Universal Fragmentation Model for Tandem Mass Spectrometry Based Molecular Structural Elucidation

Béla Paizs,*[†‡$] Daniel McGill,[†] Daniel Simon,[†] Zoltán Takáts*[†‡$]

[†]The Rosalind Franklin Institute, Harwell Campus, Didcot, UK; [$]deShape ltd. London, UK; [‡]Faculty of Medicine, Department of Metabolism, Digestion and Reproduction, Imperial College London, South Kensington Campus, London, UK

ABSTRACT: Tandem mass spectrometry based structural elucidation is currently hampered by our limited understanding of fragmentation chemistry. Here we present the Universal Fragmentation Model (UFM) that is based on gas-phase ion chemistry and modelling and is capable of predicting high-quality fragmentation pathways, structures and energetics for general molecules. We demonstrate that UFM can interpret fragmentation chemistries dominated by complex rearrangements.

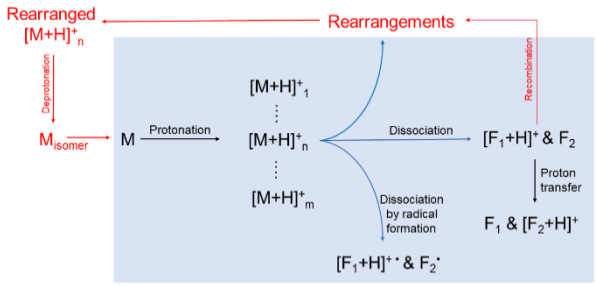
Modern "omics" technologies like proteomics,[1] metabolomics,[2] lipidomics,[3] and glycomics,[4] rely heavily on the identification of various molecular species in complex mixtures by mass spectrometry (MS). Information used for structural elucidation is obtained as accurate masses of intact (single stage MS) and fragmented species (generated in tandem mass spectrometry (MS/MS)), and molecular and fragment collision cross-sections (CCS) if MS is coupled to ion mobility spectrometry (IMS). Modern MS&MS/MS&IMS experiments hyphenated to liquid/gas chromatography produce large datasets which can be processed only by using automated tools implementing chemoinformatics[5] strategies.

To decipher the structural information encoded in MS&MS/MS&IMS datasets the MS community has devoted substantial efforts to understanding fine details of the chemical processes happening in mass spectrometers. Generations of chemists have studied the dissociations of energized ions to develop fragmentation rules/laws which could be used to explain, and subsequently predict, the chemistries involved. Typically, these studies involve detailed characterization of the energetics and kinetics of dissociation for carefully selected sets of model compounds, with the aim of developing concise fragmentation models – a well-defined set of fragmentation laws and rules - which could then be used to explain data recorded for unknown compounds to elucidate their molecular structures.

For some molecular classes and fragmentation techniques these efforts have been largely successful in terms of linking fragmentation chemistry to structures, but accurate prediction of dissociation chemistry remains elusive for arbitrary molecules. For example, in electron impact (EI) MS[6] structural information can be deduced in a straightforward fashion if fragments are formed by direct bond cleavages[6]. However, it also became evident early on that the formation of numerous fragments cannot be explained by considering only bond fission; instead, rather complex rearrangements[6] can take place. These rearrangement pathways alter primary molecular connectivity and can lead to fragments with structures not directly deducible from the original analyte. These are notoriously difficult to predict; consequently, EI-MS spectra are undercharacterised despite long years of dedicated research to assign structure without referring to spectral libraries.

Due to the relatively uniform chemical space (amino acid residues connected by fragile amide bonds) peptide fragmentation chemistry is well understood, and existing peptide fragmentation models[7] can explain basic fragmentation characteristics for protonated peptides. In metabolomics, however, the targeted chemical space is vastly more complex. While the dissociations of some metabolites can be described by straightforward bond cleavages, it is often the case that rearrangements dominate the fragmentation patterns of even simple molecules.[8] As a consequence of the chemical complexity involved, no *fully predictive* (i.e. accurately predicting the fragments based on the structure) fragmentation model has been proposed for metabolomics so far.[9] Consequently, metabolite identification is currently a major bottleneck hindering further development and applications of metabolomics technologies. Recent attempts to address this problem range from simple bond-cutting approaches to sophisticated machine learning based technologies.[10]

To generate accurate and robust fragmentation patterns we introduce here a comprehensive fragmentation model for mass spectrometry-based metabolomics, referred to as the Universal Fragmentation Model (UFM). The UFM flow chart is demonstrated for general molecules (M) being charged by protonation ($[M+H]^+$) in Scheme 1.



Scheme 1. UFM flow chart for protonated molecules.

Typical metabolites have a number of protonation sites; the corresponding protomers ($[M+H]_1^+$, … $[M+H]_n^+$,… $[M+H]_m^+$) are all individually considered by UFM for a variety of reactions (indicated by blue arrows in Scheme 1). First, likely fragmentation channels incorporating heterolytic bond fission for each protomer ($[M+H]_n^+$) are predicted, which primarily result in a charged $[F_1+H]^+$ and a neutral F_2 fragment. Typically, each protomer could be broken down via several fragmentation channels. The charged fragments from these channels are either observed in the mass spectrometer leading to build up the MS/MS spectrum, or - under low energy collision conditions – the ion-molecule complex formed by F_1, F_2 and the ionizing proton might undergo intermolecular proton transfer to form charged $[F_2+H]^+$ fragments, as well as recombination reactions to form isomers of $[M+H]_n^+$. In some cases, the charged $[F_1+H]^+$ fragment becomes unstable and further dissociates to form another $[F'_1+H]^+$ charged fragment and another neutral F_3.

Formation of radical fragments[11] from $[M+H]_n^+$ via homolytic bond cleavage is also considered. Under typical CID conditions, the fragmentation chemistry of the majority of metabolites is dominated by heterolytic bond cleavages, leading to closed shell ions and neutrals. However, in some cases (for example some aromatic systems), homolytic bond cleavage and radical fragment formation might be significant. This occurs typically if no low-energy heterolytic fragmentation pathway is available and/or a particularly stable radical fragment is formed. Currently, UFM considers the formation of radical fragments if no low-energy heterolytic bond cleavage is available for $[M+H]_n^+$.

Beyond fragment recombination, a number of additional *rearrangement chemistries* are considered by UFM to explain complex fragmentation patterns. For each $[M+H]_n^+$ protomer, various hydride transfer (HT), ring opening, ring closure, and functional group isomerization pathways are predicted, leading to rearranged $[M+H]_n^+$ species - these are then deprotonated (often via a number of deprotonation pathways, leading to a number of neutral tautomers) to generate isomers (M_{isom}) of the original neutral M.

The blue-background area of the UFM flow chart in Scheme 1 summarizes all the chemistries considered for a given neutral M, while the area outside depicts reactions adopted to generate rearrangements and isomers of M. Iterative regeneration of the blue background area for each rearranged isomer (M_{isom}) of the original neutral M guarantees treatment of even multi-step complex rearrangement chemistries in the UFM framework. Furthermore, by executing the flow chart for neutrals derived from fragmentation products (like F_1 or F_2), one can generate not only primary but also higher order fragments, and can compute theoretical MS^n spectra as determined by actual experimental conditions.

Due to the complex structural manipulations involved, manual implementation of the UFM can be error-prone and time consuming, even for small molecules. To make the model generally applicable to various practical problems, we developed a software implementation termed deFrag of the UFM flow chart. deFrag generates all UFM chemistries described in Scheme 1 and produces provisional 3D structures for all species involved. These structures can then be optimized using any quantum chemistry software capable of reporting back optimized geometries, energies, and elec-

tronic properties. The latter are used to guide prediction of reactivity (bond cleavage and formation) for the investigated species, practically automatizing reaction mechanism generation for UFM. The optimized energetics are used to decide which fragmentation/rearrangement channels are most favoured, essentially implementing a simple kinetic model in deFrag.

Currently, deFrag reads 2D structures as .mol files, uses PM6-D3H4[12] as implemented in Mopac[12] for quantum chemical calculations, and performs a limited potential energy search using distance geometry[13] and geometry optimizations for all investigated species. Charge-directed pathways are predicted using Mayer's method[14], while charge-remote reactions (e.g. H_2-losses) are generated from a pre-stored library. Geometry optimization of numerous charged systems (like –OH protonated species) leads to other (dissociated or rearranged) species; deFrag considers the last intact geometries of such systems from the optimizations. F_1/F_2 fragment ratios in dissociating post-reaction complexes are approximated based on computed proton affinity values. deFrag can be conveniently tailored to specific MS conditions by setting thresholds for ring formation/opening, HT&PT, etc. It also provides energetic and mechanistic information on multistep reactions in a concise manner. Feasibility of a particular reaction step is decided by the energetics of the reactant and product and the type (PT vs. HT vs. dissociation vs. etc) of the transition, and the presence and characteristics of competing pathways.

Figure 1 presents experimental and theoretical UFM data for protonated phenylalanine (**1**), a typical metabolite with aromatic and aliphatic parts featuring common functional groups. Panels A and B depict the low-energy CID spectra and energy-dependent break-down graphs of $[\mathbf{1}+H]^+$ with main fragments at m/z 120 and 103, and further fragments at m/z 149, 131, and 77 as observed in a Waters Xevo G2-S qTOF instrument. UFM calculations (Panels C and D) indicate that formation of fragment *m/z* 149.060 (NH_3 loss, cleaved bond indicated by dotted line, E_P and E_D stand for the corresponding protomer and dissociated energetics (kcal/mol)) is initiated from the energetically most favored $[\mathbf{1}+H]^+$ structure (**1**_166, Panel D), while formation of fragment **1**_120 requires PT to the carboxyl OH, loss of water&CO. This multi-step pathway is still more favored than loss of NH3 due to the high reactivity of the OH-protonated species and the high energy of the NH_3-loss product (**1**_149). Elimination of benzene from a C-protonated species is clearly disfavored as compared to loss of water or ammonia. The primary fragment **1**_120 undergoes PT, eliminating ammonia to form **1**_103, which in turn eliminates ethyne to form **1**_77 while **1**_149 undergoes water-loss to form **1** 131. The fragmentation chemistry of $[\mathbf{1}+H]^+$ can be completely explained by PTs and direct bond cleavages, and the UFM results obtained in cca. 20 min on a laptop computer are in full agreement with earlier literature data.[15]

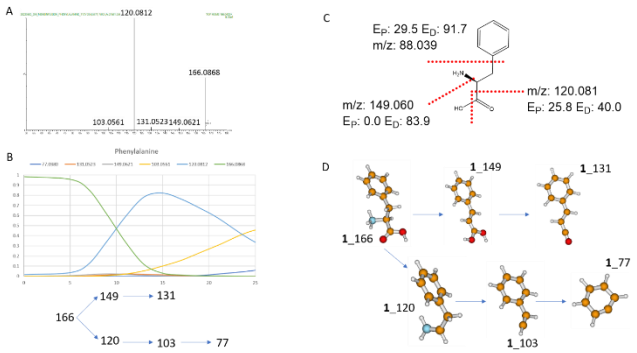


Figure 1. MS/MS (QqToF CID spectrum in panel A, and energy-resolved breakdown curves and associated fragmentation tree in panel B) and UFM predicted fragmentation data for protonated phenylalanine (**1**). For further details, see text.

To further demonstrate the power of UFM, Figure 2 presents experimental and UFM data for protonated 2-(hydroxymethyl)-N-methyl benzamide (**2**), a phenylalanine isomer. Panels A and B depict the low-energy CID spectra and energy-dependent break-down graphs of $[\mathbf{2}+H]^+$ dominated by fragments at *m/z* 148, 119, and 91. Chemical intuition suggests that $[\mathbf{1}+H]^+$ fragments primarily via the amide nitrogen protonated structure, leading to loss of methyl-amine (cleavage of an amide bond) and formation of *m/z* 135. This fragment, however, is not observed in the spectra; rather, $[\mathbf{2}+H]^+$ fragments primarily through $[\mathbf{2}+H\text{-}H_2O]^+$ at *m/z* 148 (Panel B), which is rather unexpected since this dissociation involves rupture of an aliphatic $C(sp^3)$-OH bond that is, for peptides (e.g. in Ser

side chains), less preferred than amide bond cleavage. UFM explains the prevalence of *m/z* 148 over *m/z* 135 by acknowledging that the ionizing proton can be easily mobilized to the hydroxymethyl group from the global minimum amide O protonated species (**2**_166, panel D), while its mobilization to the amide N happens through a high-energy 4-center PT. After elimination of water, the resulting CH_2^+ can be stabilized by the adjacent phenyl group via conjugation in **1**_148a (Panel D). Furthermore, the nearby amide oxygen can initiate formation of a new five-membered ring, providing extra stabilization as in **1**_148b. The energy-dependent breakdown graph (Panel B) indicates $[M+H-H_2O]^+$ fragments further to *m/z* 119 by elimination of CH_3N as an imine (CH_2=NH); this actually happens *after* hydride transfer from the methyl group to the charged carbon in **2**_148b. The resulting bicyclic **2**_119 first undergoes ring-opening and then elimination of protonated CO, which loses its charging proton back to the C_7H_6 fragment under typical low-energy collision conditions. It is important to stress here that the fragmentation of even a simple molecule like **2** can defy expectations: *all of its fragments are formed via rearrangements*.

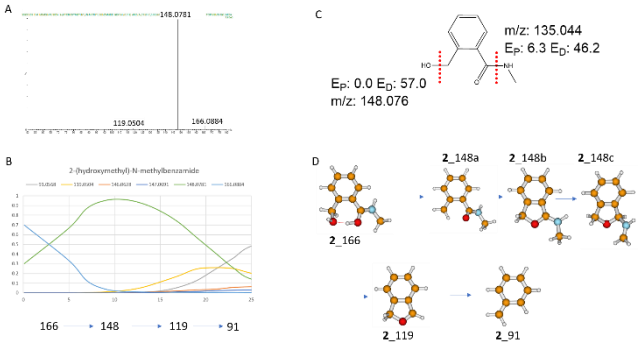


Figure 2. MS/MS and UFM predicted fragmentation data for protonated 2-(hydroxymethyl)-N-methyl benzamide (**2**). For further details, see text.

UFM predicts that the only relatively low-energy *direct* fragmentation channel for protonated 2-(4-methoxyphenyl) acetamide – another phenylalanine isomer - $[\mathbf{3}+H]^+$ is loss of ammonia after PT; however, the resulting *m/z* 149.060 fragment is not observed in the experimental spectra. On the other hand, fragmentation of $[\mathbf{3}+H]^+$ is dominated by loss of 45.0226 (CH_3NO), while also displaying a minor water-loss peak at *m/z* 148.078 - both of these fragments can be formed only via rearrangement pathways. UFM predicts that protonation of the aromatic ring is energetically feasible (< 10 kcal/mol), and the resulting **3**_166b structure can undergo a number of ring formation reactions, including a pathway leading to **3**_166c where the amide N attacks a ring carbon adjacent to the protonation site. Further proton and hydride transfers lead to **3**_166f, which is energetically more favored than the original amide oxygen protonated 2-(4-methoxyphenyl) acetamide. **3**_166f can undergo PT and opening up of the five-membered ring to form **3**_166g, which can eliminate formamide after PT to generate fragment **3**_121. Note that all **3**_166a…**3**_166g isomers are energetically more favored than the products on the ammonia-loss channel. Furthermore, the **3**_166b -> **3**_166c cyclization eliminates the NH2 group that could potentially be eliminated as ammonia to form *m/z* 149.060. This definitively explains the dominant formamide loss observed for of $[\mathbf{3}+H]^+$. The minor water-loss fragment can also be conveniently explained from **3**_166f. While for $[\mathbf{2}+H]^+$ rearrangements happened to the primary and further fragments, for $[\mathbf{3}+H]^+$ the parent structure undergoes a main structural rearrangement.

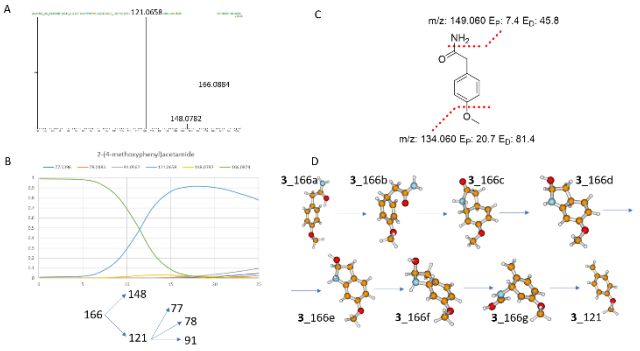


Figure 3. MS/MS and UFM predicted fragmentation data for protonated 2-(4-methoxyphenyl) acetamide (**3**).

The SI showcases MS/MS data and corresponding UFM-based analysis for four additional phenylalanine isomers, demonstrating a remarkable range of chemistries and great potential of this technology for structure elucidation in MS/MS. We are currently testing UFM for compounds included in the CASMI contests[16] and for a number of ionization modes.

ASSOCIATED CONTENT

Supporting Information. MS/MS and UFM predicted fragmentation data for four additional phenylalanine isomers along with Cartesian coordinates and energies of species depicted on Figures 1-3 are available as Supporting Information. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Bela Paizs, The Rosalind Franklin Institute, Building R113, Rutherford Appleton Laboratory, Harwell Campus, Didcot, Oxfordshire, OX11 0QX, Bela.Paizs@rfi.ac.uk

Zoltan Takats, Department of Metabolism, Digestion and Reproduction, Imperial College London, South Kensington Campus, London, UK, Z.Takats@imperial.ac.uk

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

ABBREVIATIONS

HT, Hydride transfer; PT, Proton transfer, QqToF, quadrupole/Time of Flight.

REFERENCES

(1) Aebersold, R.; Mann, M. *Nature* **2003**, *422*, 198–207.

(2) (a) Fiehn, O. *Plant. Mol. Biol.* **2002**, *48*, 155–171. (b) Patti, G.; Yanes, O.; Siuzdak, G. *Nat. Rev. Mol. Cell Biol.* **2012**, *13*, 263–269.

(3) Holčapek, M.; Liebisch, G.; Ekroos, K. *Anal. Chem.* **2018**, *90*, 4249–4257.

(4) Zaia. J. *Chem. Biol.* **2008**, *15*, 881 —892.

(5) Neumann, S.; Böcker, S. *Anal. Bioanal. Chem.* **2010**, *398*, 2779–2788.

(6) (a) McLafferty, F. W.; Turecek, F. Interpretation of Mass Spectra, 4th edition, 1993. (b) McLafferty F. W. Anal. Chem. **1959**, *31*, 82–87.

(7) (a) Wysocki V. H.; Tsaprailis, G.; Smith, L. L.; Breci, L. A.; *J. Mass Spectrom.* **2000**; *35*, 399-406. (b) Dongré, A. R.; Jones, J. L.; Somogyi, Á.; Wysocki, V. H. *J. Am. Chem. Soc.* **1996**, *118*, 8365–8374. (c) Paizs, B.; Suhai S. *Mass Spectrom. Rev.* **2005**, *24*, 508-48.

(8) van Tetering, L.; Spies, S.; Wildeman, Q.D.K. *et al. Commun. Chem.* **2024**, *7*, 30.

(9) (a) Singh, A. *Nat. Methods* **2020**, *17*, 24. (b) Singh, A. *Nat. Methods* **2023**, *20*, 33.

(10) (a) Hill, A. W.; Mortishire-Smith, R. J. *Rapid Commun. Mass Spectrom.* **2005**, *19*, 3111– 3118. (b) Wolf, S.; Schmidt, S.; Muller-Hannemann, M.; Neumann, S. *BMC Bioinformatics* **2010**, *11*, 148. (c) Allen, F.; Greiner, R.; Wishart, D. *Metabolomics* **2015**, *11*, 98–110. (d) Grimme, S. *Angew. Chem., Int. Ed.* **2013**, *52*, 6306–6312 (e) Tsugawa, H.; Kind, T.; Nakabayashi, R.; Yukihira, D.; Tanaka, W.; Cajka, T.; Saito, K.; Fiehn, O.; Arita, M. *Anal. Chem.* **2016**, *88*, 7946–7958. (f) Ridder, L.; Hooft, J.J.J.v.d.; Verhoeven, S. *Mass Spectrom.* **2014**, *3*, 0033. (g) Bocker, S.; Duhrkop, K. *J. Cheminformatics* **2016**, *8*, 5. (h) Duhrkop, K.; Shen, H.; Meusel, M.; Rousu, J.; Bocker, S. *Proc. Nat. Ac. Sci.* **2015**, *112*, 12580–12585 (2015). (i) Peironcely, J. E.; Rojas-Cherto, M.; Tas, A.; Vreeken, R.; Reijmers, T.; Coulier, L.; Hankemeier, T. *Anal. Chem.* **2013**, *85*, 3576–3583 (j) Brouard, C.; Shen, H.; Duhrkop, K.; d'Alche-Buc, F.; Bocker, S.; Rousu, J. *Bioinformatics* **2016**, *32*, i28– i36, (k) Heinonen, M.; Shen, H.; Zamboni, N.; Rousu, J. *Bioinformatics* **2012**, *28*, 2333– 2341, (l) Stravs, M.A.; Duhrkop, K.; Bocker, S.; Zamboni, N. *Nat. Methods* **2022**, *19*, 865–870. (m) Cao, L.; Guler, M.; Tagirdzhanov, A.; Lee, Y. Y.; Gurevich, A.; Mohimani, H. *Nat. Commun.* **2021**, *12*, 3718. (n) Murphy, M.; Jegelka, S.; Fraenkel, E.; Kind, T.; Healey, D.; Butler, T. arXiv 2023, arXiv:2301.11419.

(11) Xing, S.; Huan, T. *Anal. Chim. Acta* **2022**, *1200*, 339613,

(12) (a) Stewart, J. J. P. *J. Mol. Modeling*, **2007**, *13*, 1173–1213. (b) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *J. Chem. Phys.* **2010**, *132*, 154104. (c) MOPAC2016. Stewart, J.J.P, Stewart Computational Chemistry, Colorado Springs, CO, USA

(13) Blaney J. M.; Dixon J. S. Reviews in Computational Chemistry; Lipkowitz K. B.; Boyd D. B.; Eds.; John Wiley & Sons, Inc.: Hoboken, NJ, **1994**; *5*, 299–335.

(14) Somogyi A.; Wysocki V.H.; Mayer I. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 704-17.

(15) El Aribi, H.; Orlova, G.; Hopkinson, A. C.; Siu, K. W. M. *J. Phys. Chem. A* **2004**, *108*, 3844–3853.

(16) Schymanski, E.L.; Ruttkies, C.; Krauss, M. *et al. J. Cheminform.* **2017**, *9*, 22.