Predictive mass spectrometry from quantum-mechanical fragmentation and intensity modelling

Victor Posligua[1] Jeremy Tan[2] Stella Sing Yee Lee[2] Yihui Chong[2] Bingquan Shen[2] Stephen G. Dale[1,3]
[1]*Institute for Functional Intelligent Materials, National University of Singapore, Block S9, 4 Science Drive 2, 117544 Singapore*

[2]*DSO National Laboratories, 12 Science Park Drive, 118225 Singapore*

[3]*Department of Materials Science and Engineering, National University of Singapore, 9 Engineering Drive 1, 117575 Singapore*

Correspondence to: Victor Posligua victorhp@nus.edu.sg

1. Introduction

Machine learning methods have been widely explored for chemical property prediction, including electron-impact mass spectrum (EI-MS) prediction (e.g., NEIMS[1]), but their reliability often remains limited by the availability, coverage and consistency of experimental training data [2,3].

Mass spectrometry is a comparatively data rich exception, with curated EI reference libraries containing on the order of 10^5–10^6 spectra; for example, the NIST 23 EI library reports ~394K EI spectra (with replicate spectra for some compounds)[4]. However, even these resources fall short of the enormous data requirements to train AL methods and leaves gaps across chemical space, motivating complementary data sources that can generate controlled and traceable spectra.

Here we present an end-to-end quantum mechanics (QM) pipeline that predicts full m/z-intensity spectra directly from SMILES. The approach is energy-driven: fragmentation channels are ranked and filtered using QM energetics and peak intensities are assigned using simple global rules to enable transparent benchmarking and failure mode analysis. We evaluate on alcohol reference series and a subset obtained from the Critical Assessment of Small Molecule Identification (CASMI) [5], and include QCxMS2 [6] as a state-of-the-art literature comparison.

2. Methods

2.1 Workflow

From SMILES, an initial 3D conformation is generated with RDKit/ETKDG [7, 8] and the molecular ion (charge +1) is optimised with Gaussian 16 [9] (B3LYP/6-311G(d,p), zero-point corrected energies). Initial fragmentations are generated by systematic bond cleavage; charged/neutral assignments are evaluated using an energy-based discriminator ($\Delta(\Delta E)$) to decide which fragments are expected to appear. Multi-step pathways are captured via recursive subsequent fragmentations seeded by the lowest bond dissociation energy (BDE) fragments (top-3 from the initial stage), retaining charge and reusing the optimised geometries.

2.2 Intensity models

A baseline model maps fragment BDEs to relative intensities via an exponential decay with fixed, global parameters. A single-pass refinement (one-step reweighting) then reshapes the baseline spectrum toward a smooth BDE-ordered envelope anchored by (i) the smallest-BDE fragment (base peak) and (ii) a bounded parent-ion target that decreases with the density of low-BDE channels.

2.3 Benchmarking

Agreement with experiment is quantified by entropy similarity (Jensen–Shannon divergence mapped to [0,1]) after consistent preprocessing. QCxMS2 spectra are processed with the same protocol for like-for-like comparison.

3. Results and discussion

Across our selected test set, the baseline BDE–intensity model achieves a mean entropy similarity of 0.603, which is of the same order as 0.685 obtained with the recently published literature competitor QCxMS2 under the same preprocessing and scoring protocol. Applying the one-step reweighting refinement yields a consistent upward shift in most cases: the mean increases to 0.628 (+0.024, ~+4%), with a clearer gain on the alcohol subset (0.681 → 0.718, +0.037). On the CASMI subset (restricted to non-aromatic molecules within the current validated scope), backpropagation improves the mean from 0.237 to 0.257 (+0.020), while QCxMS2 gives 0.132 under the same protocol.

4. Conclusions

While our current results show promising agreement for selected chemistries, the broader goal is to generate reliable EI-MS spectra at scale to supplement experimental libraries. Accordingly, this talk will emphasise performance trends, the main bottlenecks we are addressing and early progress toward large-scale QM data generation.

Acknowledgments

This work is supported by DSO National Laboratories (Project Agreement No. DSOCL24004 under DSOAL21180) and was executed on the NUS HPC cluster (Vanda).

References

[1] Wei, J.N. *et al.*, Rapid Prediction of Electron–Ionization Mass Spectrometry Using Neural Networks. ACS Cent. Sci. 5, 700-708, 2019.

[2] Zhu, R.L. and Jonas, E. Rapid Approximate Subset-Based Spectra Prediction for Electron Ionization–Mass Spectrometry. Anal. Chem., 95, 2653-2663, 2023.

[3] Malarvannan, M., *et al.* Assessment of computational approaches in the prediction of spectrogram and chromatogram behaviours of analytes in pharmaceutical analysis: assessment review. Futur. J. Pharm. Sci., 9, 86, 2023.

[4] Scientific Instrument Services (SIS). NIST 23 Mass Spectral Library, NIST 2023/2020/2017 Database, Agilent Format Available. https://www.sisweb.com/software/ms/nist.htm (Accessed 2026-01-15).

[5] Vaniya, A. and Feihn, O. Revisiting CASMI: Compound ID for 500 New Unknowns, Using LC/MS/MS Data, 2022.

[6] Gorges, J. and S. Grimme, QCxMS2 – a program for the calculation of electron ionization mass spectra via automated reaction network discovery. Phys. Chem. Chem. Phys., 27, 6899–6911, 2025.

[7] Bento, A.P., *et al. An open source chemical structure curation pipeline using RDKit*. Journal of Cheminformatics, 12(1), 51, 2020.

[8] Riniker, S. and Landrum, G.A. Better informed distance geometry: using what we know to improve conformation generation. J. Chem. Inf. Model., 55, 2562–2574, 2015.

[9] Frisch, M.J., *et al.* Gaussian 16 Rev. C.01. Wallingford, CT, 2016.