Benchmarking tandem mass spectra of small natural product molecules via ab initio molecular dynamics

Naythan Z. X. Yeo[1], Dillon W. P. Tay[1], Shi Jun Ang[2]

[1]Institute of Sustainability for Chemicals, Energy and Environment (ISCE2), Agency for Science, Technology and Research (A*STAR), 8 Biomedical Grove, #07-01 Neuros Building, Singapore 138665, Republic of Singapore.

[2]Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, #16-16 Connexis, Singapore 138632, Republic of Singapore.

Corresponding author(s): Dillon W. P. Tay (dillon_tay@isce2.a-star.edu.sg) and Shi Jun Ang (ang_shi_jun@ihpc.a-star.edu.sg)

ABSTRACT

Natural products have proven to be valuable, particularly in the fields of drug discovery and chemogenomics. Tandem mass spectrometry, along with reference mass spectral libraries, has been frequently used to assist the characterization of natural products present in unknown complex mixtures. As current spectral libraries only contain a small percentage of known natural products, their continual expansion is crucial for accurate molecular identification. However, doing so through experimental means is often expensive and time-consuming. This study explores the use of *ab initio* molecular dynamics simulations (AIMD) based on the lightweight GFN2-xTB semiempirical Hamiltonian, to generate mass spectra for small natural products molecules. Through this approach, more than 2,700 unique mass spectra were generated and analysed in relation to the Global Natural Products Social Molecular Networking (GNPS) database. This study found that AIMD performs relative well (mean cosine similarity score of 0.68), with improved performance observed in aromatic molecules but limitations found when applied to molecules with carboxylic acid groups. Other key findings relating to experimental and simulated conditions also led to several recommendations for future work in this area. Overall, AIMD proved to have huge potential to be used to develop a putative natural product mass spectral library.

1 INTRODUCTION

Natural products are molecules produced by living organisms,[1] and have wide ranging applications, especially in the fields of drug discovery and chemogenomics.[2,3] They can also act as bio-betters in comparison with synthetic molecules as they offer greener production routes and higher biodegradability.[4] As such, there is a tremendous potential for them to be capitalised in answering growing calls for sustainability. However, one of the main challenges faced in the identification of unknown natural products, present in complex biological mixtures, is the lack of reliable methods in elucidating their molecular structures. In the past decades, the use of tandem mass spectrometry with Electron Spray Ionisation (ESI) and Collision Induced Dissociation (CID) for structural elucidation of natural products have grown in popularity.[5]

ESI is a soft ionisation process that imparts a charge to the target molecule, and often results in little fragmentation.[6] The process usually generates a precursor ion, often with the addition or removal of a proton (which will be referred to as 'M+H' and 'M-H' respectively in this report). This precursor ion is then made to collide with a neutral gas atom or molecule (CID) where it fragments into smaller fragments. Larger fragments bend less relative to smaller fragments in the presence of an electric and/or magnetic field, which allows fragments of different masses-to-charge ratio (m/z) to be

TABLE 1: DISTRIBUTION OF MOLECULES

Ionisation Mode	**Unique Molecules**	**Aromatic Compounds***	**Carbonyls**	**Amines**	**Alcohol**	**Sulfur Containing**	**Ether**
M+H	194	71	117	116	48	21	9
M-H	133	44	93	63	51	16	2
Total	327	115	210	179	99	37	11

*Refers to compounds with at least one aromatic functional group
Note that the values for total molecules do not tally because some molecules have multiple functional groups

separated. A mass analyser then records the frequency of each fragment as the intensity of its corresponding *m/z* value, which forms a molecule's mass spectrum. This technique has an edge over other fragmentation methods as it allows for non-volatile and thermally labile natural products to be analysed as well.[7]

Mass spectral libraries are usually used in conjunction with a search algorithm to identify an unknown molecule from its mass spectrum.[8,9] However, the applicability of these tools is often constrained by the limited size of the reference database, as molecules cannot be identified if its mass spectrum is not present in the library. While mass spectral libraries are constantly being updated,[10] the acquisition of experimental mass spectral data of known molecules is time-consuming, expensive, chemically complex, and could even be dangerous.[11] In this respect, various computational methods have been developed to map molecules to their corresponding mass spectra, and vice versa, to augment these databases. They are typically based on chemical fragmentation rules,[12] machine learning (ML),[13,14] atomistic simulations,[15,16] or a combination of methods.[17]

ML algorithms can be used to perform *in silico* fragmentation or directly generate mass spectra without predicting bond-breaking probabilities.[18] Though these algorithms are advantageous in their speed, they are reliant on large datasets for training which may not be always available. Furthermore, using trained ML algorithms to predict mass spectra often have limited domains of applicability as they do not extrapolate well beyond the type of molecules that they have been trained on. Rule-based methods are also limited in the same way as generalisability is limited to known fragmentation pathways. On the other hand, atomistic simulations, through the use of *ab initio* molecular dynamics (AIMD), do not share this limitation. While it is usually slower than ML or rule-based algorithms, they are only based on physical laws. This means that it could in theory, be used to generate the mass spectrum for any arbitrary compound.[19]

Motivated by the many functional properties of natural products and the challenges faced in identifying them, this study aims to build an *in silico* natural product tandem mass spectral (MS2) library to complement existing experimental MS2 libraries. As natural product MS2 libraries like the Global Natural Product Social Molecular Networking (GNPS)[20] are relatively small (circa 30,000 unique molecules), machine learning algorithms, that are often data-hungry (usually trained with >500,000 structures), [21] will usually underfit and not generalize well. Therefore, we employ semi-empirical quantum chemistry via AIMD to bridge this gap. A total of 2708 MS2 spectra comprising of 3 different collision energies and 2 ion modes were generated and compared with known MS2 spectra from GNPS (refer to section 2.3 for more details). Both the accuracy and generalisability of these generated spectra were studied to determine the viability of using AIMD to build a high-fidelity putative natural product MS2 library.

2 METHODOLOGY

2.1 Experimental Data

Experimental mass spectral data were taken from the GNPS MS2 library (accessed on 1 April 2022),[20] which contained MS2 data for natural products. The data from the GNPS database was filtered to obtain entries with valid Simplified Molecular-Input Line-Entry System (SMILES)[22] containing less than 15 atoms. Only neutral, single-species molecules were chosen for simplicity. This filtered database contained 4,794 valid mass spectra of 246 unique molecules, which included both positive and negative ionisation modes. Majority of these entries used M+H and M-H as its precursor ion with M+Na being the next most common one (Fig. 1). Hence, this study mainly focuses on molecules with M+H or M-H as precursor ions. Molecules with other precursor ions were only briefly analysed in the overview but filtered out for in subsequent sections of the paper. The distribution in functional groups of this filtered set of moleculesis shown in Table 1. One major limitation of using this dataset is that few entries contain information on the experimental collision energy used (393 out of 4,794 entries). Furthermore, factors like chamber lengths and the inert species used for collision were not reported in the database. Since these factors affect the fragmentation that occurs[16], several assumptions had to be made in the selection of experimental spectra. This is elaborated in greater detail in sections 2.3.

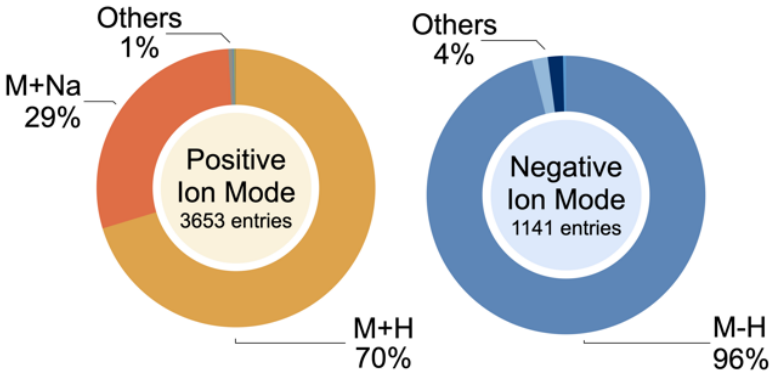


Figure 1. Distribution of different precursor ions from the GNPS. (Refer to Table S1 and Fig. S1 for a detailed breakdown).

2.2 Generation of Spectral Data

In silico fragmentation begins by the identifying different sites in molecules that can be (de-)protonated during ESI. This was done using Conformer–Rotamer Ensemble Sampling Tool (CREST)[23], a computational chemistry tool that outputs coordinates of all possible stable (de-)protonated structures. These structures of a particular molecule will be referred to as (de-)protomers. In this study, CREST was set to output structures that form within as 30kcal/mol window, for both M+H and M-H to simulate both positive and negative ion modes. Next, mass spectral peaks were generated using the quantum chemistry package, QCxMS.[16] QCxMS uses *ab initio* molecular dynamics (AIMD) simulations with the semiempirical GFN2-xTB Hamiltonian, which provides an efficient quantum mechanical description of all elements up to Z = 86 (Rn). The x in QCxMS stands for either electron ionisation (EI) or Collision Induced Dissociation (CID), but only CID was used in this study (as that is used in tandem mass spectrometry). CID was simulated with an Argon atom in a chamber of length 25cm (default QCxMS settings) while charge used was varied according to input from CREST. As fragmentation is dependent on collision energy,[16] three independent simulations for every (de-)protomer was carried out at 40, 60 and 80 eV (This will now be referred to as the simulated collision energy). A complete list of specifications for QCxMS runs is reported in Annex B.

2.3 Evaluation

Similarity Metric

To evaluate the accuracy of generated spectra, a similarity function is required as a metric to compare them with the ground-truth spectra from GNPS. Three cosine similarity functions (with different weights), Euclidean distance, absolute distance, and a composite function[13,24] was used to evaluate the similarity between two mass spectra. (Refer to Annex C for more details). It was found that the functions were generally in good agreement with each other and that the weighted cosine similarity function (weights mass 1, intensity 0.5) had the highest agreement rate among them. Therefore, the function used in this study to evaluate the various mass spectra was the weighted cosine similarity function[13] as shown in Equation (1). Note that as larger m/z peaks are more characteristic, they bear higher weightages when computing the similarity value.

$$\text{Similarity}(I_q, I_l) = \frac{\sum_{k=1}^{Mmax} {m_k I_{qk}}^{0.5} \cdot {m_k I_{lk}}^{0.5}}{\sqrt{\sum_{k=1}^{M_q} ({m_k I_{qk}}^{0.5})^2 \cdot \sum_{k=1}^{M_l} ({m_k I_{lk}}^{0.5})^2}} \quad (1)$$

Equation 1: Similarity Score between a query spectrum I_q and a library spectrum I_l, whereby M_{max} is the largest *m/z* peak in both spectra and I_{qk} and I_{lk} represent the k^{th} intensity peak in the query and library spectrum respectively.

Differences between Precursor Ions

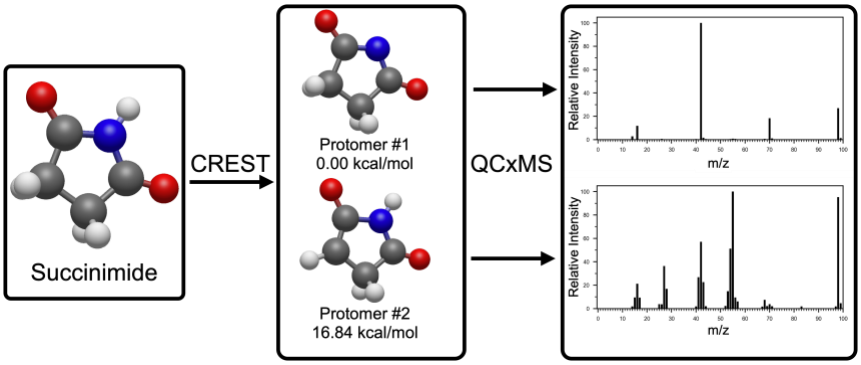


Figure 2. Generation of 2 different mass spectra of succinimide for 2 unique protomers

Each precursor ion can have different (de-)protomers that are generated by CREST which yields a different mass spectrum from QCxMS (Fig. 2). Therefore, there is a need to select a unique spectrum for each molecule. Since a protomer with a higher abundance would contribute more to the final mass spectra, a weighted average of all the different mass spectra based on their relative abundance was taken.

$$Z = \sum_i^n e^{-\frac{1}{k_B T} E_i} \quad (2)$$

Equation 2: The canonical ensemble partition function that encodes information on how the probabilities of different microstates of molecule with energy E and temperature T are partitioned (n is the number of discrete microstates).[25]

$$P_i = \frac{e^{-\frac{1}{k_B T} E_i}}{Z} \quad (3)$$

Equation 3: The probability that the system with energy E_i occupies the i^{th} microstate calculated with the partition function (2).

Each protomer's abundance was determined using the partition function (shown in equations (2) and (3)), calculated using electronic energy levels determined at the GFN2-xTB level of theory. The intensity peaks of all the different (de-)protomers' mass spectrum were then scaled by their relative abundance and summed together to form a unique mass spectrum, which I will now refer to as the Boltzmann weighted spectrum. Unless otherwise specified (in sections 3.4), this spectrum was taken as the generated spectrum for the rest of the paper.

Differences in Experimental Conditions

In the GNPS library, some molecules have multiple mass spectral entries that were likely obtained under varying experimental conditions. The weighted cosine similarity scores for these mass spectra and the generated spectra would be different. Therefore, to evaluate a molecule's performance, there is a need to obtain a unique cosine similarity score from an experimental-generated mass spectrum pair. Since the exact specifications of experimental conditions like collision energy and chamber length were largely unknown (refer to section 2.2), the cosine similarity score corresponding to the closet matching experimental-generated mass spectrum pair was taken. This was done under the assumption that the chosen pair had the closest matching conditions.

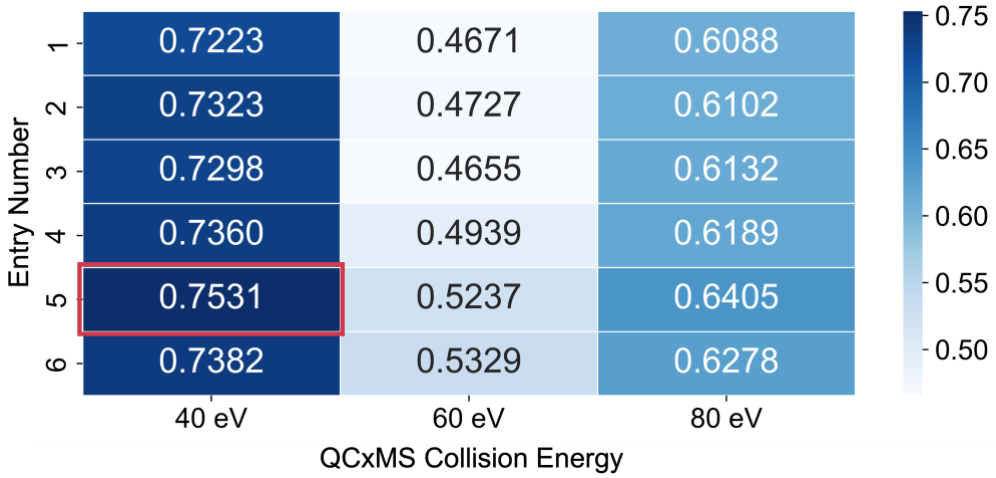


Figure 3. The cosine similarity values of homocysteine thiolactone, a molecule with six entries in GNPS. As seen here, 40 eV and entry 5 was taken as the best experimental-generated mass spectrum. Therefore, 0.7531 was taken as the similarity score for this molecule.

As an example, entry 5 of homocysteine thiolactone (Fig. 3) was assumed to match the simulated collision energy (40 eV) most closely so 0.7531 was taken as the cosine similarity score. Therefore, unless specified otherwise (section 3.4) the best mass spectrum for the different entries was taken as the experimental mass spectrum for the rest of this paper. Note that for comparisons of subsets of molecules (like precursor ions or collision energy), the best entry of that subset is taken.

2.4 Technical Details on Analysis

Through the course of this study, the following python packages have been used to aid in analysis:

- RDKit[26] was used to obtain 3D coordinates of SMILES found in the GNPS database, as well as obtain descriptors used in analysis of different molecules (section 3.2)
- xyz2mol[27] was used to obtain SMILES of 3D structures of (de-)protomers that were generated from CREST. This was used with Marvin Sketch [28] so that the 2D structures of different (de-) protomers could be visualised.
- IFG (Identify Functional Groups)[29] was a python package that outputs all functional groups found in a molecule from its SMILES.

- Pandas[30] python library was used to process data
- Matplotlib[31] and Seaborn[32] python libraries were used to plot figures in this report
- Avogadro 2[33,34] was used to generate 3D structures from xyz coordinates for figures

3 RESULTS and DISCUSSION

3.1 Overall Results

Overall, QCxMS performs well in predicting mass spectra for both M+H and M-H precursor ions as shown in Figure 4(a). There are no observable differences between the distribution of cosine similarities. Furthermore, both the mean (0.68) and median (0.73) scores of both ion modes were identical up to 2d.p. This shows that QCxMS has no bias for ion mode and that both modes are comparable in terms of accuracy. However, other precursor ions for both positive and negative ion modes performed significantly worse than M+H and M-H, with a mean of 0.28 and median of 0.16. When the precursor ion contains a different added species (defined as the charged atom or molecule that is added during ESI) like M+Na, every *m/z* value that corresponds to a fragment with that added species would be predicted wrongly. The stark contrast between accuracies of the predicted spectra shows that many of the fragments contain the added species, which results in a very poor performance for those precursor ions. Hence, following Figure 4(a), all other results have been filtered to remove entries with non-M+H and non-M-H entries to keep conditions constant.

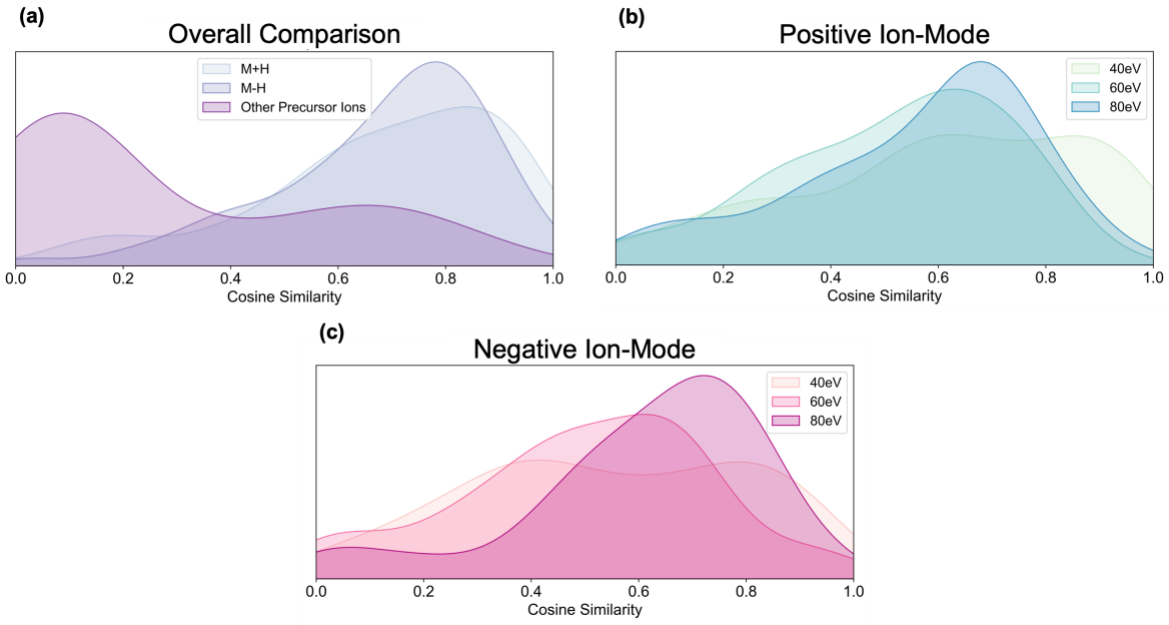


Figure 4. Distribution of cosine similarities for (a) different precursor ions, (b,c) different collision energies used in QCxMS for both ion modes.

TABLE 2: SPREAD OF COSINE SIMILARITIES FOR VARIOUS QCXMS RUN-TYPES

Collision Energy		**40 eV**	**60 eV**	**80 eV**
M+H	Mean	0.61	0.52	0.56
	Median	0.64	0.55	0.62
M-H	Mean	0.54	0.50	0.61
	Median	0.55	0.51	0.67

When comparing different simulated collision energies, 60eV performed worst for both M+H and M-H with the lowest mean and median scores. 40 eV performed slightly better than 80 eV for positive ion modes, but slightly worse for negative ion mode (Table 2). In both ion modes, 40 eV has a bi-modal distribution while 60 and 80 eV only as a single mode (Fig 4(b) and (c)). The effects of collision energies are discussed in greater detail in section 3.3, whereby both experimental and simulated collision energies were studied in greater detail.

3.2 Molecular Analysis

Next, molecules were classified based on the functional groups present in them (Table 1), resulting in two noticeable trends being identified.

Aromatic Rings

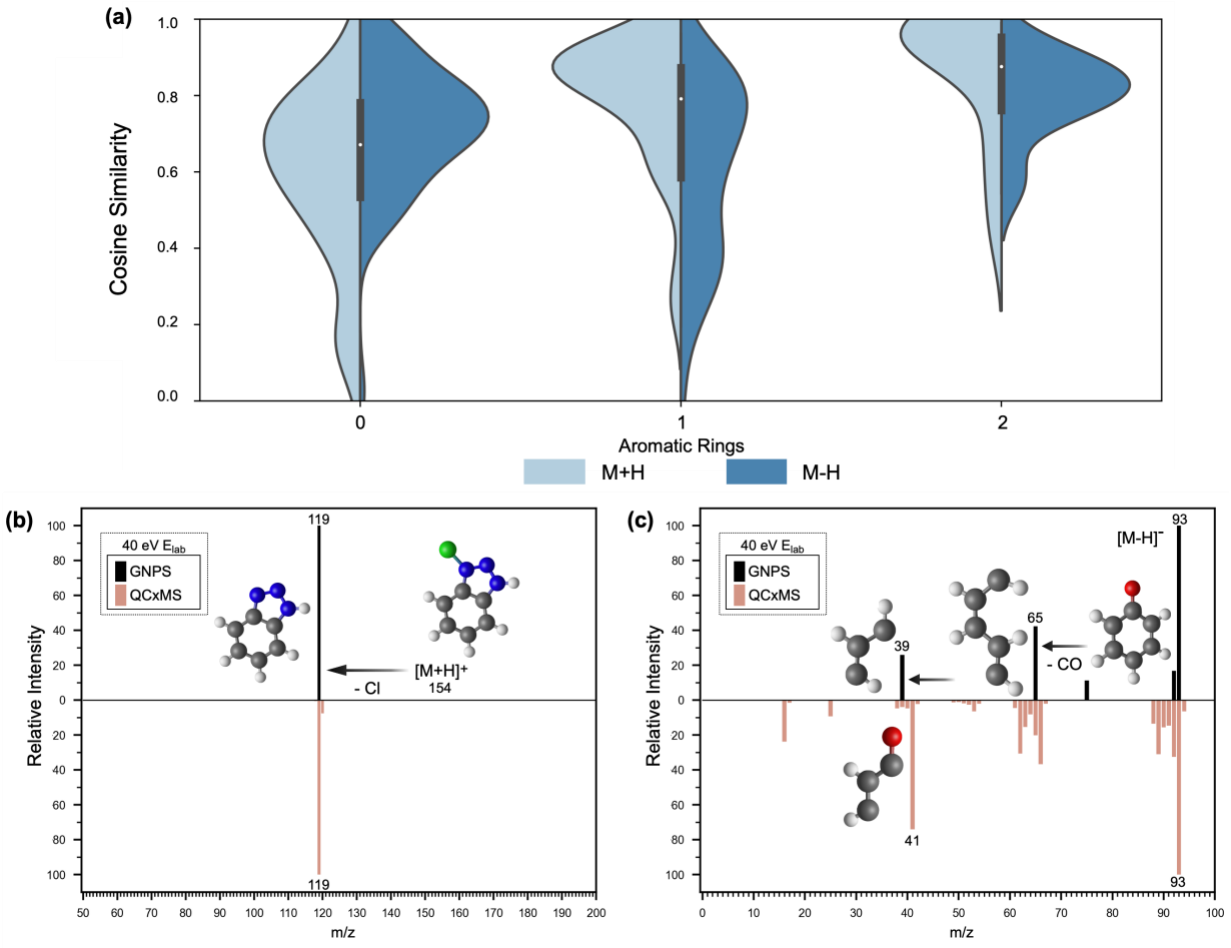


Figure 5. (a) Distribution of cosine similarity scores of both M+H and M-H ions for molecules containing Aromatic Rings. (b,c) Comparison of experimental spectra from GNPS (black, top) and generated spectra from QCxMS (red, inverted) of 1-Chlorobenzotriazole (b) and Phenol (c). The mass spectra here show the weighted average of different protomers. The 3D structures of significant fragments and precursor ions were displayed beside their peaks. Only protomer 1 at 0.00kcal/mol was displayed as the precursor ion as the contributions of other protomers were not significant (<1%).

As shown in Fig. 5(a), the presence of aromatic rings improves the accuracy of QCxMS significantly for M+H, with the mean score increasing from 0.59 (no rings) to 0.79 (one ring) and 0.88 (two rings). In M-H, the mean score first decreases from 0.72 (no rings) to 0.60 (one ring) but increases to 0.80 (two rings). While the trend is not as clearly reflected in M-H, molecules with 2 aromatic rings still had the highest mean score. The standard deviation for all the entries is less than 0.23 (Table S2). Due to resonance[35], aromatic ring structures present in the molecule rarely fragment upon collision with

a neutral gas atom. QCxMS can predict peaks corresponding to these unfragmented ring with a high degree of accuracy. This results in the increased performance for aromatic molecules. For example, in Fig. 5(b), QCxMS accurately predicted that the Cl-N bond in 1-Chlorobenzotriazole's precursor ion would undergo homolytic bond cleavage, with the stable ring system remaining intact. Note that chlorine is not seen in this mass spectra as it was uncharged when fragmented and thus not detected.

However, this trend was not clearly observed for molecules with a single aromatic ring in negative ion modes (M-H). This stabilising effect of rings was not as prominent here as fragmentation of the ring structures was more commonly observed. For example, in Fig. 5(c), there were multiple steps in the fragmentation of the Phenolate precursor ion, in which the ring did not remain intact. It is worth nothing here that while aromaticity limits fragmentation of rings, such a correlation is not clearly observed with the total number of fragments (Fig. S2) in the mass spectrum. This is because the number of fragments depend on other factors like the kinds of bond cleavage that occur.[5,6] Therefore, conclusions drawn above were based on empirical observations of various mass spectra plotted. Another interesting observation noted was that aromatic rings also affect trends seen in other descriptors. For example, there is a positive correlation between the number of amines present in molecule and the distribution of cosine similarity scores (Fig. S3(a)). However, this can be attributed to the increased percentage of aromatic molecules when the number of amines increases (Fig. S3(b)). The inverse corelation is also observed with the number of rotatable bonds present in a molecule (Fig. S4). Like before, this is related to aromatic rings as rings do not contain rotatable bonds. Aromatic rings were proposed to be the causal factor in these trends as it had the soundest theoretical backing and was a common factor among them.

Carboxylic Acid

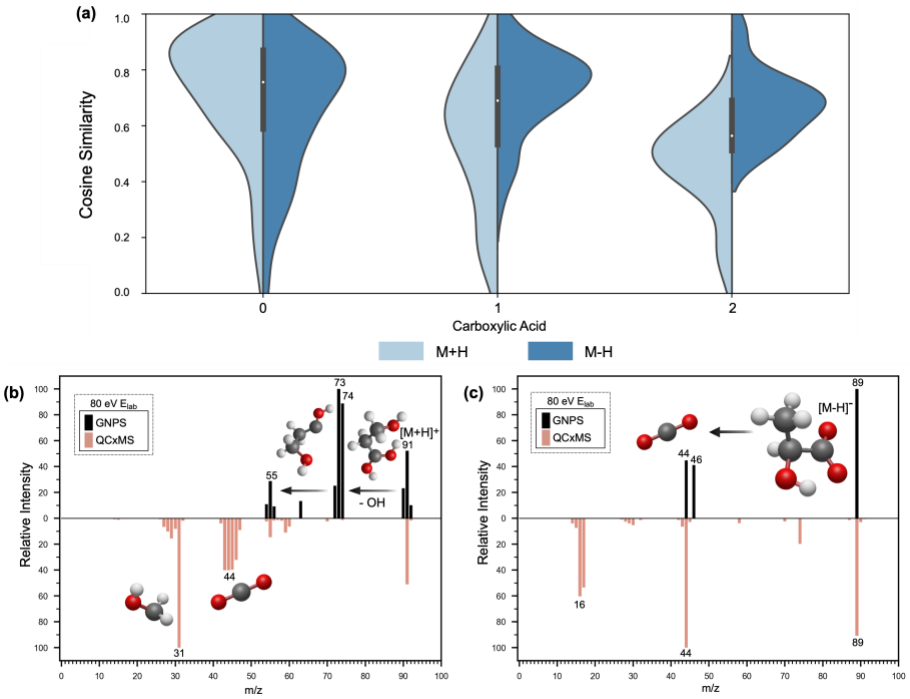


Figure 6. (a) Distribution of cosine similarity scores of both M+H and M-H ions for molecules containing Carboxylic Acid groups. (b,c) Comparison of experimental spectra from GNPS (black, top) and generated spectra from QCxMS (red, inverted) of hydroxypropionic acid (b) and Lactate (c). The mass spectra here show the weighted average of different protomers. The 3D structures of significant fragments and precursor ions were displayed beside their peaks. Only protomer 1 at 0.00kcal/mol was displayed as the precursor ion as the contributions of other protomers were not significant (<1%).

Fig. 6(a) shows an inverse relationship between the number of carboxylic acid groups in a molecule and the cosine similarity score for M+H, with the mean decreasing from 0.78 (no carboxylic acid) to 0.55 (one carboxylic acid) to 0.47 (two carboxylic acids). However, for M-H, the trend is not clear as the mean changes from 0.65 (no carboxylic acid) to 0.73 (one carboxylic acid) to 0.67 (two carboxylic acids). Like before, the standard deviations are all less than 0.23 (refer to Table S2 for all the data). On the surface, this correlation appears to be a result of a relationship with aromatic rings because there exists an inverse relationship between the number of carboxylic acid groups and the percentage of molecules that are aromatic (Fig. S5). When there are no carboxylic acid groups, 60% of molecules are aromatic. This percentage drops to 20% and then to 0% for 1 and 2 carboxylic acid groups respectively. Furthermore, like before, the trend is prominent for M+H but not M-H. However, if the correlation observed were simply dependent on aromatic rings, the distribution of similarity scores for molecules with 2 carboxylic acids should resemble the distribution of all non-aromatic molecules. This is not true for M+H as the mean for those two classes differ greatly by 0.12 (0.47 for molecules with 2 carboxylic acid groups but 0.59 for those with no aromatic rings).

While aromaticity could play a factor in this trend, the evidence shown above points to additional limitations regarding carboxylic acid groups. It is noted empirically that for M+H precursor ions that contain carboxylic acid groups, many *m/z* values are often predicted wrongly by QCxMS (as opposed to only intensities being predicted wrongly). For example, in Fig 6(b), experiments suggest a cleavage of the C-O bond in the hydroxypropionic acid precursor ion, resulting in peaks of *m/z* =74 (loss of OH). However, QCxMS instead predicted a major peak at *m/z* = 44 and 31, corresponding to CO_2^+ and CH_3O^+, revealing the limitations of the semi-empirical GFN2-xTB method. This could point to the need for more expensive methods to accurately predict the mass spectrum of such molecules.

For M-H, this limitation is not observed, with the mean scores only differing by 0.02 for molecules with 0 and 2 carboxylic acid groups (0.31 difference for M+H). Furthermore, unlike M+H, the distribution of scores for all non-aromatic molecules and molecules with 2 carboxylic acid groups are very similar (only differing by 0.05). This could be due to carboxylic acids being de-protonated at the same spot most of the time. QCxMS displays its capabilities here in accurately predicting the fragments for the carboxylate anion. For example, as seen in Fig. 6(c), two major peaks in lactate's mass spectrum of *m/z* = 44 and 89 were accurately predicted.

Other analysis

Five other functional groups, as well as 13 other RDKit 2D descriptors were investigated here. However, QCxMS appears to be relatively unbiased to these factors as they generally yielded poor correlations. Refer to Annex D for the plots and data of all the descriptors.

3.3 Collision Energy

For a fair comparison between different experimental and simulated collision energies, we further shortlist molecules with a wide range of collision energies labelled for analysis. Therefore, while there were 88 unique molecules with their collision energies (in eV) labelled, only 20 unique molecules for M+H and 15 molecules for M-H were chosen. (Refer to annex E for the full list of molecules). These molecules were chosen because they contained labelled experimental spectra for the same fixed range of collision energies. For M+H, the spectra varied from 20-70 eV at regular intervals of 10 eV, while for M-H, the spectra varied from 10-40 eV at regular intervals of 15 eV. The average values for each experimental and simulated collision energy were plotted below in Fig. 7.

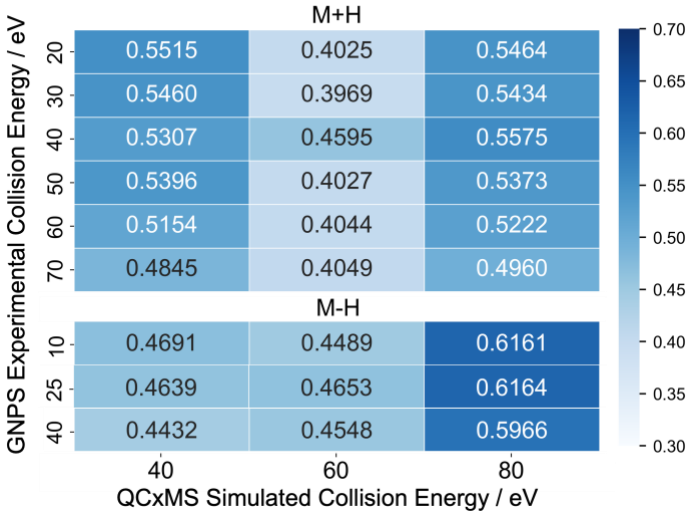


Figure 7. Annotated heatmap showing variation of average cosine similarities across collision energies in experimental and simulated mass spectra

The results shown in Fig. 7 do not point conclusively to any trends in collision energy and cosine similarity scores. Rather, it suggests that collision energy does not play that big of a part in affecting the accuracy of QCxMS for this set of molecules with less than 15 atoms. This is supported by the low standard deviation between the various cosine similarity scores. When comparing scores due to variations in experimental collision energy, the standard deviation is 0.023 (M+H) and 0.011 (M-H). For variations due to simulated collision energy, the standard deviation is 0.069 (M+H) and 0.088 (M-H). Relative to a score that ranges from 0 to 1, this variation is not very significant.

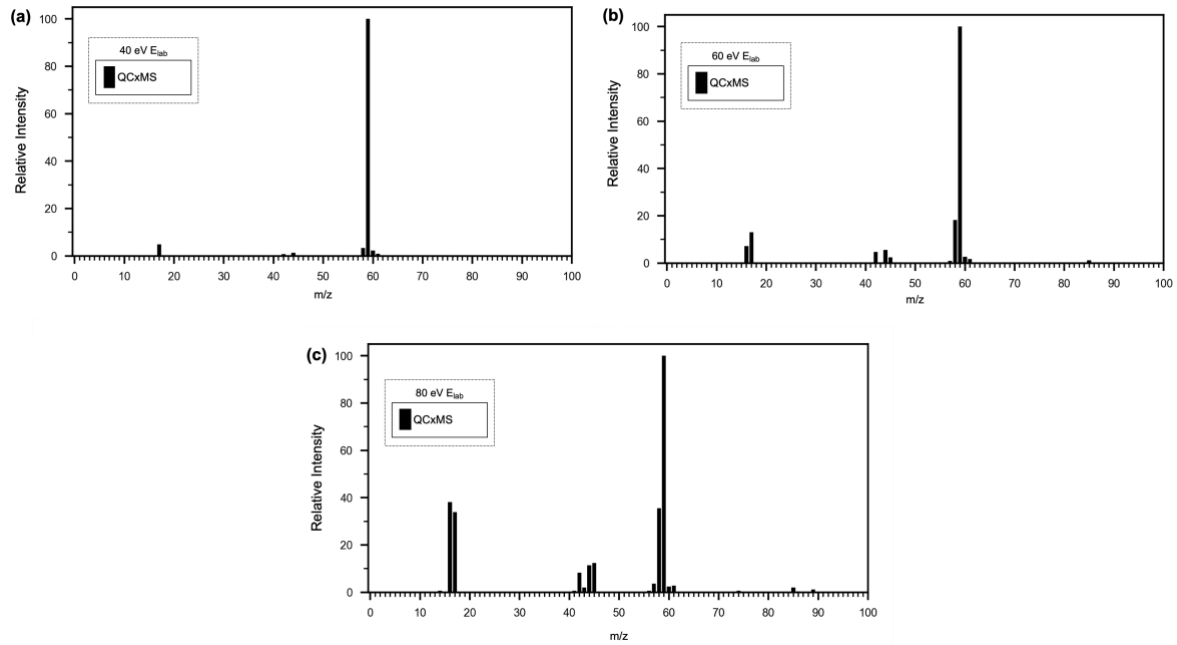


Figure 8. Variation of generated spectra of malonic acid (M-H) for collision energies of (a) 40 eV, (b) 60eV and (c) 80eV

This is likely because mass spectra of different collision energies often share similar peaks with varying intensities. For example, the predicted mass spectra of malonic acid shown in Fig. 8 has prominent peaks corresponding to *m/z* = 17 and 59 ($OH^-$and CH_2COOH^-). Those two peaks are present in all three variations of the simulated collision energy, which

results in the mass spectra being relatively similar to each other. This can be accounted for by the limited number of fragmentation pathways for this set of molecules with 15 atoms.

An interesting point noted from Fig. 7 is that QCxMS always predicts better when the simulated collision energy is higher than experimental one. A similar findings was also reported by Koopman et al.,[16,36] where the generated spectra were matched with experimental spectra that were approximately 15-30 eVs lower in collision energy.

3.4 Different (de-)protomer mass spectrum

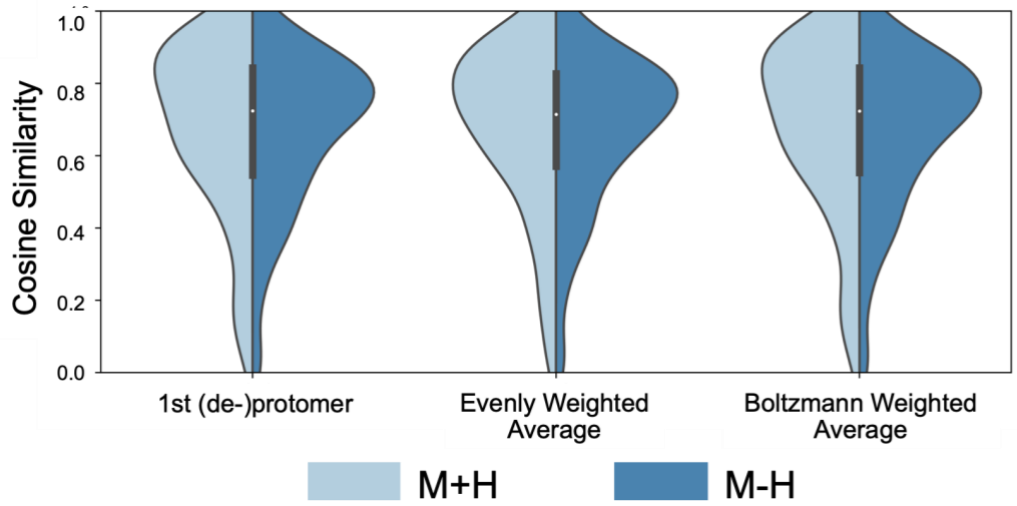


Figure 9. Distribution of cosine similarity scores when taking the 1^{st}(de-)protomer's mass spectrum (lowest energy) , an evenly weighted average or the Boltzmann weighted average mass spectrum as the generated spectrum.

In this paper, the Boltzmann weighted mass spectrum was taken as the generated mass spectrum because it (in theory) allows for the most accurate depiction of tandem mass spectrometry. Two other ways of identifying a unique mass spectrum from different (de-)protomers were explored here. The first method is to take the first (de-) protomer's (corresponding to the lowest GFN2-xTB energy) mass spectrum as the true mass spectrum. The second method takes the average of all the different mass spectra generated from each (de-)protomer by scaling each intensity peak evenly. This means that all the mass spectra contribute equally to the final spectrum regardless of its relative abundance. (This will now be referred to as the evenly weighted mass spectrum). It is seen in Fig. 9 that there are no significant differences between using the three ways to select a unique generated mass spectrum. The mean cosine similarity for all three methods were identical for M-H (up to 2d.p.), and only differed by 0.01 for M+H (evenly weighted mass spectrum performed worse than the other two by 0.01). Refer to Table S3 for the data. This can be attributed to the fact that the mass spectra generated for (de-)protomer structures of the same molecule and conditions are very similar. The average cosine similarity between (de-)protomer structures is 0.91 (here the mean was calculated across both ion modes and all simulated collision energies). This results in the distribution of cosine similarity scores for the three methods being almost identical.

3.5 Recommendations for Mass Spectral Library

In the development of a putative natural product mass spectral library, considering only small molecules (molecules with fewer than 15 atoms), the following two recommendations are made.

- Perform calculations to simulate the mass spectrum with one collision energy. Performing multiple calculations at different energy levels to match different experimental collision energies is not recommended as the impact on collision energy on small molecules is not very significant. Since a higher simulated collision energy (relative to

experimental collision energy) has a slightly better performance, 80 eV is recommended. Furthermore, using 80 eV would usually result in peaks with larger intensity at lower *m/z* values which will make the spectrum more characteristic and hence could be better for matching in a mass spectral library.

- Only consider the lowest energy (de-)protomer structure when simulating mass spectrum since the performance of it is almost identical to calculating mass spectra for all structures. This is expected to significantly reduce the total amount of computational time.

4 CONCLUSIONS

A comprehensive study on the use of *ab initio* molecular dynamics (AIMD) to generate putative mass spectra of small natural product molecules was performed. 2,708 unique mass spectra across three collision energies and two ion modes were generated and benchmarked against experimental spectra from GNPS. Overall, AIMD has shown good potential in generating accurate mass spectra, with the mean cosine similarity score being 0.68. We found that there were no significant differences in the performance for positive and negative ion modes. Another key finding is that the AIMD spectra of molecules containing aromatic rings has higher accuracy as aromatic structures are often un-fragmented. Limitations in the predictive power of AIMD for molecules with carboxylic acid groups were also identified. In addition, it was found that variations in collision energy and differences due to different (de-) protomer structures did not affect the accuracy of our predictions.

These findings led to several recommendations for developing a natural product mass spectral library. To further improve the reliability of a generated mass spectral library, a benchmarking work of using various approximated Hamiltonians can be explored to mitigate limitations in molecules that contain carboxylic acid groups. Such a study will be useful in identifying more efficient and accurate methods for putative natural product mass spectral library generation. Furthermore, the use of machine learning (ML) could also be employed to improve spectra accuracies. For example, the generated spectra from AIMD could be passed through a spectra-to-spectra model to map them to the ground experimental truths. It is envisaged that the synergy between AIMD and ML techniques will allow efficient building of high-fidelity putative MS2 spectra library and allow scientists to discover functional natural products from biological mixtures with an unprecedented speed.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge funding support from the Agency for Science, Technology and Research (A*STAR), Singapore (#21719). This work was supported by the A*STAR Computational Resource Centre through the use of its high performance computing facilities.

REFERENCES

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

[2] Lachance, H., Wetzel, S., Kumar, K., & Waldmann, H. Charting, navigating, and populating natural product chemical space for drug discovery. *Journal of Medicinal Chemistry* **2012**, *55* (13), 5989–6001.

[3] Khan R. A. Natural products chemistry: The emerging trends and prospective goals. *Saudi Pharmaceutical Journal* **2018**, *26* (5), 739–753.

[4] Atanasov, A.G., Zotchev, S.B., Dirsch, V.M., *et al*. Natural products in drug discovery: advances and opportunities. *Nature Reviews Drug Discovery* **2021**, *20*, 200-216.

[5] Demarque, D. P., Crotti, A. E., Vessecchi, R., Lopes, J. L., & 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.

[6] Steckel, A., & Schlosser, G. An Organic Chemist's Guide to Electrospray Mass Spectrometric Structure Elucidation. *Molecules (Basel, Switzerland)* **2019**, *24* (3), 611.

[7] Ho, C. S., Lam, C. W., Chan, M. H., Cheung, R. C., Law, L. K., Lit, L. C., Ng, K. F., Suen, M. W., & Tai, H. L. Electrospray ionisation mass spectrometry: principles and clinical applications. *The Clinical Biochemist. Reviews* **2003**, *24* (1), 3–12.

[8] Stein, S. Mass spectral reference libraries: An ever-expanding resource for chemical identification. *Analytical Chemistry* **2012**, *84* (17), 7274–7282.

[9] Smith, J. S., & Thakur, R. A. Mass spectrometry. *Food Science Texts Series* **2010**, 457–470.

[10] King, E., Overstreet, R., Nguyen, J., & Ciesielski, D. Augmentation of MS/ms libraries with spectral interpolation for improved identification. *Journal of Chemical Information and Modeling* **2022**, *62* (16), 3724–3733.

[11] Chernicharo, F. C. S., Modesto-Costa, L., & Borges, I. Molecular dynamics simulation of the electron ionization mass spectrum of tabun. *Journal of Mass Spectrometry* **2020**, *55* (6).

[12] Zhou, J., Weber, R. J., Allwood, J. W., Mistrik, R., Zhu, Z., Ji, Z., Chen, S., Dunn, W. B., He, S., & Viant, M. R. HAMMER: automated operation of mass frontier to construct in silico mass spectral fragmentation libraries. *Bioinformatics (Oxford, England)* **2014**, *30* (4), 581–583.

[13] 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.

[14] 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.

[15] Grimme, S. Towards first principles calculation of electron impact mass spectra of molecules. *Angewandte Chemie International Edition* **2013**, *52* (24), 6306–6312.

[16] 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* (7), 1735–1751.

[17] Wang, F., Liigand, J., Tian, S., Arndt, D., Greiner, R., & Wishart, D. S. CFM-ID 4.0: More accurate ESI-MS/MS spectral prediction and compound identification. *Analytical Chemistry* **2021**, *93* (34), 11692–11700.

[18] Nguyen, D. H., Nguyen, C. H., & Mamitsuka, H. Recent advances and prospects of computational methods for metabolite identification: a review with emphasis on machine learning approaches. *Briefings in Bioinformatics* **2018**, *20* (6), 2028–2043.

[19] 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).

[20] Wang, M., Carver, J., Phelan, V. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology* **2016**, *34* (8), 828–837.

[21] Stravs, M.A., Dührkop, K., Böcker, S. *et al.* MSNovelist: de novo structure generation from mass spectra. *Nature Methods* **2022,** *19* (7), 865–870.

[22] Weininger, D. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling* **1988**, *28* (1), 31–36.

[23] Grimme, S. Exploration of chemical compound, conformer, and reaction space with meta-dynamics simulations based on tight-binding quantum chemical calculations. *Journal of Chemical Theory and Computation* **2019**, *15* (5), 2847–2862.

[24] Stein, S. E., & Scott, D. R. Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry* **1994**, *5* (9), 859–866.

[25] Schmitz, K. S. Classical statistical mechanics. In *Physical Chemistry*; Elsevier: Amsterdam, 2017; pp 559–632.

[26] RDKit: Open-source cheminformatics. https://www.rdkit.org, Version = 2020.09.1.0

[27] Kim, Y., & Kim, W. Y. Universal structure conversion method for organic molecules: From atomic connectivity to three-dimensional geometry. *Bulletin of the Korean Chemical Society* **2015**, *36* (7), 1769–1777.

[28] Cherinka, B., Andrews, B. H., Sánchez-Gallego, J., *et al.* (2019). Marvin: A tool kit for streamlined access and visualization of the SDSS-IV manga data set. *The Astronomical Journal* **2019**, *158* (2), 74.

[29] Identify Functional Groups: https://github.com/wtriddle/IFG

[30] Virtanen, P., Gommers, R., Oliphant, T. E., *et al.* SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **2020**, *17* (3), 261-272.

[31] Hunter, J. D. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* **2007**, *9* (3), 90–95.

[32] Waskom, M. Seaborn: Statistical data visualization. *Journal of Open Source Software* **2021**, *6* (60), 3021.

[33] Avogadro 2: an open-source molecular builder and visualization tool. Version 1.91.0 http://avogadro.cc/

[34] 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).

[35] Aihara, J. Why Aromatic Compounds Are Stable. *Scientific American* **1992**, *266* (3), 62–69.

[36] Koopman, J., & Grimme, S. Calculation of mass spectra with the QCXMS method for negatively and multiply charged molecules. *Journal of the American Society for Mass Spectrometry* **2022**, *33* (12), 2226–2242.

Annex A: Supplementary Figures

TABLE S1: BREAKDOWN OF PRECURSOR IONS

Positive Ion Mode						
Precursor Ion	M+H	M+Na	2M+H	M-H+2Na	M-H2O+H	2M+Na
Count	2568	1053	16	7	7	2
Negative Ion Mode						
Precursor Ion	M-H	2M-2H+Na	2M-H	M+acetate	-	
Count	1096	21	20	4	-	

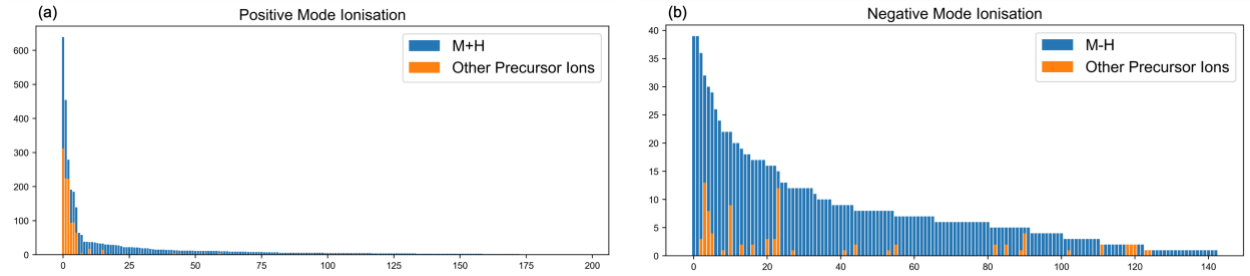


Figure S1. Distribution of number of entries for different molecules, with blue representing entries with either M+H or M-H as the precursor ion and orange representing entries with other precursor ions for (a) positive mode and (b) negative mode of ionisation.

TABLE S2: DATA FOR MOLECULAR ANALYSIS

Functional Group / Descriptors		**Aromatic Rings**			**Carboxylic Acid**		
Count		0	1	2	0	1	2
M+H	**Mean**	0.59	0.79	0.88	0.73	0.55	0.47
	Median	0.63	0.86	0.95	0.78	0.61	0.50
	StDev	0.22	0.17	0.17	0.21	0.23	0.15
M-H	**Mean**	0.72	0.60	0.80	0.65	0.73	0.67
	Median	0.73	0.65	0.82	0.72	0.76	0.68
	StDev	0.16	0.23	0.11	0.23	0.14	0.12

Functional Group / Descriptors		**Amines**					**Rotatable Bonds**			
Count		0	1	2	3	4/5	0	1	2	3
M+H	**Mean**	0.58	0.72	0.77	0.88	0.83	0.77	0.62	0.51	0.35
	Median	0.64	0.75	0.82	0.92	0.89	0.84	0.66	0.53	0.38
	StDev	0.24	0.19	0.13	0.12	0.17	0.20	0.22	0.18	0.15
M-H	**Mean**	0.66	0.66	0.68	0.74	0.82	0.64	0.76	0.70	0.54
	Median	0.69	0.73	0.71	0.76	0.83	0.70	0.79	0.70	0.54
	StDev	0.20	0.22	0.16	0.12	0.08	0.23	0.13	0.13	0.04

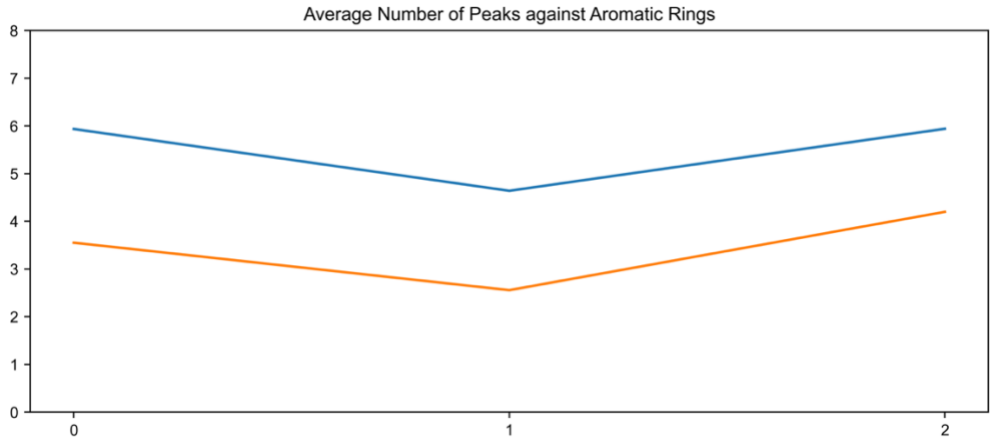


Figure S2. Plot showing the average number of peaks (defined as an *m/z* value with intensity above 1) for M+H (blue) and M-H (orange). A decreasing trend between aromatic rings and lesser peaks is not observed here.

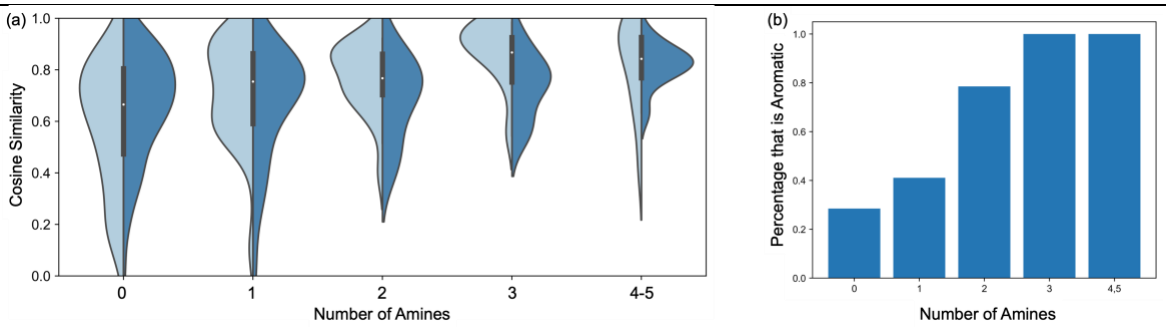


Figure S3. (a) The distribution of cosine similarity against number of amines (see table S2 for data) (b) Percentage of molecules that are aromatic plotted for different number of Amines in a molecule, the increasing trend shown in (a) is due to an increase in aromatic compounds.

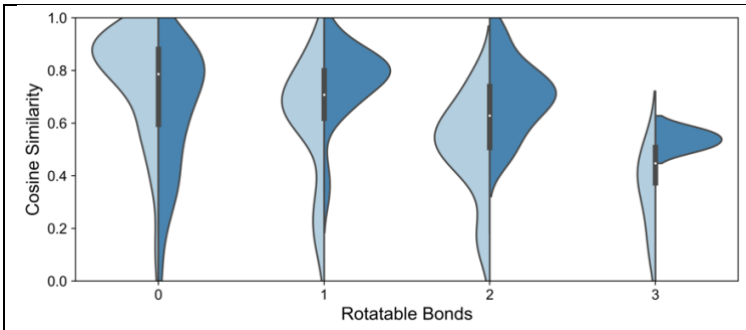


Figure S4. The distribution of cosine similarity against number of rotatable bonds (see table S2 for data)

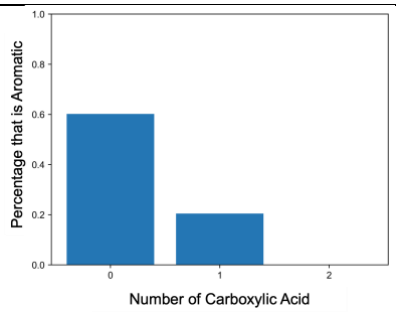


Figure S5. The percentage of molecules that are aromatic plotted for different number of carboxylic acids in a molecule

TABLE S3: DATA FOR SECTION 3.4

	Boltzmann Weighted		**Evenly Weighted**		**First (de-)protomer**	
	M+H	M-H	M+H	M-H	M+H	M-H
Mean	0.68	0.67	0.67	0.67	0.68	0.67
Median	0.73	0.71	0.71	0.72	0.73	0.72
Standard Deviation	0.24	0.21	0.23	0.21	0.24	0.21

Annex B: QCxMS Details

Breakdown of run-types

Collision Energy / eV	Type	Succeeded	SCF convergence failure
40	M+H	621	4
	M-H	261	5
60	M+H	759	3
	M-H	261	5
80	M+H	618	4
	M-H	261	5
Total	-	2781	26

QCxMS settings:

QC Program : xTB

QC Level : GFN2-xTB

Dispersion : D4

M+ Ion charge(charge) : 1,-1

total traj. (ntraj) : 25 x number of atoms

time steps (tstep) : 0.50 fs

sim. time / MD (tmax) : 0.75 ps

Initial temp. (tinit) : 500.00 K

---------- CID settings ----------

Collision Gas : Ar

E (LAB) : 40.00, 60.00, 80.00 eV

Activation Run - Type : General

Gas Pressure (PGas) : 0.132 Pa

Gas Temp. (TGas) : 300.00 K

Cell length (lchamb) : 0.250 m

Annex C: Evaluation Metrics

To evaluate all the different metrics, all the molecules in GNPS database were compared. The goal of this study was to evaluate if the metrics agree with each other in terms of which entry had the highest score.

The 6 functions[13,24] that were investigated are the following:

Cosine Similarity:

$$\frac{(\sum W_L W_U)^2}{\sum {W_L}^2 \sum {W_U}^2}$$

Euclidean Distance:

$$(1 + \frac{\sum (W_L - W_U)^2}{\sum {W_U}^2})^{-1}$$

Absolute Distance:

$$(1 + \frac{\sum |W_L - W_U|}{\sum W_U})^{-1}$$

Composite:

$$\frac{N_U F_D + N_{L\&U} F_R}{N_U + N_{L\&U}}$$

Whereby, the following terms are defined as follows:

L = library spectrum, taken to be generated spectrum for QCxMS

U = unknown/query spectrum, taken to be experimental spectrum from GNPS

$W = (m/z\ value)^{MW} * (Peak\ Intensity)^{IW}$, MW and IW are unique constants (below)

$N = Number\ of\ Peaks$

$F_D = Cosine\ Similiarity\ Score$

$F_R = \frac{1}{N_{L\&U}} \sum_i^{L\&U} (\frac{W_{L,i}}{W_{L,i-1}} \frac{W_{U,i-1}}{W_{U,i}})^{IW}$, Ratio of peaks term

With the following values taken for MW and IW:

	MW	IW
Cosine Similarity 1	1	0.5
Cosine Similarity 2	3	0.6
Cosine Similarity 3	0.5	0.5
Euclidean Distance	2	0.6
Absolute Distance (Ratio of peaks term)	0	1
Absolute Distance (Cosine Similarity Term)	3	0.5

Here, 335 unique molecules and ion modes were studied. The cosine similarity scores of the Boltzmann weighted spectrum and all experimentally generated mass spectrum were calculated for every entry in GNPS. Molecules and ion modes with only 1 entry were not considered. The entry that has the highest similarity score was recorded as the best entry. The metrics are said to agree if they compute the same entry to be the best entry for a particular molecule and ion mode.

As seen in Fig. C1, all 6 metrics agree 39% of the time (130/335). This good agreement shows that the metrics used were all generally suitable to evaluate the similarity of two mass spectra.

To determine the 'best' of the 6 metrics, the number of times a metric agrees falls within the 'majority' of metrics is recorded. The majority is defined as the highest agreement possible for a particular molecule and ion mode. For example, for agreement 6, all metrics would fall in the 'majority'. For agreement 5, 5 of the metrics fall in the 'majority' while one does not. For molecules and ion modes with multiple best entries (for eg two metrics point to entry 1, another two point to entry 2), then both are taken as the 'majority'. In Fig.C2, it is also seen that cosine similarity 1 agrees with the 'majority' 91% of the time, therefore, it is the chosen metric.

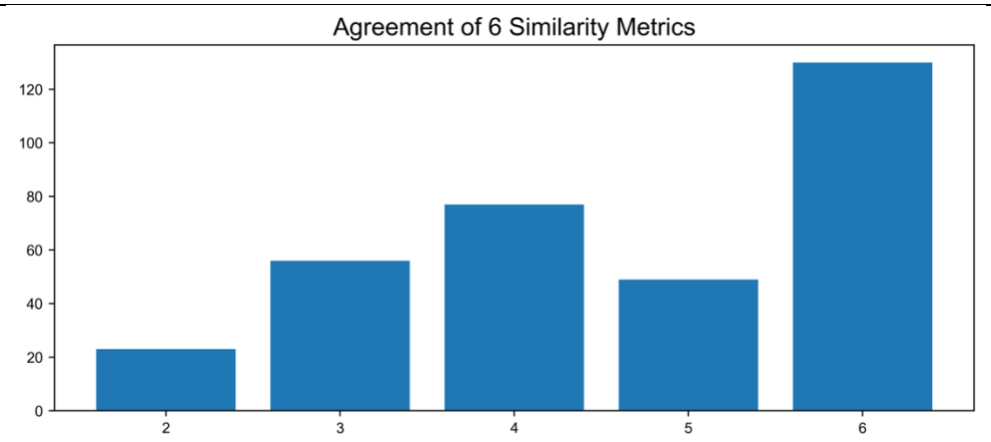


Figure C1: Agreement of all 6 Similarity Metrics

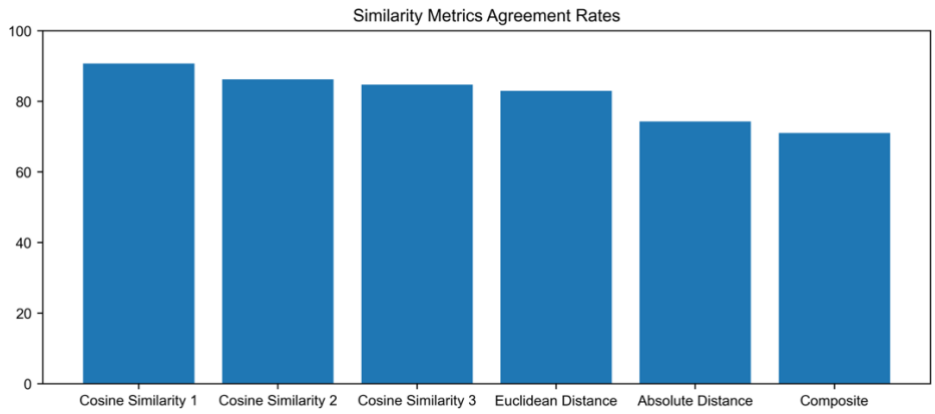


Figure C2: Agreement rates of all 6 Similarity Metrics

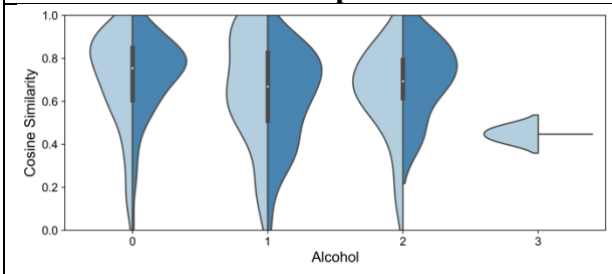
Annex D: Other Descriptors

The sign (R) beside the descriptor indicates that the values have been rounded and binned into 4 distinct bins for the sake of the plot. Here, the mean of the bin is reported. For integer values like number of atoms, the values have been binned and the exact range is reported. If the standard deviation is not reported, it means that that value only has one entry.

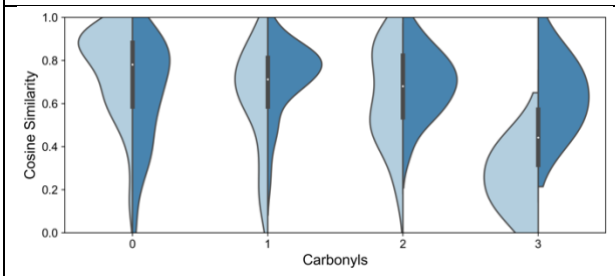
Legend (Precursor Ion Colour):

- M+H
- M-H

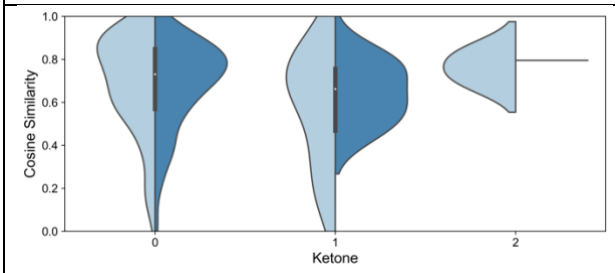
Other Functional Groups



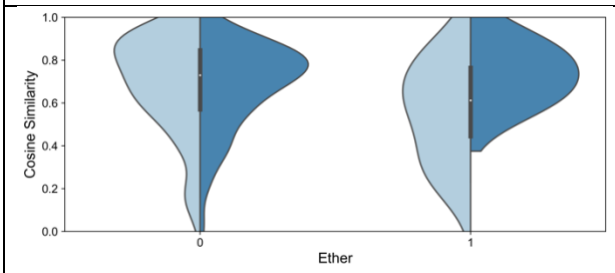
Alcohol		0	1	2	3
M+H	Mean	0.69	0.67	0.64	0.45
	Median	0.75	0.67	0.65	0.45
	StDev	0.22	0.25	0.20	0.04
M-H	Mean	0.72	0.61	0.71	0.45
	Median	0.76	0.67	0.72	0.45
	StDev	0.18	0.21	0.16	-



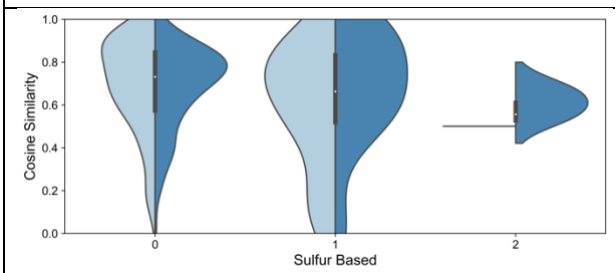
Carbonyls		0	1	2	3
M+H	Mean	0.76	0.62	0.64	0.26
	Median	0.80	0.66	0.65	0.26
	StDev	0.19	0.24	0.22	0.16
M-H	Mean	0.64	0.73	0.69	0.63
	Median	0.69	0.757	0.70	0.63
	StDev	0.24	0.15	0.15	0.17



Ketone		0	1	2
M+H	Mean	0.69	0.58	0.76
	Median	0.73	0.67	0.76
	StDev	0.23	0.26	0.09
M-H	Mean	0.68	0.64	0.80
	Median	0.73	0.61	0.80
	StDev	0.20	0.14	-



Ether		0	1
M+H	Mean	0.69	0.56
	Median	0.73	0.60
	StDev	0.23	0.23
M-H	Mean	0.68	0.74
	Median	0.73	0.74
	StDev	0.20	0.15

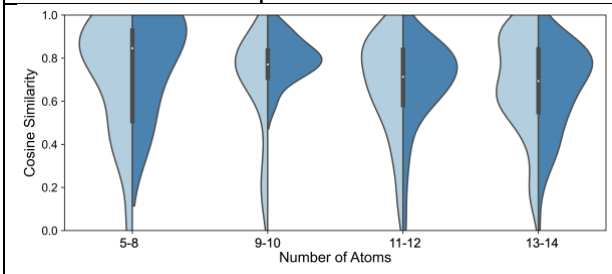


Sulfur Based		0	1	2
M+H	Mean	0.69	0.58	0.50
	Median	0.73	0.62	0.50
	StDev	0.22	0.29	-
M-H	Mean	0.68	0.68	0.61
	Median	0.73	0.69	0.61
	StDev	0.19	0.27	0.08

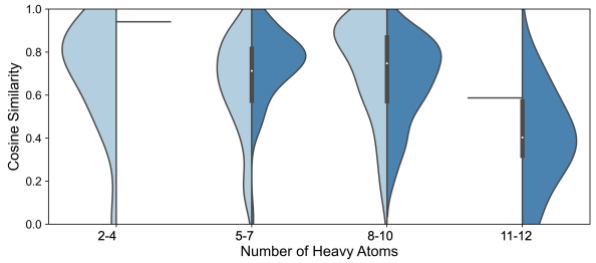
*Refers to functional groups that contain Sulfur

Other RDKit 2D Descriptors

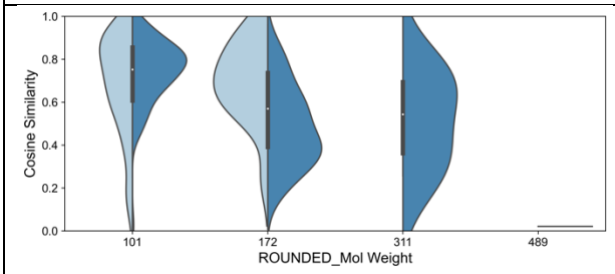
Refer below for explanations of what some these descriptors mean



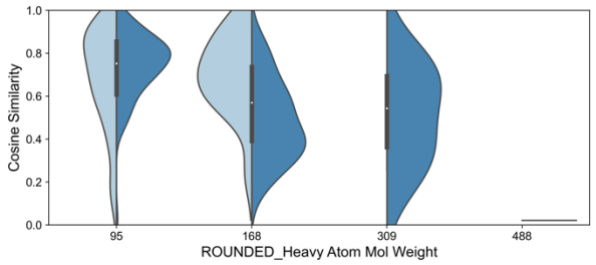
Number of Atoms		5-8	9-10	11-12	13-14
M+H	Mean	0.71	0.68	0.68	0.68
	Median	0.83	0.75	0.71	0.69
	StDev	0.28	0.23	0.22	0.23
M-H	Mean	0.81	0.80	0.68	0.66
	Median	0.94	0.79	0.72	0.70
	StDev	0.23	0.10	0.19	0.20



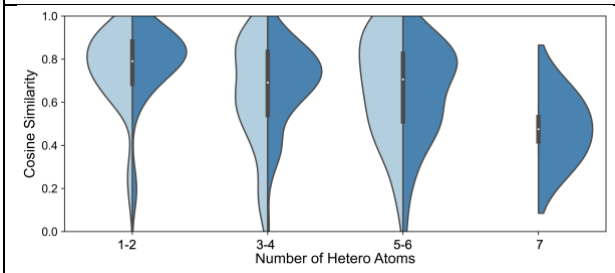
Number of Heavy Atoms		2-4	5-7	8-10	11-12
M+H	Mean	0.72	0.64	0.74	0.59
	Median	0.76	0.67	0.81	0.59
	StDev	0.24	0.22	0.23	-
M-H	Mean	0.94	0.74	0.66	0.38
	Median	0.94	0.78	0.69	0.40
	StDev	-	0.16	0.19	0.22



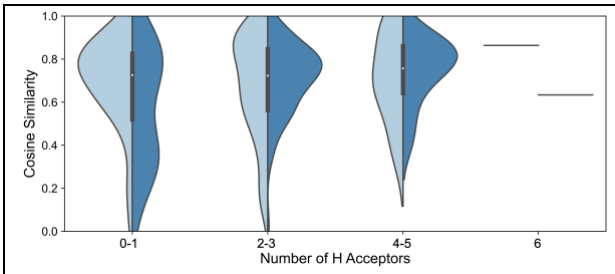
Mol Weight (R)		101	172	311	489
M+H	Mean	0.68	0.70	-	-
	Median	0.73	0.71	-	-
	StDev	0.23	0.19	-	-
M-H	Mean	0.74	0.48	0.51	0.02
	Median	0.78	0.42	0.54	0.02
	StDev	0.15	0.18	0.22	-



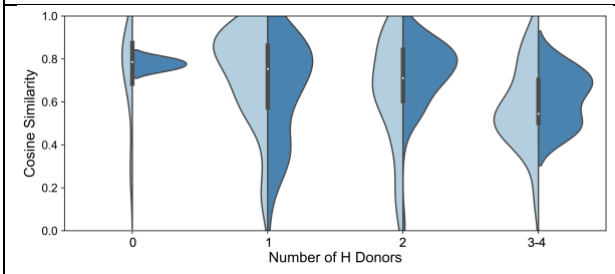
Heavy Atom Mol Weight (R)		95	168	309	488
M+H	Mean	0.68	0.70	-	-
	Median	0.73	0.71	-	-
	StDev	0.23	0.19	-	-
M-H	Mean	0.74	0.48	0.51	0.02
	Median	0.78	0.42	0.54	0.02
	StDev	0.15	0.18	0.22	-



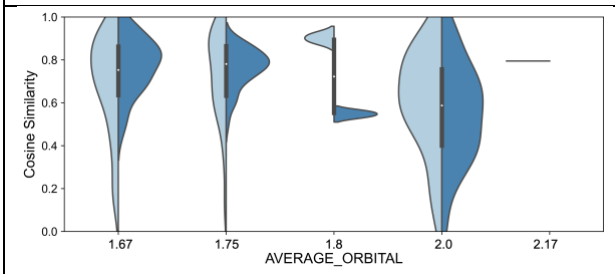
Number of Hetero Atoms		1-2	3-4	5-6	7
M+H	Mean	0.75	0.63	0.68	-
	Median	0.78	0.65	0.71	-
	StDev	0.18	0.25	0.24	-
M-H	Mean	0.79	0.68	0.64	0.48
	Median	0.82	0.71	0.70	0.48
	StDev	0.17	0.18	0.22	0.16



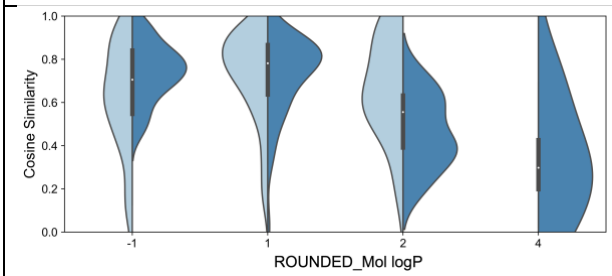
Number of H Acceptors		0-1	2-3	4-5	6
M+H	Mean	0.70	0.67	0.72	0.86
	Median	0.75	0.71	0.71	0.86
	StDev	0.21	0.24	0.21	-
M-H	Mean	0.57	0.71	0.75	0.63
	Median	0.62	0.74	0.79	0.63
	StDev	0.27	0.16	0.15	-



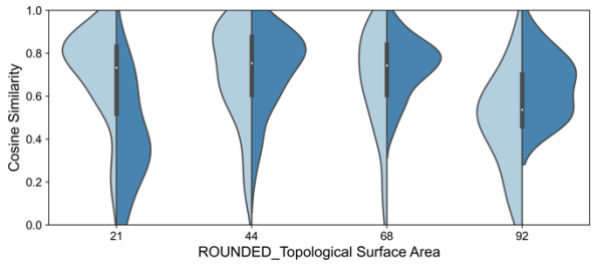
Number of H Donors		0	1	2	3-4
M+H	Mean	0.73	0.73	0.65	0.58
	Median	0.79	0.78	0.66	0.53
	StDev	0.21	0.21	0.25	0.19
M-H	Mean	0.78	0.64	0.73	0.62
	Median	0.78	0.70	0.76	0.63
	StDev	0.03	0.23	0.15	0.12



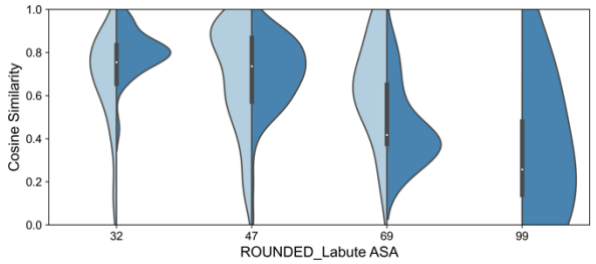
Average Orbital		1.67	1.75	1.8	2.0
M+H	Mean	0.69	0.72	0.90	0.60
	Median	0.73	0.75	0.90	0.62
	StDev	0.23	0.21	0.02	0.24
M-H	Mean	0.77	0.76	0.55	0.55
	Median	0.80	0.79	0.55	0.57
	StDev	0.13	0.09	0.02	0.23
M+H 2.17: One entry cosine similarity 0.79					



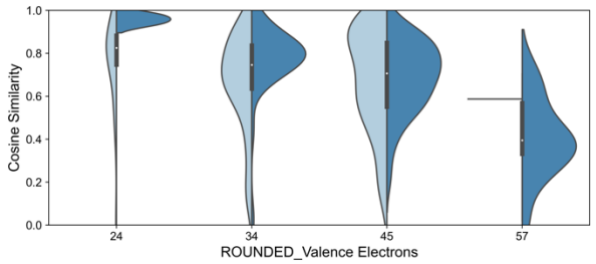
Mol logP (R)		-1	1	2	4
M+H	Mean	0.63	0.73	0.67	-
	Median	0.64	0.78	0.62	-
	StDev	0.24	0.21	0.23	-
M-H	Mean	0.74	0.72	0.46	0.33
	Median	0.75	0.78	0.40	0.30
	StDev	0.13	0.18	0.16	0.28



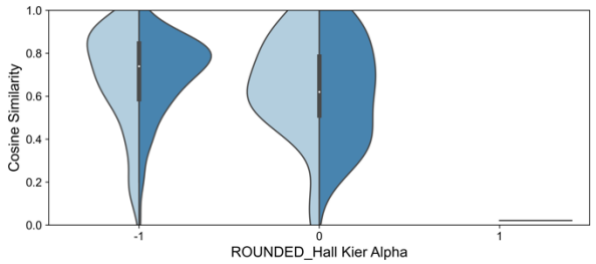
TPSA (R)		21	44	68	92
M+H	Mean	0.73	0.70	0.67	0.48
	Median	0.78	0.72	0.71	0.51
	StDev	0.19	0.23	0.23	0.21
M-H	Mean	0.41	0.73	0.73	0.65
	Median	0.39	0.78	0.76	0.67
	StDev	0.23	0.18	0.13	0.14



Labute ASA (R)		32	47	69	99
M+H	Mean	0.67	0.69	0.65	-
	Median	0.72	0.75	0.64	-
	StDev	0.21	0.24	0.23	-
M-H	Mean	0.80	0.70	0.45	0.33
	Median	0.80	0.71	0.39	0.26
	StDev	0.10	0.16	0.17	0.35



Valence Electrons (R)		24	34	45	57
M+H	Mean	0.75	0.66	0.68	0.59
	Median	0.82	0.71	0.73	0.59
	StDev	0.21	0.22	0.24	-
M-H	Mean	0.96	0.77	0.69	0.40
	Median	0.96	0.79	0.70	0.39
	StDev	0.03	0.16	0.16	0.18



Hall Kier Alpha (R)		-1	0	1
M+H	Mean	0.69	0.65	-
	Median	0.73	0.62	-
	StDev	0.23	0.24	-
M-H	Mean	0.70	0.63	0.02
	Median	0.75	0.67	0.02
	StDev	0.18	0.23	-

Descriptor	**Meaning**
Average Orbital	The average orbital is calculated by taking the average of the orbitals (1,2,3) of every atom in the molecule
Mol logP	Mol logP represents molecular lipophilicity, which is calculated by the octanol-water partition coefficient
TPSA	TPSA is the Topological Polar Surface Area
Labute ASA	Labute ASA is the Approximate Surface Area
Hall Kier Alpha	Represents the electrotopological-state of a molecule

Annex E: eV Data

M+H

eV range: 20,30,40,50,60,70

SMILES	Name
C[C@H](N)C(=O)O	L-Alanine
NCC(=O)O	Glycine
CN(C)C	Trimethylamine
Oc1ccccn1	2-pyridone
CC(C(=O)O)C(=O)O	Methylmalonic acid
NCCS(=O)O	Hypotaurine
c1ncc2[nH]cnc2n1	Purine
N[C@@H](CS)C(=O)O	L-Cysteine
NCCS(=O)(=O)O	Taurine
O=C(O)c1cccnc1	Niacin
Nc1ccnc(=O)[nH]1	Cytosine
C[C@H](O)C(=O)O	L-Lactic acid
Oc1ncnc2nc[nH]c12	Hypoxanthine
O=C1CCNC(=O)N1	Dihydrouracil
CNC(=N)N	Methylguanidine
O=C(O)c1ccccn1	Picolinic acid
NC1CCSC1=O	DL-Homocysteine thiolactone
O=CNCC(=O)O	N-Formylglycine
O=C(O)c1ccc[nH]1	Pyrrole-2-carboxylic acid
CC(=O)CC(=O)O	Acetoacetic acid

M-H

eV range: 10,25,40

CC(C(=O)O)C(=O)O	Methylmalonic acid
C[C@H](O)C(=O)O	L-Lactic acid
O=C(O)CC(=O)O	Malonic acid
O=C(O)CCC(=O)O	Succinic Acid
O=C(O)CC(=O)C(=O)O	Oxalacetic acid
O=C(O)/C=C/C(=O)O	Fumaric acid
O=C(O)[C@H](O)CO	D-Glyceric acid
O=C(O)/C=C\\C(=O)O	Maleic acid
O=C(O)c1ccco1	2-Furoic acid
CCCC(=O)O	Butyric Acid
O=C(O)CO	Glycolic acid
O=CC(=O)O	Glyoxylic acid
CCC(=O)C(=O)O	2-Oxobutanoic acid
C[C@@H](O)C(=O)O	D-Lactic acid
O=C(O)C(=O)CS	3-Mercaptopyruvic acid