
| 2.7.1 <i>Programs to Calculate Fragments.....</i>                               | <i>11</i> |
| 2.7.2 <i>Assessing Predicted Spectra.....</i>                                   | <i>12</i> |
| 2.7.3 <i>Ranking Structural Candidates .....</i>                                | <i>15</i> |
| 2.8 ADDITIONAL CRITERIA .....                                                   | 16        |
| 2.8.1 <i>Retention Indices .....</i>                                            | <i>16</i> |
| 2.8.2 <i>Partitioning Coefficients .....</i>                                    | <i>16</i> |
| 2.8.3 <i>Steric Energy .....</i>                                                | <i>17</i> |
| 2.8.4 <i>EDA Specific Information .....</i>                                     | <i>18</i> |
| <b>3 STRUCTURE GENERATION AND UNKNOWN SPECTRA.....</b>                          | <b>20</b> |
| 3.1 METHODS .....                                                               | 20        |
| 3.2 RESULTS.....                                                                | 23        |
| 3.2.1 <i>Example with a single spectrum.....</i>                                | <i>23</i> |
| 3.2.2 <i>Overall Results.....</i>                                               | <i>26</i> |
| 3.2.3 <i>Providing ‘lines of evidence’.....</i>                                 | <i>28</i> |
| 3.2.4 <i>Use of toxicity ‘classifiers’ .....</i>                                | <i>31</i> |
| 3.3 DISCUSSION.....                                                             | 31        |
| 3.4 OUTCOMES .....                                                              | 34        |
| <b>4 MASS SPECTRAL FRAGMENT PREDICTION.....</b>                                 | <b>35</b> |
| 4.1 METHODS TO COMPARE MASS SPECTRAL FRAGMENT PREDICTION .....                  | 35        |
| 4.1.1 <i>Program Settings and Abbreviations .....</i>                           | <i>35</i> |
| 4.1.2 <i>File and Input Preparation.....</i>                                    | <i>36</i> |
| 4.1.3 <i>Comparison.....</i>                                                    | <i>37</i> |
| 4.2 RESULTS OF FRAGMENT PREDICTION AND COMPARISON .....                         | 39        |
| 4.3 RESULTS OF SPECIFIC EXAMPLES .....                                          | 43        |
| 4.3.1 <i>Specific Example 1: C<sub>3</sub>H<sub>5</sub>O<sub>2</sub>Cl.....</i> | <i>43</i> |
| 4.3.2 <i>Specific Example 2: C<sub>5</sub>H<sub>12</sub>S<sub>2</sub>.....</i>  | <i>46</i> |
| 4.3.3 <i>Specific Example 3: C<sub>7</sub>H<sub>6</sub>Cl<sub>2</sub>O.....</i> | <i>49</i> |
| 4.3.4 <i>Using Classifiers to Eliminate Structure Candidates.....</i>           | <i>53</i> |
| 4.4 DISCUSSION OF FRAGMENT PREDICTION .....                                     | 54        |
| 4.4.1 <i>Does a High Match Value Mean a Better Ranking? .....</i>               | <i>54</i> |
| 4.4.2 <i>Match Value versus Assignment Quality Index .....</i>                  | <i>55</i> |
| 4.4.3 <i>Candidate Inclusion/Exclusion .....</i>                                | <i>57</i> |
| 4.4.4 <i>Which Program, Which Settings? .....</i>                               | <i>57</i> |
| 4.5 IMPLICATIONS AND CONCLUSIONS .....                                          | 58        |
| <b>5 STRATEGIES FOR STRUCTURE ELUCIDATION.....</b>                              | <b>60</b> |
| 5.1 CALCULATION OF MOLECULAR FORMULA .....                                      | 61        |
| 5.1.1 <i>Programs to Calculate Molecular Formulae .....</i>                     | <i>62</i> |
| 5.1.2 <i>Comparison of Molecular Formula Calculations .....</i>                 | <i>63</i> |



|           |                                                                                               |            |
|-----------|-----------------------------------------------------------------------------------------------|------------|
| 5.2       | RETENTION INDICES AND C <sub>12</sub> H <sub>10</sub> O <sub>2</sub> ISOMERS.....             | 66         |
| 5.2.1     | <i>Retention Index Measurement and Prediction.....</i>                                        | 67         |
| 5.2.2     | <i>Results and Discussion of Retention Index Measurements .....</i>                           | 68         |
| 5.3       | MATCH VALUE COMPARISON WITH C <sub>12</sub> H <sub>10</sub> O <sub>2</sub> ISOMERS.....       | 70         |
| 5.4       | STERIC ENERGY AS AN EXCLUSION CRITERION.....                                                  | 73         |
| 5.4.1     | <i>Calculation of Steric Energy.....</i>                                                      | 73         |
| 5.4.2     | <i>Steric Energy Distribution.....</i>                                                        | 73         |
| 5.5       | METHOD FOR STRUCTURE GENERATION AND PROGRESSIVE ELIMINATION .....                             | 76         |
| 5.5.1     | <i>Automatic Data Processing.....</i>                                                         | 77         |
| 5.6       | STRUCTURE GENERATION RESULTS FOR C <sub>12</sub> H <sub>10</sub> O <sub>2</sub> ISOMERS ..... | 78         |
| 5.6.1     | <i>Classifier Assessment .....</i>                                                            | 78         |
| 5.6.2     | <i>Combining Substructural Classifiers and Exclusion Criteria.....</i>                        | 80         |
| 5.7       | DISCUSSION OF STRUCTURE ELIMINATION STRATEGIES .....                                          | 82         |
| 5.7.1     | <i>Substructure Classifier Use.....</i>                                                       | 82         |
| 5.7.2     | <i>Boiling Point/Lee Retention Index Restriction .....</i>                                    | 83         |
| 5.7.3     | <i>Partitioning Coefficient (log K<sub>ow</sub>) .....</i>                                    | 83         |
| 5.7.4     | <i>Steric Energy .....</i>                                                                    | 83         |
| 5.7.5     | <i>Spectral Match.....</i>                                                                    | 84         |
| 5.8       | IMPLICATIONS AND CONCLUSIONS .....                                                            | 84         |
| <b>6</b>  | <b>SUCCESSFUL UNKNOWN IDENTIFICATION IN EDA STUDIES .....</b>                                 | <b>86</b>  |
| 6.1       | TENTATIVE IDENTIFICATION OF BITTERFELD GROUNDWATER CONTAMINANTS .....                         | 86         |
| 6.2       | DICLOFENAC AND TRANSFORMATION PRODUCTS .....                                                  | 89         |
| 6.3       | EDA OF ELBE RIVER WATER USING BLUE RAYON AS A PASSIVE SAMPLER .....                           | 92         |
| 6.3.1     | <i>Methods.....</i>                                                                           | 92         |
| 6.3.2     | <i>Results - General .....</i>                                                                | 93         |
| 6.3.3     | <i>BRIA1_19.208.....</i>                                                                      | 94         |
| 6.3.4     | <i>BRIA1_25.410.....</i>                                                                      | 98         |
| 6.3.4     | <i>Discussion.....</i>                                                                        | 103        |
| <b>7</b>  | <b>SUMMARY AND FUTURE WORK.....</b>                                                           | <b>105</b> |
| <b>8</b>  | <b>ACKNOWLEDGEMENTS .....</b>                                                                 | <b>108</b> |
| <b>9</b>  | <b>REFERENCES.....</b>                                                                        | <b>109</b> |
| <b>10</b> | <b>APPENDIX .....</b>                                                                         | <b>114</b> |
|           | APPENDIX 1: ADDITIONAL TABLES .....                                                           | 114        |
|           | APPENDIX 2: LIST OF SCRIPTS .....                                                             | 122        |

## List of Figures

|            |                                                                            |     |
|------------|----------------------------------------------------------------------------|-----|
| Figure 1:  | Scheme of Effect-Directed Analysis                                         | 4   |
| Figure 2:  | Example Mass Spectra                                                       | 7   |
| Figure 3:  | Spectrum of 2-propyn-1-ol                                                  | 13  |
| Figure 4:  | Processing of mass spectra using MOLGEN-MS and NIST                        | 21  |
| Figure 5:  | Unknown spectrum at 10.875 min, $\log K_{ow} = 4.37-4.85$                  | 21  |
| Figure 6:  | Comparison of experimental and predicted isotope peaks                     | 24  |
| Figure 7:  | All mathematically possible structures of formula $C_4H_2Cl_4$             | 25  |
| Figure 8:  | Predicted data for Run 1 and Run 3 for the unknown in Figure 5             | 26  |
| Figure 9:  | Number of structures without, with classifiers and with filtering criteria | 27  |
| Figure 10: | Unknown spectrum at 5.80 min, $\log K_{ow} = 2.72-3.20$ and NIST match     | 28  |
| Figure 11: | Unknown spectrum at 25.38 min, $\log K_{ow} = 4.10-4.37$ and NIST match    | 29  |
| Figure 12: | All cyclic structures for the unknown in Figure 11                         | 30  |
| Figure 13: | The SA2 substructure and the resulting structures generated                | 31  |
| Figure 14: | Scheme: matching structures to spectrum using fragmentation patterns       | 38  |
| Figure 15: | Average RRP for different programs and settings                            | 42  |
| Figure 16: | Spectra and structures for 6 NIST matches for $C_3H_5O_2Cl$                | 44  |
| Figure 17: | Matrices of the 6 $C_3H_5O_2Cl$ structures and spectra                     | 45  |
| Figure 18: | Spectra and structures for 11 NIST matches for $C_5H_{12}S_2$              | 47  |
| Figure 19: | Matrices of the 11 $C_5H_{12}S_2$ structures and spectra                   | 48  |
| Figure 20: | Spectra and structures for 12 NIST matches for $C_7H_6Cl_2O$               | 50  |
| Figure 21: | Matrices of the 12 $C_7H_6Cl_2O$ structures and spectra                    | 51  |
| Figure 22: | Using substructures and analytical properties to eliminate structures      | 60  |
| Figure 23: | Structures of IQ, 1NP and 2HA                                              | 63  |
| Figure 24: | Steric energy distribution of 1000 molecules                               | 74  |
| Figure 25: | Steric energy of the 29 $C_{12}H_{10}O_2$ isomers                          | 75  |
| Figure 26: | Database-independent identification strategy for GC-EI-MS data             | 76  |
| Figure 27: | Compound identification using GC-EI-MS and structure generation            | 78  |
| Figure 28: | Number of structures with (a) 95 % and (b) additional classifiers          | 79  |
| Figure 29: | Number of structures remaining following progressive exclusion             | 80  |
| Figure 30: | Influence of filtering criteria shown using Structure 15                   | 81  |
| Figure 31: | Influence of filtering criteria shown using Structure 9                    | 81  |
| Figure 32: | Selected tentative identifications from the Bitterfeld groundwater EDA     | 88  |
| Figure 33: | Unknown spectrum of a diclofenac transformation product                    | 90  |
| Figure 34: | Precursor-based substructure and resulting two structures generated        | 91  |
| Figure 35: | Comparison of unknown spectrum with CPBA standard                          | 92  |
| Figure 36: | Predicted data for 137 structures for BR1A1_19.208                         | 97  |
| Figure 37: | Predicted data for 561 structures for BR1A1_20.410                         | 100 |

## List of Tables

|           |                                                                                                           |     |
|-----------|-----------------------------------------------------------------------------------------------------------|-----|
| Table 1:  | Mutagenicity toxicophores and the representative substructures                                            | 18  |
| Table 2:  | <i>Daphnia magna</i> structural alerts and the representative substructures                               | 19  |
| Table 3:  | Classifier information for the unknown in Figure 5                                                        | 24  |
| Table 4:  | Abbreviations used to describe programs and settings used in Section 4                                    | 36  |
| Table 5:  | Match values for the correct structure for 27 spectra using all programs                                  | 40  |
| Table 6:  | RRPs for the correct structure for 27 spectra                                                             | 40  |
| Table 7:  | Match value quantiles calculated for 1000 randomly-selected spectra                                       | 41  |
| Table 8:  | Average parameters for all programs and settings for different data sets                                  | 42  |
| Table 9:  | MVs and AQIs for Specific Example 1                                                                       | 44  |
| Table 10: | Number of fragments predicted for 11 structures of Specific Example 2                                     | 49  |
| Table 11: | Average fragment data for all structures from Specific Example 3                                          | 52  |
| Table 12: | Average MVs, AQIs and RRP for the three Specific Examples                                                 | 53  |
| Table 13: | Total number of candidates and the RRP with and without classifiers                                       | 53  |
| Table 14: | Example programs for calculation of molecular formula                                                     | 62  |
| Table 15: | Calculation of the molecular formula for IQ, 1NP and 2HA                                                  | 64  |
| Table 16: | Calculation of the molecular formula of 1,2-dichloroethane                                                | 65  |
| Table 17: | 29 C <sub>12</sub> H <sub>10</sub> O <sub>2</sub> isomers retrieved from the NIST database                | 66  |
| Table 18: | Measured and predicted KRIs and LRIs for purchased C <sub>12</sub> H <sub>10</sub> O <sub>2</sub> isomers | 69  |
| Table 19: | MVs and RRP for 29 C <sub>12</sub> H <sub>10</sub> O <sub>2</sub> isomers                                 | 71  |
| Table 20: | The two ‘correct’ structures for Structure 16                                                             | 71  |
| Table 21: | Energy quantiles calculated for ChemBio3D and MOLGEN-QSPR                                                 | 74  |
| Table 22: | Selected tentative identifications from the Bitterfeld groundwater EDA                                    | 87  |
| Table 23: | Peaks of interest in the acid and neutral fractions of sample BR1                                         | 94  |
| Table 24: | NIST matches for BR1A1_19.208                                                                             | 94  |
| Table 25: | Predicted data for the final two candidates for BR1A1_19.208                                              | 97  |
| Table 26: | NIST matches for BR1A1_20.410                                                                             | 98  |
| Table 27: | Predicted data for the final eleven candidates for BR1A1_20.410                                           | 102 |

## List of Equations

|              |                                                          |    |
|--------------|----------------------------------------------------------|----|
| Equation 1:  | Match value (MV)                                         | 13 |
| Equation 2:  | Example match value calculation                          | 14 |
| Equation 3:  | Cosine dot-product ( $F_D$ )                             | 14 |
| Equation 4:  | Abundance window ( $W$ )                                 | 14 |
| Equation 5:  | Stein and Scott Composite Equation                       | 15 |
| Equation 6:  | $F_R$ term for Equation 5                                | 15 |
| Equation 7:  | Relative Ranking Position (RRP)                          | 15 |
| Equation 8:  | General Retention Index (RI)                             | 16 |
| Equation 9:  | Determination of $\log K_{ow}$ from capacity factor $k'$ | 17 |
| Equation 10: | Calculation of the capacity factor $k'$                  | 17 |
| Equation 11: | Ring and Double Bond count (RDB) for MOLGEN-MS           | 22 |

## 1 Introduction

The identification of unknown compounds in complex environmental samples remains a major analytical challenge. Despite ever-increasing numbers of databases and entries within them, these still only cover a fraction of all compounds produced commercially. Furthermore, many compounds are transformed when they reach the environment, such that metabolites and transformation products can be present in greater concentrations than the original precursor compound [1, 2]. Environmental investigations, including Effect-Directed Analysis (EDA), often involve gas chromatography coupled with mass spectrometry (GC-MS) due to the availability of well-established mass spectral libraries for tentative identification of compounds. If the spectrum is not present in the database, however, there is no widely accepted alternative for compound identification other than structure elucidation by hand, a robust but very time-consuming method. This thesis aims to integrate analytical and computer tools to identify unknown spectra measured using electron impact mass spectrometry (EI-MS) and apply this especially to the case of toxicant identification in EDA.

The major concepts applied in this work are reviewed in Section 2 (Background), including EDA, mass spectrometric techniques, the use of structure generation and substructure identification as an alternative to database searching and information on the additional criteria used for candidate selection. This material is presented in part in a review article [3] and book chapter [4]. The following three sections are presented in the approximate chronological order with which the work was performed, as the results of Section 3 stimulated the investigation performed in Section 4, which in turn prompted the incorporation of additional criteria covered in Section 5. Section 3 explores the use of structure generation combined with mass spectral classifiers, predicted mass spectra and compound partitioning behaviour to identify unknown compounds [5]. Unknown spectra from a groundwater EDA performed by C. Meinert [6] were used to develop these ideas. As a study by Kerber et al. [7] showed that the predicted spectra used to rank the candidates in Section 3 was not especially selective, the investigation presented in Section 4 was conducted using additional programs with more advanced spectral fragmentation prediction. Three programs were tested in terms of computer-based EI-MS fragment prediction and the influence of the results on structure identification. The results were published in 2009 [8] and showed that more advanced fragmentation prediction actually had a detrimental effect on the selectivity. Thus, additional information was needed to improve candidate selection. Section 5 comprises an evaluation of retention behaviour, boiling point prediction and steric energy calculations to improve candidate selection above those in Section 3. The resulting automated sequence of data evaluation for structural identification/elimination based on GC-EI-MS

was evaluated on known spectra [9]. Section 6 contains examples where the methods developed have been applied to identify unknown compounds and toxicants in Effect-Directed Analysis, including selected results from the groundwater EDA performed by C. Meinert [6], a study performed by T. Schulze, S. Weiss et al. [1] and a river water EDA performed by C. Gallampois [10-12]. This thesis concludes with Summary and Future Work.

This work would not have been possible without the data and contributions of fellow co-workers. Where not adequately represented by citations or in the Acknowledgements, explicit contributions are detailed where relevant in the text.

## 2 Background

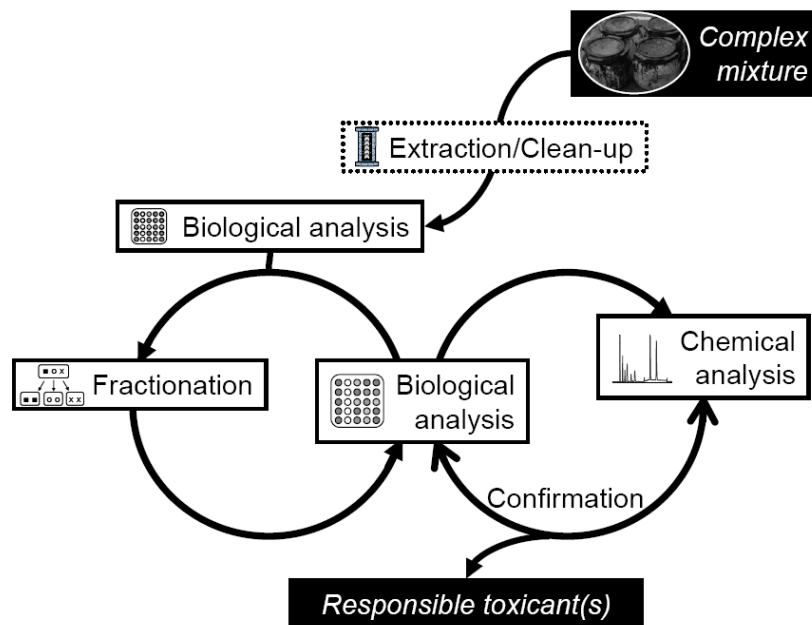
### 2.1 *Effect-Directed Analysis*

Effect-Directed Analysis (EDA) combines biological and chemical analysis to isolate and identify compounds in complex samples (e.g. waters, sediments, waste slurries, biota) that are toxicologically significant to specific organisms. Biological assays are used to assess the toxicity, while progressive fractionation steps are used to reduce sample complexity, until toxic (or active) fractions are suitable for detailed chemical analysis. As not all biotests or bioassays used can be directly linked to toxicity, the term ‘active’ fraction can be used as an alternative expression to represent fractions of interest. Following chemical analysis, which ideally results in the identification of compounds present in the active fractions, quantification and finally confirmation steps are required to ensure that the identified compounds are actually responsible for the toxicity observed [13]. Here, confirmation involves both analytical and toxicological confirmation. Analytical confirmation generally requires the measurement of a standard compound to compare with the unknown, where possible also using an orthogonal analytical technique. Toxicological confirmation involves testing of the identified substances with the same biotest or bioassay as the sample to determine whether these are responsible for the sample activity.

The processes involved in EDA are shown in Figure 1. The final step of toxicant identification and confirmation remains one of the challenges of undertaking a successful EDA study and often involves gathering evidence as to the identity of a compound (the weight-of-evidence approach), rather than a clear yes/no answer [14]. Without the identification of significant unknowns, comprehensive confirmation of the observed effects cannot take place. Several EDA studies published recently confirm this situation, reviewed in [3] and references within.

The fractions from EDA often remain relatively complex even after extensive fractionation, ruling out time-efficient manual evaluation of peaks of interest. The nature of the samples, clean-up and fractionation procedures results in samples suitable for analysis using chromatographic techniques coupled with mass spectrometry, but often not suitable for other analytical techniques that may yield additional structural information such as nuclear magnetic resonance (NMR) or infra-red (IR) techniques, due to the amount and purity of compounds present [14]. Thus, the identification of unknown compounds generally relies heavily on the information gained from one or more chromatographic separations followed by mass spectrometry (MS) and incorporation of as much information gained from the biotests and separation as possible [3]. Due to the

complexity of information and chromatograms, computer tools are increasingly necessary to support the analyst in the elucidation of the unknown structures.



**Figure 1: Scheme of Effect-Directed Analysis (modified from [13]). Substances identified using chemical analysis are confirmed analytically and toxicologically to determine responsible toxicants.**

## 2.2 Analytical Techniques used in EDA

Gas chromatography coupled with mass spectrometry (GC-MS) is generally a common analytical starting-point for EDA studies [13], due to the relative availability and ease-of-use of the method. While the majority of earlier EDA studies were based almost entirely on GC-MS (reviewed for example in [13, 15, 16]), recent studies also use GC-MS as a starting point for analysis [3], followed with additional analysis to take advantage of recent developments in chromatographic-mass spectroscopic techniques. The mass spectra generated from hard ionisation techniques (e.g. electron impact MS, EI-MS), often used in combination with GC, are often regarded as a ‘fingerprint’, as this technique can produce many fragments and therefore often unique and/or easily identifiable spectra. As a result, comparatively extensive databases of EI-MS spectra are available, e.g. the NIST [17] and Wiley [18] databases contain between them approximately 667,000 unique spectra in the latest versions. The reproducibility of the method is also conducive to the identification of substructures within the spectra of unknowns. The major disadvantage of GC-MS-based techniques is the limited range of compounds that can be analysed successfully. Volatile and stable compounds, or compounds derivitised to increase their volatility can be measured, however low volatility, thermally unstable or polar compounds cannot be analysed [19].

Liquid chromatography (LC)-based methods are used increasingly in the structure elucidation of toxicants, drug impurities, degradation products, metabolites and biologically active compounds in natural products (e.g. [20-24]), as they are suitable for a wider range of compounds including polar, thermo-labile and high molecular mass compounds. A recent study screening 89 drugs and their metabolites in clinical samples using LC-MS and GC-MS revealed that LC-MS could be used to identify most of the compounds (94 %), compared with GC-MS (64% of the compounds) [22]. Substructures associated with mutagenicity of 4337 environmentally-relevant compounds, reported by Kazius et al. [25] and shown in Table 1 (page 18), cover broad ranges of compounds not amenable to GC methods; hence these compounds would not be detected during mutagenicity-based EDA using GC methods alone. LC methods offer a variety of phases to achieve selective and efficient separation without derivatisation, avoiding the associated increase in spectral complexity (further information in [3]).

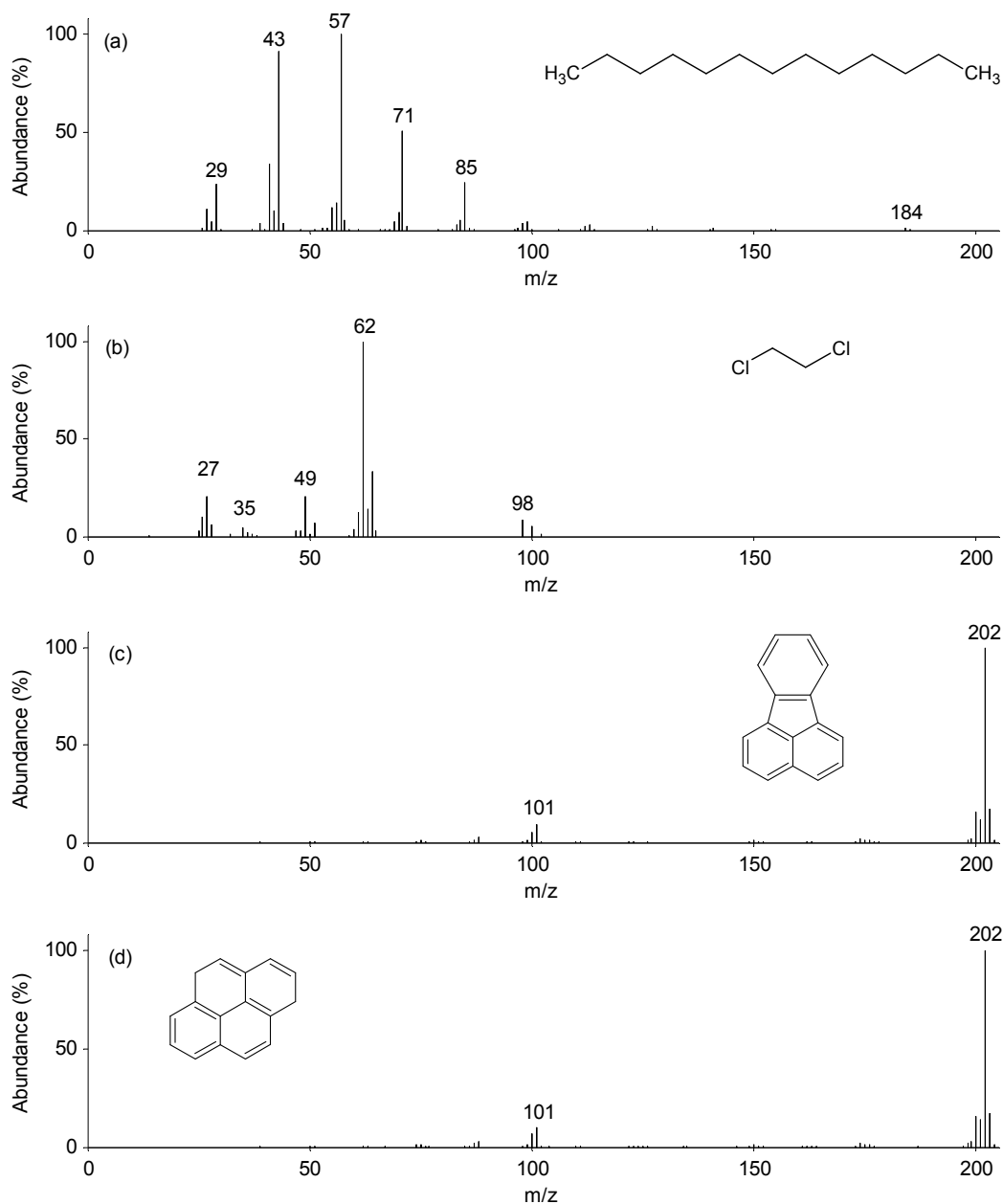
Although both chromatographic techniques are coupled to mass spectrometry, the resulting spectra are very different. While the hard ionisation spectra (e.g. EI-MS) yield reproducible fragmentation patterns, this is sometimes at the expense of the molecular weight ion, 'M'. In contrast, the mass spectra generated with soft ionisation techniques such as Chemical Ionization, Desorption Ionization and Atmospheric Nebulisation Ionization, often combined with LC, generally only yield molecular mass information without significant fragmentation. This is very useful in combination with high accuracy mass spectrometry, while tandem ( $MS^2$  or MS/MS) or even multi-stage ( $MS^n$ ) spectra are used to give additional fragmentation information. These spectra are generally less reproducible than the spectra produced by EI-MS [4], which are generally reproducible across different instruments and laboratories. As a result of this and the generally good availability of programs for EI-MS interpretation, this work concentrates mainly on the interpretation of GC-EI-MS spectra. As alternative higher accuracy MS-based methods are becoming more established, with many new programs released in the last year, the possible extension of the methods developed here to high accuracy  $MS^n$  techniques is discussed in Section 7. The determination of the molecular formula based on exact mass is also discussed briefly in Section 5.1.

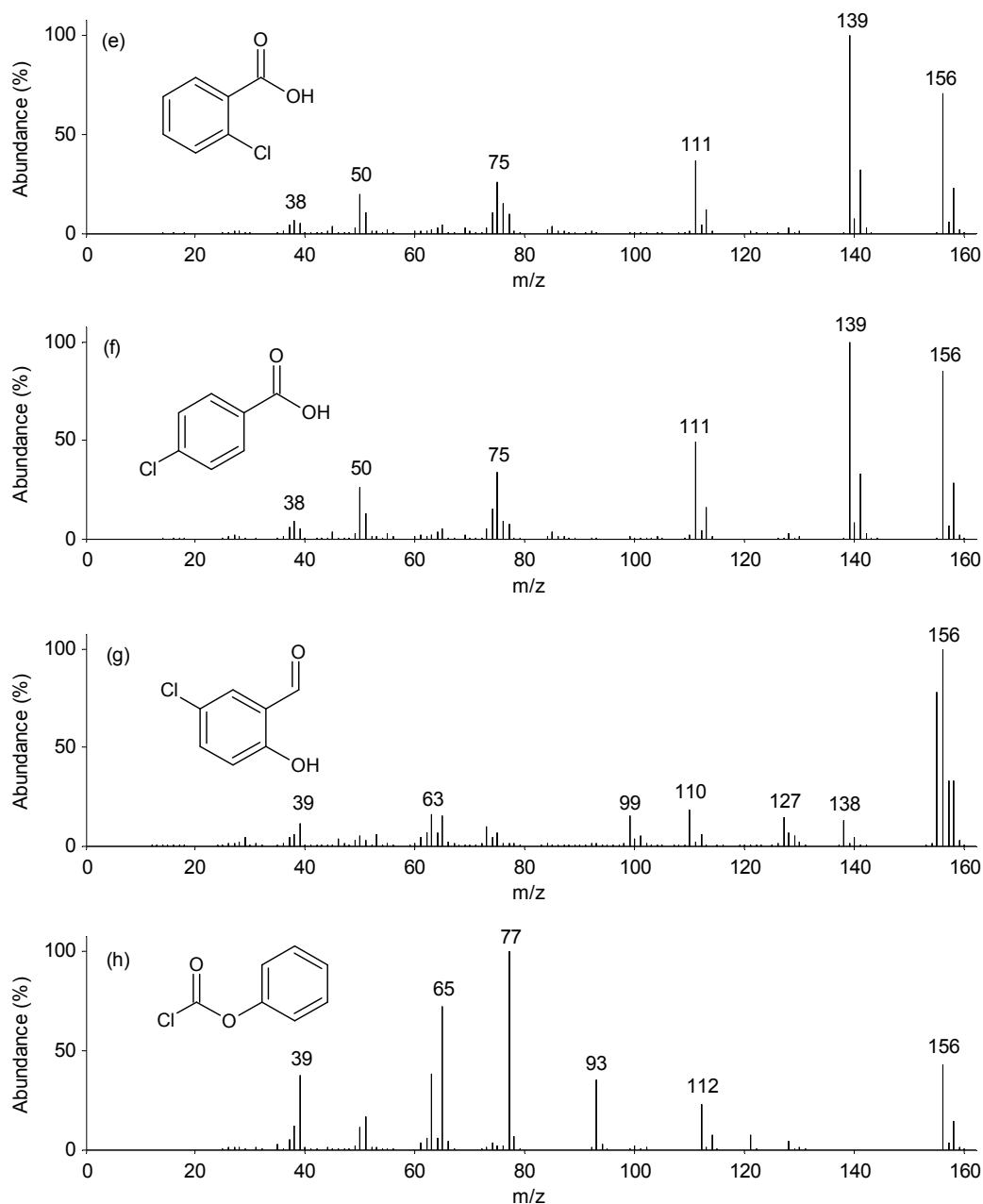
### ***2.3 Database Searching and Mass Spectra***

Most GC-EI-MS instruments come with software linked either to one of the commercial databases or to their own internal database, such that implementation of database searches is very easy for the user, often as simple as one click on the chromatogram. During a search, the measured mass spectrum is compared with those in the database, generating, in the case of NIST [17], a match factor, reverse match factor and a probability that this spectrum is the 'right' match. The match factor and reverse match



factor give an indication of how well the mass peaks (reported as  $m/z$ , the mass-to-charge ratio) and their magnitudes match, excluding and including  $m/z$  not measured in the experimental spectrum, respectively. Similar spectra should thus have very high values for both. The probability, however, is relative to all other spectra in the database and is thus more subjective. If the experimental spectrum is distinctive and very similar to only one spectrum in the database (e.g. Figure 2(b)), a match is usually associated with a high probability, however if there are similar spectra for different compounds, only a low probability is possible because the match could be one of several spectra. This is often seen for isomers, e.g. polycyclic aromatic hydrocarbons (PAHs), Figure 2(c) & (d), substituted aromatics, e.g. Figure 2(e) & (f) and alkanes, Figure 2(a).





**Figure 2: Example mass spectra. (a) Tridecane, showing the typical alkane chain pattern. (b) 1,2-dichloroethane as an example of a distinct spectrum for the given formula. (c) Fluoranthene and (d) pyrene are almost identical spectra for different compounds of the same formula. Spectra (e) 2-chlorobenzoic acid and (f) 4-chlorobenzoic acid differ from each other only in peak magnitudes, while compounds with the same formula (g) 5-chloro-2-hydroxy-benzaldehyde and (h) phenyl-chloro-carbonate have distinctly different fragmentation patterns in the spectra. Spectral data from the NIST 2005 database [17], spectrum numbers 61976, 114952, 228362, 227992, 228871, 228878, 231382 and 292166, respectively.**

If there are no exact matches but some similar spectra, it is also possible to have a high probability of a match for spectra that do not match well visibly, because the probability compares with the spectra available (see for example Figure 11, page 29).

The down side of database searching based on EI-MS spectra is that the spectra are not necessarily unique (e.g. Figure 2(c) and (d)), while the mass peak ratios in spectra measured on different type of instruments (e.g. quadrupole versus ion trap) can vary dramatically, such that measurement differences can exceed differences between spectra of similar compounds. This may even happen for instruments of the same type when tuned for maximal sensitivity at a specific mass or wrong mass range [4]. Differences in mass peak abundance also affect the database search results, as the search takes both mass peak presence and its relative abundance into account (see Equations 3 to 6). Furthermore, the searching algorithms used are such that the results are quite trustworthy for compounds within the database but are less reliable for spectra outside the database domain. Although 200,000 spectra in a database (NIST 2005) seems an impressive number, this is a tiny sub-fraction of the number of different chemicals produced (estimated at around 8,400,000 in 2006 [26]), let alone their breakdown products and metabolites. Online databases are becoming more common and comprehensive (ChemSpider, for example, has over 59 million entries [27]), but these are not generally linked with spectral information. It is also estimated that the molecular ion ('M') peak may be missing from up to 30% of all EI-MS spectra [28], which makes estimation of even the molecular weight of unknown compounds challenging. Thus, while GC-MS coupled with a database search is a good starting point, more information is generally necessary for identifying all compounds of interest in an EDA study.

Further information about databases for exact mass data,  $MS^n$  spectra and databases including unknown compounds can be found in [4].

## 2.4 Identification of the Molecular Formula

A number of strategies are available for calculating the molecular formula based on unit accuracy (i.e. masses reported to 1 Da) mass spectra, based mainly on the presence of isotope peaks for either the molecular ion ('M' peak) or fragments. Only a few, low molecular weight compounds have only one formula possible for a given unit resolution molecular weight. If the 'M' peak and its isotopes are sufficiently abundant, calculation of the molecular formula is relatively straightforward using for example MolForm (part of MOLGEN-MS [29]) or ACD Formula Generator (part of the MS Manager Suite [30]). Searching for the formula based purely on mass difference is not sufficient to isolate the correct formula for low resolution data, as the mass is only determined within one Dalton and therefore the closeness to the integer value is not indicative of composition.

In some EI-MS spectra, the isotope peaks of the 'M' peak are not available, or of such low abundance that calculations based on these would be inaccurate (see Figure 2(a) and (b)). An alternative is to calculate the formula based on the isotope peaks of fragment signals, also using the above-mentioned programs. For formula calculations based on the

'M' peak, only odd-electron ions need to be considered (as M has lost one electron to form  $M^{+}$ ), but both odd and even electron ions need to be considered for calculations based on the fragment peaks to account for the possible loss of atoms, not just an electron. An additional calculation module based on the whole spectrum has also been implemented in MOLGEN-MS, called "ElCoCo" [29]. These different strategies were compared as part of this work and are included in Section 5.1.

Calculation of the molecular formula based on accurate mass (i.e. reported to 0.001 Da or less) is generally performed by an assessment of the closeness of the formula exact mass to the measured mass, rather than by matching the isotope patterns. The exact masses of the elements are used to determine the combination of elements closest to the mass measured, taking into account the number of electrons in the measured ion. The isotope patterns are then used subsequently in formula selection, if needed. Further information is available for example in [31-33].

## 2.5 *Structure Generation*

Once the molecular formula (or possible formulae) has been determined, structure generation can be used to determine the compound identity. Structure generation provides a completely different approach to dealing with unknown compounds. This approach is database independent, i.e. the outcome does not depend on the number or quality of spectra within a database. All possible structure candidates for the given formula(e) are considered, which allows the user to see how many structures fit the data, rather than taking the first (or 'best') match and thereby overlooking other possible candidates merely because they were absent from the database. However, sufficient substructural information is required (e.g. based on MS fragmentation patterns) to avoid the generation of thousands to millions of possible structures. In general, the larger the molecular formula, the greater the number of structural possibilities, so where insufficient information is available from the spectrum, candidate selection and elimination become critical to a successful tentative identification.

Structure generation itself requires only a molecular formula (or even several) as a bare minimum. For this simplest case, basic structure generation programs can be used. One such program is MOLGEN (several versions available, e.g. [34-36]), which generates all mathematically possible structures for the input formula(e). Although it may be interesting to know how many structures could be possible for an unknown, this is not generally very useful for identification purposes and the number of possible structures can in fact be overwhelming. Thus, including as much information as possible into the structure generation procedure becomes necessary to avoid data overload. The method presented in this thesis is primarily based on this principle, i.e. the inclusion of as much analytical information as possible from the EDA study to allow the identification of

unknown compounds. The different information and methods included in this study are described in the next sub-sections.

## 2.6 *Substructure Classifiers*

Substructure identification based on common patterns within EI-MS spectra has been under development for many years, such that simple EI-MS interpretation is incorporated in most undergraduate Analytical Chemistry courses. Several books detailing MS interpretation are available (e.g. [37]) and a review of the general concepts of substructure identification and early programs (not limited to MS) is given in [38]. Although manual interpretation of the mass spectrum is still essential as part of the data evaluation process, this thesis aims to optimise the choice and sequence of available programs to maximise the identification accuracy while performing tasks in a streamlined manner. The focus is therefore on assessing the ability of various programs to perform this task, rather than an evaluation of the rules themselves explicitly, covered in Sections 3, 5.6 and 5.7.

Several years ago, Varmuza, Werther and co-workers developed database-independent substructure identifiers (or ‘classifiers’) [39, 40]. These classifiers assign percentage likelihood to the presence or absence of given substructures, based on the experimental spectrum. A training set of 300 spectra was used for each ‘classifier’ (150 with, 150 without the desired feature), with a testing set of another 300 spectra (again 150 with and 150 without the desired feature). Due to their database-independence, these substructural classifiers are now included in a number of programs such as the freeware AMDIS [41] for spectral deconvolution and MOLGEN-MS [29], which combines EI-MS interpretation and structural generation in one program. 160 of these database-independent classifiers are currently implemented in MOLGEN-MS. The search for substructures is performed by both programs in a very short time frame. If performing structure generation using MOLGEN-MS, the classifier information is automatically loaded into the next stage of the program (classifier selection and structure generation, see Figure 4, page 21), whereas information from AMDIS can be included manually into structure generation.

The NIST database [17] also incorporates a substructure search, which assigns probabilities to the presence/absence of substructures based on the experimental spectrum and spectra within the database using the nearest neighbour approach [42, 43]. Both the 2005 and 2008 NIST versions include 541 substructures. Additionally, NIST estimates the number of chlorine and bromine present and suggests possible molecular weight, with probabilities. The substructure information includes chemical elements as well as Ring and Double Bond counts (RDB, see Equation 11), both of which are useful for structure generation purposes.

The substructures present or absent can be used during structure generation to limit the number of structures generated. Substructures likely to be present with a given probability are included on a “good list” (i.e. included in the structure generation), whereas substructures likely to be absent with a given probability are excluded from structure generation via the “bad list”. MOLGEN-MS already performs this automatically for the Varmuza classifiers and this was extended to include the NIST substructural information as part of this work (see Section 5).

## 2.7 *Assessment of Spectral Match*

### 2.7.1 *Programs to Calculate Fragments*

An important method to match structures to an unknown spectrum is to predict the mass spectra of the potential structures and compare this with the experimental spectrum. General EI-MS fragmentation rules have been developed and published over several years (e.g. [37]) and have been incorporated into a number of programs, both commercial (e.g. Mass Frontier [44] and ACD MS Fragmenter, part of the MS Manager Suite [30]) and research-based releases (e.g. MOLGEN-MS [29], MASSIS [45], MASSIMO [46] and EPIC [47]). These have been extended in the commercial releases to incorporate other forms of MS, such that the predictive capacities also extend to protonation and deprotonation, cluster ion formation, alkali metal adducts and chemical ionisations. Again, the work presented here focused primarily on the generation of EI-MS fragments.

Three programs were assessed for the prediction of EI-MS fragments as part of this study. MOLGEN-MSF [48] (an advanced, command-line version of the spectral prediction module of MOLGEN-MS) uses general mass spectral fragmentation rules [7] but can also accept additional fragmentation mechanisms as optional input during calculation. Mass Frontier, developed by HighChem and marketed by Thermo Scientific Inc. [44] generates the predicted mass spectral fragments according to general (basic) fragmentation rules, to specific library rules (either from a user library or the library provided by HighChem), or both. The library provided with the software contains 19,000 mechanisms taken from the literature [49]. MS Fragmenter, part of the MS Manager package from Advanced Chemistry Development Inc. (ACD) [30] assigns generated fragments for a given structure to the given spectrum via the AutoAssignment option [50]. The output is a table of fragments and the ‘Assignment Quality Index’ (AQI), which summarises the percent of the spectrum assigned by the calculated fragments in terms of the total ion chromatogram, TIC.

The performance of an earlier version of MOLGEN-MSF was assessed [7] using 100 randomly-selected spectra from the NIST database [17] (1998 version) and generating all constitutional isomers matching the molecular formula of the spectrum. The results

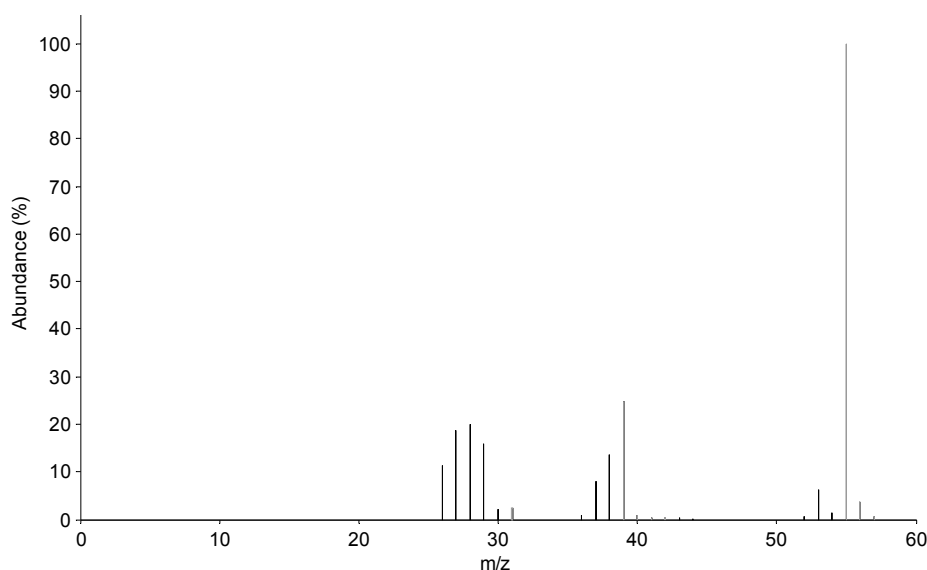
indicated that the use of general fragmentation rules alone was insufficient (in terms of accuracy) for automatic structure elucidation, but the authors suggested that the use of more sophisticated programs for virtual fragmentation may improve the ranking results [7]. Establishing whether this was the case was one of the aims of the comparison undertaken in this work (see Section 4). The assessment was based on the original study for comparability; the basic calculations are described in Section 2.7.2 and Section 4.

Recent studies focusing on high accuracy or tandem mass spectroscopic methods refer to both Mass Frontier and ACD MS Fragmenter for use in structure elucidation [51, 52], however these were based on a limited set of candidates such as matching database entries [51] or a given set of precursor ions [52], rather than all possible structures. Although alternative approaches exist to match structure to spectrum [52], software packages such as MASSIS [45], MASSIMO [46] and EPIC [47] were not available to us. The software FiD [52] for accurate mass tandem MS data shows promising first results compared with Mass Frontier and it may be interesting to investigate this approach further in future studies where accurate mass data is available. MetFrag [53] has also been developed recently for accurate mass data; like FiD it is based on bond dissociation energies rather than fragmentation reactions. Neither program has been validated so far for unit mass data.

### 2.7.2 *Assessing Predicted Spectra*

One way of assessing whether a proposed structure could match a mass spectrum is to calculate the possible mass spectral fragments resulting from the structure and match these, in terms of both occurrence and magnitude, to the fragment masses appearing in the experimental mass spectrum. This is shown in Figure 3, where experimental spectrum of 2-propyn-1-ol is shown in black overlaid by the predicted (calculated) spectrum using MOLGEN-MSF in grey (e.g. peaks at  $m/z = 31, 39, 40, 55, 56$ ). As the experimental spectrum is used to assign peak magnitudes to the predicted spectrum, fragments that are not present in the experimental spectrum are assigned a magnitude of zero and are thus not displayed (see Equation 1).

In Mass Frontier [44], the fragments calculated are used to generate a ‘bar code’ spectrum for the structure, i.e. all peaks are assigned an abundance of 100 %, shown for example in Figure 14. This can then be compared with the experimental spectrum via a graphical user interface (GUI). Although this visual comparison is user-friendly, this is not conducive to the objective evaluation of large datasets. The creators of Mass Frontier noted the difficulty in predicting energies and barriers in fragmentation to explain their use of bar code spectra rather than attempting to predict fragment magnitudes [49].



**Figure 3:** NIST spectrum (number 63617) of 2-propyn-1-ol (solid lines) with the fragments predicted by MOLGEN-MSF overlaid in dashed grey lines. Predicted spectrum magnitudes are assigned from the experimental spectrum (see Equation 1), so predicted fragments that are not present in the experimental spectrum are not displayed.

In contrast, the Assignment Quality Index (AQI) calculated by the ACD MS Manager [30] attempts to encompass both the presence and potential magnitude of predicted fragment peaks. According to the user manual, the AQI is defined as the ratio of the experimental total ion chromatogram (TIC) to the TIC calculated by MS Manager and is designed to estimate the coincidence of experimental and predicted peaks, the presence or absence of associated isotope peaks and the quality of fragment prediction [50]. The manual also states that it is not a “pure” assessment tool and can be both over- and understated for certain fractions. A more detailed description of the AQI or the actual calculation is not available, nor is a better reference. Hence, although calculation of the AQI was possible for fragments generated within MS Manager, calculation for fragments predicted using other programs was not possible, nor was a full evaluation of the AQI. As a result, the AQI has to be treated as a black-box calculation.

Kerber et al. [7] introduced the concept of a match value (MV) to assess the match of the predicted spectrum to the experimental spectrum. The magnitude of the experimental fragments is assigned to the calculated fragments, rather than attempting to predict the magnitudes. This is defined mathematically, including a derivation [7] and expansion (in German) [54]. The current version of MOLGEN-MSF uses a slightly modified version of the MV, shown in Equation 1:

$$\text{Equation 1} \quad MV = 1 - \sqrt{\frac{\sum_m (I(m) - x(m)I(m))^2}{\sum_m (I(m))^2}}$$



where  $MV$  is the match value,  $m$  is the mass to charge ( $m/z$ ) ratio of the fragment,  $I(m)$  is the intensity of the experimental mass spectral peak at  $m$  (scaled to the base peak to a value between 0 and 1) and  $x(m)$  indicates the presence/absence of predicted fragments such that  $x(m) = 0$  if there is no predicted fragment for  $m$  and  $x(m) = 1$  if there is a predicted fragment for  $m$ .

As an example, if the experimental spectrum in Figure 3 is simplified by taking only the peaks with abundance above 20 % to 28(20 %) 39(30 %) 55(100 %) and the predicted fragments 39 and 55 (excluding isotope peaks), the  $MV$  is:

**Equation 2** 
$$MV = 1 - \sqrt{\frac{(0.2 - 0)^2 + (0.3 - 0.3)^2 + (1 - 1)^2}{(0.2^2 + 0.3^2 + 1^2)}} = 1 - \sqrt{\frac{0.04}{1.13}} = 0.8119$$

Using the  $MV$ , the magnitude of the experimental mass spectral peaks is used to assess the fragment prediction, giving higher weighting to fragments predicted for larger peaks compared with smaller peaks. Thus, a structure which only produces one fragment that coincides with the base peak (100 % abundance) of the spectrum can have a higher match value than a structure with many fragments predicted that only represent minor peaks of the mass spectrum. This definition allowed the assessment of fragments predicted with Mass Frontier and ACD MS Manager as well as MOLGEN-MSF. The implementation of this is discussed in Section 4.

Although the work presented in this thesis used the value above, several alternative strategies are available to calculate the match of two mass spectra. Stein and Scott [55] assessed five different searching algorithms for accuracy in terms of matching library spectra, finding that a dot-product algorithm with additional weighting for the mass and intensity of the peaks was the best performer. This takes into account the concept that a peak of higher mass and intensity is more characteristic than peaks with lower mass or intensity. The equations for the best results are as follows:

**Equation 3** 
$$F_D = \frac{(\sum W_L W_U)^2}{\sum W_L^2 \sum W_U^2}$$

**Equation 4** 
$$W = I(m)^{0.6} m^3$$

where  $F_D$  refers to the dot product (cosine) used,  $W$  the abundance window (scaled according to intensity  $I(m)$  of the peak  $m$  (defined as above) and  $L$  and  $U$  refer to the library and unknown spectra, respectively.

Part of the study by Stein and Scott involved optimisation of the dot-product beyond the weighted intensity, into generation of a new composite algorithm that outperformed the optimised dot-product slightly [55]. The additional term, shown in the equations below, considers neighbouring peak intensities as well.

**Equation 5**

$$Composite = \frac{(N_U F_D + N_{L\&U} F_R)}{(N_U + N_{L\&U})}$$

**Equation 6**

$$F_R = \frac{1}{N_{L\&U}} \sum_i^{L\&U} \left( \frac{W_{L,i}}{W_{L,i-1}} \frac{W_{U,i-1}}{W_{U,i}} \right)^n$$

Here  $n = 1$  or  $-1$  when the term is less than or greater than unity, respectively and  $L\&U$  refers to a peak in both the library and unknown spectrum. The increase in performance resulting from the additional term in the composite algorithm compared with the optimised dot-product algorithm was relatively minor compared with the differences between the other algorithms, as was the effect of the optimisation of the dot-product weighting (in total 3 % improvement in the percent of molecules with the correct rank). Furthermore, as this compares the peak intensities of the unknown and library spectrum, while the MV assigns the experimental intensity to the predicted fragments, optimisation of an intensity term will not improve the performance of MV. As a result, investigations in this study were only conducted in terms of the MV defined in Equation 1.

### 2.7.3 Ranking Structural Candidates

Once the MVs are calculated for all structures, the ranking of the correct structure with respect to the other constitutional isomers needs to be determined. Kerber et al. [7] defined the relative ranking position (RRP), shown in Equation 7:

**Equation 7**

$$RRP = \frac{1}{2} \left( 1 + \frac{BC - WC}{TC - 1} \right)$$

where  $BC$  denotes the number of better candidates, i.e. those with a higher match value than the correct structure,  $WC$  denotes the number of worse candidates and  $TC$  denotes the total number of candidate structures. The RRP ranges from 0 to 1, where  $RRP = 0$  if the correct structure is ranked first (i.e.  $BC = 0$ ),  $RRP = 0.5$  if  $BC = WC$  and  $RRP = 1$  if the correct structure is ranked last ( $WC = 0$ ). This definition means that structures with the same MV will also have the same RRP, avoiding the necessity of performing statistics on the ‘best case’ or ‘worst case’ rank (see e.g. [53], where conclusions were based on a worst case rank). The value of the RRP is especially clear for cases where all molecules end up with the same MV: a best case rank would represent the result overly positive (i.e. the correct structure is ranked number 1) and the worst case rank too negatively (the correct structure was the worst rank), whereas the RRP is 0.5, neither good nor bad.

## 2.8 Additional Criteria

### 2.8.1 Retention Indices

The retention time of a compound used to provide additional information for structure identification in chromatography. This is typically expressed as a retention index (RI), which is less instrument- and parameter-dependent. Two common indices used for GC-MS are available, with the application domain varying according to the columns used and compounds investigated. The Kovat's RI (KRI) is based on the C<sub>6</sub>- to C<sub>36</sub>-alkanes, while the Lee RI (LRI) is based on four two- to five-ring PAHs [56]. The general relationship is shown in Equation 8:

**Equation 8**

$$RI_x = 100 \left( n + \frac{T_n - T_x}{T_{n+1} - T_n} \right)$$

where  $RI_x$  is the retention index of compound  $x$  with retention time  $T_x$  and  $n$  refers to the number of carbon atoms of the alkane for KRI and the number of rings in the PAH for LRI. The retention times  $T_n$  and  $T_{n+1}$  are selected to bracket  $T_x$  (i.e.  $T_n < T_x < T_{n+1}$ ), otherwise extrapolation is needed to calculate the RI.

RIs calculated from experimental data with the appropriate standards can be compared with documented RIs of known compounds and thus provide additional evidence for structural identity where both the MS and RIs match well. Generally, a good spectral match coupled with a good RI match is considered as sufficient for identification of a compound, although confirmation of the identity using an orthogonal analytical method is preferable. A common error window for the KRI is  $\pm 20$  [4, 57]. For structures where no standards are available (e.g. when using structure generation to propose candidates), the RI needs to be predicted, which generally requires specialised knowledge and a large computation effort [58]. A generalised prediction of KRI based on compound class is available for all compounds within the NIST database [59], including error intervals which can be several hundred units. Eckel and Kind correlated the LRI with boiling point (BP) data [58] and found that 95 % of the compounds investigated had BPs within the range (LRI-10) and (LRI+50) °C. As a result, they concluded that unknowns with measured BPs outside the range (LRI-10 and (LRI+50) could be eliminated from consideration. BP predictions are more widely available than RI calculations and hence can also be applied to unknown structures, e.g. using EPISuite<sup>TM</sup> [60], which was used by Eckel and Kind and also here, see Section 5.2.

### 2.8.2 Partitioning Coefficients

Partitioning coefficients can also be used to eliminate structure candidates with very different properties to the unknown compound. The octanol-water partitioning coefficient ( $K_{ow}$ , typically expressed in the logarithmic form  $\log K_{ow}$ ) is a general starting point,

rating the hydrophilicity or lipophilicity of the compound. If the compound was found in a water sample, for example, highly lipophilic candidate structures (high  $\log K_{ow}$ ) would not be expected and could thus be eliminated.

Where fractionation in EDA is undertaken using reverse phase high performance liquid chromatography (RP-HPLC) using columns packed with a stationary phase containing long hydrocarbon chains (e.g. C<sub>8</sub>, C<sub>18</sub>) [61], a correlation between the retention time and the  $\log K_{ow}$  can be used to estimate the  $\log K_{ow}$  range of each fraction and hence a range for the compounds within each fraction. The general relationship is shown in Equation 9. The parameters  $A$  and  $B$  are determined by linear regression of the logarithmic capacity factor  $k'$  of several standard compounds, calculated according to Equation 10 [61]:

**Equation 9**  $\log K_{ow} = A + B \log k'$

**Equation 10**  $k' = \frac{t_R - t_0}{t_0}$

where  $t_R$  is the retention time of the standard compound and  $t_0$  the average time of solvent molecules passing through the system, obtained using an unretained organic reference compound (e.g. thiourea). Once  $A$  and  $B$  have been determined, the  $\log K_{ow}$  range for each fraction can be calculated, which can be used to assign a  $\log K_{ow}$  range for peaks within that fraction. Similarly to the boiling point calculation, the  $\log K_{ow}$  for each candidate structure can also be calculated using EPISuite™ [60]. The determination of the  $\log K_{ow}$  using RP-HPLC is generally considered for compounds with  $\log K_{ow}$  between 0 and 6 [61], with larger deviations in measured values for more hydrophobic compounds possible (e.g.  $\pm 0.5$  units), depending on measurement conditions [62].

### 2.8.3 Steric Energy





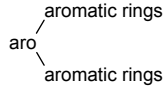
Structure generation can often lead to the generation of several candidates that, although mathematically possible, are highly unlikely to exist in complex samples for stability reasons. Although some restrictions can be introduced during structure generation to prevent the generation of these structures (e.g. exclusion of cycles containing only three or four atoms), the elimination of other structures, e.g. ‘bridging structures’ formed during generation of polycyclic compounds, is almost impossible. These molecules are, however, energetically unfavourable and could thus be eliminated from consideration using a simple energy calculation. This was trialled by T. Kind ([63], p.49, in German) during the generation of PAH structures based on the MM2 algorithm [64]. The choice of the MM2 algorithm is a compromise between speed and accuracy; while more accurate algorithms have been developed, the MM2 is fast and was sufficiently accurate to assist in the elimination of ‘unlikely’ structures [63, 65]. Thus this was the algorithm of choice for these studies. The calculation itself is based on the heat of formation, reported as kcal/mol. The standard deviation between experimental and calculated values for the

original algorithm was reported as 0.42 kcal/mol (c.f. the reported experimental error of 0.40 kcal/mol) and this compared favourably with other models at that time [64]. The MM2 calculation is integrated within the ChemBio3D Suite [66] and a force field approach to calculate the steric energy, similar to the MM2 approach, is used in MOLGEN-QSPR [67]. The results are also reported in kcal/mol. These programs were assessed in Section 5.4 of this thesis.

#### 2.8.4 EDA Specific Information

Toxicity information, based for example on the biotests used during EDA, can also be used to identify potentially excess-toxic compounds from candidate structures. High toxicity in a fraction may indicate the presence of an excess toxic compound [68]. Certain substructures can be associated with excess toxicity in some biotests, including the toxicophores for mutagenicity-based testing [25], shown in Table 1 and the structural alerts for *Daphnia magna* [68], shown in Table 2.

**Table 1: Toxicophores and the representative substructure based on a mutagenic assay. Table adapted from [25].**

| Toxicophore                                | Substructure                                                                        | Toxicophore                | Substructure                                                                          |
|--------------------------------------------|-------------------------------------------------------------------------------------|----------------------------|---------------------------------------------------------------------------------------|
| aromatic nitro                             |   | aromatic amine             | $\text{H}_2\text{N}-\text{aro}$                                                       |
| three-membered heterocycle                 |  | nitroso                    | $\text{N}=\text{O}$                                                                   |
| unsubstituted heteroatom-bonded heteroatom |  | azo-type                   | $\text{N}=\text{N}$                                                                   |
| aliphatic halide                           |  | polycyclic aromatic system |  |

Where these substructures are present, it is possible that the compound may exhibit excess toxicity (i.e. toxicity above the baseline level that can be predicted based on the  $\log K_{ow}$  alone) and may therefore be of interest in the outcome of an EDA study. Those compounds without the substructures are less likely to exhibit excess toxicity and are therefore potentially less likely to contribute to effects observed in an active fraction. As the effect is also dependent on the concentration, not just the inherent toxicity (narcotic or excess), this is a rather subjective criterion.

**Table 2: Structural alerts and representative substructures for *Daphnia magna* biotest. Modified from [68]. Explanation of the letters R, X, Y and Z are given in [68]**

| Structural Alert                     | Substructure | Structural Alert                                     | Substructure |
|--------------------------------------|--------------|------------------------------------------------------|--------------|
| $\alpha,\beta$ -unsaturated carbonyl |              | Simple carbamates                                    |              |
| $\alpha,\beta$ -unsaturated nitrile  |              | Thiocarbamates                                       |              |
| 1,1-halogenated-C=C                  |              | Dithiocarbamates                                     |              |
| Phosphorothionates                   |              | Rhodanin(e) derivatives                              |              |
| Thiophosphonates                     |              | Thiourea, methyl or dimethyl thiourea                |              |
| Phosphonates                         |              | N-alkyl thiourea                                     |              |
| Aliphatic thiols                     |              | Cyclic thiourea derivatives                          |              |
| Isothiocyanates                      |              | Primary or secondary anilines, no ortho substituents |              |
| Thiocyanates                         |              | Imide derivatives                                    |              |

In targeted EDA studies, for example degradation studies, compound precursor information can also be used to eliminate candidate structures. This is likewise a subjective strategy, dependent very much on the EDA study in progress.

The concepts discussed above were combined and implemented in various stages of the work, described in the next sections, with the aim to improve the identification of unknown compounds measured using GC-EI-MS based techniques, specifically in EDA studies.

### 3 Structure Generation and Unknown Spectra

This section explores the application of structure generation to identify unknown compounds, using spectra from a groundwater EDA study. This work was published in [5]:

Schymanski, E. L., Meinert, C., Meringer, M. and Brack, W. (2008) The Use of MS Classifiers and Structure Generation to Assist in the Identification of Unknowns in Effect-Directed Analysis, *Analytica Chimica Acta*, 615 (2), 136-147.

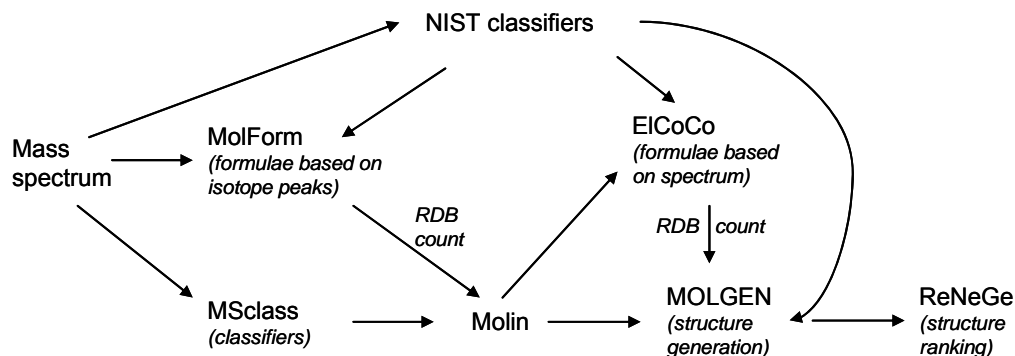
The methods, results and ensuing discussion are presented in this chapter, followed by Outcomes (Section 3.4), putting this into context with the rest of the thesis. The results of the groundwater EDA were published by Meinert et al. [6]; part of this paper including a comparison between database search identification and the structure generation approach is included in Section 6.1.

#### 3.1 Methods

The unknown spectra used in this section originate from an EDA carried out on a groundwater sample from Bitterfeld, Germany. Reverse-phase high performance liquid chromatography (RP-HPLC) was used for fractionation, with a C18 HD column (21x250 mm, Nucleosil 100-5, Macherey-Nagel, Düren, Germany), 15 mmol, pH3 phosphate buffer-acetonitrile as the mobile phase (20-80) and a flow of 0.5 mL min<sup>-1</sup>. Gas chromatography coupled with a mass spectrometry (GC-MS) was used to collect mass spectra (Model 6890 N, detector MSD 5973, Agilent Technologies, Waldbronn, Germany) with a HP-5MS capillary column (30 m x 0.25 mm I.D., 0.25 µm film, 5 % phenylmethyl-siloxane, Agilent Technologies) and temperature program 50 °C (held for 1 min), 5 °C min<sup>-1</sup> to 300 °C (held for 10 min). Mass spectra used in this study are reported by their log  $K_{ow}$  range and their retention time.

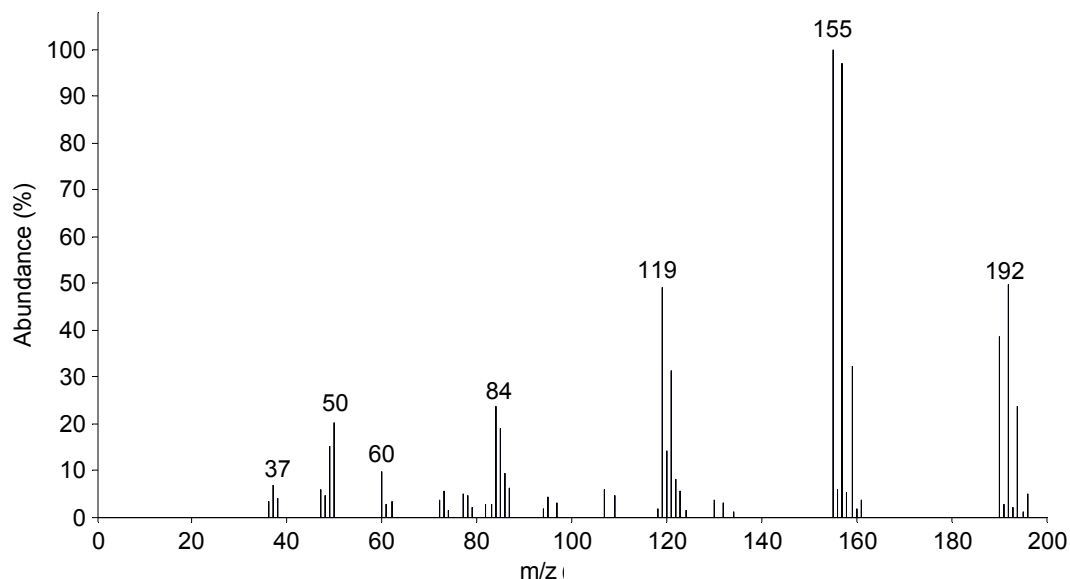
The Automated Mass Spectral Deconvolution and Identification System (AMDIS) was used to isolate mass spectra from experimental chromatograms for further analysis [41]. Mass spectra selected using AMDIS were saved in the MSP format (see e.g. [69] for the file description) and converted into a format suitable for import into MOLGEN-MS using the MATLAB [70] script ‘msp2csv’ (see Appendix 2). The NIST05 MS database [17] was used, via AMDIS, to perform a library search on each spectrum and to retrieve substructural information, the probabilities of presence of Cl and Br as well as probabilities regarding possible molecular weights of the compound.

Processing of the mass spectra in MOLGEN-MS took place over several steps, as shown in Figure 4.



**Figure 4: Processing of mass spectra using MOLGEN-MS modules and NIST.**

In the first stage, the substructures identified as present or absent to a given probability (default value 95 %) were classified using the module ‘MSclass’, based on the classifiers developed by Varmuza and co-workers [39]. Following this, the MOLGEN-MS module ‘MolForm’ was used to calculate possible molecular formulae where isotope peaks were present for the ‘M’ peak (e.g. see Figure 5). The possible formulae were ranked by MolForm, based on their fit to the isotope peaks of the molecular ion. Information on the presence or absence of elements, where available from NIST or Varmuza classifiers, was added to this calculation.



**Figure 5: Unknown spectrum with retention time 10.875 min, fraction  $\log K_{ow} = 4.37$ -4.85. Peak at  $m/z = 190$  is the ‘M’ peak; isotope peaks at  $m/z = 191$  to 196.**

The classifier information and molecular formula(e), where available from MolForm, were then loaded into the MOLGEN-MS module ‘Molin’, where ‘yes’ and ‘no’ classifier entries can be modified and are automatically checked for consistency with the given formula. Where no isotope peaks were present, or where the MolForm results were



unclear, possible molecular formulae were then calculated using the module ‘ElCoCo’ (short for *Elemental Composition Computation*). ElCoCo calculates and ranks formulae by how well they explain the experimental spectrum (not just the isotope peaks of the molecular ion), taking classifier information into account. As with MolForm, additional elemental composition information from NIST or MSclass (e.g. absence of S, at least 2 Cl present) can be incorporated into the calculations.

For MolForm and ElCoCo calculations with no clear result (common for formulae with many heteroatoms present), an additional criterion based on the Ring and Double Bond count (RDB) was used. The ‘ring and double bond equivalent’ (RDBE) concept (see for example [32]) was adapted to suit the conditions in MOLGEN-MS. Equation 11 was developed on the 11 elements present in MOLGEN-MS: C, H, Br, Cl, F, I, N, O, P, S and Si, with the following default fixed valencies: H, Br, Cl, F, I = 1; O, S = 2; N = 3; Si, C = 4 and P = 5:

$$\text{Equation 11} \quad RDB = \frac{(2 + 3n_P + 2(n_C + n_{Si}) + n_N) - (n_H + n_{Br} + n_{Cl} + n_F + n_I)}{2}$$

where  $n_x$  represents the number of atoms of element ‘X’ in the molecular formula. The user can input alternative valencies for P and S (with a correspondingly modified version of Equation 11), as these atoms have multiple allowable valence states in MOLGEN-MS. At this stage MOLGEN-MS is not able to recognise multiple valence states of N, although a ‘workaround’ using MOLGEN 3.5 [34] is possible. Generation of structures allowing multiple S and P valence states within one program run generally led to post-processing problems (e.g. ‘disappearing’ hydrogen atoms during file conversions), which were largely avoided by defining one valence state up-front and, where necessary, repeating program runs with different valence state combinations. The RDB count in Equation 11 was incorporated into the Matlab [70] script ‘RDB\_count’ (see Appendix 2) to eliminate formulae calculated with MolForm or ElCoCo that were inconsistent with the NIST or Varmuza classifiers (e.g. if a benzene ring with RDB = 4 is classified as ‘present’, formulae with RDB < 4 can be eliminated).

Following this, the classifier information and formula(e) were loaded into the ‘MOLGEN’ window, where ‘yes’ classifiers became ‘good list’ entries (substructures present) and ‘no’ classifiers ‘bad list’ entries (substructures forbidden in generated structures). Classifier information from NIST was incorporated at this stage, if not at the Molin stage, by manually adding additional ‘good list’ or ‘bad list’ entries. Additional restrictions (such as ‘permanent bad list structures’ or ‘cycles with greater than 5 members only’) can also be added here. All possible structures for the entered formulae within the given restrictions are calculated based on the MOLGEN 4.0 kernel [35].

In the final module ‘ReNeGe’ (*Reaction Network Generator*), theoretical mass spectral fragment ions are calculated for each of the generated structures and compared mathematically with the experimental spectrum to calculate a match value for each structure. These match values are then used to establish a ranking of the structure candidates. Structures were then viewed and exported, together with the match values, for further processing.

Post-ranking processing of the structures was conducted in Matlab, using the OpenBabel freeware package [71] to convert structure file formats as well as the EPISuite<sup>TM</sup> programs MPBPWin and Kowwin [60] to calculate melting point (MP) and boiling point (BP) and the logarithm of the octanol-water partitioning coefficient ( $\log K_{ow}$ ) values, respectively. Data summaries and plots were generated in Matlab to store and interpret the data. The scripts are listed in Appendix 2, example plots in Figure 8.

An additional Matlab script was written to check molecular formulae for compatibility with structural alerts identified by von der Ohe et al. [68], shown in Table 2. The output of this script is a list of structural alerts (if any) that match the formula, therefore allowing the user to add additional ‘good list’ structures to identify compounds exhibiting excess toxicity that match the mass spectrum. The script is listed in Appendix 2. MOLGEN-QSPR [67] can also be used to search generated structures for the exact structural alerts (including those in Table 1 and Table 2), outputting a list with the number of each substructure present in each structure.

## 3.2 Results

### 3.2.1 Example with a single spectrum

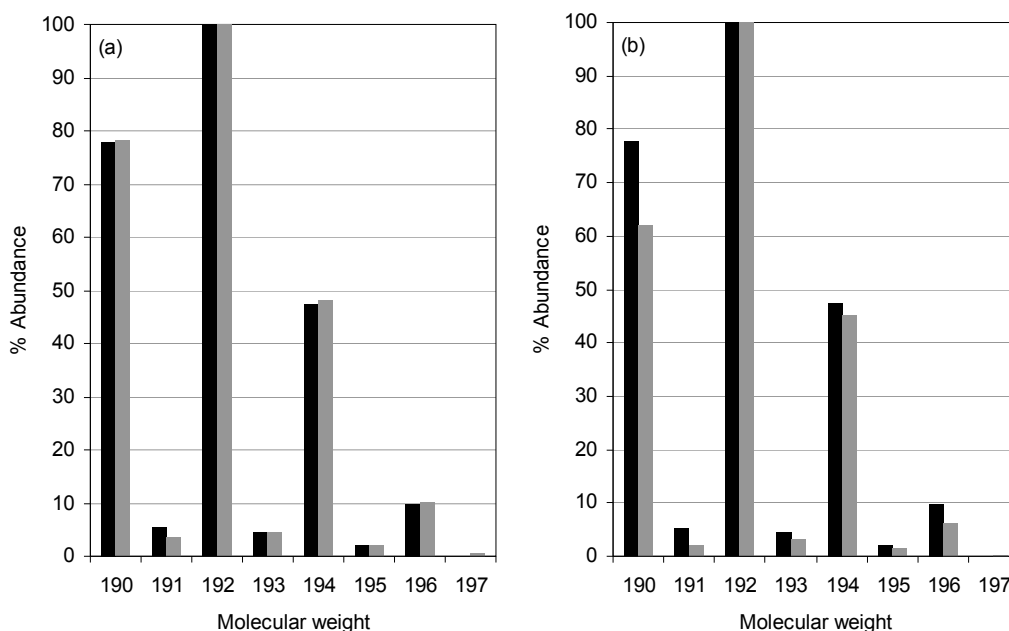
The unknown spectrum in Figure 5 is used here to demonstrate the method described above. The classifier information from MOLGEN-MS (Varmuza classifiers) and NIST for this spectrum is presented in Table 3.

Input of the formula classifier information (F, N, O, S, Si, P = 0; MW = 190) into ‘MolForm’ resulted in generation of 14 possible formulae (data not shown). The formula  $C_4H_2Cl_4$  was clearly the best-matching formula, both in terms of the isotope peaks and the classifier information. The next-best match,  $C_3H_5Br_1Cl_2$ , has markedly different isotope peaks, as shown in Figure 6.

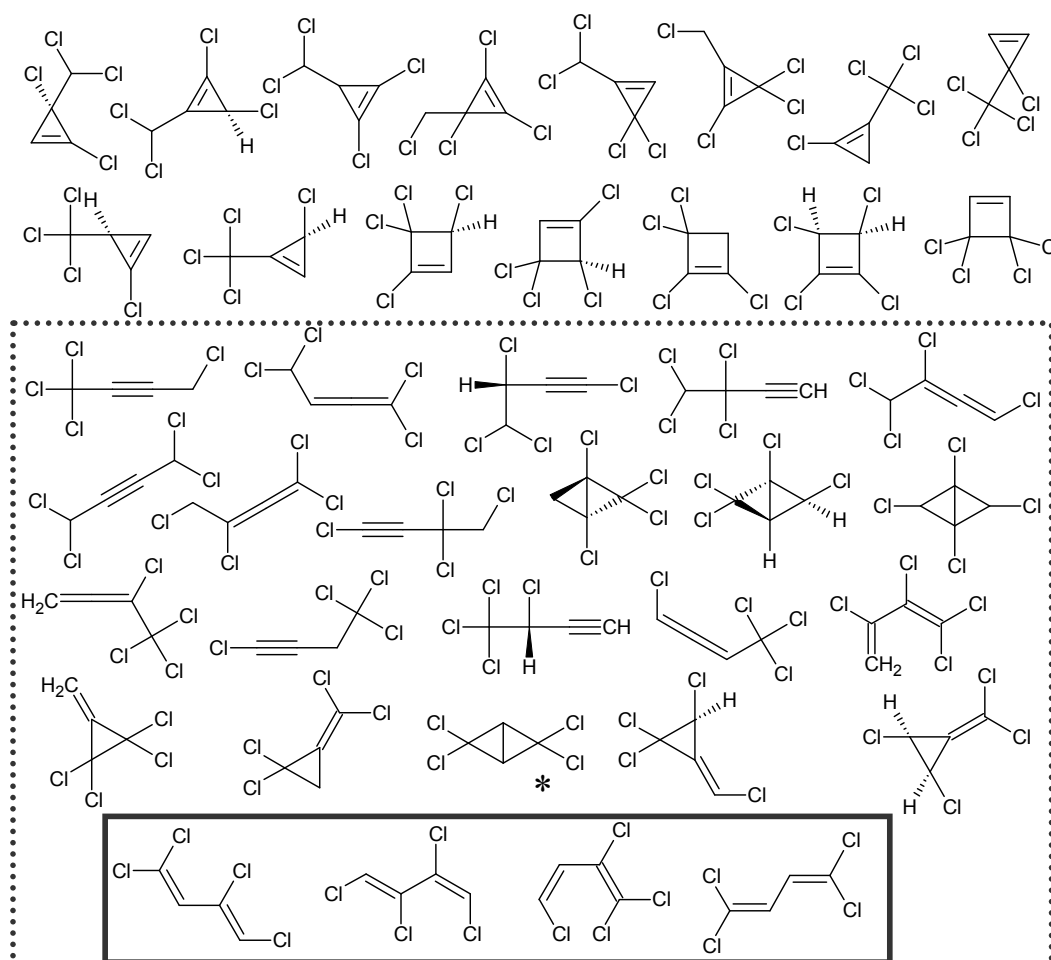
**Table 3: Classifier information for the unknown spectrum with retention time 10.875 min, log  $K_{ow}$  4.37–4.85. NIST and Varmuza classifiers to 95 % precision.**

| Formula Classifiers       |                                                                                                                                                                                                                                                                                                                |
|---------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Varmuza                   | Elements present: Cl (99 %)<br>Elements absent: $\geq 2$ Si (96 %)                                                                                                                                                                                                                                             |
| NIST                      | Molecular weight estimation: 190 (98 %); 192 (0 %)<br>Chlorine/Bromine information: Cl = 4 & Br = 0 (90 %)<br>Elements present: C, H & Halogen (99 %)<br>Elements absent: O, Si (99 %), N, P, S (98 %), F (97 %)                                                                                               |
| Substructure Classifiers* |                                                                                                                                                                                                                                                                                                                |
| Varmuza                   | Yes: None<br>No: C=C present in a ring                                                                                                                                                                                                                                                                         |
| NIST                      | Yes: C=C-C=C (99 %)<br>No: condensed ring, CH <sub>2</sub> , saturated compound, tertiary C in ring (99 %), C=C in ring, quaternary carbon, C only ring (98 %), unsubstituted C=C in chain, conjugated C=C in ring (96 %), bicyclic compound, 3-6 membered rings, exactly one double or triple C-C bond (95 %) |

\* Classifiers consistent with molecular formula only (others omitted for clarity)

**Figure 6: Comparison of the experimental isotope peaks of the M peak (black) with the top two predicted formulae (grey); (a)  $C_4H_2Cl_4$ ; (b)  $C_3H_5Br_1Cl_2$ .**

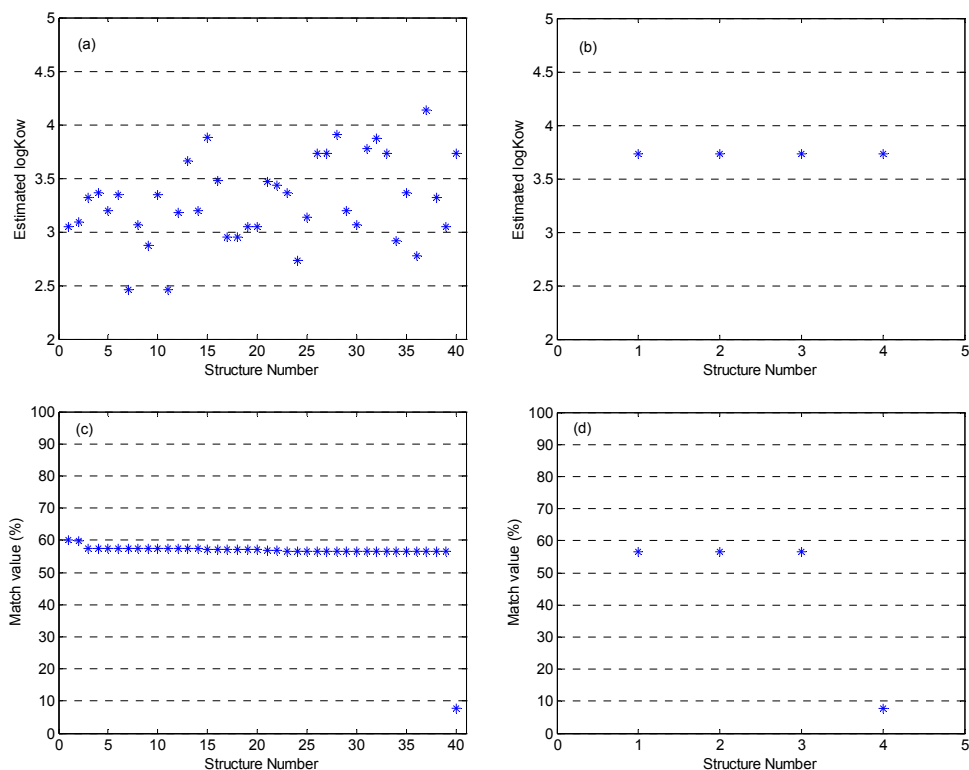
The program was then run three times with different substructure classifier combinations. Run 1 contained no classifier information, to calculate all (mathematically) possible structures for the given formula. Run 2 incorporated the Varmuza classifier information at the 95 % precision level, while Run 3 contained both the Varmuza and NIST classifier information at the 95 % precision level. 40 molecules were generated in Run 1. 25 molecules were generated with the addition of the Varmuza classifier ‘no C=C present in ring’ in Run 2. In Run 3, the NIST classifier ‘C=C-C=C’ was added to the MOLGEN good list and CH<sub>2</sub> to the bad list, resulting in generation of 4 matching molecules. The structures generated in the three runs are shown in Figure 7.



**Figure 7:** All (mathematically) possible structures generated in Run 1 for the formula  $C_4H_2Cl_4$  (with no classifiers, 40 structures). Structures within the dotted line are those generated in Run 2 (Varmuza classifiers, 25 structures). Structures within the solid line are those generated in Run 3 (Varmuza and NIST classifiers, 4 structures). \* indicates the structure with the highest  $\log K_{ow}$ .

Following generation of the structures, the estimated  $\log K_{ow}$  and match value to the experimental spectrum were calculated for each structure of each run. The results for Runs 1 (40 structures) and 3 (4 structures) are shown in Figure 8. In this case, the use of the Varmuza and NIST classifiers reduced the number of structures ten-fold, from 40 to 4. The four remaining structures have the same estimated  $\log K_{ow}$  values (3.73), which prevents the use of  $\log K_{ow}$  as a filtering criterion. However, one structure (1,1,4,4-tetrachloro-1,3-butadiene) has a significantly lower match value than the other structures, leaving three structures on which to focus confirmation studies. The estimated  $\log K_{ow}$  of 3.73 is considered close enough to the fraction range (4.37–4.85) when considering the errors associated both with the correlation of retention time and  $\log K_{ow}$  in the fractionation and the prediction of  $\log K_{ow}$  from the structure SMILES code in EPISuite<sup>TM</sup>. In this example, filtering the structures according to the  $\log K_{ow}$  prior to the use of classifiers and spectrum match value would identify Structure 37 of Figure 7(a),

with  $\log K_{ow} = 4.14$  as the structure closest to the fraction  $\log K_{ow}$  range. This structure is indicated by a star in Figure 7.



**Figure 8:** (a) Partitioning coefficient data for Run 1 (no classifiers); (b) Partitioning coefficient data for Run 3 (Varmuza and NIST classifiers, right); (c) mass spectrum match value for Run 1 and (d) mass spectrum match value for Run 3. Structure number is in order of highest (Structure 1) to lowest match value.  $\log K_{ow}$  values estimated using EPISuite<sup>TM</sup>.

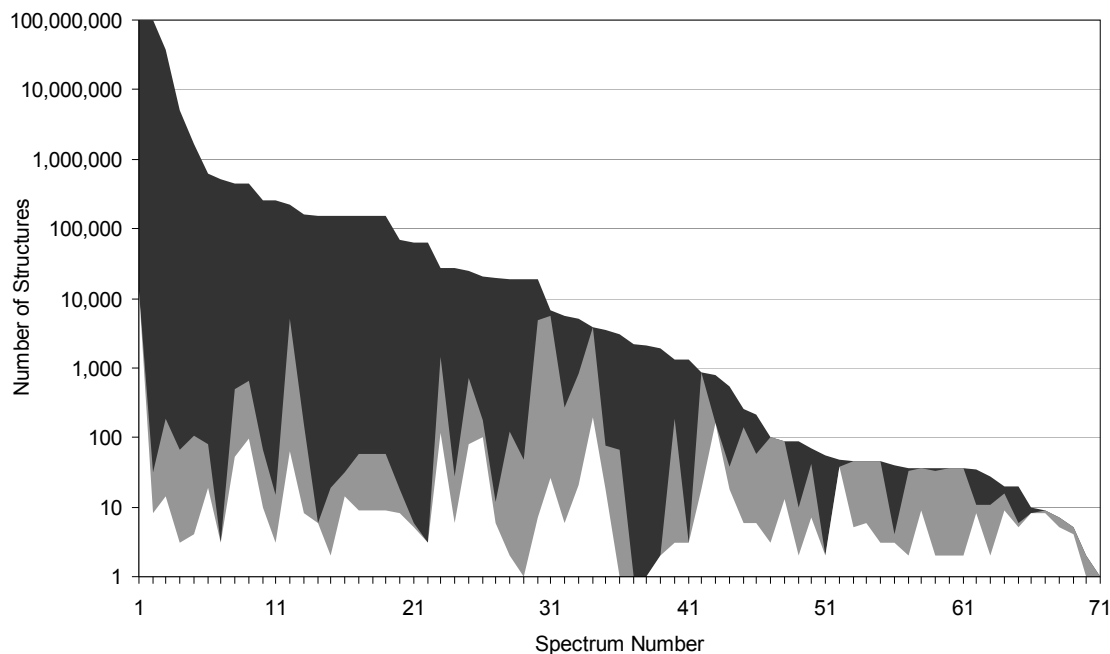
### 3.2.2 Overall Results

A total of 71 mass spectra with one molecular formula clearly matching the spectrum (from either the MolForm or ElCoCo steps in MOLGEN-MS) were selected from the Bitterfeld groundwater EDA spectra for inclusion in this study. The methods explained here have also been applied successfully to spectra with more than one potential molecular formula; however these have not been included here for clarity of presentation.

MOLGEN 3.5 was used for Run 1 to generate all possible structures for the given molecular formula, i.e. with no classifiers. MOLGEN-MS was used for Runs 2 and 3 to generate all possible structures matching the Varmuza classifiers only and the Varmuza and NIST classifiers together, respectively. The classifiers (both Varmuza and NIST classifiers) were used to the 95 % precision level. MOLGEN 3.5 was used for Run 1 as this is a much quicker calculation than using MOLGEN-MS; cases where MOLGEN 3.5 was also used for Run 2 and 3 are indicated in Appendix 1 (Table A1). The number of structures possible for each spectrum ranged between 1 (for S<sub>8</sub>) and >100,000,000 (for C<sub>18</sub>H<sub>35</sub>NO and C<sub>10</sub>H<sub>10</sub>O<sub>4</sub>). The number of structures generated using Varmuza classifiers

only ranged between 0 and >50,000 (MOLGEN-MS limits set to 50,000; MOLGEN 3.5 limits set to 100,000,000). The number of structures generated matching Varmuza and NIST classifiers ranged between 1 and 13,033 for  $C_{18}H_{35}NO$  (>100,000,000 matching structures without classifiers).

The difference between the number of structures generated in Run 1 (no classifiers) and Run 3 (Varmuza and NIST classifiers) for each spectrum is shown in Figure 9 by the black and grey areas, respectively. This figure clearly shows the order of magnitude reduction in structure numbers as a result of the mass spectral classifiers, especially for those molecular formulae with over 10,000 possible structures without classifiers (Run 1). Only 34 % (24 of 71) of the spectra had fewer than 100 possible structures generated in Run 1, compared with 70 % (50 spectra) in Run 3. Similarly, only 42 % (30 spectra) had fewer than 1000 possible structures generated in Run 1, compared with 91 % (65 spectra) in Run 3.



**Figure 9: Semi-logarithmic area plot of the number of structures generated in Run 1 (without classifiers, black) and Run 3 (Varmuza and NIST classifiers, grey). The white area represents the number of structures from Run 3 which fit the additional ‘filtering criteria’ of mass spectral match value and  $\log K_{ow}$ . The numbers in the x axis correspond with the spectrum number in Table A1 (see Appendix 1). Note that the first two MOLGEN data points represent >100,000,000 structures (calculation set to abort at 100,000,000). Data sorted by number of structures generated in Run 1.**

Further reduction in the structure numbers was achieved by applying additional ‘filtering criteria’; in this case the estimated  $\log K_{ow}$  and mass spectrum match value (MV, calculated with the MOLGEN-MS ‘ReNeGe’ module) for each structure. Structures with an estimated  $\log K_{ow}$  widely outside the fraction  $\log K_{ow}$  range (taking into account errors in estimation and calculation, as mentioned above) were excluded, as were structures with spectrum match values significantly lower than other structures. The effect of

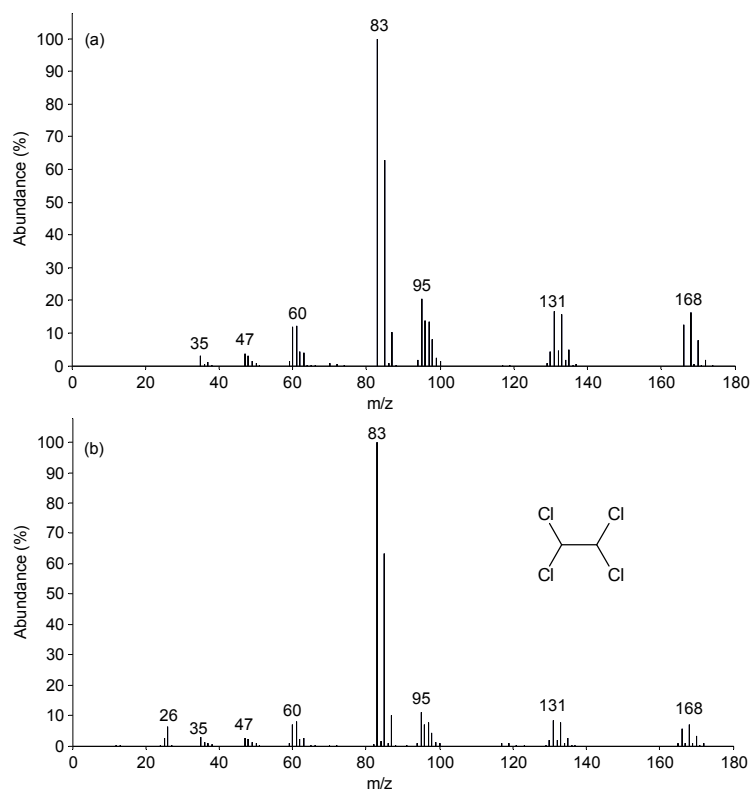
additional filtering criteria is signified by the white area in Figure 9. This demonstrates how the filtering criteria were used to reduce the number of likely matching structures further, often by an order of magnitude again.

These additional criteria reduced the number of matching structures below 10 for 52 of the 71 spectra (73 %), below 100 for all but 5 spectra (93 %) and below 1000 for all but 1 spectrum (99 %). This figure should serve as an example of the effect the filtering criteria has on the number of structures; the use of the filtering criteria is discussed further below due to the subjective nature of their use.

### 3.2.3 Providing 'lines of evidence'

In addition to the generation of all possible structures matching a mass spectrum, the method above can also be used to gain more information about possible structures and even to analyse whether a library match is in fact the 'best' or 'correct' match. Some examples are presented below.

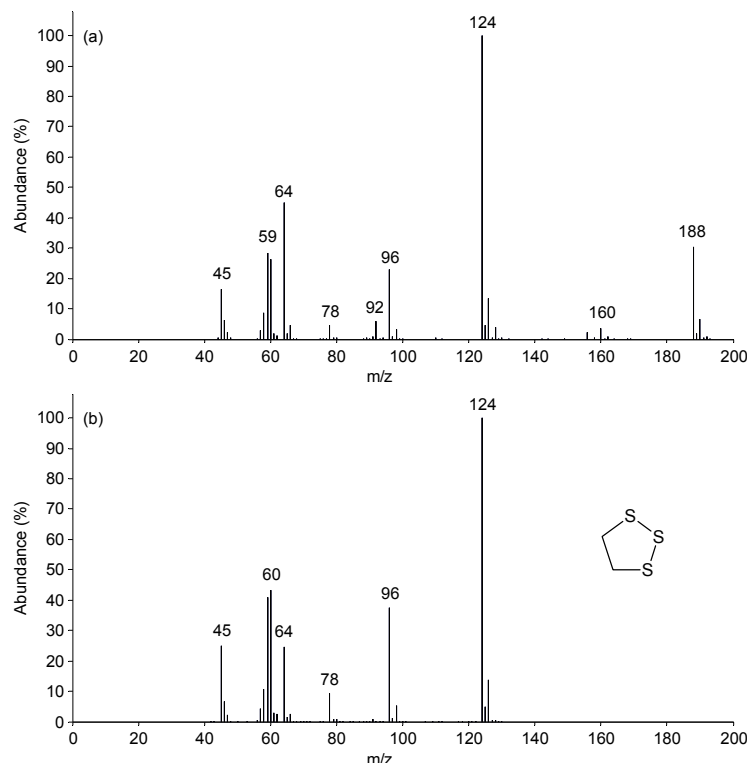
The spectrum shown in Figure 10 is a very close match to the NIST spectrum of 1,1,2,2-tetrachloroethane (TCA), with a library match value of 98.0 %. But how many compounds not in NIST could also match this spectrum?



**Figure 10:** (a) Mass spectrum of compound from fraction with  $\log K_{ow}=2.72-3.20$ , peak at 5.80 min. (b) Mass spectrum of the NIST match, 1,1,2,2-tetrachloroethane.

Running the spectrum through the method described above, the matching formula is clearly identified as  $C_2H_2Cl_4$  and only two structures are possible for this formula. Of these, one has a high MV (1,1,2,2-TCA, 73 %) and the other one a very low MV (1,1,1,2-TCA, 3 %), indicating that 1,1,2,2-TCA is most likely the correct match. The NIST spectrum for 1,1,1,2-TCA (two dominating peak groups at 117 & 131, not shown) is distinctly different to 1,1,2,2-TCA (one dominating peak group at 83), supporting this conclusion. As no other possible formulae were generated, the method described here provides significant supporting evidence to confirm that the library match is, in fact, the most likely match for the unknown compound – a useful feature where no further information is available to confirm the identity of the unknown.

The unknown shown in Figure 11 below was found to match the spectrum of the 1,2,3-trithiolane (MW = 124) with 95.9 % match value, despite the presence of a peak group at 188 in the unknown spectrum, which remained following deconvolution using AMDIS.



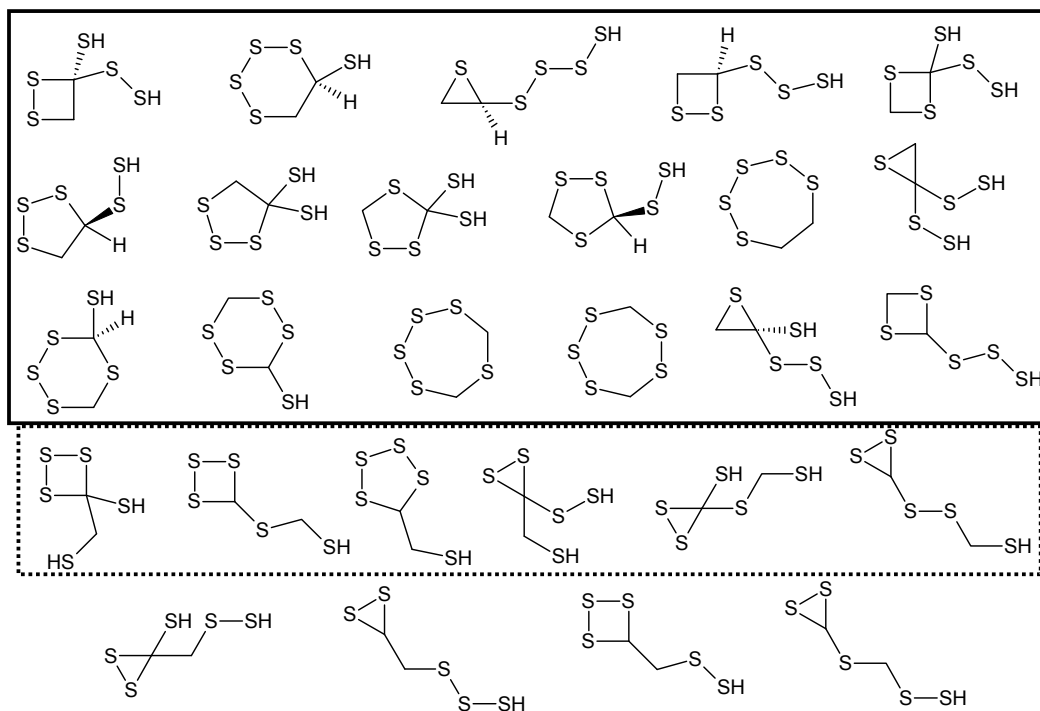
**Figure 11: (a) Mass spectrum of compound from fraction with  $\log K_{ow}$ =4.10-4.37, peak at 25.38 min. (b) Mass spectrum of the NIST match, 1,2,3-trithiolane.**

The classifiers indicated a C, H and S compound, with formula  $C_2H_4S_5$ , which is equivalent to the formula from the top NIST match ( $C_2H_4S_3$ ) with two additional sulphur atoms. The classifiers indicate that S-S (99 %), S-CH<sub>2</sub> in ring (99 %) and SH-CH<sub>2</sub> (95 %) were present, with CH<sub>3</sub> absent (99 %). Using MOLGEN-MS with the Varmuza classifier precision set at 95 %, 10 molecules were generated, of which 6 contained rings. However, none have S-CH<sub>2</sub> within the ring structure, as shown in Figure 12 (structures



within dashed box). Removing the SH-CH<sub>2</sub> classifier (95 %) results in the generation of 27 structures containing rings (shown in Figure 12), of which 17 have S-CH<sub>2</sub> in the ring (structures within the solid box of Figure 12). Note that the presence of the S-CH<sub>2</sub> classifier rules out, for example, 1,2,3-trithiolate-4,5-dithiol, the NIST match with a thiol group attached to each carbon. It is also possible, for example, to restrict the type of rings generated. In this case restricting the structure generation to structures with 5 or more membered rings eliminates the 16 three and four membered rings of the 27 structures shown in Figure 12.

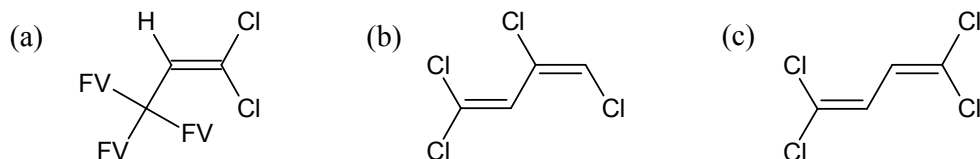
The structures shown in Figure 12 have match values between 3 % and 56 %. Of the structures with S-CH<sub>2</sub> in the ring (structures within the solid line), three have match values above 50 %; the first three structures (from the left) on the top row in Figure 12. Although perhaps the second structure, 1,2,3,4-tetrathiane-5-thiol, would intuitively be considered more likely (as it is a 6-membered ring rather than 3- or 4-membered ring), the predicted log *K<sub>ow</sub>* for this structure (1.63) is much lower than that of the other two structures (both 2.79), which are closer to the fraction log *K<sub>ow</sub>*. The choice of structure to pursue confirmation studies is therefore somewhat subjective in this case.



**Figure 12:** All cyclic structures generated for the unknown from fraction with log *K<sub>ow</sub>*=4.10-4.37, peak at 25.3769 minutes with 99% classifiers (S-CH<sub>2</sub>, S-S present, CH<sub>3</sub> absent). The solid box encloses cyclic structures with S-CH<sub>2</sub> connection within the ring. The dashed box encloses cyclic structures generated with addition of the SH-CH<sub>2</sub> classifier (95 % precision). Structures outside the boxes contain neither SH-CH<sub>2</sub> nor S-CH<sub>2</sub> in the ring and are therefore least likely to match the experimental mass spectrum. Structures are approximately arranged by match value within the groups, from higher (top left) to lower match value (bottom right).

### 3.2.4 Use of toxicity ‘classifiers’

If we take the unknown spectrum from Figure 5 with calculated formula  $C_4H_2Cl_4$ , scanning through the structural alerts identified by von der Ohe et al. [68] reveals one possible structural alert for this formula, ‘SA2: carbon-carbon bonds activated by two halogens’ (see Table 2 and Figure 13(a)). Adding this substructure to the MOLGEN-MS calculation identifies that two of the final four compounds (see Figure 7) may possibly exhibit excess toxicity, 1,1,4,4-tetrachloro-1,3-butadiene and 1,1,3,4-tetrachloro-1,3-butadiene, both of which are in the NIST library and shown in Figure 13(b) and (c).



**Figure 13:** (a) The SA2 substructure added to the good list of MOLGEN-MS. ‘FV’ denotes ‘free valence’. (b) & (c) The two structures generated using the SA2 structural alert for the unknown compound with retention time 10.875 min, fraction  $\log K_{ow}=4.37-4.85$ .

## 3.3 Discussion

As shown in the above results, the combination of mass spectral classifiers with a structure generation program adds a powerful dimension to the interpretation of a mass spectrum where clear identification of the unknown compound is not possible. The method presented can be used to identify ‘most probable’ matches for target confirmation or to provide ‘lines of evidence’ as to the identity of an unknown where the library match value was too low. This can also be used to identify how many possible compounds could match a spectrum, for example in the case where only one spectrum of a given formula is in the library or where several isomers with similar spectra are possible (e.g. substituted aromatic compounds). The method is of most use where little information is available from the database search.

The combination of good isotope peaks (enabling formula calculation using MolForm) and the NIST elemental classifiers was very effective in identifying valid formulae for each spectrum. Ensuring that all isotope peaks remain during spectral deconvolution (e.g. using AMDIS) is also important to prevent bias against inclusion of carbon atoms during calculation of the formulae, as some isotope peaks (especially the  $M+1$  peak) were often removed at the ‘default’ deconvolution settings. The IsoPat function in MOLGEN-MS provides a quick and useful ‘visual reality check’ between the predicted formula and spectral isotope peaks, while a quick calculation of the ‘RDB’ (ring and double bond) count is also useful to sort out formulae with unrealistic RDB counts. Cases where MolForm did not lead to clear results were generally compounds with many heteroatoms (e.g. where S, N and O were all present) as there are many possible combinations of C, H, S, N and O for a given molecular weight. The use of high resolution mass

spectroscopy (HR-MS) to record an accurate molecular weight would help to avoid (or at least reduce) this problem. The MolForm module of MOLGEN-MS can also perform molecular formula calculations based on a high resolution mass spectrum.

The time for structure generation is another important factor in the use of these programs. MOLGEN 3.5 is a very quick structure generator, for instance generating 19,054 possible structures for  $C_3H_9PO_3S$  in 0.11 seconds (Intel Core<sup>TM</sup>2 Duo 1.83 GHz, 1.00 GB RAM) without classifiers. The structure generation in MOLGEN-MS, based on the MOLGEN 4.0 kernel, is much slower and for  $C_3H_9PO_3S$  including classifier entries, generated 10,530 structures in 328.5 seconds. Although it is possible to use MOLGEN 3.5 to generate structures with classifier entries, each classifier has to be entered manually, which is tedious and increases the risk of entering an incorrect classifier. Furthermore, the preceding (e.g. Molin, MolForm, ElCoCo) and subsequent steps (e.g. ReNeGe) are no longer automated as in MOLGEN-MS, leaving MOLGEN-MS as the user-friendly alternative at this stage.

The time for structure generation is highly dependent on the molecular formula used and the classifier entries (both good list and bad list) and can be reduced most effectively by adding more 'good list' or 'yes' classifier entries. The results shown in Figure 9 were generated using the 95 % classifier probability; however for some spectra this is clearly insufficient to reduce the number of structures. For Spectrum 1 in Figure 9, more than 10,000 possible structures were generated using the classifier results. Additionally, the number of structures generated for Spectra 31, 34 and 42 using classifiers was not reduced significantly from those generated by MOLGEN without classifiers. Accepting classifiers with lower probability provides further restrictions and speeds up calculation, but also reduces the probability that the correct structure is among those generated. Although tests by Varmuza et al. [39] on the classifiers used in MOLGEN-MS showed that 'yes-answers are sometimes wrong, however, no-answers are almost always correct', the absence of classifier results, especially 'yes' classifiers, can lead to the generation of >10,000 structures (see Figure 9), which makes further analysis very difficult (for instance generating match values for >10,000 structures can take several hours to days). The interactive interface in MOLGEN-MS, which allows the addition or removal of 'good list' and 'bad list' structures, makes it very easy to adjust classifier entries and generate structures with different combinations of classifiers, so that program runs are not just restricted to the 95 % classifier level, but can be adjusted by the user (for example, see Figure 7 for structures generated using different classifier combinations). As stated by Benecke et al. [34] "the clever use of macroatoms is very important for the speed of the generation". As a consequence, computing several runs with different 'yes' classifiers, even if they are possibly wrong, is much more efficient in limiting structure

numbers (and computation time) than running MOLGEN-MS with ‘bad list’ structures only and generating more molecules.

The incorporation of the NIST classifiers at the Molin or MOLGEN steps (see Figure 4) is also an important improvement to the prediction powers of MOLGEN-MS, as demonstrated in Figure 7. The NIST classifiers are not only more numerous but also, in general, more specific than the classifiers in MOLGEN-MS. Again, the interactive interface in MOLGEN-MS makes it very easy to adjust classifier entries and generate structures with different combinations of classifiers.

While the combination of Varmuza and NIST classifiers is effective in reducing the number of structures to more manageable levels compared with structure generation alone, the development of additional post-generation filtering criteria is, at this stage, crucial in further reducing structure numbers and to aid in discriminating between ‘likely’ and ‘unlikely’ structures. As shown in Figure 9, the spectrum match value (MV) and the partitioning coefficient ( $\log K_{ow}$ ) of the fraction can be used as filtering criteria. Use of these criteria is, however, very subjective as the distribution of match values and partitioning coefficient data is very different for each set of structures generated. In the data presented in Figure 8, for instance, it is easy to exclude the one structure with a mass spectral match value below 10 %, as all other structures have a MV above 50 %. In many cases, however, there is no distinct ‘cut-off’ between molecules with high and low MVs, whereas for other runs all molecules have either very high or very low MVs, which eliminates the possibility of setting a cut-off at, for instance, 50 % for every run. Taking only the 10 or 20 % of structures with the highest match value, on the other hand, could eliminate too many structures, as would be the case for the data in Figure 8(d). Taking the top 10 % (in this case the top structure) eliminates two other structures with a very similar match value and  $\log K_{ow}$ .

Similar issues also arise with the  $\log K_{ow}$  filtering criterion. The spectrum shown in Figure 10 was recorded in the fraction with  $\log K_{ow} = 2.72\text{--}3.20$  and had two possible structures, 1,1,2,2-tetrachloroethane (estimated  $\log K_{ow} = 2.19$ , MV = 73 %) and 1,1,1,2-tetrachloroethane (estimated  $\log K_{ow} = 2.93$ , MV = 3 %). Excluding all structures outside the exact  $\log K_{ow}$  range of the fraction would leave only 1,1,1,2-tetrachloroethane, which has a low match value and, as already mentioned, a very different mass spectrum. Adding a nominal buffer of  $\log K_{ow} \pm 1$  (i.e. one order of magnitude) to allow for errors in estimation includes both structures and allows filtering by match value to distinguish which is the most ‘likely’ match. Again this ‘buffer’ can be subjective as certain compounds, especially polar molecules, are known to elute in fractions which do not correspond to their  $\log K_{ow}$ , due to other interactions with the column. Finizio et al. [72] report, for instance, a deviation of less than 0.5 log units for  $\log K_{ow}$  of pesticides determined from RP-HPLC methods and the experimental shake flask value, while

Eadsforth and Moser [73] report deviations of  $\pm 1$  log units for sufficiently soluble, non-polar substances. In some cases a buffer of  $\log K_{ow} \pm 2$  or more may be necessary, especially for compounds with many functional groups. Just as different bioassays are used in EDA to determine the different types of sample toxicity, a combination of several different filtering criteria are needed to suit all cases. The incorporation of toxicity classifiers, as shown in Figure 13, was presented as an additional method to identify possible excessively toxic compounds that match the spectrum.

As the spectra used in the generation of these results were real ‘unknowns’, it is impossible to further test the results. However, the use of unknown spectra was essential to retain objectivity and avoid prejudice in the selection of classifier entries, especially from NIST, in the cases where classifiers ‘clashed’. A method to automatically select NIST classifiers and check for clashes, as currently performed by the Molin module of MOLGEN-MS for the Varmuza classifiers, avoids this issue and was implemented in Section 5. An assessment of MOLGEN-MS using spectra from the NIST database (and therefore on ‘known’ compounds) has already been undertaken by Kerber et al. [7] and it was not our aim to replicate such a study here.

### 3.4 Outcomes

The outcomes of the method development based on unknown spectra show that the combination of NIST and Varmuza classifiers is instrumental in reducing the number of generated structures per spectrum down to a ‘manageable’ data set. The usefulness of the mass spectral match value to select candidates was not clear in many cases. As a result of this, the use of alternative programs to predict mass spectra and potentially improve the selectivity of this value was assessed, presented in Section 4.

The results presented above (e.g. Figure 9) also show that additional criteria are still needed to further reduce the number of possible matching structures and to select the ‘correct’ structures from the incorrect in many cases. Furthermore, as the inclusion of NIST classifiers was shown to be beneficial, incorporating this information automatically will improve the usability of MOLGEN-MS and help ensure unbiased selection of classifiers for structure generation. These two issues are addressed in Section 5, where additional filtering criteria are introduced and tested, along with the automatic classifier loading, on known spectra.

## 4 Mass Spectral Fragment Prediction

The mass spectral prediction component of MOLGEN-MS, used to generate the match value (MV) and sort the candidates generated in Section 3, appeared to vary dramatically with the spectrum in terms of selectivity. This made the results difficult to judge for unknown spectra. Kerber et al. also observed that the general fragmentation rules alone were insufficient for automatic structure elucidation but suggested that the use of more sophisticated programs for virtual fragmentation may improve the ranking results [7].

As a result, an assessment of three fragmentation programs was undertaken, Mass Frontier [44], ACD MS Fragmenter [30] and MOLGEN-MSF [48]. MOLGEN-MSF was used in place of MOLGEN-MS, as this is an advanced command-prompt version of the mass spectral matching algorithms, de-coupled from the structure generation features of MOLGEN-MS. This assessment was published in 2009 [8]:

Schymanski, E., Meringer, M. and Brack, W. (2009) Matching Structures to Mass Spectra using Fragmentation Patterns – Are the Results as Good as they Look?, *Analytical Chemistry*, 81 (9), 3608-3617.

The programs and concepts used are covered in Section 2 and are not repeated here. This section covers the methods used to assess the fragmentation patterns, the results (both general and with specific examples), discussion and finally implications in terms of structural elucidation.

### 4.1 Methods to Compare Mass Spectral Fragment Prediction

#### 4.1.1 Program Settings and Abbreviations

Generation of fragments for all candidate structures using Mass Frontier [44] was implemented using the Batch Processing function, with the MOLGEN 3.5 SDF file containing all candidate structures as input. The default settings recommended on the Mass Frontier website were used for the generation of fragments [49], included in Appendix 1, Table A2. Four settings combinations were used: 3 step fragment generation with general fragmentation rules only, 5 step generation with general fragmentation rules and 3 and 5 step generation with both general and library fragmentation rules. The output was one structure (SDF) file per input structure, which contained the fragments generated.

The generation of fragments and Assignment Quality Indices using ACD MS Manager [30] was implemented using Extended Macro Processing, which requires upload of one copy of the spectrum per structure for processing, combined with an MSP import macro (provided by ACD upon purchase) and the SDF structure file containing all candidates. The AutoAssignment settings recommended by ACD [50] were used as standard settings,

given in Appendix 1, Table A3. The only variation in the settings was 3 or 5 step fragment generation, to assess the impact of more fragmentation steps on the results. The output of the Extended Macro was a “Table of Fragments” file for each structure and a summary file containing the Assignment Quality Index of each structure to the spectrum.

MOLGEN-MSF [48] was used to generate fragments and the corresponding match values for an input SDF file and spectrum, using the fragmentation rules outlined in Section 2.7.1 and [7]. In most cases the output was limited to an output file containing the match values of each structure to the spectrum.

To simplify the presentation of the results, each program and the corresponding settings were assigned a short name. This name and the explanation are given in Table 4.

**Table 4: Abbreviations used to describe the programs and settings used in this chapter. Unless stated otherwise, results presented from each program are presented in terms of the match value given in Equation 1, while those with ‘AQI’ suffix are expressed in the ACD Assignment Quality Index.**

| Abbreviation | Description                                                                                                       |
|--------------|-------------------------------------------------------------------------------------------------------------------|
| MSF          | MOLGEN-MSF fragmentation                                                                                          |
| MF_3st       | Mass Frontier, 3 step fragmentation, general fragmentation rules only                                             |
| MF_3st_wLib  | Mass Frontier, 3 step fragmentation, general and library fragmentation rules                                      |
| MF_5st       | Mass Frontier, 5 step fragmentation, general fragmentation rules only                                             |
| MF_5st_wLib  | Mass Frontier, 5 step fragmentation, general and library fragmentation rules                                      |
| ACD_3st      | ACD MS Manager AutoAssignment, 3 step fragmentation                                                               |
| ACD_5st      | ACD MS Manager AutoAssignment, 5 step fragmentation                                                               |
| ACD_3st_AQI  | ACD MS Manager AutoAssignment, 3 step fragmentation, results expressed in terms of the ‘Assignment Quality Index’ |
| ACD_5st_AQI  | ACD MS Manager AutoAssignment, 5 step fragmentation, results expressed in terms of the ‘Assignment Quality Index’ |

#### 4.1.2 File and Input Preparation

As the match value calculation (Equation 1) was already implemented in the code of MOLGEN-MSF to assess the fragments predicted by MOLGEN-MSF, the first step in program comparison was to get MOLGEN-MSF to accept fragments generated by Mass Frontier and ACD MS Manager as input instead of those fragments generated by MOLGEN-MSF. This was achieved for Mass Frontier by reading the fragment structure files sequentially, along with a correction for an anomaly in the fragment files generated. For ACD MS Manager, the ‘Table of Fragment’ output of the Extended Macro procedure included a fragment formula column for each structure-spectrum combination. This was extracted and exported into sequentially numbered text files, together with a nominal fragment multiplicity set to 1, using Matlab [70]; the associated Matlab script is listed in Appendix 2. In both cases the input into MOLGEN-MSF included the original spectrum, typically in the MSP [69] format.

Although all programs used in this investigation could read both MSP (spectrum) and SDF [74, 75] (structure) files, some modifications to the file formats were necessary to ensure these files could be read correctly or at all. These modifications included:

- MOLGEN 3.5 SDF files: ‘M END’ lines (with two spaces between the M and END) have to be added to each structure above the ‘\$\$\$\$’ separator to ensure all structures are read in Mass Frontier and ACD imports. See Appendix 2.
- ACD provided an MSP import macro upon purchase of the software to allow import of MSP files. The file ending has to be in small letters and the MSP file needs to be in the format consistent with the NIST database. See Appendix 2.
- The ‘Name’ line of the MSP file imported into MOLGEN-MSF should be written as shown and not in capital letters.

The file manipulations mentioned above were conducted as part of this work, whereas programming changes to MOLGEN-MSF itself were the work of M. Meringer.

#### 4.1.3 Comparison

The electron impact mass spectra considered in this study were retrieved from the NIST 2005 Mass Spectral Database [17] by spectrum number and saved in the Mass Spectral Transfer (MSP) format [69]. All programs were assessed, where possible, using the 100 randomly-selected spectra from the previous study [7], to ensure comparability. Spectra no longer available in the 2005 NIST database were recovered from the archive of the Kerber et al. study for consistency. Minor adjustments to the MSP format were made, where necessary, for import into the different programs (see Section 4.1.2 above). The much longer calculation times for ACD MS Manager and Mass Frontier with library reactions required a reduction in the data set to spectra with less than 500 constitutional isomers (41 spectra) and 200 constitutional isomers (27 spectra), respectively.

The candidate structures (constitutional isomers) for each spectrum and for the three specific examples used in this Section were generated using MOLGEN 3.5 [34], with no restrictions unless indicated otherwise. The molecules were saved in the MDL SDF format [74, 75], hereafter referred to as ‘SDF format’. Specific details regarding the generation of fragments by each program are given in Section 4.1.2 above.

The match value (see Equation 1) was used in this study to compare the results generated by all programs, as it requires only the input of the fragments generated by each program and uses the peaks from the experimental spectrum to assign magnitudes. Once match values are calculated for each structure, the relative ranking position (Equation 7) can be calculated, to assess the accuracy or selectivity of the different programs. The basic concept, from generation of structures through to calculation of the relative ranking position, is presented in Figure 14.



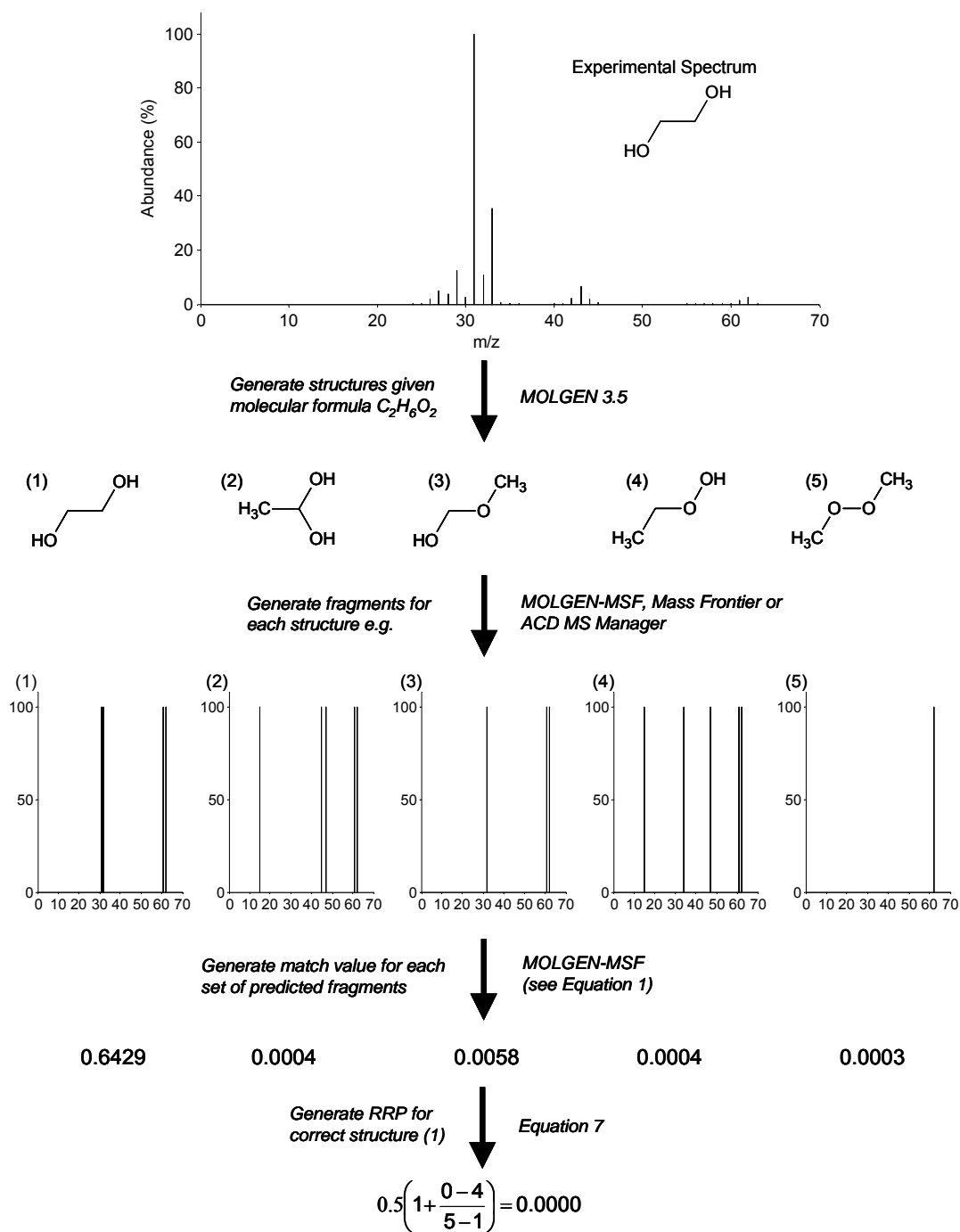


Figure 14: Schematic for matching candidate structures to an experimental spectrum using fragmentation patterns. All possible structures are generated for the formula from the experimental spectrum; fragments are predicted for each structure; the match value is calculated to match the fragments to the experimental spectrum; finally the match values are used to determine the number of ‘better’ and ‘worse’ candidates for calculation of the relative ranking position (RRP) of the correct structure.

The code of MOLGEN-MSF was extended to read the Mass Frontier and ACD outputs to enable consistent match value calculation for all programs. The input for match value calculations also included the experimental spectrum, typically in the MSP format (MOLGEN-MSF also accepts alternative input formats). The ‘Assignment Quality Index’ (ACD MS Manager) was calculated for all ACD runs, to compare with the match value. As mentioned in Section 2.7.1, the definition of the Assignment Quality Index was not sufficient to attempt to reproduce this calculation for fragments generated using Mass Frontier and MOLGEN-MSF.

Kerber et al. [7] also used simple statistics to assess the results of structure fragmentation and ranking. They defined confidence intervals to indicate how many structures need to be considered for a given spectrum to ensure inclusion of the correct structure with a certain probability, using an independent random selection of 1000 structure-spectrum pairs. For these spectra, the match values were calculated only for the correct structure, not all constitutional isomers. The same 1000 spectra from the previous study were used to calculate confidence intervals here, for all program and settings combinations, except Mass Frontier with library fragmentation (due to the long calculation time required for the large structures included in the 1000 spectra).

The  $p$ -quantile  $q_p$  was defined as a number such that  $p(1000)$  of the 1000 MV pairs is less than or equal to  $q_p$ , where  $0 < p < 1$ . For example, if  $q_{0.1} = 0.14949$ , 100 of the 1000 spectra have MVs less than or equal to 0.14949 and 900 have MVs greater than 0.14949. Taking this case, to be sure that the correct structure is included in the selected structures with reliability 0.9 (equivalent to 90 %), then all structures with  $MV > 0.14949$  should be included in the selected structures. The definition and explanation of the confidence interval concept can be found in Kerber et al. [7].

## 4.2 Results of Fragment Prediction and Comparison

The results for MOLGEN-MSF and Mass Frontier 3 and 5 step fragmentations using only the general fragmentation rules for the 100 randomly-selected spectra are given in Appendix 1, Table A4. The match values for the correct structure calculated for all program and settings combinations for the reduced data set of 27 spectra (those spectra with less than 200 structures) are given in Table 5. The relative ranking positions for the same data set are presented in Table 6.

As mentioned above, an alternative to assessing the programs is to do this in relation to the number of candidate structures that have to be considered to ensure that the correct structure is likely to be included with a fixed probability. The parameter  $q_p$ , defined above, is such that consideration of all structures with match value above  $q_p$  means that the correct structure is present with probability  $(1-p)$ , so to be 90 % sure that the correct

structure is present,  $q_{0.1}$  should be used. The quantiles calculated for the different programs and settings are presented below in Table 7. Due to the size of the molecules included in some of the 1000 spectra, this calculation could not be performed using Mass Frontier with library fragmentation mechanisms.

**Table 5: Match values of the correct structure calculated for 27 spectra for all programs and settings. The program abbreviations are as given in Table 4, the spectrum number corresponds with those in Table A2 and TC refers to the number of structures (total candidates).**

| No.              | Formula                                         | TC  | Match Value (MV) of Correct Structure |              |              |              |              |              |              |
|------------------|-------------------------------------------------|-----|---------------------------------------|--------------|--------------|--------------|--------------|--------------|--------------|
|                  |                                                 |     | MSF                                   | MF_3st       | MF_5st       | MF_3st_wLib  | MF_5st_wLib  | ACD_3st      | ACD_5st      |
| 4                | C <sub>7</sub> H <sub>14</sub>                  | 56  | 0.2631                                | 0.2668       | 0.7077       | 0.2899       | 0.7762       | 0.9375       | 0.9644       |
| 10               | CN <sub>3</sub> F <sub>5</sub>                  | 11  | 0.0000                                | 0.0551       | 0.2280       | 0.0551       | 0.2280       | 0.0551       | 0.0551       |
| 13               | CH <sub>5</sub> SiBr                            | 2   | 0.0366                                | 0.0293       | 0.0293       | 0.0293       | 0.0293       | 0.5205       | 0.5358       |
| 15               | C <sub>5</sub> H <sub>11</sub> Br               | 8   | 0.0595                                | 0.1389       | 0.1539       | 0.5014       | 0.6043       | 0.9450       | 0.9729       |
| 19               | C <sub>2</sub> H <sub>3</sub> NO                | 26  | 0.1454                                | 0.0010       | 0.0010       | 0.0010       | 0.0942       | 0.4821       | 0.4821       |
| 34               | C <sub>11</sub> H <sub>24</sub>                 | 159 | 0.5614                                | 0.5511       | 0.5511       | 0.8990       | 0.8995       | 0.9530       | 0.9763       |
| 35               | C <sub>8</sub> H <sub>16</sub>                  | 139 | 0.1416                                | 0.1367       | 0.1382       | 0.1367       | 0.6799       | 0.8560       | 0.9527       |
| 37               | C <sub>9</sub> H <sub>20</sub>                  | 35  | 0.5628                                | 0.5628       | 0.5628       | 0.8330       | 0.8332       | 0.9252       | 0.9435       |
| 40               | C <sub>5</sub> H <sub>13</sub> N                | 17  | 0.8367                                | 0.8350       | 0.8407       | 0.8644       | 0.8648       | 0.9812       | 0.9836       |
| 42               | C <sub>8</sub> H <sub>14</sub> O                | 32  | 0.0528                                | 0.0125       | 0.0601       | 0.1775       | 0.7176       | 0.9623       | 0.9826       |
| 45               | C <sub>5</sub> H <sub>12</sub> O <sub>2</sub>   | 69  | 0.2624                                | 0.0256       | 0.0326       | 0.1453       | 0.1617       | 0.8386       | 0.8413       |
| 50               | C <sub>2</sub> H <sub>6</sub> O <sub>2</sub>    | 5   | 0.6429                                | 0.6307       | 0.6307       | 0.8634       | 0.8658       | 0.6755       | 0.6755       |
| 52               | C <sub>5</sub> H <sub>6</sub>                   | 40  | 0.4656                                | 0.3690       | 0.3690       | 0.5321       | 0.5321       | 0.6303       | 0.6303       |
| 54               | C <sub>8</sub> H <sub>17</sub> Cl               | 89  | 0.0592                                | 0.0363       | 0.2249       | 0.3877       | 0.4864       | 0.9151       | 0.9639       |
| 59               | C <sub>4</sub> H <sub>12</sub> N <sub>2</sub>   | 38  | 0.7545                                | 0.7566       | 0.7566       | 0.7733       | 0.8614       | 0.9944       | 0.9945       |
| 60               | C <sub>3</sub> H <sub>3</sub> Cl <sub>3</sub>   | 8   | 0.0019                                | 0.6502       | 0.6502       | 0.6521       | 0.6521       | 0.7820       | 0.7980       |
| 61               | C <sub>5</sub> H <sub>13</sub> N                | 17  | 0.5151                                | 0.7369       | 0.7369       | 0.8148       | 0.8285       | 0.9554       | 0.9593       |
| 66               | C <sub>2</sub> H <sub>7</sub> P                 | 2   | 0.1597                                | 0.1597       | 0.1597       | 0.1597       | 0.1597       | 0.5337       | 0.5337       |
| 68               | C <sub>5</sub> H <sub>13</sub> NO               | 149 | 0.6480                                | 0.6499       | 0.8028       | 0.9135       | 0.9149       | 0.9148       | 0.9408       |
| 72               | C <sub>4</sub> H <sub>11</sub> NO               | 56  | 0.7706                                | 0.7712       | 0.7724       | 0.8502       | 0.9241       | 0.9929       | 0.9929       |
| 73               | C <sub>6</sub> H <sub>10</sub>                  | 77  | 0.0896                                | 0.6213       | 0.6213       | 0.6383       | 0.6383       | 0.8586       | 0.8680       |
| 74               | C <sub>2</sub> NF <sub>3</sub>                  | 5   | 0.4977                                | 0.6830       | 0.6830       | 0.6830       | 0.6830       | 0.7857       | 0.7857       |
| 80               | C <sub>3</sub> H <sub>7</sub> NO                | 84  | 0.6177                                | 0.1824       | 0.1824       | 0.6206       | 0.6313       | 0.9766       | 0.9803       |
| 81               | C <sub>3</sub> H <sub>7</sub> O <sub>2</sub> Br | 38  | 0.0992                                | 0.1001       | 0.4454       | 0.2888       | 0.8528       | 0.9804       | 0.9804       |
| 84               | C <sub>8</sub> H <sub>16</sub>                  | 139 | 0.6245                                | 0.6023       | 0.6091       | 0.7566       | 0.7679       | 0.8287       | 0.9804       |
| 96               | C <sub>3</sub> H <sub>4</sub> O                 | 13  | 0.6550                                | 0.0270       | 0.0270       | 0.0270       | 0.0270       | 0.7350       | 0.7350       |
| 97               | C <sub>4</sub> H <sub>5</sub> OCl               | 175 | 0.0255                                | 0.6993       | 0.7014       | 0.9056       | 0.9094       | 0.9371       | 0.9800       |
| <b>Averages:</b> |                                                 |     | <b>0.354</b>                          | <b>0.381</b> | <b>0.433</b> | <b>0.511</b> | <b>0.616</b> | <b>0.813</b> | <b>0.833</b> |

**Table 6: The relative ranking positions (RRPs) calculated for the reduced set of 27 spectra using all programs and settings. Abbreviations are consistent with Table 4 and Table 5.**

| No. | Formula                           | TC  | Relative Ranking Position (RRP) of Correct Structure |        |        |             |             |         |         |
|-----|-----------------------------------|-----|------------------------------------------------------|--------|--------|-------------|-------------|---------|---------|
|     |                                   |     | MSF                                                  | MF_3st | MF_5st | MF_3st_wLib | MF_5st_wLib | ACD_3st | ACD_5st |
| 4   | C <sub>7</sub> H <sub>14</sub>    | 56  | 0.6273                                               | 0.6091 | 0.2273 | 0.7273      | 0.6364      | 0.1455  | 0.3727  |
| 10  | CN <sub>3</sub> F <sub>5</sub>    | 11  | 0.5000                                               | 0.3000 | 0.1000 | 0.3000      | 0.1000      | 0.8000  | 0.8000  |
| 13  | CH <sub>5</sub> SiBr              | 2   | 0.5000                                               | 1.0000 | 1.0000 | 1.0000      | 1.0000      | 1.0000  | 1.0000  |
| 15  | C <sub>5</sub> H <sub>11</sub> Br | 8   | 0.5714                                               | 0.4286 | 0.4286 | 0.2857      | 0.0000      | 0.0000  | 0.5000  |
| 19  | C <sub>2</sub> H <sub>3</sub> NO  | 26  | 0.0800                                               | 0.1600 | 0.2800 | 0.2000      | 0.0800      | 0.5600  | 0.6400  |
| 34  | C <sub>11</sub> H <sub>24</sub>   | 159 | 0.6582                                               | 0.6646 | 0.6646 | 0.3956      | 0.4873      | 0.7215  | 0.5443  |

| No.              | Formula                                         | TC  | Relative Ranking Position (RRP) of Correct Structure |              |              |              |              |              |              |
|------------------|-------------------------------------------------|-----|------------------------------------------------------|--------------|--------------|--------------|--------------|--------------|--------------|
|                  |                                                 |     | MSF                                                  | MF_3st       | MF_5st       | MF_3st_wLib  | MF_5st_wLib  | ACD_3st      | ACD_5st      |
| 35               | C <sub>8</sub> H <sub>16</sub>                  | 139 | 0.6957                                               | 0.8116       | 0.8696       | 0.9638       | 0.9203       | 0.3877       | 0.4167       |
| 37               | C <sub>9</sub> H <sub>20</sub>                  | 35  | 0.4853                                               | 0.2647       | 0.2647       | 0.3971       | 0.2647       | 0.5294       | 0.4118       |
| 40               | C <sub>5</sub> H <sub>13</sub> N                | 17  | 0.0625                                               | 0.0625       | 0.0000       | 0.0625       | 0.5000       | 0.2188       | 0.4375       |
| 42               | C <sub>6</sub> H <sub>14</sub> O                | 32  | 0.5645                                               | 0.7581       | 0.5806       | 1.0000       | 0.1290       | 0.1613       | 0.1452       |
| 45               | C <sub>5</sub> H <sub>12</sub> O <sub>2</sub>   | 69  | 0.0147                                               | 0.1544       | 0.4853       | 0.0882       | 0.6912       | 0.5221       | 0.5294       |
| 50               | C <sub>2</sub> H <sub>6</sub> O <sub>2</sub>    | 5   | 0.0000                                               | 0.1250       | 0.1250       | 0.0000       | 0.0000       | 0.2500       | 0.2500       |
| 52               | C <sub>5</sub> H <sub>6</sub>                   | 40  | 0.6026                                               | 0.2949       | 0.3718       | 0.0513       | 0.2308       | 0.5128       | 0.7564       |
| 54               | C <sub>8</sub> H <sub>17</sub> Cl               | 89  | 0.1591                                               | 0.1705       | 0.2386       | 0.3011       | 0.1932       | 0.8977       | 0.9432       |
| 59               | C <sub>4</sub> H <sub>12</sub> N <sub>2</sub>   | 38  | 0.1216                                               | 0.0676       | 0.1622       | 0.5405       | 0.2838       | 0.0811       | 0.5676       |
| 60               | C <sub>3</sub> H <sub>3</sub> Cl <sub>3</sub>   | 8   | 0.5714                                               | 0.7143       | 0.7143       | 0.5714       | 0.7143       | 1.0000       | 1.0000       |
| 61               | C <sub>5</sub> H <sub>13</sub> N                | 17  | 0.8125                                               | 0.6250       | 0.6250       | 0.7500       | 0.7500       | 0.6875       | 0.9375       |
| 66               | C <sub>2</sub> H <sub>7</sub> P                 | 2   | 1.0000                                               | 1.0000       | 1.0000       | 1.0000       | 1.0000       | 1.0000       | 1.0000       |
| 68               | C <sub>5</sub> H <sub>13</sub> NO               | 149 | 0.0372                                               | 0.1858       | 0.2027       | 0.0912       | 0.3514       | 0.7466       | 0.2669       |
| 72               | C <sub>4</sub> H <sub>11</sub> NO               | 56  | 0.2364                                               | 0.2182       | 0.2182       | 0.5455       | 0.5455       | 0.5091       | 0.5364       |
| 73               | C <sub>6</sub> H <sub>10</sub>                  | 77  | 0.9342                                               | 0.5000       | 0.6316       | 0.4276       | 0.8092       | 0.7763       | 0.7829       |
| 74               | C <sub>2</sub> NF <sub>3</sub>                  | 5   | 0.0000                                               | 0.0000       | 0.0000       | 0.0000       | 0.0000       | 0.0000       | 0.0000       |
| 80               | C <sub>3</sub> H <sub>7</sub> NO                | 84  | 0.0482                                               | 0.6566       | 0.7048       | 0.1325       | 0.0663       | 0.4217       | 0.3133       |
| 81               | C <sub>3</sub> H <sub>7</sub> O <sub>2</sub> Br | 38  | 0.0000                                               | 0.0541       | 0.0000       | 0.2703       | 0.0000       | 0.0405       | 0.3514       |
| 84               | C <sub>8</sub> H <sub>16</sub>                  | 139 | 0.0362                                               | 0.0507       | 0.1957       | 0.0435       | 0.0072       | 0.6993       | 0.3768       |
| 96               | C <sub>3</sub> H <sub>4</sub> O                 | 13  | 0.0000                                               | 0.1667       | 0.2500       | 0.3333       | 0.4167       | 0.5000       | 0.5000       |
| 97               | C <sub>4</sub> H <sub>5</sub> OCl               | 175 | 0.1810                                               | 0.0920       | 0.2586       | 0.0172       | 0.1236       | 0.8621       | 0.0632       |
| <b>Averages:</b> |                                                 |     | <b>0.352</b>                                         | <b>0.375</b> | <b>0.393</b> | <b>0.389</b> | <b>0.382</b> | <b>0.520</b> | <b>0.535</b> |

**Table 7: Quantiles  $q_p$  calculated for 1000 randomly-selected spectra for Mass Frontier with general fragmentation rules, MOLGEN-MSF and ACD MS Manager, based on match values for all programs and the Assignment Quality Index (AQI) for ACD results. Program and setting abbreviations are in Table 4.**

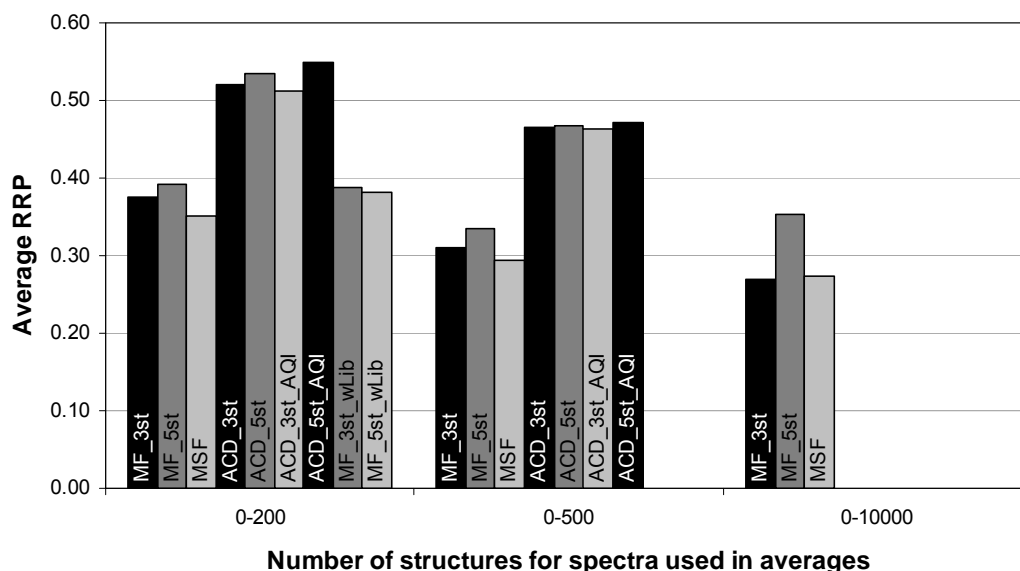
|      | Match Values |        |        |         |         | Assignment Quality Index (%) |         |
|------|--------------|--------|--------|---------|---------|------------------------------|---------|
|      | MF_3st       | MF_5st | MSF    | ACD_3st | ACD_5st | ACD_3st                      | ACD_5st |
| $p$  | $q_p$        | $q_p$  | $q_p$  | $q_p$   | $q_p$   | $q_p$                        | $q_p$   |
| 0.01 | 0.0046       | 0.0294 | 0.0043 | 0.0222  | 0.0281  | 18.1                         | 18.1    |
| 0.05 | 0.0377       | 0.1715 | 0.0343 | 0.2306  | 0.2632  | 46.0                         | 48.9    |
| 0.1  | 0.0852       | 0.2516 | 0.0774 | 0.3574  | 0.3745  | 56.7                         | 59.9    |
| 0.2  | 0.1647       | 0.3683 | 0.1768 | 0.5307  | 0.5536  | 68.8                         | 72.1    |
| 0.3  | 0.2530       | 0.4767 | 0.2678 | 0.6645  | 0.6983  | 77.7                         | 80.0    |
| 0.4  | 0.3240       | 0.5521 | 0.3434 | 0.7636  | 0.7921  | 83.8                         | 86.4    |
| 0.5  | 0.3956       | 0.6333 | 0.4403 | 0.8290  | 0.8533  | 88.2                         | 90.7    |
| 0.6  | 0.4922       | 0.6918 | 0.5331 | 0.8835  | 0.9004  | 91.7                         | 93.7    |
| 0.7  | 0.5810       | 0.7535 | 0.6162 | 0.9175  | 0.9339  | 94.7                         | 95.8    |
| 0.8  | 0.6795       | 0.8071 | 0.7099 | 0.9487  | 0.9595  | 96.6                         | 97.7    |
| 0.9  | 0.7822       | 0.8636 | 0.8074 | 0.9730  | 0.9791  | 98.6                         | 99.0    |
| 0.95 | 0.8387       | 0.8987 | 0.8663 | 0.9830  | 0.9881  | 99.5                         | 99.7    |
| 0.99 | 0.9027       | 0.9428 | 0.9327 | 0.9930  | 0.9962  | 100.0                        | 100.0   |

Several other parameters that were presented by Kerber et al. [7] were calculated for each spectrum to assist in the interpretation of the program fragmentation ability. The average values of these parameters are presented in Table 8.

The average relative ranking positions calculated for the different programs and settings, for spectra with 0-200, 0-500 and 0-10,000 structures are plotted in Figure 15.

**Table 8: Average parameters for each program and settings for data sets of 0-10,000 structures (100 spectra), 0-500 structures (41 spectra) and 0-200 structures (27 spectra). Parameters averaged over the number of spectra in the data set (n\_spec). Average\_MV denotes the average of all match values for each structure per spectrum, MV\_corr\_struct the match value (or AQI, where indicated) of the correct structure, BC, EC and WC the number of candidates with better, equal and worse match values, per spectrum, respectively, RRP denotes the relative ranking position for each spectrum and C90 denotes the number of candidates selected such that the correct structure is present with probability 90 %. BC, EC, WC and C90 were divided by the total number of candidates (TC) prior to averaging to create the ratios presented here.**

|             | Average_MV | n_spec | MV_corr_str | BC/TC | EC/TC | WC/TC | RRP    | C90/TC |
|-------------|------------|--------|-------------|-------|-------|-------|--------|--------|
| MF_3st      | 0.2725     | 100    | 0.4622      | 0.25  | 0.04  | 0.71  | 0.2685 | 0.64   |
| MF_5st      | 0.3963     | 100    | 0.5581      | 0.33  | 0.04  | 0.63  | 0.3527 | 0.58   |
| MSF         | 0.2463     | 100    | 0.4317      | 0.25  | 0.05  | 0.70  | 0.2734 | 0.60   |
| MF_3st      | 0.2608     | 41     | 0.4267      | 0.26  | 0.09  | 0.65  | 0.3097 | 0.61   |
| MF_5st      | 0.3109     | 41     | 0.4905      | 0.29  | 0.08  | 0.63  | 0.3345 | 0.45   |
| MSF         | 0.2244     | 41     | 0.3719      | 0.24  | 0.11  | 0.65  | 0.2943 | 0.55   |
| ACD_3st     | 0.7949     | 41     | 0.8599      | 0.41  | 0.09  | 0.50  | 0.4645 | 0.91   |
| ACD_5st     | 0.8319     | 41     | 0.8805      | 0.36  | 0.20  | 0.44  | 0.4679 | 0.91   |
| ACD_3st_AQI | 85.0       | 41     | 89.0        | 0.39  | 0.12  | 0.48  | 0.4627 | 0.90   |
| ACD_5st_AQI | 88.2       | 41     | 91.0        | 0.33  | 0.26  | 0.41  | 0.4720 | 0.90   |
| MF_3st      | 0.2769     | 27     | 0.3811      | 0.30  | 0.13  | 0.57  | 0.3754 | 0.57   |
| MF_3st_wLib | 0.4428     | 27     | 0.5111      | 0.34  | 0.09  | 0.58  | 0.3887 | -      |
| MF_5st      | 0.3066     | 27     | 0.4325      | 0.32  | 0.12  | 0.56  | 0.3926 | 0.45   |
| MF_5st_wLib | 0.5305     | 27     | 0.6157      | 0.32  | 0.10  | 0.57  | 0.3815 | -      |
| MSF         | 0.2328     | 27     | 0.3537      | 0.28  | 0.16  | 0.57  | 0.3519 | 0.49   |
| ACD_3st     | 0.7672     | 27     | 0.8131      | 0.45  | 0.12  | 0.43  | 0.5197 | 0.89   |
| ACD_5st     | 0.8084     | 27     | 0.8331      | 0.42  | 0.20  | 0.38  | 0.5349 | 0.90   |
| ACD_3st_AQI | 82.3       | 27     | 85.3        | 0.42  | 0.16  | 0.42  | 0.5121 | 0.86   |
| ACD_5st_AQI | 85.9       | 27     | 87.3        | 0.40  | 0.27  | 0.34  | 0.5480 | 0.87   |



**Figure 15: Average relative ranking positions (RRPs) for the different programs and settings, taken over spectra with 0-200, 0-500 and 0-10,000 structures.**

This shows that the average relative ranking position is larger (i.e. worse) for spectra with few possible structures, compared with all spectra. This trend, which was apparent in all programs, indicates that the ranking success of the match value (and AQI) is generally worse for spectra with few candidate structures.

### 4.3 Results of Specific Examples

The selection of specific examples, in addition to the randomly-selected spectra presented above, provides additional insight into the performance of each program by allowing the consideration of phenomena specific to certain structures. Three examples are used here to evaluate the use of fragmentation to match structures to their mass spectra. The formulae were selected based on the presence of several spectra for the given formula in the NIST database, where some of the spectra were clearly different from others (specifically containing different peak groups, not just different magnitudes of peaks). The examples were also chosen for the low number of possible structures (<100), to aid in interpretation and presentation of results.

#### 4.3.1 Specific Example 1: $C_3H_5O_2Cl$

The first formula,  $C_3H_5O_2Cl$ , contains two oxygens and one Ring or Double Bond equivalent (RDB), consistent with molecules such as carboxylic acids, keto ethers, esters or cyclic ethers and alcohols, with significant differences in fragmentation possibilities. The number of possible molecules generated using MOLGEN is 84 (excluding those structures with O-Cl bonds; including O-Cl bonds results in 110 structures). The six NIST spectra with this formula (excluding stereoisomers) are shown in Figure 16. The compounds are: 2-chloropropanoic acid (1), ethyl chloroformate (2), chloromethyl acetate (3), 3-chloropropanoic acid (4), methyl 2-chloroacetate (5) and 2-methoxyacetyl chloride (6). The PubChem [76] identities are 11734, 10928, 69366, 7899, 7295 and 96623, respectively.

The match values for the correct structure for a given spectrum are listed in Table 9. Instead of an additional table, the ranking results of the six spectra are presented Figure 17. The idea of this figure is to compare the ranking of the six 'known' molecules with each other, to see how well each program (with different settings) matches the structure and spectrum. If the programs match the correct structure and spectrum pair, the pattern of the top ranked structure for each spectrum, indicated by a cross, should be in a diagonal from top left to bottom right, indicated by the bold outlined squares in each matrix.

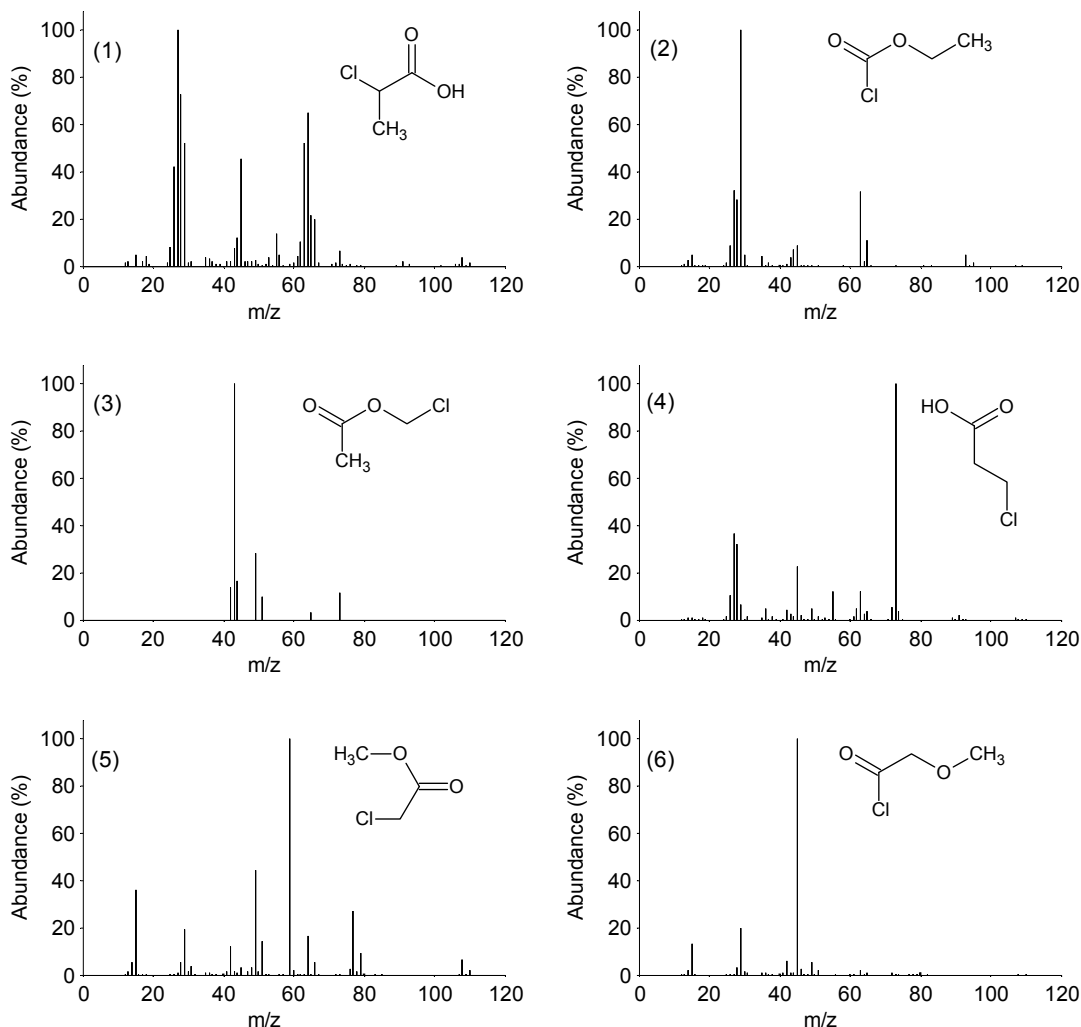
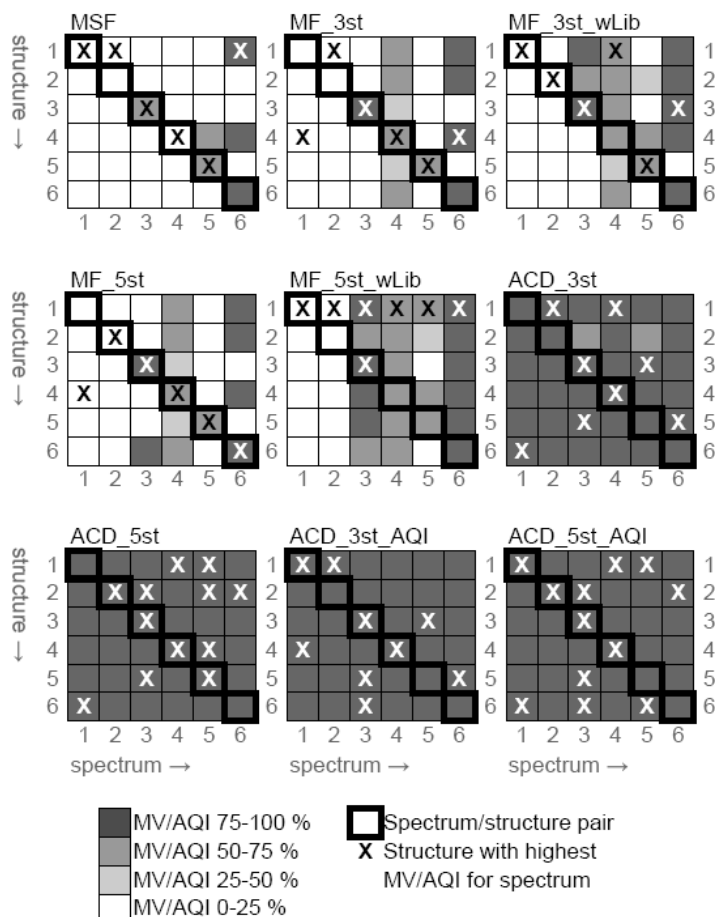


Figure 16: The spectra and structures of the six molecules with formula  $C_3H_5O_2Cl$  in the NIST database.

Table 9: Match values and Assignment Quality Indices of the actual molecule to the spectrum from NIST, predicted by the various programs with different settings. Abbreviations are given in Table 4.

| Spec. | Match Values |        |             |        |             |         |         | Assignment Quality Indices (%) |                 |
|-------|--------------|--------|-------------|--------|-------------|---------|---------|--------------------------------|-----------------|
|       | MSF          | MF_3st | MF_3st_wLib | MF_5st | MF_5st_wLib | ACD_3st | ACD_5st | ACD_3st_AQI (%)                | ACD_5st_AQI (%) |
| 1     | 0.1776       | 0.0938 | 0.1832      | 0.0938 | 0.1895      | 0.9640  | 0.9669  | 97.8                           | 98.4            |
| 2     | 0.0018       | 0.0472 | 0.0479      | 0.0492 | 0.0508      | 0.9550  | 0.9554  | 96.5                           | 96.8            |
| 3     | 0.6754       | 0.8119 | 0.9670      | 0.8600 | 0.9670      | 0.9884  | 0.9884  | 100.0                          | 100.0           |
| 4     | 0.0264       | 0.5493 | 0.5513      | 0.5494 | 0.5648      | 0.9877  | 0.9883  | 98.9                           | 99.1            |
| 5     | 0.6206       | 0.6206 | 0.6211      | 0.6336 | 0.6372      | 0.6336  | 0.6372  | 97.1                           | 97.5            |
| 6     | 0.7556       | 0.7559 | 0.7559      | 0.7625 | 0.7627      | 0.9786  | 0.9791  | 96.9                           | 97.4            |
| Avg.  | 0.376        | 0.480  | 0.521       | 0.492  | 0.529       | 0.918   | 0.919   | 97.9                           | 98.2            |



**Figure 17:** Matrices of the six  $C_3H_5O_2Cl$  structures (rows) and spectra (columns). The bold squares indicate the structure-spectrum pair (i.e. Structure 1 matches Spectrum 1). The crosses indicate the structure with the highest match value of the six structures, for a given spectrum, such that each column has at least one cross. More than one cross for a spectrum (column) indicates two or more structures with the highest match value or Assignment Quality Index. The shading indicates the approximate match value, as shown in the legend. The program abbreviations are given in Table 4.

The matrices (Figure 17) show a couple of interesting features for these small structures. With the exception of MF\_5st\_wLib, the results for Mass Frontier and MOLGEN-MSF are relatively accurate and comparable, picking the correct structure of the six spectra three to five times. Although MF\_5st (general reactions only) was the most accurate for these six structures, including the library fragmentation changed the situation dramatically, selecting the correct molecule only twice and additionally giving Structure 1 the highest match value for all runs. In comparison with the Mass Frontier and MOLGEN-MSF results, the distribution of results from ACD is far more chaotic, with often several structures selected as the best match. The ACD match values and Assignment Quality Indices were much higher than for Mass Frontier and MOLGEN-MSF (shown by the shading in Figure 17), but all structures had high match values, not



just the correct ones – a fact that does not reflect the differences in the structures and spectra shown in Figure 16.

These results were reflected in the relative ranking positions calculated for the programs for all 84 possible structures. The average relative ranking positions calculated for the six spectra for MOLGEN-MSF and Mass Frontier ranged between 0.0412 (MF\_3st), meaning the correct structure is in the top 4 %, to 0.2279 (MF\_5st\_wLib), with MOLGEN-MSF in the middle (0.1486). In contrast, the relative ranking positions for ACD ranged between 0.3404 (ACD\_3st) and 0.3936 (ACD\_5st), such that the correct structure is only in the top 34 and 40 % of all structures, reflecting the lack of specificity demonstrated in the matrices in Figure 17.

Comparing the matrices with the data included in Table 9 also indicates some counter-intuitive rankings for the match values calculated for the actual molecules. Although MOLGEN-MSF only predicts a match value of 0.0264 for Spectrum 4, the MSF matrix indicates that this structure was correctly identified for this spectrum, i.e. this match value was higher than the match value for the other five spectra. In contrast, although Structure 3 has an Assignment Quality Index of 100 % for Spectrum 3, the ACD\_5st\_AQI matrix in Figure 4.5 shows that at least 3 other spectra also had an Assignment Quality Index of 100 %, so that this is less selective than the much lower match value generated by MOLGEN-MSF for Spectrum 4. Additionally, although the correct structure for Spectrum 6 has match value close to 0.98 for the 3 and 5 step ACD calculations, this structure is not identified correctly for this spectrum. This shows that the match value of the correct structure gives little information about the ranking position of this structure in relation to all other possible structures.

#### 4.3.2 Specific Example 2: $C_5H_{12}S_2$

The second formula,  $C_5H_{12}S_2$ , includes dithiols, alkyl thiols or disulfides, with the main differences in the mass spectra resulting from differences in alkyl substitutions and symmetry of the structures. The number of possible structures generated in MOLGEN using divalent sulphur is 69. There are 11 NIST spectra with this formula, shown in Figure 18. The structures are: tert-butyl methyl disulfide (1), n-butyl methyl disulfide (2), 2,2-bis(methylthio)propane (3), ethyl n-propyl disulfide (4), methyl sec-butyl disulfide (5), 4-methylthio-1-butanethiol (6), 1,5-pentanedithiol (7), 1-methylethyl ethyl disulfide (8), bis(ethylthio)methane (9), 1,3-bis(methylthio)propane (10) and methyl isobutyl disulfide (11). The PubChem Compound Identities for these compounds are 141968, 521941, 525428, 35349, 522263, 525500, 70236, 521477, 78108, 141161 and 522262, respectively.

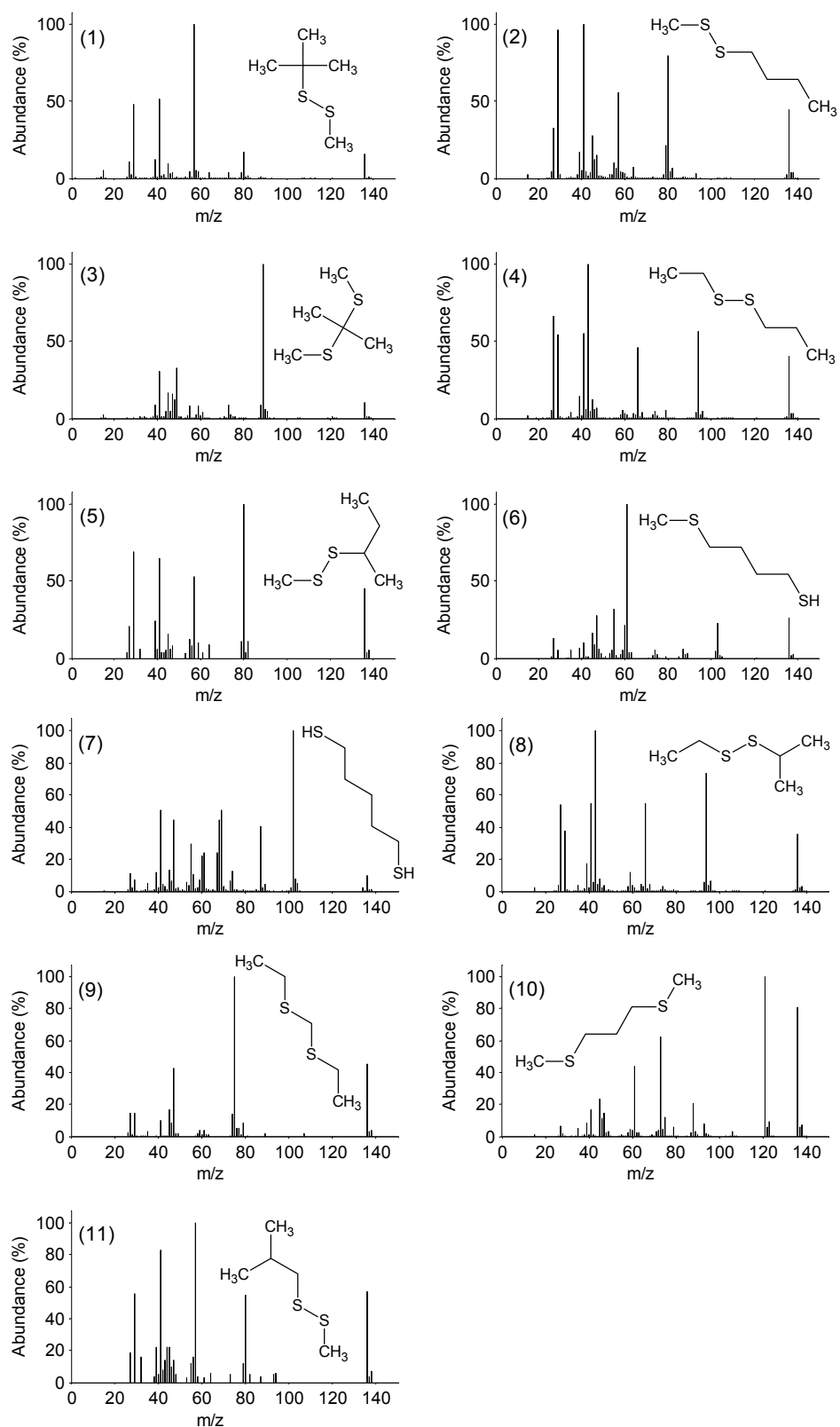
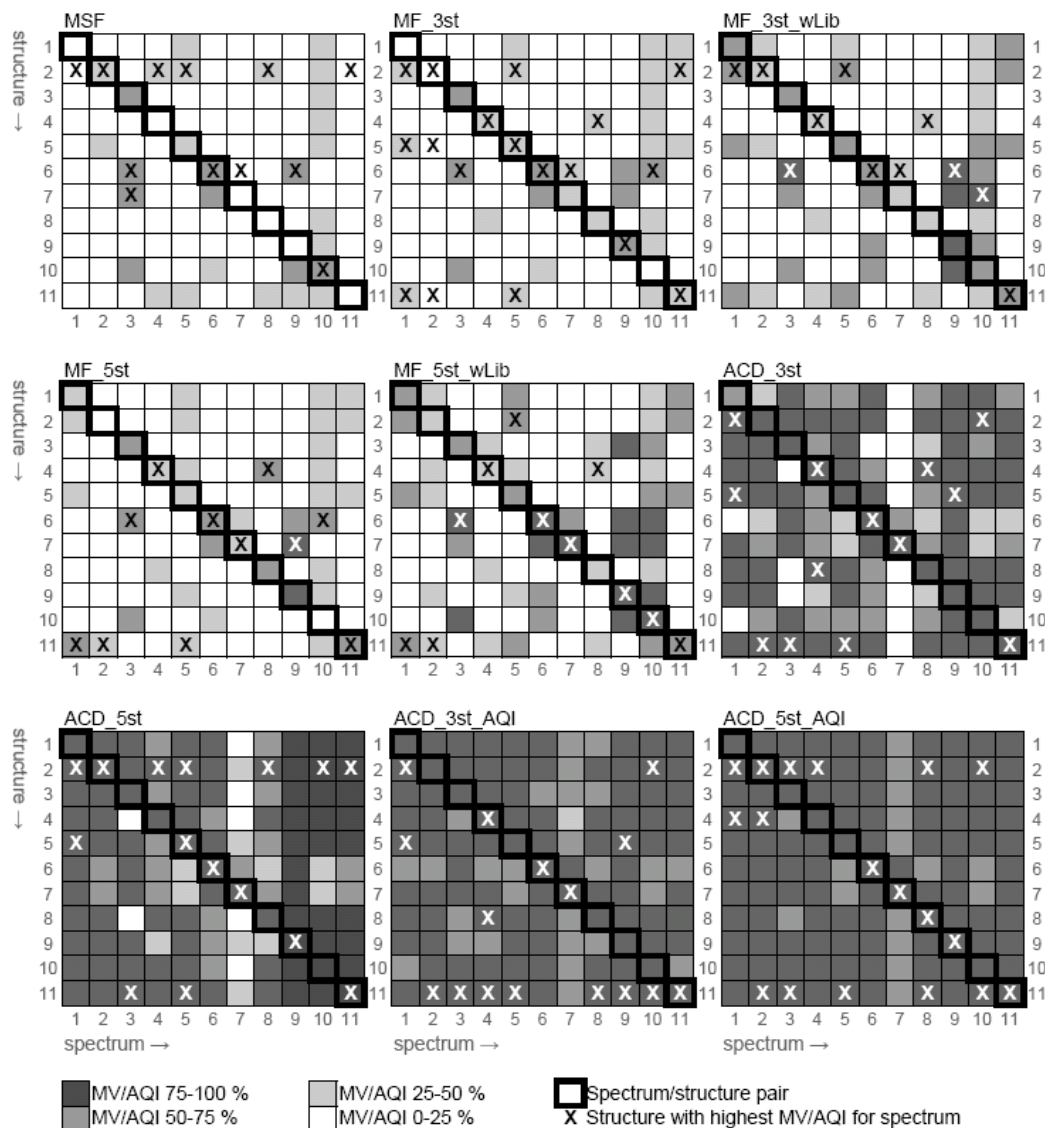


Figure 18: The spectra and structures of the eleven molecules with formula  $C_5H_{12}S_2$  in the NIST database.

The matrices, which compare the rank of the correct structure with the other ‘known’ structures (from NIST spectra) are presented in Figure 19. The match values were again generally much higher for the ACD calculations than for MOLGEN-MSF or Mass Frontier.



**Figure 19:** Matrices of the 11  $C_5H_{12}S_2$  structures (rows) and spectra (columns). The bold outlined squares indicate the structure-spectrum match (i.e. Structure 1 matches Spectrum 1). The crosses indicate the structure with the highest match value of the 11 structures, for a given spectrum, such that each column has at least one cross. More than one cross for a spectrum (column) indicates two or more structures with the same match value (MV) or Assignment Quality Index (AQI). The shading indicates the approximate match value, as shown in the legend. The naming nomenclature is as in Table 4.

The matrices demonstrate an interesting trend, with several programs showing a bias towards certain structures, including Structure 2 (MSF, MF\_3st, ACD\_5st), Structure 6 (MSF, MF\_3st) and Structure 11 (MF\_3st, MF\_5st, ACD, especially in the Assignment Quality Indices), indicated by many crosses in each row. What is also interesting with

these matrices is the structures that are not selected (no crosses in a row) – Structures 1 and 3 are never selected (either correctly or falsely), whereas Structure 10 is only selected twice. The favouring of Structures 2 and 6 could be because these are the least symmetrical chain molecules (i.e. with the most possible fragments), while Structures 1 and 3 are branched structures with few possible fragments.

The average number of fragments predicted for each structure (averaged over the 11 spectra) for each program and settings combination are presented in Table 10. The average number of fragments generated for Structures 2 and 6 (46.8 and 45.4, respectively) are much greater than the average (39.9), while the number of fragments generated for Structures 1 and 3 (32.5 and 31.0, respectively) are much lower.

**Table 10: Number of fragments predicted for Structures 1-11, averaged over all spectra, for each program and settings combination. The bottom row contains the average number of fragments generated for all structures and spectra for each program, whereas the final column contains the average number of fragments generated for that structure, over all programs. Program abbreviations are given in Table 4.**

| Structure      | MSF         | MF_3st      | MF_3st<br>_wLib | MF_5st      | MF_5st<br>_wLib | ACD_3st     | ACD_5st     | Average     |
|----------------|-------------|-------------|-----------------|-------------|-----------------|-------------|-------------|-------------|
| 1              | 21.6        | 29.0        | 33.0            | 29.0        | 35.0            | 37.4        | 42.5        | 32.5        |
| 2              | 38.8        | 40.6        | 48.6            | 48.2        | 64.3            | 41.5        | 45.5        | 46.8        |
| 3              | 22.5        | 20.6        | 24.3            | 20.6        | 49.6            | 37.2        | 42.5        | 31.0        |
| 4              | 28.1        | 39.5        | 45.4            | 48.2        | 54.5            | 38.9        | 42.6        | 42.4        |
| 5              | 29.7        | 39.7        | 50.8            | 43.4        | 58.1            | 41.0        | 44.9        | 43.9        |
| 6              | 36.1        | 43.0        | 53.0            | 46.8        | 63.3            | 35.0        | 40.6        | 45.4        |
| 7              | 30.9        | 29.3        | 49.3            | 37.3        | 64.0            | 39.0        | 41.5        | 41.6        |
| 8              | 24.0        | 37.4        | 39.4            | 37.4        | 41.4            | 38.7        | 42.8        | 37.3        |
| 9              | 21.5        | 28.3        | 45.8            | 38.1        | 59.4            | 37.0        | 43.3        | 39.0        |
| 10             | 31.7        | 20.7        | 41.2            | 20.7        | 60.1            | 36.8        | 44.5        | 36.5        |
| 11             | 31.5        | 35.4        | 42.5            | 44.3        | 55.4            | 42.3        | 46.0        | 42.5        |
| <b>Average</b> | <b>28.8</b> | <b>33.0</b> | <b>43.0</b>     | <b>37.6</b> | <b>55.0</b>     | <b>38.6</b> | <b>43.3</b> | <b>39.9</b> |

#### 4.3.3 Specific Example 3: $C_7H_6Cl_2O$

The third formula,  $C_7H_6Cl_2O$ , has 155,987 possible structures, although considering only those with a benzene ring present reduces this to 49 structures. There are 12 spectra in NIST with this formula, shown in Figure 20. These spectra include 2,6-, 3,4-, 3,5- and 2,4-dichloro-1-benzylalcohol, 2,4-, 2,6-, 2,3- and 3,5-dichloro-1-methoxybenzene, 4-chloro-1-chloromethoxybenzene and three dichloro-methyl phenol isomers. The PubChem Compound Identities for these compounds are, in the order indicated in Figure 20: 27156, 15728, 11119, 16127, 43236, 15684, 16126, 88801, 15292, 17077, 83521 and 36588. This formula was chosen to assess the ability of the different programs to discern between aromatic substitution isomers, as we had encountered difficulties identifying unknown spectra with this formula.

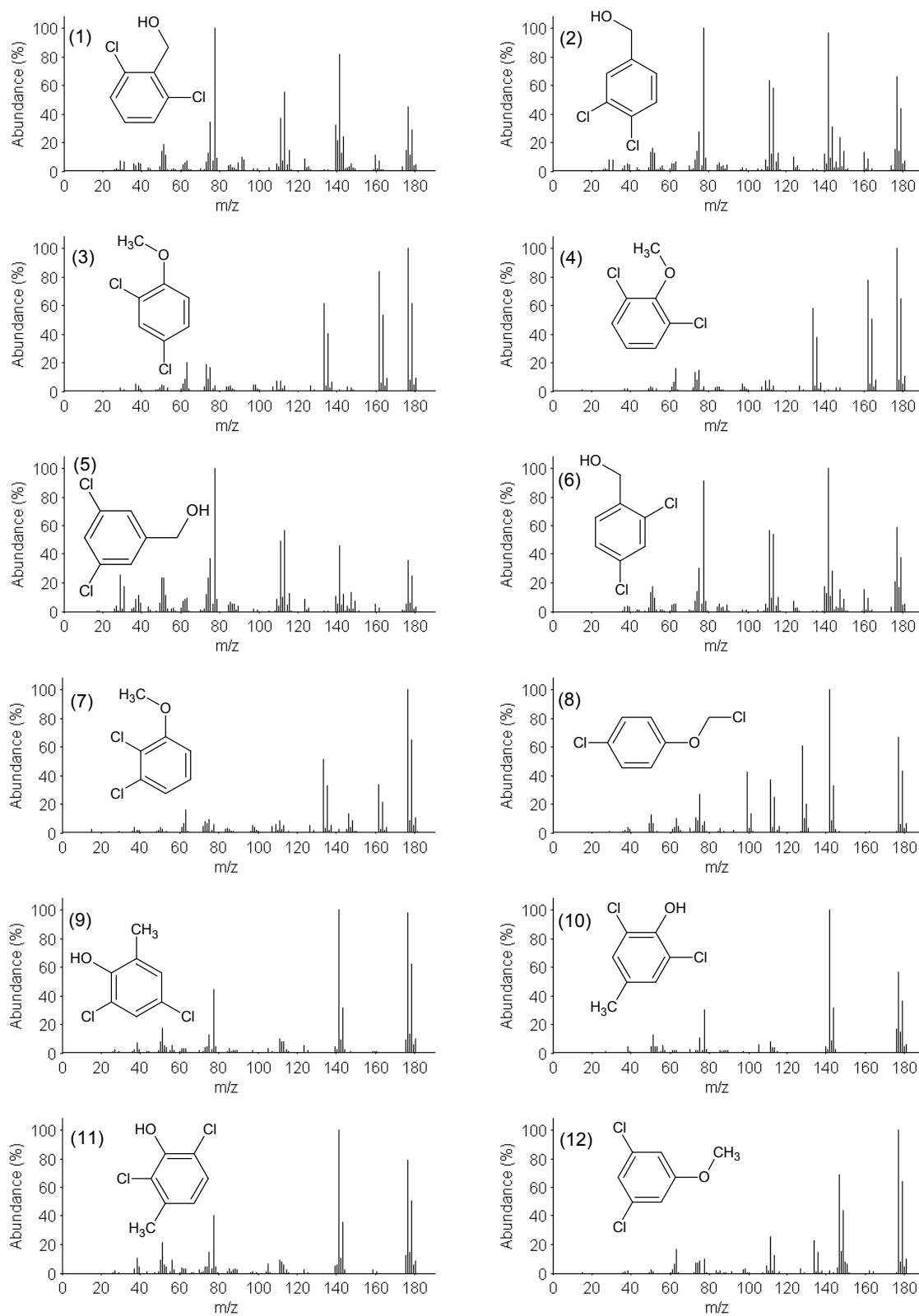
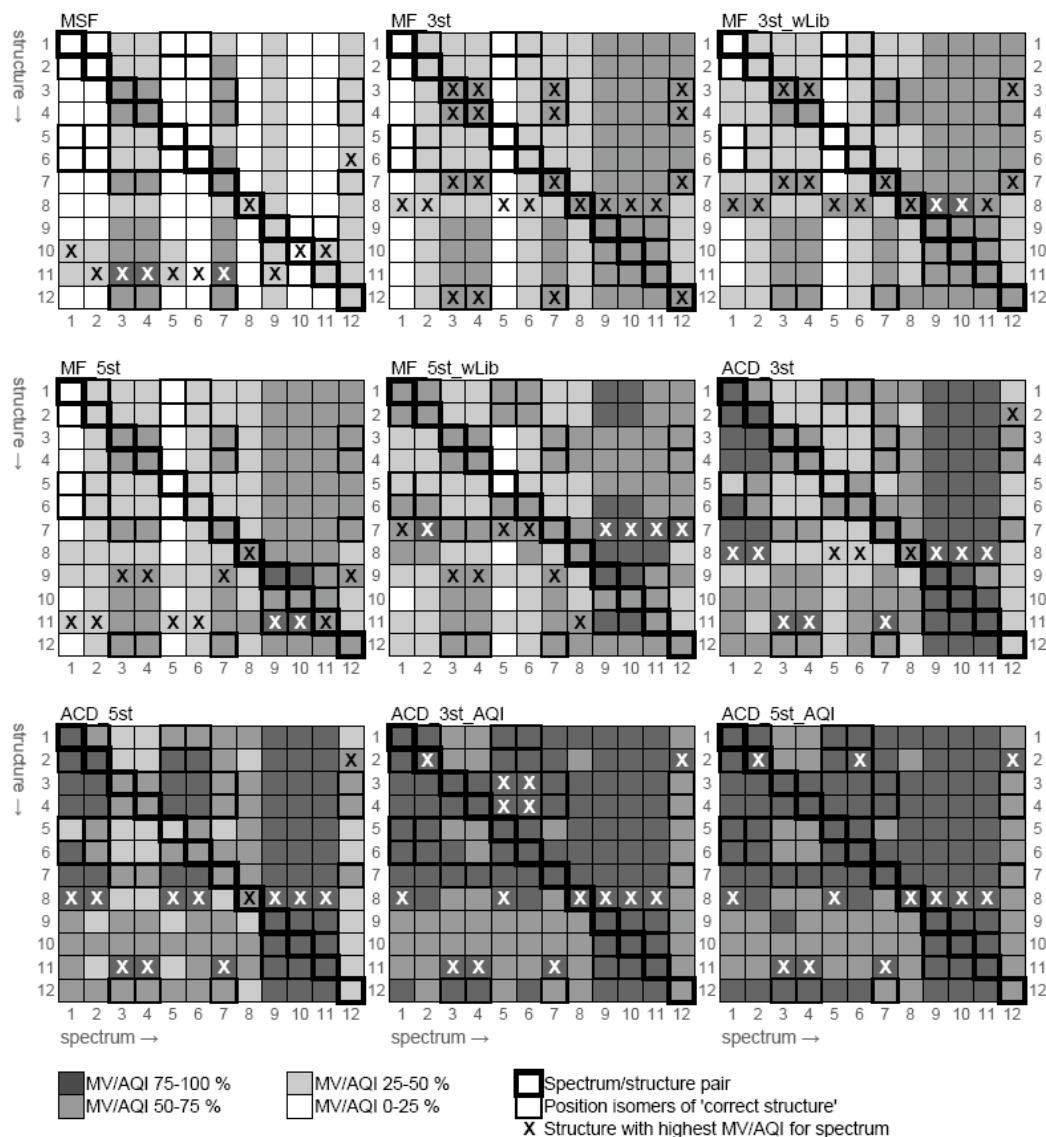


Figure 20: Spectra and structures for the 12 NIST spectra with formula  $C_7H_6Cl_2O$ .

The matrices from this example are shown in Figure 21, demonstrating the influences that the choice of a program and/or settings within the program can have on the outcome of structure ranking. Almost all of these cases show a bias in the program towards one structure above the others, but the actual structure ranked highest changes with minor modifications to program settings, especially for Mass Frontier. Structures 7, 8 and 11 feature very strongly in the results for Specific Example 3, with Structure 8 being the only structure picked correctly in all runs except MF\_5st\_wLib (where no molecule was picked correctly).



**Figure 21:** Matrices of the 12  $C_7H_6Cl_2O$  structures (rows) and spectra (columns). The abbreviations are given in Table 4.

Data on the number of fragments predicted ( $n_{pred}$ ), the number of fragments present in the experimental spectrum ( $n_{pres}$ ) and the number of fragments predicted that were not in the experimental spectrum ( $n_{add}$ ), in all cases averaged for each structure over all

calculations, are presented in Table 11. Structures 7 and 11, two of the three common matches, show a significantly higher number of fragments than the average (72.8 and 71.8, compared with the average 67 fragments). In contrast, Structure 8 does not have an above-average number of predicted fragments when compared with the other structures (65.3 fragments, compared with the average 67 fragments). The key features for Structure 8 are the ratios  $n\_pres/n\_pred$  and  $n\_add/n\_pred$ , where the previous is much higher than the average (0.71 compared with 0.61) and the former much lower (0.29 compared with 0.39). Interestingly, this data also shows that the number of predicted fragments for Structure 11 is significantly higher than for Structures 9 and 10 (71.8, compared with 65.6 and 62.9, respectively), with a similar trend in the number of fragments present, explaining the preferential selection of Structure 11 above 9 and 10, even though all are substitutional isomers. This may be due to molecular symmetry – Structures 9 and 10 have a degree of symmetry in them, whereas all substituents on Structure 11 are on one side, leaving more fragmentation possibilities open.

**Table 11: Average fragment data for each structure in Specific Example 3, over all program runs.  $n\_pred$  = number of predicted fragments,  $n\_pres$  = number of predicted fragments present in experimental spectrum,  $n\_add$  = number of predicted fragments not present in experimental spectrum.  $n\_pres/n\_pred$  indicates the ratio of present to predicted fragments,  $n\_add/n\_pred$  the ratio of additional fragments to predicted fragments.**

| Structure      | $n\_pred$    | $n\_pres$    | $n\_pres/n\_pred$ | $n\_add$     | $n\_add/n\_pred$ |
|----------------|--------------|--------------|-------------------|--------------|------------------|
| 1              | 66.42        | 48.62        | 0.61              | 24.92        | 0.39             |
| 2              | 66.51        | 48.63        | 0.61              | 25.03        | 0.39             |
| 3              | 71.36        | 50.92        | 0.58              | 28.62        | 0.42             |
| 4              | 69.01        | 49.67        | 0.59              | 27.08        | 0.41             |
| 5              | 61.26        | 43.75        | 0.59              | 24.52        | 0.41             |
| 6              | 68.06        | 49.60        | 0.61              | 25.85        | 0.39             |
| 7              | 72.83        | 51.96        | 0.59              | 29.22        | 0.41             |
| 8              | 65.32        | 51.62        | 0.71              | 19.18        | 0.29             |
| 9              | 65.63        | 46.64        | 0.61              | 26.58        | 0.39             |
| 10             | 62.93        | 45.21        | 0.61              | 24.80        | 0.39             |
| 11             | 71.76        | 52.36        | 0.63              | 27.17        | 0.37             |
| 12             | 63.85        | 44.56        | 0.57              | 27.00        | 0.43             |
| <b>Average</b> | <b>67.08</b> | <b>48.63</b> | <b>0.61</b>       | <b>25.83</b> | <b>0.39</b>      |

The average match values and Assignment Quality Indices calculated using ACD for the correct structure in this example, shown in Table 12, were much lower than the averages over the randomly-selected spectra (shown in Table 8), for example an average match value of 0.694 compared with 0.860 over 41 spectra for ACD\_3st. The relative ranking positions for MOLGEN-MSF and Mass Frontier as well as ACD were also much worse than the average relative ranking positions presented in Table 8, which partially explains our problems with identifying aromatic structures using match values. MF\_3st was the only calculation to successfully group any of the substitutional isomers together with the highest match values (for the dichloro-methoxy-benzene isomers), while including the library reactions split the isomer grouping to the detriment of the overall results.

**Table 12: Average match values, Assignment Quality Indices and relative ranking positions for the three specific examples, listed according to the program and settings used. Abbreviations are given in Table 4.**

|             | Specific<br>Example 1<br>$C_3H_5O_2Cl$ | Specific<br>Example 2<br>$C_5H_{12}S_2$ | Specific<br>Example 3<br>$C_7H_6Cl_2O$ | Specific<br>Example 1<br>$C_3H_5O_2Cl$ | Specific<br>Example 2<br>$C_5H_{12}S_2$ | Specific<br>Example 3<br>$C_7H_6Cl_2O$ |
|-------------|----------------------------------------|-----------------------------------------|----------------------------------------|----------------------------------------|-----------------------------------------|----------------------------------------|
|             | Match Values                           |                                         |                                        | Relative Ranking Position              |                                         |                                        |
| MSF         | 0.3762                                 | 0.2618                                  | 0.2879                                 | 0.1486                                 | 0.2861                                  | 0.5026                                 |
| MF_3st      | 0.4798                                 | 0.3945                                  | 0.4743                                 | 0.0412                                 | 0.1858                                  | 0.3047                                 |
| MF_3st_wLib | 0.5211                                 | 0.5378                                  | 0.4823                                 | 0.1265                                 | 0.2045                                  | 0.3585                                 |
| MF_5st      | 0.4920                                 | 0.4365                                  | 0.4963                                 | 0.0904                                 | 0.2273                                  | 0.4618                                 |
| MF_5st_wLib | 0.5287                                 | 0.6532                                  | 0.6305                                 | 0.2279                                 | 0.1965                                  | 0.3498                                 |
| ACD_3st     | 0.9179                                 | 0.9274                                  | 0.6940                                 | 0.3404                                 | 0.2848                                  | 0.3984                                 |
| ACD_5st     | 0.9192                                 | 0.9654                                  | 0.7102                                 | 0.3936                                 | 0.1598                                  | 0.4314                                 |
|             | Assignment Quality Index (%)           |                                         |                                        | Relative Ranking Position              |                                         |                                        |
| ACD_3st_AQI | 97.87                                  | 96.10                                   | 83.68                                  | 0.3635                                 | 0.3249                                  | 0.3837                                 |
| ACD_5st_AQI | 98.20                                  | 98.23                                   | 84.55                                  | 0.3865                                 | 0.2079                                  | 0.4115                                 |

#### 4.3.4 Using Classifiers to Eliminate Structure Candidates

The use of substructure classifiers from NIST and MOLGEN-MS in reducing the number of candidate structures for a spectrum (as in Section 3) is demonstrated based on known spectra. The results of Specific Example 1 (Section 4.3.1) were used to assess the use of the match values alone to identify the correct structure compared with the use of mass spectral classifiers followed by the match value (i.e. as performed in MOLGEN-MS). This is presented in Table 13.

**Table 13: The number of constitutional isomers and the relative ranking both without and with the consideration of mass spectral classifiers (Varmuza and NIST classifiers) for spectra from Specific Example 1 ( $C_3H_5O_2Cl$ ). MOLGEN-MSF was used to calculate the match values. TC: total number of candidates, BC: number of candidates with a higher match value, EC: number of candidates with match value equal to that of the correct structure (EC=1 if only the correct structure has that match value) and RRP = relative ranking position (see Equation 7). The spectrum numbers are given in Figure 16.**

| Spectrum | Without Classifiers |    |    |        | With Classifiers (95 % probability) |    |    |        |
|----------|---------------------|----|----|--------|-------------------------------------|----|----|--------|
|          | TC                  | BC | EC | RRP    | TC                                  | BC | EC | RRP    |
| (1)      | 84                  | 4  | 1  | 0.0482 | 2                                   | 0  | 1  | 0.0000 |
| (2)      | 84                  | 34 | 1  | 0.4096 | 1                                   | 0  | 1  | -      |
| (3)      | 84                  | 16 | 5  | 0.2169 | 1                                   | 0  | 1  | -      |
| (4)      | 84                  | 5  | 2  | 0.0663 | 19                                  | 3  | 2  | 0.1944 |
| (5)      | 84                  | 2  | 1  | 0.0241 | 1                                   | 0  | 1  | -      |
| (6)      | 84                  | 12 | 1  | 0.1446 | 1                                   | 0  | 1  | -      |

This table shows clearly, even for this small example, that the use of classifiers is instrumental in reducing the number of candidate structures prior to fragment generation and hence improving the chance of identifying the correct structure and limiting the number of other structures with higher match values. In four of the six spectra, the use of classifiers reduces the data set from 84 molecules to 1, in each case the correct structure. This means a greatly-reduced data set for identification purposes and in some cases a



relatively robust tentative identification, when no (or few) other molecules are possible for the given classifiers.

## 4.4 Discussion of Fragment Prediction

### 4.4.1 Does a High Match Value Mean a Better Ranking?

It is a feature of human nature to automatically react positively to a structure with a high match value and negatively to a structure with a low match value. We hope that the data presented here enables readers to question this automatic response, as it is clear that a positive reaction to a high match value can lead to a false sense of expectation regarding identification of a structure. Taking Spectrum 5 from Table A2, although MF\_5st has the highest match value (0.726), it has the worst RRP (0.501), compared with MF\_3st (MV=0.637, RRP=0.185) and MSF (MV=0.719, RRP=0.123). This also demonstrates another important fact, that the match values between the programs and settings are not directly comparable. In this case a very small difference in the match value between MF\_5st and MSF masks a large difference in the structure ranking, where for MSF only 12 % of the other possible structures have higher match values, while for MF\_5st, 50 % of the possible structures have higher match values. This makes selection of the ‘correct’ structure based on match values alone (i.e. setting a match value ‘threshold’) challenging. The quantiles presented in Table 7 also indicate that including the correct structure within the list of possible structures at a sufficiently high probability (e.g. 90 %) means taking such a low match value that almost all structures are included.

As the variation in individual examples is huge, the averages for the different program, setting and spectrum combinations can be used to draw some general conclusions (see Table 5 for the averages over 27 spectra and Figure 15 for the RRP). The average match value for the correct structure for MSF (0.354) and MF\_3st (0.381) are lower than most people’s ‘positive response limits’, while that for ACD\_5st (0.833) is significantly higher. However, despite the high match values assigned by ACD\_5st, this program setting combination demonstrated the least selectivity, with the worst average relative ranking position over the 27 spectra, at 0.535. This relative ranking position means that, on average, over 53 % of the constitutional isomers have a greater match value than the correct structure. If the match values had been assigned randomly to all structures, the average relative ranking position for the correct structure would be 0.50, meaning that ACD\_5st is actually, on average over these 27 spectra, slightly worse than randomly assigning match values to each structure. In contrast, MSF and MF\_3st, despite having low average match values, have the best relative ranking positions, at 0.352 and 0.375, respectively, over the 27 spectra. MF\_5st, although producing higher match values than MF\_3st (due to the calculation of more fragments), experiences a corresponding loss of selectivity, with the average relative ranking position increasing relative to MF\_3st for all

averages (Table 6). The Mass Frontier calculations with library reactions were also less selective than MF\_3st. Thus, although the use of additional settings increases the match values in all cases (as one would expect when the number of fragments increases) this is accompanied in all cases with a loss of predictive selectivity, demonstrated by the increasing relative ranking position. Although the simpler settings may miss many specific fragmentation pathways for the correct structure, it is clear that on average the additional fragmentation pathways increase the specific fragmentations for the ‘incorrect’ molecules more than for the correct molecules.

Figure 15 shows that this trend is consistent also for comparison over the smaller data sets, for instance including spectra with up to 500 structures improves the ACD\_3st and ACD\_5st relative ranking positions to below 0.5 (0.465 and 0.468, respectively), but this is still significantly higher than the comparable average relative ranking positions for MF\_3st and MSF (0.29 and 0.31, respectively). The increase in the average relative ranking position with decreasing number of structures can be explained by considering the variation in the structures. For a small data set (e.g. Spectrum 61,  $C_5H_{13}N$  with 17 possible structures), there are few variations in the combination of atoms in generating the structures, which corresponds to a decrease in the different fragmentation possibilities between the structures and thus decreases the probability of being able to use fragmentation to distinguish the structures successfully. Contrarily, large sets of structures have, generally, more combinational possibilities, greater numbers of possible fragmentation pathways and hence greater differences between the match values predicted for the structures.

#### 4.4.2 Match Value versus Assignment Quality Index

Another issue shown in this section is the danger in using ‘black-box’ indicators. The prediction of energies and barriers in the creation of fragments is difficult and is at this stage not sufficiently investigated to allow for incorporation into spectrum-structure match [28, 49]. The match value calculation, by taking the magnitude of the peaks from the experimental mass spectrum (see Equation 1), does not attempt in any way to predict the abundance of fragments, but instead uses the only information available (experimental) and provides a compromise solution while the prediction of fragment intensity remains in its infancy. The match value can also be calculated for any set of fragments, as long as this can be exported from the program generating the fragments in some way.

In contrast, while the ACD Assignment Quality Index attempts to incorporate the magnitude as well as presence of fragmentation peaks in the mass spectrum, the results included here, as well as several not included, show that this value should be regarded with some scepticism. The quantiles calculated for the 1000 randomly-selected spectra

(Table 7) gives a general demonstration of the Assignment Quality Index distribution. This shows that 95 % of the structures will have an Assignment Quality Index above 46 % (3 step) or 48.9 % (5 step fragmentation). This corresponds with match values calculated from the ACD results of 0.23 and 0.26, indicating that only 23-26 % of the experimental spectrum (abundance) is covered by the ACD fragments counted in the Assignment Quality Index. Likewise the 0.01 (99 %) quantile is 100 %, implying that for a spectrum with 10,000 constitutional isomers, on average 1 % or 100 structures will have an Assignment Quality Index of 100 %, which makes selection between these candidates impossible and, given the average relative ranking position for the ACD calculations, the top 1 % of structures is extremely unlikely to include the correct candidate structure anyway.

The discrepancies between the ACD results processed with the match values in comparison with the Assignment Quality Index results are also demonstrated in Specific Examples 1 and 2 (see Figure 17 and Figure 19) and, to a lesser degree, in Specific Example 3 (Figure 21). In these figures it is apparent that the Assignment Quality Index changes the relative ranking of the structures compared with the match value, in some cases with significant differences in the top candidate selection (see especially the increased bias towards Structure 11 in Figure 19). As the Assignment Quality Index is not clearly defined [50], we are not able to shed light on the nature of the differences. The developers themselves also offer a few words of caution regarding this index, adjacent to its description [50], see Section 2.7.2.

Another reason to exercise caution when interpreting the results of ACD is the presentation of only the fragments present in the experimental spectrum, not all fragments calculated. The absence of predicted fragments that are not in the experimental spectrum results in the loss of a crucial additional interpretation tool. Both MOLGEN-MSF and Mass Frontier perform fragmentation calculations independent of the mass spectrum, fragmenting each structure according to the given rules/settings and then comparing these results with the experimental spectrum. Although the match value is only calculated on those fragments present in the experimental spectrum, alternative outputs in MOLGEN-MSF include the export of all fragments, such that this information is still recoverable to the user, both for MOLGEN-MSF and Mass Frontier inputs. As the ACD calculation requires input of the experimental spectrum from the beginning, it appears that the ‘additional’ fragments (i.e. those not present in the experimental spectrum) are filtered out of the results before these are presented to the user. No possible adjustment to the settings was found to export all fragments generated, rather than just those present in the experimental spectrum. Given that the ACD match values are significantly higher than those for MOLGEN-MSF and Mass Frontier, it is likely that the

ACD program calculated many more fragments, both present and absent, than either MOLGEN-MSF or Mass Frontier, but this cannot be confirmed at this stage.

#### 4.4.3 *Candidate Inclusion/Exclusion*

The results shown in this section highlight the problems associated with considering a limited subset of constitutional isomers and using the assigned fragments to prove (or disprove) the match of the structure to spectrum. Several examples above have shown that in many cases the ‘correct’ molecule can have a very low match values, or that several other molecules can have much higher match values, such that distinguishing ‘correct’ from ‘incorrect’ is very difficult based on the match value or fragmentation patterns alone. Even the best program and setting combinations (Mass Frontier with 3 step fragmentation and MOLGEN-MSF) can only reduce the number of possible candidates (on average) to 27 % of all possible molecules for that spectrum’s molecular formula, although to ensure inclusion of the correct structure with 90 % certainty, many more molecules have to be included in most cases (expressed by the quantiles in Table 7). While consideration of all possible structures allows at least an objective overview of the match value range, consideration of a limited subset (e.g. only those structures in a database) is unlikely to give the full distribution of match values and could result in incorrect selection of an apparently good match. Even for spectra with a small number of possible structures (say 100), the inclusion of a significant percentage of the total possible candidates can result in consideration of over 30 candidate structures, which is already impractical for rapid identification/confirmation purposes. Substructural identification prior to structure generation can be used to reduce the number of candidate structures prior to calculation of match values, shown in Table 13, Section 4.3.4.

#### 4.4.4 *Which Program, Which Settings?*

The time and ease of calculation plays a significant role in the selection of program and settings. Taking Spectrum 97 (Table A2) as an example (with 175 candidate structures), although the MSF, MF\_3st and MF\_5st calculations were all completed in under 1 minute (Intel Core™ 2 Duo 1.83 GHz, 1.00 GB RAM), ACD\_3st needed 28 minutes, ACD\_5st 62 minutes, MF\_3st\_wLib 184 minutes and MF\_5st\_wLib 796 min (13 hrs, 16 min). An attempt to apply Mass Frontier with 3 step fragmentation and library rules to Spectrum 1 (1902 candidates) resulted in a 2 day, 7 hour calculation, despite the same settings without library rules taking only 3 minutes. The time of calculation also varied significantly with the molecular formula, not just the number of candidates. Spectrum 34 (C<sub>11</sub>H<sub>24</sub>), with a relatively simple formula and hence fragmentation patterns, needed < 1 min for MSF, MF\_3st and MF\_5st, 12 min for ACD\_3st, 49 min for ACD\_5st, 5 hrs 17 min for MF\_3st\_wLib and 19 hrs 16 min for MF\_5st\_wLib for 159 candidate structures. In contrast, Spectrum 68 (C<sub>5</sub>H<sub>13</sub>NO), with two heteroatoms and more complicated

fragmentations, needed less time for all calculations (for 149 candidate structures, not 159) except MF\_5st\_wLib, which took 32 hrs and 24 minutes, a significant increase compared with Spectrum 34 and an average of 13 minutes per structure, for a relatively small compound. Although these calculations could probably be sped up by installing the software on more powerful machines (e.g. a central server), commercial software typically has certain licence restrictions associated with it and correspondingly higher licence costs for network licences (ACD) and it is, to the best of our knowledge, not possible to install Mass Frontier over a network at this stage.

Other realistic considerations when choosing a program are accessibility. MOLGEN-MS is available for purchase from [www.molgen.de](http://www.molgen.de). MOLGEN-MSF was especially compiled for this study and is available for interested parties upon request. Contact details and the user manual are also available from [www.molgen.de](http://www.molgen.de). ACD MS Manager is readily available for purchase, with the exact cost depending on the licence required, additional modules purchased and employment status (student, education, commercial). Mass Frontier is less accessible than the ACD MS Manager to the interested user, as it is bound to the Thermo Scientific Xcalibur software and at this stage can only be run as single licence software on a machine with Xcalibur installed and is also more expensive than ACD MS Manager. Whichever program is chosen, users should treat the results with caution and consider additional information, not just match values, to confirm the identity of tentatively identified compounds, prior to purchase of standards for confirmation.

Note: MOLGEN-MS, MOLGEN-MSF, ACD MS Manager and Mass Frontier all contain many more settings and features that were not considered here. Additional fragmentation settings not considered could influence the outcomes considerably.

#### ***4.5 Implications and Conclusions***

Despite the hope expressed in 2006 [7] in relation to the improvement in fragmentation and match value calculation by using more sophisticated computer programs, this study indicates that the desired improvements are not yet a reality. The results presented here show convincingly that the simplest and quickest of the program and settings combinations (Mass Frontier with 3 step fragmentation and MOLGEN-MSF) are still the most effective in terms of ranking the correct structure relative to other constitutional isomers based on electron impact mass spectra, despite the lower match values. Longer calculation times with more fragmentation steps or including library reactions to produce higher match values generally resulted in a decreased selectivity.

The specific examples used demonstrate the bias of all programs towards certain structures, even for different mass spectra with the same molecular formula. This bias

can change significantly with minor changes in program settings and is often related to the number of fragments predicted in total, not just fragments present in the mass spectrum. Specific Examples 2 and 3 show that the bias is often towards the more asymmetrical molecules (with a greater number of possible fragments) and away from the symmetrical molecules (with fewer possible fragments resulting from the symmetry). This means that asymmetrical molecules will be selected more often, whether correct or incorrect. This has implications for the selection of candidates to investigate during unknown investigations.

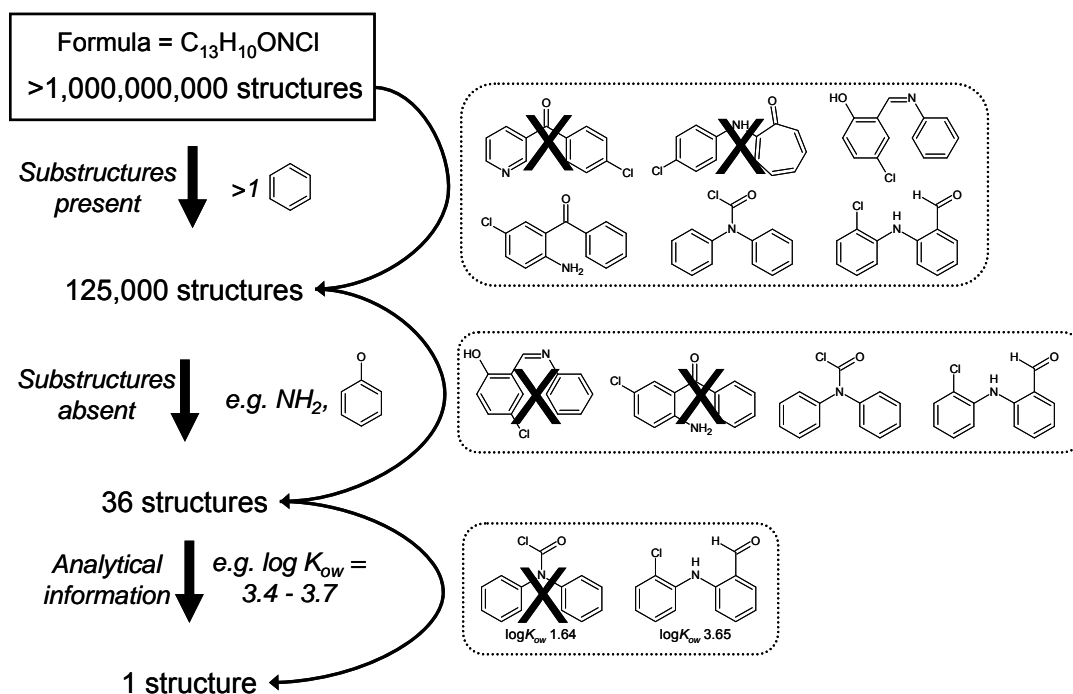
Although it is apparent that the number of fragments produced by the programs has some influence on the program bias towards some structures, we have not yet been able to incorporate this information in any meaningful way to result in a positive impact on the relative ranking position. An alternative to improve the assessment of the fragment-spectrum relationship would be to focus efforts on the prediction of fragment intensity, despite the inherent problems involved.

As this study was unable to identify a combination of program settings to improve the relative ranking position significantly above the 0.27 reported previously [7], the conclusion that the generation of fragments and match values alone are not sufficient, at this stage, to allow for computer-aided structure elucidation (CASE) via electron impact mass spectra remains the same. This leaves CASE via MS significantly behind that of other analytical techniques such as NMR (see for example the recent review [77], which details much better success rates using other software developed by ACD).

Incorporating additional information in candidate selection and developments in high resolution and tandem MS techniques opens up new windows to improve CASE via MS (see e.g. Sections 3 and 5 here). The alternative strategy for matching structures and spectra based on combinatorial structures and energies rather than fragmentation prediction, implemented in the software FiD [52] shows promising results for tandem MS data and should be investigated further. As this strategy does not appear to support low resolution data, it was not considered here. Instead, additional criteria need to be incorporated into candidate selection to improve CASE for GC-EI-MS data. This is explored in Section 5.

## 5 Strategies for Structure Elucidation

The outcomes of the method development based on unknown spectra (Section 3) showed that the combination of NIST and Varmuza substructure classifiers are instrumental in reducing the number of structures generated per spectrum down to a ‘manageable’ data set, but that additional criteria are needed to further reduce the number of possible matching structures and to separate the ‘correct’ structures from the incorrect ones. This is shown schematically in Figure 22, where first the substructures present and then those absent reduce the number of possible structures down to 36, while the analytical information results in the identification of one matching structure.



**Figure 22:** The use of substructures and analytical properties to eliminate candidate structures using structure generation methods. The structures included are for demonstration purposes. The number of structures was calculated using MOLGEN 3.5 [34], substructure information using MOLGEN-MS [29] and NIST [17], with  $\log K_{ow}$  calculated using Kowwin (EPISuite™ [60]). Each group of structures represents the reduction of candidate numbers from top to bottom (indicated by the curved arrows), using the criteria within the curve of the arrow.

It is clear from the results and discussion in Section 4 that mass spectral fragmentation and the resulting match values are not (at this stage) sufficiently reliable to be used as an exclusive criterion to eliminate significant numbers of candidates directly following structure generation. Thus, additional criteria were considered to improve the method explored in Section 3.

This section goes through the whole process of identifying candidates from a GC-EI-MS spectrum using structure generation techniques, building on the results from the previous two sections. The strategies investigated were considered not only in terms of

effectiveness but also in terms of method automation. Firstly, Section 5.1 addresses the question of molecular formula determination, without which structure generation cannot be performed. This material formed part of a book chapter, published in 2011:

Schymanski, E., Schulze, T., Hermans, J., Brack, W. (2011). Chapter 8: Computer Tools for Structure Elucidation in EDA, *Handbook of Environmental Chemistry: Effect-Directed Analysis of Complex Environmental Contamination*, Vol. 15, Ed. W. Brack, Springer-Verlag Berlin Heidelberg.

Sections 5.2, 5.3 and 5.4 then cover the use of retention indices, match values and steric energy as exclusion criteria, based on 29  $C_{12}H_{10}O_2$  spectra from the NIST database and measurement data. The method for structure generation and progressive elimination of candidates, based on these results, is presented in Section 5.5, including details concerning automation of the work flow. Section 5.6 covers the results for the 29  $C_{12}H_{10}O_2$  spectra, which are then discussed in Section 5.7, followed by implications and conclusions in Section 5.8. This work formed a manuscript, published in 2011:

Schymanski, E., Meringer, M. and Brack, W. (2011) Automated Strategies to Identify Compounds on the Basis of GC/EI-MS and Calculated Properties, *Analytical Chemistry*, 83, 903-912.

The  $C_{12}H_{10}O_2$  spectra contain a number of potentially environmentally relevant compounds with different functional groups and several substitution isomers. The method proposed here was subsequently tested on unknown spectra, which is presented along with other successful examples of unknown identification in Section 6.

### 5.1 Calculation of Molecular Formula

One of the major difficulties with EI-MS, mentioned in Section 2.2, is the determination of the molecular ion and subsequently the molecular formula. For EI-MS spectra, successful determination of the molecular formula depends generally on the presence of isotope peaks for the molecular ion ('M' peak), as only a few, low molecular weight compounds have only one formula possible for a given molecular weight. However, the molecular ion may be missing in up to 30% of EI-MS spectra [28], while many more spectra have absent or only very small isotope peaks, resulting in inaccurate calculations. This means that selection of the correct molecular formula can be a challenge - and the number of candidate structures increases with every formula if multiple formulae need to be considered. A few options exist to counter this problem, such as inclusion of substructure information from NIST, MOLGEN-MS or AMDIS to restrict element information, the ring or double bond count (RDB, see Equation 11, Section 3) or using isotope patterns from fragment peaks to generate partial molecular formulae. The NIST



substructure information contains relatively reliable RDB indications. Furthermore, if it is known that certain substructures are present (e.g. benzene, RDB=4), formulae with incompatible RDBs can be eliminated (e.g. those with RDB<4).

Softer ionisation techniques help avoid the problem of molecular formula determination as these result in less fragmentation, such that the ‘M’ peak is present in the spectrum. In combination with accurate mass measurement, this can reduce the number of possible formulae more effectively than the isotope patterns in EI-MS. This section gives a brief overview of the programs and strategies available to calculate the molecular formula from both low and high accuracy mass data and offers a quick comparison of these strategies.

### 5.1.1 Programs to Calculate Molecular Formulae

A selection of different programs available to assist in the prediction of the molecular formula for both low and high resolution data is given in Table 14. Only the first two programs are based on both mass accuracy and the isotope patterns, the others on mass accuracy only. The ‘Odd/Even Ion Display’ is relevant for comparing results of different methods. For GC-EI-MS, Odd Electron Ions (OEIs) are created when the molecule loses (or gains) an electron (e.g. the ‘M’ peak), whereas Even Electron Ions (EEIs) are created by the loss or gain of atom(s), not just electrons, i.e. fragments. LC-based ionisation methods often result in the addition or removal of H from ‘M’ and thus generally produce EEIs; fragments can however be EEIs or OEIs. The availability of RDB information (either as a filter or display) is also indicated in the table.

**Table 14: Example programs for calculating molecular formulae from MS data. The Mass/Isotope column indicates whether the isotope pattern or mass difference is used to distinguish candidates, odd/even ion display whether the program displays the ion information, RDB whether it takes RDB information as input to restrict the generated candidates.**

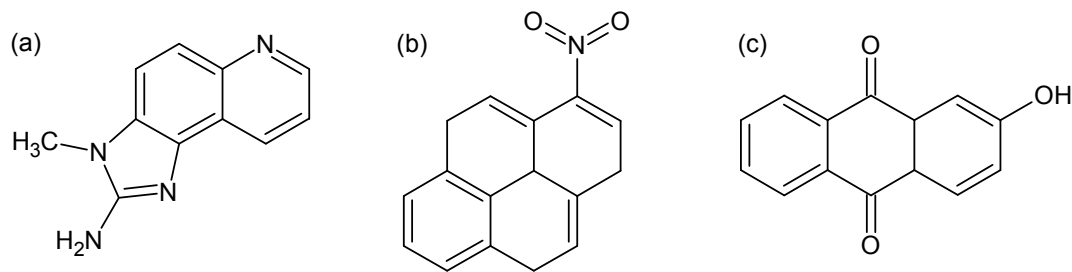
| Program                | Mass / Isotope | Odd/Even Ion Display | RDB | Elements                 | Availability                              |
|------------------------|----------------|----------------------|-----|--------------------------|-------------------------------------------|
| MolForm                | both           | yes                  | no  | C,H,N,O,S,Si,P,Br,Cl,F,I | purchase of MOLGEN-MS [29]                |
| ACD Formula Generator  | both           | yes                  | yes | max. 10, user selection  | purchase of MS Manager or similar [30]    |
| ChemCalc.org           | mass           | no                   | yes | C,H,N,O,Br,Cl,F,I        | free, online [78]                         |
| NIST Formula Generator | mass           | yes                  | yes | max. 10, user selection  | purchase as part of database [17]         |
| Xcalibur               | mass           | no                   | yes | all, user selection      | purchase or supplied with instrument [79] |
| Uni Cambridge          | mass           | no                   | no  | C,H,N,O or all.          | free, online [80]                         |

The number of elements considered by the software and the ability of the user to select these elements influences the outcome of calculations considerably. While some programs restrict their selection to selected elements only (e.g. NIST Formula Generator and MOLGEN-MS with 10 or 11 elements respectively), others offer no restrictions at all. Both strategies have their advantages and disadvantages. Naturally, the more elements are included, the greater the possibilities, so choosing which elements to

include or exclude becomes critical. Excluding possible elements too early without taking this information from the spectrum or other analytical data reduces the number of possible formulae but also the objectivity of the unknown determination, whereas including all the possible elements results in the generation of possibly hundreds of incompatible and highly unlikely formulae.

### 5.1.2 Comparison of Molecular Formula Calculations

The number of formulae and the rank of the correct formula calculated for three compounds, shown in Figure 23, were obtained using the programs listed in Table 14 and the different strategies discussed here. The structures were 2-amino-3-methyl-3H-imidazo(4,5-f)quinoline (IQ), 1-nitropyrene (1NP) and 2-hydroxy-9,10-anthraquinone (2HA). All calculations considered C, H, N and O only for consistency between programs. Exact mass calculations were based on a measured mass of 199.0989 (IQ), 248.0724 (1NP) and 225.0560 Da (2HA), recorded during standard measurement by C. Gallampoio on LC-MS high resolution, high accuracy Orbitrap with Atmospheric Pressure Chemical Ionisation (APCI) in positive mode. The calculated exact masses are 199.0978, 248.0706 and 225.0546 Da, with differences of 5.4, 7.2 and 6.1 ppm, respectively. Isotope patterns were taken from EI-MS spectra retrieved from the NIST database (spectrum numbers 138963 (IQ), 101258 (1NP) and 132776 (2HA)). Substructural information from the NIST database was used to restrict element numbers in the isotope peak calculations. All calculations took less than 1 sec. MolForm calculations considered odd electron ions only.



**Figure 23:** Structures of (a) 2-amino-3-methyl-3H-imidazo(4,5-f)quinoline (IQ), (b) 1-nitropyrene (1NP) and (c) 2-hydroxy-9,10-anthraquinone (2HA).

The results, shown in Table 15, show clearly that the exact mass calculations were the most successful, with the correct formula ranked first in one third of all calculations, with the worst ranked 6<sup>th</sup>. The results for 1NP indicate the influence that the slightly larger measurement error (compared with the other two compounds) has on the ranking of the correct formula. These results are likely to be even better with higher accuracy data (error margins below 5 ppm are generally achievable on the Orbitrap).

**Table 15: Calculation of the molecular formula of IQ, 1NP and 2HA (see Figure 23) using selected programs (see Table 14). ‘N. Formulae’ indicates the number of possible formulae generated for the data by the given program, whereas ‘Rank of Correct’, in brackets, indicates where the correct formula was in relation to the others, based on the sorting criterion (exact mass, isotope pattern, whole spectrum match – see sub-headings).**

|                                                                |                               | <b>IQ</b> | <b>1NP</b> | <b>2HA</b> |
|----------------------------------------------------------------|-------------------------------|-----------|------------|------------|
| <b>Exact Mass</b>                                              |                               |           |            |            |
| ACD                                                            | N. Formulae (Rank of Correct) | 17 (1)    | 26 (5)     | 24 (2)     |
| Cambridge                                                      | N. Formulae (Rank of Correct) | 60 (2)    | 60 (5)     | 60 (4)     |
| ChemCalc                                                       | N. Formulae (Rank of Correct) | 268 (1)   | 515 (4)    | 394 (1)    |
| MolForm                                                        | N. Formulae (Rank of Correct) | 7 (1)     | 14 (4)     | 12 (1)     |
| NIST                                                           | N. Formulae (Rank of Correct) | 2 (1)     | 5 (3)      | 3 (2)      |
| Xcalibur                                                       | N. Formulae (Rank of Correct) | 10 (2)    | 6 (6)      | 4 (3)      |
| <b>Isotope Pattern ‘M’ Peak</b>                                |                               |           |            |            |
| ACD                                                            | N. Formulae (Rank of Correct) | 187 (117) | 330 (98)   | 266 (44)   |
| MolForm                                                        | N. Formulae (Rank of Correct) | 76 (4)    | 126 (8)    | 116 (13)   |
| <b>Isotope Pattern ‘M’ Peak with Substructural Information</b> |                               |           |            |            |
| ACD                                                            | N. Formulae (Rank of Correct) | 32 (11)   | 2 (1)      | 9 (4)      |
| MolForm                                                        | N. Formulae (Rank of Correct) | 12 (2)    | 2 (2)      | 9 (5)      |
| <b>Whole EI-MS Spectrum Calculation</b>                        |                               |           |            |            |
| EICoCo                                                         | N. Formulae (Rank of Correct) | 3 (1)     | 7 (6)      | 7 (3)      |

The use of the isotope pattern of the ‘M’ peak alone is clearly insufficient to isolate the correct formula for these compounds (ranked between 4<sup>th</sup> and 117<sup>th</sup> of all possible formulae). The inclusion of substructural information from the NIST database search of the EI-MS spectra, however, improved the calculation substantially, with the correct formula ranked first or second in three of six cases and 11<sup>th</sup> in the worst case. The ‘M’ peak isotope pattern combined with substructural information was better at identifying the formula for 1NP, where the exact mass had the higher error. The number of formulae generated was also significantly lower when including substructural information, from over 100 matching formulae in most cases to a maximum of 32 formulae. The use of the ‘whole spectrum’ in the calculation by the EICoCo module of MOLGEN-MS (see Section 3 and [29]) produced results similar to the exact mass calculation.

Table 16 shows an example calculation for 1,2-dichloroethane, where the ‘M’ isotope peaks are of low abundance (see Figure 2(b), where the strong peaks at 62 dominate the ‘M’ peak at 98). The table includes a calculation based on the fragment isotope peaks at 62 as well as on the ‘M’ peak at 98, using MolForm. For the sake of comparison, the ‘whole spectrum calculation’ used in EICoCo is also included. All calculations were performed without any restriction to element numbers, based on C, H, O, N, S, Si, P, Cl, Br, F, I.

**Table 16: Calculation of the formula for 1,2-dichloroethane (Figure 2(b)) based on fragment isotope peaks (left) and the ‘M’ peak using MolForm (middle), sorted based on isotope pattern match, including all ions. The column to the right shows a calculation based on the whole spectrum using ElCoCo, sorted according to match value. The correct formulae are bolded.**

| <b>MolForm</b>                                       | <b>MolForm</b>                                  | <b>ElCoCo</b>                                        |
|------------------------------------------------------|-------------------------------------------------|------------------------------------------------------|
| Partial formula based on fragment isotopes at m/z=62 | Formula based on isotopes at 'M' peak, m/z=98   | Formula based on whole spectrum MW=98, precision 95% |
| <b>C<sub>2</sub>H<sub>3</sub>Cl</b>                  | Cl <sub>2</sub> N <sub>2</sub>                  | <b>C<sub>2</sub>H<sub>4</sub>Cl<sub>2</sub></b>      |
| CHCIN                                                | CCl <sub>2</sub> O                              | C <sub>2</sub> H <sub>7</sub> ClS                    |
| H <sub>6</sub> Si <sub>2</sub>                       | CH <sub>2</sub> Cl <sub>2</sub> N               | C <sub>2</sub> H <sub>4</sub> ClFO                   |
| H <sub>2</sub> SSi                                   | <b>C<sub>2</sub>H<sub>4</sub>Cl<sub>2</sub></b> |                                                      |
| CH <sub>6</sub> OSi                                  | H <sub>3</sub> ClSSi                            |                                                      |
| (top 5 of 40)                                        | (top 5 of 183)                                  | (only 3 results)                                     |

In this case the correct partial formula (shown in bold type) is top of the list for m/z=62, along with the whole spectrum calculation, while the complete formula is only fourth on the list for m/z=98. This shows clearly that the whole spectrum calculation can yield better results than calculations based on the isotope peaks of the ‘M’ peak alone, as can generation of a partial formula. However, the partial formula either needs to be completed by the user or used as input into a calculation similar to those shown in Table 15 to determine the rest of the formula.

Although the calculation based on accurate mass is straightforward, the six programs used in Table 15 came up with quite different results for the same input masses and elements. As for the calculations based on isotope patterns, restriction of the numbers of elements can be critical in reducing the number of candidate formulae to manageable levels. Any fragmentation information available can also be used to determine which elements are likely or unlikely. Examples include the neutral loss of NO from nitro-PAHs or the ionization of acidic groups in negative ion mode compared with the basic groups in positive mode [3]. Furthermore, if there is isotope information available, a combination match value can be calculated with some programs, based on match to the exact mass and to the isotope pattern [29, 33]. More details on molecular formula determination for exact mass data are given elsewhere (e.g. [31, 32]).

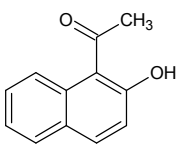
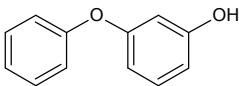
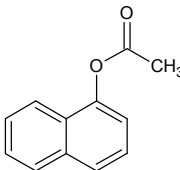
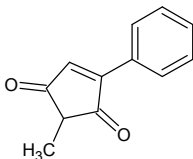
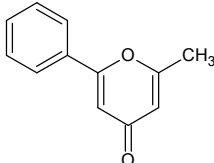
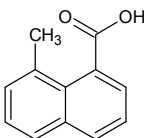
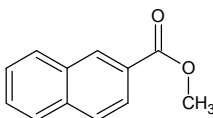
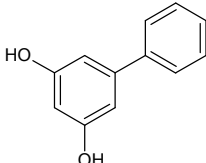
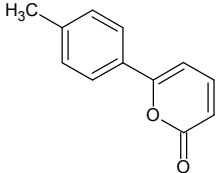
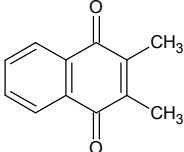
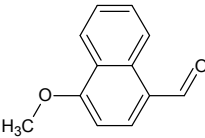
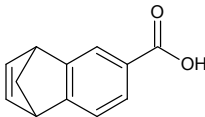
Once the molecular formula is determined, structure generation can begin in earnest. The rest of Section 5 describes the extension and automation of the structure generation method based on EI-MS spectral interpretation and the incorporation of additional analytical information, developed in Section 3. Firstly, two additional strategies for the selection or elimination of structural candidates were evaluated, a retention-boiling point correlation [58] and the use of steric energy. The automated strategy is then presented and tested on a case study using known spectra of formula C<sub>12</sub>H<sub>10</sub>O<sub>2</sub>.

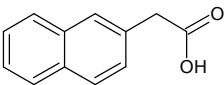
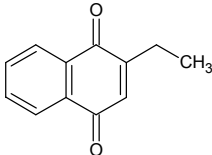
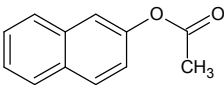
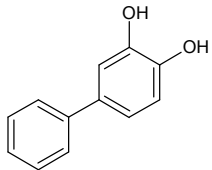
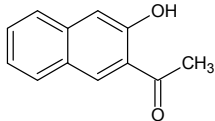
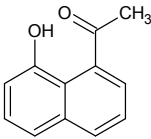
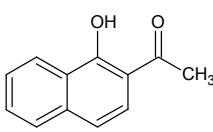
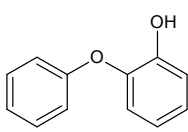
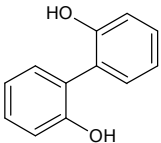
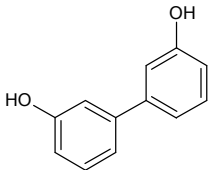
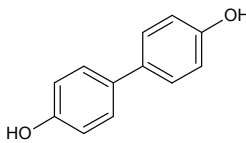
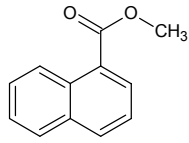
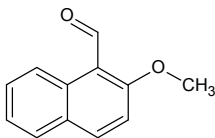
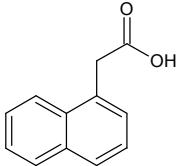
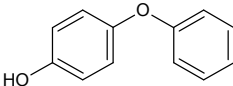
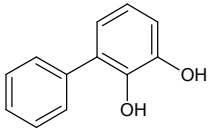
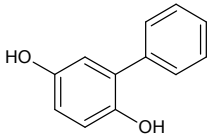
## 5.2 Retention Indices and $C_{12}H_{10}O_2$ Isomers

As introduced in Section 2.8.1, the retention index is a common starting point in confirming the identity of a compound. Here, several compounds of formula  $C_{12}H_{10}O_2$  were used to investigate the potential of using predicted retention indices to select or eliminate structure candidates during unknown identification. 29 different substances with the formula  $C_{12}H_{10}O_2$  were found in the NIST database, shown in Table 17. These compounds include a number of potentially environmentally-relevant compounds with different functional groups as well as several different substitution isomers. Of those 29, 14 are compounds with different functionality (numbers 1-14, Table 17) and the remaining 15 are substitutional isomers of the 14.

Of the 29 structures, 19 were commercially available to generate experimental retention index data. These are marked in Table 17. The compounds were obtained from ABCR (1, 24), Aldrich (11, 22), Alfa Aesar (2, 7, 21, 23, 25, 27, 29), Applichem (3), Fluka (28), Chembridge (5), Key Organics Ltd. (20), Merck (26), MP Biochemicals (13, 15) and Sigma (19).

**Table 17: 29  $C_{12}H_{10}O_2$  isomers retrieved from the NIST database, with CAS reference numbers and NIST spectrum numbers. Compounds marked with an asterisk were purchased to obtain chromatographic retention data.**

|                                                                                     |                                                                                     |                                                                                      |                                                                                       |
|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|
|  |  |  |  |
| 1) CAS RN: 574-19-6*<br>NIST Spec No: 211553                                        | 2) CAS RN: 713-68-8*<br>NIST Spec No: 73401                                         | 3) CAS RN: 830-81-9*<br>NIST Spec No: 229657                                         | 4) CAS RN: 56542-38-2<br>NIST Spec No: 150846                                         |
|  |  |  |  |
| 5) CAS RN: 1013-99-6<br>NIST Spec No: 243240                                        | 6) CAS RN: 19310-98-6<br>NIST Spec No: 242268                                       | 7) CAS RN: 2459-25-8*<br>NIST Spec No: 210940                                        | 8) CAS RN: 7028-41-3<br>NIST Spec No: 78313                                           |
|  |  |  |  |
| 9) CAS RN: 21421-61-4<br>NIST Spec No: 223345                                       | 10) CAS RN: 2197-57-1<br>NIST Spec No: 243287                                       | 11) CAS RN: 15971-29-6*<br>NIST Spec No: 236057                                      | 12) CAS RN: 63509-76-2<br>NIST Spec No: 187294                                        |

|                                                                                     |                                                                                    |                                                                                     |                                                                                      |
|-------------------------------------------------------------------------------------|------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|
|    |   |   |   |
| 13) CAS RN: 581-96-4<br>NIST Spec No: 134247                                        | 14) CAS RN: 5409-32-5<br>NIST Spec No: 215834                                      | 15) CAS RN: 1523-11-1*<br>NIST Spec No: 232580                                      | 16) CAS RN: 92-05-7<br>NIST Spec No: 221584                                          |
|    |   |   |   |
| 17) CAS RN: 17056-93-8<br>NIST Spec No: 211560                                      | 18) CAS RN: 18528-55-7<br>NIST Spec No: 211556                                     | 19) CAS RN: 711-79-5*<br>NIST Spec No: 7918                                         | 20) CAS RN: 2417-10-9*<br>NIST Spec No: 164567                                       |
|    |   |   |   |
| 21) CAS RN: 1806-29-7*<br>NIST Spec No: 155948                                      | 22) CAS RN: 612-76-0*<br>NIST Spec No: 221583                                      | 23) CAS RN: 92-88-6*<br>NIST Spec No: 113417                                        | 24) CAS RN: 2459-24-7*<br>NIST Spec No: 24851                                        |
|   |  |  |  |
| 25) CAS RN: 5392-12-1*<br>NIST Spec No: 108144                                      | 26) CAS RN: 86-87-3*<br>NIST Spec No: 292184                                       | 27) CAS RN: 831-82-3*<br>NIST Spec No: 236479                                       | 28) CAS RN: 1133-63-7*<br>NIST Spec No: 162517                                       |
|  |                                                                                    |                                                                                     |                                                                                      |
| 29) CAS RN: 1079-21-6*<br>NIST Spec No: 230993                                      |                                                                                    |                                                                                     |                                                                                      |

### 5.2.1 Retention Index Measurement and Prediction

Estimated Kovat's Retention Index (KRI) data for the 29  $C_{12}H_{10}O_2$  isomers were retrieved from the NIST database. The estimated values are calculated based on group contributions and compound class [59] and are reported with 50 % and 95 % confidence intervals.

For the purchased structures, experimental Kovat's and Lee Retention Indices were calculated according to Equation 8. GC-MS (Model 6890 N, detector MSD 5973, Agilent Technologies, Waldbronn, Germany) analysis was performed using a HP-5MS capillary column (30 m  $\times$  0.25 mm I.D., 0.25  $\mu$ m film, 5 % phenylmethylsiloxane, Agilent Technologies) and temperature program 70  $^{\circ}$ C (held for 4 min.), 3  $^{\circ}$ C/min. to 300  $^{\circ}$ C (held for 20 min.). The calibration mix  $C_8$ - $C_{36}$  from Supelco (46827U) was used to

calculate the Kovat's Retention Index (KRI), whereas the standard EPA-PAH mix from Ermsdorfer was used to record PAHs to calculate the Lee Retention Index (LRI).

Measured LRIs were used to generate an expected boiling point range according to Eckel and Kind [58], who correlated the boiling point and LRI of 370 compounds and found that 95 % of compounds had boiling points within the range (LRI -10) to (LRI + 50) °C. As the boiling point of compounds can be predicted for generated structures using EPISuite<sup>TM</sup>, this expected boiling point range can be used as an additional criterion to eliminate structures, with addition of the error associated in the boiling point prediction [58], given as 20.4 K average absolute error [81].

### 5.2.2 Results and Discussion of Retention Index Measurements

The measured Kovat's and Lee retention indices (KRI and LRI, respectively) for the purchased substances are given in Table 18, along with the predicted KRI taken from NIST and the predicted boiling point range from the LRI. The boiling point calculated by EPISuite for each compound is also included. Four of the 19 compounds purchased (Structures 2, 22, 23 and 26) could not be detected with the GC-MS method used in this study, as they were too polar.

The errors in the indices were calculated using the standards, i.e. KRI values were calculated for the PAHs and LRI values for the alkanes to determine the variation between the RI values. Determination of the indices over 25 measurements resulted in a LRI standard deviation between 0.032 for dodecane and 0.528 for octadecane (total of 7 compounds). For the KRI, the standard deviation ranged from 0.115 for naphthalene to 3.355 for fluorene (total of 10 compounds). The KRI errors are well within the error window of  $\pm 20$  units used, for example, to confirm compound identity within MODELKEY [57]. As the LRI values will be used in terms of candidate exclusion, the worst case error of 0.53 (LRI) and 3.4 (KRI) were taken as the error margin of the measurements.

Combining the errors associated with the LRI (0.53) and the boiling point prediction using EPISuite<sup>TM</sup> MPBPWin (20.4 K), leads to a cumulative error of  $\pm 21$  K for the LRI – BP correlation, such that the relationship becomes  $BP = (LRI-31) - (LRI+71)$  °C. This predicted range, along with the predicted BP from EPISuite<sup>TM</sup> is also shown in Table 18. All compounds in Table 18 are distinguishable from one another using the measured retention indices within the error margins of 0.53 (LRI) and 3.4 (KRI). Several pairs of compounds elute closely, however, (3 & 20, 1 & 19, 21 & 28, 11 & 27) and are within the  $\pm 20$  window for KRI (no such window for LRI is reported to the best of our knowledge). Of these pairs, some are distinguishable by the mass spectrum (3 & 20, 11 & 27), while others are not (e.g. 1 & 19, where differences in compound spectra are of the same magnitude as differences in replicate spectra within NIST), which would mean

these compounds would need to be separated on alternative chromatographic systems to ensure correct separation and identification of the isomers.

**Table 18: Measured Kovat's and Lee Retention Indices (KRI, LRI, respectively) for purchased compounds of formula  $C_{12}H_{10}O_2$ . The structure number corresponds with those in Table 17. Values are sorted from lowest to highest retention index. RIs calculated according to Equation 8. The worst case error margins (see text) are  $\pm 3.4$  (KRI) and  $\pm 0.53$  (LRI).**

| Structure Number | Measured KRI | Measured LRI | Predicted KRI (NIST) | Predicted BP range (LRI-31)-(LRI+71) °C | Predicted BP (EPISuite™) °C |
|------------------|--------------|--------------|----------------------|-----------------------------------------|-----------------------------|
| 20               | 1568.9       | 266.8        | 1664 $\pm$ 382       | 236.4 - 337.2                           | 315                         |
| 3                | 1575.8       | 267.9        | 1611 $\pm$ 201       | 237.5 - 338.3                           | 302                         |
| 15               | 1602.0       | 272.4        | 1611 $\pm$ 201       | 242.0 - 342.8                           | 302                         |
| 24               | 1625.5       | 275.6        | 1611 $\pm$ 201       | 245.2 - 346.0                           | 302                         |
| 7                | 1647.8       | 279.2        | 1611 $\pm$ 201       | 248.8 - 349.6                           | 302                         |
| 1                | 1690.5       | 286.2        | 1801 $\pm$ 382       | 255.8 - 356.6                           | 338                         |
| 19               | 1696.9       | 287.4        | 1801 $\pm$ 382       | 257.0 - 357.8                           | 338                         |
| 21               | 1729.2       | 291.9        | 1808 $\pm$ 301       | 261.5 - 362.3                           | 354                         |
| 28               | 1732.1       | 292.9        | 1808 $\pm$ 301       | 262.5 - 363.3                           | 354                         |
| 2                | 1775.6       | 298.4        | 1664 $\pm$ 382       | 268.0 - 368.8                           | 315                         |
| 25               | 1814.2       | 305.2        | 1722 $\pm$ 382       | 274.8 - 375.6                           | 320                         |
| 27               | 1830.7       | 307.5        | 1664 $\pm$ 382       | 277.1 - 377.9                           | 315                         |
| 11               | 1836.5       | 308.6        | 1722 $\pm$ 382       | 278.2 - 379.0                           | 320                         |
| 29               | 1930.1       | 323.8        | 1808 $\pm$ 301       | 293.4 - 394.2                           | 354                         |
| 5                | 2095.8       | 350.9        | 1597 $\pm$ 382       | 320.5 - 421.3                           | 302                         |

The range of the measured values for all isomers was 520 for KRI and 83 for the LRI. The prediction error margins are of similar magnitude, between 201 and 382 for KRI and 102 for the LRI. This confirms previous conclusions that general prediction of RIs is still insufficient to identify single structures but may be useful as a filter [43]. The results also confirm the weakness of prediction based on group contributions, discussed by Stein et al [59], where all isomers have the same predicted values despite large discrepancies in reality (e.g. Structures 21 and 29). This is one of the largest sources of error in the KRI prediction based on group contribution [59].

Furthermore, despite the very large errors associated with the predicted KRI reported in NIST, the measured KRI of one compound (Structure 5) is well outside the 95 % confidence interval of the predicted value (2095.8 compared with  $1597+382=1979$ ). The same structure is also outside the error margins used for the predicted LRI. This shows the lack of applicability of these predicted values outside the compound domain used to form the predictions, even with such large error margins. As such, these criteria may be of limited value in structure exclusion in some cases. This is shown for example in Section 6.3 for phthalimide. More accurate prediction algorithms for smaller subsets of compounds exist (e.g. [43] and references within) and these could be used on a case-by-case basis to improve candidate selection using retention prediction if sufficient information about the compound is available (e.g. compound class, functional groups). This information is, however, often not available during unknown identification,



especially during effect-directed analysis (EDA) where a lack of sample often prevents additional analyses that may yield functional group information. Furthermore, compounds with different combinations of functional groups can often occur together when using structure generation techniques for unknown identification, unless sufficient substructure classifiers are available to cover all functional groups present in the molecules.

### 5.3 Match Value Comparison with $C_{12}H_{10}O_2$ Isomers

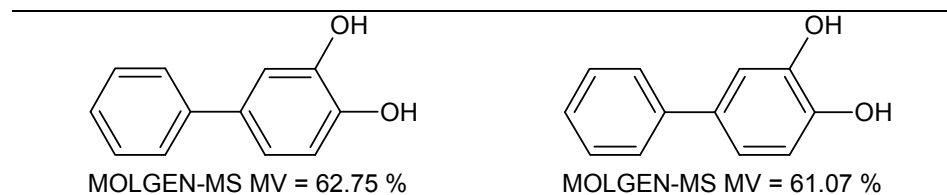
A match value comparison was conducted for the  $C_{12}H_{10}O_2$  isomers, following the method in Section 4. As the simplest program settings were identified in Section 4 as the most effective in terms of candidate separation, only these have been reported in Table 19. The calculations were performed with MOLGEN-MSF [48] using the default reaction settings, as well as Mass Frontier [44] and ACD Fragmenter [30], both with 3 step fragmentation, on each of the 29 molecules and the 29 spectra. More details about the settings used are given in Section 4.

In contrast with the previous study, which indicated that the results of MSF and Mass Frontier were generally comparable and ACD generally slightly worse, for this example Mass Frontier was clearly the better. This is shown using the relative ranking position (RRP). Mass Frontier had a much lower average relative ranking position (0.216) compared with ACD (0.303) and MSF (0.366). Multiplying the relative ranking position (RRP) by 29 gives the average place or rank of the correct structure relative to the other 28 structures based on mass spectral match only. The correct molecule is on average in 6<sup>th</sup>, 9<sup>th</sup> or 11<sup>th</sup> place for Mass Frontier, ACD and MSF calculations, respectively. When it comes to identifying the ‘correct’ structure in first place based on match value, the three programs were very similar, with each matching three or four of the structures correctly (except MSF aromatic, see below). This is in stark comparison with the spectral match within the NIST database, which identifies correct spectrum using the library search in all 29 cases with a very high probability (lowest value 80.8 %, see Table A3, Appendix 1). This confirms the conclusions in Section 4, that spectral match based on predicted fragments alone is insufficient to identify candidate structures, even from compounds of the same formula but with very different mass spectra.

**Table 19: Match Value (MV) and Relative Ranking Position (RRP) calculation for the 29 C<sub>12</sub>H<sub>10</sub>O<sub>2</sub> isomers and their respective spectra, using the programs MOLGEN-MSF (MSF), Mass Frontier (MF) and ACD. The structure numbers correspond with Table 17. MV (0 = poor match, 100 % = perfect match) and RRP (0 = correct molecule is best match, 1 = correct molecule is worst match) calculations are defined in Equation 1 and Equation 7, respectively.**

| <i>Structure Number</i> | <i>MSF MV (%)</i> | <i>MSF RRP</i> | <i>MSF aromatic MV (%)</i> | <i>MSF aromatic RRP</i> | <i>MF 3 step MV (%)</i> | <i>MF 3 step RRP</i> | <i>ACD 3 step MV (%)</i> | <i>ACD 3 step RRP</i> |
|-------------------------|-------------------|----------------|----------------------------|-------------------------|-------------------------|----------------------|--------------------------|-----------------------|
| 1                       | 63.4              | 0.125          | 62.3                       | 0.054                   | 62.5                    | 0.054                | 65.1                     | 0.214                 |
| 2                       | 63.1              | 0.625          | 56.6                       | 0.839                   | 60.3                    | 0.125                | 65.9                     | 0.714                 |
| 3                       | 1.8               | 0.536          | 1.0                        | 0.196                   | 70.2                    | 0.089                | 75.0                     | 0.429                 |
| 4                       | 34.2              | 0.393          | 31.9                       | 0.000                   | 29.7                    | 0.036                | 88.9                     | 0.000                 |
| 5                       | 31.7              | 0.768          | 31.2                       | 0.143                   | 77.6                    | 0.000                | 87.5                     | 0.107                 |
| 6                       | 23.7              | 0.107          | 21.2                       | 0.125                   | 49.7                    | 0.054                | 80.7                     | 0.071                 |
| 7                       | 82.4              | 0.036          | 30.4                       | 0.018                   | 79.2                    | 0.018                | 94.5                     | 0.000                 |
| 8                       | 92.6              | 0.464          | 90.5                       | 0.714                   | 90.5                    | 0.679                | 91.0                     | 0.821                 |
| 9                       | 48.4              | 0.036          | 3.4                        | 0.036                   | 14.1                    | 0.000                | 78.5                     | 0.071                 |
| 10                      | 65.1              | 0.661          | 64.2                       | 0.661                   | 73.6                    | 0.107                | 95.3                     | 0.000                 |
| 11                      | 67.1              | 0.268          | 29.7                       | 0.232                   | 29.8                    | 0.125                | 68.8                     | 0.071                 |
| 12                      | 72.8              | 0.000          | 53.4                       | 0.107                   | 57.5                    | 0.089                | 77.9                     | 0.107                 |
| 13                      | 86.6              | 0.107          | 70.8                       | 0.089                   | 71.0                    | 0.018                | 96.8                     | 0.000                 |
| 14                      | 25.9              | 0.500          | 21.3                       | 0.036                   | 37.6                    | 0.071                | 77.2                     | 0.143                 |
| 15                      | 1.6               | 0.554          | 0.9                        | 0.161                   | 80.1                    | 0.089                | 80.6                     | 0.571                 |
| 16                      | 62.7              | 0.018          | 60.5                       | 0.500                   | 60.5                    | 0.661                | 64.5                     | 0.464                 |
| 17                      | 57.4              | 0.143          | 57.2                       | 0.054                   | 57.2                    | 0.089                | 58.9                     | 0.250                 |
| 18                      | 82.6              | 0.143          | 82.4                       | 0.054                   | 82.4                    | 0.089                | 83.5                     | 0.357                 |
| 19                      | 68.7              | 0.179          | 68.3                       | 0.054                   | 69.5                    | 0.054                | 71.8                     | 0.214                 |
| 20                      | 64.7              | 0.357          | 53.7                       | 0.839                   | 64.1                    | 0.000                | 70.6                     | 0.429                 |
| 21                      | 39.4              | 1.000          | 36.5                       | 0.911                   | 37.2                    | 0.875                | 42.0                     | 0.964                 |
| 22                      | 74.0              | 0.732          | 72.9                       | 0.357                   | 73.0                    | 0.375                | 77.2                     | 0.607                 |
| 23                      | 72.4              | 0.964          | 71.9                       | 0.500                   | 71.9                    | 0.679                | 75.4                     | 0.786                 |
| 24                      | 68.2              | 0.000          | 37.8                       | 0.018                   | 57.5                    | 0.018                | 73.5                     | 0.036                 |
| 25                      | 33.8              | 0.821          | 22.0                       | 0.696                   | 23.2                    | 0.839                | 54.7                     | 0.107                 |
| 26                      | 86.1              | 0.000          | 68.1                       | 0.089                   | 68.2                    | 0.054                | 86.4                     | 0.036                 |
| 27                      | 80.2              | 0.446          | 69.7                       | 0.964                   | 77.4                    | 0.125                | 84.6                     | 0.214                 |
| 28                      | 81.7              | 0.196          | 75.4                       | 0.446                   | 75.5                    | 0.375                | 79.5                     | 0.393                 |
| 29                      | 79.0              | 0.446          | 69.0                       | 0.411                   | 69.0                    | 0.482                | 71.6                     | 0.607                 |
| <b>Avg</b>              | <b>59.0</b>       | <b>0.366</b>   | <b>48.8</b>                | <b>0.321</b>            | <b>61.0</b>             | <b>0.216</b>         | <b>76.5</b>              | <b>0.303</b>          |
| <b>RRP*29</b>           |                   | <b>10.6</b>    |                            | <b>9.3</b>              |                         | <b>6.3</b>           |                          | <b>8.8</b>            |
| <b>Correct Matches</b>  |                   | <b>3</b>       |                            | <b>1</b>                |                         | <b>3</b>             |                          | <b>4</b>              |

**Table 20: The two ‘correct’ structures for Structure 16 using MOLGEN-MS.**



The results given in Table A3 for Structures 16, 21, 22, 23, 28 and 29 show that aromatic doublets are calculated during structure generation involving the biphenyl moiety (22 structures generated compared with 12 unique structures). An example for Structure 16 is given in Table 20. Although there is an automatic filtering of aromatic doublets incorporated within MOLGEN-MS, the biphenyl moiety confuses this, being two separate aromatic rings. Table 20 and Table A3 also show that the different doublets actually have different match values, a fact that can affect the ranking of the candidate structures by match value. Unlike MOLGEN-MS, when calculating with MSF the aromatic bonds can be defined explicitly prior to fragment calculation. As a result, the MSF calculation was repeated with explicit definition of the aromatic bonds within the molecules. This is also reported in Table 19 (column 'MSF aromatic'). Inclusion of the explicit aromatic bonds reduced the match values for MSF but improved the relative ranking position slightly, at the expense of only identifying one compound correctly as opposed to three or four for the other calculations. All substitutional isomers had the same match value (data not shown). The low match values and low selectivity with MSF can be attributed to the lack of defined reactions with aromatic bonds. As it is possible to define these reactions, the inclusion of such reactions would probably result in a significant increase in the match value and constitute a significant improvement to MOLGEN-MSF.

The poor performance of MSF (and hence MOLGEN-MS) for some isomers can be seen quickly for example in Structures 3 and 15 (1- and 2-naphthyl acetate), 5 and 6, where the match values are significantly less than for Mass Frontier or ACD. Structures 4 and 14 have relatively low values for MSF and Mass Frontier, while 25 is low for all programs. For Structures 3 and 15, this was due to one main fragment that was missing,  $m/z = 144$ , which was the base peak of both spectra. Inclusion of this reaction in MOLGEN-MS and MOLGEN-MSF would likewise improve the results for these compounds considerably.

Despite the results here showing that MOLGEN-MSF was not the best performer for this example, the overall results in Section 4 indicate that MOLGEN-MSF is generally on par with Mass Frontier. As this calculation is relatively fast and inbuilt in MOLGEN-MS, the MOLGEN-MS match values were used further in the overall work flow, rather than Mass Frontier. An option for final candidate selection would be to perform additional calculations using Mass Frontier to complement the MOLGEN-MS match values, once less likely candidates have been excluded, to speed up the Mass Frontier calculation times.

## 5.4 *Steric Energy as an Exclusion Criterion*

As mentioned briefly in Section 2.8.3, the steric energy of a molecule has been used to exclude energetically unfavourable molecules from consideration (see [63], p. 47). This was tested further here, using two programs to calculate the steric energy based on a heat of formation calculation. This was evaluated on known molecules to determine possible limits to apply for structure elimination (presented here), while the use for structure elimination is covered in Section 5.6.

### 5.4.1 *Calculation of Steric Energy*

Two independent programs, ChemBio3D from CambridgeSoft [66] and MOLGEN-QSPR [67], were used to calculate the steric energy. ChemBio3D is based on the MM2 algorithm and calculates an enthalpy of formation (kcal/mol), while MOLGEN-QSPR uses force-field mechanics to calculate the steric energy [65] (also kcal/mol). A batch function was programmed for ChemBio3D using python [82] and ChemScript [66], which exported the SMILES code and energy. This batch function requires a SDF file with 2D coordinates already defined. In contrast, MOLGEN-QSPR works with all SDF files (with and without 2D placement). Following SDF import and the explicit addition of hydrogen atoms, 3D optimisation was undertaken (using 5 iterations) and molecular descriptor representing the steric energy was calculated, with the results exported as structure number and energy. As the calculations started from random placements, resulting in occasional anomalies where no chemically relevant energy minimum was found, these calculations were repeated three times, with the minimum of the three runs taken as the final result.

In order to have a baseline interpretation of the energy values of both programs, energy values were calculated for a random set of 1000 molecules (the same as used here in Section 4 and previously [7]) and used to generate quantiles. These quantiles, defined in Section 4, were used to assess the energy distribution of molecules known to exist in reality. Briefly,  $p$  refers to the probability and  $q_p$  the quantile associated with that probability. To include all structures likely to exist (in terms of energy) with a probability of 90 %, in this case  $q_{90}$  should be considered.

### 5.4.2 *Steric Energy Distribution*

The quantiles for the steric energy calculation are shown in Figure 24 and given in Table 21. 90 % of the molecules had energy below 231.24 kcal/mol for ChemBio3D and 429.0 kcal/mol for MOLGEN-QSPR, which could be used as an inclusion criterion following structure generation.

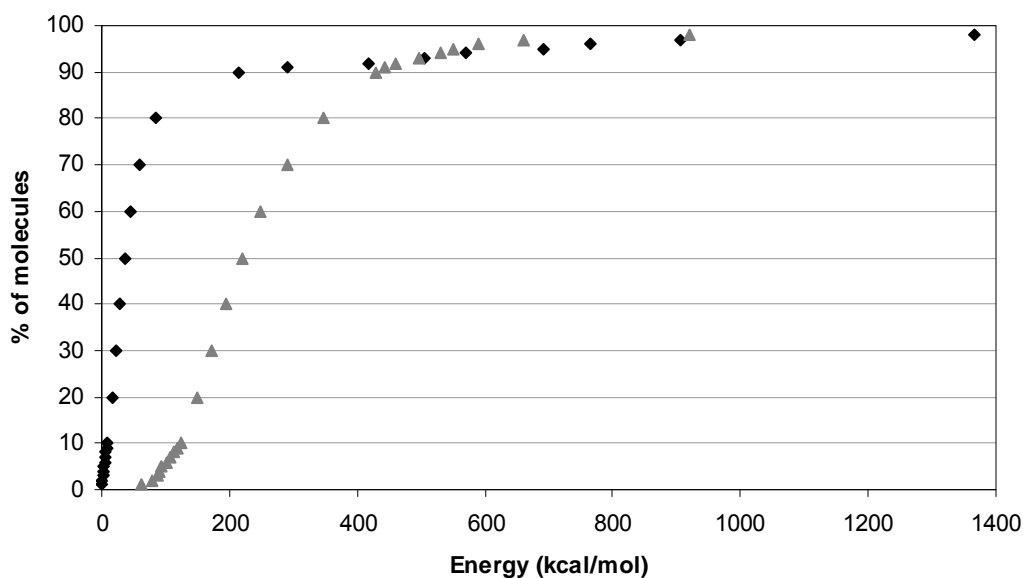


Figure 24: Steric energy distribution of 1000 randomly selected molecules, in kcal/mol. Data is given in Table 21. Black diamonds represent ChemBio3D results using the batch function, grey triangles MOLGEN-QSPR.

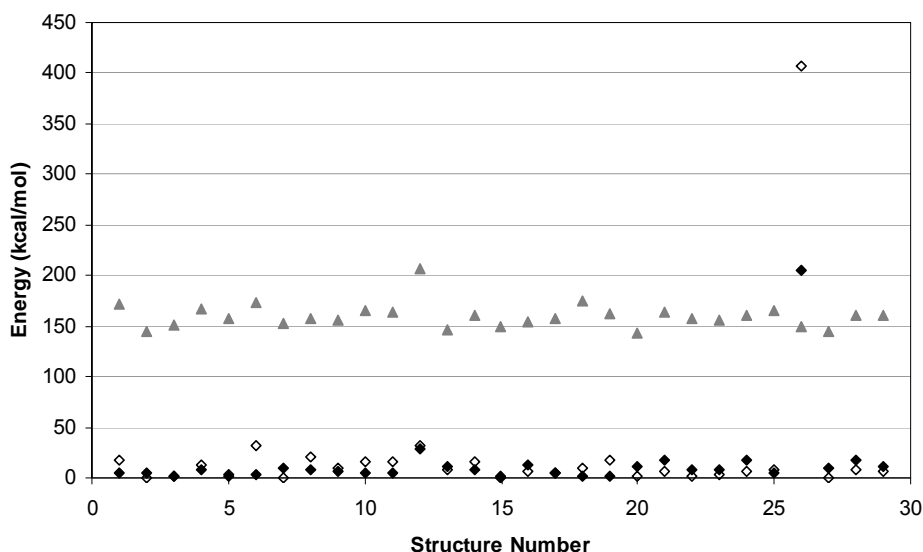
Table 21: Energy quantiles calculated from MM2 energy values generated for 1000 randomly-selected compounds using ChemBio3D [66] and MOLGEN-QSPR [67].

| $p$  | $q_p$<br>ChemBio3D | $q_p$<br>MOLGEN-QSPR<br>(3 × 5 iterations) |
|------|--------------------|--------------------------------------------|
| 0.01 | 0.39               | 61.6                                       |
| 0.02 | 1.38               | 77.7                                       |
| 0.03 | 2.09               | 86.5                                       |
| 0.04 | 3.17               | 90.3                                       |
| 0.05 | 3.64               | 94.3                                       |
| 0.06 | 4.68               | 102.4                                      |
| 0.07 | 5.91               | 107.2                                      |
| 0.08 | 6.43               | 112.8                                      |
| 0.09 | 7.65               | 117.9                                      |
| 0.1  | 8.19               | 123.5                                      |
| 0.2  | 15.69              | 150.1                                      |
| 0.3  | 22.46              | 173.4                                      |
| 0.4  | 28.88              | 195.5                                      |
| 0.5  | 36.96              | 219.3                                      |
| 0.6  | 45.89              | 249.0                                      |
| 0.7  | 58.46              | 289.4                                      |
| 0.8  | 83.94              | 346.3                                      |
| 0.9  | 213.24             | 429.0                                      |
| 0.91 | 290.41             | 444.5                                      |
| 0.92 | 416.85             | 460.7                                      |
| 0.93 | 506.11             | 497.7                                      |
| 0.94 | 570.96             | 530.7                                      |
| 0.95 | 692.74             | 550.7                                      |
| 0.96 | 764.13             | 589.7                                      |
| 0.97 | 905.12             | 661.8                                      |
| 0.98 | 1367.26            | 919.3                                      |
| 0.99 | 1399.29            | 12493000                                   |

The steep increase in the energy for the last 10 % of molecules was not entirely expected, given that the 1000 molecules are known to exist with a spectrum measured in the NIST database. The energies of the 10 % excluded cover a much wider range than the 90 % included, especially for ChemBio3D.

The quantiles for MOLGEN-QSPR, especially the 99 % quantile, can be improved using further program repetitions and iterations. The values presented here, calculated from three times 5 iterations, are consistent with the method used to generate the energy values in Sections 5.6 and 6.3.

The energy values calculated using the two programs for the  $C_{12}H_{10}O_2$  isomers are shown in Figure 25. As a form of quality control of the ChemBio3D batch processing, these 29 values were also calculated manually via the user interface, which does not require pre-definition of the 2D coordinates.



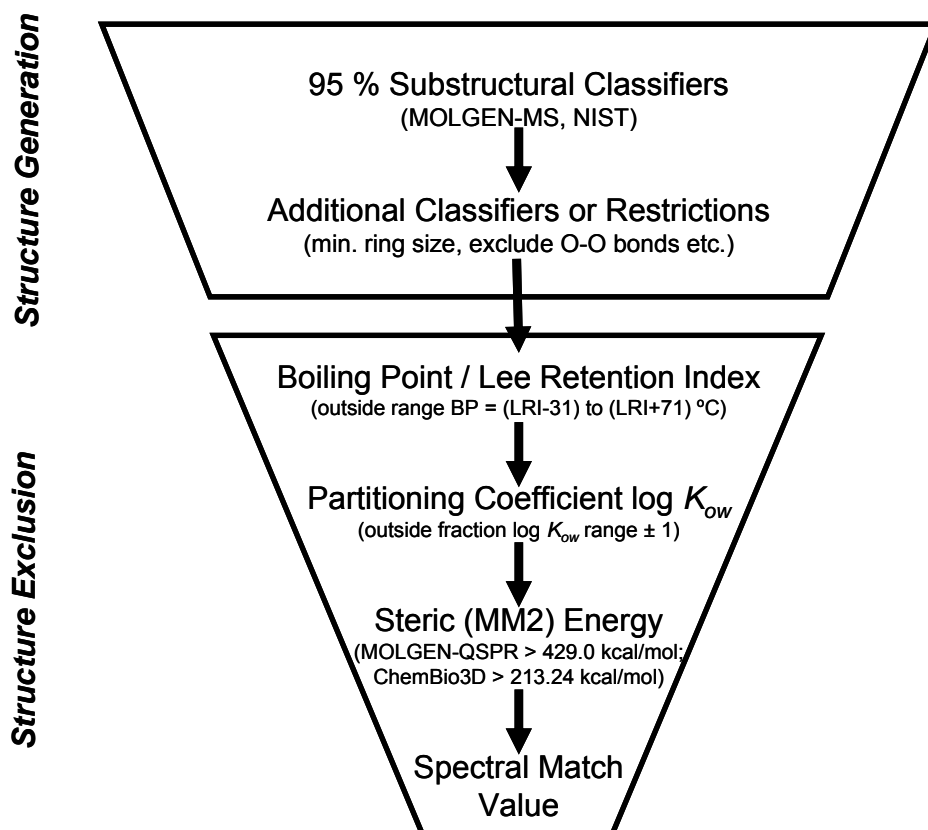
**Figure 25:** Steric energy of the 29  $C_{12}H_{10}O_2$  isomers. Black diamonds: MM2 optimisation via the ChemBio3D user interface, hollow diamonds: MM2 optimisation via ChemBio3D batch processing and grey triangles: Energy optimisation (minimum of three times 5 iterations) using MOLGEN-QSPR.

The values shown in Figure 25 are all well below the 90 % quantiles in Table 21, except Structure 26 for the ChemBio3D calculation. Otherwise the programs show a relatively similar distribution of energies, although the energy values are very different, in accordance with Figure 24. The user interface and batch values for ChemBio3D show slight differences (larger for Structure 26) as the calculations start from a different 2D placement. The batch runs with ChemBio3D were almost identical, as molecules were optimised from the same 2D placement. As MOLGEN-QSPR starts from a random placement after each file import, the results varied slightly for each run. The average of the three runs is displayed.

Figure 25 confirms that the choice of the 90 % cut-off for energy is appropriate. Even the clear outlier, Structure 26, is within the 90 % energy limit of 213.24 for ChemBio3D, while the MOLGEN-QSPR values are well below the  $q_{90}$  of 429.0. Using the  $q_{80}$  would exclude Structure 26 for ChemBio3D, while a less conservative value, e.g.  $q_{95}$ , would result in the inclusion of unlikely structures (see for example Table 27, page 102).

### 5.5 Method for Structure Generation and Progressive Elimination

The knowledge gained from the study based on unknown spectra (Section 3) and the two sections above (5.2, 5.3) can now be compiled to form an overall method for the processing of GC-EI-MS spectra with the aim of identifying the compound measured without relying on a database spectrum search. This is shown schematically in Figure 26.



**Figure 26:** The strategy for the database-independent identification of unknown compounds based on GC-EI-MS spectra presented in this section, including example exclusion criteria.

The following information flow was applied. Two cases were considered for structure generation. Firstly, substructure classifiers were selected according to the default settings of 95 % precision (e.g. Section 3, [40]). As shown in Section 3, however, these default settings were not always sufficient to reduce the number of candidates generated, such that the program limits were reached. Thus, a second case was introduced, either including additional classifiers (e.g. those less than 95 % precision) or, in the case where

no classifiers with reasonable precision were available, alternative restrictions to limit the number of structures generated. This case is subsequently termed ‘modified classifier selection’. Following classifier selection, structure generation and calculation of the match values was conducted using MOLGEN-MS. Scripts developed as part of this work were then used to calculate the boiling point and log  $K_{ow}$  values via EPISuite<sup>TM</sup> and the steric energy using ChemBio3D and MOLGEN-QSPR, for all candidates generated. Finally, the script summarised this calculated data into a table and plots for application of the exclusion criteria.

The exclusion criteria were applied in the order shown in Figure 26. Firstly, the BP-LRI criterion was used for those compounds where the measured LRI was available, excluding those compounds with estimated boiling points outside the range (LRI-31) – (LRI+71) °C. Following this, the partitioning coefficient was used to eliminate compounds outside the estimated log  $K_{ow} \pm 1$  (for EDA fractions this becomes the estimated fraction log  $K_{ow}$  range  $\pm 1$ ). The error margin of  $\pm 1$  is the range reported to include 96.5 % of all predicted values from EPISuite<sup>TM</sup> Kowwin. The steric energy criterion was then used to exclude all compounds with ChemBio3D energies above 213.24 kcal/mol and MOLGEN-QSPR energies above 429.0 kcal/mol. Finally, the spectral match value was used last, as this is a very example-specific criterion and is best applied following the other criteria rather than before.

### 5.5.1 Automatic Data Processing

The MSP file, saved from NIST or AMDIS for each structure, was converted into CSV format for import into MOLGEN-MS using a Matlab [70] script. The NIST substructural information was also processed using a script with optional user input to include additional substructures. The output was saved such that it is automatically uploaded into the classifier selection stage of MOLGEN-MS. The scripts are listed in Appendix 2.

Spectral processing using MOLGEN-MS is demonstrated pictorially in Figure 27 and is described in greater detail in Section 3.1. This method was followed here, with the exception that the NIST substructural information is now included automatically. Briefly, the substructural classifier information from NIST and MOLGEN-MS was used to limit the elements present and/or absent based on the spectrum to enable calculation of the molecular formulae. These were selected based on the deviation of the calculated isotope pattern from the experimental pattern. The ring and double bond count (RDB) was incorporated into formula selection using a Matlab script implementing Equation 11 (Section 3.1). Following calculation and selection of the desired formula, MOLGEN-MS automatically filters the substructural classifiers from both programs to ensure compatibility with the formula. The substructural classifier probability was either left at the default 95 %, or additional classifiers of lower precision were considered, depending



on the case (see above). The structures were then generated and fragmented within MOLGEN-MS to assigned match values according to Equation 1.

Once the spectral interpretation with MOLGEN-MS was complete, the input and output files were listed in a summary file. A Matlab script read this summary file and generated the additional exclusion criteria as described above, to assist in selection of the most likely matching structures. The script coordinates file conversions (e.g. from SDF [74] into SMILES) and inputs into the EPI Suite<sup>TM</sup> [60] programs MPBPWin (boiling point and melting point calculation) and Kowwin (octanol-water partitioning coefficient or log  $K_{ow}$  calculation) and saves the output in the current directory, including a tab-separated table and figures containing all estimated data.

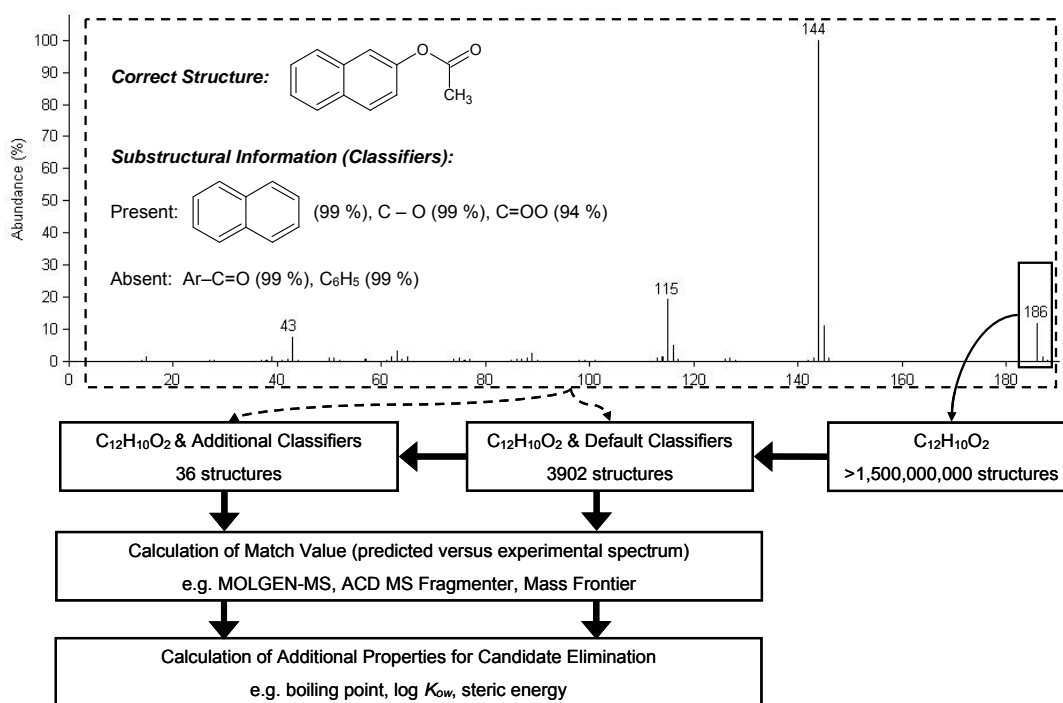


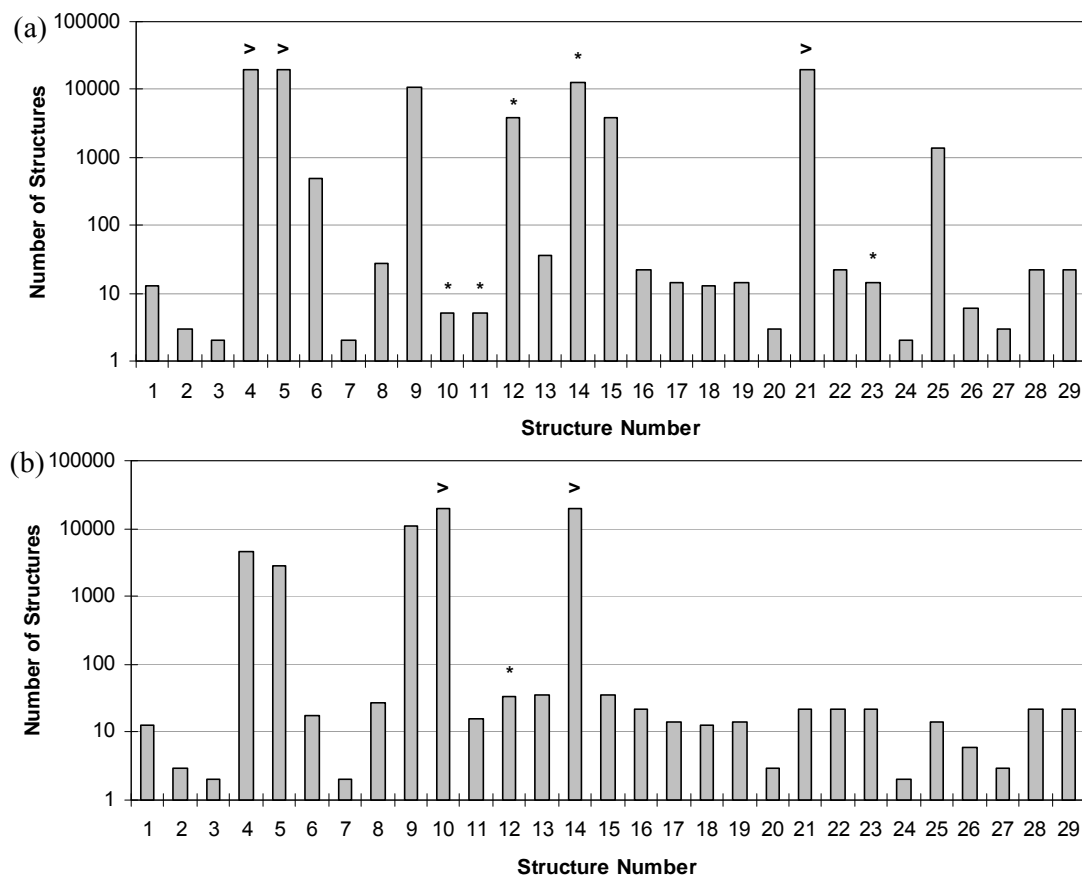
Figure 27: The information flow used to identify compounds based on GC-EI-MS and structure generation, showing the two cases used here, ‘default classifiers’ (substructural information from NIST and MOLGEN-MS with 95 % or greater precision) and ‘modified classifiers’ (additional substructural information with lower than 95 % precision, additional restrictions etc.). Spectrum taken from NIST [17].

## 5.6 Structure Generation Results for $C_{12}H_{10}O_2$ Isomers

### 5.6.1 Classifier Assessment

As mentioned above, two cases were considered for the inclusion of substructure information (classifiers). As well as using the default 95 % precision, a modified classifier selection (i.e. including those associated with a probability of less than 95 % or removing clashing classifiers) was considered in this example to assess the benefits of restricting structure numbers versus lowering the probability. The general scheme applied

is shown in Figure 27. The results are included in Appendix 1, Table A5, which shows the substructural classifiers used, the number of structures calculated, calculation time, match value (MV) range and the MV and place of the correct structure. Where multiple runs were needed, the modified classifier selection is also included. The results using 95 % classifier selection (i.e. no additional user input) and modified classifier selection (e.g. additional classifiers, removing classifiers preventing generation of the correct molecule) are presented in Figure 28(a) and (b), respectively.



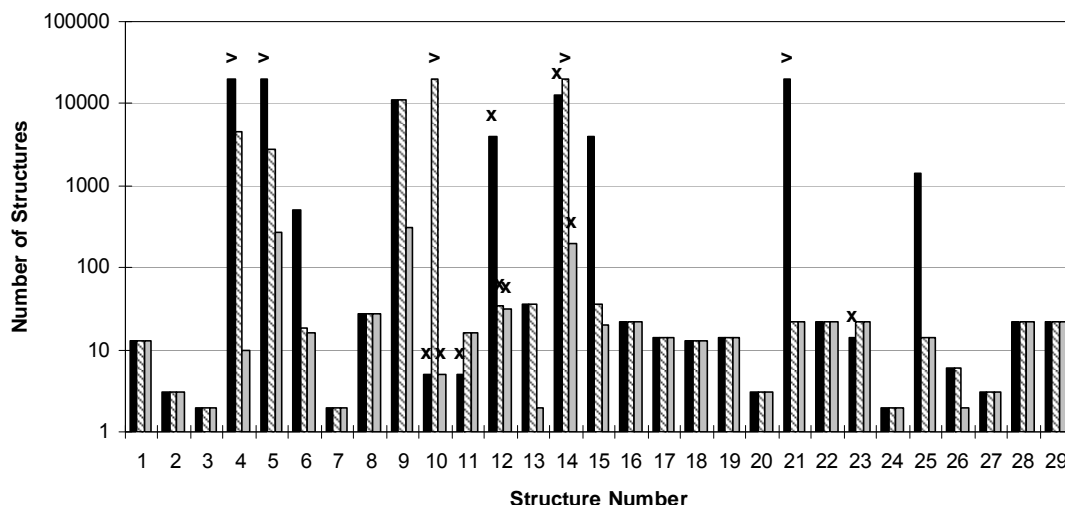
**Figure 28: (a) Number of structures generated with 95 % classifiers (b) Number of structures generated with modified classifiers (see Table A3, Appendix 1). \* indicates that the correct structure is not present, while > indicates that the calculation limit of 20,000 was reached.**

The modified classifier selection improves the results significantly, with all but 5 of the 29 structures down to less than 100 possible candidates, compared with 9 based on the 95 % classifiers. Furthermore, only 1 compound no longer has the correct structure present in the structure space (Structure 12), compared with 5 for the 95 % classifiers.

The use of the additional classifiers reduces the possible structures for 19 of 29 compounds to a group of substitutional isomers (e.g. compounds 1-3, 7-8, 11, 15-25 and 27-29). However for at least five of the remaining 10 compounds, several thousand structures are still possible, necessitating the use of the exclusion criteria to identify the more likely candidates.

### 5.6.2 Combining Substructural Classifiers and Exclusion Criteria

The structures from the two cases above (default and modified classifier selection) were processed using the exclusion criteria described above. The results of these exclusion steps are shown in Figure 29 for the modified classifier selection.



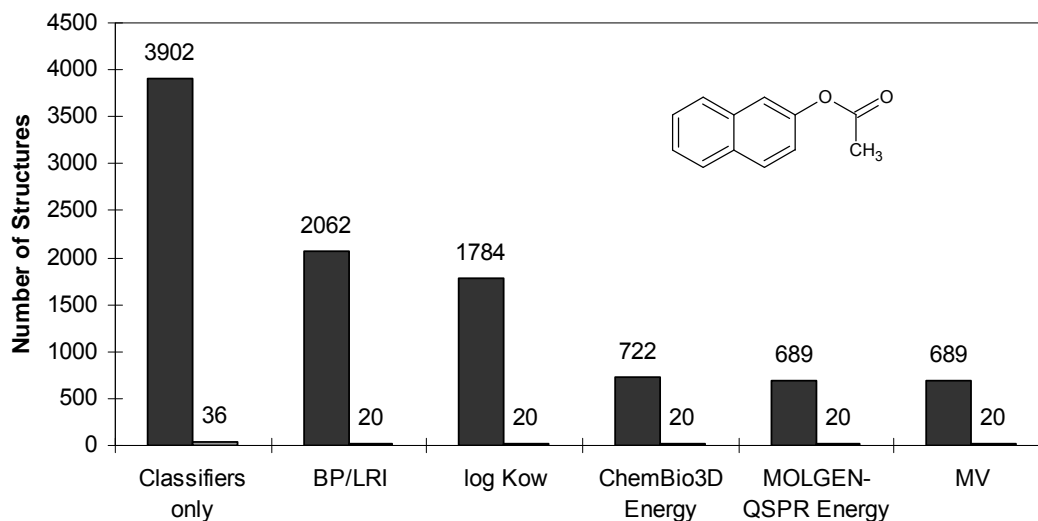
**Figure 29:** Number of structures remaining in consideration following generation using 95 % classifiers (black columns), additional classifiers (striped columns) and the additional classifiers followed by stepwise application of the exclusion criteria described in Figure 26 (grey columns). Crosses indicate that the correct structure is absent, while > indicates that the calculation limit of 20,000 was exceeded. The filtering step was performed on the 95 % classifier runs for Structures 10 and 14, as removal of incorrect classifiers led to generation beyond the 20,000 limit. As a result, the correct structure is absent.

This plot shows clearly that the exclusion criteria are effective in eliminating at least one order of magnitude of structures in several cases (note the logarithmic scale), including Structures 4, 5, 9, 13 and 14. In most cases where the exclusion criteria had no effect, the classifiers had already reduced the structures down to substitution isomers (see above).

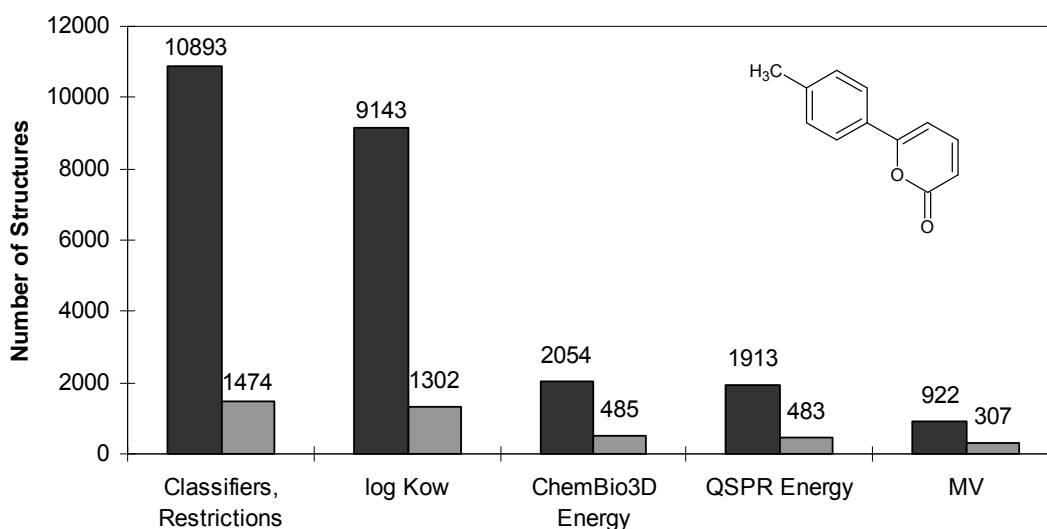
The relative influence of the classifiers and the different filtering criteria is shown in Figure 30 for Structure 15. Although the filtering criteria are effective in reducing the number of structures from 3,902 to 689, the use of the additional classifier is much more effective in reducing structure numbers in this case (3,902 to 36, reduced further to 20 after exclusion).

In Figure 31, the influence of a different strategy to reduce structure numbers is shown, using Structure 9. 10,893 structures were generated with classifiers >95 % precision. As no additional classifiers were available, only structures with 5 or more atoms in a ring (not the default 3 or more) were generated. This strategy is not always optimal, as this can lead to elimination of some possible compounds (e.g. epoxides, cyclo-propyl groups), however the drastic reduction in the number of structures generated is demonstrated clearly and such a reduction is necessary in examples with insufficient classifiers to come up with candidates as a first step. The likelihood of eliminating the

‘correct’ structure as a result is generally minimal but case-dependant. In this case, the exclusion criteria reduce the number of candidates further, such that 307 structures remain at the final step, compared with 922 without the additional restrictions. The correct structure remained in consideration in both cases. In some examples (e.g. Structure 15), the classifiers themselves restrict the generation sufficiently and the use of this additional criterion is not necessary.



**Figure 30:** The influence of additional classifiers and exclusion criteria on the number of structures for consideration, based on Structure 15 (insert). The black columns indicate the number of structures remaining following generation using the default classifiers (95 % precision), followed by stepwise application of the exclusion criteria. The grey columns indicate the same steps using additional classifiers (94 % precision in this case).



**Figure 31:** The number of structures generated for Structure 9 (see insert) using >95 % precision classifiers (black columns) followed by stepwise application of the filtering criteria and using the ‘cycles containing 5 or more atoms’ criterion prior to filtering (grey columns). No LRI data was available for Structure 9.

## 5.7 Discussion of Structure Elimination Strategies

The results presented in the section above show that the exclusion criteria following structure generation using MOLGEN-MS are successful in reducing the number of candidates as a result of structure generation by several orders of magnitude in many cases. Some issues with each of the criteria are discussed briefly below.

### 5.7.1 Substructure Classifier Use

The use of substructural classifiers is clearly shown in Figure 30 as the key to success in limiting the number of structures generated for a given molecular formula. Although the exclusion criteria can reduce the number of structures dramatically following generation, exclusion by using the correct substructural information prior to generation is still more effective. The choice of which classifiers to use in some examples can be ‘trial and error’ and may require the consideration of several different structure generation runs to avoid exclusion of candidates for an unknown spectrum.

Problems can occur in some cases when combining the NIST and Varmuza classifiers in MOLGEN-MS. When many similar classifiers are identified for a spectrum, it is best to manually review the list prior to structure generation to ensure some classifiers do not restrict the structure generation in undesired ways. For example, the presence of  $\text{Ar-C=OCH}_2$  instead of the  $\text{C=OCH}_2$  classifier in combination with the naphthalene substructure excludes the generation of 2-naphthalene acetic acid for Structure 13 (see Table A3) due to the arrangement of the double bonds. Similarly, the presence of the classifier ‘ph-O’ (interpreted as  $\text{C}_6\text{H}_5\text{-O}$ ) restricts the generated structures correctly for Structures 2, 20 and 27 but excludes the correct structures for Structures 22 and 23.

A feature of the example used here is that biphenyl and naphthalene often both appear with probability 95 %, leaving the choice of classifier up to the user (only one is possible for the formula  $\text{C}_{12}\text{H}_{10}\text{O}_2$ ). As this example is conducted with known spectra, it was possible to choose the classifiers to get the right answer – this option would need to be covered using multiple runs with different classifier combinations for real unknown determination, as it can be difficult to determine which of the clashing classifiers to use. Recording a UV spectrum of the sample during liquid chromatographic (LC) measurements (e.g. during fractionation in the case of EDA studies) could provide supplementary information to select which of the classifiers applies in this case.

Although the problem of aromatic doublets can be solved using the ‘remove aromatic doublets’ option in the MOLGEN programs, as the different structures have different match values, this filtering may affect the ranking of the molecules. Instead using the ‘assign aromatic bonds’ in MOLGEN-MSF removes the problem of different Kekule structures representing the same aromatic compound, by generating the same match values for both. The inclusion of fragmentations involving aromatic bonds will be an

important extension to MOLGEN-MSF to address this problem. This will not be a problem for structures without the biphenyl moiety present.

The results shown in Figure 29 reveal that the correct structures are missing for Structures 10, 12 and 14. Structures 10 and 14 are quinones, whereas Structure 12 has a bridging CH<sub>2</sub> group over a naphthalene ring. Although Structure 12 is relatively unusual, Structures 10 and 14 are not unusual compounds but removal of the clashing classifiers restricting generation of the correct structure resulted in over 20,000 molecules. The development of substructural classifiers specifically for quinones and similar compounds would improve the ability to identify these structures correctly. Improvement to the classifiers overall would also improve the applicability of this method to spectra further outside the domain of the NIST database.

#### 5.7.2 Boiling Point/Lee Retention Index Restriction

The retention index data presented in this study shows that the experimentally measured values are very useful and accurate in identifying the correct isomers. The prediction of KRI, included for example in the NIST database, is still too inaccurate to exclude a significant number of candidates (see Table 18) and does not cover all experimental values despite the large error margins. Although the same also applies for the LRI, where the boiling point-LRI relationship developed by Eckel and Kind [58] results in a very large ‘inclusion window’, this has proved to be a useful exclusion criterion in all examples presented here where the LRI was available. Care still needs to be taken when using this relationship, however, as shown for Structure 5 and further below in Section 6.3.4, where both compounds are outside the error margins despite the large range.

#### 5.7.3 Partitioning Coefficient ( $\log K_{ow}$ )

As shown in Section 3 and again here, the  $\log K_{ow}$  is a useful exclusion criterion despite having a relatively high error associated with the values. This is especially relevant where the LRI-based exclusion is not available. Taking a less conservative error margin would increase the exclusion rate based on this criterion, as would improvements to the prediction of  $\log K_{ow}$ . If the compound class of the unknown is clear, more specific calculations of  $\log K_{ow}$  could be used, rather than the very general calculations available in EPISuite<sup>TM</sup> [60]. The disadvantage of more specific calculations is that this would have to be applied on a compound specific basis and is therefore no longer applicable for an automated strategy designed for any unknown organic contaminant.

#### 5.7.4 Steric Energy

The use of the steric energy has proved to be a surprisingly versatile and effective exclusion criterion in this study. The calculation with MOLGEN-QSPR is very quick (seconds up to minutes for several thousand structures, Intel Core<sup>TM</sup> 2 Duo 1.83 GHz,

1.00 GB RAM) and provides similar information in a much shorter timeframe compared with ChemBio3D, despite the discrepancy in values. The quantiles show that molecules known to exist can have very high energies, however the majority have low energies. Furthermore, molecules that were very sterically hindered (e.g. one-atom bridge bonds over PAH systems) had energies far above the 90 % quantile used here as the exclusion criterion. The combination of both programs resulted in the exclusion of slightly more candidates than one alone.

These results are in agreement with a previous study investigating the energy of known (natural or synthesised) C<sub>6</sub>H<sub>6</sub> isomers [65], which also found that force field steric energy calculation in MOLGEN-QSPR was sufficiently accurate to describe the energy of structures compared with high-level enthalpy of formation calculation, with much lower computational effort.

Improvements to this criterion could be made by taking a bigger random set of molecules known to exist, to generate more accurate quantiles, or to restrict the data sets to include only compounds containing certain elements, a certain range in the number of atoms or other properties, such that a more accurate energy distribution for compounds of given formula could be estimated. As the molecular formula is known prior to structure generation, this could represent a simple improvement to the exclusion criterion presented here, which is currently very conservative.

Another factor not considered in the steric energy calculation is for example the activation energy. It is possible that this calculation excludes molecules that are energetically unfavourable but stable due to high activation energy required for the formation of breakdown products. This could be especially relevant for some environmental contaminants that have been produced despite unfavourable reaction conditions to take advantage of specific properties.

#### 5.7.5 *Spectral Match*

As identified in Section 4, the use of predicted spectra to assess the spectral match appears to be a very subjective selection criterion, depending on the quality of fragment prediction. The match value still forms an important part of the filtering process shown in Figure 26, as this value can be used to generate a rank of remaining candidates in terms of the original measured data.

### 5.8 *Implications and Conclusions*

The methods outlined in this section are very effective in reducing the number of structures generated for a given compound based on data generated during GC-EI-MS analysis. Further improvement to prediction techniques and the incorporation of other potential exclusion criteria could strengthen this further (e.g. including toxicity prediction

or additional partitioning properties). The applicability of this method to identify unknown compounds is shown in Section 6.

Although based on GC-EI-MS, several principles described here can also be extrapolated to liquid chromatographic (LC) and/or to MS based on other ionisation techniques. The limited availability of substructural classifiers, shown to be the critical first step, are likely to be the limiting factor in developing a similar method for other ionisation techniques.



## 6 Successful Unknown Identification in EDA Studies

The following section gives examples for the successful identification of unknowns in selected EDA studies. Firstly, Section 6.1 contains examples of tentative identifications resulting from the groundwater EDA used as the source of unknown spectra for Section 3. The results of the EDA study were published in 2010 by C. Meinert et al. [6].

Meinert, C., Schymanski, E., Küster, E., Kühne, R., Schüürmann, G. and Brack, W. (2010). Application of preparative capillary gas chromatography (pcGC), automated structure generation and mutagenicity prediction to improve effect-directed analysis of genotoxicants in a contaminated groundwater, *Environmental Science and Pollution Research*, 17, 885-897.

Section 6.2 contains the identification of a transformation product of diclofenac, confirmed both analytically and toxicologically and published by T. Schulze et al. [1]. Only the identification steps performed as part of this work are described here.

Schulze, T., Weiss, S., Schymanski, E., von der Ohe, P.C., Schmitt-Jansen, M., Altenburger, R., Streck, G. and Brack, W. (2010). Identification of a phytotoxic photo-transformation product of diclofenac using effect-directed analysis, *Environmental Pollution*, 158, 1461-1466.

A further two examples are presented in Section 6.3, from a river water EDA study performed by C. Gallampois et al. [11]. The identification steps performed on the GC-MS data are detailed below and form part of a manuscript in preparation [10].

Schymanski, E. L., Gallampois, C., Brack, W. (2010). Identification of Unknown GC-EI-MS Spectra in Elbe EDA Study, *in preparation*.

### 6.1 Tentative Identification of Bitterfeld Groundwater Contaminants

The spectra from the groundwater EDA from Bitterfeld, used during the method development in Section 3 are used again here to perform tentative identification of compounds present based on the methods outlined in Section 3 combined with a NIST mass spectral database search. The methods are included above and in [6] and are not repeated here. Instead, a representative selection of the tentatively identified compounds resulting from the structure generation approach is presented along with the results of the database search to demonstrate the use of the structure generation approach.

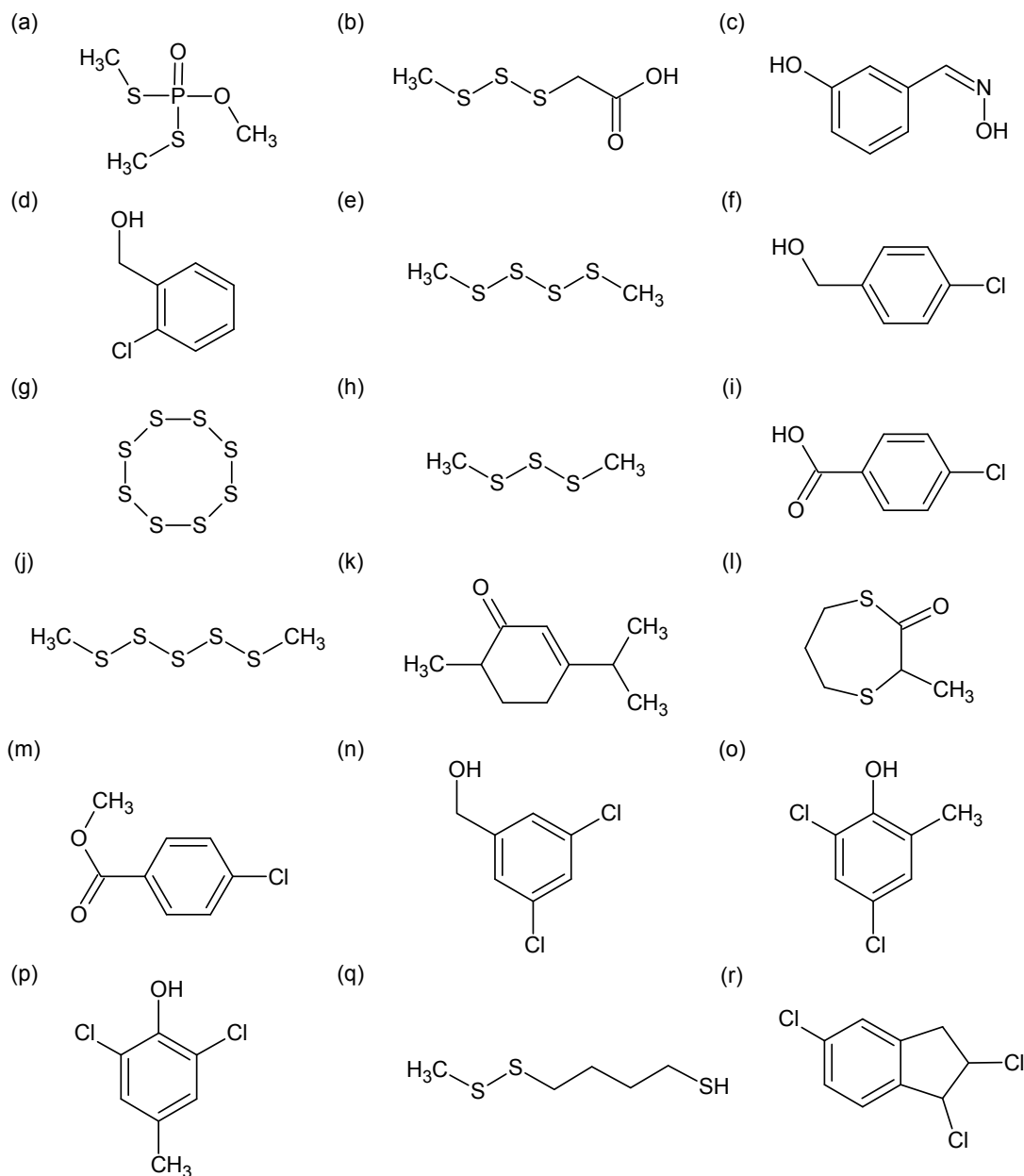
A total of 150 spectra were obtained from several fractions (4, 5, 6, 7, 9, 11 and 12) as well as eleven sub-fractions of fraction 8 (8.01 through to 8.11). Of these, 42 (28 %) could be tentatively identified using the NIST database search alone, considering spectra with a match probability above 65 %.

71 of those 150 spectra had a clear molecular formula based on calculations using MolForm or ElCoCo (see Section 3 and Section 5.1) and could be used for structure generation. These are the same spectra used in Section 3. From these 71, 32 of the compounds identified using NIST were confirmed using structure generation methods, while 20 additional peaks were tentatively identified using the methods from Section 3 alone. This resulted in a total of 62 tentative identifications, covering 41 % of the total spectra. A selection of the tentative identifications, taken from Fractions 4, 5, 6 and 8.09, is given in Table 22.

**Table 22: Selected tentative identifications from the Bitterfeld groundwater EDA. NIST matches are given where the probability was > 65 %. The MOLGEN-MS match refers to the tentative identification from structure generation methods, the following columns contain the total number of matches within the data ranges shown. The letters refer to the structures shown in Figure 32. Compounds highlighted in grey were confirmed as genotoxicants [6]. CAS Numbers (where available) are given in square brackets.**

| Fraction | NIST Match                                             | MOLGEN-MS Match                                                | No. of Matches | MV Range | log K <sub>ow</sub> Range |
|----------|--------------------------------------------------------|----------------------------------------------------------------|----------------|----------|---------------------------|
| 4        | O,S,S-Trimethyldithiophosphate [22608-53-3] <b>(a)</b> | O,S,S-Trimethyldithiophosphate [22608-53-3] <b>(a)</b>         | 4              | 50-82    | 0.50-1.70                 |
|          | unknown                                                | (Methyltrisulfanyl) acetic acid [not available] <b>(b)</b>     | 1              | 72       | 1.13                      |
|          | 3-Hydroxybenzaldehyde oxime [22241-18-5] <b>(c)</b>    | 3-Hydroxybenzaldehyde oxime [22241-18-5] <b>(c)</b>            | 6              | 44-46    | 1.00-1.50                 |
| 5        | 2-Chloro benzene methanol [17849-38-6] <b>(d)</b>      | o-, m- or p-Chloro benzene methanol                            | 3              | 15-45    | 1.72                      |
|          | Dimethyltetrasulfide [5756-24-1] <b>(e)</b>            | Dimethyltetrasulfide [5756-24-1] <b>(e)</b>                    | 5              | 64-68    | 1.50-2.40                 |
|          | unknown                                                | o-, m- or p-Chloro benzene methanol (p-) [873-76-7] <b>(f)</b> | 3              | 18-19    | 1.72                      |
|          | Cyclic octaatomic sulfur [10544-50-0] <b>(g)</b>       | Cyclic octaatomic sulfur [10544-50-0] <b>(g)</b>               | 1              | 11       | 0.23                      |
| 6        | Dimethyltrisulfide [3658-80-8] <b>(h)</b>              | Dimethyltrisulfide [3658-80-8] <b>(h)</b>                      | 8              | 49-67    | 1.50-2.00                 |
|          | unknown                                                | 2-, 3- or 4-Chlorobenzoic acid [74-11-3] <b>(i)</b>            | 3              | 73-75    | 2.52                      |
|          | Dimethylpentasulfide [7330-31-6] <b>(j)</b>            | Dimethylpentasulfide [7330-31-6] <b>(j)</b>                    | 8              | 20-23    | 1.20-2.40                 |
| 8.09     | unknown                                                | 3-Isopropyl-6-methylcyclohex-2-en-1-one [499-74-1] <b>(k)</b>  | 3              | 80-85    | 2.94                      |
|          | unknown                                                | 3-Methyl-1,4-dithiepan-2-one [72018-97-4] <b>(l)</b>           | 4              | 50-58    | 2.60-2.90                 |
|          | unknown                                                | o-, m- or p- Methyl chloro-benzoate [1126-46-1] <b>(m)</b>     | 3              | 77-78    | 2.47                      |
|          | unknown                                                | 3,5-Dichlorobenzene methanol [60211-57-6] <b>(n)</b>           | 6              | 40-42    | 2.36                      |
|          | 2,4-Dichloro-6-methylphenol [1570-65-6] <b>(o)</b>     | Dichloromethylphenol or Chloro(chloromethyl)phenol             | 9              | 78-80    | 2.90-3.40                 |
|          | 2,6-Dichloro-4-methylphenol [2432-12-4] <b>(p)</b>     | Dichloromethylphenol                                           | 9              | 78-79    | 2.90-3.40                 |
|          | 2,6-Dichloro-4-methylphenol [2432-12-4] <b>(p)</b>     | Dichloromethylphenol                                           | 9              | 78-79    | 2.90-3.40                 |
|          | unknown                                                | 4-(Methyldisulfanyl)butane-1-thiol [not available] <b>(q)</b>  | 2              | 60-68    | 3.23/3.29                 |
|          | unknown                                                | Trichloroindane <b>(r)</b>                                     | 6              | 45-60    | 4.47/4.94                 |

As the tentative identifications from structure generation methods generally included several possible isomers, this is indicated by the number of matches and the ranges of mass spectral match value and log  $K_{ow}$  given in the table. The full listing of compounds is given in Table 1 in Meinert et al. [6]. The compounds corresponding to the letters in Table 22 are given in Figure 32.



**Figure 32: Selected tentatively identified compounds from the Bitterfeld groundwater EDA. The letters correspond to the entries in Table 22.**

The results in Table 22, as a selection of compounds from [6], shows that the structure generation approach combined with substructural classifiers provides a strong line of evidence in structure identification or confirmation if no standards are available and also helps to suggest possible structures if the library search does not produce a satisfactory

match. Twenty of the original 150 compounds would not have been identified otherwise, as these were not contained within the NIST database, or not considered a match with a sufficiently high probability. Nine of these are included in the table above, structures (b), (f), (i), (k), (l), (m), (n), (q) and (r) in Figure 32. The exclusion of candidates based on the  $\log K_{ow}$  criterion was a valuable part of the method, to prevent consideration of too many structures.

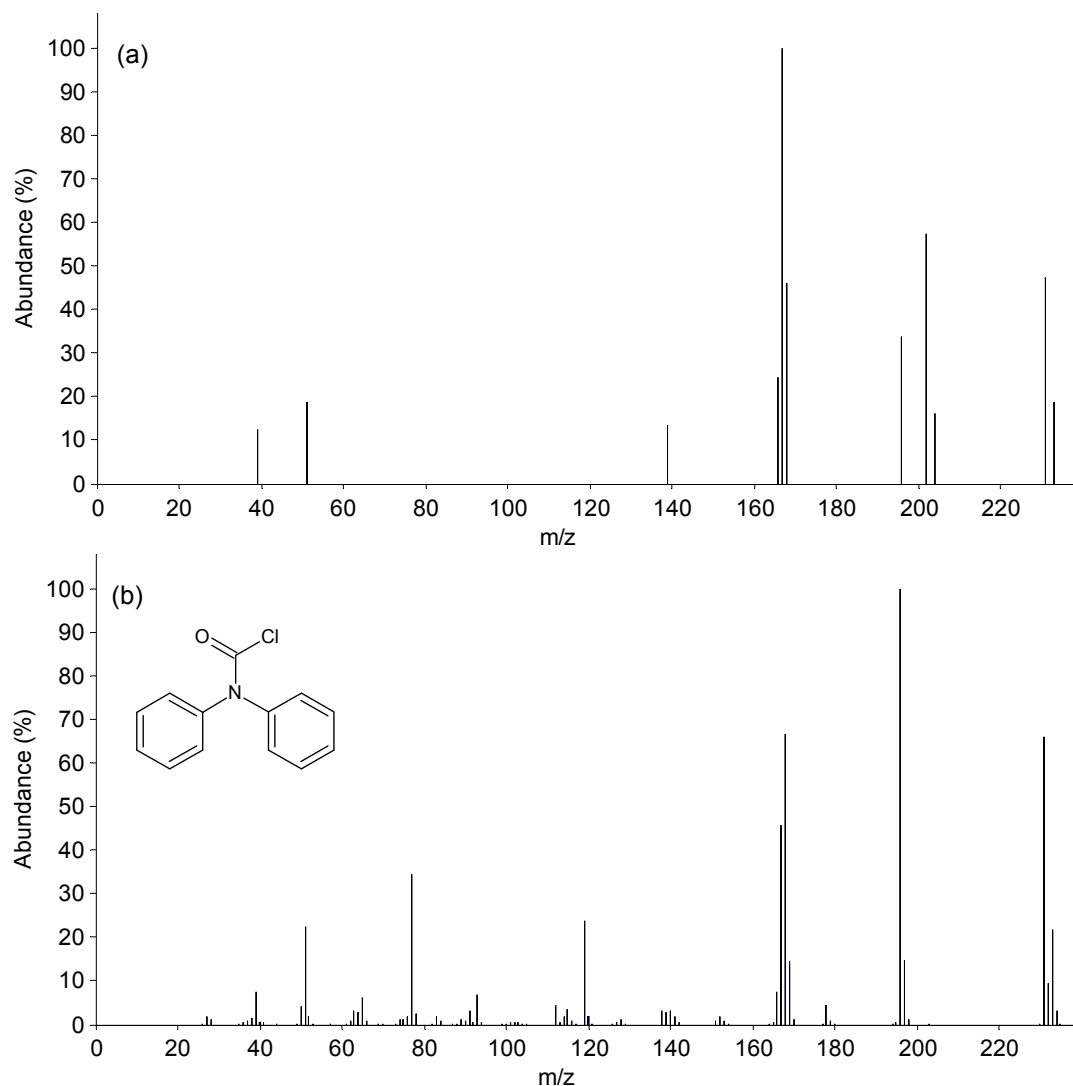
The compounds shown in Figure 32 cover the range of compounds found in the Bitterfeld groundwater quite well. The groundwater was characterised with many thiophosphate compounds, oxy-chloro aromatic compounds and alkyl sulfide chains. As shown in Table 22 with the grey highlighting (compounds (i) and (o) in Figure 32), some of the compounds identified were found to be genotoxic experimentally by C. Meinert et al. [6], whereas other compounds (e.g. compound (a) in Figure 32) were found to be non-genotoxic.

As several of the tentatively identified structures were not available for purchase, mutagenicity prediction was used by the authors to potentially select genotoxic candidates, based on Kazius et al. [25] (see Table 1) using the software ChemProp [83, 84]. Ten structures were predicted to be mutagenic, 19 non-mutagenic and 29 were outside the applicability domain of the model [6]. Confirmation of the results experimentally showed that three experimentally determined genotoxics were predicted as non-mutagenic by the model, whereas three predicted mutagens were revealed to be non-genotoxic [6]. This was attributed in part by the authors to the differences between the *umuC* genotoxicity assay and the Ames fluctuation mutagenicity test (*S. typhimurium*) endpoints, despite their comparable sensitivity [6]. As a result, the mutagenicity prediction could not be used to narrow down the list of tentatively identified compounds into a list of candidate genotoxics in this case.

## 6.2 Diclofenac and Transformation Products

Diclofenac is an anti-inflammatory pharmaceutical product, commonly detected in surface and groundwater systems as well as in waste water treatment plant effluents [1]. Photolysis is the most important transformation pathway of diclofenac (see references within [1]), however a previous study found that transformation products were more toxic towards green algae than the parent compound, with a maximum toxicity between 31 and 53 hours of sunlight irradiation [85]. To determine which of the transformation products was responsible for the enhanced toxicity of irradiated diclofenac towards the green algae *Scenedesmus vacuolatus*, EDA was performed, initially by S. Weiss, on a diclofenac solution exposed to sunlight.

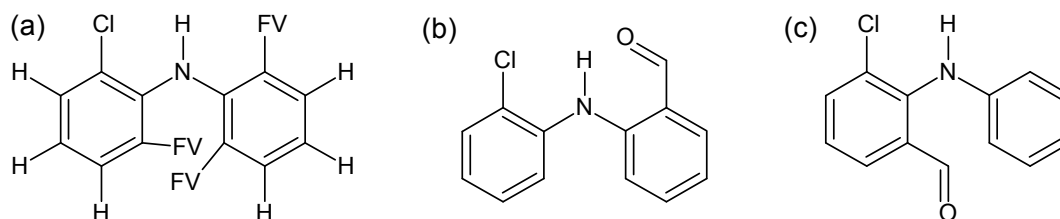
Reversed-phase HPLC fractionation during the EDA yielded several fractions with approximate  $\log K_{ow}$  ranges. Only one of these fractions showed enhanced toxicity towards the algae, with a  $\log K_{ow}$  range between 3.4 and 3.7. A detailed look at the chromatogram revealed only one peak. The corresponding spectrum of interest, albeit of rather low quality, shown in Figure 33, with the spectrum of the closest NIST match, diphenyl carbamic chloride (DCC, 48.1 % probability) below it. Although the spectra show some similarities, the peak group at 202 is missing in the DCC spectrum, while the peak groups around 80 and 120 are missing in the unknown spectrum. Furthermore, the  $\log K_{ow}$  of DCC, 1.64, was far outside the  $\log K_{ow}$  range of 3.4 to 3.7 estimated for the fraction, providing more evidence that this was not the compound detected.



**Figure 33:** (a) Unknown spectrum of a diclofenac transformation product at 16.675 minutes,  $\log K_{ow}$  3.4-3.7 and (b) the spectrum of the closest NIST match, diphenyl carbamic chloride (DCC, see insert). DCC spectrum retrieved from the NIST database.

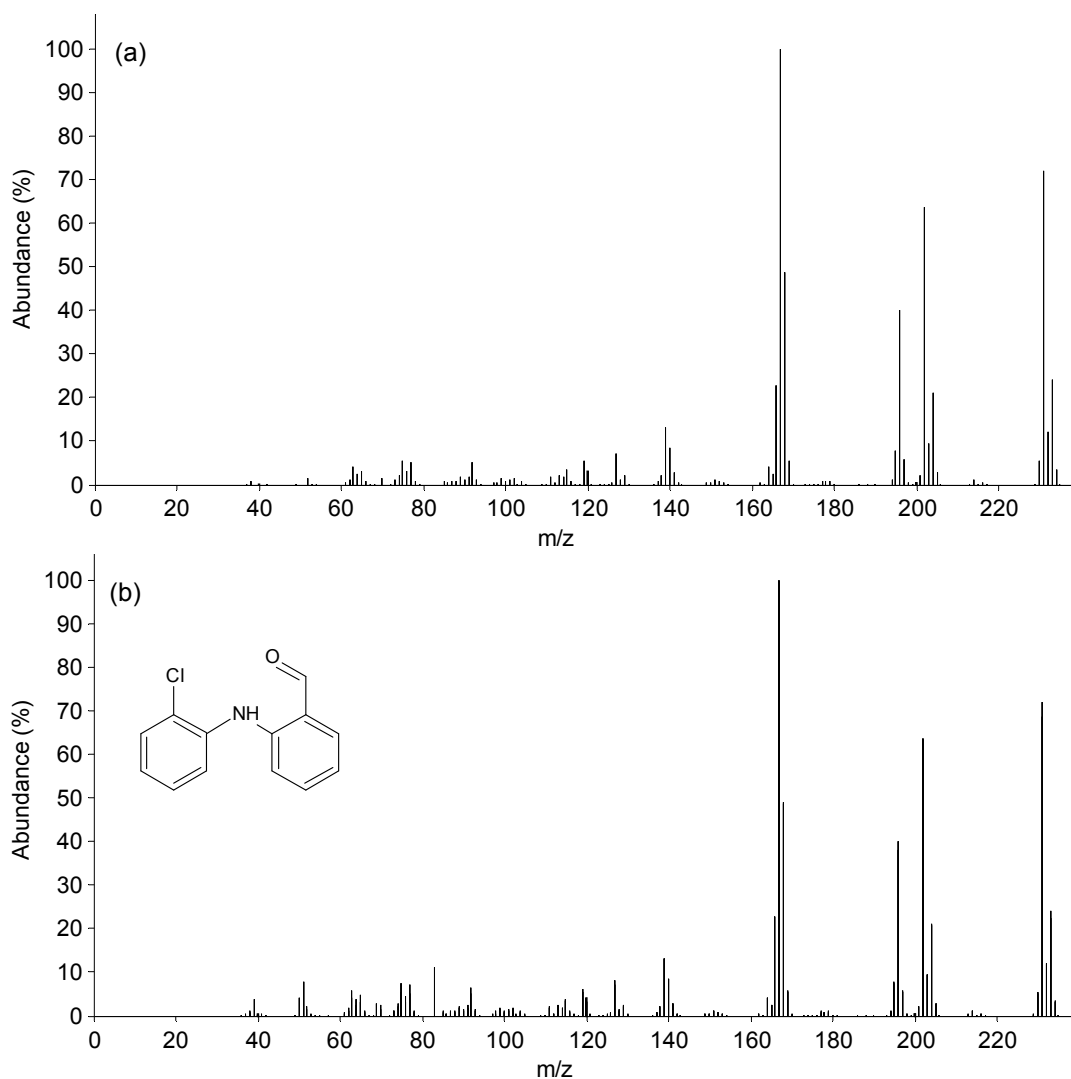
Identification of this unknown spectrum using the methods developed in Section 3 was undertaken. Substructural information retrieved from the NIST database and MOLGEN-

MS revealed the presence of more than one aromatic ring, one chlorine and 0-2 oxygens. Substructures absent included Ar-O,  $\text{CH}_2/3$ , ethers, OH,  $\text{NCH}_3$ ,  $\text{NH}(\text{CH}_2)$ , C-O,  $\text{C}=\text{OO}$  and  $\text{NH}_2$ . Combining these restrictions led to a molecular formula of  $\text{C}_{13}\text{H}_{10}\text{ClNO}$  and a reduction in possible structures from over 1,000,000,000 (based on the formula alone) down to 36 possible structures (see Figure 22 for a schematic representation of this process). However, 36 structures are still too many for candidate selection. As the precursor is known in this study, this can be incorporated in the structure generation. Aguera et al. [86] published a detailed study on transformation products of diclofenac, which revealed structural similarities in all transformation products. This information was condensed into one substructure for addition to the MOLGEN-MS good list, shown in Figure 34(a). The subsequent structure generation resulted in only two candidates, shown in Figure 34(b) and (c). Both structures have a predicted  $\log K_{ow}$  value of 3.65 (EPISuite™ Kowwin), which is within the fraction range of 3.4 to 3.7. As diclofenac has two chlorines on one aromatic ring and the acetic acid group on the other aromatic ring, the structure shown in Figure 34(b) was considered more likely.



**Figure 34:** (a) Diclofenac-specific substructure added to MOLGEN-MS. FV refers to free valence, while the hydrogens are shown explicitly as this formed a part of the substructure. (b) and (c) The two structures generated as a result of this substructure restriction in combination with the other substructure classifiers.

The compound (b), 2-[(2-chlorophenyl)amino]benzaldehyde (CPAB) was synthesised by SYNCHEM GmbH, Felsberg, Germany. The mass spectrum of the standard, shown in Figure 35(b), contained the same main peak groups as the original unknown. Confirmation analysis performed by T. Schulze et al. [1], including a repeat of the EDA to obtain a better quality unknown spectrum, revealed a good match between the synthesised standard and the re-isolated unknown spectrum (Figure 35(a)), with a NIST match value of 989 (of 1000) and KRI values of 1981.0 and 1980.8 for the unknown and standard, respectively. This compound was also confirmed as the one responsible for the enhanced toxicity of the transformed diclofenac towards the green algae *S. vacuolatus* [1].



**Figure 35:** (a) Spectrum of the re-isolated unknown and (b) spectrum of 2-[(2-chlorophenyl)amino] benzaldehyde (CPAB), shown in insert.

### 6.3 EDA of Elbe River Water using Blue Rayon as a Passive Sampler

The methods described in Section 5, specifically Figure 26, were used to identify potential mutagens isolated during the EDA of river water collected from Pardubice (Czech Republic) using the passive sampler blue rayon [87, 88]. The EDA method development and analysis was performed by C. Gallampois et al. [11, 12]; the results of the identification based on available GC-MS data are presented in this section [10].

#### 6.3.1 Methods

GC-EI-MS analysis was conducted on various fractions obtained using semi-preparative liquid chromatography during an EDA of a river water sample collected using a blue rayon passive sampler. Although the compounds adsorbed by blue rayon and

fractionation methods are more suitable for LC-MS/MS analysis methods, GC-MS analysis was performed where sufficient sample was available to support the identification efforts. The Ames fluctuation test [89] was used to assess the mutagenicity of the sample and the resulting fractions. Further details on the EDA method are given in [11].

GC-MS (Model 6890 N, detector MSD 5973, Agilent Technologies, Waldbronn, Germany) analysis was performed using a HP-5MS capillary column (30 m  $\times$  0.25 mm I.D., 0.25  $\mu$ m film, 5 % phenylmethylsiloxane, Agilent Technologies) and temperature program 70 °C (held for 4 min.), 3 K/min. to 300 °C (held for 20 min.). The calibration mix C8-C36 from Supelco (46827U) was used to calculate the Kovat's Retention Index (KRI), whereas the standard EPA-PAH mix from Ermsdorfer was used to record PAHs to calculate the Lee Retention Index (LRI). The spectra of all samples were recorded first without and then with addition of the standards to avoid the loss of peaks of interest.

AMDIS [41] was used to deconvolute the spectra and identify peaks in the sample by comparing the fractionated sample with the respective fractionated blank extract of blue rayon. Deconvolution settings were medium. KRI and LRI data were calculated according to Equation 8. MSP files for the selected deconvoluted spectra were saved from AMDIS and submitted to a NIST library search. As for case study spectra, NIST substructural information was printed to PDF and exported to text format for automatic upload into MOLGEN-MS. The MSP file from AMDIS was converted to CSV for import into MOLGEN-MS using a Matlab script (see Appendix 2). All other spectral processing was as described in Section 5.

### 6.3.2 Results - General

The EDA on the blue rayon samples from Pardubice involved several steps. The Ames test results on the first fractionation step using solid phase extraction (SPE) and ionic exchange revealed that the acidic and neutral fractions were active, while the basic fraction showed no significant mutagenicity [11]. As the blue rayon is designed to sample planar compounds with three or more aromatic rings, the acid and neutral fractions were analysed with GC-MS methods by C. Gallampois and M. Heinrich to determine if any compounds would be detected at all. Processing of the results with AMDIS after subtraction of the blank revealed a few peaks of interest in both the acidic and neutral fractions, shown in Table 23. Corresponding LC-MS/MS measurements by C. Gallampois after blank subtraction revealed thousands of peaks of interest, justifying firstly the need for further fractionation and secondly confirming that most compounds in the samples can not be detected using GC-MS.



**Table 23: Peaks of interest in the acidic and neutral fractions.**

| Sample      | Spectrum (m/z)              | Top NIST match (Probability)                                             |
|-------------|-----------------------------|--------------------------------------------------------------------------|
| BR1A_18.970 | 50 76 104 148               | Phthalic acid (55.0 %)                                                   |
| BR1A_24.966 | 50 76 104 147               | Phthalimide (40.0 %)                                                     |
| BR1A_54.591 | 57 190 242 283 339 395 410  | 4,4'-[(1-methylethylidene)bis(4,1-phenyleneoxy)]bis-benzenamine (66.3 %) |
| BR1A_67.976 | 128 156 256 284 302         | No clear match                                                           |
| BR1A_69.432 | 110 160 250 284 319         | 4-(1,3-bis-(4-chlorobenzyl)-imidazolindin-2-yl)-pyridine (57.1 %)        |
| BR1N_10.386 | 42 58 83 98 140 155         | 2,2,6,6-tetramethyl-4-piperidone (65.0 %)                                |
| BR1N_26.926 | 41 57 74 91 163 175 191 206 | 2,4-di-tert-butylphenol (60.6 %)                                         |
| BR1N_56.120 | 77 94 170 251 362           | Diphenyl-2-ethylhexyl-phosphate (83 %)                                   |

The first sub-fraction of the acidic fraction, BR1A1, was analysed as this showed medium to high mutagenicity and calculations during method validation indicated that this fraction may have the compounds most likely to be detectable using GC-MS analysis [11]. Two peaks of interest were present following deconvolution, BR1A1\_19.208 and BR1A1\_25.410, corresponding with the first two compounds in Table 23. No other peaks of interest were found. Although several active sub-fractions of the neutral fraction were also analysed, no peaks of interest were seen (results not shown).

As the fraction BR1A1 was obtained using HPLC with a polymeric C18 column, the log Kow range of the fraction can be determined. A linear regression with 7 standards yielded a log Kow range for Fraction A1 (0-10 min) of -0.15 to 1.35. This range is larger than usual for fractionation during EDA due to the use of the polymeric C18 column.

### 6.3.3 BR1A1\_19.208

The data for the first unknown is shown in Table 24. From a quick glance at the retention index data, the first match appears to fit the data the best.

**Table 24: NIST match data for the first unknown BR1A1\_19.208.**

| Compound                           | Spectrum (m/z)                         | Match  | MW  | KRI                        |
|------------------------------------|----------------------------------------|--------|-----|----------------------------|
| BR1A1_19.208                       | 74(12) 76(57) 104(100) 148(20)         |        |     | 1320 (exp)                 |
| Phthalic anhydride (CAS 85-44-9)   | 50(43) 74(20) 76(89) 104(100) 148(34)  | 44.6 % | 148 | 1443 ± 382                 |
| Phthalic acid (CAS 88-99-3)        | 50(38) 74(19) 76(77) 104(100) 148(22)  | 41.2 % | 166 | 1620 ± 220<br>1917*, 1617* |
| Phthalamic acid (CAS 88-97-1)      | 17 (19) 50(41) 76(86) 104(100) 148(16) | 8.4 %  | 165 | 1673 ± 382                 |
| Monoethylphthalate (CAS 2306-33-4) | 50(43) 74(18) 76(81) 104(100) 148(14)  | 4.6 %  | 194 | 1629 ± 382                 |

\* indicates experimental values listed in the NIST database.

The method outlined in Section 5 was used to confirm this indication, as follows:

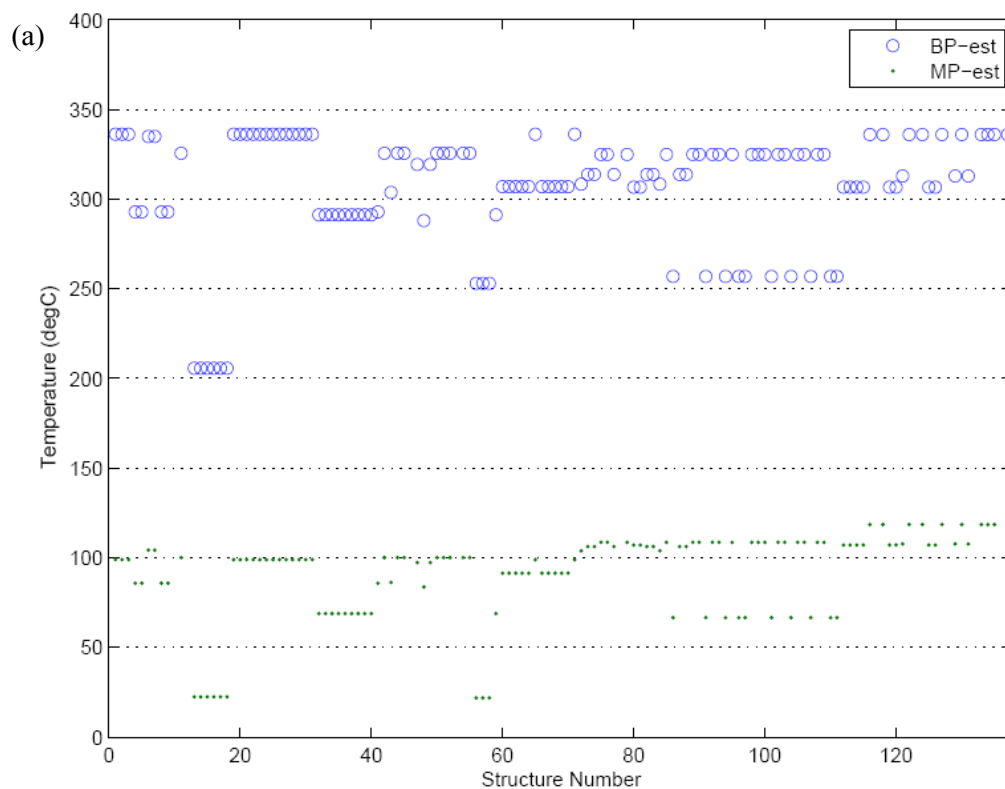
NIST and MOLGEN-MS classifier information revealed the substructure 'ArC=OO' (aryl ester) was present (99 %) and RDB ≥ 5. Aryl-oxygen bonds were absent, as were aryl-saturated carbon bonds and CH<sub>2</sub> or CH<sub>3</sub> groups.

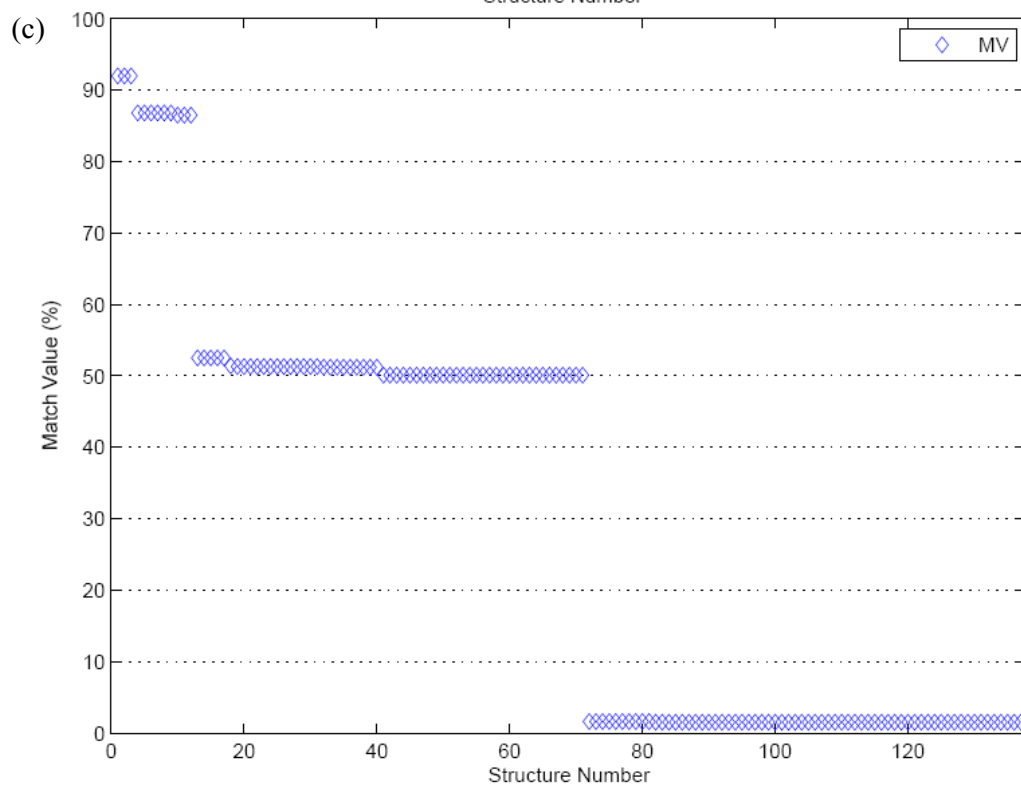
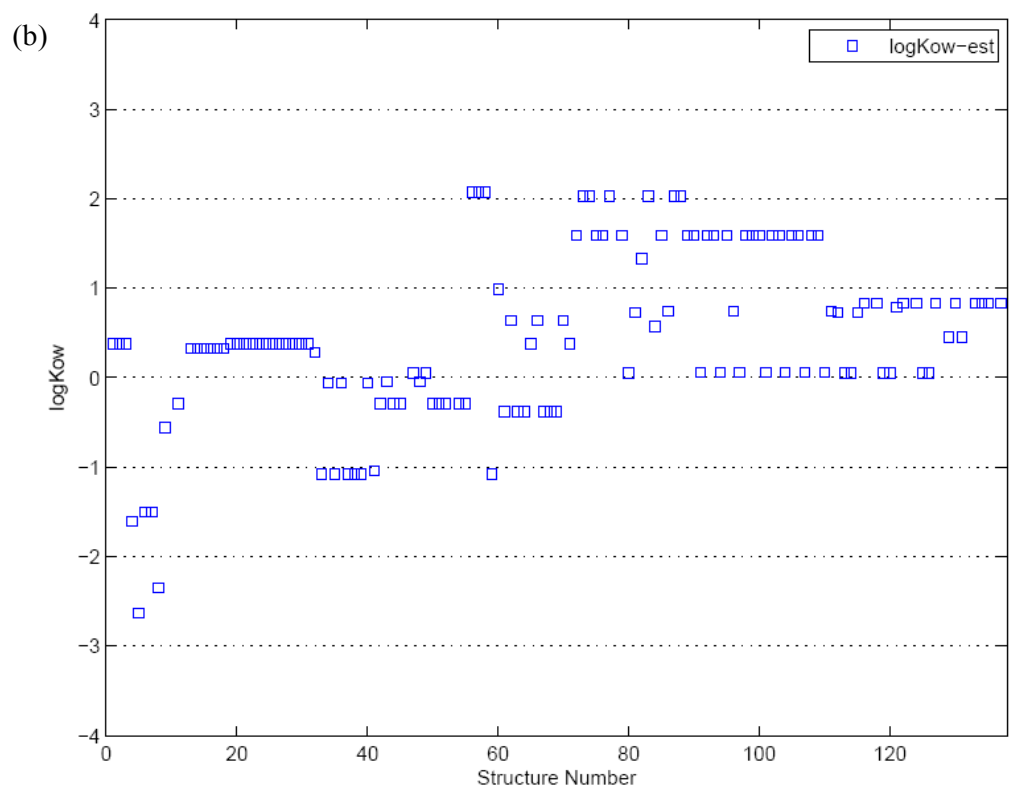
Combining the substructure, RDB and element information left compounds with the elements C ≥ 7, O = 2-4 and H with RDB = 6-8. Entering this information into MOLGEN-MS MolForm resulted in the generation of three compatible molecular

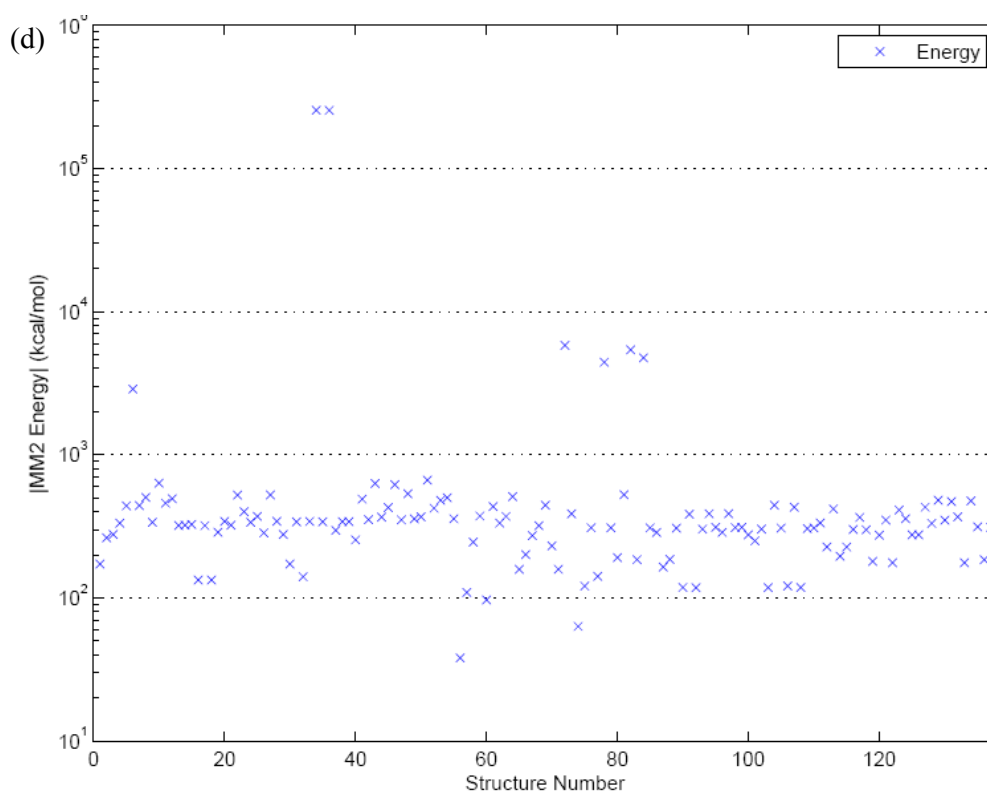
formulae, ranked according to the isotope peak deviation (lower deviation indicates better match):  $C_9H_8O_2$  (RDB 6, deviation 4.7),  $C_8H_4O_3$  (RDB 7, deviation 5.7) and  $C_7O_4$  (RDB 8, deviation 6.8). All three formulae were used with the classifier information to generate 137 possible structures using MOLGEN-MS, in 3.2 sec, with MVs between 1.5 and 92.0 %. Although the last formula is highly unlikely, this is accounted for during structure generation.

The predicted data for all structures is given in Figure 36, including melting and boiling point data (a),  $\log K_{ow}$  (b), match values (c) and energy (d). Structures are sorted according to MV (plot (c)).

The LRI calculated for BR1A1\_19.208 was 224.7. The correlation developed by Eckel & Kind [58] with the additional error margin defined above means that structures with boiling points outside the range (LRI -31) to (LRI + 71) °C, i.e. 193.7 to 295.7 °C, can be eliminated from consideration. Using this as the first selection criterion results in the elimination of 108 of the 137 possible structures, leaving 29 candidates. Three of these 29 candidates had a  $\log K_{ow}$  below the inclusion range of -1.15 to 2.35, such that 26 candidates remain. All but three of these remaining 26 compounds had energies above 213.24 kcal/mol calculated with ChemBio3D. These remaining three structures are shown in Table 25, together with the calculated parameter range for all structures and the inclusion criteria.

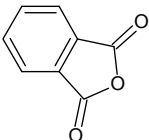
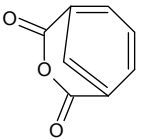
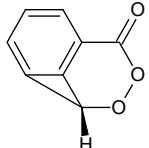






**Figure 36:** Predicted data for 137 structures generated for BRA1\_19.208 according to the substructural information. (a) Melting and boiling point data (b)  $\log K_{ow}$  (c) MOLGEN-MS match value (MV) and (d) ChemBio3D steric energy. Structures sorted according to descending MV.

**Table 25:** Predicted data for the final three candidates for BR1A1\_19.208, including the calculated parameter range for all structures and the exclusion criteria.

| Structure                                                                           | MV     | BP          | $\log K_{ow}$ | Energy (ChemBio3D) | Energy (MOLGEN-QSPR) |
|-------------------------------------------------------------------------------------|--------|-------------|---------------|--------------------|----------------------|
|  | 50.1   | 253         | 2.07          | 38                 | 117                  |
|  | 50.1   | 253         | 2.07          | 109                | 254                  |
|  | 51.2   | 291         | 0.28          | 140                | 276                  |
| Parameter range (all structures)                                                    | 1.5-92 | 206-336     | -2.63-2.07    | 38-256,000         | 117-817              |
| Inclusion range                                                                     | > 50 % | 193.7-295.7 | -1.15 to 2.35 | < 213.24           | < 429                |

LC-MS/MS analysis (LTQ Orbitrap, APCI positive) of BR1A1 revealed a small peak of  $m/z[M+H]^+ = 149.0229$  at retention time 3.04 minutes, corresponding with the formula of the selected candidates (calculated  $[M+H]^+ = 149.0233$ , monoisotopic mass = 148.0160), which was not detected in any blanks [12]. As the top candidate in Table 25 was considered most likely, the corresponding analytical standard, phthalic anhydride, was purchased (Sigma Aldrich) and measured using the same GC-MS method. Phthalic anhydride was detected at 18.9843 minutes, with a very similar mass spectrum to the unknown BR1A1\_19.208. The KRI and LRI were calculated as 1320.2 and 224.0, respectively, compared with the unknown 1320 and 224.7, resulting in a confirmation of the unknown compound as the first structure in Table 25, phthalic anhydride. Confirmation LC-MS/MS analysis of phthalic anhydride revealed a peak at 3.06 minutes of mass 149.0226. Thus the presence of phthalic anhydride in the sample is confirmed using two analytical techniques.

#### 6.3.4 BR1A1\_25.410

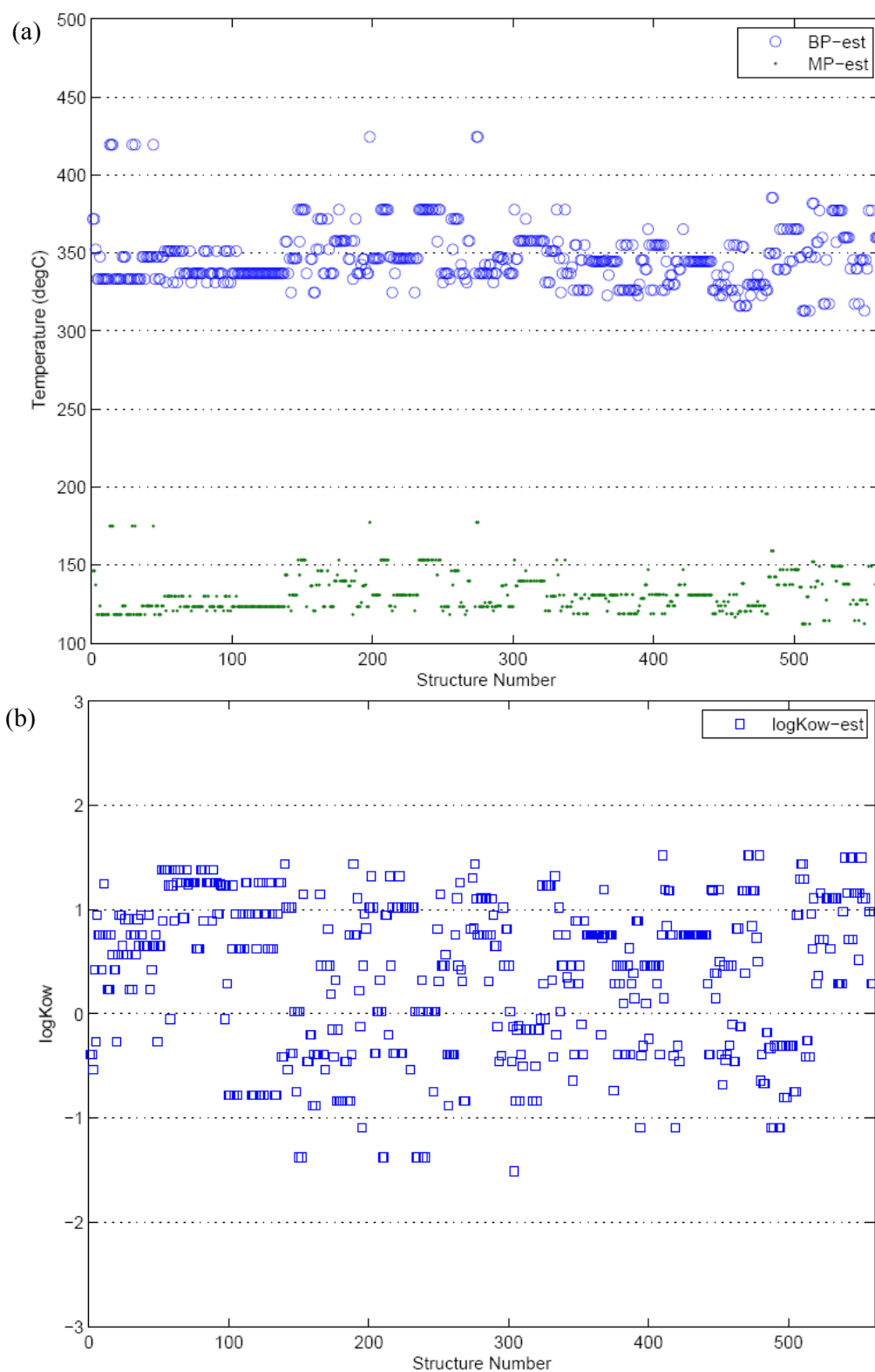
The data for the second unknown is shown in Table 26. From a quick glance at the retention index data, either of the first two compounds could match the unknown, while the larger molecules further down have much larger KRIs.

**Table 26: NIST match data for the unknown BR1A1\_25.410.**

| Compound                                              | Spectrum                               | Match  | MW  | KRI            |
|-------------------------------------------------------|----------------------------------------|--------|-----|----------------|
| BR1A1_25.410                                          | 76(47) 103 (24) 104(52) 147(100)       |        |     | 1472 (exp)     |
| Phthalimide (CAS 85-41-6)                             | 50(20) 76(55) 103(16) 104(57) 147(100) | 61.4 % | 147 | 1381 $\pm$ 382 |
| $\alpha$ -cyanobenzoic acid (CAS 3839-22-3)           | 50(46) 76(99) 103(37) 104(74) 147(100) | 22.4 % | 147 | 1428 $\pm$ 382 |
| Hydroxymethylphthalimide (CAS 118-29-6)               | 50(38) 76(100) 103(29) 104(59) 147(69) | 6.22 % | 177 | 1781 $\pm$ 382 |
| N-(2-acetamidoethylthio) phthalimide (CAS 25158-14-9) | 50(59) 76(100) 103(35) 104(75) 147(92) | 4.3 %  | 264 | 2488 $\pm$ 382 |
| Phthalimidomethyl-3-methoxy-benzoate (NIST 224824*)   | 50(56) 76(100) 103(40) 104(66) 147(83) | 4.1 %  | 311 | 2667 $\pm$ 382 |

\* CAS Number not available, NIST spectrum number provided instead.

The method from Section 5 was used to identify the unknown. NIST and MOLGEN-MS classifier information revealed the substructure ‘Ar-C=O-N<’ (aromatic amide) and ‘NHC=O’ (amide) were present (99 %) and RDB  $\geq 5$ . Combining the substructure, RDB and element information left compounds with the elements C  $\geq 7$ , N  $\geq 1$ , O  $\geq 2$ , H and S with RDB = 7-8. Entering this information into MOLGEN-MS MolForm resulted in the generation of two compatible molecular formulae, ranked according to the isotope peak deviation (lower deviation indicates better match): C<sub>7</sub>H<sub>1</sub>N<sub>1</sub>O<sub>3</sub> (RDB 8, deviation 0.38) and C<sub>8</sub>H<sub>5</sub>N<sub>1</sub>O<sub>2</sub> (RDB 7, deviation 0.54). Both formulae were used with the classifier information to generate 561 possible structures (all with formula C<sub>8</sub>H<sub>5</sub>N<sub>1</sub>O<sub>2</sub>) using MOLGEN-MS, with MVs between 39.6 and 89.1 %. The predicted data for all structures is given in Figure 37.



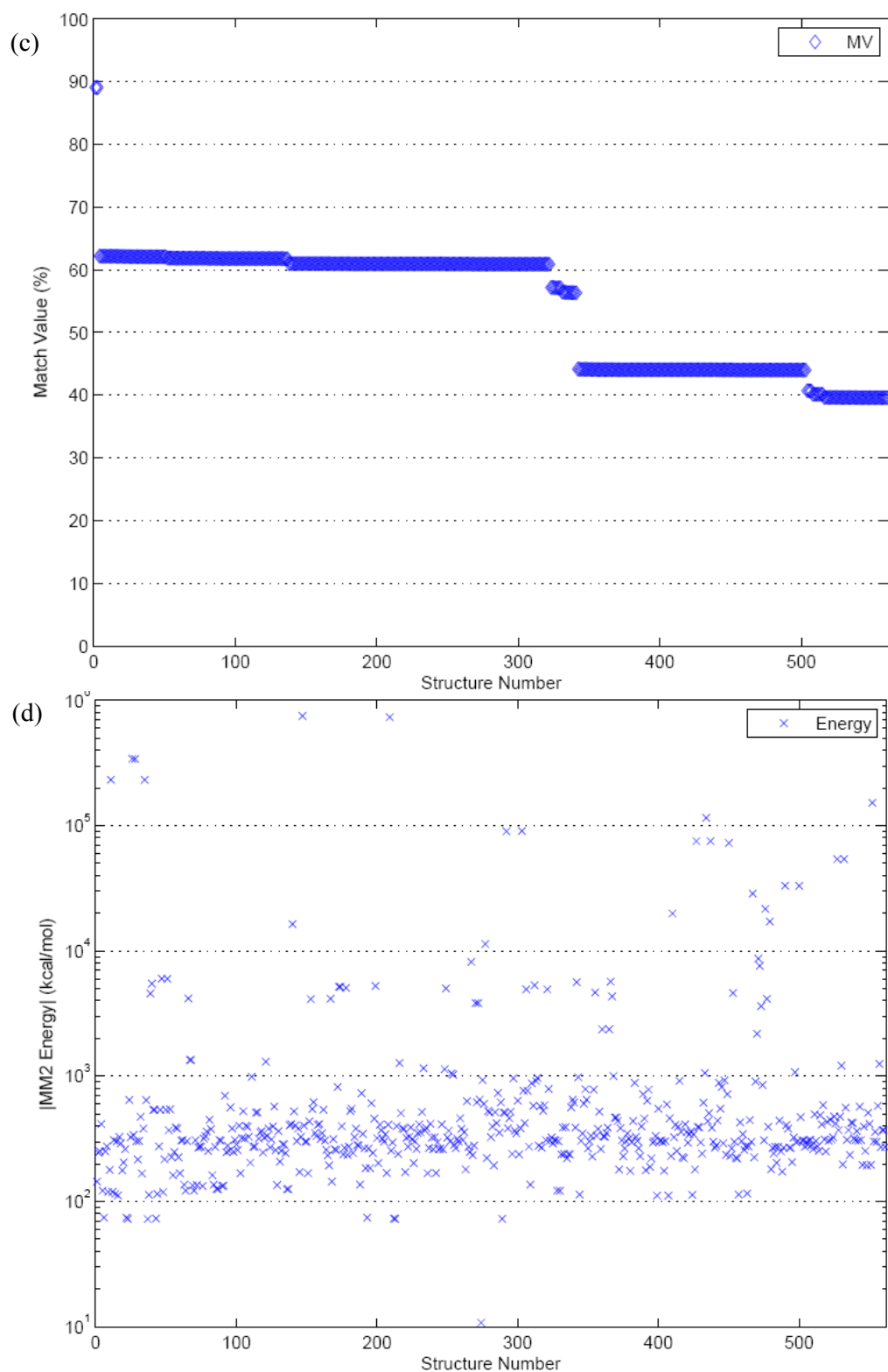


Figure 37: Predicted data for 561 structures generated for BRA1\_25.410 according to the substructural information. (a) Melting and boiling point data (b)  $\log K_{ow}$  (c) MOLGEN-MS match value (MV) and (d) ChemBio3D steric energy. Structures sorted according to descending MV.

The range of predicted data for the generated structures is much lower than in the previous example. The LRI calculated for BR1A1\_25.410 was 251.0, i.e. an inclusion range of BP = 220.0 to 322.0 °C. Only 14 of the predicted structures are within this range, however all of these 14 compounds had energies above the ChemBio3D cut-off of 213.24 kcal/mol. As this would leave no structures in consideration, instead no structures were eliminated based on the BP-LRI correlation. The reason for this discrepancy in values is discussed in Section 6.3.4.

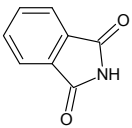
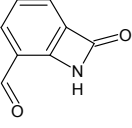
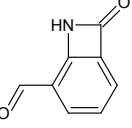
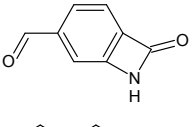
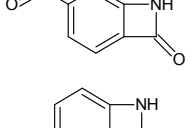
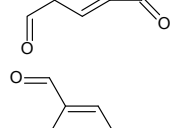
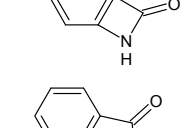
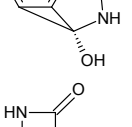
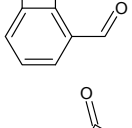
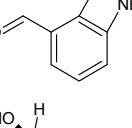
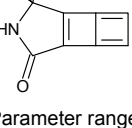
The log  $K_{ow}$  of the generated structures ranged from -1.51 to 1.52, with no values calculated for 31 structures. Those structures with log  $K_{ow}$  below -1.15 (the lower end of the exclusion range, including error margin) or not calculated were excluded, such that 521 structures remain. Those structures without calculated values were excluded as EPISuite™ generally only failed to predict data for very strange compounds in our experience. Applying the energy criteria developed in Section 5.4, 80 structures remain in consideration with ChemBio3D energies below 213.24 kcal/mol. A further 7 can be eliminated with MOLGEN-QSPR energies above 429 kcal/mol, such that 73 structures remain following all elimination steps.

Figure 37(c) shows a distinct break in the MVs between 44 and 57 %. The MV can thus be used to eliminate all structures with MV below 56 %, leaving 60 candidates for further consideration. A look at Figure 37(d) shows one compound with a much lower energy than all others (almost on the x-axis, Structure 278), followed by a few with energies below 100 kcal/mol. A similar pattern was observed using MOLGEN-QSPR calculations; one candidate had a much lower energy than all others (132 kcal/mol), another 9 structures had energies below 200, while for the other structures the values were significantly larger. These structures with lower energies are given in Table 27. An additional structure had a much higher MV (89.1 %) than the other structures and despite having a ChemBio3D energy above 100, this is also included in Table 27.

Looking at the structures shown in Table 27, the structure with the lowest energy would generally be considered more likely than the other structures. This corresponds with the top NIST match, phthalimide. Phthalimide (Riedel) was measured using the same GC-MS methods with the Kovat's and Lee RI standards. A peak at 25.464 minutes was detected, with KRI = 1474 and LRI = 251.3, compared with the unknown KRI = 1472 and LRI = 251.0. The mass spectrum also matches well with the unknown spectrum. In addition, the peak  $m/z[M-H]^- = 146.0249$  was detected in LC-MS/MS (LTQ Orbitrap, APCI negative) analysis of BR1A1 at 4.80 min, consistent with the standard phthalimide retention time of 4.85 with signal  $m/z[M-H]^- = 146.0254$  and the calculated exact mass of  $[M-H]^- = 146.0248$  and monoisotopic mass = 147.0320 [12]. As a result, the presence of phthalimide in BR1A1 is confirmed using structure generation methods, GC-MS analysis and LC-MS/MS analysis.



**Table 27: Predicted data for the final selected 11 candidates for BR1A1\_20.410. Entries sorted according to the ChemBio3D energy. The BP inclusion range is shown in brackets as this was not applied here (see text).**

| Structure                                                                           | MV          | BP            | logK <sub>ow</sub> | Energy<br>(ChemBio3D) | Energy<br>(MOLGEN-QSPR) |
|-------------------------------------------------------------------------------------|-------------|---------------|--------------------|-----------------------|-------------------------|
|    | 60.88       | 424.5         | 1.30               | 10.8                  | 131.8                   |
|    | 62.06       | 347.6         | 0.95               | 72.4                  | 184.1                   |
|    | 62.11       | 347.6         | 0.95               | 72.4                  | 184.1                   |
|    | 60.88       | 347.6         | 0.95               | 72.6                  | 184.4                   |
|   | 60.93       | 347.6         | 0.95               | 72.6                  | 184.4                   |
|  | 60.93       | 347.6         | 0.95               | 72.9                  | 184.4                   |
|  | 62.06       | 347.6         | 0.95               | 72.9                  | 184.4                   |
|  | 60.93       | 346.2         | 0.22               | 74.1                  | 338.4                   |
|  | 62.11       | 347.6         | 0.95               | 74.1                  | 184.3                   |
|  | 62.17       | 347.6         | 0.95               | 74.1                  | 184.3                   |
|  | 89.08       | 371.79        | -0.39              | 143.7                 | 329.5                   |
| Parameter range<br>(all structures)                                                 | 39.59-89.08 | 313.0-424.5   | -1.51-1.52         | 10.8-749,800          | 131.8-1439              |
| Inclusion range                                                                     | > 57 %      | (220.0-322.0) | -1.15 to 2.35      | < 213.24              | < 429                   |

Mutagenicity testing of both identified substances using the standards in the Ames fluctuation test revealed no mutagenicity associated with either compound [12]. Thus both compounds are confirmed analytically but are not responsible for the toxicity of the sample. As no other peaks of interest were found in the GC-MS chromatogram, the toxicants responsible for the effects are either below the detection limits or not detectable using GC-MS methods.

#### 6.3.4 Discussion

The method proposed and tested in Section 5 has been shown here to be very effective in excluding false structural candidates and leading to the tentative identification of the structures that were finally confirmed as being present in the sample using two orthogonal analytical techniques. Although the experimental KRI and LRI information gave an indication which of the NIST matches was more likely, the relatively low spectral match probabilities and lack of experimental KRI values in NIST meant that a pure database search did not provide enough information for a tentative identification in these cases. The use of structure generation, substructure classifiers and exclusion criteria has proven to be a valuable method of providing additional evidence for a tentative identification.

The boiling point-LRI relationship developed by Eckel and Kind [58] was once again a useful exclusion criterion for the first example, phthalic acid, but not for the second example, phthalimide. This failure warranted further investigation. Within the original study, Eckel and Kind tabulate seven compounds with LRI 50 points lower than the boiling point (i.e. outside their proposed range), benzoic acid, 2-phenyl indole, 1,3-benzenediol, 9-nitroanthracene, 1,4-dinitrobenzene, 4-nitroaniline and 3-nitroaniline. All molecules are aromatic and the majority also contain nitrogen. No LRI was previously available for phthalimide, however a comparison of boiling point data using a few different common sources revealed a very large discrepancy in predicted values (all in °C) for phthalimide: 424.5 (EPISuite<sup>TM</sup> [60]), 359 (ACD/Labs within ChemSpider [27]), 308 (SPARC [90]) and 366 (experimental, listed in ChemSpider [27]). These values show a range of over 100 K, indicating that the boiling point is an unreliable criterion for this compound or even class of compounds. Eckel and Kind also reported a relationship to calculate the boiling point from the LRI [58], which gives BP = 267 °C for phthalimide, which is within the range shown in Table 27.

The structures demonstrated in Table 25 and Table 27 show that the energy values, like the mass spectral match values, appear to be very example specific. Although the energies of all structures are below the 90 % quantile (see Section 5.4), those which would be considered visually to be unfavourable and less likely to exist have a significantly higher energy than the confirmed structures. Thus it is likely that a less

conservative exclusion value could be used, for example use of the 80 % quantile would have excluded all but the two confirmed structures based on ChemBio3D calculations – although the 80 % MOLGEN-QSPR quantile would not have excluded any of these structures. A counter-example is already present, however. Figure 25 shows that use of the 80 % quantile to exclude structures based on energy would exclude the correct structure for Structure 26 from consideration in the  $C_{12}H_{10}O_2$  example (Section 5.4). As mentioned above, a bigger set of randomly selected molecules or more iterations in the energy calculation may improve the MOLGEN-QSPR results. Furthermore, an investigation into the correlation with the number of atoms, elements or other structural properties may lead to the development of more specific energy regressions which could then be used to determine appropriate energy boundaries based on the molecular formula. As the calculations are very quick, this would not come at a high computational cost.

The application of this method on more samples will give further evidence for the use of the energy criterion coupled with the other criteria for the progressive exclusion of candidates.

Although subsequent mutagenicity testing showed that the compounds identified here were not associated with the mutagenicity in the sample, the only two significant peaks of interest detected in the mutagenic samples based on GC-MS measurements were confirmed analytically following identification using the methods developed during this study. Identification efforts of mutagenic compounds based on LC-MS/MS measurements are continuing and will be presented elsewhere [12].

## 7 Summary and Future Work

The methods developed throughout this thesis provide a viable alternative to database searching in order to identify unknown compounds measured using GC-EI-MS. Although developed specifically for the spectra resulting from EDA studies, these methods are applicable to any GC-EI-MS data set and thus fill a gap in spectral identification where spectra are not present in a given database. The methods have been used successfully to identify three unknown compounds (Sections 6.2 and 6.3) and also provided indications for the identities of over 52 unknowns (Section 6.1 and [6]). As such, this provides a valuable enhancement to the identification of unknown compounds in complex environmental samples based on GC-EI-MS.

The strength of the method here lies in the combination of information from different sources and the compilation of this in a central summary file, to allow easy selection or exclusion of candidates based on the actual analytical information. The use of mass spectral substructure classifiers was shown to be essential in limiting the number of candidates generated up front. The partitioning information ( $\log K_{ow}$ ) obtained during RP-HPLC fractionation of EDA samples was a valuable exclusion criterion, as was the Lee Retention Index – Boiling Point (LRI-BP) correlation. These criteria often led to the reduction of candidates following structure generation with substructure information by an order of magnitude or more, despite the relatively large error margins associated with the EPISuite<sup>TM</sup> predictions. The incorporation of more accurate retention behaviour prediction would improve the exclusion of candidates, however a compromise between broadly applicable models and more accurate but limited-domain models needs to be found when using structure generation techniques. Focus on more chromatographically-relevant parameters such as the Linear Solvation Energy Relationships (LSERs) developed by Abraham (e.g. [91-93]) may likewise improve candidate selection. Improvements to parameter prediction are the subject of current research (e.g. [94-96]). The criterion based on steric energy was instrumental in reducing structure candidates and specifically eliminating those one would consider highly unlikely visually. The results presented in Section 6.3 also indicate that the energy criterion could be optimised further in the future, for example by correlating these values with the number of atoms or elements present or other structural features.

Improvements to the use of mass spectral fragment prediction in candidate selection would strengthen the overall method further. The results presented in this work show that it is a very subjective criterion, useful in some cases but not in others and that incorporating additional fragmentation mechanisms during fragmentation prediction decreases the selectivity of the fragments, despite the increase in spectral match. It is possible that instead of adding more reaction mechanisms to generate fragments, a

different strategy in fragment prediction is needed. The programs FiD [52] and MetFrag [53], designed for accurate mass tandem MS data, use bond dissociation energies, rather than fragmentation rules to predict fragments. Initial results for both are promising compared with Mass Frontier, however a very limited evaluation of FiD on unit mass data by the authors showed a distinct drop in selectivity due to the dramatic rise in candidate fragments [52]. Validation of this approach compared with rule-based fragment prediction could lead to an improvement in candidate selection, for both unit mass and accurate mass measurements.

Another potential improvement to candidate selection based on mass spectral fragments could involve optimisation of the match value used (see Equation 1). The results of Stein and Scott [55], see Equation 4, indicate that a power in the mass term can be used to weight high  $m/z$  peaks compared with lower ones, with positive effect on library search results. However, attempts to optimise the match value for the predicted-experimental spectrum match thus far has no great effect on the outcomes (see e.g. [54], in German).

The incorporation of toxicity information in the process to assist in either selection of candidates exhibiting effect in samples (see Section 6.1 and [6]) or for structure elimination is still difficult. The toxicity is dependant on many factors including the biotest used, the concentration of the substance, the partitioning behaviour and potential substructures associated with excess toxicity. The results from Meinert et al. [6] confirm that more work is needed to improve the incorporation of toxicity information into unknown identification. Many quantitative structure-activity relationships (QSARs) have been published for different activities (e.g. [25, 68, 97]) and connecting the information from the biotests used in EDA into the identification of unknown structures potentially responsible for the effects will be an important step forward.

The results of Section 6.3 confirm the ideas reviewed in Section 2 that many compounds of relevance in the environment cannot be analysed using GC-EI-MS. An extension of the methods developed here to LC-MS<sup>(n)</sup> based systems would be of distinct advantage in the identification of environmentally relevant compounds. One of the major hurdles blocking progress in this direction is the lack of substructural classifiers for non-EI-MS techniques. Although exact mass data provides information about the substructures present, this is not yet associated with the probabilities that formed an integral part of the methods described in this thesis. On the other hand, the different fragmentation strategies available for accurate mass tandem MS data (e.g. MetFrag and FiD) may provide a viable alternative to the rule-based fragmentation prediction evaluated in Section 4.

A possible enhancement to the GC-MS methods could be the incorporation of softer ionisation and higher accuracy methods for GC to provide complementary information supporting the identification. The use of MS<sup>(n)</sup> techniques, coupled either to GC or LC,

avoid the problem of molecular formula determination, which is a common pitfall in the identification of unknowns based on EI-MS alone. Accurate mass GC-MS data would also allow application and thus verification of the different fragmentation strategies above. The work presented here clearly shows that the more information that is available for use in candidate selection, the better the chances of getting a reliable identification at the end.

The current method provides a valuable contribution to the identification of unknown organic contaminants of environmental significance. The areas for future research as well as an extension into LC-based techniques will strengthen the use of structure generation techniques for unknown identification further.

## 8 Acknowledgements

Firstly I would like to thank Dr. Werner Brack for providing me with the opportunity to work in the active and multi-disciplinary Department for Effect-Directed Analysis at the Helmholtz Centre for Environmental Research (UFZ), Leipzig, Germany, with a reputation far exceeding the Department size. His continuous encouragement and support in participating in all facets of scientific life has made this whole thesis possible.

Secondly, my thanks go to Prof. Dr. Gerrit Schüürmann, Department of Ecological Chemistry, UFZ and Faculty for Chemistry and Physics at the Technischen Universität Bergakademie Freiberg for the opportunity to submit my thesis to TU Freiberg.

Special thanks go to Markus Meringer for the continuing development of the MOLGEN programs in his free time and the constant willingness to answer questions and help develop little ideas into results. We would not have got this far without his help. Thank you also to Prof. Adalbert Kerber and co-workers for the development and use of the programs.

To the co-authors of the papers that formed the sections of this thesis, thank you for your efforts (in alphabetical order): Mahmoud Bataineh, Christine Gallampois, Kai-Uwe Goss, Jos Hermans, Conny Meinert, Tobias Schulze and Sara Weiss. To those who contributed measurements, data, ideas and comments to the papers and parts of this thesis, likewise thank you: Marion Heinrich, Martin Krauss, Stan Schymanski, Peter von der Ohe and Nadin Ulrich. Thank you also goes to the anonymous reviewers whose feedback improved the manuscripts.

The work in this thesis was supported by the European Commission through the Integrated Project MODELKEY (Contract-No. 511237-GOCE). Thanks go especially to the MODELKEY Project Manager, Dr. Michaela Hein and the members of Sub-project KEYTOX for their feedback and encouragement, as well as to all other MODELKEYs.

To all those within the Department of Effect-Directed Analysis who are not already mentioned above, thank you for making the Department a great place to be: Andrea, Angela, Anja, Britta, Christine H., Cynthia, Eva, Fabian, Georg, Ines, Ivonne, Karsten, Katrin, Mareen, Margit, Nicole, Rene, Steffi R., Steffi II, Thomas, Urte and all other guests. To my office mates, building companions, fellow UDO members and UFZ employees, who are too many to name: you know who you are – thank you.

The many contributors to open source software and information such as OpenBabel and Wikipedia should also be acknowledged.

Finally, thanks to my family and friends for their support, especially Stan and Angus for sacrificing countless hours together for ‘the thesis’.

## 9 References

1. Schulze, T., Weiss, S., Schymanski, E., von der Ohe, P.C., Schmitt-Jansen, M., Altenburger, R., Streck, H.-G. and Brack, W. (2010). Confirmation of identity and phytotoxicity of a photo-transformation product of diclofenac. *Environmental Pollution*, 158(5) p. 1461-1466.
2. Helbling, D.E., Hollender, J., Kohler, H.P.E., Singer, H. and Fenner, K. (2010). High-Throughput Identification of Microbial Transformation Products of Organic Micropollutants. *Environmental Science & Technology*, 44(17) p. 6621-6627.
3. Schymanski, E., Bataineh, M., Goss, K.-U. and Brack, W. (2009). Integrated Analytical and Computer Tools for Structure Elucidation in Effect-Directed Analysis. *TrAC Trends in Analytical Chemistry*, 28(5) p. 550-561.
4. Schymanski, E., Schulze, T., Hermans, J. and Brack, W. (2011) *Chapter 8: Computer Tools for Structure Elucidation in EDA*, in *Handbook of Environmental Chemistry: Effect-Directed Analysis of Complex Environmental Samples*, Vol. 15, Ed. W. Brack, Springer-Verlag, Germany.
5. Schymanski, E.L., Meinert, C., Meringer, M. and Brack, W. (2008). The use of MS classifiers and structure generation to assist in the identification of unknowns in effect-directed analysis. *Analytica Chimica Acta*, 615(2) p. 136-147.
6. Meinert, C., Schymanski, E., Kuster, E., Kuhne, R., Schuurmann, G. and Brack, W. (2010). Application of preparative capillary gas chromatography (pcGC), automated structure generation and mutagenicity prediction to improve effect-directed analysis of genotoxicants in a contaminated groundwater. *Environmental Science and Pollution Research*, 17(4) p. 885-897.
7. Kerber, A., Meringer, M. and Rücker, C. (2006). CASE via MS: Ranking structure candidates by mass spectra. *Croatica Chemica Acta*, 79(3) p. 449-464.
8. Schymanski, E., Meringer, M. and Brack, W. (2009). Matching Structures to Mass Spectra using Fragmentation Patterns - Are the results as good as they look? *Analytical Chemistry*, 81(9) p. 3608-3617.
9. Schymanski, E., Meringer, M. and Brack, W. (2011) Automated Strategies to Identify Compounds on the Basis of GC/EI-MS and Calculated Properties. *Analytical Chemistry*, 83(3) p.903-912.
10. Schymanski, E., Gallampois, G. and Brack, W. Identification of Unknown GC-EI-MS Spectra in Elbe EDA Study. *In preparation*.
11. Gallampois, C., Bataineh, M. and Brack, W. Development of a reverse-phase fractionation method for planar polar mutagens in river waters. *In preparation*.
12. Gallampois, C., Schymanski, E., Bataineh, M., Krauss, M. and Brack, W. Development of an HPLC-ESI/APCI-Orbitrap-MS/MS method and strategy to identify planar polar mutagens in river waters. *In preparation*.
13. Brack, W. (2003). Effect-directed analysis: a promising tool for the identification of organic toxicants in complex mixtures? *Analytical and Bioanalytical Chemistry*, 377(3) p. 397-407.
14. Brack, W., Schmitt-Jansen, M., Machala, M., Brix, R., Barcelo, D., Schymanski, E., Streck, G. and Schulze, T., (2008). How to confirm identified toxicants in effect-directed analysis. *Analytical and Bioanalytical Chemistry*, 390(8) p. 1959-1973.
15. Hewitt, L.M. and Marvin, C.H. (2005). Analytical methods in environmental effects-directed investigations of effluents. *Mutation Research-Reviews in Mutation Research*, 589(3) p. 208-232.
16. Brack, W., Klammer, H.J.C., de Ada, M.L. and Barcelo, D. (2007). Effect-directed analysis of key toxicants in European river basins - A review. *Environmental Science and Pollution Research*, 14(1) p. 30-38.
17. NIST/EPA/NIH. (2005) *NIST Mass Spectral Library*. National Institute of Standards and Technology, US Secretary of Commerce: USA.
18. Wiley. (2006) *Wiley Registry of Mass Spectral Data 8th Edition*. Wiley: New York, USA.
19. Liao, W.T., Draper, W.M. and Perera, S.K. (2008). Identification of unknowns in atmospheric pressure ionization mass spectrometry using a mass to structure Search Engine. *Analytical Chemistry*, 80(20) p. 7765-7777.
20. Thurman, E.M., Ferrer, I., Zweigenbaum, J.A., Garcia-Reyes, J.F., Woodman, M. and Fernandez-Alba, A.R. (2005). Discovering metabolites of post-harvest fungicides in citrus with liquid chromatography/time-of-flight mass spectrometry and ion trap tandem mass spectrometry. *Journal of Chromatography A*, 1082(1) p. 71-80.



21. Kosjek, T. and Heath, E. (2008). Applications of mass spectrometry to identifying pharmaceutical transformation products in water treatment. *TrAC Trends in Analytical Chemistry*, 27(10) p. 807-820.
22. Sauvage, F.L., Saint-Marcoux, F., Duretz, B., Deporte, D., Lachatre, G. and Marquet, P. (2006). Screening of drugs and toxic compounds with liquid chromatography-linear ion trap tandem mass spectrometry. *Clinical Chemistry*, 52(9) p. 1735-1742.
23. Polettini, A., Gottardo, R., Pascali, J.P. and Tagliaro, F. (2008). Implementation and performance evaluation of a database of chemical formulas for the screening of pharmaco/toxicologically relevant compounds in biological samples using electrospray ionization-time-of-flight mass spectrometry. *Analytical Chemistry*, 80(8) p. 3050-3057.
24. Hao, H.P., Cui, N., Wang, G.J., Xiang, B.R., Liang, Y., Xu, X.Y., Zhang, H., Yang, J., Zheng, C.N., Wu, L., Gong, P. and Wang, W. (2008). Global Detection and Identification of Nontarget Components from Herbal Preparations by Liquid Chromatography Hybrid Ion Trap Time-of-Flight Mass Spectrometry and a Strategy. *Analytical Chemistry*, 80(21) p. 8187-8194.
25. Kazius, J., McGuire, R. and Bursi, R. (2005). Derivation and validation of toxicophores for mutagenicity prediction. *Journal of Medicinal Chemistry*, 48(1) p. 312-320.
26. Muir, D.C.G. and Howard, P.H. (2006). Are there other persistent organic pollutants? A challenge for environmental chemists. *Environmental Science & Technology*, 40(23) p. 7157-7166.
27. Royal Society of Chemistry. *ChemSpider* <http://www.chemspider.com>. Accessed 16/03/2010.
28. Lehotay, S.J., Mastovska, K., Amirav, A., Fialkov, A.B., Martos, P.A., Kok, A.d. and Fernández-Alba, A.R. (2008). Identification and confirmation of chemical residues in food by chromatography-mass spectrometry and other techniques. *TrAC Trends in Analytical Chemistry*, 27(11) p. 1070-1090.
29. Kerber, A., Laue, R., Meringer, M. and Varmuza, K., (2001). MOLGEN-MS: Evaluation of low resolution electron impact mass spectra with MS classification and exhaustive structure generation. *Advances in Mass Spectrometry*, 15 p. 939-940.
30. ACD (2007) *MS Manager*. Version 11.01, Advanced Chemistry Development, Inc, Toronto, Canada.
31. Kind, T. and Fiehn, O. (2006). Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics*, 7.
32. Kind, T. and Fiehn, O. (2007). Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics*, 8.
33. Meringer, M., Reinker, S., Zhang, J. and Muller, A. (2011). MS/MS Data Improves Automated Determination of Molecular Formulas by Mass Spectrometry. *Match - Communications in Mathematical and in Computer Chemistry*, 65(2) p. 259-290.
34. Benecke, C., Grüner, T., Kerber, A., Laue, R. and Wieland, T. (1997). MOLEcular structure GENeration with MOLGEN, new features and future developments. *Fresenius Journal of Analytical Chemistry*, 359(1) p. 23-32.
35. Kerber, A., Laue, R., Grüner, T. and Meringer, M. (1998). Molgen 4.0. *Match - Communications in Mathematical and in Computer Chemistry*, 37 p. 205-208.
36. Gugisch, R., Kerber, A., Kohnert, A., Laue, R., Meringer, M., Rücker, C. and Wassermann, A. *MOLGEN 5.0*, [www.molgen.de](http://www.molgen.de). Accessed 02/08/2010.
37. McLafferty, F.W. and Turecek, F. (1993), *Interpretation of Mass Spectra*. Mill Valley, California. USA. University Science Books.
38. Munk, M.E. (1998). Computer-based structure determination: Then and now. *Journal of Chemical Information and Computer Sciences*, 38(6) p. 997-1009.
39. Varmuza, K., Stancl, F., Lohninger, H. and Werther, W. (1996). Short Communication: Automatic recognition of substance classes from data obtained by gas chromatography/mass spectroscopy. *Laboratory Automation and Information Management*, 31 p. 225-230.
40. Varmuza, K. and Werther, W., (1996). Mass spectral classifiers for supporting systematic structure elucidation. *Journal of Chemical Information and Computer Sciences*, 36(2) p. 323-333.
41. NIST (2005) *Automated Mass Spectral Deconvolution and Identification System (AMDIS)*. National Institute of Standards and Technology (NIST), US Department of Defense: USA.
42. Stein, S.E. (1995). Chemical Substructure Identification by Mass-Spectral Library Searching. *Journal of the American Society for Mass Spectrometry*, 6(8) p. 644-655.
43. Kind, T. and Fiehn, O. (2010). Advances in structure elucidation of small molecules using mass spectrometry. *Bioanalytical Reviews*, 2 p.23-60.
44. HighChem (2007) *Mass Frontier*. Version 5.0. HighChem Ltd./Thermo Scientific, Bratislava, Slovakia.

45. Fan, B.T., Chen, H.F., Petitjean, M., Panaye, A., Doucet, J.P., Xia, H.R. and Yuan, S.G. (2005). New strategy of mass spectrum simulation based on reduced and concentrated knowledge databases. *Spectroscopy Letters*, 38(2) p. 145-170.
46. Gasteiger, J., Hanebeck, W. and Schulz, K.P. (1992). Prediction of Mass-Spectra from Structural Information. *Journal of Chemical Information and Computer Sciences*, 32(4) p. 264-271.
47. Hill, A.W. and Mortishire-Smith, R.J. (2005). Automated assignment of high-resolution collisionally activated dissociation mass spectra using a systematic bond disconnection approach. *Rapid Communications in Mass Spectrometry*, 19 p. 3111-3118.
48. Meringer, M. (2009) *MOLGEN-MSF*, [www.molgen.de/documents/MolgenMsf.pdf](http://www.molgen.de/documents/MolgenMsf.pdf). M. Meringer, Munich, Germany.
49. HighChem (2007) *Mass Frontier User Information* <http://www.highchem.com/massfrontier/mass-frontier.html>. Accessed 11/11/2010.
50. ACD (2007), *ACD/MS Manager & Processor Reference Manual (Version 11.0)*: Advanced Chemistry Development, Inc, Toronto, Canada.
51. Hill, D.W., Kertesz, T.M., Fontaine, D., Friedman, R. and Grant, D.F. (2008). Mass spectral metabonomics beyond elemental formula: Chemical database querying by matching experimental with computational fragmentation spectra. *Analytical Chemistry*, 80(14) p. 5574-5582.
52. Heinonen, M., Rantanen, A., Mielikainen, T., Kokkonen, J., Kiuru, J., Ketola, R.A. and Rousu, J. (2008). FiD: a software for ab initio structural identification of product ions from tandem mass spectrometric data. *Rapid Communications in Mass Spectrometry*, 22(19) p. 3043-3052.
53. Wolf, S., Schmidt, S., Müller-Hannemann, M. and Neumann, S. (2010). In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics*, 11 p. 148.
54. Meringer, M. (2004), *Mathematical Models for Combinatorial Chemistry and Molecular Structure Elucidation*. Berlin, Germany: Logos-Verlag.
55. Stein, S.E. and Scott, D.R. (1994). Optimization and Testing of Mass-Spectral Library Search Algorithms for Compound Identification. *Journal of the American Society for Mass Spectrometry*, 5(9) p. 859-866.
56. Rostad, C.E. and Pereira, W.E. (1986). Kovats and Lee Retention Indexes Determined by Gas Chromatography/Mass Spectrometry for Organic-Compounds of Environmental Interest. *Journal of High Resolution Chromatography & Chromatography Communications*, 9(6) p. 328-334.
57. Thomas, K., Balaam, J., Brack, W., Brix, R., Hamers, T., Hermans, J., Lamoree, M., Leonards, P., Muusse, M., Schulze, T., Schymanski, E., Streck, G. and Weiss, J. Effects-directed analysis: tools for the improved detection of unknowns. *In preparation*.
58. Eckel, W.P. and Kind, T. (2003). Use of boiling point-Lee retention index correlation for rapid review of gas chromatography-mass spectrometry data. *Analytica Chimica Acta*, 494(1-2) p. 235-243.
59. Stein, S.E., Babushok, V.I., Brown, R.L. and Linstrom, P.J. (2007). Estimation of Kovats retention indices using group contributions. *Journal of Chemical Information and Modeling*, 47(3) p. 975-980.
60. USEPA, (2007) *Estimation Program Interface (EPI) Suite (TM)*, V3.20. United States Environmental Protection Agency, USA.
61. OECD (2004). Guideline for the testing of chemicals 117. Partition coefficient (n-octanol/water) - High performance liquid chromatography (HPLC) method.
62. Paschke, A., Manz, M. and Schuurmann, G. (2001). Application of different RP-HPLC methods for the determination of the octanol/water partition coefficient of selected tetrachlorobenzyltoluenes. *Chemosphere*, 45(6-7) p. 721-728.
63. Kind, T. (2003) *Combination of GC-MS and Chemometrics for the Analysis of Compounds in Complex Environmental Samples (in German)*, Thesis, Chemistry and Mineralogy, University of Leipzig: Leipzig, Germany.
64. Allinger, N.L. (1977). Conformational-Analysis .130. MM2 - Hydrocarbon Force-Field Utilizing V1 and V2 Torsional Terms. *Journal of the American Chemical Society*, 99(25) p. 8127-8134.
65. Kerber, A., Laue, R., Meringer, M. and Rücker, C. (2005). Molecules in silico: Potential versus known organic compounds. *Match-Communications in Mathematical and in Computer Chemistry*, 54(2) p. 301-312.
66. CambridgeSoft (2007) *ChemBio3D*. Version Ultra 11.0. CambridgeSoft, USA.
67. Kerber, A., Laue, R., Meringer, M. and Rucker, C. (2004). Molgen-QSPR, a software package for the study of quantitative structure property relationships. *Match-Communications in Mathematical and in Computer Chemistry*, 51 p. 187-204.
68. von der Ohe, P.C., Kuhne, R., Ebert, R.U., Altenburger, R., Liess, M. and Schuurmann, G., (2005). Structural alerts - A new classification model to discriminate excess toxicity from narcotic

- effect levels of organic compounds in the acute daphnid assay. *Chemical Research in Toxicology*, 18(3) p. 536-555.
69. D'Arcy, P. and Mallard, W.G. (2004) *AMDIS - User Guide*. National Institute of Standards and Technology (NIST), U.S. Department of Commerce, USA.
  70. MathWorks (2006) *MATLAB*. The MathWorks Inc. USA.
  71. OpenBabel (2007) *Open Babel*. <http://openbabel.sourceforge.net>. Accessed 11/11/2010.
  72. Finizio, A., Vighi, M. and Sandroni, D. (1997). Determination of N-octanol/water partition coefficient (Kow) of pesticide critical review and comparison of methods. *Chemosphere*, 34(1) p. 131-161.
  73. Eadsforth, C.V. and Moser, P. (1983). Assessment of Reverse-Phase Chromatographic Methods for Determining Partition-Coefficients. *Chemosphere*, 12(11-1) p. 1459-1475.
  74. Dalby, A., Nourse, J.G., Hounshell, W.D., Gushurst, A.K.I., Grier, D.L., Leland, B.A. and Laufer, J. (1992). Description of Several Chemical-Structure File Formats Used by Computer-Programs Developed at Molecular Design Limited. *Journal of Chemical Information and Computer Sciences*, 32(3) p. 244-255.
  75. Symyx Technologies, Inc. (2007) *CTFile Formats*. <http://www.symyx.com/downloads/public/ctfile/ctfile.pdf>. Accessed 02/03/2009.
  76. NCBI. (2010) *PubChem* <http://pubchem.ncbi.nlm.nih.gov/>. National Center for Biotechnology Information, USA. Accessed 16/03/2010.
  77. Elyashberg, M.E., Williams, A. and Martin, G.E. (2008). Computer-assisted structure verification and elucidation tools in NMR-based structure elucidation. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 53(1-2) p. 1-104.
  78. Krompiec, M. and Patiny, L. (2003) *ChemCalc* <http://www.chemcalc.org>. Accessed 17/03/2010.
  79. ThermoFisher (2008) *Xcalibur 2.0.7 SP1*. Thermo Fisher Scientific Inc.
  80. Goodman, J.M. (2009) *University of Cambridge Molecular Formula Search* <http://www-jmg.ch.cam.ac.uk/tools/magnus/EadFormW.html>. Accessed 17/03/2010.
  81. Stein, S.E. and Brown, R.L. (1994). Estimation of Normal Boiling Points from Group Contributions. *Journal of Chemical Information and Computer Sciences*, 34(3) p. 581-587.
  82. Python (2006) *Python* [www.python.org](http://www.python.org). Python Software Foundation.
  83. Schüürmann, G., Kühne, R., Kleint, F., Ebert, R.-U., Rothenbacher, C. and Herth, P. (1997) *A software system for automatic chemical property estimation from molecular structure*, in *QSAR in Environmental Science*, F. Chen and G. Schüürmann, Editors., SETAC Press: Pensacola, FL; USA. p. 93-114.
  84. Schüürmann, G., Ebert, R.U., Nendza, M., Dearden, J.C., Paschke, A. and Kühne, R. (2007) *Prediction of fate-related compound properties*, in *Risk Assessment of Chemicals: An Introduction*, K. van Leeuwen and T. Vermeire, Editors., Springer Science: Dordrecht. p. 375-426.
  85. Schmitt-Jansen, M., Bartels, P., Adler, N. and Altenburger, R. (2007). Phytotoxicity assessment of diclofenac and its phototransformation products. *Analytical and Bioanalytical Chemistry*, 387(4) p. 1389-1396.
  86. Aguera, A., Estrada, L.A.P., Ferrer, I., Thurman, E.M., Malato, S. and Fernandez-Alba, A.R. (2005). Application of time-of-flight mass spectrometry to the analysis of phototransformation products of diclofenac in water under natural sunlight. *Journal of Mass Spectrometry*, 40(7) p. 908-915.
  87. Sakamoto, H. and Hayatsu, H. (1990). A simple method for monitoring mutagenicity of river water. Mutagens in Yodo river system, Kyoto. *Bulletins of Environmental Contamination and Toxicology*, 44 p. 521-528.
  88. Hayatsu, H. (1992). Cellulose bearing covalently linked copper phthalocyanine trisulphonate as an adsorbent selective for polycyclic compounds and its use in studies of environmental mutagens and carcinogens. *Journal of Chromatography A*, 597 p. 37-56.
  89. Perez, S., Reifferscheid, G., Eichhorn, P. and Barcelo, D. (2003). Assessment of the mutagenic potency of sewage sludges contaminated with polycyclic aromatic hydrocarbons by an ames fluctuation assay. *Environmental Toxicology and Chemistry*, 22(11) p. 2576-2584.
  90. SPARC (2010) *SPARC Online Calculator* <http://sparc.chem.uga.edu/sparc/>. Accessed 08/09/2010.
  91. Abraham, M.H. (1993). Hydrogen-Bonding .27. Solvation Parameters for Functionally-Substituted Aromatic-Compounds and Heterocyclic-Compounds, from Gas-Liquid-Chromatographic Data. *Journal of Chromatography*, 644(1) p. 95-139.

- 
92. Abraham, M.H., Chadha, H.S., Whiting, G.S. and Mitchell, R.C. (1994). Hydrogen-Bonding .32. An Analysis of Water-Octanol and Water-Alkane Partitioning and the Delta-Log-P Parameter of Seiler. *Journal of Pharmaceutical Sciences*, 83(8) p. 1085-1100.
  93. Abraham, M.H., Ibrahim, A. and Zissimos, A.M. (2004). Determination of sets of solute descriptors from chromatographic measurements. *Journal of Chromatography A*, 1037(1-2) p. 29.
  94. Schwobel, J., Ebert, R.U., Kuhne, R. and Schuurmann, G. (2009). Modeling the H Bond Donor Strength of -OH, -NH, and -CH Sites by Local Molecular Parameters. *Journal of Computational Chemistry*, 30(9) p. 1454-1464.
  95. Schwobel, J., Ebert, R.U., Kuhne, R. and Schuurmann, G. (2009). Prediction of the Intrinsic Hydrogen Bond Acceptor Strength of Chemical Substances from Molecular Structure. *Journal of Physical Chemistry A*, 113(37) p. 10104-10112.
  96. Schwobel, J., Ebert, R.U., Kuhne, R. and Schuurmann, G. (2009). Prediction of the Intrinsic Hydrogen Bond Acceptor Strength of Organic Compounds by Local Molecular Parameters. *Journal of Chemical Information and Modeling*, 49(4) p. 956-962.
  97. Cronin, M.T.D., Aptula, A.O., Dearden, J.C., Duffy, J.C., Netzeva, T.I., Patel, H., Rowe, P.H., Schultz, T.W., Worth, A.P., Voutzoulidis, K. and Schuurmann, G. (2002). Structure-based classification of antibacterial activity. *Journal of Chemical Information and Computer Sciences*, 42(4) p. 869-878.

## 10 Appendix

### *Appendix 1: Additional Tables*

|                                                                                                          |     |
|----------------------------------------------------------------------------------------------------------|-----|
| Table A1: Data for Figure 9, Section 3.2.2                                                               | 115 |
| Table A2: Program settings used for Mass Frontier ('Reaction Restrictions' menu)                         | 117 |
| Table A3: AutoAssignment settings used for ACD MS Fragmenter (MS Manager)                                | 117 |
| Table A4: Match Values and RRP's for 100 Spectra, Section 4.2                                            | 118 |
| Table A5: Summary of MOLGEN-MS results for the 29 C <sub>12</sub> H <sub>10</sub> O <sub>2</sub> isomers | 120 |

**Table A1: Data for Figure 9, Section 3.2.2. The unknown spectra were sorted by the number of structures generated in Run 1 and then assigned a spectrum number. The molecular formula was that calculated as the best fit using MOLGEN-MS. Further details are in the text.**

| Spectrum Number | Molecular Formula                                             | Run 1<br>No classifiers | Run 2<br>Varmuza<br>classifiers 95 % | Run 3<br>Varmuza & NIST<br>classifiers (95 %) | Number of<br>NIST spectra<br>with formula |
|-----------------|---------------------------------------------------------------|-------------------------|--------------------------------------|-----------------------------------------------|-------------------------------------------|
| 1               | C <sub>18</sub> H <sub>35</sub> NO                            | >100,000,000            | >50,000                              | 13,033                                        | 6                                         |
| 2               | C <sub>10</sub> H <sub>10</sub> O <sub>4</sub>                | >100,000,000            | >50,000                              | 32                                            | 50                                        |
| 3               | C <sub>12</sub> H <sub>10</sub>                               | 37,720,012              | 1                                    | 188                                           | 6                                         |
| 4               | C <sub>8</sub> H <sub>7</sub> ClO <sub>2</sub>                | 5,160,746               | 3                                    | 67                                            | 27                                        |
| 5               | C <sub>9</sub> H <sub>8</sub> Cl <sub>4</sub>                 | 1,678,835               | 435                                  | 106                                           | 2                                         |
| 6               | C <sub>8</sub> H <sub>10</sub> S <sub>2</sub>                 | 607,376                 | 80                                   | 80                                            | 5                                         |
| 7               | C <sub>7</sub> H <sub>5</sub> ClO <sub>2</sub>                | 507,196                 | >50,000                              | 3                                             | 10                                        |
| 8               | C <sub>10</sub> H <sub>16</sub> O                             | 452,458                 | >50,000                              | 483                                           | 217                                       |
| 9               | C <sub>10</sub> H <sub>16</sub> O                             | 452,458                 | >50,000                              | 648                                           | 217                                       |
| 10              | C <sub>7</sub> H <sub>5</sub> Cl <sub>3</sub> O               | 255,964                 | 0 / 65                               | 65                                            | 7                                         |
| 11              | C <sub>7</sub> H <sub>5</sub> Cl <sub>3</sub> O               | 255,964                 | 0 / 65                               | 15                                            | 7                                         |
| 12              | C <sub>5</sub> H <sub>7</sub> Cl <sub>2</sub> NOS             | 218,339                 | 33,882                               | 5,054                                         | 0                                         |
| 13              | C <sub>4</sub> H <sub>11</sub> O <sub>2</sub> PS <sub>2</sub> | 157,770                 | 64                                   | 156                                           | 1                                         |
| 14              | C <sub>7</sub> H <sub>6</sub> Cl <sub>2</sub> O               | 155,987                 | 0 / 31                               | 6                                             | 13                                        |
| 15              | C <sub>7</sub> H <sub>6</sub> Cl <sub>2</sub> O               | 155,987                 | 19                                   | 19                                            | 31                                        |
| 16              | C <sub>7</sub> H <sub>6</sub> Cl <sub>2</sub> O               | 155,987                 | 0 / 26 / 31                          | 31                                            | 31                                        |
| 17              | C <sub>7</sub> H <sub>6</sub> Cl <sub>2</sub> O               | 155,987                 | 31 / 26                              | 57                                            | 31                                        |
| 18              | C <sub>7</sub> H <sub>6</sub> Cl <sub>2</sub> O               | 155,987                 | 31 / 26                              | 57                                            | 31                                        |
| 19              | C <sub>7</sub> H <sub>6</sub> Cl <sub>2</sub> O               | 155,987                 | 31 / 26                              | 57                                            | 31                                        |
| 20              | C <sub>8</sub> H <sub>10</sub> S                              | 69,669                  | 18                                   | 18                                            | 21                                        |
| 21              | C <sub>7</sub> H <sub>7</sub> ClO                             | 62,643                  | 0                                    | 6                                             | 14                                        |
| 22              | C <sub>7</sub> H <sub>7</sub> ClO                             | 62,643                  | 0                                    | 3                                             | 14                                        |
| 23              | C <sub>3</sub> H <sub>9</sub> O <sub>2</sub> PS <sub>2</sub>  | 27,776                  | 12,391                               | 1,471                                         | 2                                         |
| 24              | C <sub>3</sub> H <sub>9</sub> O <sub>2</sub> PS <sub>2</sub>  | 27,776                  | 12                                   | 27                                            | 2                                         |
| 25              | C <sub>10</sub> H <sub>16</sub>                               | 24,938                  | 726                                  | 726                                           | 161                                       |
| 26              | C <sub>6</sub> H <sub>8</sub> OS                              | 20,610                  | 6,182                                | 174                                           | 9                                         |
| 27              | C <sub>6</sub> H <sub>3</sub> Cl <sub>3</sub> O               | 19,969                  | 0 / 12                               | 12                                            | 6                                         |
| 28              | C <sub>3</sub> H <sub>9</sub> O <sub>3</sub> PS               | 19,054                  | 18,511                               | 122                                           | 2                                         |
| 29              | C <sub>3</sub> H <sub>9</sub> O <sub>3</sub> PS               | 19,054                  | 10,530                               | 48                                            | 2                                         |
| 30              | C <sub>3</sub> H <sub>9</sub> OPS <sub>3</sub>                | 19,054                  | 13,131                               | 4,932                                         | 1                                         |
| 31              | C <sub>4</sub> H <sub>8</sub> O <sub>2</sub> S <sub>2</sub>   | 6,795                   | 12                                   | 5,540                                         | 4                                         |
| 32              | C <sub>5</sub> H <sub>8</sub> Cl <sub>2</sub> O <sub>2</sub>  | 5,459                   | 5,351                                | 264                                           | 14                                        |
| 33              | C <sub>3</sub> H <sub>6</sub> O <sub>2</sub> S <sub>3</sub>   | 5,031                   | 2,170                                | 831                                           | 0                                         |
| 34              | C <sub>3</sub> H <sub>7</sub> NOS <sub>2</sub>                | 3,838                   | 71                                   | 3,811                                         | 0                                         |
| 35              | C <sub>6</sub> H <sub>12</sub> Cl <sub>2</sub> O <sub>2</sub> | 3,576                   | 231                                  | 75                                            | 4                                         |
| 36              | C <sub>4</sub> H <sub>9</sub> NOS                             | 3,095                   | 155                                  | 68                                            | 6                                         |
| 37              | C <sub>6</sub> H <sub>6</sub> O                               | 2,237                   | 0 / 1                                | 1                                             | 4                                         |
| 38              | C <sub>4</sub> H <sub>3</sub> Cl <sub>3</sub> O <sub>2</sub>  | 2,080                   | 1                                    | 1                                             | 1                                         |
| 39              | C <sub>5</sub> H <sub>6</sub> S <sub>2</sub>                  | 1,938                   | 1,936                                | 2                                             | 3                                         |
| 40              | C <sub>3</sub> H <sub>6</sub> O <sub>2</sub> S <sub>2</sub>   | 1,333                   | 1,311                                | 183                                           | 1                                         |
| 41              | C <sub>6</sub> H <sub>4</sub> Cl <sub>2</sub>                 | 1,323                   | 3                                    | 3                                             | 3                                         |
| 42              | C <sub>10</sub> H <sub>20</sub>                               | 852                     | 852                                  | 851                                           | 110                                       |
| 43              | C <sub>13</sub> H <sub>28</sub>                               | 802                     | 162                                  | 162                                           | 71                                        |
| 44              | C <sub>4</sub> H <sub>7</sub> F <sub>3</sub> S <sub>2</sub>   | 551                     | 38                                   | 38                                            | 2                                         |
| 45              | C <sub>3</sub> H <sub>6</sub> S <sub>4</sub>                  | 263                     | 263*                                 | 139                                           | 1                                         |
| 46              | C <sub>5</sub> H <sub>6</sub> Cl <sub>4</sub>                 | 217                     | 217                                  | 57                                            | 0                                         |
| 47              | C <sub>3</sub> H <sub>6</sub> S <sub>3</sub>                  | 102                     | 69                                   | 69                                            | 3                                         |

| Spectrum Number | Molecular Formula                               | Run 1<br>No classifiers | Run 2<br>Varmuza<br>classifiers 95 % | Run 3<br>Varmuza & NIST<br>classifiers (95 %) | Number of<br>NIST spectra<br>with formula |
|-----------------|-------------------------------------------------|-------------------------|--------------------------------------|-----------------------------------------------|-------------------------------------------|
| 48              | C <sub>4</sub> H <sub>10</sub> S <sub>3</sub>   | 88                      | 88                                   | 88                                            | 3                                         |
| 49              | C <sub>2</sub> H <sub>4</sub> S <sub>5</sub>    | 88                      | 88*                                  | 10                                            | 1                                         |
| 50              | C <sub>5</sub> H <sub>12</sub> S <sub>2</sub>   | 69                      | 42                                   | 42                                            | 11                                        |
| 51              | C <sub>2</sub> H <sub>4</sub> O <sub>2</sub> S  | 55                      | 15                                   | 2                                             | 1                                         |
| 52              | C <sub>2</sub> H <sub>4</sub> S <sub>4</sub>    | 48                      | 48*                                  | 38*                                           | 2                                         |
| 53              | C <sub>4</sub> H <sub>4</sub> Cl <sub>4</sub>   | 45                      | 45                                   | 45                                            | 0                                         |
| 54              | C <sub>4</sub> H <sub>4</sub> Cl <sub>4</sub>   | 45                      | 45                                   | 45                                            | 0                                         |
| 55              | C <sub>4</sub> H <sub>4</sub> Cl <sub>4</sub>   | 45                      | 45                                   | 45                                            | 0                                         |
| 56              | C <sub>4</sub> H <sub>2</sub> Cl <sub>4</sub>   | 40                      | 25                                   | 4                                             | 2                                         |
| 57              | C <sub>4</sub> H <sub>3</sub> Cl <sub>5</sub>   | 37                      | 37                                   | 33                                            | 0                                         |
| 58              | C <sub>4</sub> H <sub>3</sub> Cl <sub>5</sub>   | 37                      | 37                                   | 37                                            | 0                                         |
| 59              | C <sub>4</sub> H <sub>3</sub> Cl <sub>5</sub>   | 37                      | 37                                   | 33                                            | 0                                         |
| 60              | C <sub>4</sub> H <sub>3</sub> Cl <sub>5</sub>   | 37                      | 37                                   | 37                                            | 0                                         |
| 61              | C <sub>4</sub> H <sub>3</sub> Cl <sub>5</sub>   | 37                      | 37                                   | 37                                            | 0                                         |
| 62              | C <sub>2</sub> H <sub>6</sub> S <sub>5</sub>    | 35                      | 28*                                  | 11                                            | 1                                         |
| 63              | C <sub>3</sub> H <sub>6</sub> S <sub>3</sub>    | 28                      | 28                                   | 11                                            | 1                                         |
| 64              | C <sub>2</sub> H <sub>6</sub> S <sub>4</sub>    | 20                      | 16                                   | 16                                            | 1                                         |
| 65              | C <sub>2</sub> H <sub>6</sub> S <sub>4</sub>    | 20                      | 16                                   | 6                                             | 1                                         |
| 66              | C <sub>2</sub> H <sub>6</sub> S <sub>3</sub>    | 10                      | 8                                    | 8                                             | 1                                         |
| 67              | C <sub>3</sub> H <sub>5</sub> BrCl <sub>2</sub> | 9                       | 9                                    | 9                                             | 2                                         |
| 68              | C <sub>3</sub> H <sub>4</sub> Cl <sub>2</sub>   | 7                       | 7                                    | 7                                             | 7                                         |
| 69              | C <sub>3</sub> H <sub>3</sub> Cl <sub>5</sub>   | 5                       | 5                                    | 5                                             | 1                                         |
| 70              | C <sub>2</sub> H <sub>2</sub> Cl <sub>4</sub>   | 2                       | 2                                    | 2                                             | 2                                         |
| 71              | S <sub>8</sub>                                  | 1                       | 1*                                   | 1*                                            | 1                                         |

Notes: (a) > indicates program exceeded limits or generation aborted. For Figures, '>' is taken as '='

(b) 2 or more numbers in one entry indicates different runs for different classifier combinations

(c) \* indicates calculations done using MOLGEN 3.5

**Table A2: Program settings used for Mass Frontier ('Reaction Restrictions' menu).**

| Tab                     | Active Options                                                                                                                                                                                                             |
|-------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Knowledge Base          | 'General Fragmentation Rules ' OR 'Both'<br>Fragmentation Library Options: Active records only, Ignore General Frag. Rules in library reactions, Charge Localization Concept only                                          |
| Ionization and Cleavage | Ionization Method: M <sup>•+</sup> Electron Impact (EI)<br>Ionization on: Non-bond. el (-e-), Pi bond ( $\pi$ ), Sigma bond ( $\sigma$ )<br>Cleavage: Alpha ( $\alpha$ ), Inductive (i)                                    |
| H-Rearrangement         | In Odd-Electron Ion (rH <sub>A</sub> ): Hydrogen transfer from atom: Steric optimal (Recommended)<br>In Even-Electron Ion: Hydrogen transfer from atom: $\alpha$ , $\beta$ (rH <sub>B</sub> ), $\gamma$ (rH <sub>C</sub> ) |
| Resonance               | Resonance Reactions: Electron Sharing (es), Charge Stabilization (cr), Radical Isomerisation (rr)<br>Display Resonance Reactions: No (Recommended)                                                                         |
| Additional              | Allowed on Aromatic System: Ionization, Stabilization, Cleavage<br>Hydrogen Radical Lost: No (Recommended)<br>Allowed Carbo – cation/anion: Primary, Secondary, Tertiary                                                   |
| Sizes                   | Reaction Steps: Max Number: 3 or 5 (depending on 'variation')<br>Mass Range: From 30 to 3000 m/z (values under 30 not accepted)<br>Reactions Limit: Value 10000.                                                           |

**Table A3: AutoAssignment settings used for ACD MS Fragmenter (MS Manager).**

| Tab                    | Active Options                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
|------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Reaction               | Positive Ions - Ionization Type: electron ionization, $\sigma$ -ionization<br>Distonic Ions Formation - hydrogen shift, double bonds cleavage, triple bonds cleavage, saturated rings cleavage<br>Common Reactions – aromatic bonds cleavage, resonance reactions, rings formation, hydride shift<br>Maximum 10,000 fragments generated on each step<br>Number of fragmentation steps: 3 or 5 (default 3)                                                                                         |
| Spectrum               | Mass Range: 1 to 1,000<br>Relative Abundance Range: 0 to 100                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| Specific Fragmentation | Heterolytic Cleavage – odd-electron ions, $\beta$ -distonic with radical site shift, even-electron ions<br>Homolytic Cleavage – odd-electron ions, include H loss, $\delta$ -distonic with radical site shift, even-electron ions, include H loss, protonated ions with hydrogen shift<br>Hydrogen Rearrangements – 1,3-shift, 1,5-shift, McLafferty rearrangement<br>Neutral Losses: NL before ionization, NL after ionization, 1,3-shift, 1,4-shift, 1,5-shift, ejection from cyclic structures |



**Table A4: Match values (Equation 1) of the correct structure and the relative ranking positions (Equation 7) calculated for the 100 spectra for Mass Frontier and MOLGEN-MSF. Program and settings abbreviations are as set out in Table 4. NIST\_no denotes the spectrum number in the NIST database, m the molecular mass (Da) and TC the total number of candidates (constitutional isomers) for that formula. Note here the match value reported ranges between 0 and 1 and is not expressed as percent (default output from MOLGEN-MS, for example).**

| No. | NIST_No. | Formula                                          | m   | TC   | Match Value of correct structure |        |        | Relative Ranking Position |        |        |
|-----|----------|--------------------------------------------------|-----|------|----------------------------------|--------|--------|---------------------------|--------|--------|
|     |          |                                                  |     |      | MF_3st                           | MF_5st | MSF    | MF_3st                    | MF_5st | MSF    |
| 1   | 61627    | C <sub>9</sub> H <sub>16</sub>                   | 124 | 1902 | 0.8320                           | 0.8701 | 0.8320 | 0.2278                    | 0.4645 | 0.2146 |
| 2   | 26708    | C <sub>8</sub> H <sub>17</sub> N                 | 127 | 2258 | 0.5597                           | 0.6247 | 0.5257 | 0.4411                    | 0.5713 | 0.4993 |
| 3   | 113790   | C <sub>9</sub> H <sub>20</sub> O                 | 144 | 405  | 0.1760                           | 0.6250 | 0.1844 | 0.3156                    | 0.2351 | 0.2042 |
| 4   | 158384   | C <sub>7</sub> H <sub>14</sub>                   | 98  | 56   | 0.2668                           | 0.7077 | 0.2631 | 0.6091                    | 0.2273 | 0.6273 |
| 5   | 38909    | C <sub>10</sub> H <sub>18</sub>                  | 138 | 5568 | 0.6374                           | 0.7264 | 0.7193 | 0.1846                    | 0.5017 | 0.1230 |
| 6   | 61924    | C <sub>10</sub> H <sub>20</sub>                  | 140 | 852  | 0.1360                           | 0.8125 | 0.1022 | 0.6193                    | 0.0987 | 0.5834 |
| 7   | 60708    | C <sub>8</sub> H <sub>12</sub>                   | 108 | 2082 | 0.5877                           | 0.6153 | 0.6787 | 0.1547                    | 0.4166 | 0.1528 |
| 8   | 1911     | C <sub>6</sub> H <sub>12</sub> O <sub>2</sub>    | 116 | 1313 | 0.5206                           | 0.6231 | 0.5593 | 0.0152                    | 0.0206 | 0.0122 |
| 9   | 61640    | C <sub>13</sub> H <sub>28</sub>                  | 184 | 802  | 0.6678                           | 0.6678 | 0.6678 | 0.1298                    | 0.1298 | 0.1298 |
| 10  | 4617     | CN <sub>3</sub> F <sub>5</sub>                   | 149 | 11   | 0.0551                           | 0.2280 | 0.0000 | 0.3000                    | 0.1000 | 0.5000 |
| 11  | 194167   | C <sub>4</sub> H <sub>8</sub> N <sub>2</sub> O   | 100 | 6754 | 0.2239                           | 0.2951 | 0.4254 | 0.2218                    | 0.4118 | 0.0252 |
| 12  | 186524   | C <sub>6</sub> H <sub>9</sub> OBr                | 176 | 3703 | 0.1769                           | 0.6350 | 0.1639 | 0.2458                    | 0.1348 | 0.2212 |
| 13  | 38120    | CH <sub>3</sub> SiBr                             | 124 | 2    | 0.0293                           | 0.0293 | 0.0366 | 1.0000                    | 1.0000 | 0.5000 |
| 14  | 146109   | C <sub>4</sub> H <sub>2</sub> N <sub>2</sub> FCI | 132 | 6393 | 0.7602                           | 0.8579 | 0.5112 | 0.2671                    | 0.0070 | 0.1814 |
| 15  | 73456    | C <sub>5</sub> H <sub>11</sub> Br                | 150 | 8    | 0.1389                           | 0.1539 | 0.0595 | 0.4286                    | 0.4286 | 0.5714 |
| 16  | 61694    | C <sub>9</sub> H <sub>14</sub>                   | 122 | 7244 | 0.2973                           | 0.4415 | 0.3327 | 0.1344                    | 0.3412 | 0.2620 |
| 17  | 42198    | C <sub>6</sub> H <sub>11</sub> OBr               | 178 | 1115 | 0.7514                           | 0.8727 | 0.8207 | 0.0768                    | 0.0197 | 0.0242 |
| 18  | 109982   | C <sub>4</sub> H <sub>7</sub> SiCl <sub>3</sub>  | 188 | 729  | 0.5188                           | 0.5398 | 0.5156 | 0.0034                    | 0.0076 | 0.0357 |
| 19  | 120      | C <sub>2</sub> H <sub>3</sub> NO                 | 57  | 26   | 0.0010                           | 0.0010 | 0.1454 | 0.1600                    | 0.2800 | 0.0800 |
| 20  | 154091   | C <sub>8</sub> H <sub>14</sub>                   | 110 | 654  | 0.5546                           | 0.5546 | 0.3005 | 0.1838                    | 0.6631 | 0.7833 |
| 21  | 71109    | C <sub>6</sub> H <sub>14</sub> N <sub>2</sub>    | 114 | 2338 | 0.6679                           | 0.7005 | 0.7105 | 0.0548                    | 0.3128 | 0.0276 |
| 22  | 162833   | C <sub>10</sub> H <sub>18</sub>                  | 138 | 5568 | 0.4108                           | 0.6497 | 0.6199 | 0.1536                    | 0.1638 | 0.1038 |
| 23  | 249757   | C <sub>5</sub> H <sub>9</sub> N                  | 83  | 313  | 0.1136                           | 0.2502 | 0.3053 | 0.7548                    | 0.4615 | 0.5128 |
| 24  | 3238     | C <sub>5</sub> H <sub>10</sub> O <sub>2</sub> S  | 134 | 4560 | 0.1134                           | 0.1134 | 0.1124 | 0.1481                    | 0.5782 | 0.1743 |
| 25  | 113090   | C <sub>8</sub> H <sub>14</sub>                   | 110 | 654  | 0.6907                           | 0.6907 | 0.7081 | 0.2726                    | 0.5735 | 0.1937 |
| 26  | 63698    | C <sub>3</sub> H <sub>4</sub> N <sub>2</sub> O   | 84  | 1371 | 0.1169                           | 0.1169 | 0.2012 | 0.3376                    | 0.7588 | 0.1394 |
| 27  | 74975    | C <sub>6</sub> H <sub>12</sub> O <sub>3</sub>    | 132 | 6171 | 0.5343                           | 0.7865 | 0.5441 | 0.1289                    | 0.3659 | 0.1330 |
| 28  | 185578   | C <sub>5</sub> H <sub>10</sub> O <sub>4</sub>    | 134 | 5841 | 0.8586                           | 0.8673 | 0.8350 | 0.0284                    | 0.3959 | 0.1500 |
| 29  | 61113    | C <sub>10</sub> H <sub>20</sub>                  | 140 | 852  | 0.7630                           | 0.8772 | 0.8581 | 0.1716                    | 0.0159 | 0.0546 |
| 30  | 160559   | C <sub>4</sub> H <sub>13</sub> NP <sub>2</sub>   | 137 | 396  | 0.7472                           | 0.7476 | 0.1320 | 0.0177                    | 0.0557 | 0.3823 |
| 31  | 46389    | C <sub>5</sub> H <sub>10</sub> O <sub>3</sub>    | 118 | 1656 | 0.8403                           | 0.8795 | 0.8252 | 0.0302                    | 0.1970 | 0.0483 |
| 32  | 46612    | C <sub>9</sub> H <sub>18</sub> O                 | 142 | 4745 | 0.6880                           | 0.8471 | 0.7695 | 0.1184                    | 0.2661 | 0.0470 |
| 33  | 105465   | C <sub>7</sub> H <sub>16</sub> Si                | 128 | 889  | 0.7817                           | 0.8250 | 0.8254 | 0.0270                    | 0.0270 | 0.0028 |
| 34  | 61433    | C <sub>11</sub> H <sub>24</sub>                  | 156 | 159  | 0.5511                           | 0.5511 | 0.5614 | 0.6646                    | 0.6646 | 0.6582 |
| 35  | 113438   | C <sub>8</sub> H <sub>16</sub>                   | 112 | 139  | 0.1367                           | 0.1382 | 0.1416 | 0.8116                    | 0.8696 | 0.6957 |
| 36  | 215368   | C <sub>6</sub> H <sub>10</sub> O                 | 98  | 747  | 0.1115                           | 0.1171 | 0.0633 | 0.6649                    | 0.8881 | 0.8780 |
| 37  | 20664    | C <sub>9</sub> H <sub>20</sub>                   | 128 | 35   | 0.5628                           | 0.5628 | 0.5628 | 0.2647                    | 0.2647 | 0.4853 |
| 38  | 62859    | C <sub>8</sub> H <sub>14</sub>                   | 110 | 654  | 0.4286                           | 0.5076 | 0.4424 | 0.1271                    | 0.4824 | 0.1639 |
| 39  | 69684    | C <sub>11</sub> H <sub>24</sub> O                | 172 | 2426 | 0.4653                           | 0.5631 | 0.4863 | 0.0237                    | 0.3841 | 0.0089 |
| 40  | 629      | C <sub>5</sub> H <sub>13</sub> N                 | 87  | 17   | 0.8350                           | 0.8407 | 0.8367 | 0.0625                    | 0.0000 | 0.0625 |
| 41  | 152851   | C <sub>4</sub> H <sub>7</sub> O <sub>2</sub> Cl  | 122 | 487  | 0.6093                           | 0.6093 | 0.2142 | 0.0062                    | 0.0391 | 0.0123 |
| 42  | 114082   | C <sub>6</sub> H <sub>14</sub> O                 | 102 | 32   | 0.0125                           | 0.0601 | 0.0528 | 0.7581                    | 0.5806 | 0.5645 |
| 43  | 196609   | C <sub>5</sub> H <sub>11</sub> NO <sub>2</sub>   | 117 | 6418 | 0.5334                           | 0.5689 | 0.5374 | 0.2637                    | 0.5587 | 0.2139 |
| 44  | 204405   | C <sub>9</sub> H <sub>14</sub>                   | 122 | 7244 | 0.6228                           | 0.8422 | 0.6002 | 0.1267                    | 0.2815 | 0.3232 |
| 45  | 28546    | C <sub>5</sub> H <sub>12</sub> O <sub>2</sub>    | 104 | 69   | 0.0256                           | 0.0326 | 0.2624 | 0.1544                    | 0.4853 | 0.0147 |
| 46  | 113901   | C <sub>9</sub> H <sub>16</sub>                   | 124 | 1902 | 0.5049                           | 0.5084 | 0.4485 | 0.0468                    | 0.4424 | 0.1915 |
| 47  | 193841   | C <sub>6</sub> H <sub>16</sub> OSi               | 132 | 425  | 0.9357                           | 0.9357 | 0.9344 | 0.3066                    | 0.3597 | 0.2382 |
| 48  | 604      | C <sub>4</sub> H <sub>6</sub> O <sub>2</sub>     | 86  | 263  | 0.4864                           | 0.4864 | 0.4876 | 0.0573                    | 0.3416 | 0.0573 |
| 49  | 73972    | C <sub>9</sub> H <sub>21</sub> NO                | 159 | 7769 | 0.9124                           | 0.9125 | 0.9312 | 0.3589                    | 0.6582 | 0.0397 |
| 50  | 63639    | C <sub>2</sub> H <sub>6</sub> O <sub>2</sub>     | 62  | 5    | 0.6307                           | 0.6307 | 0.6429 | 0.1250                    | 0.1250 | 0.0000 |

| No.              | NIST_No. | Formula                                                      | m   | TC   | Match Value of correct structure |              |              | Relative Ranking Position |              |              |
|------------------|----------|--------------------------------------------------------------|-----|------|----------------------------------|--------------|--------------|---------------------------|--------------|--------------|
|                  |          |                                                              |     |      | MF_3st                           | MF_5st       | MSF          | MF_3st                    | MF_5st       | MSF          |
| 51               | 135135   | C <sub>4</sub> H <sub>8</sub> NOCl                           | 121 | 1371 | 0.5566                           | 0.5874       | 0.3839       | 0.0015                    | 0.0358       | 0.0182       |
| 52               | 63008    | C <sub>5</sub> H <sub>6</sub>                                | 66  | 40   | 0.3690                           | 0.3690       | 0.4656       | 0.2949                    | 0.3718       | 0.6026       |
| 53               | 61471    | C <sub>13</sub> H <sub>28</sub>                              | 184 | 802  | 0.6317                           | 0.6317       | 0.6488       | 0.3514                    | 0.3514       | 0.3439       |
| 54               | 60569    | C <sub>8</sub> H <sub>17</sub> Cl                            | 148 | 89   | 0.0363                           | 0.2249       | 0.0592       | 0.1705                    | 0.2386       | 0.1591       |
| 55               | 41785    | C <sub>8</sub> H <sub>16</sub> O                             | 128 | 1684 | 0.0428                           | 0.8527       | 0.7469       | 0.5157                    | 0.0523       | 0.0657       |
| 56               | 66064    | C <sub>9</sub> H <sub>14</sub>                               | 122 | 7244 | 0.2592                           | 0.3820       | 0.2072       | 0.4283                    | 0.7126       | 0.6575       |
| 57               | 160476   | C <sub>6</sub> H <sub>10</sub> O                             | 98  | 747  | 0.8907                           | 0.8907       | 0.8891       | 0.0369                    | 0.2895       | 0.1723       |
| 58               | 73870    | C <sub>8</sub> H <sub>12</sub>                               | 108 | 2082 | 0.4548                           | 0.4548       | 0.4457       | 0.1033                    | 0.4121       | 0.3263       |
| 59               | 108516   | C <sub>4</sub> H <sub>12</sub> N <sub>2</sub>                | 88  | 38   | 0.7566                           | 0.7566       | 0.7545       | 0.0676                    | 0.1622       | 0.1216       |
| 60               | 4169     | C <sub>3</sub> H <sub>3</sub> Cl <sub>3</sub>                | 144 | 8    | 0.6502                           | 0.6502       | 0.0019       | 0.7143                    | 0.7143       | 0.5714       |
| 61               | 46224    | C <sub>5</sub> H <sub>13</sub> N                             | 87  | 17   | 0.7369                           | 0.7369       | 0.5151       | 0.6250                    | 0.6250       | 0.8125       |
| 62               | 158830   | C <sub>7</sub> H <sub>9</sub> Br                             | 172 | 2732 | 0.3146                           | 0.3291       | 0.0058       | 0.0776                    | 0.3103       | 0.6175       |
| 63               | 61715    | C <sub>8</sub> H <sub>14</sub>                               | 110 | 654  | 0.5806                           | 0.6036       | 0.5212       | 0.0904                    | 0.3706       | 0.2657       |
| 64               | 1123     | C <sub>4</sub> H <sub>4</sub> O <sub>3</sub>                 | 100 | 1073 | 0.0492                           | 0.8124       | 0.0681       | 0.3125                    | 0.0401       | 0.2360       |
| 65               | 156613   | C <sub>9</sub> H <sub>22</sub> NP                            | 175 | 9663 | 0.9099                           | 0.9259       | 0.0004       | 0.1936                    | 0.3801       | 0.7810       |
| 66               | 176      | C <sub>2</sub> H <sub>7</sub> P                              | 62  | 2    | 0.1597                           | 0.1597       | 0.1597       | 1.0000                    | 1.0000       | 1.0000       |
| 67               | 114550   | C <sub>7</sub> H <sub>14</sub> O                             | 114 | 596  | 0.5770                           | 0.8119       | 0.5535       | 0.1387                    | 0.1084       | 0.1160       |
| 68               | 214253   | C <sub>5</sub> H <sub>13</sub> NO                            | 103 | 149  | 0.6499                           | 0.8028       | 0.6480       | 0.1858                    | 0.2027       | 0.0372       |
| 69               | 70751    | C <sub>7</sub> H <sub>19</sub> N <sub>3</sub>                | 145 | 4238 | 0.6554                           | 0.6566       | 0.6480       | 0.2424                    | 0.6613       | 0.0775       |
| 70               | 62909    | C <sub>6</sub> H <sub>12</sub> O                             | 100 | 211  | 0.4573                           | 0.4573       | 0.4758       | 0.2333                    | 0.4476       | 0.3143       |
| 71               | 37206    | C <sub>7</sub> H <sub>13</sub> N                             | 111 | 3809 | 0.3332                           | 0.7841       | 0.3557       | 0.4791                    | 0.2428       | 0.3339       |
| 72               | 229049   | C <sub>4</sub> H <sub>11</sub> NO                            | 89  | 56   | 0.7712                           | 0.7724       | 0.7706       | 0.2182                    | 0.2182       | 0.2364       |
| 73               | 19272    | C <sub>6</sub> H <sub>10</sub>                               | 82  | 77   | 0.6213                           | 0.6213       | 0.0896       | 0.5000                    | 0.6316       | 0.9342       |
| 74               | 831      | C <sub>2</sub> NF <sub>3</sub>                               | 95  | 5    | 0.6830                           | 0.6830       | 0.4977       | 0.0000                    | 0.0000       | 0.0000       |
| 75               | 114407   | C <sub>7</sub> H <sub>12</sub>                               | 96  | 222  | 0.6961                           | 0.7391       | 0.7955       | 0.4095                    | 0.3824       | 0.0226       |
| 76               | 5393     | C <sub>4</sub> H <sub>6</sub> O <sub>2</sub> Cl <sub>2</sub> | 156 | 1131 | 0.0317                           | 0.2480       | 0.0292       | 0.2283                    | 0.0867       | 0.1792       |
| 77               | 30409    | C <sub>5</sub> H <sub>18</sub> Si <sub>3</sub>               | 162 | 521  | 0.2269                           | 0.2269       | 0.0000       | 0.4981                    | 0.4981       | 0.9788       |
| 78               | 60785    | C <sub>9</sub> H <sub>20</sub> O                             | 144 | 405  | 0.6521                           | 0.6668       | 0.4782       | 0.0012                    | 0.2067       | 0.2500       |
| 79               | 72642    | C <sub>9</sub> H <sub>22</sub> N <sub>2</sub>                | 158 | 4994 | 0.7650                           | 0.7753       | 0.7536       | 0.3966                    | 0.6845       | 0.3615       |
| 80               | 118272   | C <sub>3</sub> H <sub>7</sub> NO                             | 73  | 84   | 0.1824                           | 0.1824       | 0.6177       | 0.6566                    | 0.7048       | 0.0482       |
| 81               | 108346   | C <sub>3</sub> H <sub>7</sub> O <sub>2</sub> Br              | 154 | 38   | 0.1001                           | 0.4454       | 0.0992       | 0.0541                    | 0.0000       | 0.0000       |
| 82               | 26687    | C <sub>8</sub> H <sub>14</sub>                               | 110 | 654  | 0.3234                           | 0.3234       | 0.3170       | 0.6662                    | 0.9479       | 0.8392       |
| 83               | 113772   | C <sub>7</sub> H <sub>14</sub> O                             | 114 | 596  | 0.1485                           | 0.2307       | 0.1535       | 0.6235                    | 0.6513       | 0.6782       |
| 84               | 1614     | C <sub>8</sub> H <sub>16</sub>                               | 112 | 139  | 0.6023                           | 0.6091       | 0.6245       | 0.0507                    | 0.1957       | 0.0362       |
| 85               | 107506   | C <sub>9</sub> H <sub>19</sub> F                             | 146 | 211  | 0.2495                           | 0.2495       | 0.2998       | 0.0024                    | 0.0024       | 0.0000       |
| 86               | 98625    | C <sub>6</sub> H <sub>14</sub> Si                            | 114 | 314  | 0.7622                           | 0.7622       | 0.7426       | 0.2460                    | 0.2588       | 0.0831       |
| 87               | 1908     | C <sub>6</sub> H <sub>12</sub> O <sub>2</sub>                | 116 | 1313 | 0.1238                           | 0.2130       | 0.2368       | 0.5103                    | 0.7066       | 0.3925       |
| 88               | 134724   | C <sub>3</sub> H <sub>4</sub> NSBr                           | 165 | 480  | 0.3427                           | 0.3427       | 0.1458       | 0.0073                    | 0.0574       | 0.2443       |
| 89               | 50930    | C <sub>9</sub> H <sub>18</sub>                               | 126 | 338  | 0.8575                           | 0.9079       | 0.3775       | 0.0697                    | 0.0653       | 0.1795       |
| 90               | 64555    | C <sub>5</sub> H <sub>10</sub> N <sub>2</sub>                | 98  | 2668 | 0.6097                           | 0.6391       | 0.6093       | 0.1348                    | 0.1877       | 0.1562       |
| 91               | 113750   | C <sub>9</sub> H <sub>20</sub> O                             | 144 | 405  | 0.1184                           | 0.6512       | 0.1261       | 0.1361                    | 0.2017       | 0.0668       |
| 92               | 114530   | C <sub>8</sub> H <sub>16</sub> O                             | 128 | 1684 | 0.1901                           | 0.2923       | 0.2104       | 0.1771                    | 0.4593       | 0.0850       |
| 93               | 61453    | C <sub>12</sub> H <sub>24</sub>                              | 168 | 5513 | 0.8410                           | 0.8544       | 0.1715       | 0.0170                    | 0.1189       | 0.1776       |
| 94               | 37233    | C <sub>9</sub> H <sub>16</sub>                               | 124 | 1902 | 0.0783                           | 0.2076       | 0.1734       | 0.8133                    | 0.8301       | 0.3062       |
| 95               | 60877    | C <sub>12</sub> H <sub>24</sub>                              | 168 | 5513 | 0.7620                           | 0.8824       | 0.7680       | 0.0889                    | 0.0274       | 0.0751       |
| 96               | 63617    | C <sub>3</sub> H <sub>4</sub> O                              | 56  | 13   | 0.0270                           | 0.0270       | 0.6550       | 0.1667                    | 0.2500       | 0.0000       |
| 97               | 72945    | C <sub>4</sub> H <sub>5</sub> OCl                            | 104 | 175  | 0.6993                           | 0.7014       | 0.0255       | 0.0920                    | 0.2586       | 0.1810       |
| 98               | 113601   | C <sub>12</sub> H <sub>24</sub>                              | 168 | 5513 | 0.6791                           | 0.7110       | 0.6541       | 0.0149                    | 0.0474       | 0.0127       |
| 99               | 52322    | C <sub>5</sub> H <sub>13</sub> N <sub>3</sub>                | 115 | 4054 | 0.2284                           | 0.5986       | 0.3751       | 0.6339                    | 0.5216       | 0.2846       |
| 100              | 215367   | C <sub>6</sub> H <sub>8</sub> O                              | 96  | 1623 | 0.1968                           | 0.2693       | 0.3200       | 0.3924                    | 0.6874       | 0.5953       |
| <b>Averages:</b> |          |                                                              |     |      | <b>0.462</b>                     | <b>0.558</b> | <b>0.432</b> | <b>0.269</b>              | <b>0.353</b> | <b>0.273</b> |

**Table A5: Summary of MOLGEN-MS results for the 29 C<sub>12</sub>H<sub>10</sub>O<sub>2</sub> isomers. Classifier codes are explained within NIST and MOLGEN-MS.**

|    | NIST Prob. (%) | Class. Prob. (%) | Classifiers added (+) / removed (-)                                            | Struct. Gen. | Calc. time (s) | MV range (%) | MV / Place of correct structure |
|----|----------------|------------------|--------------------------------------------------------------------------------|--------------|----------------|--------------|---------------------------------|
| 1  | 87.4           | 95               |                                                                                | 13           | 544.2          | 63.2-63.5    | 63.4 / 10                       |
| 2  | 92.8           | 95               | -naphth (MSGSL)<br>-ar-OHs (MSBL)                                              | 3            | 1.0            | 63.1 (all)   | 63.1 / eq. 1                    |
| 3  | 81.4           | 95               | -arC (MSBL)<br>-aromaph (MSBL)                                                 | 2            | 1.4            | 1.8 (all)    | 1.8 / eq. 1                     |
| 4  | 93.6           | 95               |                                                                                | >20,000      | 94.1           | n/c          |                                 |
|    |                | 95 (86)*         | +ph-C, +C=O (NISTGL)                                                           | 4637         | 24.4           | 24.9-76.7    | 34.23 / eq. 1377                |
|    |                | 95 (86)*         | +ph-C, +C=O (NISTGL)<br>Min. Cycle = 5                                         | 204          | 10.4           | 24.9-76.7    | 34.2 / eq. 22                   |
| 5  | 88.1           | 95               |                                                                                | >20,000      | 86.0           | n/c          |                                 |
|    |                | 95 (82)*         | +C=O, COC, C6H5 (NISTGL), -CCH3-r, CH2-r (NISTBL)                              | 2814         | 11.0           | 31.5-90.9    | 31.7 / eq. 2739                 |
|    |                | 95 (82)*         | +C=O, COC, C6H5 (NISTGL), -CCH3-r, CH2-r (NISTBL)<br>Min. Cycle = 5            | 400          | 4.8            | 31.6-81.5    | 31.7 / 368                      |
| 6  | 95.7           | 95               | -biphenyl (MSGSL)<br>-ArCH2 (MSBL)                                             | 496          | 55.5           | 17.1-25.7    | 25.2 / 15                       |
|    |                | 95 (94)*         | -biphenyl (MSGSL)<br>-ArCH2 (MSBL)<br>+C=O (NISTGL)                            | 18           | 4.7            | 19.5-25.3    | 25.2 / 12                       |
| 7  | 89.9           | 95               |                                                                                | 2            | 0.5            | 82.47-82.52  | 82.52 / 1                       |
| 8  | 95.4           | 95               |                                                                                | 0            | -              | -            | -                               |
|    |                | 95               | -Ar-C (NISTBL)<br>-ArOHs (MSBL)                                                | 27           | 0.7            | 92.1-93.4    | 92.57 / eq. 8                   |
| 9  | 98.2           | 95               |                                                                                | 0            | -              | -            | -                               |
|    |                | 95               | -C6H5 (MSGSL)                                                                  | 10,893       | 7301.1         | 1.27-83.36   | 48.37 / eq. 523                 |
|    |                | 95               | -C6H5 (MSGSL)<br>Min cycle = 5                                                 | 1474         | 1262.6         | 1.27-80.51   | 48.37 / eq. 156                 |
|    |                | 95               | -C6H5 (MSGSL)<br>Min cycle = 6                                                 | 533          | 585.4          | 1.32-80.51   | 48.37 / eq. 61                  |
| 10 | 96.8           | 95               |                                                                                | 5            | 0.9            | 65.00-67.51  | Not present                     |
|    |                | 95               | -biphenyl (MSGSL)                                                              | >20,000      | 15.0           | n/c          |                                 |
|    |                | 95 (90)*         | -biphenyl, (MSGSL)<br>+ArCring(MSGSL)*                                         | >20,000      | 92             | n/c          |                                 |
|    |                | 95 (90)*         | -biphenyl (MSGSL)<br>+ArCring(MSGSL)*<br>Min cycle 6                           | >20,000      | 500            | n/c          |                                 |
|    |                | 95 (89)*         | -biphenyl (MSGSL)<br>+ArCring(MSGSL)*<br>+C=O, +ArC=O (NISTGL)*                | >20,000      | 630            | n/c          |                                 |
|    |                | 95 (89)*         | -biphenyl (MSGSL)<br>+ArCring(MSGSL)*<br>+C=O, +ArC=O (NISTGL)*<br>Min cycle 6 | >20,000      | 800            | n/c          |                                 |
| 11 | 96.7           | 95               |                                                                                | 5            | 1.1            | 66.26-67.00  | Not present                     |
|    |                | 95 (85)*         | -biphenyl(MSGSL)<br>+naphth, +OCH3, +C=O (NISTGL)                              | 16           | 2.5            | 66.35-67.45  | 67.42 / 3                       |
| 12 | 97.8           | 95               |                                                                                | 3,943        | 116.9          | 18.85-88.78  | Not present                     |
|    |                | 95 (90)*         | +C=OO(NISTGL)                                                                  | 34           | 1.5            | 19.08-72.92  | Not present                     |
| 13 | 92.3           | 95               |                                                                                | 0            | 169.4          | -            | -                               |
|    |                | 95               | -et-est(MSGSL), -badlist                                                       | 9            | 421.1          | 10.1-86.6    | 86.56 / 1 but isomers absent    |
|    |                | 95               | -et-est(MSGSL), -badlist<br>-ArCH2C=O; +CH2C=O                                 | 36           | 0.6            | 10.07-86.56  | 86.56 / 1                       |

|    | NIST Prob. (%) | Class. Prob. (%) | Classifiers added (+) / removed (-)                                                      | Struct. Gen. | Calc. time (s) | MV range (%)    | MV / Place of correct structure |
|----|----------------|------------------|------------------------------------------------------------------------------------------|--------------|----------------|-----------------|---------------------------------|
| 14 | 96.9           | 95               |                                                                                          | 0            | -              | -               | -                               |
|    |                | 95               | -COC(NISTBL)                                                                             | 12,913       | 64.2           | 16.98-52.93     | Not present                     |
|    |                | 95               | -COC(NISTBL)                                                                             | 807          | 7.7            | 17.21-49.67     | Not present                     |
|    |                | 95 (79)*         | Min cycle 5<br>+ArC=O(NISTGL)<br>-CH3C=OO, -CH3C=O (MSGI)<br>-COC(NISTBL)<br>Min cycle 6 | >20,000      | 370            | n/c             |                                 |
| 15 | 80.8           | 95               |                                                                                          | 3902         | 60.6           | 0.6-3.88        | 1.60 / eq. 1862                 |
|    |                | 95 (94)*         | C=OO(NISTGL)                                                                             | 36           | 1.5            | 0.65-3.83       | 1.60 / eq. 33                   |
| 16 | 97.7           | 95               | -naph(MSGI)                                                                              | 22           | 0.6            | 60.79-62.75     | 62.75 / eq. 1<br>61.071 / 18    |
|    |                |                  | -ArC(NISTBL)                                                                             |              |                |                 |                                 |
|    |                |                  | -phen-OHs(MSGI)                                                                          |              |                |                 |                                 |
| 17 | 88.0           | 95               |                                                                                          | 0            | -              | -               | -                               |
|    |                | 95               | All but naph, CH3C=O                                                                     | 14           | 0.6            | 57.339-57.459   | 57.459 / 2                      |
| 18 | 92.2           | 95               |                                                                                          | 0            | -              | -               | -                               |
|    |                | 95               | -clashing BL entries                                                                     | 13           | 381.8          | 82.523-82.651   | 82.523 / 13                     |
| 19 | 83.4           | 95               |                                                                                          | 0            | -              | -               | -                               |
|    |                | 95               | -clashing BL entries                                                                     | 13           | 116.6          | 68.669-68.861   | Not present                     |
|    |                | 95               | -all but naph, C=OCH3                                                                    | 14           | 1.1            | 68.666-68.861   | 68.666 / 14                     |
| 20 | 96.7           | 95               |                                                                                          | 0            | -              | -               | -                               |
|    |                | 95               | - all BL entries                                                                         | 3            | 3.9            | 64.669          | 64.669 / eq. 1                  |
| 21 | 94.1           | 95               | Min cycle 6                                                                              | >20,000      | 82.7           | -               | -                               |
|    |                | 95 (89)*         | Ar-Ar (NIST GL)                                                                          | 22           | 0.4            | 38.747 – 44.924 | 39.705 / 20 38.747 / 22         |
| 22 | 87.7           | 95               |                                                                                          | 0            | -              | -               | -                               |
|    |                | 95               | -all but naph, ArOH from GL, phenOH from BL                                              | 3602         | 78.4           | 72.78-86.59     | Not present                     |
|    |                | 95               | -PhO, naph from GL, all BL                                                               | 22           | 0.4            | 73.574-76.027   | 74.673 / 4<br>73.988 / 12       |
| 23 | 90.8           | 95               | -all but ArOH, naph from GL, ArC, phenOH from BL                                         | 14           | 2.9            | 72.356-73.013   | Not present                     |
|    |                | 95 (94)*         | - naph, +biphenyl                                                                        | 22           | 1.1            | 72.437 – 76.697 | 72.437 / 22                     |
| 24 | 86.0           | 95               |                                                                                          | 0            | -              | -               | -                               |
|    |                | 95               | -CH2, me-est, C=OCH3 from BL                                                             | 2            | 0.2            | 66.916 – 66.939 | 66.916 / 2                      |
| 25 | 96.7           | 95               | -CH2 (NISTBL)                                                                            | 1387         | 56.2           | 32.15-57.975    | 33.836 / eq. 1267               |
|    |                | 95 (86)*         | +ArCHO, OCH3 (NISTGL)<br>-CH2 (NISTBL)                                                   | 14           | 11.5           | 33.423-33.836   | 33.836 / eq. 1                  |
| 26 | 92.9           | 95               |                                                                                          | 0            | -              | -               | -                               |
|    |                | 95               | -et-est                                                                                  | 1            | 538.3          | 85.464          | Not present                     |
|    |                | 95               | -ArC, ArCH2C=O<br>+CH2C=O                                                                | 6            | 29.6           | 5.615 – 85.464  | 85.083 / 2                      |
| 27 | 94.4           | 95               | -phenOH, ArOR (BL)                                                                       | 3            | 4.0            | 80.242 – 80.243 | 80.243 / eq. 1                  |
| 28 | 91.5           | 95               | -naph (MSGI), ArC, phenOH from BL                                                        | 22           | 6.3            | 79.06-81.69     | 81.68 / 2<br>79.52 / 11         |
| 29 | 92.1           | 95               | -naph (GL), -ArC, phenOH (BL)                                                            | 22           | 8.0            | 77.57 - 82.315  | 79.63 / 5<br>79.005 / 6         |

\* bracketed value indicates Yes substructure classifier level for additional classifiers. 'eq.' refers to equal ranks, i.e. where more than one structure has the same match value. GL refers to good list, BL to bad list.

## Appendix 2: List of Scripts

|                                |                                                                                                                                                                 |
|--------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| ACDassign.m                    | Prepare MSP files for fragment generation using ACD                                                                                                             |
| ACDfrag_eval.m                 | Evaluates fragments generated using ACD                                                                                                                         |
| ACDtxt_gen.m                   | Conversion of ACD Table of Fragments for MOLGEN-MSF                                                                                                             |
| add_M_END.m                    | Function file to add “M END” lines to SDF files                                                                                                                 |
| bulk_NIST_plot_bar.m           | Plot multiple MSP files at once                                                                                                                                 |
| frag_eval.m                    | Evaluate fragments and compare with original spectrum                                                                                                           |
| frag_eval_bulk.m               | frag_eval for processing of multiple files                                                                                                                      |
| frag_sum.m                     | Summarises information generated by frag_eval_bulk.m                                                                                                            |
| inp_file_read.m                | Function file to read MOLGEN-MS “INP” files                                                                                                                     |
| kow_EPI.m                      | Function to generate logKow via EPISuite from SMILES files                                                                                                      |
| MFprep.m                       | Perpare sdf files for Mass Frontier batch processing                                                                                                            |
| MolForm_ElCoCo<br>_file_read.m | Function file to read output files from MolForm or ElCoCo for<br>summary file                                                                                   |
| MM2Energy.py                   | Python script to calculate ChemBio3D MM2 Energy                                                                                                                 |
| mpbp_file.m                    | Function to generate melting and boiling point data via<br>EPISuite from SMILES files                                                                           |
| mpr_file_read.m                | Function to read MOLGEN3.5 MRP files                                                                                                                            |
| MS_peak_summary.m              | Read AMDIS/NIST MSP files and scale peaks to 100 %                                                                                                              |
| ms_plot.m                      | Create Spectrum plot for MSP or CSV files                                                                                                                       |
| ms_summary2.m                  | <i>Main Script:</i> Generation of data summary files and plots. Needs<br>input summary text file, e.g. “shortname_runX_summary.txt”                             |
| MSclass_read.m                 | Function to read MSclass files for summary file                                                                                                                 |
| msp2csv.m                      | Conversion of MSP files to CSV for MOLGEN-MS.                                                                                                                   |
| msp2csv_batch.m                | msp2csv for conversion of multiple files                                                                                                                        |
| msp_read.m                     | Function to read MSP files and return list of peaks                                                                                                             |
| NIST_msp_split.m               | Split multi-spectra MSP files into one MSP per spectrum                                                                                                         |
| NISTclass.m                    | Read NIST classifier information for input to MOLGEN-MS.<br>Needs input files “NIST_classifier_list_8col_noEdit.txt” or<br>“NIST_classifier_list_8col_test.txt” |
| NISTinfo.m                     | Read and output compound information from NIST MSP files                                                                                                        |
| NISTmsp_check.m                | Function to check/convert MSP format for ACD                                                                                                                    |
| python_E2.m                    | Function to generate ChemBio3D steric energy values                                                                                                             |
| rank_MSF.m                     | Generate match values for given structures and spectrum                                                                                                         |
| RDB_count.m                    | Calculation of the ring and double bond count (Equation 11)<br>from MolForm/ElCoCo output.                                                                      |
| struct_alert.m                 | Determine whether a ‘structural alert’ is present for a formula.<br>Needs “struct_alerts_daphnia.csv”                                                           |
| tms.m                          | Function to separate trimethylsilyl groups from a given formula                                                                                                 |

These scripts can be found in original format (\*.m, \*.py etc) and as PDF (\*.pdf) from <http://www.ufz.de/index.php?en=14839> and [www.molgen.de](http://www.molgen.de) (coming soon) with a sub-set also available from <http://pubs.acs.org>.