Title

Theoretical and experimental explorations on compound identification in metabolomics

Author
Wang, Shunyang

Publication Date
2022

Theoretical and experimental explorations on compound identification in metabolomics

By

Shunyang Wang

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Chemistry and Chemical Biology

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Prof. Oliver Fiehn, Chair

Prof. Dean J. Tantillo

Prof. Lee-Ping Wang

2022

Acknowledgements

Firstly, I would like to thank my parents – Dongjie Wang and Zhimin Wang, and my family for their love and support. I would also like to thank my girlfriend Yingxin Su and our lovely birds: Lemon & Lime, CoCo, Bai & Hui, Ginger and Bean for their warmly accompany all the time. I would like to thank professors from Chemistry Department, Lee-Ping Wang, Andrew J. Fisher, C. William McCurdy for their mentorship and help especially during the qualifying exam and dissertation. Many of my fellow Fiehn Lab and Tantillo group members were fantastic mentors and great friends during my time at UC Davis, and I am especially grateful to Clayton Bloszies, Zhitao Feng, Jesi Lee, Yuanyue Li, Jeannette Martins, Delilah Milner, Elys Paola Rodriguez, Tong Shen, Nathaniel Troup (and his nicest family), Arpana Vaniya, Luis Valdiviez for sharing the good time, helping me in the hard time and making me feel at home all the time! I have a unique graduate student journey because I had the most and best mentors: I want to thank Ivana Blazenovic for introducing me to the metabolomics field, for always helping me and being on my side; Tobias Kind for his help and advice in both life and science, physically and emotionally, for all the wonderful times we spent together; Dean J. Tantillo for his support, trust and encouragement, for guidance in quantum chemistry, for being a great role model; finally, a massive thank to Oliver Fiehn for his encyclopedic scientific knowledge, leadership and optimism, for his comprehensive help in research and writing, for all the inspiring discussions! I could not have done it without you, and I learnt what makes an excellent scientist from all of you! I still owe my thanks to many people, but I can't spend all the pages here. Thus, I would like to thank the decision four years ago to travel 6,000 miles to make this great adventure and thank fate for letting us meet!

Thesis Abstract

The high-throughput ability and sensitivity of mass spectrometry make it the most popular platform for a metabolomics study. Thus, identifying small molecules from mass spectra plays a central role in metabolomics. There are many computational techniques for mass spectra raw data processing, including feature detection, peak alignment, and mass spectral deconvolution. This dissertation focuses on the conversion between spectral and structural information. It is a spontaneous thought to identify compound structure from mass spectra by searching the query spectra against a reference library with similarity score. However, this approach is limited by the availability of reference spectra and standard compounds. To bridge the gap, different computation tools based on fragmentation trees are developed to help annotate spectra. The other thought is to generate in-silico spectra from the structural information. Several compound databases, such as PubChem, KEGG, HMDB and CHEBI can be the source of structural information. Machine learning and heuristic approaches are trained from spectral knowledge to generate in-silico spectra. However, all those approaches cannot predict beyond the known data. As a first-principles method, Quantum chemistry modeling is only based on the rules of quantum mechanics and can help us explore the unknown space of metabolites. Yet, the quantum chemical simulation on molecules over 600 Da is too expensive to get accurate results. Chemical ionization with methane reagent gas can help identify the molecular ion species and help the unknown metabolite identification.

This dissertation describes computational studies using molecular dynamics on in-silico mass spectra of small molecule generation. Chapter 1 provides an overview of applications of quantum chemistry to generate in-silico mass spectra. Chapter 2 showcases the performance of quantum chemistry simulations on small molecules and probes the parameter space to find potential improvements to the existing method. The conformational flexibility effect is also explored and has no correlation with prediction accuracy. Chapter 3 describes a new workflow to include chemical derivation, especially silylation in the quantum chemistry calculation. Different compound classes including organic acid, alcohol, amide, amine and thiols are simulated and compared against experimental mass spectra. The molecular dynamics trajectories are also investigated to find missed fragmentations from rearrangements. Chapter 4 provides a new algorithm that introduces excited state calculation into the molecular dynamics prediction of mass spectra. The new algorithm can predict more fragmentation reactions that are missing in previous studies and as a result, the mass spectral similarity scores are increased, and simulation is more accurate. Chapter 4 also discusses the limitations of molecular dynamics simulation time and the lack of rearrangement reactions. Chapter 5 provides another routine from the experimental side to help unknown metabolites identification with methane chemical ionization and quadrupole-time of flight mass spectrometry.

Go, get you home, you fragments![1]

[1] *Coriolanus: Act 1, Scene 1, William Shakespeare*

Table of Contents

Chapter 1: An Introduction to Quantum-Mechanical Calculations for Mass spectrometry

1.1. Quantum Chemistry

With the development of computer science, quantum chemistry becomes an available tool with sufficient computational efficiency and chemical accuracy. The core idea of quantum chemistry is to find approximate solution of the many-electron Schrödinger equation. The Hartree-Fock theory and density functional theory (DFT) are the earliest and most popular quantum chemistry methods in 1920s. Since then, quantum chemistry has involved in chemical sciences by predicting and interpreting experimental observations, such as NMR, IR, UV, and Raman spectra. In contrast to those applications, no straightforward procedure exists for quantum-chemistry-based prediction of mass spectra. For example, prediction of IR and Raman vibrational spectra became possible by 1965 using simple force fields and in the late 1970s using ab initio calculations. While single fundamental fragmentations can be predicted with the help of quantum chemistry, the cascade of reactions and rearrangements resulting from multiple reaction pathways, and most importantly the m/z peak abundances from complex molecules, have been highly difficult to deduce.

In a major breakthrough and one of the most important discoveries in computational MS, Grimme published the Quantum Chemistry Electron Ionization MS (QCEIMS) program for the first principle calculation of 70 eV mass spectra in 2013.[1] QCEIMS is discussed in more detail in the next section.

1.2. Electron ionization (EI)

Electron ionization (EI) MS (70 eV) is an established analytical technique and is commonly coupled to GC for analysis of small molecules below 400 Da. Electrons are emitted from a heated filament and focused on gaseous neutral molecules. When the accelerated electrons hit the neutral molecule, radical cations are formed, and another electron is ejected. The vibrationally excited carbocations then undergo further bond dissociations and fragmentations on a very fast time scale. The smaller mostly singly charged fragment ions are then accelerated towards a detector and recorded as spectral signals. The ionization efficiency at 70 eV is the highest and most molecules can be ionized at this energy, allowing for creation of reproducible mass spectra.[2] The power of GC-MS lies in the fact that the instrument industry has subsequently standardized the EI source energies to 70 eV, resulting in the availability of reproducible spectra and available databases to search[3]. Gas chromatography coupled to tandem mass spectrometry (GC-MS/MS) has not reached a breakthrough yet, due to the more complex instrumentation and missing MS/MS spectral databases for spectral matching[4].

Historically, the interpretation of EI-derived spectra depended on statistical rate theory[5-9] and investigation of kinetic processes, especially work based on quasi-equilibrium theory (QET)[10] and Rice-Ramsperger-Kassel-Marcus (RRKM)[11-14] theory which can be used to predict rate constants. Many of the classical investigations of 70 eV radical cations or anions are limited to single ion species or specific molecules due to the complexity of fragmentation and rearrangement reactions.

The main disadvantage of traditional QET/RRKM approaches is that rate calculations are based on the selection of specific ion transition states and activated complexes on the PES. With increasing atom numbers, the complexity of the reaction space rises exponentially and would require a priori knowledge of reaction pathways that are not always available[15]. Methods such as the global reaction route mapping (GRRM) strategy[16], the AutoMeKin software[17] or the Chemical Trajectory Analyzer (ChemTraYzer) software[18] have been developed to systematically and automatically explore the reaction space[19].

The QCEIMS approach published by Grimme in 2013 combines Born–Oppenheimer molecular dynamics (BOMD), a type of AIMD, with statistical sampling to predict 70 eV mass spectra[1]. In contrast to other methods, QCEIMS is purely based on physical and chemical principles and can calculate mass spectra from any given molecule. Using a combination of ab initio molecular dynamics (AIMD) and stochastic sampling across hundreds of reaction pathways, the correct m/z value of ions and their associated abundances can be predicted. More excitingly, all reaction trajectories are retained and allow for a "look inside" the reaction processes of a mass spectrometer, which then makes it possible to investigate all fragmentations and rearrangements individually. To achieve a balance between efficiency and accuracy, QCEIMS can calculate on various levels of theory, including semiempirical models OM2/OM3[20], DFTB+[21], GFNn-xTB[22] and several DFT methods. The complex relaxation processes from the electronically excited state of the precursor ion are modeled by limiting the reaction on ionic ground state PES. The impact excess energy is converted to kinetic energy by a heating process, during which the atomic velocities are scaled to a preset impact excess energy value. Such a simple electronic structure can handle the fragmentation reactions, and can its ability to give a reasonable result is one of the key innovations of QCEIMS. [1]

The QCEIMS software is coupled with several independent software packages such as ORCA[23-24], TurboMole[25], MOPAC[26], MNDO99[27-28] and DFTB+[21]. Most importantly, the latest independent and therefore stand-alone version of QCEIMS directly implements the GFN-XTB method. This allows for simple installation and practical use of QCEIMS in any research environment with access to HPC. The only required input is a chemical structure. Because the GFNn-xTB[22] methods are parameterized to elements up to Z=86, they are applicable to the most common molecules and therefore provide calculations of 70 eV mass spectra with metalloids such as silicone[29-30]. This is important, because the trimethylsilyl group (TMS) is often used during GC-MS derivatization experiments[31].

One of the advantages of QCEIMS is that reaction pathways are automatically recorded as MD trajectories during the simulation. This allows for comprehensive investigation of the fragmentation mechanism. However, the confirmation of such reactions would require comprehensive investigations, because for any given reaction, a multitude of possibilities exists. In the original paper[1], Grimme found that most of the primary fragmentations occur within 2–3 ps, while secondary fragmentation reactions take much longer but are important in larger systems. Many well-known reaction pathways in MS are accurately reported by QCEIMS, including α-cleavage[15], McLafferty rearrangement, retro-Diels-Alder[32] reaction and CO loss[1]. For molecules with several tautomers, a combination of initial conditions based on Boltzmann population can be used to improve simulation accuracy.[32]

In 2016, Cautereels et al. described a different method for the calculation of 70 eV mass spectra using empirical rules for limiting the number of fragmentations along the PES based on DFT calculations[33]. The rules include observations of bond strengths, bond cleavages (that are thermodynamically controlled) and 1,4-rearrangements and McLafferty rearrangements (that are kinetically controlled).[34] The procedure includes conformational sampling and calculation of Boltzmann weights including the calculation of the most stable radical cations. Homolytic and heterolytic fragmentation pathways are calculated under observation of the heuristic rules. Final peak abundances are determined based on a formula that includes the average of energies of the fragmentation pathways and specific fragments. Such an approach could become very useful in the future for detailed investigations of reaction pathways using classical transition state theory.

Similar to the evaluation of machine-learning prediction methods such as CFM-ID,[35] quantum chemical models have to be rigorously tested by comparing theoretical predictions against experimental reference spectra.[36] Similarity match scores and compound rankings should be reported.[3] This can be done with the National Institute of Standards and Technology (NIST) MS Search program and the NIST and MassBank of North America (MoNA) mass spectral databases[4].

1.3. QCEIMS computational costs and accuracy

The QCEIMS protocol contains three types of quantum mechanical calculations: energy/force calculations to generate the potential energy surface for MD, molecular orbital (MO) calculations to determine internal excess energies and ionization potential (IP) calculations of each fragment to generate the statistical charges. The original version of QCEIMS utilizes DFT methods for MO and IP calculations, whereas the energy/force calculations for the time-consuming MD steps use the OM2/OM3[37] orthogonal corrected semiempirical methods.

For example, the simulation of the 70 eV EI mass spectrum of anisole (C7H8O, MW=108.057 Da) requires 1.2 million individual MD steps and 82 minutes of computational time on 16 CPU cores at the OM2 level. The choice of the underlying method significantly affects simulation speed. The GFNn-xTB methods[29-30] will be 10–20 times slower than the semiempirical OM2[20] simulations, while purely DFT-based MD can be 100 or more times slower than the semiempirical methods.

The computational cost for semiempirical methods is usually much smaller than ab initio methods (OM2/PM6[38] < GFNn-xTB < DFT), whereas in terms of accuracy, we see the opposite trend with

DFT being the most accurate method (DFT > GFNn-xTB > OMx/PMx). Interestingly, chemical bonds are more easily dissociated in semiempirical simulations relative to the more accurate DFT simulations[39]. Therefore, more accurate calculations of the PES may require even more simulation steps with longer fragmentation processes.

Because DFT methods are closer to the ‘exact’ PES, they should be used as a reference in evaluating more approximate models,[39] but their increased computational cost puts them out of reach for simulating EI mass spectra of larger molecules. We are optimistic that GPU-accelerated implementations of DFT methods in software such as TeraChem[40-41] or Fermions++ may lead to fast high-accuracy simulations[42].

However, the usage of fractional occupation number weighted densities[43] can reproduce some properties of multireference wavefunctions, making it a possible low-cost alternative for treating multireference systems. On the other hand, when the energy gap between the excited state and ground state goes to zero, the Born-Oppenheimer approximation and single reference methods used in QCEIMS can break down. The treatment of highly excited electronic states using multi-reference methods,[44-46] such as the states accessed during QCEIMS, is under active investigation and can guide the development of improved simulation approaches in the future.

The accuracy of predicted in silico spectra has to be evaluated against diverse and large number of experimentally measured spectra.[3] QCEIMS (with OM2/OM3) performs on the same accuracy level as the best available machine learning algorithms such as CFM-ID[35]. The QCEIMS method also has the advantage that any given molecule can be calculated. The reason is that machine learning methods require experimental training data, while QCEIMS as an ab initio method is only based on physical and chemical principles. The most important question for practitioners is the practical use of algorithms in daily research applications. Currently, it is not possible to calculate most compounds with high similarity match scores (>850). It is also not yet possible to determine the quality of predictions in advance due to the stochastic nature of the computations. It is foreseeable that with improved accuracy of future versions of QCEIMS and related methods, a wide range of in silico spectra can be used for training in machine leaning to allow for even faster simulation of in silico mass spectra from all known compounds.

1.4. Coupling EI to other spectroscopic methods

While GC-MS mass spectra at 70 eV can give structural insights, it is not possible to fully interpret all mass spectra because in many cases the molecular ion is not observed and following individual fragmentations is not directly possible. Techniques such as chemical ionization or cold EI can help increase the stability and abundance of the molecular ion.[47] Furthermore, integrating parallel analysis techniques such as IR, Raman, and UV will allow for easier structure-to-spectrum identification using quantum mechanical calculations of optical spectra[48]. In such a case, MS and optical spectroscopy experiments are performed in parallel and the resulting spectra can be investigated theoretically using quantum chemistry methods or QM/MM[49]. For example, threshold photoionization mass spectra can be acquired with photoelectron photoion coincidence (PEPICO) spectroscopy and can be coupled with DFT calculations to gain insights into fragmentation behavior.[50-52] In particular, coupling MS with IR multiple-photon dissociation spectroscopy (IRMPD) seems to be an excellent way for interpreting dissociation pathways by combining experiments with quantum chemical calculations.[53] While such instrumental setups are complex and expensive, they show the possibilities of instrumental integration with quantum mechanical computations. Such techniques, while discussed here in detail for EI, can also be coupled to other methods such as ESI and CID MS/MS.

1.5. Reference

1. Grimme, S., Towards First Principles Calculation of Electron Impact Mass Spectra of Molecules. *Angewandte Chemie International Edition* **2013,** *52* (24), 6306-6312.

2. Gross, J. H., *Mass Spectrometry: A Textbook*. Springer: Berlin Heidelberg, 2011.

3. Stein, S., Mass spectral reference libraries: an ever-expanding resource for chemical identification. **2012**.

4. Kind, T.; Tsugawa, H.; Cajka, T.; Ma, Y.; Lai, Z.; Mehta, S. S.; Wohlgemuth, G.; Barupal, D. K.; Showalter, M. R.; Arita, M., Identification of small molecules using accurate mass MS/MS search. *Mass spectrometry reviews* **2018,** *37* (4), 513-532.

5. Armentrout, P. B.; Ervin, K. M.; Rodgers, M. T., Statistical Rate Theory and Kinetic Energy-Resolved Ion Chemistry: Theory and Applications. *The Journal of Physical Chemistry A* **2008,** *112* (41), 10071-10085.

6. Tureček, F.; Julian, R. R., Peptide Radicals and Cation Radicals in the Gas Phase. *Chemical Reviews* **2013,** *113* (8), 6691-6733.

7. Rennie, E. E.; Cooper, L.; Shpinkova, L. G.; Holland, D. M. P.; Shaw, D. A.; Guest, M. F.; Mayer, P. M., Methyl t-Butyl Ether and Methyl Trimethylsilyl Ether Ions Dissociate near Their Ionization Thresholds: A TPES, TPEPICO, RRKM, and G3 Investigation. *The Journal of Physical Chemistry A* **2009,** *113* (20), 5823-5831.

8. Tsyshevsky, R. V.; Garifzianova, G. G.; Shamov, A. G.; Khrapkovskii, G. M., Fragmentation reactions in the 1-nitropropane radical cation induced by γ-hydrogen shift: Ab initio study. *International Journal of Mass Spectrometry* **2014,** *369*, 36-43.

9. Solano, E. A.; Mayer, P. M., A complete map of the ion chemistry of the naphthalene radical cation? DFT and RRKM modeling of a complex potential energy surface. *The Journal of Chemical Physics* **2015,** *143* (10), 104305.

10. Rosenstock, H. M.; Wallenstein, M. B.; Wahrhaftig, A. L.; Eyring, H., Absolute Rate Theory for Isolated Systems and the Mass Spectra of Polyatomic Molecules. *Proc Natl Acad Sci U S A* **1952,** *38* (8), 667-678.

11. Rice, O. K.; Ramsperger, H. C., THEORIES OF UNIMOLECULAR GAS REACTIONS AT LOW PRESSURES. *Journal of the American Chemical Society* **1927,** *49* (7), 1617-1629.

12. Kassel, L. S., Studies in Homogeneous Gas Reactions. I. *The Journal of Physical Chemistry* **1928,** *32* (2), 225-242.

13. Marcus, R. A.; Rice, O. K., The Kinetics of the Recombination of Methyl Radicals and Iodine Atoms. *The Journal of Physical Chemistry* **1951,** *55* (6), 894-908.

14. Marcus, R. A., Unimolecular Dissociations and Free Radical Recombination Reactions. *The Journal of Chemical Physics* **1952,** *20* (3), 359-364.

15. Bauer, C. A.; Grimme, S., How to Compute Electron Ionization Mass Spectra from First Principles. *The Journal of Physical Chemistry A* **2016,** *120* (21), 3755-3766.

16. Maeda, S.; Harabuchi, Y.; Ono, Y.; Taketsugu, T.; Morokuma, K., Intrinsic reaction coordinate: Calculation, bifurcation, and automated search. *International Journal of Quantum Chemistry* **2015,** *115* (5), 258-269.

17. Vázquez, S. A.; Otero, X. L.; Martinez-Nunez, E., A trajectory-based method to explore reaction mechanisms. *Molecules* **2018,** *23* (12), 3156.

18. Döntgen, M.; Przybylski-Freund, M.-D.; Kröger, L. C.; Kopp, W. A.; Ismail, A. E.; Leonhard, K., Automated discovery of reaction pathways, rate constants, and transition states using reactive molecular dynamics simulations. *Journal of chemical theory and computation* **2015,** *11* (6), 2517-2524.

19. Maeda, S.; Ohno, K.; Morokuma, K., Systematic exploration of the mechanism of chemical reactions: the global reaction route mapping (GRRM) strategy using the ADDF and AFIR methods. *Phys. Chem. Chem. Phys.* **2013,** *15* (11), 3683-3701.

20. Weber, W.; Thiel, W., Orthogonalization corrections for semiempirical methods. *Theoretical Chemistry Accounts* **2000,** *103* (6), 495-506.

21. Cui, Q.; Elstner, M., Density functional tight binding: values of semi-empirical methods in an ab initio era. *Phys. Chem. Chem. Phys.* **2014,** *16* (28), 14368-14377.

22. Grimme, S.; Bannwarth, C.; Shushkov, P., A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements (Z = 1–86). *Journal of Chemical Theory and Computation* **2017,** *13* (5), 1989-2009.

23. Neese, F., Software update: the ORCA program system, version 4.0. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2018,** *8* (1), e1327.

24. Neese, F., The ORCA program system. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2012,** *2* (1), 73-78.

25. Furche, F.; Ahlrichs, R.; Hättig, C.; Klopper, W.; Sierka, M.; Weigend, F., Turbomole. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2014,** *4* (2), 91-100.

26. Stewart, J. J., MOPAC: a semiempirical molecular orbital program. *Journal of computer-aided molecular design* **1990,** *4* (1), 1-103.

27. Dral, P. O.; Wu, X.; Thiel, W., Semiempirical quantum-chemical methods with Orthogonalization and dispersion corrections. *Journal of chemical theory and computation* **2019,** *15* (3), 1743-1760.

28. Dewar, M. J. S.; Thiel, W., Ground states of molecules. 38. The MNDO method. Approximations and parameters. *Journal of the American Chemical Society* **1977,** *99* (15), 4899-4907.

29. Ásgeirsson, V.; Bauer, C. A.; Grimme, S., Quantum chemical calculation of electron ionization mass spectra for general organic and inorganic molecules. *Chemical Science* **2017,** *8* (7), 4879-4895.

30. Koopman, J.; Grimme, S., Calculation of Electron Ionization Mass Spectra with Semiempirical GFNn-xTB Methods. *ACS Omega* **2019,** *4* (12), 15120-15133.

31. Zaikin, V.; Halket, J. M., *A handbook of derivatives for mass spectrometry*. IM publications: 2009.

32. Bauer, C. A.; Grimme, S., Automated Quantum Chemistry Based Molecular Dynamics Simulations of Electron Ionization Induced Fragmentations of the Nucleobases Uracil, Thymine, Cytosine, and Guanine. *European Journal of Mass Spectrometry* **2015,** *21* (3), 125-140.

33. Cautereels, J.; Claeys, M.; Geldof, D.; Blockhuys, F., Quantum chemical mass spectrometry: ab initio prediction of electron ionization mass spectra and identification of new fragmentation pathways. *Journal of mass spectrometry* **2016,** *51* (8), 602-614.

34. Morton, T. H., Neutral products from gas phase rearrangements of simple carbocations. *Adv. Gas-Phase Ion Chem* **2001,** *4*, 213-256.

35. Allen, F.; Pon, A.; Greiner, R.; Wishart, D., Computational prediction of electron ionization mass spectra to assist in GC/MS compound identification. *Analytical chemistry* **2016,** *88* (15), 7689-7697.

36. Spackman, P. R.; Bohman, B.; Karton, A.; Jayatilaka, D., Quantum chemical electron impact mass spectrum prediction for de novo structure elucidation: Assessment against experimental reference data and comparison to competitive fragmentation modeling. *International Journal of Quantum Chemistry* **2018,** *118* (2), e25460.

37. Dral, P. O.; Wu, X.; Spörkel, L.; Koslowski, A.; Thiel, W., Semiempirical quantum-chemical orthogonalization-corrected methods: Benchmarks for ground-state properties. *Journal of chemical theory and computation* **2016,** *12* (3), 1097-1120.

38. Stewart, J. J., Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. *J Mol Model* **2013,** *19* (1), 1-32.

39. Bauer, C. A.; Grimme, S., Elucidation of Electron Ionization Induced Fragmentations of Adenine by Semiempirical and Density Functional Molecular Dynamics. *The Journal of Physical Chemistry A* **2014,** *118* (49), 11479-11484.

40. Giorgi, G., Mass Spectrometry and Tandem Mass Spectrometry: An Overview. In *Detection of Chemical, Biological, Radiological and Nuclear Agents for the Prevention of Terrorism*, Springer: 2014; pp 17-31.

41. Titov, A. V.; Ufimtsev, I. S.; Luehr, N.; Martinez, T. J., Generating efficient quantum chemistry codes for novel architectures. *Journal of chemical theory and computation* **2013,** *9* (1), 213-221.

42. Kussmann, J. r.; Ochsenfeld, C., Hybrid CPU/GPU integral engine for strong-scaling ab initio methods. *Journal of chemical theory and computation* **2017,** *13* (7), 3153-3159.

43. Bauer, C. A.; Hansen, A.; Grimme, S., The Fractional Occupation Number Weighted Density as a Versatile Analysis Tool for Molecules with a Complicated Electronic Structure. *Chemistry (Weinheim an der Bergstrasse, Germany)* **2017,** *23* (25), 6150-6164.

44. Roos, B.; Taylor, P.; Siegbahn, P., A Complete Active Space Scf Method (casscf) Using a Density-Matrix Formulated Super-Ci Approach. *Chem. Phys.* **1980,** *48* (2), 157-173.

45. Andersson, K.; Malmqvist, P.; Roos, B., 2nd-Order Perturbation-Theory with a Complete Active Space Self-Consistent Field Reference Function. *J. Chem. Phys.* **1992,** *96* (2), 1218-1226.

46. Hirao, K., Multireference Moller-Plesset Method. *Chem. Phys. Lett.* **1992,** *190* (3-4), 374-380.

47. Keshet, U.; Goldshlag, P.; Amirav, A., Pesticide analysis by pulsed flow modulation GCxGC-MS with Cold EI—an alternative to GC-MS-MS. *Analytical and bioanalytical chemistry* **2018,** *410* (22), 5507-5519.

48. Buchalter, S.; Marginean, I.; Yohannan, J.; Lurie, I. S., Gas chromatography with tandem cold electron ionization mass spectrometric detection and vacuum ultraviolet detection for the comprehensive analysis of fentanyl analogues. *Journal of Chromatography A* **2019,** *1596*, 183-193.

49. Jacovella, U.; da Silva, G.; Bieske, E. J., Unveiling New Isomers and Rearrangement Routes on the C7H8+ Potential Energy Surface. *The Journal of Physical Chemistry A* **2019,** *123* (4), 823-830.

50. Wu, X.; Zhou, X.; Hemberger, P.; Bodi, A., A guinea pig for conformer selectivity and mechanistic insights into dissociative ionization by photoelectron photoion coincidence: fluorocyclohexane. *Phys. Chem. Chem. Phys.* **2020**.

51. Candian, A.; Bouwman, J.; Hemberger, P.; Bodi, A.; Tielens, A. G., Dissociative ionisation of adamantane: a combined theoretical and experimental study. *Phys. Chem. Chem. Phys.* **2018,** *20* (8), 5399-5406.

52. Majer, K.; Signorell, R.; Heringa, M. F.; Goldmann, M.; Hemberger, P.; Bodi, A., Valence Photoionization of Thymine: Ionization Energies, Vibrational Structure, and Fragmentation Pathways from the Slow to the Ultrafast. *Chemistry–A European Journal* **2019,** *25* (62), 14192-14204.

53. Bouwman, J.; Horst, S.; Oomens, J., Spectroscopic characterization of the product ions formed by electron ionization of adamantane. *ChemPhysChem* **2018,** *19* (23), 3211-3218.

Chapter 2: Predicting in-silico electron ionization mass spectra using quantum chemistry

2.1. Abstract

Compound identification by mass spectrometry needs reference mass spectra. While there are over 102 million compounds in PubChem, less than 300,000 curated electron ionization (EI) mass spectra are available from NIST or MoNA mass spectral databases. Here, we test quantum chemistry methods (QCEIMS) to generate in-silico EI mass spectra (MS) by combining molecular dynamics (MD) with statistical methods. To test the accuracy of predictions, in-silico mass spectra of 451 small molecules were generated and compared to experimental spectra from the NIST 17 mass spectral library. The compounds covered 43 chemical classes, ranging up to 358 Da. Organic oxygen compounds had a lower matching accuracy, while computation time exponentially increased with molecular size. The parameter space was probed to increase prediction accuracy including initial temperatures, the number of MD trajectories and impact excess energy (IEE). Conformational flexibility was not correlated to the accuracy of predictions. Overall, QCEIMS can predict 70 eV electron ionization spectra of chemicals from first principles. Improved methods to calculate potential energy surfaces (PES) are still needed before QCEIMS mass spectra of novel molecules can be generated at large scale.

2.2. Introduction

Mass spectrometry is the most important analytical technique to detect and analyze small molecules. Gas chromatography coupled to mass spectrometry (GC/MS) is frequently used for such molecules and has been standardized with electron ionization (EI) at 70 eV more than 50 years ago [1]. Yet, current mass spectral libraries are still insufficient in breadth and scope to identify all chemicals detected: there are only 306,622 EI-MS compound spectra in the NIST 17 mass spectral database [2], while PubChem has recorded 102 million known chemical compounds of which 14 million are commercially available. That means there is a large discrepancy between compounds and associated reference mass spectra [3]. For example, less than 30% of all detected peaks can be identified in GC-MS based metabolomics [4]. To solve this problem, the size and complexity of MS libraries must be increased. Several approaches have been developed to compute 70 eV mass spectra, including machine learning [5-6], reaction rule-based methods [7] and a method based on physical principles, the recently developed quantum chemical software Quantum Chemical Electron Ionization Mass Spectrometry (QCEIMS). [8]

While empirical and machine learning methods depend on experimental mass spectral data for development, quantum chemical methods only consider physical laws. Thus, in principle, QCEIMS can compute spectra for any given compound structure. Yet, approximations and parameter estimations are needed to allow predictions in a timely manner, reducing the accuracy of QCEIMS predictions. QCEIMS uses Born–Oppenheimer Molecular Dynamics (MD) to calculate fragment ions within picosecond reaction times with femtosecond intervals for the MD trajectories. A statistical sampling process is used to count the number of observed fragments and to derive the peak abundances for each observed ion [9] (Figure 2-1).

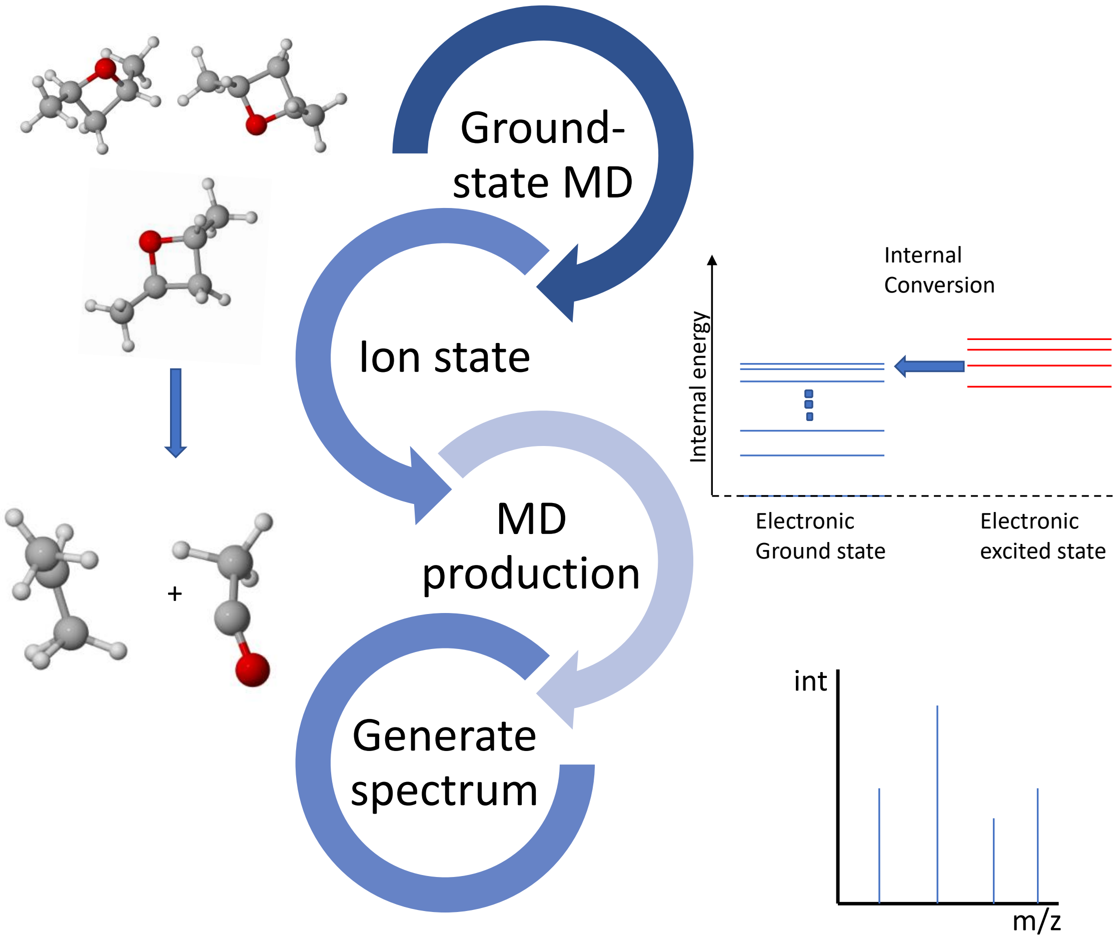


Figure 2-1 Workflow of QCEIMS. (1) generating conformers by equilibrium molecular dynamics; (2) ionizing each neutral starting structure by assigning impact excess energy (IEE) to kinetic energy; (3) generating EI fragments by parallel molecular dynamics; (4) assigning charges on each fragment using ionization potential (IP) energies and peak intensity counts, then assembling fragments to obtain summary spectra.

It is unclear how reliable QCEIMS predictions are because the methods have not yet been tested on hundreds of compounds. MS matching accuracy is neither easily predictable nor quantifiable, because theoretical and experimental EI mass spectra have not been compared on a large scale. To test how structural constraints affect prediction accuracies, we utilized the QCEIMS method to predict spectra of 451 compounds with different molecular flexibility, sizes and chemical classes.

2.3. Methods

2.3.1. Molecular structure preparation

We used ChemAxon's [10] MarvinView and MarvinSketch (v18.23) to manipulate structures. First, small molecules were manually chosen from the NIST 17 mass spectral database. 3-D coordinates were generated using the Merck Molecular Force Field (MMFF94) [11] with Avogadro (v1.2.0) [12] in Molfiles (*.mol) format. We used OpenBabel (v2.3.90) [13] to convert structures to the TurboMole format (*.tmol) as required by the QCEIMS (v2.16) program. We used the QCEIMS plotms program to export JCAMP-DC mass spectra. External additional conformers were generated independently by conformational search packages, including GMMX from Gaussian[14], the conformer generator in ChemAxon's MarvinSketch and by using RDKit [15] (v2019.03.1).

2.3.2. Parallel cluster calculation with QCEIMS

We utilized the QCEIMS program for in-silico fragmentation with the following parameters: 70 eV ionization energy, 500K initial temperature and 0.5 femtosecond (fs) time steps. For molecular dynamics, we used the semiempirical OM2 method [16] (Quantum-Chemical-Orthogonalization-Corrected Method) using the MNDO99 (v2013) [17] software. The impact excess energy (IEE) satisfied the Poisson type distribution. The Orca software (3.0.0) [18] was employed to calculate the vertical SCF ionization potential at the PBE0 [19] - D3 [20] /SV(p) [21] level.

We conducted QCEIMS calculations on cluster nodes equipped with two Intel Xeon E5-2699Av4 CPUs, 44 cores and 88 threads in total, operated at 2.40 GHz. Each node was equipped with 128 GByte RAM and a 240 GByte Intel DCS3500 datacenter grade SSD. In order to conduct and monitor the calculation process, we developed a SLURM job script to submit batch jobs. While the initial ground state molecular dynamics simulation is only single-threaded, all subsequent calculations were massively paralleled. Because QCEIMS executes multiple trajectory calculations at once, we oversubscribed the parallel number of CPU threads to be used to 66 (instead of 44) during QCEIMS production runs. Such a CPU oversubscription is possible, because molecular dynamics (OM2 with MNDO99) and density functional theory (DFT) calculations are executed in a heterogeneous way by different programs [8]. The speed advantage of using more threads than CPU cores available was confirmed with benchmarks.

2.3.3. Similarity score evaluation

QCEIMS generated several outputs and logging files, including the in-silico mass spectrum in JCAMP exchange format (*.jdx), structures of fragments (*.xyz) and molecular dynamics trajectories (*.xyz). We then used experimental mass spectra from the NIST17 database as references to compare with our computational results. In GC-MS, mass spectral similarity scores (0 to 1000) describe how well experimental spectra match recorded library spectra [22-23]. Here we used the same principle for QCEIMS-generated spectra as input. Similarity scores below 500 are usually not considered for annotation of compounds. While similarity scores above 700 may represent true matches, only scores above 850 are regularly used for direct compound identifications in GC-MS experiments [24]. Here we used two different kinds of similarity scores (see equations 1-3):

$$Cos\ = \sqrt{\frac{(\Sigma I_U I_L)^2}{\Sigma I^2{}_L\, \Sigma I^2{}_U}} \tag{1}$$

$$Dot = \sqrt{\frac{(\Sigma W_U W_L)^2}{\Sigma W^2{}_L\, \Sigma W^2{}_U}} \tag{2}$$

$$W = [Peak\ Intensity]^m [Mass]^n \tag{3}$$

We wrote a Python (v3.6) script to read mass spectra and analyze the similarity by (a) cosine similarity (*Cos*, eq 1) (b) weighted dot-product similarity (*Dot*, eq 2); with the test data, we set the parameters as m=0.6 and n=3. Our method calculates very similar values as implemented in the NIST MS Search program (see Supporting Information). To validate some of our simulations, we also used MassFrontier 7.0 [7] to generate fragmentation pathways and compared them with the mechanisms found from our trajectories. MassFrontier can predict fragmentation pathways from general fragmentation rules and mechanisms recorded in its literature database.

2.3.4. Flexibility analysis

To describe molecular flexibility, we used two molecular descriptors: the number of rotatable bonds (RBN) [25] and Kier flexibility index (PHI) [26]. The RBN is the number of bonds for which rotation around themselves is expected to be associated with low (< 5 kcal/mol) barriers, excluding ring bonds and amide bonds. The Kier flexibility index is a structure-based property calculated from atom numbers, rings, branches and covalent radii. With fewer rotatable bonds and lower Kier flexibility index, the molecule has less conformational flexibility. The software AlvaDesc [27] (v1.0.8) is utilized to compute these properties. We used both Microsoft EXCEL for Mac and Matplotlib (v3.1.1) to analyze and visualize the data.

2.4. Results

2.4.1. Comparison of in-silico and experimental spectra of example molecules

Following the general workflow, we first tested the QCEIMS software on two trajectories for a simple molecule, 3-cyclobutene-1,2-dione (Figure 2-2). The observed fragment ions yielded an excellent weighted dot-product similarity score of 972 and a cosine similarity of 839. When analyzing the trajectories to show the fragmentation pathways, we found clear evidence of the mechanisms by which the three main product ions observed in the experimental mass spectrum were produced (m/z 82, 54, 26), i.e., molecular ion, a neutral loss of carbon monoxide $[M-CO]^+$ and loss of another carbon monoxide to yield $[M-2CO]^+$ (Figure 2-2). Trajectory 2 lasted only 402 fs until the maximum of three fragments per trajectory was achieved (set in the QCEIMS source code), while trajectory 1 lasted 656 fs, because the initial two fragments reached a stable state and did not fragment further for a long time. The QCEIMS predictions also agreed with mechanisms predicted by the heuristic rule-based commercial MassFrontier software, showing first an α-cleavage followed by a CO molecule loss. This simple example shows that QCEIMS can generate correct molecular fragments and predict reasonable reaction mechanisms.

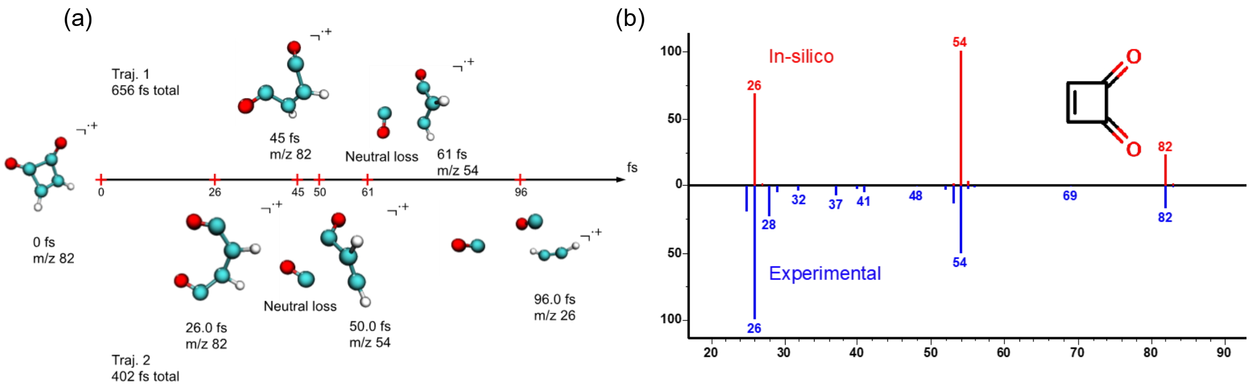


Figure 2-2. Example for correctly predicting experimental EI mass spectra through molecular dynamics.

(a) Fragmentation trajectories of 3-cyclobutene-1,2-dione to generating EI fragment m/z 54 (upper panel) and m/z 26 (lower panel) (b) Quantum chemistry molecular dynamics in-silico spectrum (upper panel) versus experimental mass spectrum (lower panel

Here we show six molecules (Figure 2-3 a-f) as examples for QCEIMS predicted spectra versus experimental library spectra (Table 1). These examples demonstrate that QCEIMS yields different prediction accuracies. The examples also show different degrees of molecular flexibility. For each molecule, spectra showed specific characteristics that are here explained in brief.

Table 2-1 Mass spectral similarities of QCEIMS simulations against experimental spectra for select compounds

Name	InChIKey (short)*	M.W.**	RBN	PHI	Dot	Cos
2,4-Dimethyl-oxetane	KPPWZEMUMPFHEX	86.07	2	2.64	414	729
2-Nonene	IICQZTQZQSBHBY	126.27	5	7.52	789	762
2-Propynyloxy Benzene	AIQRJSXKXVZCJO	132.06	0	1.17	379	426
Furan	YLQBMQCUIZJEEH	68.08	0	0.55	988	806
1,8-Nonadiene	VJHGSLHHMIELQD	124.25	6	7.05	163	713
Adamantane	ORILYTVJVMAKLC	136.13	0	1.18	883	678

* first 14-characters of full InChIKey; **M.W. is the molecular weight in Daltons (Da); RBN (rotatable bond number) and PHI (Kier flexibility index) are rigidity descriptors and Dot and Cos are mass spectral similarity scores.

2,4-dimethyl-oxetane (Figure 3a): With a weighted dot-product score of 417, this spectrum represents a low-quality in-silico prediction. We need to clarify that, for simplicity, we only calculated the spectrum of cis-2,4-dimethyl-oxetane, while its reference spectrum in NIST 17 mass spectral library contains no stereochemistry information because neither EI-MS nor chromatography technology can easily differentiate diastereomers. The experimental spectrum showed a low-intensity $[M]^{\cdot+}$ at m/z 86 and initial neutral losses of a methyl-group and water (m/z 71 and m/z 68). QCEIMS did not predict these initial losses. Indeed, the high number of experimental fragment ions suggest that this molecule splits readily along multiple reaction pathways, most likely through breaking the molecular ether-bonds that subsequently break into smaller fragments. The main fragment ions at *m/z* 42 and m/z 44 were correctly predicted by QCEIMS as $C_3H_6^+$ and $C_2H_4O^+$ but not by the rule-based software MassFrontier. This case suggests that quantum mechanics-based simulations can produce novel reaction pathways that are absent from rule-base software predictions.

2-Nonene (Figure 3b): The in-silico spectrum of 2-nonene was highly similar to the experimental spectrum with dot-product match of 789. The main fragment ion at m/z 55 and the $[M]^{\cdot+}$ at m/z 126 were very well reproduced. However, ion abundances of $[M-1]^+$, $[M-2]^+$, $[M-3]^+$ and $[M-4]^+$ were overestimated. In QCEIMS, these ions resulted from loss of several atomic or molecular hydrogens, suggesting that these bonds were fragmented more easily under semiempirical methods [23] than under experimental conditions.

Aromatic systems (Figure 3c and 3d): Both 2-propynyloxy benzene and furan were aromatic oxygen-containing molecules with low PHI values (1.71 and 0.55, respectively). Although the presence of most fragment ions was correctly predicted by QCEIMS for both molecules, dot-product similarity scores were radically different with a dot-product of 379 for 2-propynyloxy benzene and a dot-product similarity of 988 for furan). For 2-propynyloxy benzene, this low matching score was caused by the absence of an experimental $[M]^{\cdot+}$ at m/z 132 that was largely overestimated in the in-silico spectrum. The fragmentation base ion (at 100% intensity) at m/z 93 represents the stable phenol ion and a neutral loss of C_3H_3, while the experimentally observed fragment at m/z 95 was missed in the QCEIMS prediction. At the same time, the presence of the $C_3H_3^+$ product ion at m/z 39 (and a neutral loss of a phenol moiety) was overestimated by the

QCEIMS method. This result suggests that the QCEIMS method needs further optimization in predicting the correct assignment of cation stability and assignment of the molecule with the lowest ionization energy in the fragmentation process (Stevenson's rule [28]).

1,8-nonadiene (Figure 3e): For this molecule, a great disagreement between the cosine similarity score of 713 and the weighted dot-product of 163 was observed. The weighted dot-product emphasizes high m/z ions that are penalized if missing in spectral matching. Again, QCEIMS overestimated the abundance of the molecular ion $[M]^{\cdot+}$ and of several atomic or molecular hydrogens from it. In addition, QCEIMS underestimated a neutral methyl loss (to m/z 109) and a neutral loss of ethylene (to m/z 96). To capture all potential fragmentations in QCEIMS such as the missed ethylene loss, more accurate PES estimates are needed.

Adamantane (Figure 3f): Adamantane is a well-known inflexible molecule. Our QCIEMS simulations correctly predicted the structure of the m/z 79 product ion as protonated benzene, proved by an independent publication of an infrared multiphoton dissociation spectrum [29] and DFT computations [30]. In comparison, the rule-based MassFrontier generated less reasonable fragment molecules that included cyclopropyl-moieties. The QCEIMS results showed that the m/z 93 product ion is likely associated with both *ortho*- and *para*-protonated toluene, in accordance with infrared multiphoton dissociation spectrum results [29]. These instances highlight the ability of QCEIMS to predict non-obvious mechanisms, such as rearrangements from sp^3 hybrid carbons to aromatic system.

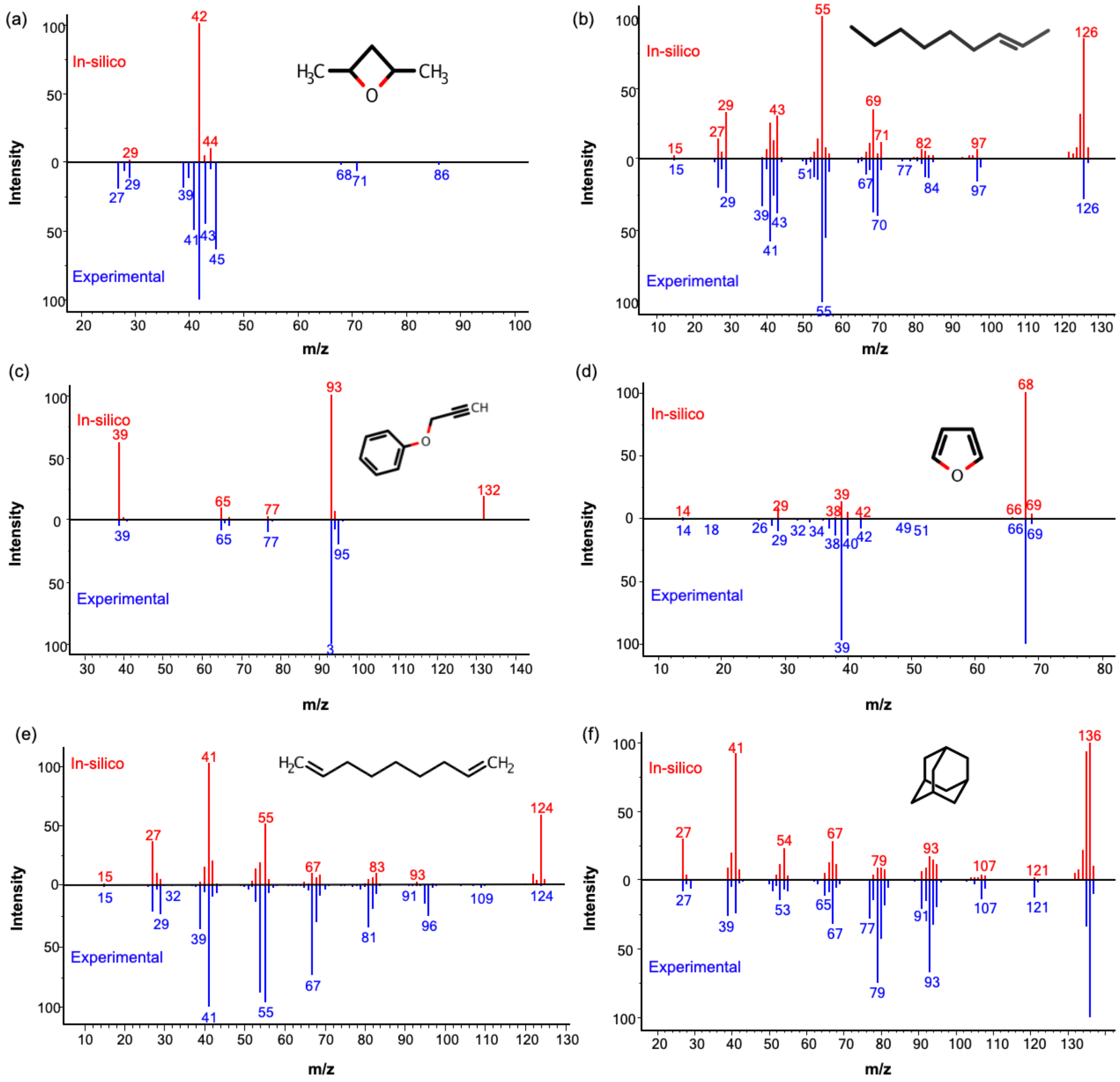


Figure 2-3. Examples for comparing experimental 70 eV EI mass spectra (lower panels) to QCEIMS in-silico mass spectra (upper panels) for six small molecules.

2.4.2. Probing the QCEIMS parameter space

A number of parameters can be chosen in the QCEIMS software, including the number of trajectories, impact excess energy per atom and initial temperatures. Other parameters such as the type of energy distribution and maximum MD time were excluded because they were already optimized during the development of QCEIMS [8]. We used OM2 because other semiempirical methods had been shown previously to perform worse [8]. For each molecule we chose one conformer and performed QCEIMS simulations with different parameter settings. By repeating QCEIMS simulations 50 times, we confirmed that identical mass spectra were obtained when using the same conformer under the same parameter settings. We changed parameter settings for 2,4-dimethyl-oxetane, 2-nonene and adamantane.

(1) Number of trajectories (ntraj)

In molecular dynamics, different reaction trajectories must be explored to cover possible routes of independent fragmentations across the energy surface. Each trajectory requires computational time, and therefore, the number of trajectories should be as low as possible. However, it is not clear a priori how many trajectories sufficiently cover the chemical reaction space and allow convergence to a consensus spectrum. By default, the QCEIMS program automatically calculates the number of trajectories by multiplying the number of atoms by 25. We explored this default value ranging from 8 to 1000 trajectories per atom for the different molecules, yielding up to 15,000 trajectories in total (Figure 2-4a). For each of the three molecules, the difference between the best and the worst similarity score differed only by 10% or less. None of the three molecules had improved similarity scores with higher number of trajectories. Indeed, it appeared that increasing the number of trajectories might lead to slightly lower dot-product similarity scores as observed for 2-nonene and adamantane, possibly due to a higher contribution of rare fragmentation reactions that lead to low abundant fragment ions that negatively impact similarity to experimental spectra. We concluded that the default value of 25 trajectories per atom number in a molecule was reasonable.

(2) Impact excess energy per atom (ieeatm)

Next, we tested the impact excess energy (IEE) that is introduced by the colliding electron in electron ionization as vibrational energy into the molecules. The default value (ieeatm) in QCEIMS software is set at 0.6 eV per atom on the basis of previous OM2 tests [31]. At the beginning of each molecular dynamics simulation the molecule is heated by increasing the atom velocities until the impact excess energy is converted to kinetic energy that leads to bond fragmentation. In other words, the collision energy is used to vibrationally excite and break the molecule. Higher impact excess energy will lead to a higher kinetic energy, causing the molecule to fragment more easily and to decrease the intensity of molecular ions. We observed that QCEIMS-simulated mass spectra contained fewer fragment ions than their experimental references. For example, the experimental spectrum of 2,4-dimethyl-oxetane (Figure 3a) has 23 product ions, while our QCEIMS simulation produced only four fragment ions plus the molecular ion peak m/z 86. We probed different internal excess energies from 0.2 to 0.8 eV (Figure 4b). With increasing IEE, more fragmentation occurs, increasing the intensity of low mass fragments, but we did not see an increase in the total number of fragments produced. Because the weighted dot-product score gives more weight to the more selective masses found at high m/z ranges, we found that higher IEE values led to decreasing similarity scores. In short, changing ieeatm did not provide a route to improve QCEIMS spectra and we kept the default value of 0.6 eV for subsequent tests.

(3) Initial temperature (tinit)

Last, we investigated the effect of temperature settings ranging from initial temperatures (tinit) of 300 to 1000 K, while keeping all other parameters at default values (Figure 4c). For 2-nonene and adamantane we found that the initial temperatures led to decreasing similarity scores, consistent with the concept that molecules under higher temperature will have more kinetic energy and tend to fragment more easily. For QCEIMS simulations, 2,4-dimethyl-oxetane generated the molecular ion m/z 86 only at low tinit of 300 K, leading to an artificially higher similarity score. As the other two tested molecules also showed their best spectrum similarities at tinit 300K, we chose this parameter value for a final test that utilized a combination of each best setting of ieeatm, ntraj and tinit for each molecule. Interestingly, these simulations did not lead to significant improvements or even to overall decreased similarity scores (see Supporting Information). Therefore, we kept the overall default parameter values for subsequent studies.

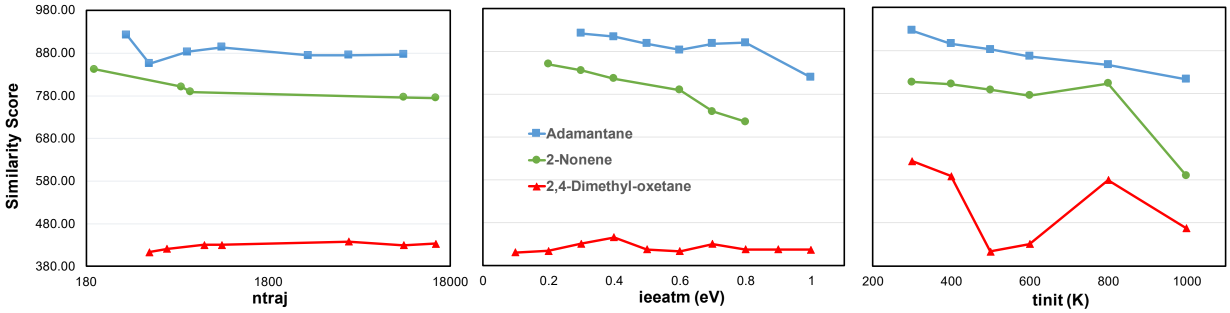


Figure 2-4. Impact of QCEIMS parameter settings on MS similarity scores comparing in-silico spectra to experimental spectra. Left panel: altering the number of trajectories (ntraj). Mid panel: altering the external energy per atom (ieeatm). Right panel: altering the initial temperature (tinit).

2.4.3. Different starting conformers as input for QCEIMS

Local minima on the potential energy surface that are related by rotations around single bonds are called conformational isomers, or conformers. In a mass spectrometer, the conformations of a large cohort of individual chemical molecules are distributed in accord with a Boltzmann distribution at a given ion source temperature. All conformers contribute to the final mass spectrum, to varying degrees related to their relative energies. Ideally, QCEIMS should cover the overall ensemble of conformers. To investigate the impact of the input conformers on the overall QCEIMS results, we selected the highly flexible 2-nonene (PHI=7.51, RBN=5) and the non-flexible adamantane (PHI=1.17, RBN=0) structures. We employed the GMMX software with the Merck Molecular Force Field (MMFF94) to generate starting conformers for individual QCEIMS simulations. For 99 simulations with different starting conformers of 2-nonene, the maximum difference between the lowest-energy and the highest-energy conformer was 2.83 kcal/mol (Figure 2-5a). For these conformers, dot-product similarity scores ranged from 719 to 824, with a median of 781 and a standard deviation of 24 (Figure 5b). Due to the rigid skeleton and inflexibility of adamantane, GMMX provided only one conformer. Therefore, we used the open source molecular dynamics package CP2K [32] to generate 50 adamantane structures with twisted or stretched bonds that yielded an overall energy range of 5.39 kcal/mol (Figure 5c). Dot-product similarity scores ranged from 849 to 948, with a median similarity of 923 and a standard deviation of 31 (Figure 5d). The examples of these very different molecules showed that QCEIMS similarity scores were independent from input conformer energies (Figure 5a, 5c). Yet, these examples also showed that for both molecules, the QCEIMS fragmentation of specific conformers can lead to quite different dot-product similarities compared to experimental mass spectra, ranging over 100 similarity score units. In addition, we found that dot-product similarities were not normally distributed (Figure 5b, 5d). Our results showed that conformational and other small structural changes may affect QCEIMS simulations. Although adamantane has only a single conformational energy minimum, even slight bond stretches or twists led to quite different mass spectral similarity scores, presumably by biasing molecular dynamics trajectories toward different regions of the potential energy surface. While the QCEIMS software automatically chooses energy-optimized conformers, we propose that a range of different conformers must be calculated to get a good estimate of average mass spectra across the conformational space.

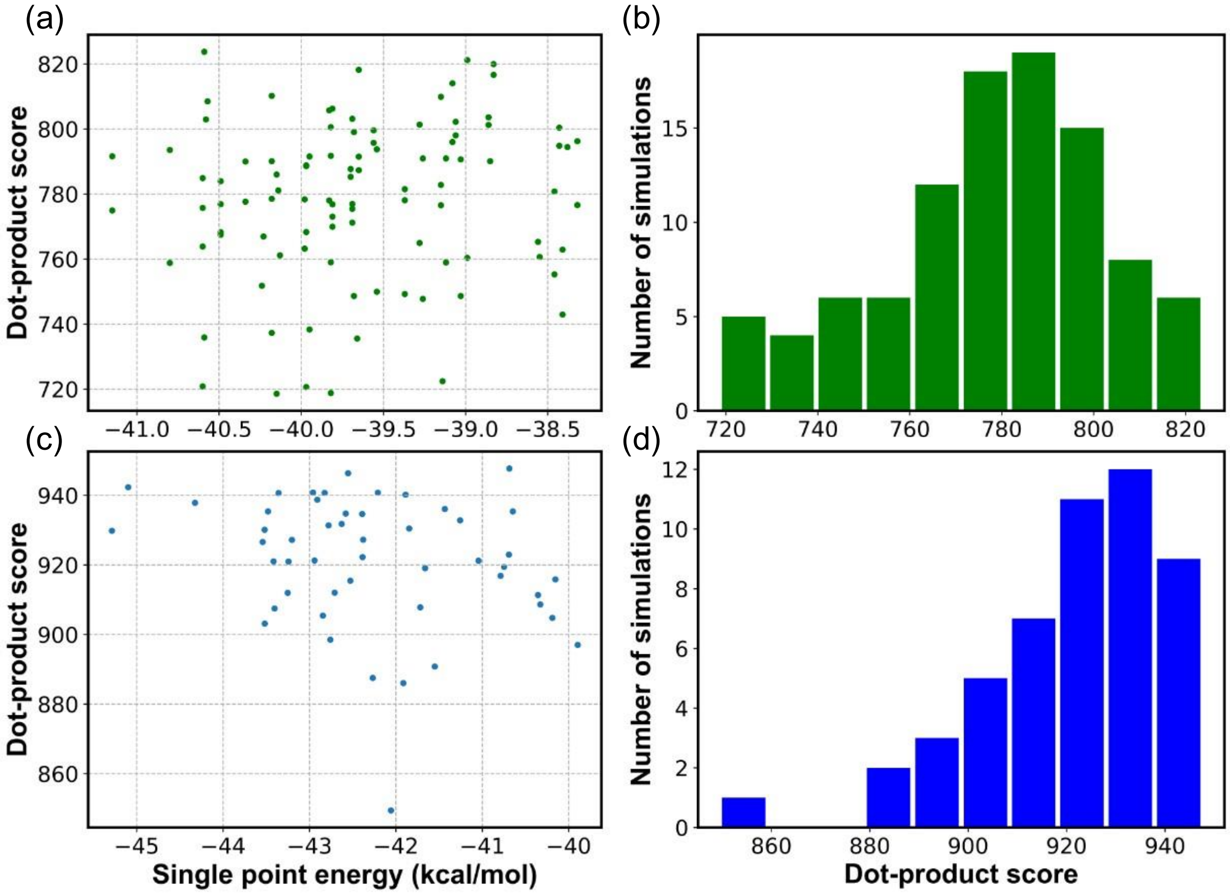


Figure 2-5. Impact of using different starting conformational isomers on MS similarity scores comparing in-silico spectra to experimental spectra. Each conformer has a specific single-point electronic energy.

Upper panels: 2-nonene conformers yielding dot-product MS similarity scores with histogram of the simulation results. Lower panels: adamantane conformers yielding dot-product MS similarity scores with histogram of the simulation results.

2.4.4. Large scale QCEIMS prediction of small molecule fragmentations

In order to be useful for experimental mass spectrometry, in-silico predictions must not only correctly explain fragmentation and rearrangement reactions for specific molecules but must also be scalable to generate spectra for hundreds, if not thousands of molecules. Here, we demonstrate the scalability of QCEIMS predictions for small molecules to systematically evaluate parameters and overall accuracies.

The OM2 method only supports carbon, hydrogen, nitrogen, oxygen and fluorine. We therefore chose 451 low molecular weight compounds containing only carbon, hydrogen, nitrogen and oxygen (CHNO). Molecular masses ranged from 26 to 368 Da with an average mass of 129 Da (see Supporting Information). For OM2, computational effort scales as $O(N^2) \sim O(N^3)$ [33], with N as number of atoms per molecule [33]. The number of single point energy calculations can be estimated to be linearly related to the number of trajectories, and thus linear to the number of atoms. On our computer system with 66 CPU threads, we achieved an average calculation time of 1.55 h per molecule (Figure 2-6a). Yet, as expected, calculation times exponentially increased with the number of atoms per molecule. For example, with more than 50 atoms, calculation times exceeded 14 hours on the system we had employed (Figure 6a).

Overall, the QCEIMS calculations across all 451 molecules yielded moderately accurate weighted-dot product similarity scores with an average of 608 (Figure 6b). Similarity scores below 500 are usually not considered for annotation of compounds. While similarity scores above 700 may represent true matches, only scores above 850 are regularly used for direct compound identifications in GC-MS experiments [24]. 47% of all molecules showed good dot-product match factors >700 and 20% of the molecules had excellent scores at >850 similarity. In comparison, lower cosine similarity scores were achieved with an average mass spectral similarity of 557 and a much higher proportion of unacceptably low scoring spectra at similarities <500 (Figure 6b). The regular cosine similarity score does not use weight functions for specific m/z values, unlike the weighted dot product score introduced in 1994 [22] that gives more weight to more specific high m/z product ions in MS fragmentation compared to less specific low m/z fragmentations based on large GC-MS library evaluations. Here, we see a similar trend for QCEIMS spectra. At this time, the Dot product score is recommended to use. The spectra with similarities < 500 is not recommended for matching, but the base peak or molecular ion peak can still provide useful structural information. In order to better apply the QCEIMS generated spectra in spectral matching purpose, a manual validation is required for each hit, regarding to the relative intensity and peak number. Additionally, a higher prediction accuracy is needed and the way to improve the accuracy has been discussed here. The related attempt is in progress in our lab.

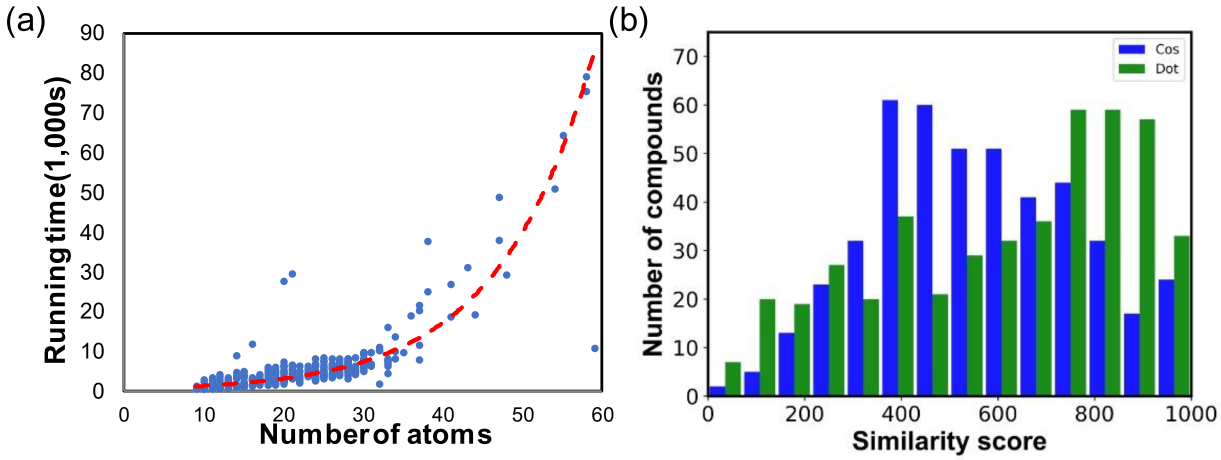


Figure 2-6. (a) Processing time of QCEIMS simulations of all 451 molecules versus the number of atoms per molecule. Red trend line: fitted exponential functions. (b) Histogram of weighted dot-product MS similarity scores against experimental spectra for all 451 molecules versus simple cosine similarity matches.

2.4.5. Molecular descriptors and prediction accuracy

Next, we tested the impact of the chemical structures themselves. We used ClassyFire [34] to classify all 451 chemicals into superclasses. We found QCEIMS predictions were significantly worse when comparing the organic oxygen superclass of 75 compounds against other superclasses with more than 50 members. Organic oxygen compounds had an average weighted dot-product of 520 whereas the 128 organoheterocyclic compounds achieved significantly better similarities of 648 at $p<0.0015$. The 100 organic nitrogen compounds yielded an average dot-product similarity of 657 at $p<0.001$ and the 62 hydrocarbons gave an average of dot-product similarity of 692 at $p<0.0001$. In conclusion, the QCEIMS method appears to perform worse for oxygen-containing organic compounds than for other major classes. For superclasses with fewer than 50 compounds, statistical tests were deemed to be not robust enough to allow such conclusions.

We also tested if rigid molecules resulted in better prediction accuracy than more flexible ones. Our hypothesis was based on an initial observation that for planar aromatic compounds such as pyridine or aniline, QCEIMS created better quality spectra than for molecules with long chain flexible structures. Our compound data set contained 295 molecules with low flexibility at Kier flexibility index (PHI) < 5 and 161 molecules with high flexibility of PHI > 6. Dot product scores varied significantly across both high-flexibility and low-flexibility molecules (Figure 2-7a). We found no relationship between flexibility and prediction accuracy. Similarly, we tested rotatable bond number (RBN) as a potential cause for prediction errors (Figure 7b). The median scores for molecules with different RBN values varied between 200-800 and did not depend on increasing RBN. This finding suggests that prediction accuracy is independent of the number of rotatable bonds. In conclusion, we could not find a correlation between flexibility and prediction accuracy at the level of simulation employed.

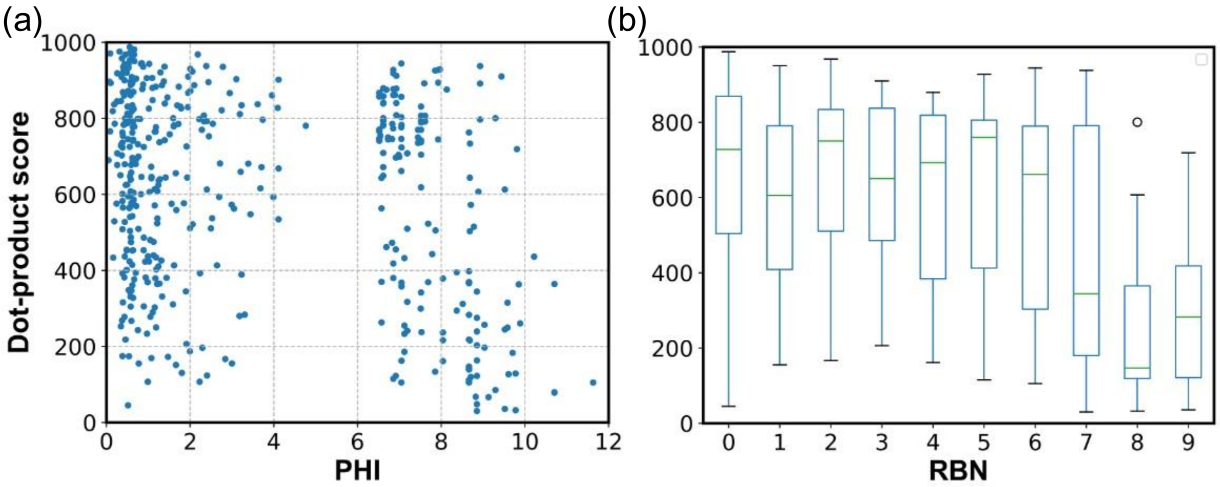


Figure 2-7. Impact of molecular flexibilities on MS similarity scores comparing in-silico spectra to experimental spectra. Influence of molecular flexibility. (a) scatter point plot of dot-product scores versus Kier flexibility index PHI; (b) boxplot of dot-product scores versus rotational bond number RBN

2.5. Conclusions

We here show that quantum chemistry calculations can be effectively used to correctly predict electron ionization fragmentation mass spectra as used in GC/MS analyses worldwide. Using QCEIMS software, mechanisms of fragmentation confirmed classic fragmentation rules. However, we found large differences in accuracy of predictions for different molecules. Changing parameters in QCEIMS was not a viable method to improve simulation results. Likely, capturing the potential energy surface accurately or even conducting the excited-state molecular dynamics [35-36] can be the key to further improving EI-MS prediction. For the first time, QCEIMS simulation was tested on hundreds of small molecules with limited computational resources within one month. We found that the superclass of organooxygen compounds performed much worse than organoheterocyclic compounds, hydrocarbons or organic nitrogen compounds. This observation may lead to future improvements in QCEIMS software as well as further inclusion of other heteroatoms in QCEIMS simulations.

2.6. References

1. Lai, Z.; Tsugawa, H.; Wohlgemuth, G.; Mehta, S.; Mueller, M.; Zheng, Y.; Ogiwara, A.; Meissen, J.; Showalter, M.; Takeuchi, K.; Kind, T.; Beal, P.; Arita, M.; Fiehn, O., Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics. *Nat Methods* **2018,** *15* (1), 53-56.

2. Stein, S., Mass Spectral Reference Libraries: An Ever-Expanding Resource for Chemical Identification. *Analytical Chemistry* **2012,** *84* (17), 7274-7282.

3. Vinaixa, M.; Schymanski, E. L.; Neumann, S.; Navarro, M.; Salek, R. M.; Yanes, O., Mass spectral databases for LC/MS- and GC/MS-based metabolomics: State of the field and future prospects. *TrAC Trends in Analytical Chemistry* **2016,** *78*, 23-35.

4. Blaženović, I.; Kind, T.; Ji, J.; Fiehn, O., Software Tools and Approaches for Compound Identification of LC-MS/MS Data in Metabolomics. *Metabolites* **2018,** *8* (2).

5. Wei, J. N.; Belanger, D.; Adams, R. P.; Sculley, D., Rapid Prediction of Electron–Ionization Mass Spectrometry Using Neural Networks. *ACS Central Science* **2019,** *5* (4), 700-708.

6. Allen, F.; Pon, A.; Greiner, R.; Wishart, D., Computational Prediction of Electron Ionization Mass Spectra to Assist in GC/MS Compound Identification. *Anal Chem* **2016,** *88* (15), 7689-97.

7. *Mass Frontier* 7.0 HighChem, Ltd.: 2011.

8. Grimme, S., Towards First Principles Calculation of Electron Impact Mass Spectra of Molecules. *Angewandte Chemie International Edition* **2013,** *52* (24), 6306-6312.

9. Bauer, C. A.; Grimme, S., How to Compute Electron Ionization Mass Spectra from First Principles. *J Phys Chem A* **2016,** *120* (21), 3755-3766.

10. (http://www.chemaxon.com), C. *Marvin was used for drawing, displaying and characterizing chemical structures, substructures and reactions*, Marvin 19.9.0; ChemAxon 2019.

11. Halgren, T. A., Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry* **1996,** *17* (5-6), 490-519.

12. Hanwell, M. D.; Curtis, D. E.; Lonie, D. C.; Vandermeersch, T.; Zurek, E.; Hutchison, G. R., Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *Journal of Cheminformatics* **2012,** *4* (1), 17.

13. O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R., Open Babel: An open chemical toolbox. *Journal of Cheminformatics* **2011,** *3* (1), 33.

14. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16 Rev. C.01*, Wallingford, CT, 2016.

15. *RDKit: Open-source cheminformatics;* *http://www.rdkit.org*, 2019.03.1; 2019.

16. Weber, W.; Thiel, W., Orthogonalization corrections for semiempirical methods. *Theoretical Chemistry Accounts* **2000,** *103* (6), 495-506.

17. Dewar, M. J. S.; Thiel, W., Ground states of molecules. 38. The MNDO method. Approximations and parameters. *Journal of the American Chemical Society* **1977,** *99* (15), 4899-4907.

18. Neese, F., The ORCA program system. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2012,** *2* (1), 73-78.

19. Perdew, J. P.; Ernzerhof, M.; Burke, K., Rationale for mixing exact exchange with density functional approximations. *The Journal of Chemical Physics* **1996,** *105* (22), 9982-9985.

20. Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H., A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *The Journal of Chemical Physics* **2010,** *132* (15), 154104.

21. Schäfer, A.; Horn, H.; Ahlrichs, R., Fully optimized contracted Gaussian basis sets for atoms Li to Kr. *The Journal of Chemical Physics* **1992,** *97* (4), 2571-2577.

22. Stein, S. E.; Scott, D. R., Optimization and testing of mass spectral library search algorithms for compound identification. *J Am Soc Mass Spectr* **1994,** *5* (9), 859-866.

23. Bauer, C. A.; Grimme, S., Elucidation of Electron Ionization Induced Fragmentations of Adenine by Semiempirical and Density Functional Molecular Dynamics. *The Journal of Physical Chemistry A* **2014,** *118* (49), 11479-11484.

24. Kind, T.; Tsugawa, H.; Cajka, T.; Ma, Y.; Lai, Z.; Mehta, S. S.; Wohlgemuth, G.; Barupal, D. K.; Showalter, M. R.; Arita, M.; Fiehn, O., Identification of small molecules using accurate mass MS/MS search. *Mass Spectrometry Reviews* **2018,** *37* (4), 513-532.

25. Veber, D. F.; Johnson, S. R.; Cheng, H.-Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D., Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *Journal of Medicinal Chemistry* **2002,** *45* (12), 2615-2623.

26. Kier, L. B., An Index of Molecular Flexibility from Kappa Shape Attributes. *Quantitative Structure-Activity Relationships* **1989,** *8* (3), 221-224.

27. Srl, A. *Alvascience Srl, alvaDesc (software for molecular descriptors calculation, https://www.alvascience.com)*, 1.0.8; 2019.

28. Stevenson, D. P., Ionization and dissociation by electronic impact. The ionization potentials and energies of formation of sec.-propyl and tert.-butyl radicals. Some limitations on the method. *Discussions of the Faraday Society* **1951,** *10* (0), 35-45.

29. Bouwman, J.; Horst, S.; Oomens, J., Spectroscopic Characterization of the Product Ions Formed by Electron Ionization of Adamantane. *Chemphyschem : a European journal of chemical physics and physical chemistry* **2018,** *19* (23), 3211-3218.

30. Candian, A.; Bouwman, J.; Hemberger, P.; Bodi, A.; Tielens, A. G. G. M., Dissociative ionisation of adamantane: a combined theoretical and experimental study. *Phys Chem Chem Phys* **2018,** *20* (8), 5399-5406.

31. Grimme, S., Towards First Principles Calculation of Electron Impact Mass Spectra of Molecules. *Angew Chem Int Edit* **2013,** *52* (24), 6306-6312.

32. Hutter, J.; Iannuzzi, M.; Schiffmann, F.; VandeVondele, J., cp2k: atomistic simulations of condensed matter systems. *WIREs Computational Molecular Science* **2014,** *4* (1), 15-25.

33. Thiel, W., Semiempirical quantum–chemical methods. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2014,** *4* (2), 145-157.

34. Djoumbou Feunang, Y.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; Greiner, R.; Wishart, D. S., ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *Journal of Cheminformatics* **2016,** *8* (1), 61.

35. Desouter-Lecomte, M.; Sannen, C.; Lorquet, J. C., A transition state theory of nonadiabatic unimolecular reactions controlled by a conical intersection. Application to the C2H+4 ion. *The Journal of Chemical Physics* **1983,** *79* (2), 894-904.

36. Nelson, T. R.; White, A. J.; Bjorgaard, J. A.; Sifain, A. E.; Zhang, Y.; Nebgen, B.; Fernandez-Alberti, S.; Mozyrsky, D.; Roitberg, A. E.; Tretiak, S., Non-adiabatic Excited-State Molecular Dynamics: Theory and

Applications for Modeling Photophysics in Extended Molecular Materials. *Chemical Reviews* **2020,** *120* (4), 2215-2287.

Chapter 3: Quantum chemical prediction of electron ionization mass spectra of trimethylsilylated metabolites

3.1. Abstract

Chemical derivatization, especially silylation, is widely used in gas chromatography coupled to mass spectrometry (GC-MS). By introducing the trimethylsilyl (TMS) group to substitute active hydrogens in the molecule, thermostable volatile compounds are created that can be easily analyzed. While large GC-MS libraries are available, the number of spectra for TMS-derivatized compounds is comparatively small. In addition, many metabolites cannot be purchased to produce authentic library spectra. Therefore, computationally generated in silico mass spectral databases need to take TMS derivatizations into account for metabolomics. The quantum chemistry method QCEIMS is an automatic method to generate electron ionization (EI) mass spectra directly from compound structures. To evaluate the performance of the QCEIMS method for TMS-derivatized compounds, we chose 816 trimethylsilyl derivatives of organic acids, alcohols, amides, amines and thiols to compare in-silico generated spectra against the experimental EI mass spectra from the NIST17 library. Overall, in-silico spectra showed a weighted dot-score similarity (1000 is maximum) of 635 compared to the NIST17 experimental spectra. Aromatic compounds yielded a better prediction accuracy with an average similarity score of 808, while oxygen-containing molecules showed lower accuracy with only an average score of 609. Such similarity scores are useful for annotation of small molecules in untargeted GC-MS based metabolomics, suggesting that QCEIMS methods can be extended to compounds that are not present in experimental databases. Despite this overall success, 37% of all experimentally observed ions were not found

in QCEIMS predictions. We investigated QCEIMS trajectories in detail and found missed fragmentations in specific rearrangement reactions. Such findings open the way forward for future improvements to the QCEIMS software.

3.2. Introduction

Gas chromatography coupled to mass spectrometry (GC-MS) requires volatile compounds for analysis. The generation of volatile derivatives from polar or thermo-labile compounds using silylation derivatization reactions is still the first choice for many modern applications.[1] The most common reagents for such applications are MSTFA (N-methyl-N-(trimethylsilyl) trifluoroacetamide), TMCS (trimethylchlorosilane), BSA (N,O-bis(trimethylsilyl)acetamide) and BSTFA (N,O-bis(trimethylsilyl) trifluoroacetamide).[2] Reactive functional groups that can be silylated with these reagents under mild conditions include alcohols, aldehydes, carboxylic acids, amines, amides, thiols, and inorganic acids. [2]

Silylation is used in many applications including medical investigations, metabolic profiling, toxicological screening, and environmental research.[3-4] All these approaches use mass spectral library matching for compound annotations and identifications. An experimental spectrum is compared against a reference spectrum in a database. The reference spectra were obtained from authentic reference compounds that underwent silylation reactions.

Licensed libraries such as Wiley or NIST20[5] contain around 5,000 TMS derivatives. Smaller TMS libraries for GC-MS[4, 6-7] are also freely available in MassBank of North America (MassBank.us), including retention indices that are used to improve automatic compound annotations. However, these libraries contain less than 3000 compounds combined, which is in stark contrast to the estimated 300,000 known natural products[8] and the more than 12 million commercially available compounds of more than 100 million known structures in PubChem. Furthermore, many silylation reactions are incomplete. While hydroxyls, thiols and carboxylic acid moieties are always completely derivatized, primary and secondary amines may be not be exhaustively derivatized.[9] Even with softer chemical ionization and accurate mass GC-MS, it is very difficult to elucidate the structures of unknown compounds.[10] To increase the size of available EI-MS libraries, mass spectra can be predicted *in silico* from molecular structures.[11] While machine learning models have been used to model TMS compound spectra, accuracy in EI-MS predictions was found to be lacking.[12-13] EI-MS spectra also can be predicted from first principles by quantum chemical modeling using QCEIMS[14-17] with the semi-empirical GFNn-xTB[18-20] method. Recent work showed that in-silico spectra generated by QCEIMS can help structure elucidation and identify unknowns.[21] However, these methods have not been tested so far on TMS-derivatized molecules. We here present data testing the performance of QCEIMS to generate theoretical mass spectra for a diverse set of compound classes using 816 TMS derivatized compounds.

3.3. Methods

3.3.1. Parallel fragmentation prediction

To test the general performance of QCEIMS for TMS-derivatized compounds (**Figure 3-1**), 816 molecules with TMS groups at <700 Da were selected from the NIST17 mass spectral database. While we used mono-TMS compounds to test the impact of TMS-derivatives on different compound classes, we also calculated nine doubly (2TMS) and nine triply (3TMS) silylated derivatives to demonstrate the extensibility of the QCEIMS method. Starting with the IUPAC International Chemical Identifier (InChI), we generated 3-D structures with the Merck Molecular Force Field (MMFF94[22]) and saved them in mol (*.mol) and TurboMole format (*.tmol) using

OpenBabel (v2.3.90).[23] We then used QCEIMS (v4.0) to generate *in silico* mass spectra, for which a new version QCxMS[24] including an EI-MS prediction module has recently been released at https://github.com/qcxms/QCxMS. Default settings for QCEIMS were used, with GFN1-xTB[18] used for force/energy calculations and IPEA parameters used for ionization potential (IP) calculations. The CYLview[25] program was used to visualize compound structures.

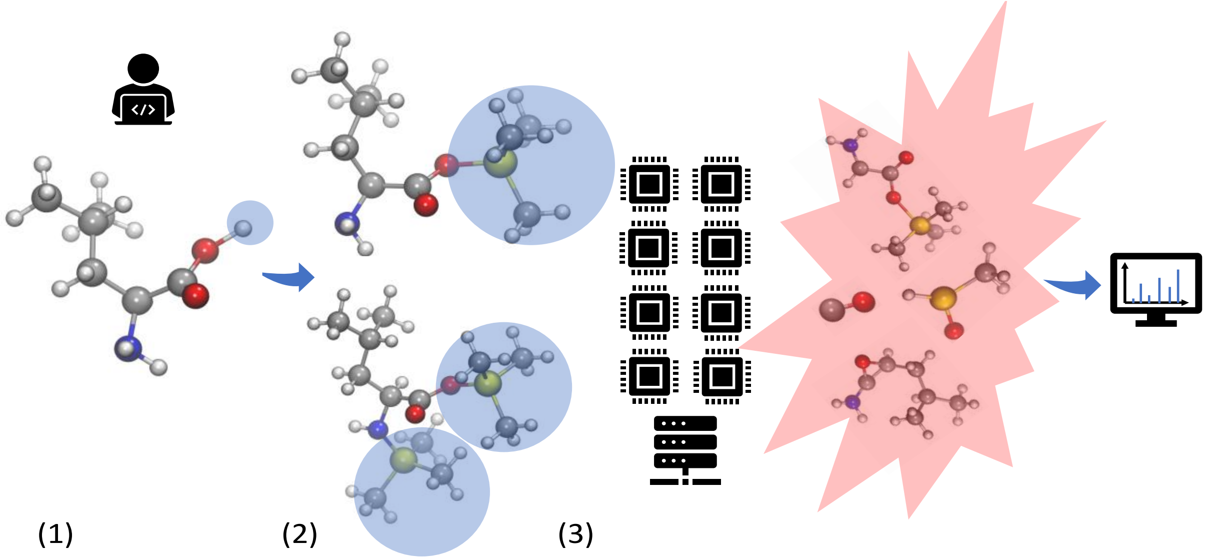


Figure 3-2. QCEIMS workflow of TMS derivatives: 1) substituting the active hydrogen of test molecules with trimethylsilyl groups; 2) generating 3D structures and initial conditions for QCEIMS; 3) parallel simulation to get fragments and *in silico* spectra

3.3.2. Substructure compound classification

Chemical compounds can be classified by substructure analysis into many different classes.[26] To evaluate the simulation accuracy on different compound classes, we here used the α-position of heteroatoms next to the silicon in TMS-groups to classify compounds. For example, if the α-heteroatom belonged to a carboxyl substructure, such compounds were annotated as acid, regardless which other functional groups were present in the molecule. A python script based on

RDKit[27] was used to classify molecules into five main compound classes: alcohols, acids, amines, amides and thiols. A detailed classification tree is presented in **Scheme S1**.

3.3.3. *In silico* accurate mass spectra

The QCEIMS program currently generates integer mass-to-charge ratio. One advantage of using quantum chemistry for MS simulations is that the type and frequency of molecule fragments are counted, while element and isotopic masses are computed. Therefore, we programmed an extension to the QCEIMS program that also incorporates accurate isotopic masses for elemental compositions (Supporting Information, Zenodo repository). Such accurate-mass in silico spectra are important when using high-resolution GC-MS instruments, which are increasingly used during structure elucidation of unknown compounds detected by GC-MS.[10, 28]

3.3.4. In-silico mass spectrum annotation

Experimental mass spectra in the NIST17 database were used as the true positive examples to evaluate the accuracy of *in silico* spectra generated by the QCEIMS process. Cosine similarity scores and modified dot-product scores were used for spectra comparison.[11]

$$Dot = \sqrt{\frac{(\Sigma W_I W_E)^2}{\Sigma W^2{}_I \, \Sigma W^2{}_E}} \qquad (1)$$

$$W = [Peak\ Intensity]^{0.5}[Mass]^3 \quad (2)$$

Where W is the mass-weighted peak intensity, the subscript *I* denotes the *in-silico* intensity and *E* denotes the experimental intensity.

MassFrontier 7.0[29] was utilized to help annotate *m/z* peak and neutral losses for all 70 eV mass spectra.

3.3.5. Accurate mass GC-MS analysis

Accurate mass spectra were acquired on an Agilent 7890A GC system with Agilent 7200 Accurate Mass Quadrupole Time-of-Flight(Q-TOF) mass spectrometer system (Agilent Technologies, Santa Clara, CA, U.S.A.). Chemicals were derivatized with 10 μL of methoxyamine hydrochloride in pyridine (20 mg/ml) to protect aldehyde- or ketone- groups, and then trimethylsilylated to increase volatility by 90 μL N-methyl-N-(trimethylsilyl)trifluoroacetamide (MSTFA). Previously published gas chromatographic conditions were used.[30] Mass spectra were obtained from *m/z* 50 to 800 at a 5 Hz scan rate in electron ionization mode with electron energy of 70 eV.

3.4. Results and Discussion

3.4.1. Trajectory analysis of in-silico predictions of fragmentations in electron ionization spectra

One advantage of first-principles simulation is that we can follow molecular dynamics (MD) trajectories during the fragmentation reactions. In this way, we can annotate observed *m/z* fragment ions with fragmentation substructures that provide insights into reaction mechanisms.[31-32] A selection of representative head-to-tail mass spectral comparisons including MD trajectories are given in the Supplemental Information (**Figure S1-S8**). Experimental mass spectra represent the likelihood and frequency of many stochastic fragmentation events. Therefore, many trajectories are combined into simulated spectra when using QCEIMS. We first exemplify this principle on few typical mass spectra from different compound classes. As example of an aliphatic acid, the head-to-tail comparison of the QCEIMS-predicted fragmentation of O-trimethylsilyl-leucine to the experimental NIST17 library spectrum (**Figure 3-2**) shows that many experimental observed ions were indeed correctly predicted by simulation. However, the ion intensities were often found to be different between predicted and experimental spectra, yielding a low dot-score MS similarity. For example, the $[M\text{-}15]^+$ fragment ion *m/z* 188 was predicted at 70% of base peak abundance, compared to the experimentally found 26% abundance. Similarly, the $[M\text{-}89]^+$ fragment ion at *m/z* 114 was predicted at 14% abundance compared to an experimental 2% abundance. Such disagreements in ion relative abundances heavily distort dot-score similarity calculations. We therefore set out to better understand the QCEIMS trajectories that led to ion formation. Relative abundances are determined by the prevalence of trajectories leading to specific fragments. QCEIMS spectra account for all charged fragments from all trajectories. We used 25 trajectories per atom for each molecule, guided by the idea that large molecules may have more options of fragmentations than smaller ones [14]. For example, the simulation of leucine-OTMS with 34 atoms accumulated a total of 850 trajectories. 27 trajectories resulted in the formation of the $[M\text{-}89]^+$ fragment ion (m/z 114 in Figure 2) with an average trajectory length of 900 fs and a median trajectory length of 857 fs. The QCEIMS method predicted two fragmentation pathways: (1) in 24 trajectories, a loss of $^{\cdot}CH_3$ was followed by a loss of $OSi(CH_3)_2$ (**Figure 2a, 2b**) and (2) in three trajectories, a loss of a $TMSO^{\cdot}$ radical was found (**Figure 2c**). For calculating the relative abundance of ions in QCEIMS spectra, the stability of ions is estimated by comparing the statistical charge of fragment to their ionization potentials, which is weighted by the Boltzmann distribution. Because of this weighting method, the same $[M\text{-}89]^+$ fragment ion in pathway (1) and (2) shows an extremely different statistical charge. The statistical charge for the 27 trajectories in pathway (1) is almost +1 while the three trajectories of pathway (2) have an average statistical charge of 0.04. In addition, we considered the impact excess energy (IEE), which denotes the residual energy introduced by the electron impact after ionizing the neutral molecule. For the 27 trajectories that generated the [M-89]$^+$ fragment ion, an average IEE of 25 eV was found. In contrary, for the 48 trajectories that stopped after loss of a methyl-group, an average IEE of 16 eV was found. This lower IEE thus led trajectories to remain at [M-15]$^+$ fragment ions without subsequent secondary fragmentations. We also found that [M-15]$^+$ fragmentations were exclusively associated with methyl-losses from the TMS-group, but not from the branched leucine carbon backbone. For the predicted [M-15]$^+$ fragment ion, an average trajectory length of 1625 fs and median trajectory of 1066 fs was found. In comparison, therefore, the [M-89]$^+$ fragment will be formed faster, but only under conditions that lead to higher impact excess energy.

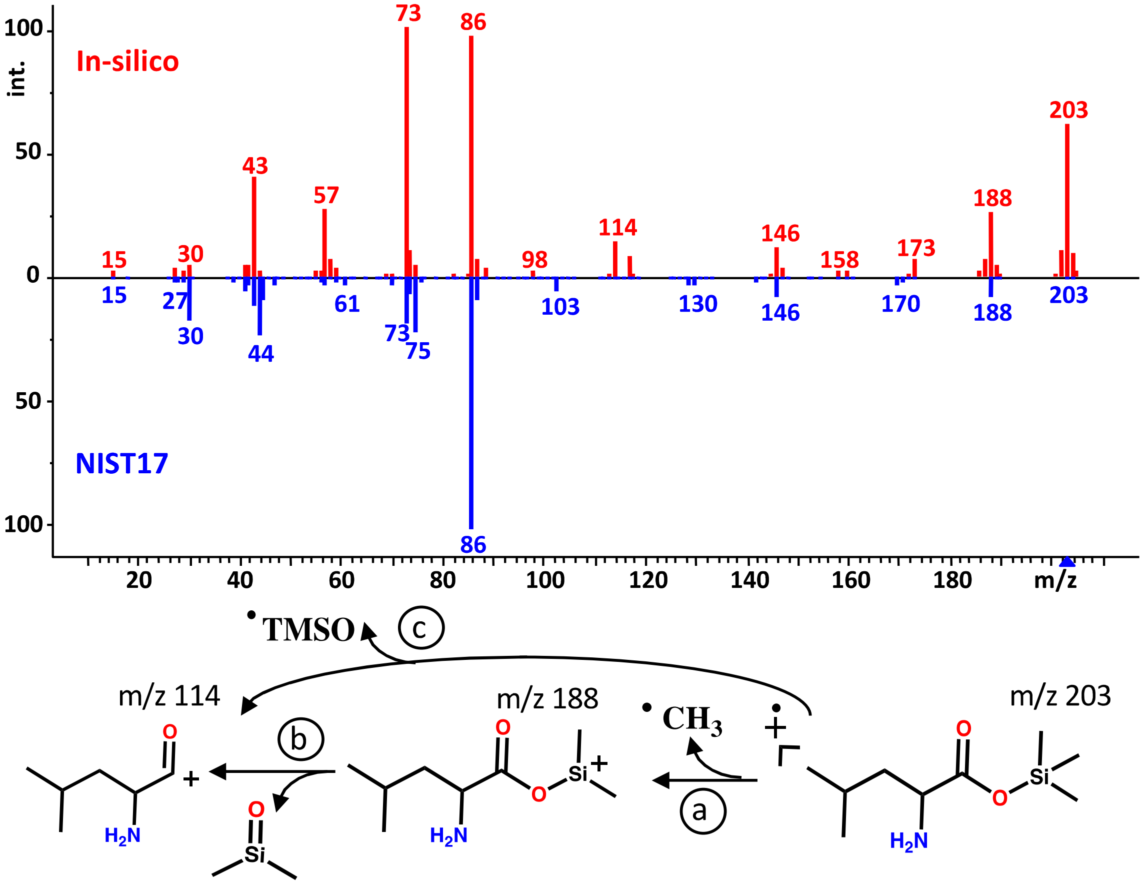


Figure 3-3. Fragmentation of leucine-OTMS modeled by QCEIMS compared to the experimental mass spectrum from the NIST17 mass spectral library. Lower panel: alternative fragmentation mechanisms (a-c) as detailed by QCEIMS trajectories. https://mona.fiehnlab.ucdavis.edu/spectra/display/MoNA040855

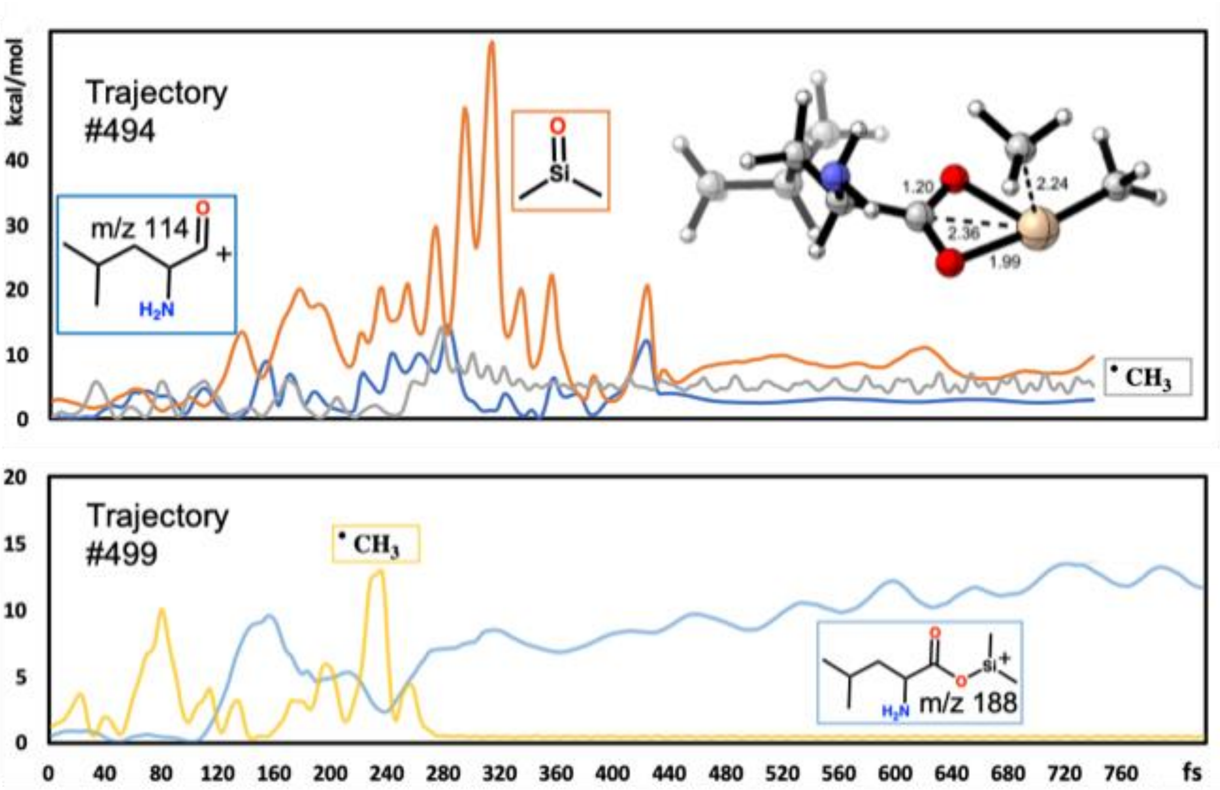


Figure 3-4. Temporal change of vibrational energy of substructures during the fragmentation of leucine-OTMS as modeled by QCEIMS for two individual trajectories. Upper panel: Trajectory #494 leading to substructure m/z 114 (blue), substructure dimethylsilanone (orange) and substructure methyl-group (grey). Inset: reaction intermediate observed at 380 fs. Lower panel: Trajectory #499 leading to substructure m/z 188 leucine-ODMS (light blue) and substructure methyl-group (yellow).

Previous papers have shown that statistical models purely based on IEE values are insufficient to predict experimental mass spectra.[33-34] Aside from the IEE, the distribution of energy within a molecule may also influence the likelihood of specific fragmentation pathways. We therefore analyzed fragmentation of leucine-OTMS from this perspective. To evaluate the effect of energy distributions, we performed an energy partition analysis[35] on trajectories #494 and #499, which yielded fragment ions *m/z* 114 ($[M-89]^+$ ion) and *m/z* 188 ($[M-15]^+$ ion) (**Figure 3**). Energy distribution plots of substructures often show oscillations throughout the trajectories, but the timing of fragmentations indeed coincide with the most drastic changes in energy distributions. For example, in the trajectory leading to the *m/z* 188 ion (**Figure 3**, lower panel), the methyl-substructure showed significant vibrational energy at 80 fs that led to bond stretching, but the actual fragmentation and generation of the methyl radical only appeared at around 240 fs. In comparison, for the *m/z* 114 ($[M-89]^+$ ion trajectory #494, the methyl radical departed at around 200 fs and, subsequently, a $OSi(CH_3)_2$ fragment departed at around 400 fs. At 380 fs, an intermediate structure was observed with a four-membered ring (**Figure 3**, upper panel insert). We also separated and validated the transition state structure for the methyl group loss in **Figure S12**. For both trajectories #494 and #499, methyl-substructures showed an increase in vibrational energy around 240-280 fs that led to fragmentation reactions. However, in trajectory #499 the vibrational energy was distributed within the leucine-ODMS substructure whereas in trajectory #494, the vibrational energy was rapidly distributed to the $OSi(CH_3)_2$ substructure. After a final energy redistribution to the dimethylsilanone substructure at around 420 fs, the loss of a neutral $OSi(CH_3)_2$ fragment occurred. In conclusion, analyses of QCEIMS trajectories, despite relying on the imperfect harmonic oscillator approximation, reveal how the distribution of vibrational energy can influence the directions of reactions and explain the prevalence of different reaction pathways.

3.4.2. Mass spectral fragmentation rules

Over decades of interpreting electron ionization mass spectra, characteristic product ions have been determined for specific molecular substructures,[36] including for trimethlysilylated compounds used in metabolomics.[28] We therefore investigated if our MD simulations correctly predicted such product ions. Here we provide detailed information on fragmentation for two molecules, with additional examples given for other compound classes in **Figure S9**. The aromatic acid trimethylsilyl-4-methoxybenzoate (**Figure 4**) was predicted to form the molecular radical ion in a higher abundance than experimentally observed. For aromatic acids and their derivatives, five product ions have been described as characteristic fragments.[28] Among these, the $[M-CH_3]^+$ and $[M-OTMS]^+$ neutral losses were accurately predicted with QCEIMS simulations (**Figure 4b**). The *m/z* 194 ion could be produced in two different ways, either as secondary methyl loss from a *m/z* 209 ion leading to a *m/z* 194.039 radical cation that was also found when we analyzed this molecule using accurate mass GC-QTOF MS (**Figure S9**), but not the alternative *m/z* 194.076 ion that would have resulted from a neutral loss of $O=CH_2$ from the 4-methoxy-group. Similarly, the *m/z* 135.045 ion was correctly predicted by QCEIMS to arise from a neutral loss of TMSO*, and not as an alternative product with *m/z* 135.024 ($C_8H_{11}Si$) that would have been formed by a literature-described four-membered ring rearrangement (**Figure 4b**).[37] These examples show that QCEIMS can produce mechanistic predictions that were experimentally verified by accurate mass GC-QTOF MS measurements.

However, the neutral loss of CO_2 from a *m/z* 209 species to form a *m/z* 165.073 fragment (**Figure 4c**) via a four-membered ring arrangement was not correctly predicted by QCEIMS. Using the rule-based MassFrontier software,[30] this ion likely originates from a rearrangement reaction in which the silicon is transferred to the benzene ring through a four-membered transition structure with CO_2 as leaving group.[37] Two arguments may explain this observation. First, high energy transition structure itself can only be accessed if a specific initial conformation is formed, similar to conformer-defined reactions simulated previously.[38] This example demonstrates that QCEIMS predictions could be improved by more comprehensive conformer sampling to correctly accommodate the probabilities alternative reaction pathways. Secondly, our simulation time was limited to a few picoseconds (10^{-12} s). Rearrangement reactions in mass spectrometry may reach a time scale of 10^{-11} ~ 10^{-6} s [39] which is too long to be simulated by molecular dynamics methods. QCEIMS predictions also correctly matched the experimental accurate mass *m/z* 107.050 for $[C_7H_7O]^+$ leading to distributed positive charge along the aromatic ring (**Figure 4d**). However, several trajectories were also detected that led to other energetically unstable structures through ring-opening reactions (**Figure 4e**). Such trajectories may contribute to incorrect predictions of relative ion intensities. We also found that the *m/z* 77 for the benzyl cation and *m/z* 92 for $C_6H_4O^{+\cdot}$ were underestimated by the simulation. These two fragments were generated by two continuous fragmentations, highlighting the importance of considering multiple step fragmentations and the length of simulation times.

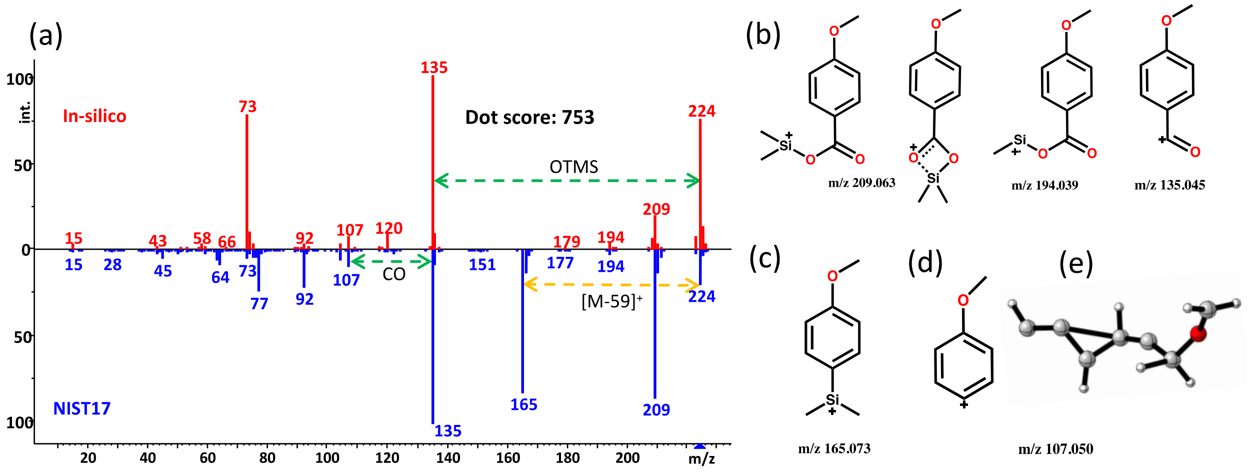


Figure 3-4. Fragmentation analysis of trimethylsilyl-4-methoxybenzoate as an example of aromatic carboxylates. (a) head-to-tail comparison of QCEIMS prediction against the experimental NIST17 library spectrum; (b) proposed structures predicted by QCEIMS simulation and validated by accurate mass GC-QTOF MS measurements; (c) proposed structure of experimentally found rearrangement product m/z 165; (d) proposed aromatic structure for fragment ion m/z 107 along with a high energy structure predicted by QCEIMS trajectories; (e) energetically unstable structures observed; In-silico spectrum available at https://mona.fiehnlab.ucdavis.edu/spectra/display/MoNA040747

In QCEIMS predictions for primary alcohols, many fragment ions correctly matched experimentally observed ions (**Figure 3-5a**): *m/z* 209 for $[M\text{-}15]^{+}$, *m/z* 103 for TMS-OCH_2^{+}, *m/z* 73 for TMS^{+} ions, and *m/z* 59 for $(CH_3)_2SiH^{+}$.[28] The characteristic *m/z* 73 TMS+ ions are generated by Si-O bond dissociations. Errors in predicting ion abundances are likely due to inaccurate estimations of the dissociation energies of oxygen-silicon bonds. QCEIMS predicted different trajectories that led to two distinct fragment structures for the *m/z* 194 peak (**Figure 5b**): $C_{11}H_{18}OSi^{+\cdot}$ (*m/z* 194.112) and $C_{10}H_{14}O_2Si^{+\cdot}$ (*m/z* 194.076) in an intensity ratio of 1:25. Nine trajectories showed a seven-membered ring rearrangement reaction en route to the *m/z* $C_{10}H_{14}O_2Si^{+\cdot}$ peak (**Figure 5c**). Both fragment ions were confirmed experimentally by high resolution GC-QTOF MS (**Figure S10**), albeit with a different relative intensity ratio of 2:9. Nevertheless, this observation shows that QCEIMS can correctly predict rearrangement reactions.

3.4.3. Average accuracy of QCEIMS predictions for different compound classes

To obtain an overview how accurate the QCEIMS approach is for predicting TMS-derivatized mass spectra for different classes of typical metabolites, we calculated spectra for a total of 816 molecules. All QCEIMS predicted spectra have been uploaded to MassBank.us. Molecules were selected by following the frequency distribution of chemical classes in the NIST database. A discussion of simulation time can be found in **Figure S11**. We summarized all structures into five major compound classes (**Table 1**) and subdivided these into aromatic and aliphatic structures by the location of the TMS-derivatized heteroatom (**Figure S1**). We had previously shown for QCEIMS predictions of underivatized molecules that mass-weighted dot score similarities were better suited than cosine scores for matching predicted to experimental spectra.[11] We found the same trend for TMS-derivatized compounds here and therefore only present the mass-weighted dot score match factors (**Table 1**). Detailed comparisons for cosine and dot score similarities are given for all 816 compounds in **Table S1**. Across all compound subclasses, dot score similarities ranged from 532 to 847 when compared to standard 70 eV spectra in the NIST17 database (**Table 1**). In addition, for 18 example molecules we showed that the QCEIMS approach can be extended to 2TMS- and 3TMS-derivatives (**Table S2**). The nine tested 2TMS-derivatives yielded an average dot-product score of 615, whereas the nine tested 3TMS-derivatives only gave an average dot-product score of 449. Short QCEIMS simulation times may become even more detrimental for predicting intramolecular rearrangements for molecules with multiple TMS groups, for example, for predicting fragments such as m/z 147 for TMS-diols[37].

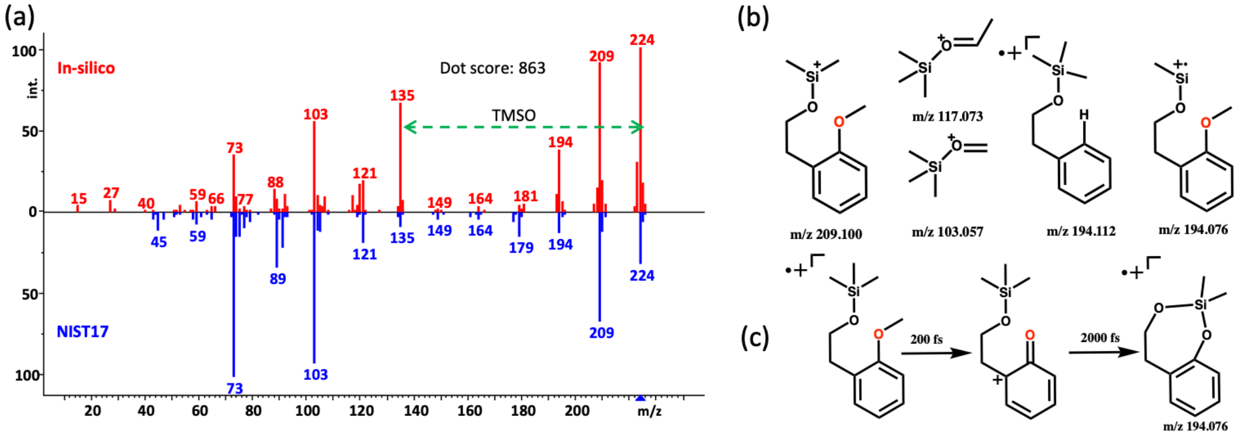


Figure 3-5. Fragmentation analysis of trimethylsilylated 2-methoxyphenylethanol as an example of primary alcohols; (a) head-to-tail comparison of QCEIMS prediction against the experimental NIST17 library spectrum; (b) examples of correctly QCEIMS predicted fragment ions; (c) seven-membered ring structure of m/z 194.076 https://mona.fiehnlab.ucdavis.edu/spectra/display/MoNA040546

Table 3-1. Matching 816 QCEIMS theoretical spectra against NIST17 experimental spectra using weighted dot-product and Jaccard similarity indices. Averages ± standard deviations are given.

Super-class	Subclass	Count	Dot score	Jaccard
acids	total	211	605 ± 183	0.51 ± 0.10
	aromatic	50	710 ± 123	0.49 ± 0.10
	aliphatic	161	572 ± 187	0.51 ± 0.10
alcohols	total	443	611 ± 224	0.53 ± 0.13
	aromatic	117	832 ± 79	0.52 ± 0.15
	aliphatic	326	532 ± 206	0.53 ± 0.13
amides	total	30	727 ± 152	0.56 ± 0.12

	aromatic	14	806 ± 34	0.59 ± 0.11
	aliphatic	16	658 ± 181	0.52 ± 0.13
amines	total	106	744 ± 186	0.58 ± 0.13
	aromatic	50	838 ± 95	0.56 ± 0.12
	aliphatic	56	661 ± 208	0.60 ± 0.13
thiols	total	26	743 ± 186	0.49 ± 0.11
	aromatic	15	847 ± 31	0.55 ± 0.04
	aliphatic	11	601 ± 217	0.41 ± 0.11

Two important differences were noted when comparing mass spectral similarity scores between experimental and QCEIMS predicted spectra across all compound classes. (1) Most aromatic compounds yielded a significantly higher similarity score than corresponding aliphatic compounds of the same class, with the exception of aromatic and aliphatic acids, which yielded comparable scores. (2) Average mass-weighted dot scores of oxygen-containing compounds (acids, alcohols) were significantly lower than other compound classes (amides, amines, thiols).

When inspecting head-to-tail comparisons of mass spectra (**Figures 3-2, 4, 5** and **S1-S8**), we found that spectra with low dot score similarities usually exhibited disagreements in the high *m/z* peak region, especially with respect to the presence and abundance of the molecular ion peak ($M^{+\cdot}$). The high *m/z* region is given especially large weight in the weighted dot-score calculation that is used in GC-MS analyses,[40] and hence, differences in $M^{+\cdot}$ abundances heavily contribute to lower scores. The radical ion produced for aromatic compounds can be stabilized through π-delocalization which leads to high ion intensities for both predicted and experimental spectra, and ultimately a high weighted dot-similarity score. When comparing the prediction errors across the different functional groups (superclasses), it was clearly noted that both alcohols and acids showed a large difference in $M^{+\cdot}$ abundances between predicted and experimental spectra. In comparison, intensities for $M^{+\cdot}$ molecular ions were more predictable for thiols and amides, and to some extent, also for amines. This finding confirms our previous results for non-silylated compounds that also had shown worse matching scores for oxygen-containing molecules compared to molecules without oxygen atoms.[11]

3.4.4. Relationship of MS-similarity score to QCEIMS spectral predictions

Overall dot-score similarities are heavily influenced by predicted ion intensities. However, the current accuracy of QCEIMS predictions can also be evaluated based on the number of ions that were correctly simulated by QCEIMS trajectories, in relation to ions that were predicted but not experimentally validated, and ions that were experimentally found but not predicted. This evaluation can be mathematically expressed by the Jaccard Index:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

where A are predicted ions, B are experimental ions, $|A \cap B|$ is the intersection of ions found in both predicted and experimental spectra, and $|A \cup B|$ is the complement of both predicted and experimental ions. Therefore, the Jaccard index ranges from 0 (if no ion is correctly predicted) to 1 (if all ions are correctly predicted). Because the generation of ions is an inherently stochastic process and as the QCEIMS model used here limited the number of tested trajectories to 25-times the number of atoms per molecule, we limited the calculation of Jaccard indices to ions that were found at more than 1% intensity of the base peak ions.

Overall, an average of 53% of all experimental ions were correctly predicted by the QCEIMS method for the 816 trimethylsilylated molecules examined (**Figure 3-6, Table 3-1**), showing that quantum chemistry for electron ionization spectra is both scalable for hundreds of molecules and can produce useful true positive rates. Interestingly, the Jaccard index shows that we have on average a higher proportion of fragment ions that were experimentally found but not QCEIMS predicted than incorrect predictions by QCEIMS that were not experimentally validated (**Figure 6**). This observation shows that a range of fragmentation reactions were not located using QCEIMS, for example, the rearrangement via a four-membered ring transition structure in **Figure 3-5** (missing ion *m/z* 165). Other reactions that heavily depend on conformational or electronic states are likely undersampled, for example, hydrogen migration reactions. When we investigated the degree of Jaccard index accuracy with respect to different substructures, no statistical difference was found (**Table 1**), unlike for overall dot-product similarities. Similarly, when we investigated the dependency of dot-score similarities of QCEIMS predicted spectra versus the Jaccard index errors, no significant impact was evident for the relative contribution of overpredicted ions or underpredicted ions.

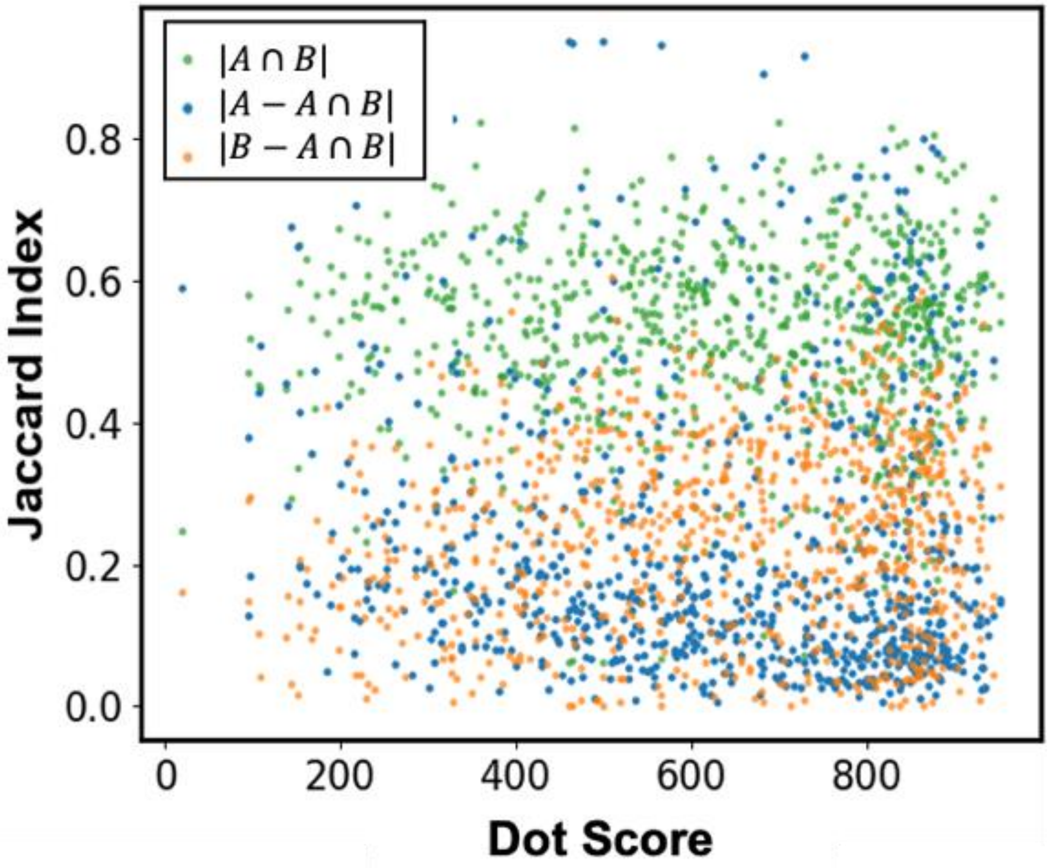


Figure 3-6. Comparison of 816 compound spectra for QCEIMS prediction versus experimental mass spectra. For each spectrum, the Jaccard similarity index was calculated giving three fractions: the intersection of correctly predicted ions (green dots), versus ions only found in experimental spectra (underpredicted, orange), or ions only found in QCEIMS predicted spectra (overpredicted, blue).

3.5. Conclusions

We presented the first large-scale application of the QCEIMS algorithm on trimethylsilylated compounds. We completed calculations for almost twice as many compounds than in a previous report on non-derivatized molecules.[11] Together, these two studies show that quantum chemistry prediction of mass spectra is now on the verge of being applicable to thousands of compounds, with the prospect of being useful for compounds that are not commercially available and not present in current MS libraries. On a single CPU thread, calculations took approximately 2.3 hours per atom or approximately 7.2 hours on a 16 CPU cluster for a molecule with 50 atoms. Calculation times increase quadratically if larger molecules are calculated. Assuming these calculations were run on 5000 nodes with molecules that do not exceed 50 atoms, we might be able to calculate spectra for 100,000 molecules within 100 days, as long as the size and complexity of molecules is similar as presented here.

To assess the accuracy of such predictions, we analyzed the fragmentation reactions for specific molecules and the MS/MS matching scores of QCEIMS predicted spectra across aliphatic and aromatic trimethylsilylated compounds. Overall, we found that QCEIMS predictions were most accurate for aromatic compounds with nitrogen-heteroatoms than for oxygen-containing aliphatic compounds. We also uncovered some challenges for this method. For example, internal vibrational energy redistribution appears to impact the selectivity between competitive reactions. While many complex rearrangements were correctly predicted, we found that some reactions with four-membered transition states were missed by QCEIMS trajectory analyses. When calculating the Jaccard Index of QCEIMS predicted spectra versus experimental reference spectra, we concluded that such missed reactions had more impact on poor MS-similarity scores than over-predicted fragment ions. Despite the necessary approximations used in the QCEIMS tool, overall matching scores showed that predicted spectra have high enough quality to be useful in mass spectrometry research, including identification of unknown compounds in untargeted screens. Future advancements in QCEIMS may explore additional conformer sampling and different atom velocities. In addition, we will test excited-state MD simulations to investigate if the inclusion of higher energy states may improve predictions in electron ionization mass spectrometry. [38, 41]

3.6. Reference

1. Fiehn, O., Extending the breadth of metabolite profiling by gas chromatography coupled to mass spectrometry. *Trends Analyt Chem* **2008,** *27* (3), 261-269.

2. Zaikin, V.; Halket, J. M., *A handbook of derivatives for mass spectrometry*. IM Publications: Chichester, 2009.

3. Kind, T.; Wohlgemuth, G.; Lee, D. Y.; Lu, Y.; Palazoglu, M.; Shahbaz, S.; Fiehn, O., FiehnLib: Mass Spectral and Retention Index Libraries for Metabolomics Based on Quadrupole and Time-of-Flight Gas Chromatography/Mass Spectrometry. *Analytical Chemistry* **2009,** *81* (24), 10038-10048.

4. Kopka, J.; Schauer, N.; Krueger, S.; Birkemeyer, C.; Usadel, B.; Bergmüller, E.; Dörmann, P.; Weckwerth, W.; Gibon, Y.; Stitt, M.; Willmitzer, L.; Fernie, A. R.; Steinhauser, D., GMD@CSB.DB: the Golm Metabolome Database. *Bioinformatics* **2004,** *21* (8), 1635-1638.

5. Stein, S., Mass Spectral Reference Libraries: An Ever-Expanding Resource for Chemical Identification. *Analytical Chemistry* **2012,** *84* (17), 7274-7282.

6. Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; Oda, Y.; Kakazu, Y.; Kusano, M.; Tohge, T.; Matsuda, F.; Sawada, Y.; Hirai, M. Y.; Nakanishi, H.; Ikeda, K.; Akimoto, N.; Maoka, T.; Takahashi, H.; Ara, T.; Sakurai, N.; Suzuki, H.; Shibata, D.; Neumann, S.; Iida, T.; Tanaka, K.; Funatsu, K.; Matsuura, F.; Soga, T.; Taguchi, R.; Saito, K.; Nishioka, T., MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry* **2010,** *45* (7), 703-714.

7. Kind, T.; Tsugawa, H.; Cajka, T.; Ma, Y.; Lai, Z.; Mehta, S. S.; Wohlgemuth, G.; Barupal, D. K.; Showalter, M. R.; Arita, M.; Fiehn, O., Identification of small molecules using accurate mass MS/MS search. *Mass Spectrometry Reviews* **2018,** *37* (4), 513-532.

8. Sorokina, M.; Steinbeck, C., Review on natural products databases: where to find data in 2020. *Journal of Cheminformatics* **2020,** *12* (1), 20.

9. Kumari, S.; Stevens, D.; Kind, T.; Denkert, C.; Fiehn, O., Applying in-silico retention index and mass spectra matching for identification of unknown metabolites in accurate mass GC-TOF mass spectrometry. *Anal Chem* **2011,** *83* (15), 5895-902.

10. Lai, Z.; Tsugawa, H.; Wohlgemuth, G.; Mehta, S.; Mueller, M.; Zheng, Y.; Ogiwara, A.; Meissen, J.; Showalter, M.; Takeuchi, K.; Kind, T.; Beal, P.; Arita, M.; Fiehn, O., Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics. *Nature Methods* **2018,** *15* (1), 53-56.

11. Wang, S.; Kind, T.; Tantillo, D. J.; Fiehn, O., Predicting in silico electron ionization mass spectra using quantum chemistry. *Journal of Cheminformatics* **2020,** *12* (1), 63.

12. Wei, J. N.; Belanger, D.; Adams, R. P.; Sculley, D., Rapid Prediction of Electron-Ionization Mass Spectrometry Using Neural Networks. *ACS Cent Sci* **2019,** *5* (4), 700-708.

13. Allen, F.; Pon, A.; Greiner, R.; Wishart, D., Computational prediction of electron ionization mass spectra to assist in GC/MS compound identification. *Analytical chemistry* **2016,** *88* (15), 7689-7697.

14. Grimme, S., Towards First Principles Calculation of Electron Impact Mass Spectra of Molecules. *Angewandte Chemie International Edition* **2013,** *52* (24), 6306-6312.

15. Bauer, C. A.; Grimme, S., First principles calculation of electron ionization mass spectra for selected organic drug molecules. *Organic & Biomolecular Chemistry* **2014,** *12* (43), 8737-8744.

16. Bauer, C. A.; Grimme, S., Elucidation of Electron Ionization Induced Fragmentations of Adenine by Semiempirical and Density Functional Molecular Dynamics. *The Journal of Physical Chemistry A* **2014,** *118* (49), 11479-11484.

17. Borges, R. M.; Colby, S. M.; Das, S.; Edison, A. S.; Fiehn, O.; Kind, T.; Lee, J.; Merrill, A. T.; Merz, K. M.; Metz, T. O.; Nunez, J. R.; Tantillo, D. J.; Wang, L.-P.; Wang, S.; Renslow, R. S., Quantum Chemistry Calculations for Metabolomics. *Chemical Reviews* **2021,** *121* (10), 5633-5670.

18. Grimme, S.; Bannwarth, C.; Shushkov, P., A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements (Z = 1–86). *Journal of Chemical Theory and Computation* **2017,** *13* (5), 1989-2009.

19. Koopman, J.; Grimme, S., Calculation of Electron Ionization Mass Spectra with Semiempirical GFNn-xTB Methods. *ACS Omega* **2019,** *4* (12), 15120-15133.

20. Bannwarth, C.; Ehlert, S.; Grimme, S., GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J Chem Theory Comput* **2019,** *15* (3), 1652-1671.

21. Schreckenbach, S. A.; Anderson, J. S. M.; Koopman, J.; Grimme, S.; Simpson, M. J.; Jobst, K. J., Predicting the Mass Spectra of Environmental Pollutants Using Computational Chemistry: A Case Study and Critical Evaluation. *J Am Soc Mass Spectr* **2021,** *32* (6), 1508-1518.

22. Halgren, T. A., Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry* **1996,** *17* (5-6), 490-519.

23. O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R., Open Babel: An open chemical toolbox. *Journal of Cheminformatics* **2011,** *3* (1), 33.

24. Koopman, J.; Grimme, S., From QCEIMS to QCxMS: A Tool to Routinely Calculate CID Mass Spectra Using Molecular Dynamics. *J Am Soc Mass Spectr* **2021,** *32* (7), 1735-1751.

25. Legault, C. Y. *CYLview*, 1.0b; Université de Sherbrooke: 2009.

26. Djoumbou Feunang, Y.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; Greiner, R.; Wishart, D. S., ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *Journal of Cheminformatics* **2016,** *8* (1), 61.

27. *RDKit: Open-source cheminformatics*, 2019.03.1; 2019.

28. Lai, Z.; Fiehn, O., Mass spectral fragmentation of trimethylsilylated small molecules. *Mass Spectrometry Reviews* **2018,** *37* (3), 245-257.

29. *Mass Frontier* 7.0 HighChem, Ltd.: 2011.

30. Lai, Z.; Kind, T.; Fiehn, O., Using Accurate Mass Gas Chromatography–Mass Spectrometry with the MINE Database for Epimetabolite Annotation. *Analytical Chemistry* **2017,** *89* (19), 10171-10180.

31. Beuck, S.; Schwabe, T.; Grimme, S.; Schlörer, N.; Kamber, M.; Schänzer, W.; Thevis, M., Unusual mass spectrometric dissociation pathway of protonated isoquinoline-3-carboxamides due to multiple reversible water adduct formation in the gas phase. *J Am Soc Mass Spectr* **2009,** *20* (11), 2034-2048.

32. Engeser, M.; Mundt, C.; Bauer, C.; Grimme, S., N-Methylimidazolidin-4-one organocatalysts: gas-phase fragmentations of radical cations by experiment and theory. *Journal of Mass Spectrometry* **2017,** *52* (7), 452-458.

33. Lorquet, J. C., Basic questions in mass spectrometry. *Organic Mass Spectrometry* **1981,** *16* (11), 469-482.

34. Lorquet, J. C., Landmarks in the theory of mass spectra. *International Journal of Mass Spectrometry* **2000,** *200* (1), 43-56.

35. Kurouchi, H.; Andujar-De Sanctis, I. L.; Singleton, D. A., Controlling Selectivity by Controlling Energy Partitioning in a Thermal Reaction in Solution. *Journal of the American Chemical Society* **2016,** *138* (44), 14534-14537.

36. Gasteiger, J.; Hanebeck, W.; Schulz, K. P., Prediction of mass spectra from structural information. *Journal of Chemical Information and Computer Sciences* **1992,** *32* (4), 264-271.

37. Harvey, D. J.; Vouros, P., MASS SPECTROMETRIC FRAGMENTATION OF TRIMETHYLSILYL AND RELATED ALKYLSILYL DERIVATIVES. *Mass Spectrometry Reviews* **2020,** *39* (1-2), 105-211.

38. Semialjac, M.; Schröder, D.; Schwarz, H., Car–Parrinello Molecular Dynamics Study of the Rearrangement of the Valeramide Radical Cation. *Chemistry – A European Journal* **2003,** *9* (18), 4396-4404.

39. Holmes, J. L., Assigning structures to ions in the gas phase. *Organic Mass Spectrometry* **1985,** *20* (3), 169-183.

40. Stein, S. E.; Scott, D. R., Optimization and testing of mass spectral library search algorithms for compound identification. *J Am Soc Mass Spectr* **1994,** *5* (9), 859-866.

41. Moss, C. L.; Liang, W.; Li, X.; Tureček, F., The Early Life of a Peptide Cation-Radical. Ground and Excited-State Trajectories of Electron-Based Peptide Dissociations During the First 330 Femtoseconds. *J Am Soc Mass Spectr* **2012,** *23* (3), 446-459.

42. Krauss, D.; Mainx, H. G.; Tauscher, B.; Bischof, P., Fragmentation of trimethylsilyl derivatives of 2-alkoxyphenols: A further violation of the 'even-electron rule'. *Organic Mass Spectrometry* **1985,** *20* (10), 614-618.

43. Arafat, E. S.; Trimble, J. W.; Andersen, R. N.; Dass, C.; Desiderio, D. M., Identification of fatty acid amides in human plasma. *Life Sci* **1989,** *45* (18), 1679-87.

44. Draffan, G. H.; Stillwell, R. N.; McCloskey, J. A., Electron impact-induced rearrangement of trimethylsilyl groups in long chain compounds. *Organic Mass Spectrometry* **1968,** *1* (5), 669-685.

45. Becke, A. D., A new mixing of Hartree–Fock and local density-functional theories. *The Journal of Chemical Physics* **1993,** *98* (2), 1372-1377.

46. Becke, A. D., Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics* **1993,** *98* (7), 5648-5652.

47. Ditchfield, R.; Hehre, W. J.; Pople, J. A., Self-Consistent Molecular-Orbital Methods. IX. An Extended Gaussian-Type Basis for Molecular-Orbital Studies of Organic Molecules. *The Journal of Chemical Physics* **1971,** *54* (2), 724-728.

48. Fukui, K., The path of chemical reactions - the IRC approach. *Accounts of Chemical Research* **1981,** *14* (12), 363-368.

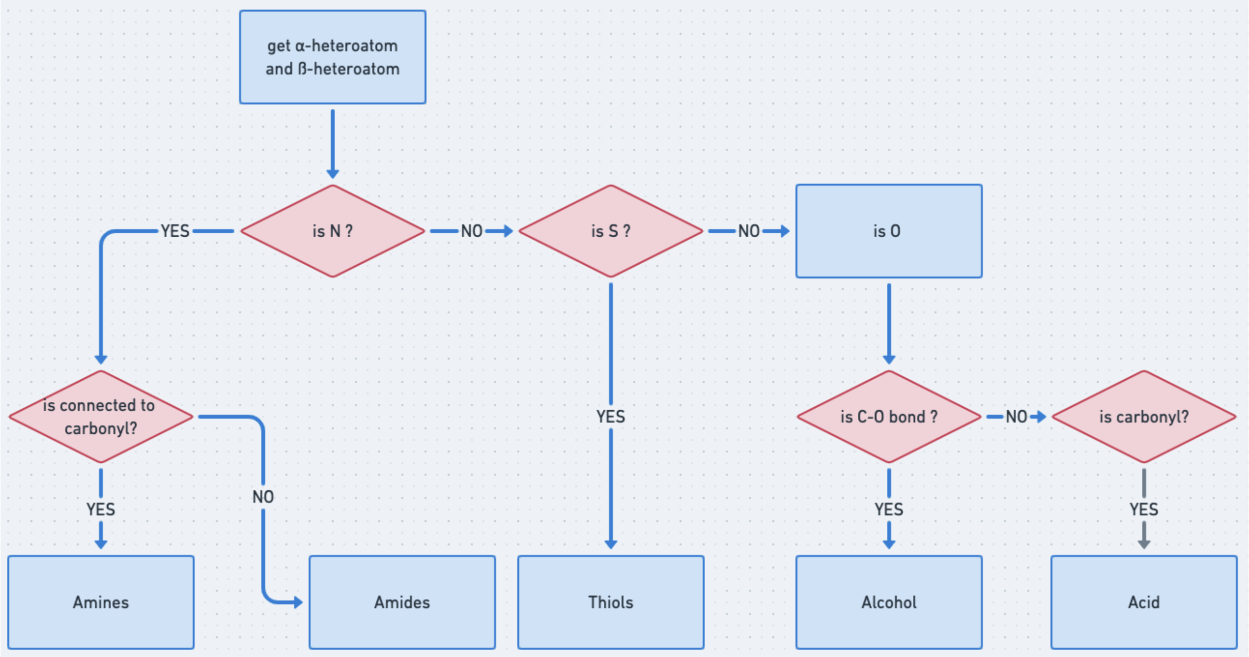
49. Sengupta, A.; Raghavachari, K., Solving the Density Functional Conundrum: Elimination of Systematic Errors To Derive Accurate Reaction Enthalpies of Complex Organic Reactions. *Organic Letters* **2017,** *19* (10), 2576-2579.

50. Schäfer, A.; Horn, H.; Ahlrichs, R., Fully optimized contracted Gaussian basis sets for atoms Li to Kr. *The Journal of Chemical Physics* **1992,** *97* (4), 2571-2577.

51. Weigend, F.; Ahlrichs, R., Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys Chem Chem Phys* **2005,** *7* (18), 3297-3305.

52. Sun, L.; Park, K.; Song, K.; Setser, D. W.; Hase, W. L., Use of a single trajectory to study product energy partitioning in unimolecular dissociation: Mass effects for halogenated alkanes. *The Journal of Chemical Physics* **2006,** *124* (6), 064313.

3.7. Supporting Information



Scheme S1. Flowchart of compound classification

Table S1. Mass spectral similarities of selected compounds in different classes and their cosine and dot-product similarity scores when compared against experimentally obtained reference mass spectra.

#	Name	Dot	Cos
305	trimethylsilyl 2-amino-4-methylpentanoate	373	697
470[a]	trimethylsilyl 4-methoxybenzoate	729	575
486[a]	2-(2-methoxyphenyl) ethoxy-trimethylsilane	571	585
491	trimethyl-(2-propoxyphenoxy) silane	724	529

587	5-pyridin-3-yl-1-trimethylsilylpyrrolidin-2-one	517	473
566	5-methyl-1-trimethylsilylindole-2,3-dione	363	596
444	1-phenyl-2-(trimethylsilylamino) propan-1-one	665	665
535	1-N-trimethylsilylnaphthalene-1,5-diamine	898	538
501	trimethyl(2-phenoxyethylsulfanyl) silane	47	454
648	methyl 2-trimethylsilylsulfanylbenzoate	840	695

a. Discussed in the paper

Table S2. Dot and Cos similarity scores of multi-TMS-derivative compounds

NAME	TMS	#ATOM	ExactMass	Dot	Cos
Methylamine, 2TMS derivative	2	31	175.1212	712	458
Hydroxylamine, 2TMS derivative	2	29	177.1005	746	483
Formamide, 2TMS derivative	2	30	189.1005	633	428
Ethanamine, 2TMS derivative	2	34	189.1369	589	513
Methoxyamine, 2TMS derivative	2	32	191.1161	524	321
Trimethylsilylpropargyl alcohol, 2TMS derivative	2	32	200.1052	327	136

Allylamine, 2TMS derivative	2	35	201.1369	674	440
Propylamine, 2TMS derivative	2	37	203.1525	438	299
1,2-Ethenediol, 2TMS derivative	2	32	204.1001	890	777
Hydroxylamine, 3TMS derivative	3	41	249.1400	159	336
Urea, 3TMS derivative	3	44	276.1509	575	151
Ethanolamine, 3TMS derivative	3	47	277.1713	300	348
3,4-Bis(trimethylsilyl)-1H-pyrazole, 3TMS derivative	3	45	284.1560	660	515
Glycine, 3TMS derivative	3	46	291.15060	550	290

3.7.1. Trajectory analysis

For example (**Figure S1**, #305, L-Leucine-TMS derivative), the molecular ion peak in the reference spectrum almost disappeared and in the in-silico spectrum it is relatively high abundant. The Figure S1b shows the fragments generated by MassFrontier and validated by our accurate in-silico mass spectrum. It is a common situation that under 70 eV most of the molecular ions are fragmented and result in a low abundant molecular ion peak, while the $[M-117]^+$ peak (loss of COOTMS$^{\bullet}$) becomes the base peak. Thus, the molecular ion has a shorter lifetime in the experiment, and we didn’t provide enough energy to break the molecules in the simulation.

The **figure S1c** scans the bond length between the CH3 and the Si under high level calculation (pbe0/6-31G(d)) on a more general model. It shows that a four-member ring generation after the loss of $^{\bullet}$CH3, which is also validated by the QCEIMS simulation.

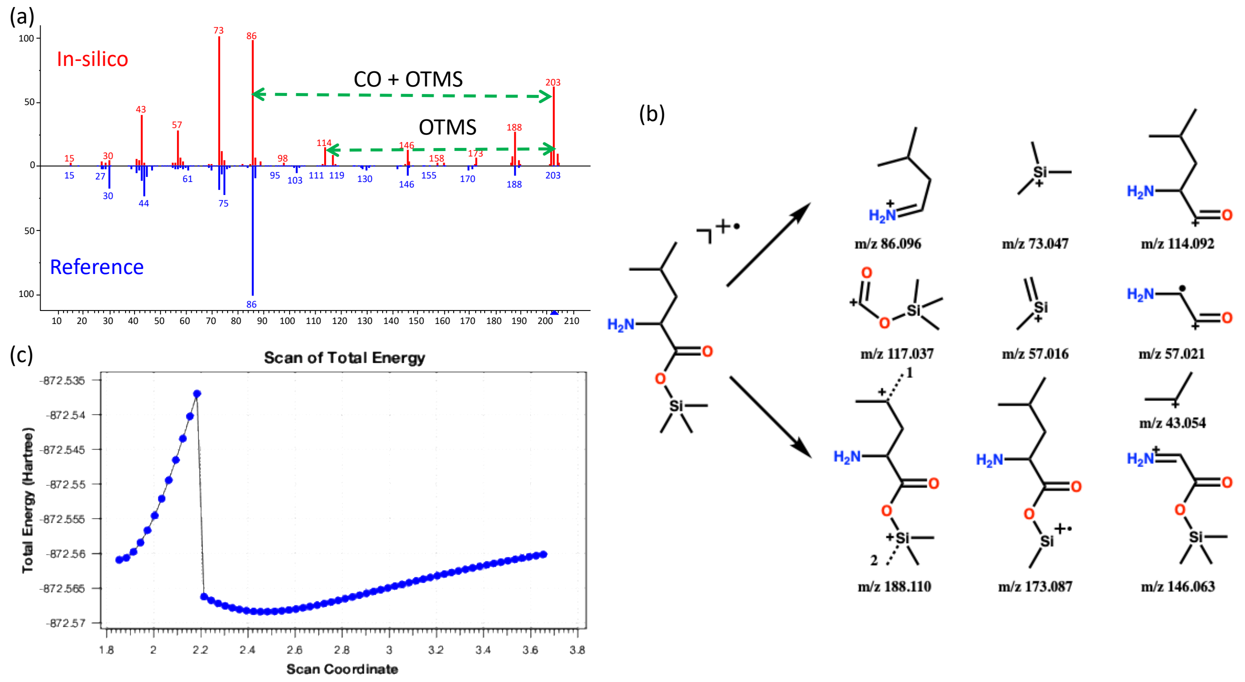


Figure S1. Fragmentation analysis of carboxylic groups; (a) head-to-tail spectrum of trimethylsilyl leucinate against reference spectrum; (b) fragments found in the simulation; (c) energy change of bond length scan

(In-silico spectrum available at https://mona.fiehnlab.ucdavis.edu/spectra/display/MoNA040855)

3.7.2. Mass spectral fragmentation rules (Part II)

Alcohols

We discussed two alcohol isomers with different aromaticity. Figure S2 shows an aromatic alcohol compound. The $[M\text{-}CH3]^+$ and *m/z* 73 peak are characteristic peaks of TMS derivatives. The fragmentations of *m/z* 194 and *m/z* 120 peak are found in the simulation and validated by the MassFrontier prediction, but those peaks are not in the experimental spectrum.

One interesting observation is peaks *m/z* 135 and *m/z* 136 (Figure S2c). There are 18 trajectories generating *m/z* 135 and one peak generating the peak *m/z* 136. Whether the hydrogen on position B migrates to position A decides the different positive charge sites of peak *m/z* 135. Rather than the Si-O-O-C four-member ring structure proposed by MassFrontier, the simulation only found a radical ion *m/z* 136 peak. But in this trajectory (TMP.642), a sequential hydrogen atom rearrangement is observed. The five-member ring in *m/z* 136 and 151 is also a typical structure of ortho-oxygen aromatic compounds. [42]

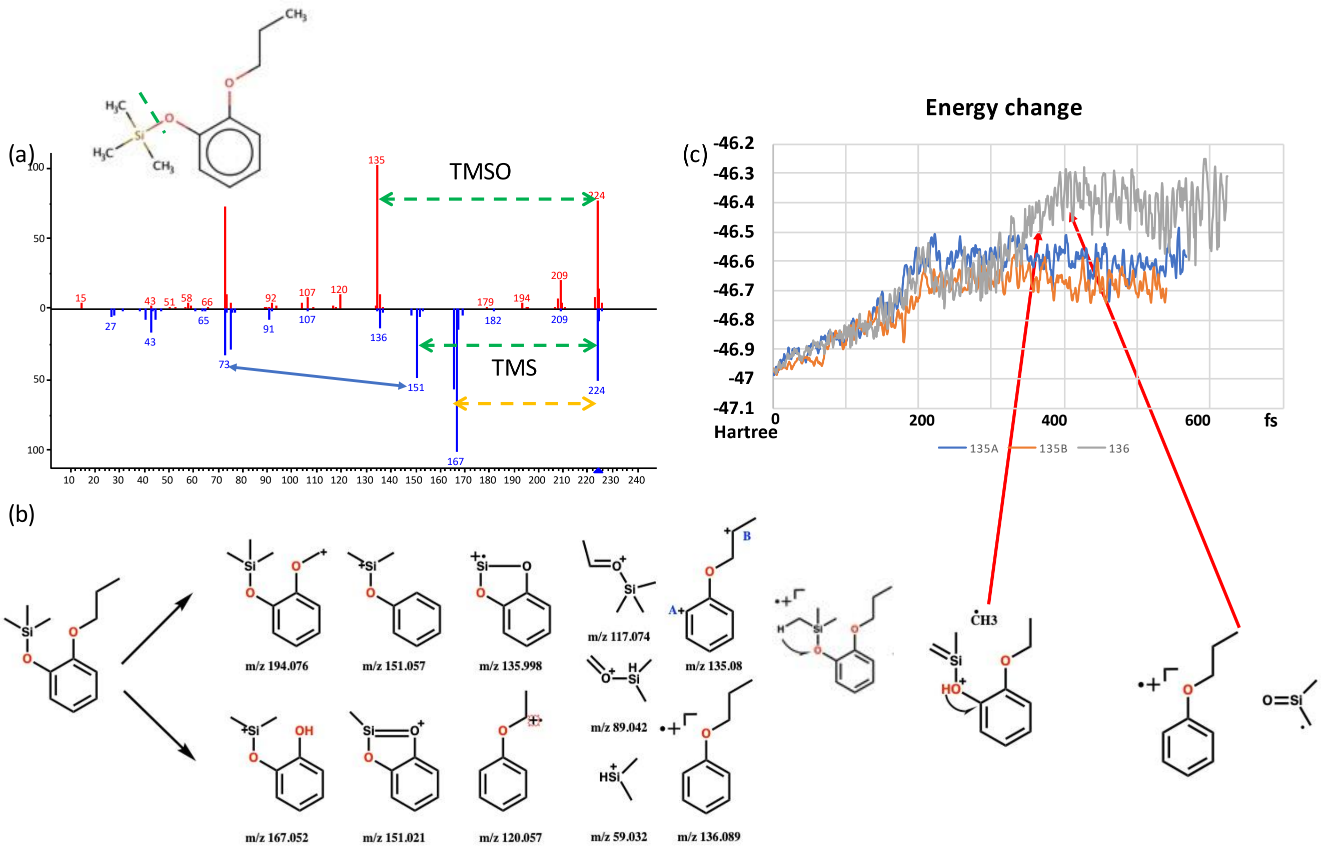


Figure S2. Fragmentation analysis of aromatic carboxylic group; (a) head-to-tail spectrum of 2-propoxyphenol, TMS derivative against reference spectrum; (b) fragments found in the simulation; (c) energy change *m/z* peak 135 and 136

(In-silico spectrum available at https://mona.fiehnlab.ucdavis.edu/spectra/display/MoNA040498)

Amides

Amides contain amine groups next to a ketone functional group. In the case which nitrogen is adjacent to an aromatic carbon, we assigned the molecules as aromatic amides. Amides have a similar structure to carboxylic acids and similar fragments in the mass spectrum are observed. [37, 43] For Aliphatic amides (#587, Figure S3), the simulation reproduced the dominant peaks (*m/z* 234, 219, 165, 145, 73) quite well, but missed the *m/z* 118 peak, which is only one hydrogen less than the peak at *m/z* 119. The generation of C8H8N$^+$ peak contains a nitrogen rearrangement and is shown in the simulation energy plots. The fragmentation reaction started with a TMS group migrating from nitrogen to oxygen (compound A). From compound B to compound C, the energy kept increasing, while it is equilibrated from C to D. We can find four main structures from the trajectory 432.

Amides TMS derivate

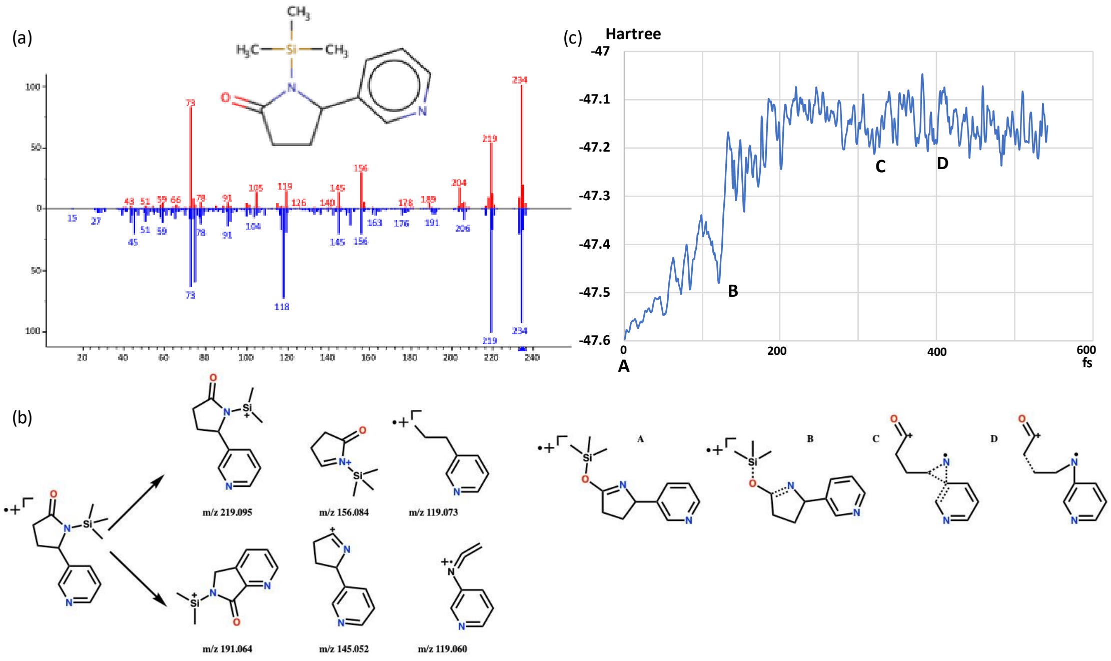


Figure S3. Fragmentation analysis of aromatic carboxylic group; (a) head-to-tail spectrum of Norcotinine, TMS derivative against reference spectrum; (b) fragments found in the simulation; (c) energy-simulation time plot

(In-silico spectrum available at https://mona.fiehnlab.ucdavis.edu/spectra/display/MoNA040305)

The example showed is not a typical aromatic amide (#566, Figure S4). The *m/z* 205 peak comes from the loss of a carbonyl group. The *m/z* 174 peak is missing both in the simulation and MassFrontier prediction. The *m/z* 118 peak is identified as $C7H4NO^{+}$ and $C8H6O^{+\bullet}$ and by their accurate masses.

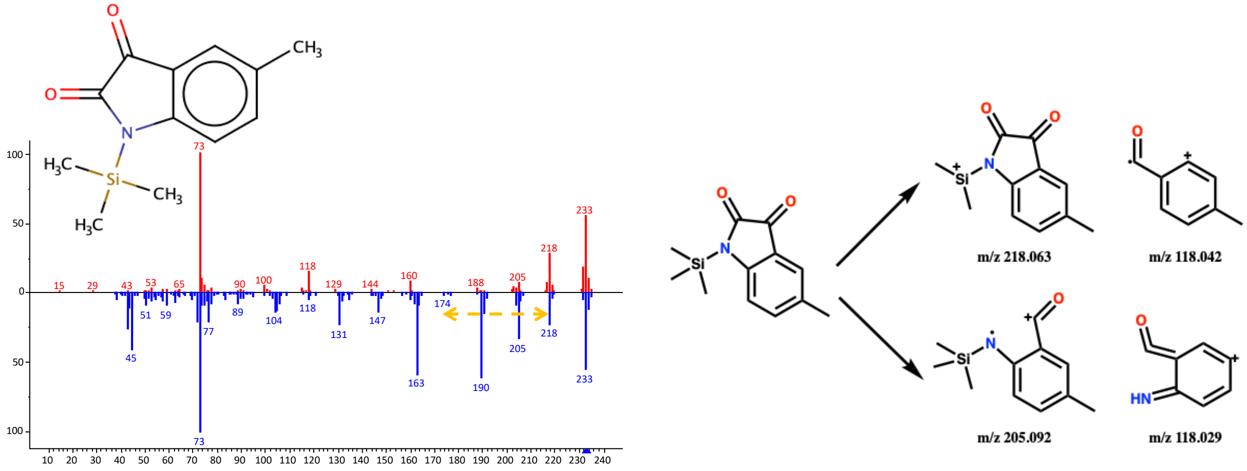


Figure S4. head-to-tail spectrum of 5-Methylisatin, TMS derivative and fragments found in simulation

(In-silico spectrum available at https://mona.fiehnlab.ucdavis.edu/spectra/display/MoNA040974)

Amines

Primary amines (#444, Figure S5) can be double or single TMS derivatives, but here we only focused on one TMS group bound to an amine group. Again, the molecular ion peak is overestimated by QCEIMS, while the other two peaks *m/z* 116 and 73 match the reference quite well. There are two fragments contributing to the *m/z* 116 peak where the C5H14NSi$^+$ peak is the dominant one existing in 130 trajectories and the C6H8N$^+$ fragment is only found in 4 trajectories. The C6H8N$^+$ is an aromatic structure with many degrees of unsaturation and the trajectory shows a SN$_2$ type rearrangement mechanism.

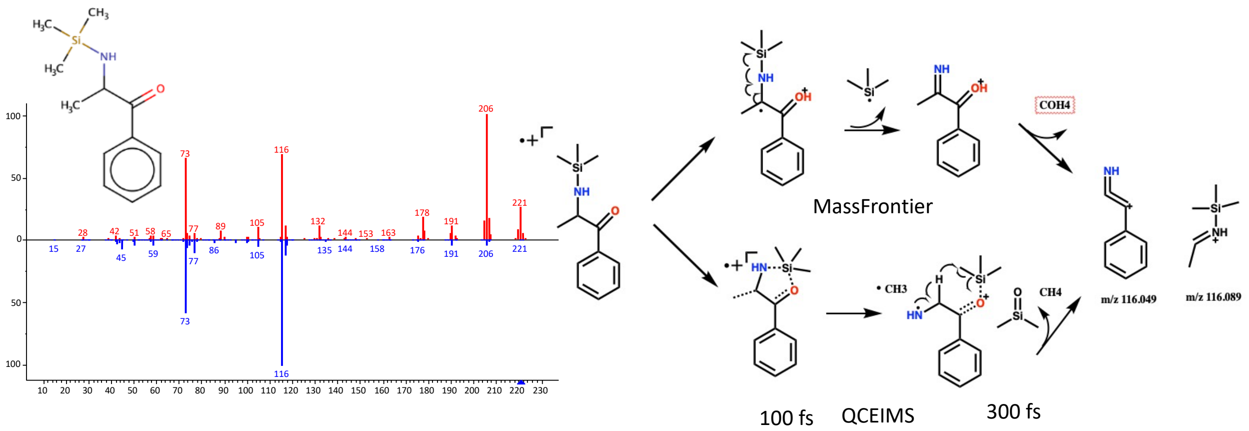


Figure S5. head-to-tail spectrum of Cathinone, TMS derivative and fragments found in simulation

(In-silico spectrum available at https://mona.fiehnlab.ucdavis.edu/spectra/display/MoNA040552)

#535(Figure S6) is an example of Aromatic amines. The Naphthalene ring made its mass spectrum simple and clean, meaning it's a good way to study the relative intensity and the reaction selectivity. The *m/z* 45, 59, 73, 215 peaks are typical TMS derivatives peaks. The *m/z* 158, 157, 142, 141 are resulted from the cation or radical cation stabilized by the naphthalene ring. It should be noted that the *m/z* 100 peak are recorded as C8H4$^{+\bullet}$ and C4H10NSi$^+$ in the simulation and lost in the MassFrontier prediction, while it is presented in the experimental reference spectrum. This reaction needs to be validated by accurate mass spectra. The abnormal structures could come from the inaccuracy of the GFN-XTB method we used, DFT level simulation should be investigated in the future.

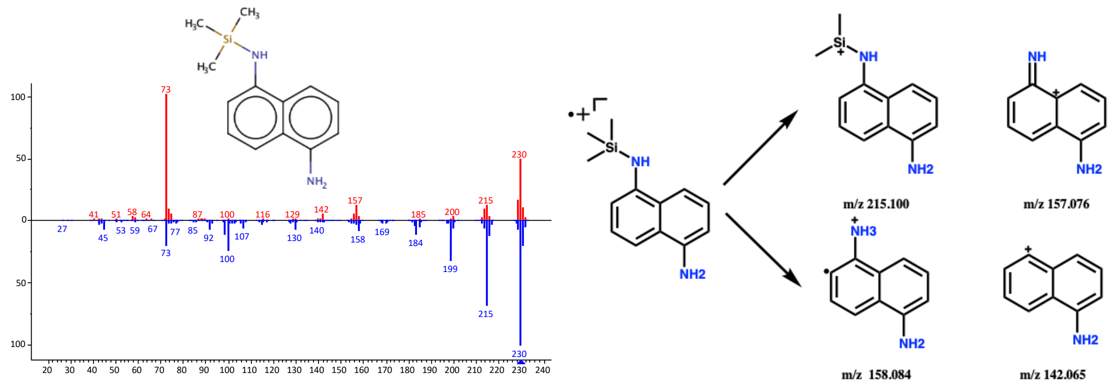


Figure S6. head-to-tail spectrum of 1,5-Diaminonaphthalene, TMS derivative and fragments found in simulation

(In-silico spectrum available at https://mona.fiehnlab.ucdavis.edu/spectra/display/MoNA040622)

Thiols

Thiol compounds exhibit many equivalents to alcohol compounds with a mass difference of 16 Da, for example, *m/z* 133, 106 and 91 [37]. However, in example (#501, Figure S7), the experimental result does not have the [M-30]+ peak (*m/z* 196). The simulation shows an *m/z* 153 peak rather than the *m/z* 151 peak. The latter comes from the loss of two hydrogen atoms and an alpha cleavage. The *m/z* 121 peak results from the loss of (CH3)$_3$SiS$^\bullet$, which is also found in other thiol compound and validated by deuterium labeling. [44] The typical peak *m/z* 106 is also missed in the reference spectrum. The in-silico spectrum has a strong *m/z* 105 peak as C3H9SiS$^+$ coming from the hydrogen atom loss from *m/z* 106. The *m/z* 94 and 93 peak also have one hydrogen difference.

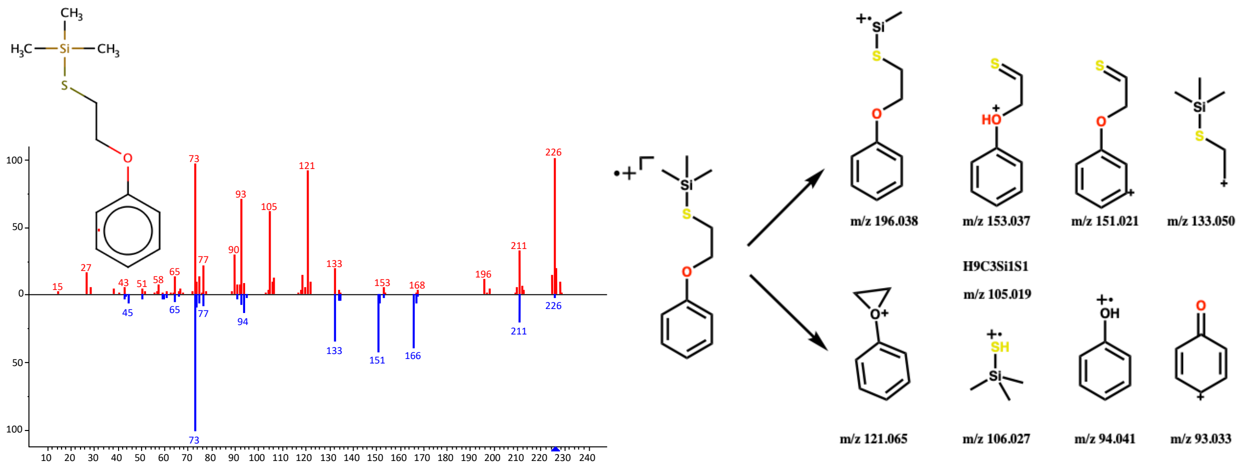


Figure S7. head-to-tail spectrum of 2-Phenoxyethanethiol, TMS derivative and fragments found in simulation

(In-silico spectrum available at https://mona.fiehnlab.ucdavis.edu/spectra/display/MoNA040782)

The aromatic thiol (#648, Figure S8) shows a higher dot product score. The molecular ion peak is missed in the experimental reference spectrum. The *m/z* 195 has two resources $C9H11OSiS^+$ (6 trajectories) and $C8H7O2SiS^+$ (32 trajectories). The *m/z* 136 peak has a high intensity in the reference spectrum and the simulation generates a fragment with two five-membered-ring. The *m/z* 89 peak is in disagreement with the simulated and the reference spectrum. It is identified as $C4H9O^+$ in the simulation and the MassFrontier software was not able to predict this reaction.

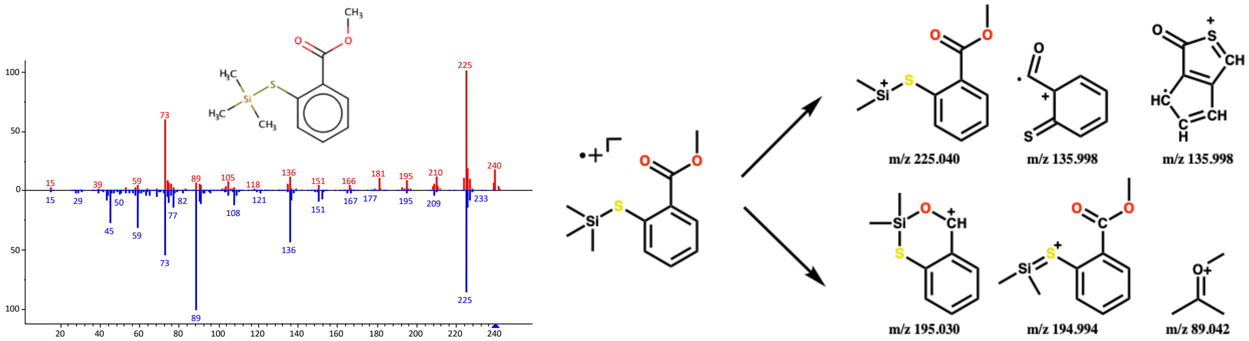


Figure S8. Head-to-tail spectrum of Methyl thiosalicylate, TMS derivative and fragments found in simulation

(In-silico spectrum available https://mona.fiehnlab.ucdavis.edu/spectra/display/MoNA040841)

3.7.3. Accurate Mass spectra

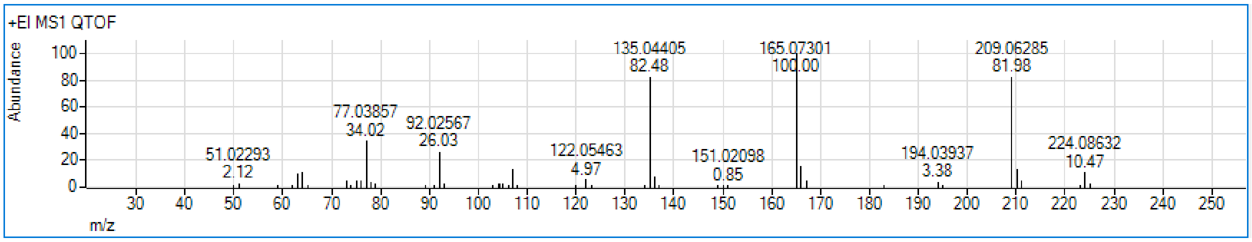


Figure S9. Accurate mass EI-MS of Trimethylsilyl 4-methoxybenzoate

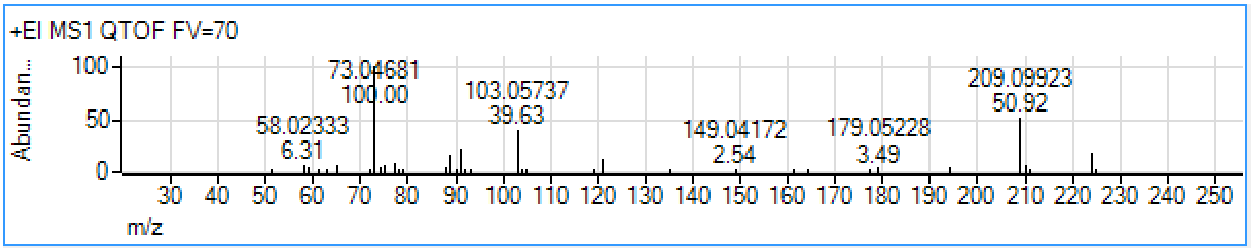


Figure S10. Accurate mass EI-MS of [2-(2-Methoxyphenyl)ethoxy](trimethyl)silane

3.7.4. Estimate Processing time of QCEIMS TMS compounds

Because the calculations are divided into trajectories, the program has a good ability of parallelization, while at the same time it's hard to analysis the average calculation time. Instead, we use the time of a single point calculation and assume that each trajectory has 5000 steps to estimate the running time of each molecule. Because we used two different type of CPUs, we only evaluate 464 molecules calculated on Intel Xeon E5-2699Av4 CPUs.

$$Estimate\ running\ time = t(single\ point) \times 5000 \times n(trajectories)$$

Compared with our previous project[11], the average calculation time increases from 1.6 h to 1.9h (on 44 CPU threads) where better PESs and larger molecules (TMS derivatives) are simulated. Because of the better parameterization and algorithm, the QCEIMS v4.0 has a computational effort scales better than O(N^2). Thus, the increase of simulation time can be bypassed by better parallelized property.

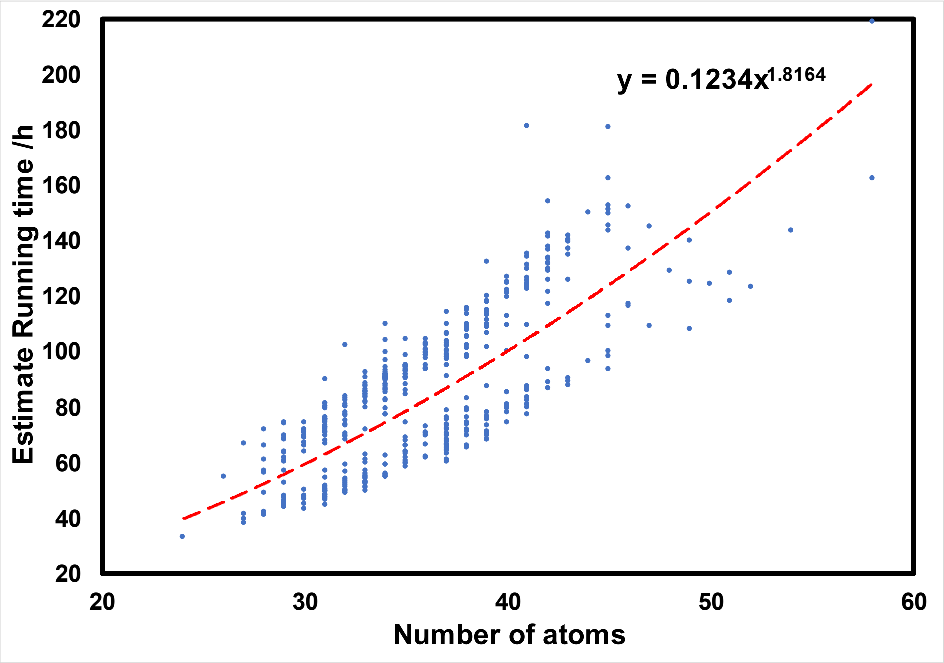


Figure S11. Processing time of 464 molecules versus the number of atoms

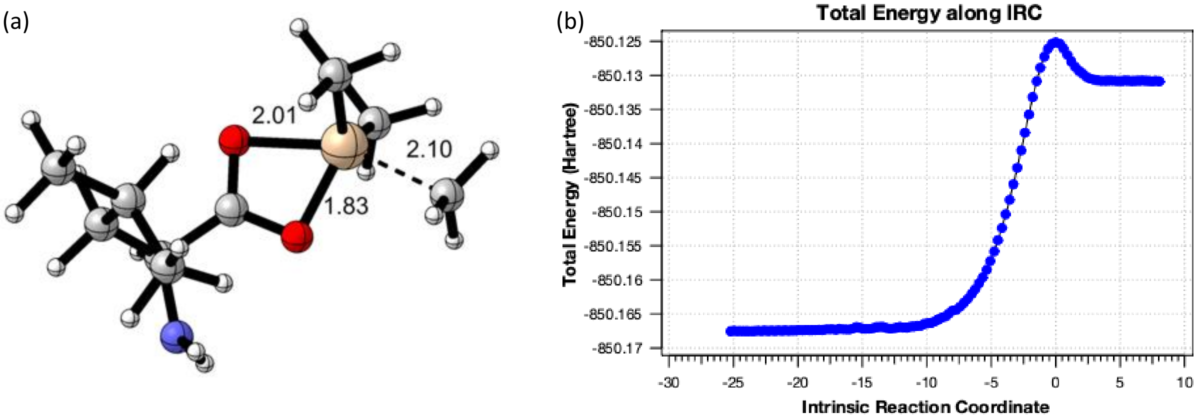


Figure S12. (a) The transition state structure of methyl group loss under B3LYP[45-46]/6-31G[47] level (b) The intrinsic reaction coordinate (IRC)[48] calculation with Gaussian09[13] proves the transition state structure linking reactants and products. The energy barrier of this reaction is 23.18 kcal/mol under wb97xd[49]/def2-TZVPP[50-51] level.

3.7.5. Energy partition analysis

The classical single trajectory dynamics approach[52] is used to calculate the translational, rotational and vibrational energies, where the molecule is treated as rigid body. The python code is adopted from Kurrouchi et al.[35] The total kinetic energy is calculated by equation (1), where i donates atom number and $v_i(t)$ is the velocity of atom i at time t.

$$E_{kin}(t) = \sum_{i=1}^{n} \frac{1}{2} m_i v_i^2(t) \ldots\ldots(1)$$

The translational energy was calculated as

$$E_{trans} = \frac{1}{2} M V_G^2(t), M = \sum_{i=1}^{n} m_i \ldots\ldots(2)$$

Where V_G is the velocity of the center of mass.

The translation motion is removed to set the original point to the center of mass:

$$r_i'(t) = r_i(t) - R_G(t), v_i'(t) = v_i(t) - V_G(t) \ldots\ldots(3)$$

The principal moment of inertia $I_x(t), I_y(t), I_z(t)$ and the rotation matrix Φ are calculated by diagonalizing the inertia tensors. Then, the atom coordinate and velocity are rotated to the principal axes of the molecule:

$$r_i''(t) = \Phi r_i'(t), v_i''(t) = \Phi v_i'(t) \ldots\ldots(4)$$

With that, the angular momentum can be calculated as:

$$L_x(t) = \sum_{i=1}^{n} m_i \left(0, r_y'', r_z''\right) \times \left(0, r_y'', r_z''\right),$$

$$L_y(t) = \sum_{i=1}^{n} m_i \,(r_x'', 0, r_z'') \times (r_x'', 0, r_z''),$$

$$L_z(t) = \sum_{i=1}^{n} m_i \,(r_x'', r_y'', 0) \times (r_x'', r_y'', 0) \ldots\ldots(5)$$

The angular velocity around the principal axes w , the rotational energy and the rotation velocity can be calculated as:

$$L_x(t) = I_x(t)\mathrm{w}_x(t),$$

$$E_{rot} = \frac{1}{2}\left\{I_x(t)\mathrm{w}_x^2(t) + I_y(t)\mathrm{w}_y^2(t) + I_z(t)\mathrm{w}_z^2(t)\right\},$$

$$v_{i,rot}(t) = \{\mathrm{w}_x(t), \mathrm{w}_y(t), \mathrm{w}_z(t) \times r_x''(t)\}\text{......}(6)$$

Once we remove the rotation movement, the vibrational energy can be calculated:

$$v_{i,vib}(t) = v_i''(t) - v_{i,rot}(t),$$

$$E_{vib} = \sum_{i=1}^{n} \frac{1}{2} m_i v_{i,rot}^2(t)\ldots\ldots(7)$$

With the energy partition analysis, we can separate the vibrational energy from the total kinetic energy in classical mechanics trajectory.

Chapter 4: Beyond the ground state: predicting electron ionization mass spectra using excited state molecular dynamics

4.1. Abstract

Here, we provide an algorithm that introduces excited states into the molecular dynamics prediction of 70 eV electron ionization mass spectra. To decide the contributions of different electronic states, the ionization cross section associated with relevant molecular orbitals were calculated by the Binary-Encounter-Bethe (BEB) model. We used a fast orthogonalization model/single and double state configuration interaction (OM2/CISD) method to implement excited states calculations and combined this with the GFN1-xTB semi-empirical model. Demonstrated by predicting the mass spectrum of urocanic acid we showed better accuracies to experimental spectra using excited state molecular dynamics than calculations that only used ground state occupation. For several histidine pathway intermediates we found that excited state corrections yielded an average of 73% more true positive ions compared to the OM2 method when matching to experimental spectra, and 16% more true positive ions compared to GFN method. Importantly, the exited state models also correctly predict several fragmentation reactions that were missing from both ground state methods. Overall, for 48 calculated molecules we found best average mass spectral similarity scores for the mixed excited state method compared to the ground state methods using either cosine-, weighted dot-score or entropy similarity calculations. Therefore, we recommend adding excited state calculations for predicting electron ionization mass spectra of small molecules in metabolomics.

4.2. Introduction

Using quantum chemical methods along with statistical methods to predict electron ionization (EI) mass spectra (MS)[1-2] has been explored for many types of molecules, including organic molecules, inorganic molecules[3], heavy metal-containing molecules[4]. The quantum chemistry based QCEIMS software [1, 3-9] can provide reasonable results and detailed reaction pathways. In a recent publication, the QCxMS software was introduced, combining electron ionization and collision-induced dissociation modelling into a single software package. [10-11] However, a shortcoming of these methods is the ground state potential energy surface (PES) may have inaccurate results even with density functional theory (DFT) methods[8, 12], which can cause missing fragment ions. For example, Wang et al. showed that only 50% of observed ions were generally captured in 681 trimethylsilylated molecules, which compromised the accuracy of QCxMS simulations. To improve predictions of relative energies of structures on PESs, alternative theoretical methods (e.g., Post-Hartree-Fock methods[13]) can be utilized. In addition, the inclusion of excited states might improve predictions by accounting for fragmentation reactions that are not accessible on the ground state PES.[14-15] The focus of the current study was to examine the impact of including excited states in the semi-empirical molecular dynamics method for prediction of 70 eV EI mass spectra.

The EI ionization process can be described as a (1e, 2e) gas phase reaction:

$$M + e^{-} \rightarrow (AB)^{+\cdot} + 2e^{-}$$

In this process, the analyte molecule M is impacted by e^{-}, which will be scattered and cause the loss of another e^{-}, resulting in the reaction complex $(AB)^{+\bullet}$, which can undergo further reactions.[16] Ionization cross sections play an important role in providing information of the EI process and the Binary-Encounter-Bethe (BEB) model[17] provides an ab-initio means of calculating ionization cross sections without any fitting parameters. The 70eV electron can traverse the molecule in a few femtoseconds, a much shorter time than the bond vibration period. Therefore, the transition from the *v*=0 vibrational state of the ground electronic state to an excited state generally obeys the Frank-Condon principle and can be modeled as “vertical ionization” [18-20]: one electron is removed from the neutral molecule with the molecular structure unchanged.

The impact energy can be divided into ionization energy and impact excess energy. According to the vertical ionization model, upon ionization, the impact excess energy is saved in highly excited vibrational modes, and this is the driving force of future fragmentation. With the increase of molecular size, energy must be distributed to more degrees of freedom (DOF), and thus more energy is needed for redistribution. [21] QCxMS assumes that the excited ion state goes through an internal conversion to the vibrationally hot ion ground state.[1] This model introduces the impact excess energy (IEE) to the nuclear DOF in a continuous time by increasing the velocity of each atom. QCxMS has been tested successfully in many cases[3-9] and it has been shown that the IEE distribution model[1] has a small effect on the simulated spectra.

We hypothesized that simulations including excited states could reveal reactions not encountered in ground state simulations. Because we focused on finding more reactions, non-adiabatic coupling between excited states[22] was not considered in this project. Instead, in the excited state molecular dynamics, the molecule starts at different states, but will jump back to the ground state once the fragmentation is detected. In addition, one must determine which methods are the most appropriate to describe relevant excited state PESs in terms of both accuracy and computing resource feasibility. We found that the orthogonal-corrected semiempirical quantum-chemical methods OMx[23] can be used for ground state mass spectral predictions.[12] To include the dynamic and static electron-correlation effects for excited states, configuration interaction (CI) and multireference

(MR) methods are needed.[24] A graphical unitary-group approach (GUGA)[25] can be combined with OMx models. In this way, we can capture the excited state correlation effects with state-based semiempirical treatments. [26-29] In this paper, we provide a prototype of a mass spectra prediction model and discuss additional method improvements for large-scale MS predictions. This model combines ground state and excited state molecular dynamics based on the BEB model.

4.3. Methods

4.3.1. Ionization cross section by BEB model

The BEB model is simplified from the binary-encounter-dipole model[17], and has wide applications, including mass spectrometer normalization, plasma modeling and material radiation effects calculation. [30] The electron impact ionization cross section for molecular orbital i (MO_i), is calculated by:

$$\sigma_i = \frac{S_i}{t_i + u_i + 1}\left[\frac{\ln t_i}{2}\left(1 - \frac{1}{t_i^2}\right) + 1 - \frac{1}{t_i} - \frac{ln\, t_i}{t_i + 1}\right]$$

$$u = \frac{U}{B}, t = \frac{T}{B}, S = \frac{4\pi a_0^2 N R^2}{B^2}$$

$$a_0 = 0.592\text{Å}, R = 13.61\ eV$$

Where T is the energy of the impact electrons; B is the electron binding energy of MO_i, U is the kinetic energy of MO_i; the occupation number N of MO_i is two for ground state molecules. The GAMESS[31] package is used to calculate the molecular orbital properties under Hartree-Fock method with 6-31G basis set. The GAUSSIAN program[32] is used to optimize the structure and calculate the molecular orbital contours. The Avogadro v1.2 software [33] was used to visualize the molecular orbitals. A python script package (https://github.com/Shunyang2018/EXMD) was developed to calculate the electron impact ionization cross section and relative ratio of each electronic states after ionization.

4.3.2. Modified QCEIMS algorithm

The QCEIMS v4.0 [1, 3-9] code was used with several modifications to conduct excited state molecular dynamics. Default settings were applied for the ground state calculations at the GFN2-xTB level[34]. For the excited state, uniform velocity scaling was enforced during the internal conversion step. The MNDO99 program [35] was used for the semiempirical OM2[23, 36] level gradient calculations for excited state molecular dynamics (MD)[25, 28, 37]. The active space is decided by the following rule set. Each π bond provides a pair of occupied/unoccupied orbitals, and each oxygen or nitrogen atom provides a lone pair orbital. Because the radical cation system is open shell, only one reference occupation is used. Restricted Open-shell Hartree-Fock (ROHF) is used, while single and double excitations are allowed for the reference configurations for simplicity. Different parameter settings of the excited state molecular dynamics were tested in Figures S8-S10. Once fragmentation is detected, the ionization potential (IP) of each fragment is calculated and partial charge is assigned by the Boltzmann distribution. Another MD simulation will be performed on the fragment with largest partial charge for secondary fragmentations. This MD simulation is for the ion ground state, where the fractional orbital occupations[38] in unrestricted OM2/SCF calculations are used. Ions larger than 15 Dalton are counted and used to generate in-silico spectra of ground state and excited states separately. Then, the excited state (D_1, D_2…) spectra are used as corrections to the ground state (D_0) spectrum per their relative ratios obtained from the BEB model, assuming that the D_0 state is ionized from the highest occupied molecular orbital (HOMO) with an ionization cross section σ_0, D_1 is from HOMO-1 with σ_1 and so on. Theoretically, D_2 and higher excited states can contribute to fragmentation reactions, but we found that D_1 state calculations already predicted most of the experimentally observed reactions. Importantly, we did not aim to theoretically or comprehensively compare all different methods, but to give practical applications on molecules that are typically encountered in metabolomics research. The OM2/CISD method had an average failure rate of 0.64 across all trajectories of all 48 molecules due to self-consistent field convergence problems (Supplement S8). For simplicity, only the two lowest electronic states (D_0, D_1) are taken as reference states and non-adiabatic crossing is neglected. Because excited state calculations are used as corrections, this approach can be extended to other higher excited states in the future.

4.4. Results

4.4.1. Urocanic Acid as a demonstration case

Urocanic acid is an intermediate of histidine catabolism.[39] We chose urocanic acid as an example because it contains both nitrogen and oxygen elements, an imidazole aromatic system, and a carboxylic acid functional group, features typical of many organic molecules of biological interest. The MOs of urocanic acid are shown in Table S1 and visualizations of the MO contours are shown in Figure S2. The HOMO, HOMO-1, HOMO-3, and HOMO-5 orbitals are π orbitals, and the HOMO-2, HOMO-4, HOMO-6, HOMO-7, HOMO-8, and HOMO-9 are n orbitals associated with lone pairs on oxygen and nitrogen. Consequently, an active space of 11 electrons and 10 orbitals (11, 10) should be sufficient for modeling the first excited state of the urocanic acid radical cation. The MO ionization cross section according to electron kinetic energy of the four highest occupied MOs is shown in Figure 4-1. Vertical ionization from the HOMO generates the D_0 state (remove one electron from HOMO), while from HOMO-1 generates the D_1 state. According to the ratio of σ at 70 eV, the ground state is significantly more likely than other states, thus the spectrum from ground state MD contributes most to the final corrected spectrum. The apex of the ionization cross section curve is slightly lower than 70 eV and shifts to 70 eV with lower energy MOs, which is consistent with the region of highest ionization efficiency. That is the basic reason why 70 eV is the classic experimental energy for electron ionization in gas chromatography-mass spectrometry.

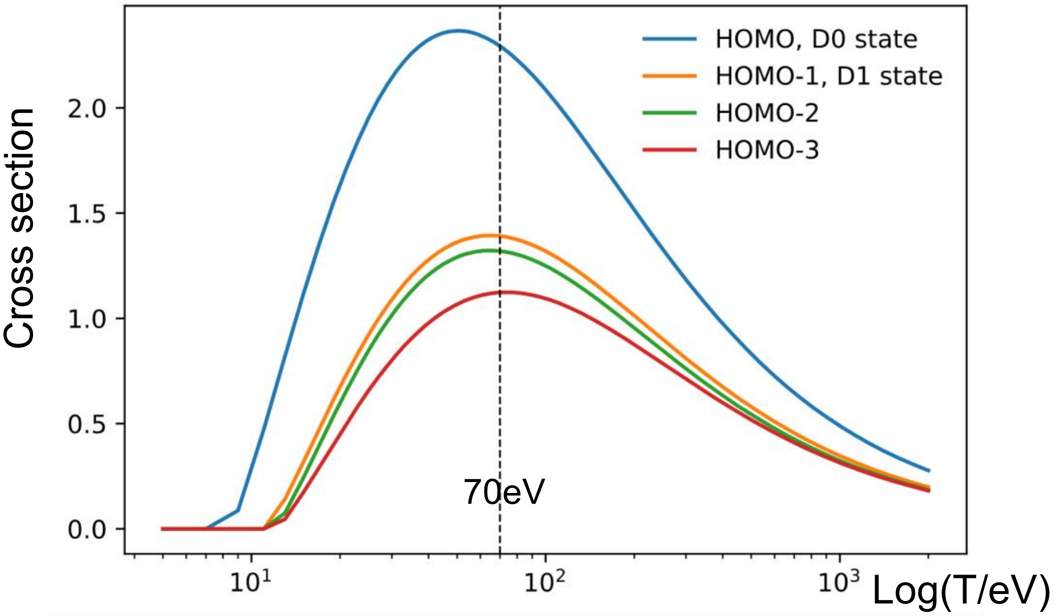


Figure 4-1. Ionization cross section of four highest molecular orbitals of Urocanic acid; dash line denotes the 70 eV kinetic energy used in EI; blue line denotes the Highest Occupied Molecular Orbital, which has the largest ionization cross section.

Different quantum chemistry methods, including GFN1-xTB, GFN2-xTB, OM2 and PBE0[40]-D3[41]/SV(P)[42] (density functional theory) were tested on urocanic acid in the ground electronic state (Figure 4-2). OM2 (Figure 2c) is one of the fasted methods available in the QCEIMS program, but it is only parameterized for five elements: C, H, O, N, F. The default GFN1-xTB method and its advanced GFN2-xTB version (Figure 2a) yielded only minute differences in predicted mass spectra when applied to the chemical urocanic acid. More GFN2-xTB calculations can be found in Support File S13. Interestingly, neither the PBE0/SV(P) method (Figure 2b) nor GFN1-xTB or GFN2-xTB correctly predicted the experimental *m/z* 45 fragment ion. The PBE0/SV(P) method requires larger computing resources. This disadvantage is amplified if hundreds of trajectories need to be calculated. In comparison to the GFN1-xTB method, the prediction of relative intensities of other fragment ions (m/z 138, 93, 39) did not improve with PBE0/SV(P) method; more false positive ions (m/z 98, 44) were captured and ions missing in simulations with the semi-empirical methods were still not found. The comparison between these three different methods showed that optimizing the potential energy surface did not solve all the problems in mass spectral predictions.

The spectrum calculated from the first excited state MD and the mixed spectrum after correction of the first excited state are compared in Figure 2c and 2d. When moving from the ground state to the excited state, the molecular ion intensity decreased dramatically, and more fragment ions and higher intensities of these ions were found in the low mass range. This change can be explained by a higher reactivity of the excited state. We used the weighted dot product score [12] to evaluate the similarity between in-silico spectra and the experimental reference spectra given in the NIST 17 mass spectral library. Although the dot product score slightly decreased from 907 to 894 with excited state correction, the details of the in-silico spectrum improved. For example, the group of fragment ions around m/z 28, 39 and 67 were better captured, giving a higher confidence when comparing spectra. The *m/z* 39 fragment ion was identified as C_2HN by Mass Spectrum Interpreter[43] software, a product that was missing in all ground state simulations. In the first excited state simulation, the *m/z* 39 ion was found in 30 out of 400 trajectories, arising from C_2HN^+ and $C_3H_3^+$. Overall, the mixed spectra predicted 62 instead of only 44 fragment ions from the D_0 spectrum. The number of true positives, i.e. predicted ions that found in experimental spectra, increased from 31 to 37. Some problems remained unsolved even when including the D_1 excited state, however. For example, the excited state MD overestimated the intensity of low mass ions, especially *m/z* 28 and 40. This problem might be solved by including more electronic states.

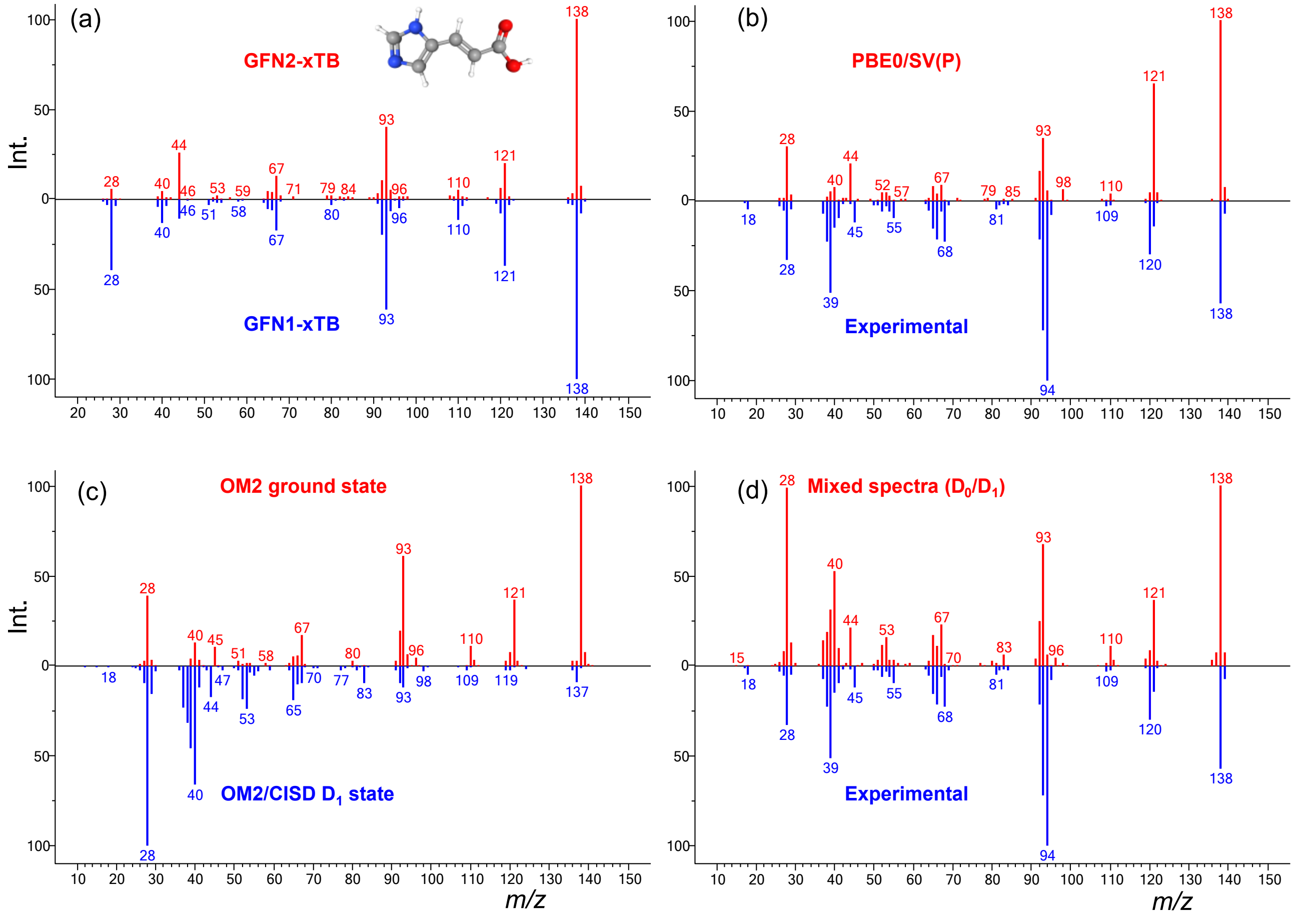


Figure 4-2. In-silico spectra of urocanic acid using different simulation methods. (a) semi-empirical level GFN2-xTB versus GFN1-xTB. (b) DFT level PBE0/SV(P) versus the experimental spectrum from the NIST17 library; (c) semi-empirical level OM2 at ground state versus the semi-empirical configuration interaction level OM2/CISD (first excited state); (d) mixed spectrum of D0 (OM2) and D1 (OM2/CISD) simulations versus the NIST17 experimental spectrum

When we compared the experimental spectrum of urocanic acid with all in-silico generated spectra by all molecular dynamic methods, we found that three significant fragment ions were missed or underestimated, especially evident when removing isotope ions (Figure S12). Here, m/z 68 was missing from all simulations, whereas *m/z* 120 and *m/z* 94 were represented much found at much lower intensity in predicted spectra compared to experimental spectra (Figure 2). All three ions shared a similar pattern: while the exact fragment ion product is missing or very low abundant, there was an abundant ion within 1 Dalton of its expected location (*m/z* ± 1). This observation implied that hydrogen rearrangements were not described correctly in the simulations. For example, the *m/z* 94 fragment ion results from a loss of CO_2, which is a product of a three-membered ring rearrangement reaction of hydrogen that is transferred to a double bond. We found that the other two fragments ions *m/z* 120, 68 were also generated by hydrogen rearrangements, as discussed later in more detail.

4.4.2. Rearrangement reaction in the unimolecular dissociation

To validate the overall performance of excited state corrections and to investigate the rearrangement reactions, we selected several molecules from the histidine biosynthetic pathway.[44] Seven of the 10 major pathway intermediates had 70eV EI mass spectra included in the NIST 17 database. Table 4-1 shows the parameters used in the calculation. On average, the ground state simulation contributed around 60% to the final mixed spectra, while the first excited state contributed around 40%. If we added the second excited state with a similar weight as D_1 state, the contribution of the ground state would be decreased to around 40%. The mass spectral prediction with the OM2 method had an average running time of 1.55 h per molecule, while the GFN1-xTB required 7.2 h per molecule, both on 16 CPU threads. The OM2/CISD simulation took about twice as long as the OM2 ground state calculation. MD on the first excited state required a similar amount of simulation time as the ground state simulation. GFN1-xTB does not allow for computations of excited state MD trajectories. The OM2 method has its own disadvantage because it is only parameterized for limited elements, C, H, N, O and F, while GFN1-xTB includes 86 elements.

Table 4-1. Parameters used for simulating ground state/ excited state ratios. The fraction of D_0 and D_1 was decided by the ionization cross section. The sum of unoccupied and occupied molecular orbital is the size of active space. The number of active electrons is calculated as the product of electrons in the molecular orbitals minus 1.

Name	Fraction D_0	Fraction D_1	Unoccupied MO	Occupied MO	Active electron
Alanine	0.58	0.42	2	4	7
Glutamic acid	0.58	0.42	2	7	13
Histidine	0.60	0.40	3	6	11
Histamine	0.60	0.40	2	5	9
Carnosine	0.59	0.41	4	11	21
1-Methyl-histidine	0.59	0.41	3	8	15
Urocanic acid	0.62	0.38	4	6	11

As shown in Figure 4-3, most ions fragments derived from alanine were correctly predicted. The in-silico spectra predicted the loss of two or three hydrogen fragments ($[M-1]^+$, $[M-2]^{+\bullet}$ ions) from the molecular ion *m/z* 89. Such hydrogen losses were not verified by the experimental spectrum and originated from the OM2 method parametrization.[12] The prediction of the histamine mass spectrum benefited from the additional ions predicted by including the D_1 excited state. However, the base ion intensity for *m/z* 82 was underestimated in both the D_0 and D_1 simulations. As in the urocanic acid example described above, this problem is a result of hydrogen rearrangement reactions.

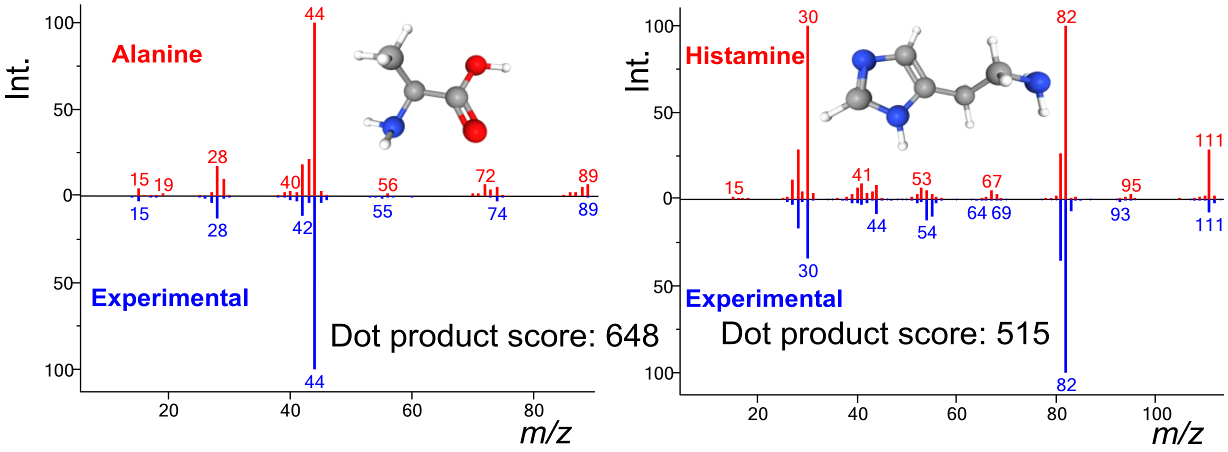


Figure 4-3. Head-to-tail spectra of alanine and histamine; Top: mixed spectra obtained by GFN1-xTB/OM2/CISD; Bottom: experimental spectra from the NIST17 library.

Head-to-tail plots for the other molecules from Table 1 can be found in Figure S4-S7. For four of the seven molecules examined, the base ions were missing (*m/z* 94 of urocanic acid, *m/z 82* of histidine and histamine, *m/z 96* of 1-methylhistidine). The rule-based Mass Frontier software (Thermo Fisher Scientific Inc.) predicted that these ions were generated by hydrogen rearrangement reactions or a mobile proton mechanism.[45-47] The MD trajectories in our calculations also supported this model; however, we observed very few occurrences of these specific hydrogen rearrangement reactions. As an example, we plotted the energy versus time for two trajectories at ground state simulations of the fragmentation of the urocanic acid radical cation (Figure 4-4). Trajectory #13 generated m/z 93 (C_5N_2 H_5^+), with a neutral loss of $CO_2H^{\bullet}$, while trajectory #318 generated m/z 94 ($C_5H_2H_6^{+\bullet}$), with a neutral loss of carbon dioxide. The difference was caused by a hydrogen rearrangement between two fragments in the latter trajectory. Although there was a mobile hydrogen in trajectory #13, it was only intramolecular rearrangement of $C_5N_2H_5^+$ fragment. To better visualize the electronic energy change, we plotted the moving average trend line per 100 fs. This trend line showed that energies stayed unchanged once the fragments were generated in trajectory #13. In trajectory #318, the energy decreased after an energy barrier was surmounted. Because unimolecular reactions do not include collision and energy exchange, the system electronic energy will not change for most homolytic bond cleavages. In other words, energy curves will stay at the same level after passing through transition states. However, rearrangement fragmentations face different conditions because forming new bonds during a rearrangement reaction decreases the system's total electronic energy level. Therefore, rearrangement reactions are usually exothermic. After the transition state is passed, products will be found at energetically lower states, and this extra energy is converted to the kinetic energy (translational energy of fragments). This is the so-called kinetic energy release (KER) process.[48-49] In the QCxMS simulation, all molecular dynamics steps are under the constant total energy (NVE ensemble), except the heating process. These two different energy change types (Figure 4) were consistent with our observations that the electronic energy converted to translational energy of fragments and caused them to depart while the total energy was conserved. This observation proved that the model captured the hydrogen rearrangement reactions as an exit channel on the ground/excited state potential energy surface (PES). Yet, relative intensities of fragment ions (selectivity of the reactions) were still underestimated.

One possible reason is the simulation time scale (on the level of a picosecond) is limited when compared to the total time that mass spectrometers use (on the level of a microsecond[50]). The reaction time scale of unimolecular rearrangement reaction in a mass spectrometer is $10^{-11} \sim 10^{-6}$ s, while a simple bond dissociation fragmentation proceeds much faster at 10^{-12}s.[51] The typical ion flight time in a quadrupole or Time-of-flight mass spectrometer is around 50 μs.[52] Our parameter settings simulated molecular dynamics for up to 5,000 femtoseconds which reproduced most reactions well but failed to account for the frequency of rearrangement reactions. But even for simple fragmentation reactions, 5 ps may not be sufficient time as we found 42 / 400 trajectories that did not yield any fragmentations within the maximum simulation time. For the PBE0/SV(P) method we found even 115 trajectories that did not yield any fragmentation, increasing the prediction of the relative abundance of the unfragmented molecular ion and thereby lowering the weighted dot similarity score by 230 units, a decrease of 15% in prediction scores. Another explanation why rearrangement reactions were missed is that such reactions might need a specific starting conformation to overcome positional barriers[53]. More comprehensive conformer sampling might overcome this problem. Yet, when we tested conformer-rotamer sampling and Wigner distribution sampling, we did not improve results for the lacking rearrangement reactions. Hence, for correctly predicting hydrogen rearrangements, the limitation of simulation time may remain as the main obstacle.

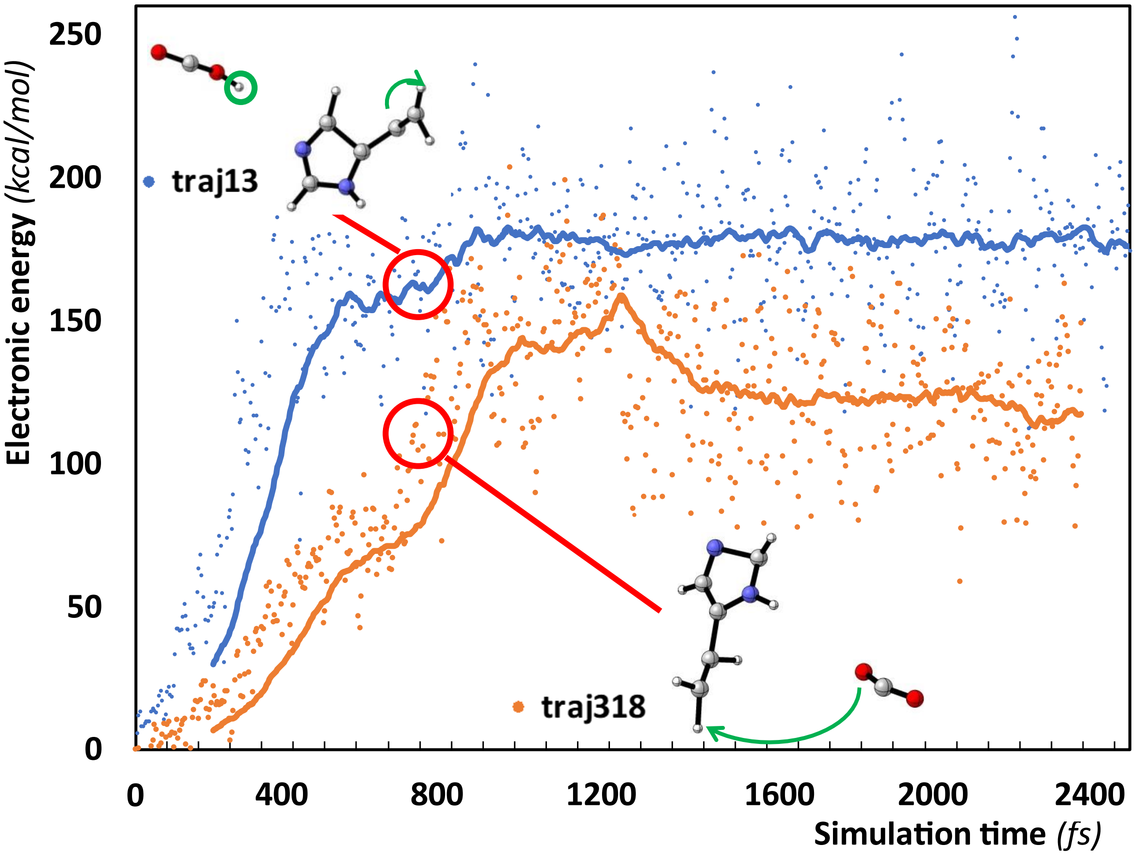


Figure 4-4. Electronic energy change of the whole system according to the simulation time of urocanic acid; (1) blue: trajectory 13: $C_5N_2H_5^+$ and neutral $CO_2H^\bullet$; (2) orange: trajectory 318 $C_5H_2H_6^{+\bullet}$ and neutral CO_2; (3) all trend lines are moving average per 100 fs; (4) insert: the structures at around 780 fs

4.4.3. Method test on small molecules

The examples discussed before do not provide a theoretical validation of the value of adding an excited-state method to the ground state method. Unfortunately, more accurate ab-initio techniques are prohibited by the size of metabolites tested here. Instead, to compare QCxMS-based ground state simulations to the new excited state method, we randomly selected 48 molecules from our previously published study[12]. We used both weighted dot-score similarity and the Jaccard index that calculated the number of true positive predicted ions divided by the number of all ions observed in combination of the in-silico and experimental spectra. For 48 calculated molecules, we found an improvement from an average dot product similarity score of 681 in the OM2 method to 724 for GFN1-xTB method and 726 using the mixed method, adding excited states to supplemented ground state simulations (Supplement S8). We have recently introduced the concept of entropy similarity scores that outperformed dot-product similarities and improved False Discovery Rates[54]. Comparing spectral entropy similarities indeed showed a significant improvement for the mixed method: the two ground state methods resulted in an average entropy similarity score of 600 whereas the mixed method gave an average score of 680 and a narrower Kernel density distribution (Figure 4-5). The OM2 method showed a much lower average Jaccard index of 0.34 than the default GFN1-xTB or the mixed method. While the average Jaccard index of these two methods were not statistically significant different, the mixed method clearly showed

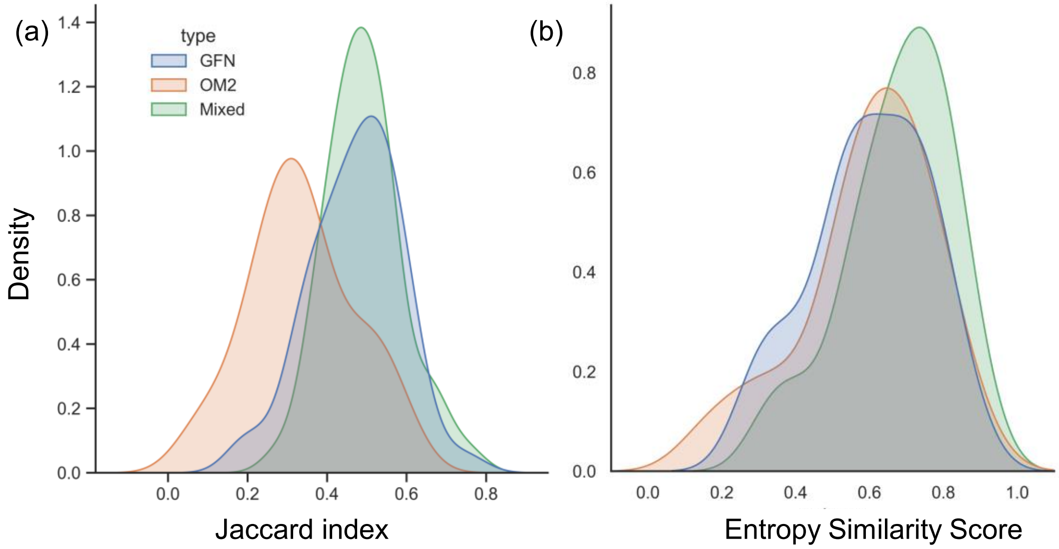


Figure 4-5. Kernel density estimate plot of (a) Jaccard index and (b) entropy similarity score of 48 small molecules; Blue: ground state spectra predicted by GFN1-xTB; Orange: ground state spectra simulated by the OM2 method; Green: mixed spectra of D_0 and D_1 using the GFN1-xTB/OM2/CISD method

less variance and a higher maximum density of the Jaccard indices of the modeled in-silico spectra in comparison to the GFN1-xTB method. On average, the mixed method yielded 16% and 73% more true positive ions in pairwise comparisons than the GFN1-xTB and OM2 ground state method, respectively. We found that the mixed method avoided extremely poor simulations that were observed for the OM2 methods with dot score < 200 with a true positive rate < 0.4 (File S13).

4.5. Conclusions

We here show for the first time that molecular dynamics can be utilized with excited state calculations by using the BEB model to scale contributions based on ionization cross sections. We provided and tested this excited state correction method for quantum chemistry molecular dynamics prediction of standard 70 eV mass spectra and showed that it improved the existing GFNn-xTB method by generating about 16% more correctly predicted fragment ions. For example, the mixed method presented here added hydrogen shift reactions that were missed by the classic methods. When comparing this mixed method with other tools such as DFT and semi-empirical methods like OM2, we found clear improvements in accuracy that came with only 20% increased computational times. Although our OM2/CISD excited state simulations discovered more fragmentation reactions than the standard model, improvement in spectra similarities to experimental spectra were limited because the GFNn-xTB method generally already yielded a high number of excellent predictions, as shown in detail for the molecule urocanic acid. However, predicting rearrangement reactions are the bottleneck of QCxMS. Because much longer simulation times may be prohibitively expensive with respect to computational costs, we propose that machine learning methods are needed to recognize rearrangement reactions. We recommend this mixed model to be used to generate in-silico electron ionization mass spectral libraries for small molecules.

4.6. Reference

1. Grimme, S., Towards First Principles Calculation of Electron Impact Mass Spectra of Molecules. *Angewandte Chemie International Edition* **2013,** *52* (24), 6306-6312.

2. Borges, R. M.; Colby, S. M.; Das, S.; Edison, A. S.; Fiehn, O.; Kind, T.; Lee, J.; Merrill, A. T.; Merz, K. M.; Metz, T. O.; Nunez, J. R.; Tantillo, D. J.; Wang, L.-P.; Wang, S.; Renslow, R. S., Quantum Chemistry Calculations for Metabolomics. *Chemical Reviews* **2021,** *121* (10), 5633-5670.

3. Ásgeirsson, V.; Bauer, C. A.; Grimme, S., Quantum chemical calculation of electron ionization mass spectra for general organic and inorganic molecules. *Chemical Science* **2017,** *8* (7), 4879-4895.

4. Koopman, J.; Grimme, S., Calculation of Electron Ionization Mass Spectra with Semiempirical GFNn-xTB Methods. *ACS Omega* **2019,** *4* (12), 15120-15133.

5. Jeroen Koopman, C. B., Vilhjalmur Ásgeirsson, Stefan Grimme QCEIMS 4.0 Installation Guide & Manual.

6. Bauer, C. A.; Grimme, S., How to Compute Electron Ionization Mass Spectra from First Principles. *The Journal of Physical Chemistry A* **2016,** *120* (21), 3755-3766.

7. Bauer, C. A.; Grimme, S., First principles calculation of electron ionization mass spectra for selected organic drug molecules. *Organic & Biomolecular Chemistry* **2014,** *12* (43), 8737-8744.

8. Bauer, C. A.; Grimme, S., Elucidation of Electron Ionization Induced Fragmentations of Adenine by Semiempirical and Density Functional Molecular Dynamics. *The Journal of Physical Chemistry A* **2014,** *118* (49), 11479-11484.

9. Bauer, C. A.; Grimme, S., Automated Quantum Chemistry Based Molecular Dynamics Simulations of Electron Ionization Induced Fragmentations of the Nucleobases Uracil, Thymine, Cytosine, and Guanine. *European Journal of Mass Spectrometry* **2015,** *21* (3), 125-140.

10. Koopman, J.; Grimme, S., From QCEIMS to QCxMS: A Tool to Routinely Calculate CID Mass Spectra Using Molecular Dynamics. *Journal of the American Society for Mass Spectrometry* **2021,** *32*.

11. Koopman, J.; Grimme, S., *Calculation of mass spectra with the QCxMS method for negatively and multiply charged molecules*. 2022.

12. Wang, S.; Kind, T.; Tantillo, D. J.; Fiehn, O., Predicting in silico electron ionization mass spectra using quantum chemistry. *Journal of Cheminformatics* **2020,** *12* (1), 63.

13. Townsend, J.; Kirkland, J. K.; Vogiatzis, K. D., Chapter 3 - Post-Hartree-Fock methods: configuration interaction, many-body perturbation theory, coupled-cluster theory. In *Mathematical Physics in Theoretical Chemistry*, Blinder, S. M.; House, J. E., Eds. Elsevier: 2019; pp 63-117.

14. Lorquet, J. C., Basic questions in mass spectrometry. *Organic Mass Spectrometry* **1981,** *16* (11), 469-482.

15. Lorquet, J. C., Landmarks in the theory of mass spectra. *International Journal of Mass Spectrometry* **2000,** *200* (1), 43-56.

16. Mark, T. D., Fundamental aspects of electron impact ionization. *International Journal of Mass Spectrometry and Ion Physics* **1982,** *45*, 125-145.

17. Kim, Y.-K.; Rudd, M. E., Binary-encounter-dipole model for electron-impact ionization. *Physical Review A* **1994,** *50* (5), 3954-3967.

18. Gross, J. H., Principles of Ionization and Ion Dissociation. In *Mass Spectrometry: A Textbook*, Gross, J. H., Ed. Springer Berlin Heidelberg: Berlin, Heidelberg, 2011; pp 21-66.

19. Franck, J.; Dymond, E. G., Elementary processes of photochemical reactions. *Transactions of the Faraday Society* **1926,** *21* (February), 536-542.

20. Condon, E., A Theory of Intensity Distribution in Band Systems. *Physical Review* **1926,** *28* (6), 1182-1201.

21. Bente, P. F.; McLafferty, F. W.; McAdoo, D. J.; Lifshitz, C., Internal energy of product ions formed in mass spectral reactions. Degrees of freedom effect. *The Journal of Physical Chemistry* **1975,** *79* (7), 713-721.

22. Nelson, T. R.; White, A. J.; Bjorgaard, J. A.; Sifain, A. E.; Zhang, Y.; Nebgen, B.; Fernandez-Alberti, S.; Mozyrsky, D.; Roitberg, A. E.; Tretiak, S., Non-adiabatic Excited-State Molecular Dynamics: Theory and Applications for Modeling Photophysics in Extended Molecular Materials. *Chemical Reviews* **2020,** *120* (4), 2215-2287.

23. Dral, P. O.; Wu, X.; Spörkel, L.; Koslowski, A.; Weber, W.; Steiger, R.; Scholten, M.; Thiel, W., Semiempirical Quantum-Chemical Orthogonalization-Corrected Methods: Theory, Implementation, and Parameters. *J Chem Theory Comput* **2016,** *12* (3), 1082-1096.

24. Gerber, R. B.; Shemesh, D.; Varner, M. E.; Kalinowski, J.; Hirshberg, B., Ab initio and semi-empirical Molecular Dynamics simulations of chemical reactions in isolated molecules and in clusters. *Physical Chemistry Chemical Physics* **2014,** *16* (21), 9760-9775.

25. Koslowski, A.; Beck, M. E.; Thiel, W., Implementation of a general multireference configuration interaction procedure with analytic gradients in a semiempirical context using the graphical unitary group approach. *Journal of Computational Chemistry* **2003,** *24* (6), 714-726.

26. Tuna, D.; Lu, Y.; Koslowski, A.; Thiel, W., Semiempirical Quantum-Chemical Orthogonalization-Corrected Methods: Benchmarks of Electronically Excited States. *J Chem Theory Comput* **2016,** *12* (9), 4400-4422.

27. Silva-Junior, M. R.; Thiel, W., Benchmark of Electronically Excited States for Semiempirical Methods: MNDO, AM1, PM3, OM1, OM2, OM3, INDO/S, and INDO/S2. *J Chem Theory Comput* **2010,** *6 5*, 1546-64.

28. Liu, J.; Thiel, W., An efficient implementation of semiempirical quantum-chemical orthogonalization-corrected methods for excited-state dynamics. *The Journal of Chemical Physics* **2018,** *148* (15), 154103.

29. Foresman, J. B.; Head-Gordon, M.; Pople, J. A.; Frisch, M. J., Toward a systematic molecular orbital theory for excited states. *The Journal of Physical Chemistry* **1992,** *96* (1), 135-149.

30. Możejko, P.; Sanche, L., Cross section calculations for electron scattering from DNA and RNA bases. *Radiation and Environmental Biophysics* **2003,** *42* (3), 201-211.

31. Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery Jr, J. A., General atomic and molecular electronic structure system. *Journal of Computational Chemistry* **1993,** *14* (11), 1347-1363.

32. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16 Rev. C.01*, Wallingford, CT, 2016.

33. Hanwell, M. D.; Curtis, D. E.; Lonie, D. C.; Vandermeersch, T.; Zurek, E.; Hutchison, G. R., Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *Journal of Cheminformatics* **2012,** *4* (1), 17.

34. Bannwarth, C.; Ehlert, S.; Grimme, S., GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J Chem Theory Comput* **2019,** *15* (3), 1652-1671.

35. Dewar, M. J. S.; Thiel, W., Ground states of molecules. 38. The MNDO method. Approximations and parameters. *Journal of the American Chemical Society* **1977,** *99* (15), 4899-4907.

36. Thiel, W., Semiempirical quantum–chemical methods. *WIREs Computational Molecular Science* **2014,** *4* (2), 145-157.

37. Tuna, D.; Spörkel, L.; Barbatti, M.; Thiel, W., Nonadiabatic dynamics simulations of photoexcited urocanic acid. *Chemical Physics* **2018,** *515*, 521-534.

38. Bauer, C. A.; Hansen, A.; Grimme, S., The Fractional Occupation Number Weighted Density as a Versatile Analysis Tool for Molecules with a Complicated Electronic Structure. *Chemistry – A European Journal* **2017,** *23* (25), 6150-6164.

39. Koltai, T.; Reshkin, S. J.; Harguindey, S., Chapter 15 - Pharmacological interventions part III. In *An Innovative Approach to Understanding and Treating Cancer: Targeting pH*, Koltai, T.; Reshkin, S. J.; Harguindey, S., Eds. Academic Press: 2020; pp 335-359.

40. Perdew, J. P.; Ernzerhof, M.; Burke, K., Rationale for mixing exact exchange with density functional approximations. *The Journal of Chemical Physics* **1996,** *105* (22), 9982-9985.

41. Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H., A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *The Journal of Chemical Physics* **2010,** *132* (15), 154104.

42. Schäfer, A.; Horn, H.; Ahlrichs, R., Fully optimized contracted Gaussian basis sets for atoms Li to Kr. *The Journal of Chemical Physics* **1992,** *97* (4), 2571-2577.

43. Center, N. M. S. D. *Mass Spectrum Interpreter* 3.4; 2019.

44. Alifano, P.; Fani, R.; Liò, P.; Lazcano, A.; Bazzicalupo, M.; Carlomagno, M. S.; Bruni, C. B., Histidine biosynthetic pathway and genes: structure, regulation, and evolution. *Microbiol Rev* **1996,** *60* (1), 44-69.

45. Cautereels, J.; Blockhuys, F., Quantum Chemical Mass Spectrometry: Verification and Extension of the Mobile Proton Model for Histidine. *Journal of The American Society for Mass Spectrometry* **2017,** *28* (6), 1227-1235.

46. Tsugawa, H.; Kind, T.; Nakabayashi, R.; Yukihira, D.; Tanaka, W.; Cajka, T.; Saito, K.; Fiehn, O.; Arita, M., Hydrogen rearrangement rules: computational MS/MS fragmentation and structure elucidation using MS-FINDER software. *Analytical chemistry* **2016,** *88* (16), 7946-7958.

47. Demarque, D. P.; Crotti, A. E. M.; Vessecchi, R.; Lopes, J. L. C.; Lopes, N. P., Fragmentation reactions using electrospray ionization mass spectrometry: an important tool for the structural elucidation and characterization of synthetic and natural products. *Natural Product Reports* **2016,** *33* (3), 432-455.

48. Holmes, J. L.; Terlouw, J. K., The scope of metastable peak observations. *Organic Mass Spectrometry* **1980,** *15* (8), 383-396.

49. Williams, D. H., A transition-state probe. *Accounts of Chemical Research* **1977,** *10* (8), 280-286.

50. Wollnik, H., Time-of-flight mass analyzers. *Mass Spectrometry Reviews* **1993,** *12* (2), 89-114.

51. Holmes, J. L., Assigning structures to ions in the gas phase. *Organic Mass Spectrometry* **1985,** *20* (3), 169-183.

52. Hesse, M., Mass Spectrometry. A Textbook. By Jürgen H. Gross. *Angewandte Chemie International Edition* **2004,** *43* (35), 4552-4552.

53. Semialjac, M.; Schröder, D.; Schwarz, H., Car–Parrinello Molecular Dynamics Study of the Rearrangement of the Valeramide Radical Cation. *Chemistry – A European Journal* **2003,** *9* (18), 4396-4404.

54. Li, Y.; Kind, T.; Folz, J.; Vaniya, A.; Mehta, S. S.; Fiehn, O., Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification. *Nat Methods* **2021,** *18* (12), 1524-1531.

4.7. Supporting Information

Table 4-S1. Molecular orbital energies (U, kinetic energy; B, electron binding energy) in eV and computed ionization cross section (σ) at 70 eV

MO	**Energy**	**U**	**B**	**σ@70eV**
HOMO	-6.84	33.90	8.52	2.29
HOMO-1	-8.92	39.84	11.36	1.39
HOMO-2	-9.45	34.25	12.07	1.32
HOMO-3	-9.55	51.61	12.23	1.12
HOMO-4	-9.66	63.49	12.52	0.99
HOMO-5	-10.68	58.01	13.35	0.93
HOMO-6	-12.42	43.59	15.30	0.81
HOMO-7	-13.84	53.54	15.83	0.71
HOMO-8	-13.96	38.98	16.20	0.76
HOMO-9	-14.41	48.56	17.10	0.64

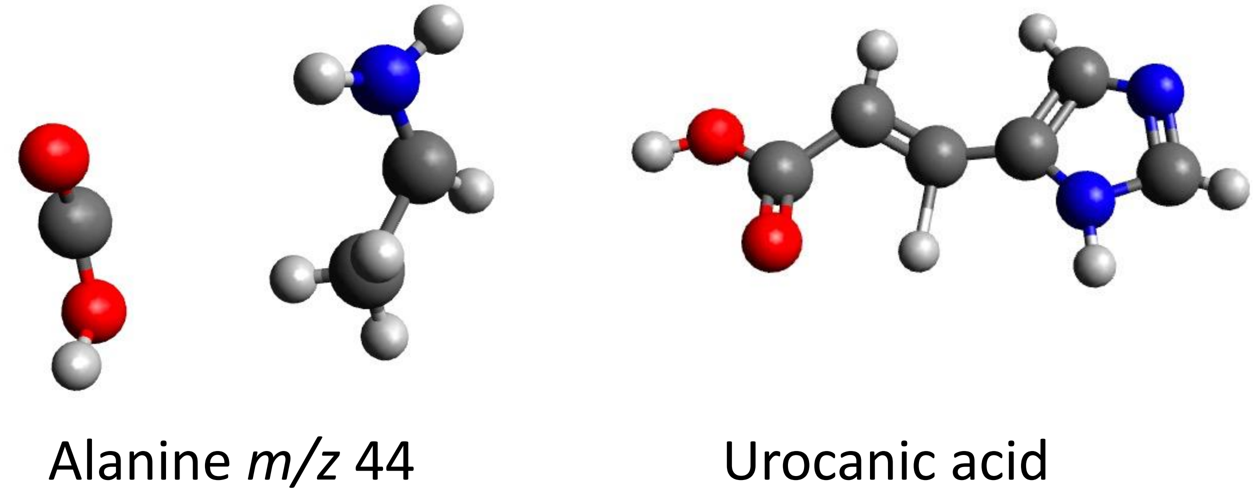


Figure 4-S1. Molecular dynamics based on vertical ionization model without excess energy (left)alanine, generating base peak *m/z* 44 (right) urocanic acid, no reaction within 2985 fs, generating the molecular ion peak

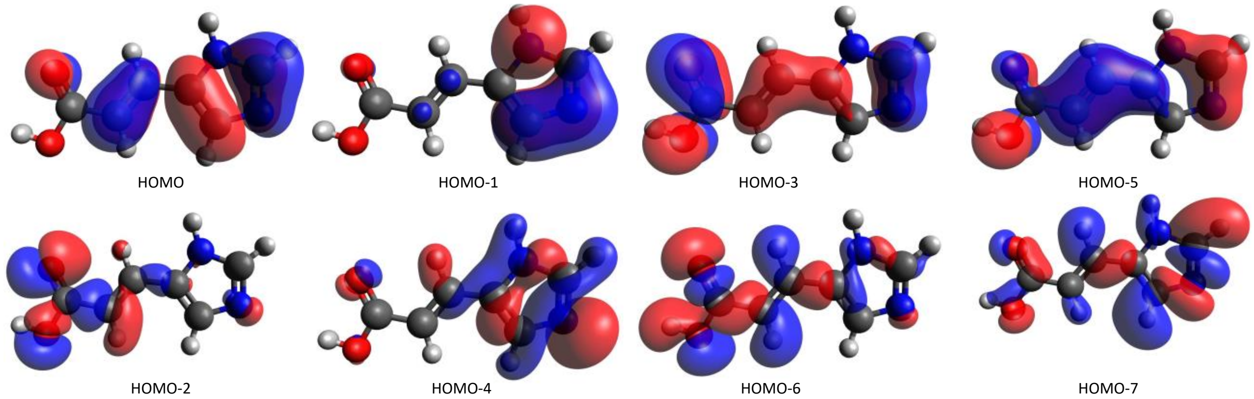


Figure 4-S2. Molecular orbitals under HF/STO-3G level: the HOMO, HOMO-1, HOMO-3, and HOMO-5 orbitals are π orbitals; the HOMO-2, HOMO-4, HOMO-6, HOMO-7 are n orbitals associated with lone pairs on oxygen and nitrogen

4.7.1. Effect of active space

To better understand the results of our excited state MD simulations, we also tested the effect of active space size. **Figure S2** shows the D_1 state spectra with different active space settings. Compared with the D_0 state spectrum, all D_1 state spectra had a very small molecular ion peak *m/z* 138. This phenomenon originates from the relative high energy of excited state structures which leads to easy fragmentation. The intensity of the molecular ion peak *m/z* 138 becomes larger, however, with the increase of active space size. A larger active space can describe the electronic structure better and lower the electronic energy. The *m/z* 67 peak is only abundant under (3,4) simulation, with 16 trajectories; with a (19,16) active space, only five trajectories lead to the same ion. Ring-opening/rearrangement of the imidazole ring gave the *m/z* 67 as $C_3H_3N_2^+$. While with active space (3,4), the orbital used in the calculation cannot include the π orbital and cause two problems: the whole molecule broke into many pieces at the very beginning of the simulation to give chain shape $C_3H_3N_2^+$; the $C_3H_3N_2^+$ undergoes a fast ring-opening reaction once generated. These two problems cause the overestimation of *m/z* 67 under this relatively small active space. The failure of (3,4) active space implies that the choose of active space is important to get the potential energy surface correctly. Overall, a larger active space can make the simulation more accurate, and we need to find a balance between active space and computational time.

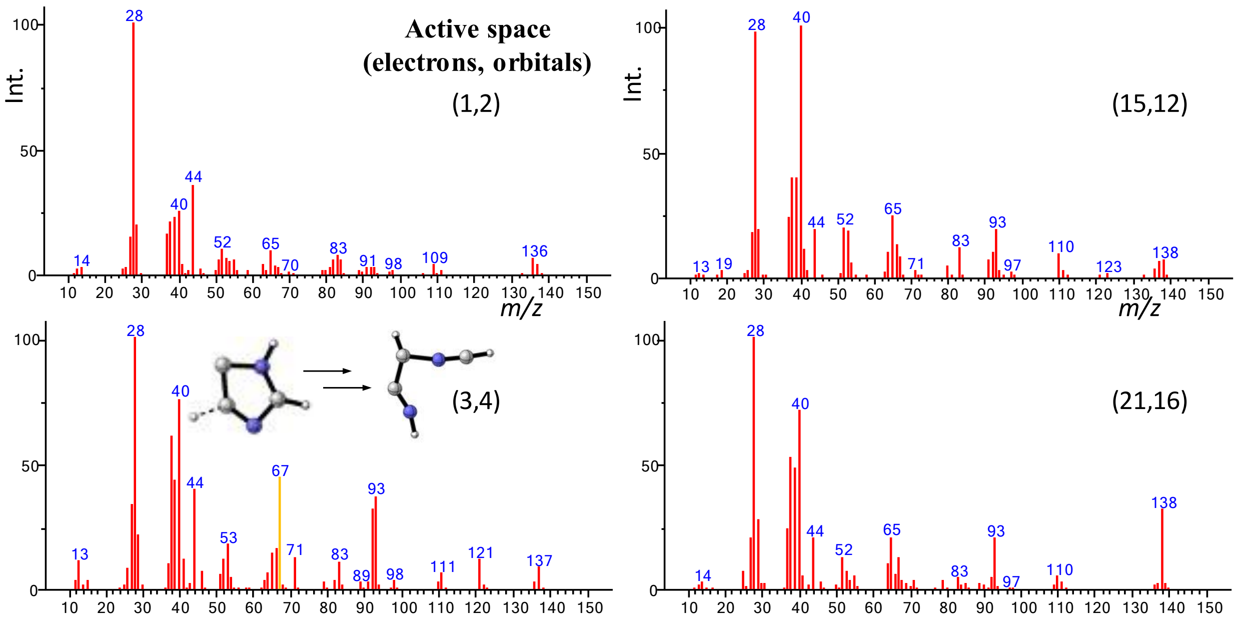


Figure 4-S3. D_1 First excited state spectra of urocanic acid with different active space sizes; Orange line: *m/z* 67 of fragment $C_3H_3N_2^+$, only observed strongly in active space (3, 4); Inside: ring break reaction taken from trajectory 54

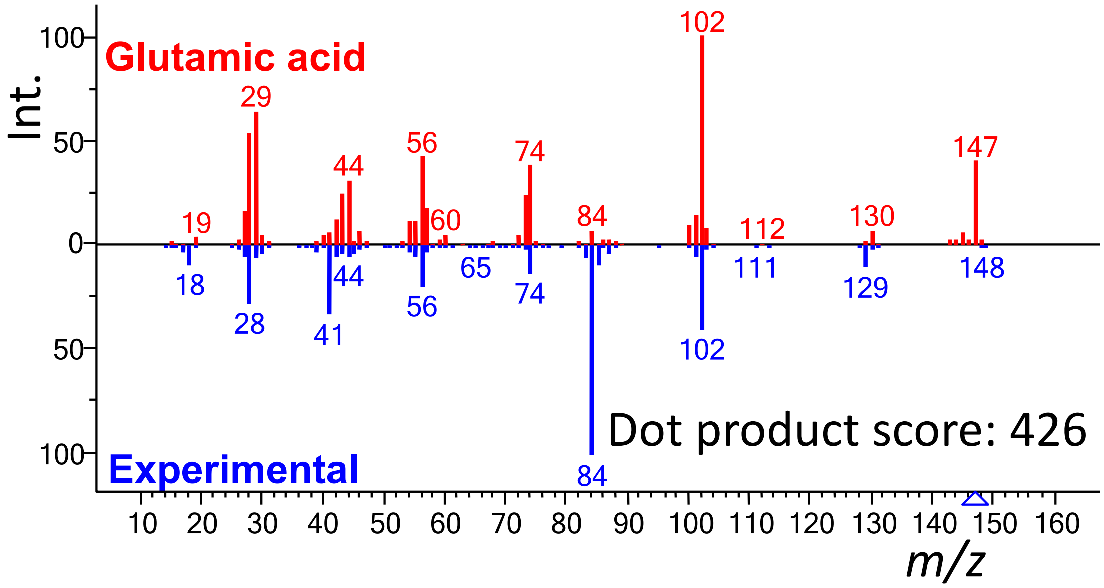


Figure 4-S4. Head-to-tail plot; Hybrid spectrum of glutamic acid against experimental spectrum from NIST17

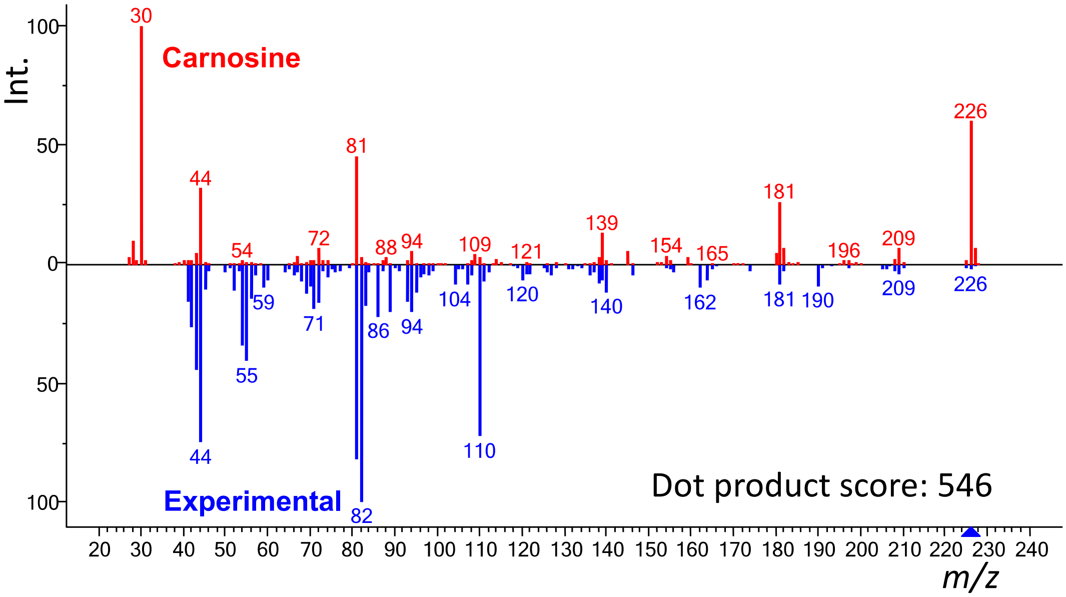


Figure 4-S5. Head-to-tail plot; Hybrid spectrum of carnosine against experimental spectrum from NIST17

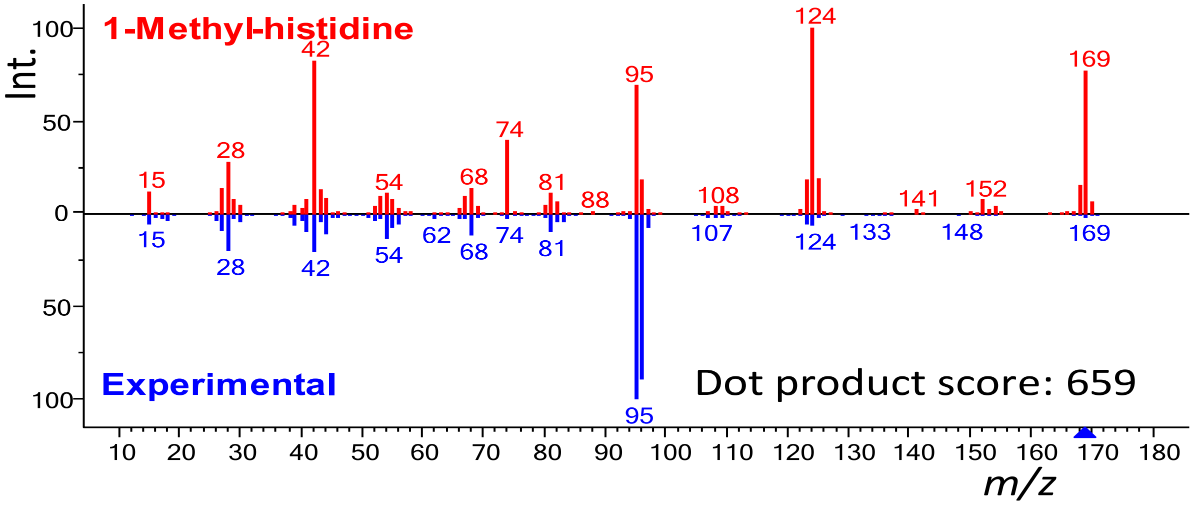


Figure 4-S6. Head-to-tail plot; Hybrid spectrum of 1-methyl-istidine against experimental spectrum from NIST17

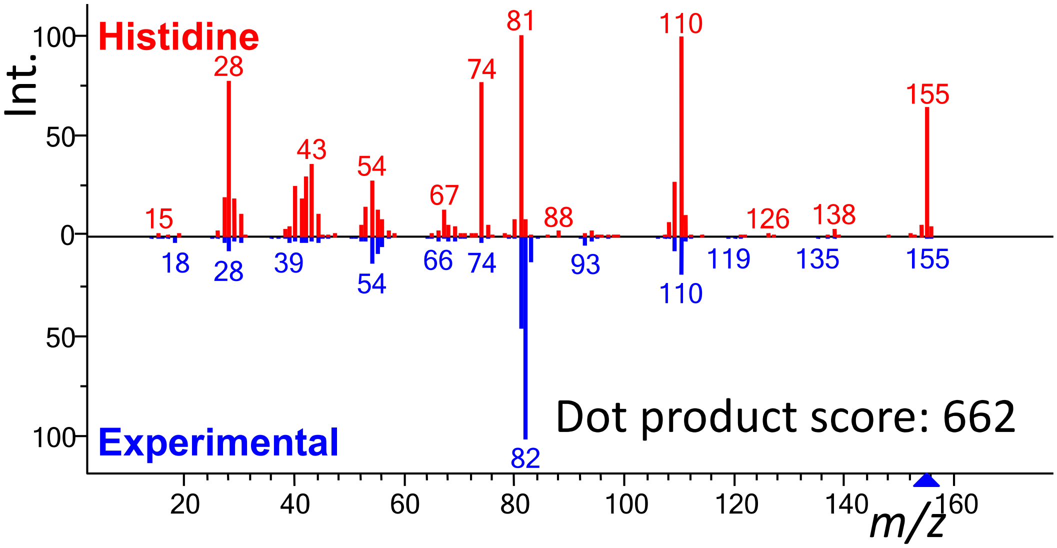
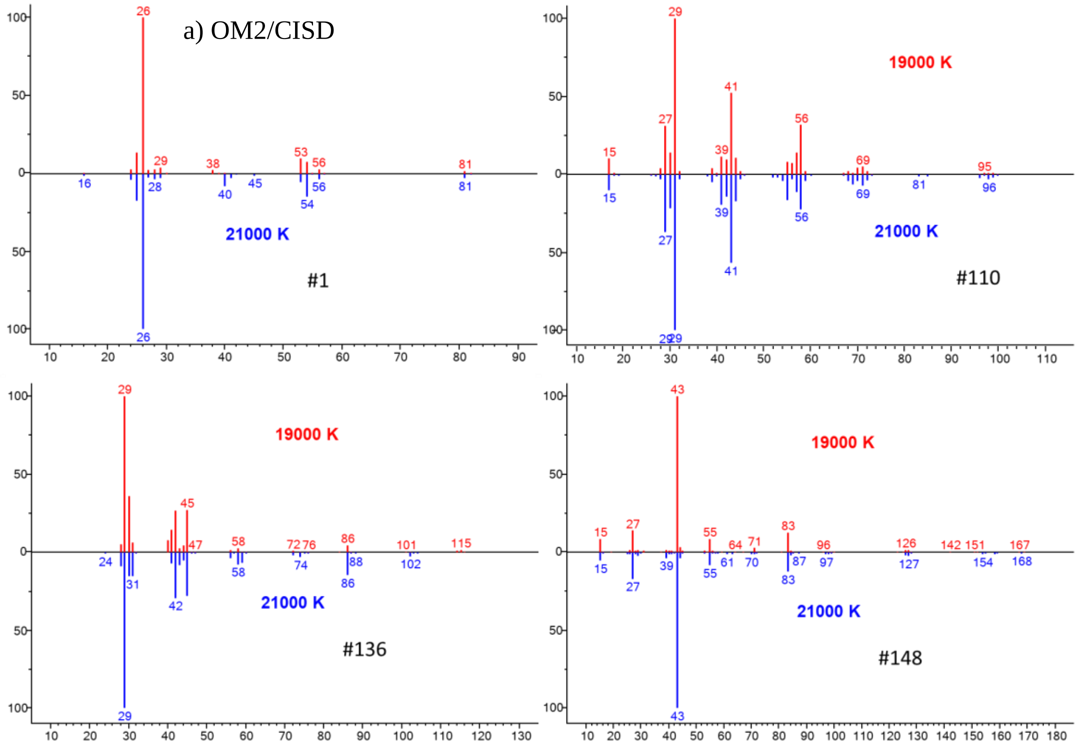


Figure 4-S7. Head-to-tail plot; Hybrid spectrum of histidine against experimental spectrum from NIST17



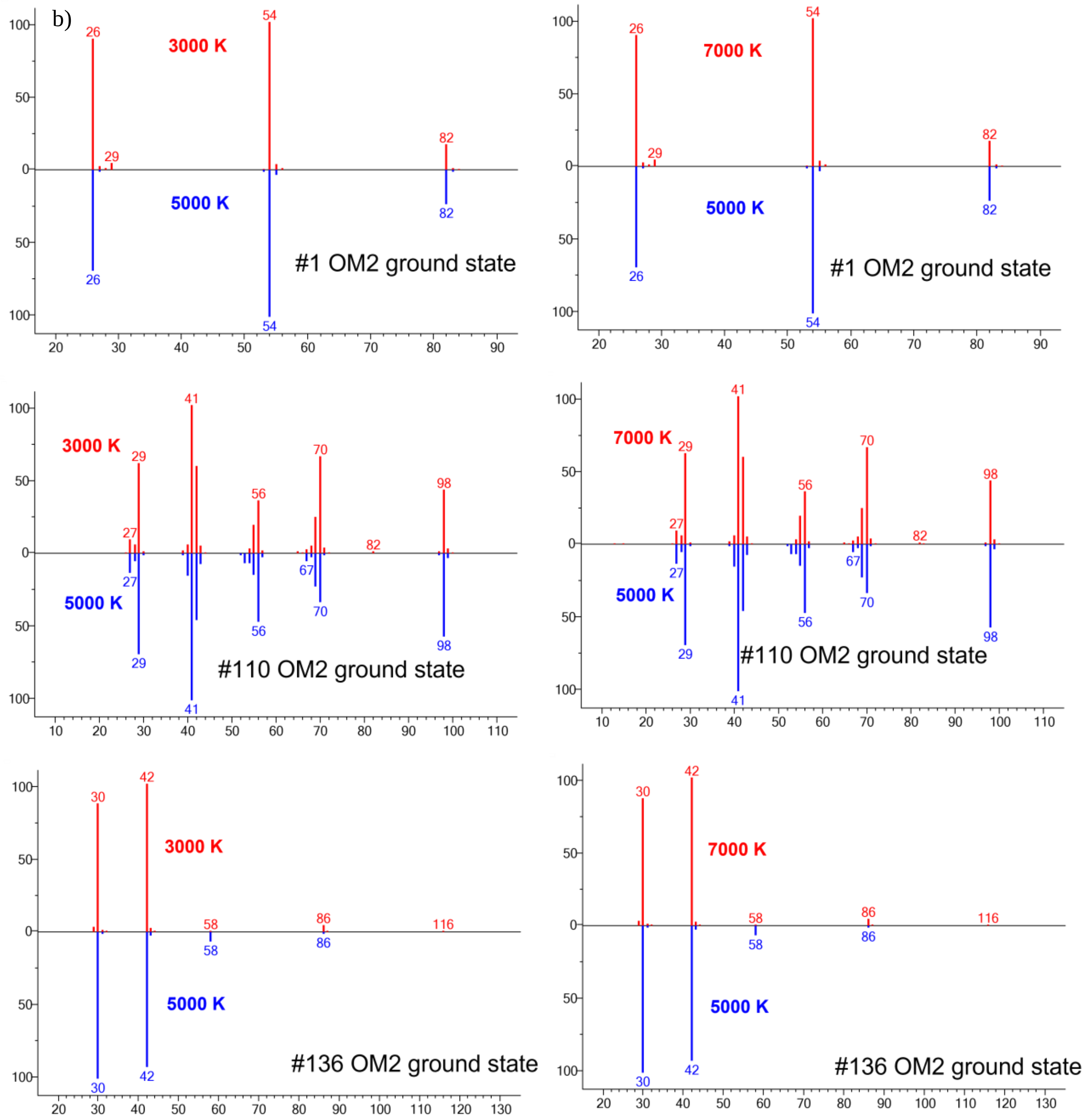

b)
3000 K
5000 K
#1 OM2 ground state
26
29
54
82
26
54
82
7000 K
5000 K
#1 OM2 ground state
26
29
54
82
26
54
82
3000 K
5000 K
#110 OM2 ground state
27
29
41
56
70
82
98
27
29
41
56
67
70
98
7000 K
5000 K
#110 OM2 ground state
27
29
41
56
70
82
98
27
29
41
56
67
70
98
3000 K
5000 K
#136 OM2 ground state
30
42
58
86
116
30
42
58
86
7000 K
5000 K
#136 OM2 ground state
30
42
58
86
116
30
42
58
86
100
50
0
50
100
10
20
30
40
50
60
70
80
90
100
110
120
130

Figure 4-S8. Head-to-tail plot; a) OM2/CISD method with different electronic temperature. b) OM2 ground state method with different electronic temperature; Different temperatures didn’t affect the results; more examples were uploaded to the Zenodo repository.

4.7.2. Parameter settings

We also noticed that spectra simulated for 3000 K were almost identical to spectra predicted for 7000 K with ion abundance changes remaining at <1%. However, when different ip-mndo methods are used, such as ip-orca applied for the 5000 K simulations, results will be different. This observation showed that the ionization potential distribution can affect the simulation but the electronic temperature in OM2 method does not impact predictions.

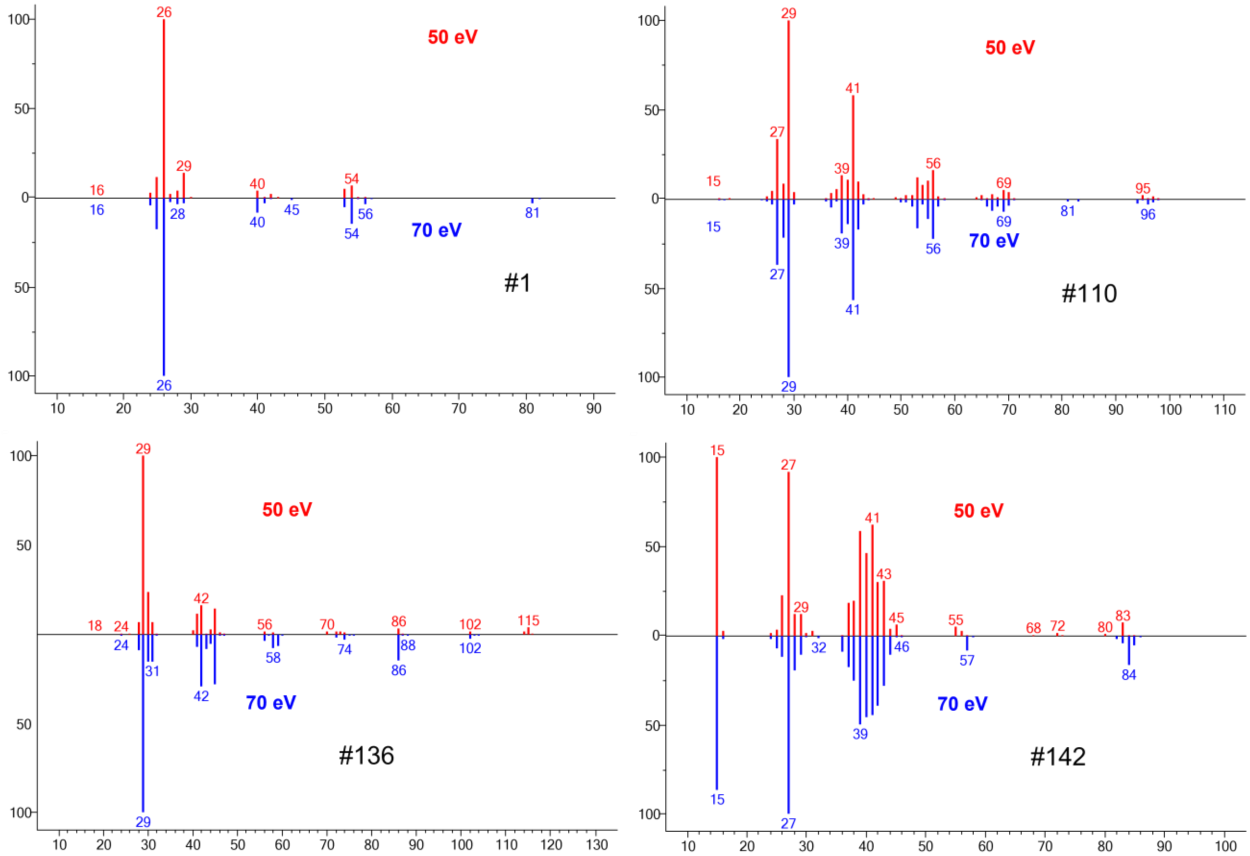


Figure 4-S9. Head-to-tail plot; OM2/CISD method with different impact energy.

Lower impact energy did not change the excited-state spectra noticeably for #148 and #110. For #136, #1 and #142, there was no trend while lower energy or 70 eV avoided some impossible neutral loss reactions. Overall, lowering the ionization energy did not improve the fragmentation predictions; more examples were uploaded to the Zenodo repository.

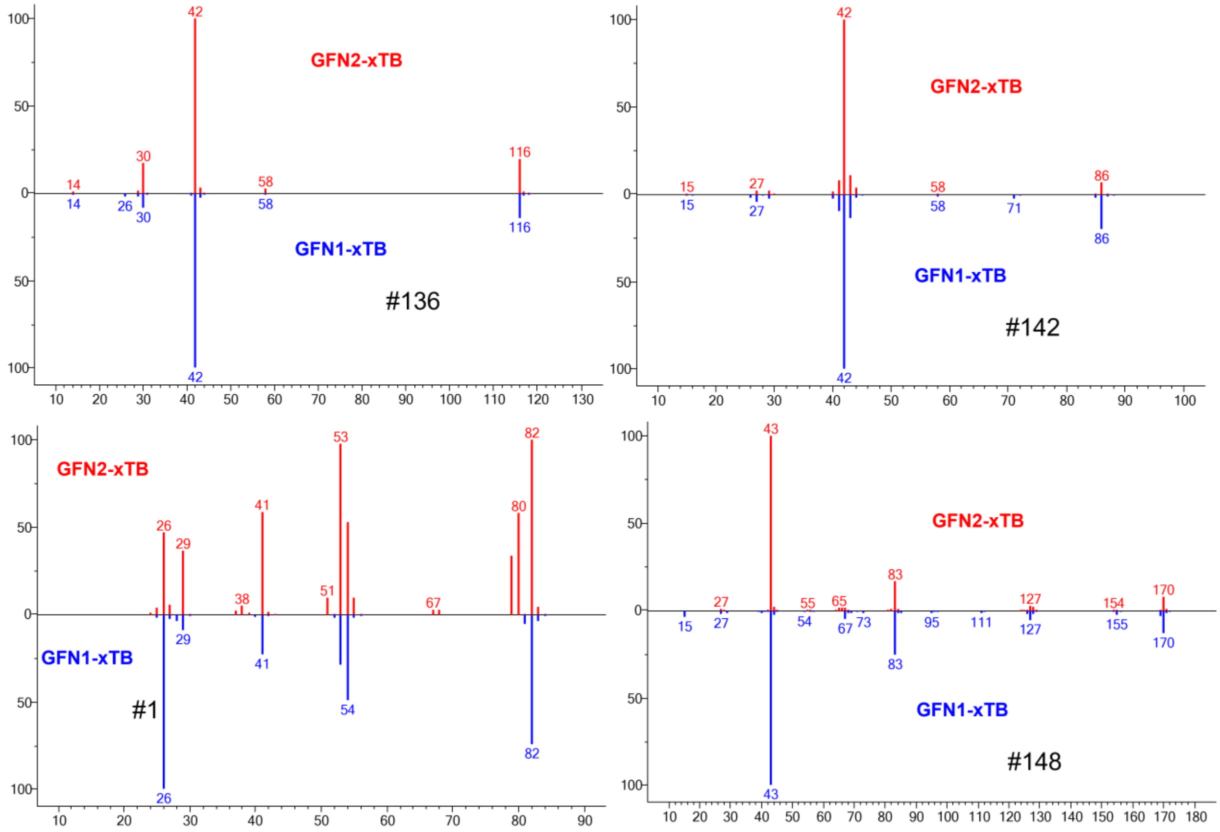


Figure 4-S10. Head-to-tail plot; GFN2-xTB and GFN1-xTB. Different temperatures didn’t affect the results; more examples were uploaded to the Zenodo repository.

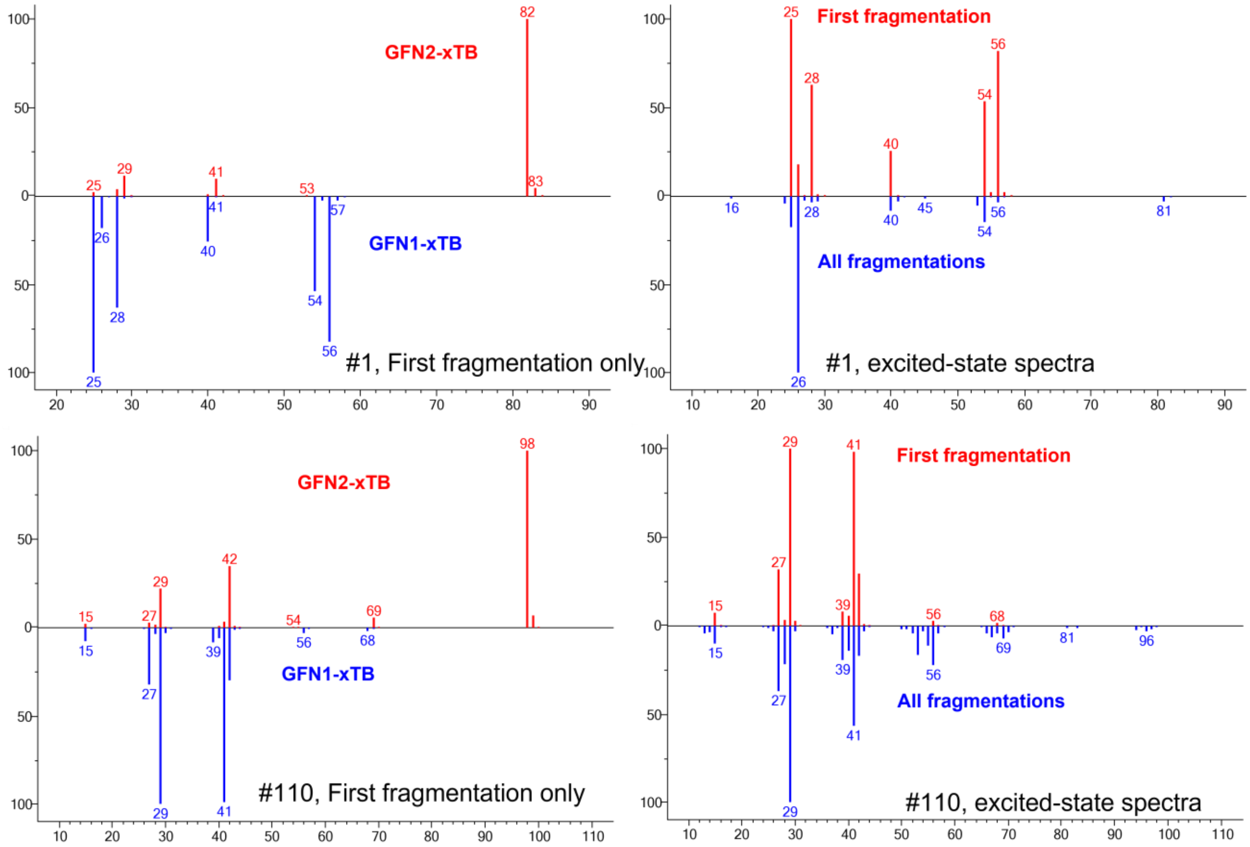


Figure 4-S11. Head-to-tail plot; comparison of first fragmentations. Different temperatures didn't affect the results; more examples including spectra without isotopic pattern and secondary fragmentations were uploaded to the Zenodo repository.

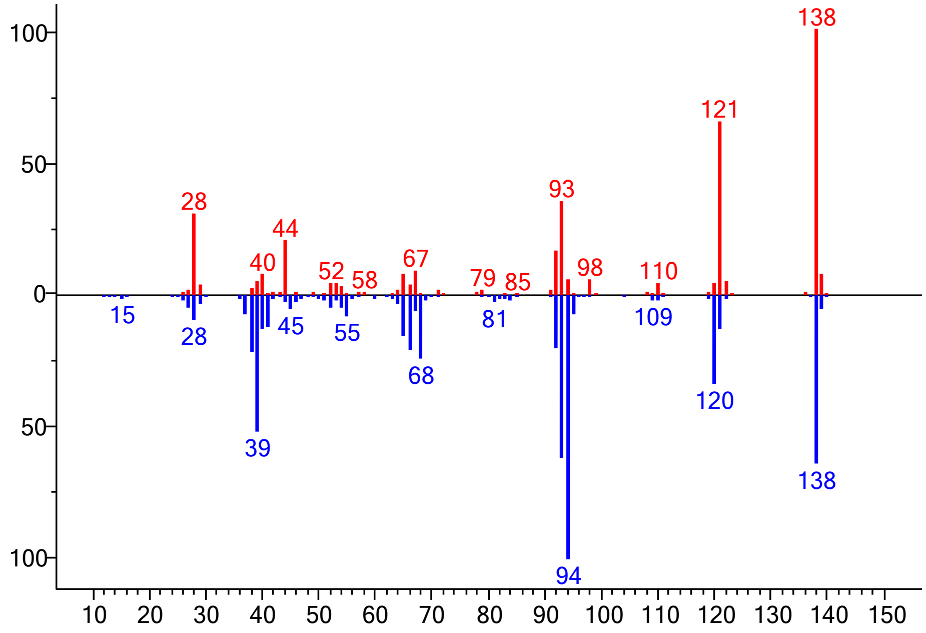


Figure 4-S12. PBE0/SV(P) spectra of Urocanic Acid without isotope ions

Chapter 5: Gas chromatography with methane chemical ionization and quadrupole-time of flight mass spectrometry obtains molecular ion species to automatically assign elemental formulas

5.1. Abstract

Gas chromatography–mass spectrometry (GC-MS) usually employs hard electron ionization, leading to extensive fragmentations that are perfectly suitable to identify compounds based on library matches. However, such spectra are less useful to structurally characterize unknown compounds that are absent from libraries, due to the lack of readily recognizable molecular ion species. We here tested methane chemical ionization on 367 trimethylsilylated (TMS) derivatized metabolites using an accurate mass quadrupole time-of-flight detector (QTOF) to determine if this approach could be used to automatically detect molecular ion species and how accurate the determination of molecular formulas from these spectra would be. For most compounds, we found a clear pattern of molecular ion adduct or fragment species. Overall, the automatic workflow correctly recognized 316 (86%) of all detected, derivatized standards. $[M-H]^+$, $[M]^+$ and $[M+H]^+$ ion species were present in all cases, but varied in intensities. Specifically, strong $[M-CH_3]^+$ fragments were observed in all 290 derivatized metabolites that were automatically recognized by , substantiated by concomitant $[M + C_2H_5]^+$ adducts in 90% of the detected species, and $[M + C_3H_5]^+$ in 84% of the cases. Together, these species formed a pattern that could be extracted automatically from GC-QTOF MS spectra. Using Sirius software, correct elemental formulas for $[M-CH_3]^+$ fragment ions were retrieved in 87% cases within the top-3 hits. In 71% of the cases,

the top-hit was found as the correct formula when hits for $[M\text{-}CH_3]^+$ and $[M]^+$ and $[M+H]^+$ molecular ion species were combined. When investigating the 51 cases for which the automatic pattern analysis failed, we found that several analytes showed a previously unknown adduct $[M + TMS]^+$, formed by rearrangement. Overall, we here demonstrate methane chemical ionization with GC-QTOF mass spectrometry as suitable avenue to identify molecular formulas for abundant unknown peaks.

5.2. Introduction

Gas chromatography–mass spectrometry (GC-MS) is a mature technology for small metabolites profiling, because of its wide coverage of chemical classes[1] and high reproducibility[2] using standardized 70 eV electron impact ionization (EI)[3]. Lower energies lead to less sensitivity and fragmentation[4-5]. Large mass spectral libraries, such as NIST EI library[6], Massbank of North America[7], and Human Metabolome Database[8]. Yet, because the EI is a hard ionization leading to the strongest ionization and fragmentation, molecular ions or readily identifiable adducts are usually low abundant or absent, especially when using trimethylsilylation (TMS).

Without knowing molecular masses of unknown metabolites, structural identifications are impossible. Alternatively, chemical ionization[3] (CI) is a softer technique than electron impact. It usually obtains molecular masses and has been successfully used for compound identifications[9]. In chemical ionization, the reagent gas molecules (usually methane, ammonia or isobutane) are first ionized and then the reagent ions ionizes neutral analyte molecules with less energy transferred due to the exothermicity of ion-molecule reactions[3]. In CI, the molecular ion has higher probability to keep entire with significantly less fragmentations than 70 eV EI.

The following complex reactions have been described in the literature to produce typical molecular ion adducts in methane CI[10-11]:

$$\mathrm{M} + \mathrm{CH}_4^+ \rightarrow \mathrm{CH}_4 + \mathrm{M}^+$$

$$\mathrm{M} + \mathrm{CH}_5^+ \rightarrow \mathrm{CH}_4 + [\mathrm{M}+\mathrm{H}]^+$$

$$\mathrm{M} + \mathrm{CH}_3^+ \rightarrow \mathrm{CH}_4 + [\mathrm{M}-\mathrm{H}]^+$$

$$\mathrm{M} + \mathrm{C}_2\mathrm{H}_5^+ \rightarrow [\mathrm{M}+\mathrm{C}_2\mathrm{H}_5]^+$$

$$\mathrm{M} + \mathrm{CH}_2^+ + 2\,\mathrm{CH}_4 \rightarrow [\mathrm{M}+\mathrm{C}_3\mathrm{H}_5]^+ + \mathrm{H}_2 + \mathrm{H}^{\cdot}$$

The formation of this series of predictable adducts can assist in automatically assigning the molecular ions in GC-chemical ionization QTOF MS. By combining GC-electron ionization MS for profiling samples and using GC-CI-QTOF MS for identifying unknown compounds, we can keep the advantage of informative spectra from EI and of molecular ion integrity from CI. In this paper, we explore the feasibility of using automatic pattern analysis for recognizing molecular ion species in GC-CI-QTOF MS and then using that information to obtain elemental formulas. We performed these analyses on a large range of metabolites under trimethylsilylation conditions, as used in untargeted GC-MS metabolomics studies.

5.3. Experimental methods

5.3.1. Data acquisition

To build a GC-CI-QTOF mass spectral test library, 1 mg of each metabolite standard was dissolved in a 1 ml methanol:water:isopropanol (5:2:2) solution. 20 µl of each standard was combined into mixtures of 20 non-isomeric compounds to minimize data acquisition time. Mixtures were evaporated to dryness and derivatized by methoximation and trimethylsilylation as published previously (ref), using O-methylhydroxylamine hydrochloride solution (Sigma-Aldrich) in pyridine and N-methyl-N-trimethylsilyltrifluoroacetamide (MSTFA; Sigma-Aldrich). Retention index markers of C8-C30 linear chain fatty acid methyl esters (FAME markers) were added to the MSTFA. 100 ul samples were transferred to autosampler vials and 1 ul was injected at 25 s splitless time (Table 1).

Table 5-1. Details of data acquisition parameters for the FiehnLib GC/MS libraries

Gas Chromatograph	Agilent 7890A GC system
Mass Spectrometer	7200 accurate mass QTOF mass spectrometer
GC column	DB5 MS column 30 m + 10 m integrated guard, 0.25 mm id, 0.25 μm film
GC parameters, injection	1 μL in 25 s splitless mode at 250 °C
GC parameters, separation	initial temperature of 60 °C with a hold time of 0.5 min, a temperature ramp of 10 °C/min to 325 °C, and a final hold time of 10 min at 325 °C.
EI ion source	temperature, 230 °C; energy, 70 eV
Chemical ionization	ion source 300 °C; CI electron energy 135 eV; CI methane gas flow rate 20%
MS parameters, tuning	autotune using FC43 (Perfluorotributylamine)
MS parameters, data acquisition	m/z 50 - 1200 at 5 Hz scan rate and 750 V detector voltage in both electron ionization (EI) mode and chemical ionization (CI) mode
MS parameters, data processing	Peak detection, deconvolution by MS-DIAL 4[9, 12]

SIRIUS+CSI:FingerID[13] was used to predict molecular formulas with default parameters unless stated otherwise. Code to evaluate prediction accuracy is available at https://github.com/Shunyang2018/EICI.

5.3.2. Sample preparation

We analyzed unknown metabolites from blueberry, strawberry and blackberry purchased fresh from a local supermarket. 20 mg of fresh berries were weighted and the other 2mg were dried down. All the samples were extracted with 1 ml EtOAC:H_2O(1:1). Two sets of samples were evaporated to dryness and derivatized by methoximation and trimethylsilylation as mentioned in the data acquisition part. GC-EI-MS spectra were obtained on a nominal mass GC-TOF mass spectrometer (Pegasus IV, Leco, MI) with a 95% dimethylpolysiloxane-5% phenyl column (rtx5-SilMS, Restek) and accurate mass spectra were acquired on an Agilent 7200 GC-CI-QTOF with the same type of polysiloxane column under the conditions given above.

5.4. Result and discussion

We first manually investigated CI mass spectra and confirmed the frequent observation of a pattern of ions derived from the molecular ion: $[M – CH_3]^+$, $[M - H]^+$, $[M]^+$, $[M + H]^+$, $[M + C_2H_5]^+$, and $[M + C_3H_5]^+$. Often, the $[M – CH_3]^+$ was observed as base peak ion (bp_, while molecular ion species $[M - H]^+$, $[M]^+$, $[M + H]^+$, were presented at variable abundance but usually at larger than 5% bp intensity, except for $[M + C_2H_5]^+$, and $[M + C_3H_5]^+$ that were mostly found at <5% bp intensity . Occasionally, additional ions were observed a lower intensity as described before[10-11] . A python script was developed to identify CI patterns by finding these isotopic ion groups and utilizing the nominal mass difference between them. (Figure 1). The molecular mass detection $[M - H]^+$, $[M]^+$, $[M + H]^+$ from the pattern recognition was used as precursor mass and combined with the CI spectrum as mgf format to be used for the SIRIUS+CSI:FingerID[13] software that is usually employed for tandem MS/MS spectra annotation. SIRIUS+CSI:FingerID was used to predict the molecular formula including Si in the list of search elements.

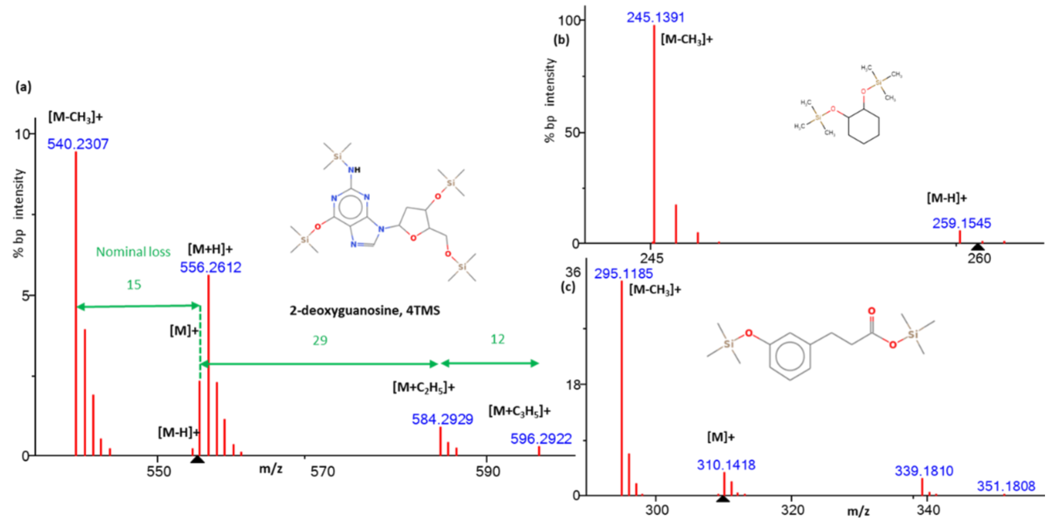


Figure 5-1. Examples of molecular ion species patterns in methane chemical ionization GC-QTOF MS. (a) CI pattern of 2'-deoxyguanosine, 4TMS, $[M + H]^+$ with $[M + C_2H_5]^+$, $[M + C_3H_5]^+$; (b) CI pattern of 1,2-cyclohexanediol, 2TMS, $[M - H]^+$; no further adducts detected; (c) CI pattern of 3-(4-hydroxyphenyl)propionic acid, 2TMS, $[M]^+$ with $[M + C_2H_5]^+$, $[M + C_3H_5]^+$

5.4.1. Overall detection rate of molecular ion species in GC-CI-QTOF MS

We probed 369 standards (Supporting Information 1) and acquired them at high concentrations in GC-methane CI-QTOF MS in mixes of 20 non-isomeric compounds. 345 TMS-derivatized versions of these compounds were detected, but 46 compounds were not detected at all even after manual curation. CI spectra were processed by the CI pattern algorithm and manually curated to find lower abundant compounds that might not have fit the algorithm pattern. Table 2 gives an overview on the diversity of chemical classes included in the mixtures using the ClassyFire software[14]. Purine and pyridines, fatty acids, indoles, carboxylic acids and hydroxy acids were well covered in CI detection, while only half of the tested organonitrogen compounds were positively identified in our tests (Table 2). Carbohydrates, classified by the ClassyFire software as organooxygen compounds, were often true negatives even in manual investigations (Table 2), most likely because these compounds bear many TMS derivative groups and for these compounds, even soft chemical ionization might lead to fragmentation of molecular ion adduct species and therefore loss of molecular ion information. Prenol lipids and steroids were also rarely detected in CI mode (Table 2), likely because of lack of ionization efficiency in CI mode compared to classic electron ionization. For most standards, retention index information was available in MassBank.us or NIST20, and hence, wide retention index windows were used to find standards within the mixtures. Here, without correct retention index data, using only accurate mass extracted ion windows led to 1.3% false positive annotations (5 compounds). We confirmed a previous report that the sensitivity of GC-MS with chemical ionization is about 20 folds lower than GC-MS electron ionization mass spectrometry.[15] This deficiency limits the use of chemical ionization in unknown metabolites identification to abundant compounds.

Table 5-2. Different chemical classes detected by CI mode(with *n*>5). Classification by ClassyFire software.

Class	**total number**	**detected by CI (%)**
Carboxylic acids and derivatives	77	83.1
Organooxygen compounds	47	55.3
Benzene and substituted derivatives	35	77.1
Fatty Acyls	29	89.7
Phenols	26	73.1
Indoles and derivatives	15	86.7
Organonitrogen compounds	11	63.6
Hydroxy acids and derivatives	9	88.9
Phenylpropanoic acids	8	87.5
Prenol lipids	8	25.0
Cinnamic acids and derivatives	7	71.4
Pyridines and derivatives	7	100.0
Steroids and steroid derivatives	7	14.3
Purine nucleosides	6	100.0

Overall, we detected 345 unique true positive standards (Supplement 1) with an average mass of 345 ± 160 Da and an average mass error for the $[M – CH_3]^+$ ion species of 0.001 ± 0.0008 Da. This data showed excellent mass accuracy for this instrument of only 2.8 ppm error that led us to expect high success rates for calculating elemental formulas. Of the molecular ion species clusters ($[M - H]^+$, $[M]^+$ and $[M + H]^+$) that were automatically detected by the algorithm, 70% had the highest intensity for $[M + H]^+$ while surprisingly many derivatives were detected with highest abundance as $[M - H]^+$ species (7%) or as $[M]^+$ species (4%) (Table 3**).** Interestingly, 14% of the $[M – CH_3]^+$ ion species were not recognized by the algorithm but were only found by manual investigations. Figure 2 shows the spectrum for 3,4-dihydroxyphenylacetic acid as an example spectrum that was rationalized manually, but that was not automatically annotated by the algorithm due to the presence of unexplained ion species above the maximum $[M + C_3H_5]^+$, here at m/z 457. In the remaining 316 cases for which we automatically found $[M – CH_3]^+$ ion species, we also detected

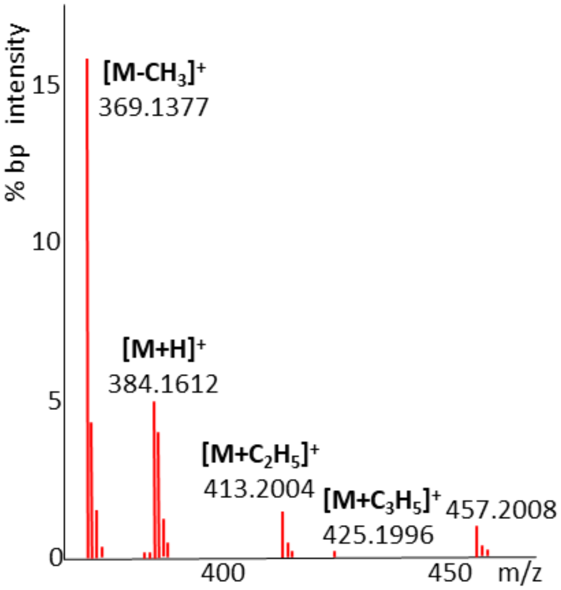


Figure 5-2. Methane CI QTOF MS spectrum of the molecular ion species region of 3,4-dihydroxyphenylacetic acid 3TMS.

corresponding $[M + C_2H_5]^+$ ion species 90% of the time, with $[M + C_3H_5]^+$ ion species detected 84% of the time. In combination, the combined pattern analysis of all signature ion species led to high confidence for the automatic detection of molecular ions in GC-QTOF MS.

Table 5-3 Count and molecular ion species of derivatized standards that were recognized automatically by the pattern algorithm

$[M]^+$	15	4%
$[M - H]^+$	24	7%
$[M + H]^+$	277	75%
Not recognized by algorithm	51	14%
total	367	Derivatized standards

Within the 51 CI spectra that did not yield automatic annotations of $[M-15]^+$ ion species, we found many examples that followed the same pattern as given in Figure 2. We rationalized these new ion species as previously unreported $[M+TMS]^+$ ions and give mass errors for three examples in Table 4. These examples unequivocally support the interpretation of these ion species, with excellent mass accuracies. Because the molecules themselves do not bear additional exchangeable, acidic protons, we conclude that these species are likely generated by intermolecular ion rearrangements of $[M]^{\cdot +}$ ions with $TMS^{\cdot}$ radicals that are cleaved from molecules within the CI reaction zone, supported by the high concentration of analyte ions used in our test cases.

Table 5-4. Examples of false negative annotations of molecular species that were missed by the automatic algorithm but rationalized as novel ion species $[M+TMS]^+$.

	observed m/z	theoretical m/z	mass error [mDa]	Ion species
3,4-dihydroxyphenylacetic acid	384.1612	384.1608	-0.4	$[M]^+$ 3TMS
	369.1377	369.1374	-0.3	$[M-CH_3]^+$ 3TMS
	413.2004	413.2000	-0.4	$[M+C_2H_5]^+$ 3TMS
	425.1996	425.2000	0.4	$[M+C_3H_5]^+$ 3TMS
	457.2088	**457.2082**	**-0.6**	**$[M+TMS]^+$ 3TMS**
phosphoric acid	315.1031	315.1033	0.2	$[M+H]^+$ 3TMS
	299.0719	299.0720	0.1	$[M-CH_3]^+$ 3TMS
	343.1345	343.1346	0.1	$[M+C_2H_5]^+$ 3TMS
	355.1342	355.1346	0.4	$[M+C_3H_5]^+$ 3TMS
	387.1428	**387.1428**	**0.0**	**$[M+TMS]^+$ 3TMS**
2,5-dihydroxyphenylacetic acid	384.1608	384.1608	0.0	$[M]^+$ 3TMS
	369.1374	369.1374	0.0	$[M-CH_3]^+$ 3TMS
	413.1995	413.2000	0.5	$[M+C_2H_5]^+$ 3TMS
	425.1985	425.2000	1.5	$[M+C_3H_5]^+$ 3TMS
	457.2082	**457.2082**	**0.0**	**$[M+TMS]^+$ 3TMS**

5.4.2. Automatic calculation of elemental formulas.

Obtaining the correct molecular formula is the starting point to identify unknown compounds in metabolomics. SIRIUS+CSI:FingerID was designed to interpret tandem mass spectrometry (MS/MS) consisting of both MS1 precursor ions and MS/MS fragment ions. SIRIUS uses fragmentation trees from mass spectral neutral loss information in addition to isotope pattern analysis to support overall calculated molecular formulas. To utilize the software for GC-CI-QTOF

MS spectra, we modified the file formats to include molecular mass information from our automatic pattern recognition algorithm. We then tested which ions were best suited to calculate correct elemental formulas in SIRIUS software, probing the most abundant [M – CH3]+ characteristic ion, the molecular ion species recognized by our pattern algorithm ([M+]+, [M-H]+ or [M+H]+) or using the isotope information in an overall combination with either molecular ion species and the [M – CH3]+ characteristic ion (Figure 3). We achieved this differentiation by either separating MS1 information as input (blue labeled ions in Figure 3), or excluding that information and only relying on the overall CI-QTOF fragment masses (green and red labeled ions in Figure 3). Surprisingly, adding isotope distribution analysis to the accurate masses for elemental formula calculations dramatically worsened the accuracy (Figure 3, Table 5) compared to calculations that did not use isotope ratio information. This result is due to the complex reactions in chemical ionization that lead to mixtures of molecular ion species and their natural isotope abundances (see Figure 3). Here, the 13C natural isotope of the [M-H]+ ion would be measured together with the 12C monoisotope ion of the [M]+ ion, as their accurate masses would be too close to be resolved with the QTOF MS instrument used here. Both species would also contribute to accurate mass and isotope abundance measurements for the [M+H]+ ion (see Figure 1). Likely for this reason, using the accurate mass of the molecular ion species with all fragment ions yielded only 60.7% correct top-hits (Table 5). In comparison, using all fragment ions, specifically with identifying the [M-CH3]+ species, gave 71.4% correct top-hits, and 87% correct hits within the top-3 ranked formulas. Here, the higher abundance of the [M-CH3]+ certainly improved measurement accuracy, but this species is also void of isotope contributions from other ion species because it can only derive from a methyl neutral loss of the corresponding 12C monoisotope [M]+ ion.

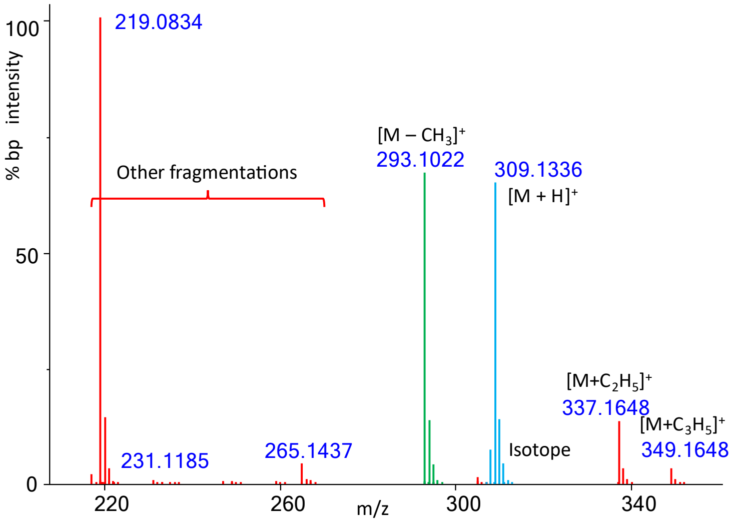


Figure 5-3. Automatic calculation of molecular formulas by Sirius/CSI:Finger ID software using CI-QTOF MS data. Example CI spectra of 2-hydroxycinnamic acid, 2 TMS. Green: [M-CH3]+ isotope cluster. Blue: molecular ion species, summarizing [M-H]+, [M]+ and [M+H]+. Red: other fragments in CI-QTOF MS spectrum.

In 8.3% of the cases, the Sirius software did not result in any hit (Figure 3, inserted table). In six of those cases, both $[M]^+$ and $[M + H]^+$ ion species returned the correct formula as top-hit. For most other cases, Sirius software did not return any hits when the formula calculation of $[M – CH_3]^+$ ions failed. Hence, in summary, more than 70% of the automatically detected molecular ion species resulted in the correct formula as top-hit when considering $[M – CH_3]^+$ species and adding [M]+ and $[M + H]^+$ ion calculations for confirmation in cases where $[M – CH_3]^+$ calculations fail. If researchers widen their search to the top-3 formula hits in compound identification workflows, more than 87% of these formulas would be expected to be correct (Figure 3, Table 5). From correct elemental formula, there are established algorithms to get 2D structure identification as previously published.

Table 5-5: Summary results for 316 TMS-derivatized compounds with automatically recognized molecular ions

Correct formula	Molecular ion species	$[M-CH_3]^+$	w/ isotope pattern
No-hit	7.9%	6.9%	12.8%
Top-10	91.7%	93.1%	87.2%
Top-5	87.6%	91.0%	83.4%
Top-3	83.1%	86.9%	78.6%
Top-hit	60.7%	71.4%	59.0%

5.5. Conclusions

369 metabolite standards were used to test the mass accuracy of a commercial GC-QTOF under methane chemical ionization, and test its suitability to automatically detect molecular ions and elemental formulas. 95% of the detected trimethylsilylated analytical standards provided high-quality CI spectra. We devised a novel algorithm that automatically searched patterns for molecular ion species, adducts and neutral loss of methyl groups and found that this algorithm correctly found 86% of all detected test molecules. Using those accurate masses in SIRIUS+CSI:FingerID software yielded 91% correct formulas in the top-5 hits, and more than 71% correct formulas retrieved as top-hit. Overall, we recommend using methane chemical ionization with GC-QTOF MS mass spectrometry as a viable route towards identification of abundant GC-MS peaks.

5.6. Reference

1. Fiehn, O., Metabolomics by Gas Chromatography-Mass Spectrometry: Combined Targeted and Untargeted Profiling. *Curr Protoc Mol Biol* **2016,** *114*, 30.4.1-30.4.32.

2. Allwood, J. W.; Erban, A.; de Koning, S.; Dunn, W. B.; Luedemann, A.; Lommen, A.; Kay, L.; Löscher, R.; Kopka, J.; Goodacre, R., Inter-laboratory reproducibility of fast gas chromatography-electron impact-time of flight mass spectrometry (GC-EI-TOF/MS) based plant metabolomics. *Metabolomics* **2009,** *5* (4), 479-496.

3. Griffiths, J., A Brief History of Mass Spectrometry. *Analytical Chemistry* **2008,** *80* (15), 5678-5683.

4. Wang, S.; Kind, T.; Bremer, P. L.; Tantillo, D. J.; Fiehn, O., Beyond the Ground State: Predicting Electron Ionization Mass Spectra Using Excited-State Molecular Dynamics. *Journal of Chemical Information and Modeling* **2022,** *62* (18), 4403-4410.

5. Leogrande, P.; Jardines, D.; Martinez-Brito, D.; de la Torre, X.; Parr, M. K.; Botrè, F., Low-energy electron ionization optimization for steroidomics analysis using high-resolution mass spectrometry. *Rapid Commun Mass Spectrom* **2021,** *35* (23), e9196.

6. NIST EI Library. https://chemdata.nist.gov/dokuwiki/doku.php?id=chemdata:start#libraries.

7. MoNA- MassBank of North America. https://mona.fiehnlab.ucdavis.edu/.

8. Wishart, D. S.; Jewison, T.; Guo, A. C.; Wilson, M.; Knox, C.; Liu, Y.; Djoumbou, Y.; Mandal, R.; Aziat, F.; Dong, E.; Bouatra, S.; Sinelnikov, I.; Arndt, D.; Xia, J.; Liu, P.; Yallou, F.; Bjorndahl, T.; Perez-Pineiro, R.; Eisner, R.; Allen, F.; Neveu, V.; Greiner, R.; Scalbert, A., HMDB 3.0--The Human Metabolome Database in 2013. *Nucleic Acids Res* **2013,** *41* (Database issue), D801-7.

9. Lai, Z.; Tsugawa, H.; Wohlgemuth, G.; Mehta, S.; Mueller, M.; Zheng, Y.; Ogiwara, A.; Meissen, J.; Showalter, M.; Takeuchi, K.; Kind, T.; Beal, P.; Arita, M.; Fiehn, O., Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics. *Nature Methods* **2018,** *15* (1), 53-56.

10. Richter, W. J.; Schwarz, H., Chemical Ionization—A Mass-Spectrometric Analytical Procedure of Rapidly Increasing Importance. *Angewandte Chemie International Edition in English* **1978,** *17* (6), 424-439.

11. Munson, M. S. B.; Field, F. H., Chemical Ionization Mass Spectrometry. I. General Introduction. *Journal of the American Chemical Society* **1966,** *88* (12), 2621-2630.

12. Tsugawa, H.; Ikeda, K.; Takahashi, M.; Satoh, A.; Mori, Y.; Uchino, H.; Okahashi, N.; Yamada, Y.; Tada, I.; Bonini, P.; Higashi, Y.; Okazaki, Y.; Zhou, Z.; Zhu, Z.-J.; Koelmel, J.; Cajka, T.; Fiehn, O.; Saito, K.; Arita, M.; Arita, M., A lipidome atlas in MS-DIAL 4. *Nature Biotechnology* **2020,** *38* (10), 1159-1163.

13. Dührkop, K.; Fleischauer, M.; Ludwig, M.; Aksenov, A. A.; Melnik, A. V.; Meusel, M.; Dorrestein, P. C.; Rousu, J.; Böcker, S., SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nature Methods* **2019,** *16* (4), 299-302.

14. Djoumbou Feunang, Y.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; Greiner, R.; Wishart, D. S., ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *Journal of Cheminformatics* **2016,** *8* (1), 61.

15. Bräkling, S.; Kroll, K.; Stoermer, C.; Rohner, U.; Gonin, M.; Benter, T.; Kersten, H.; Klee, S., Parallel Operation of Electron Ionization and Chemical Ionization for GC–MS Using a Single TOF Mass Analyzer. *Analytical Chemistry* **2022,** *94* (15), 6057-6064.

16. Fiehn, O.; Wohlgemuth, G.; Scholz, M., Setup and Annotation of Metabolomic Experiments by Integrating Biological and Mass Spectrometric Metadata. 2005; pp 224-239.

17. Kind, T.; Wohlgemuth, G.; Lee, D. Y.; Lu, Y.; Palazoglu, M.; Shahbaz, S.; Fiehn, O., FiehnLib: Mass Spectral and Retention Index Libraries for Metabolomics Based on Quadrupole and Time-of-Flight Gas Chromatography/Mass Spectrometry. *Analytical Chemistry* **2009,** *81* (24), 10038-10048.

18. BinVestigate. (http://binvestigate.fiehnlab.ucdavis.edu/).

19. Rutz, A.; Sorokina, M.; Galgonek, J.; Mietchen, D.; Willighagen, E.; Gaudry, A.; Graham, J. G.; Stephan, R.; Page, R.; Vondrášek, J.; Steinbeck, C.; Pauli, G. F.; Wolfender, J.-L.; Bisson, J.; Allard, P.-M., The LOTUS initiative for open knowledge management in natural products research. *eLife* **2022,** *11*, e70780.

20. Sorokina, M.; Merseburger, P.; Rajan, K.; Yirik, M. A.; Steinbeck, C., COCONUT online: Collection of Open Natural Products database. *Journal of Cheminformatics* **2021,** *13* (1), 2.

5.7. Supporting information

Table S1. Manual curation results of 369 CI spectra of known reference standards

	Reason	n	%
FP	Peak detected by molecular ion pattern, but accurate mass error > 5 mD and large RI difference	5	1.3
TP	Peak detected by molecular ion pattern, accurate mass	289	73.8
FN	CI ion intensity pattern too low to be used for automatic detection	14	3.6
	Additional ions at high m/z that did not fit the pattern for the algorithm	37	9.5
TN	No peak detected within expected RI windows	46	11.8

Rights and permissions

Chapter 1 is part of "Quantum Chemistry Calculations for Metabolomics"; and is reprinted with permission from American Chemical Society[1]

Chapter 2 originally appeared as "Predicting in silico electron ionization mass spectra using quantum chemistry"; and is reprinted with permission from Springer Nature[2]

Chapter 3 originally appeared as "Quantum Chemical Prediction of Electron Ionization Mass Spectra of Trimethylsilylated Metabolites"; and is reprinted with permission from American Chemical Society[3]

Chapter 4 originally appeared as "Beyond the Ground State: Predicting Electron Ionization Mass Spectra Using Excited-State Molecular Dynamics"; and is reprinted with permission from American Chemical Society[4]

Chapter 5 will be submitted to Journal of the American Society for Mass Spectrometry, with the following author list: **Shunyang Wang**, Luis Valdiviez, Yendry Carvajal Miranda, Honglian Ye, Oliver Fiehn

1. Borges, R. M.; Colby, S. M.; Das, S.; Edison, A. S.; Fiehn, O.; Kind, T.; Lee, J.; Merrill, A. T.; Merz, K. M.; Metz, T. O.; Nunez, J. R.; Tantillo, D. J.; Wang, L.-P.; **Wang, S.**; Renslow, R. S., Quantum Chemistry Calculations for Metabolomics. *Chemical Reviews* **2021,** *121* (10), 5633-5670.

2. **Wang, S**.; Kind, T.; Tantillo, D. J.; Fiehn, O., Predicting in silico electron ionization mass spectra using quantum chemistry. *Journal of Cheminformatics* **2020,** *12* (1), 63.

3. **Wang, S**.; Kind, T.; Bremer, P. L.; Tantillo, D. J.; Fiehn, O., Quantum Chemical Prediction of Electron Ionization Mass Spectra of Trimethylsilylated Metabolites. *Analytical Chemistry* **2022,** *94* (3), 1559-1566.

4. **Wang, S**.; Kind, T.; Bremer, P. L.; Tantillo, D. J.; Fiehn, O., Beyond the Ground State: Predicting Electron Ionization Mass Spectra Using Excited-State Molecular Dynamics. *Journal of Chemical Information and Modeling* **2022,** *62* (18), 4403-4410.