Quantum chemistry based prediction of electron ionization mass spectra for environmental chemicals

Helge Hecht,[*,†] Wudmir Y. Rojas,[*,†] Zargham Ahmad,[*,†] Aleš Křenek,[*,‡] Jana Klánová,[*,†] and Elliott J. Price[*,†]

†*RECETOX, Faculty of Science, Masaryk University, Kotlarska 2, Brno 602 00, Czech Republic*

‡*Institute of Computer Science, Masaryk University, Botanická 554/68a, Brno 602 00, Czech Republic*

E-mail: helge.hecht@recetox.muni.cz; wudmir.rojas@recetox.muni.cz; zargham.ahmad@recetox.muni.cz; ljocha@ics.muni.cz; jana.klanova@recetox.muni.cz; elliott.price@recetox.muni.cz

Abstract

There is a lack of experimental electron ionization high-resolution mass spectra available to assist compound identification. The *in silico* generation of mass spectra by quantum chemistry can aid annotation workflows, in particular to support the identification of compounds that lack experimental reference spectra, such as environmental chemicals. We present an open-source, semi-automated workflow for the *in silico* prediction of electron ionization high-resolution mass spectra based on the QCxMS software. The workflow was applied to predict the spectra of 367 environmental chemicals and accuracy evaluated by comparison to experimental reference spectra acquired. The

molecular flexibility, number of rotatable bonds and number of electronegative atoms of a compound were negatively correlated with prediction accuracy. Few analytes are predicted to sufficient accuracy for the direct application of predicted spectra in spectral matching workflows. The m/z values of the top 5 most abundant ions of predicted spectra rarely match ions in experimental spectra, evidencing the disconnect between simulated fragmentation pathways and empirical reaction mechanisms.

Introduction

Electron ionization (EI) mass spectra are widely used for the structural annotation of small molecules through comparison of experimentally acquired spectra with mass spectral libraries. EI is the most common ionization method coupled with gas chromatography - mass spectrometry (GC-MS). However, there is a lack of EI high-resolution mass spectral libraries available, especially for environmental chemicals[1]. Additionally, spectra acquired using high-resolution time-of-flight (TOF) and Orbitrap instruments differ from those obtained through low-resolution single quadrupole instruments. The disparity leads to relatively lower match scores when comparing high-resolution spectra to nominal mass spectra, limiting the applicability of those libraries[2,3]. Compounding this challenge, many chemicals are not readily available commercially, and existing experimental libraries exhibit limited chemical diversity.

As an alternative approach, spectra can be generated *in silico* and used for identification. Methods like machine learning (ML) or quantum chemistry (QC) calculations are employed for this purpose[4]. For example, the Neural electron-ionization mass spectrometry (NEIMS) software uses a multi-layer perceptron (MLP) architecture to generate low-resolution EI mass spectra from molecular fingerprints[5]. The specific model presented in the paper was trained in a supervised manner on the NIST/EPA/NIH Mass Spectral Library 2017 database. State of the art ML methods, such as graph neural networks (GNNs) and transformers, have demonstrated superior performance compared to MLP based models[6] and have been applied in various ways for mass spectra prediction tasks[7–9]. In particular, Zhu and Jonas[10]

utilize a representation learning approach to predict the likelihood of subformulae within a predetermined depth of bond breakages. They then leverage these probabilities to scale the ion intensities in substructure spectra derived from isotopic patterns. Their study successfully predicted EI high-resolution mass spectra, including exact peaks, for molecules from PubChem, using an artificial high-resolution dataset. However, despite these advancements, ML approaches typically require substantial amounts of high-quality training data. Consequently, the broader applicability of ML methods is hindered by the scarcity of available EI high-resolution mass spectral libraries, primarily limiting their usage to the prediction of low-resolution spectra.

In contrast, generating mass spectra through QC calculations does not rely on empirical rules or experimental data. Such approaches offer insights into fragmentation processes and reaction mechanisms. The quantum chemical electron ionization mass spectrometry (QCEIMS) software[11], later renamed QCxMS after the addition of collision induced dissociation (CID) ionization and published as open-source[12], simulates the ionization and fragmentation process by employing QC principles to generate mass spectra *in silico*. Besides *ab initio* calculations[11,13], the package supports semiempirical quantum mechanical (SQM) methods[14,15] for increased throughput. SQM modeling provides ionization potentials for high-temperature molecular dynamics (MD) and mass spectra modeling. The simplicity in parametrization, using hybrid density functional theory (DFT) reference data, enhances SQM's accessibility and efficiency. Despite potential electron delocalization overestimation, SQM excels in handling metallic systems and covalent bond dissociation, but requires corrections for optimal performance in directional electrostatic effects like halogen bonds. A standout feature is SQM's easily adaptable, element-specific parameters, enriching method versatility. Whereas DFT remains the most accurate, SQM-based MD is more than 100 times faster than pure DFT-based MD.

Several studies evaluate the performance of the QCxMS package and related SQM methods across multiple compound categories (Table 1). These include small datasets of (i) or-

ganic and inorganic small molecules used for method development validations[16,17], (ii) a restricted set of pollutants consisting of 27 halogenated compounds and 8 organophosphorus flame retardants (ODTs)[18] and (iii) purines and pyrimidines[19]. Notably, QCxMS has been recently applied to two larger datasets consisting of (iv) 451 small organic molecules[20] and (v) 816 trimethylsilylated analytes[21].

Table 1: Overview of related studies using QCxMS with SQM.

Study	Classes	Molecules	Masses	Method	Reference	Atom Types
This Study	56	367	108 - 715	GFN1-xTB	In house	C H O N F P S Cl Br Si
Wang et al.[20]	43	451	26 - 358	OM2	NIST17	C H N O
Wang et al.[17]	NA	41	55 -333	OM2/CMSID + GFN1-xTB	NIST17	C H N O F
Wang et al.[21]	10	816	115 - 299	GFN1-xTB	NIST17	C H N O Si
Schreckenbach et al.[18]	NA	35	74 - 959	GFN1-xTB	NIST	C H N O P Cl Br
Ásgeirsson et al.[16]	NA	23	86 - 505	GFN1-xTB	NIST/SDBS	C H N O P B Sb S Cl Bs Ge Te Ni Cu Cr Fe F Al Si Sn
Lee et al.[19]	12	80	120 - 207	OM2/OM3	NIST17	C H N O

These studies demonstrate QCxMS's ability to predict EI mass spectra, but the chemical space covered in the larger studies (containing $\geq$ 100 molecules) is limited to C, H, O, N and Si atoms and molecular weight $\leq$ 400 dalton (Da). Notably, previous studies do not compare the predicted spectra against experimental high-resolution spectra acquired from analytical standards but against low-resolution commercial libraries, limiting the assessment of predictive accuracy.

Concerning the expansion of QCxMS applicability, it is crucial to highlight its limited use in investigating large diverse datasets including e.g., environmental chemicals[24] (Table 2). In the present paper, we introduce an open-source workflow for large-scale prediction of EI spectra, leveraging the QCxMS software and using the extended tight-binding SQM GFN1-xTB[24] method. We demonstrate the workflow through application to predict EI

Table 2: Summary of molecular properties investigated in comparison to previous work. Row-wise maxima are printed in bold.

		This study	Wang et al. [20]	Wang et al. [17]	Wang et al. [21]	Schreckenbach et al. [18]	Ásgeirsson et al. [16]	Lee et al. [19]
atoms	mean	33.01	21.95	22.82	**33.67**	31.03	19.76	18.28
	min	12	7	8	**17**	11	6	12
	max	**80**	59	56	58	74	49	24
aromatic nitrogens	mean	0.45	0.06	0	0.13	0.03	0	**1.39**
	min	**0**	**0**	**0**	**0**	**0**	**0**	**0**
	max	3	3	0	**4**	1	0	**4**
molecular complexity	mean	0.73	0.47	0.47	0.53	0.59	0.27	**0.76**
	min	0.38	0.12	0.27	0.26	0.35	0	**0.67**
	max	**1.18**	0.80	0.77	0.84	0.85	0.69	0.82
molecular flexibility	mean	0.36	0.33	0.30	**0.63**	0.51	0.40	0.04
	min	0	0	0	**0.22**	0	0	0
	max	0.85	0.86	0.69	**0.91**	0.90	0.89	0.23
rotatable bonds	mean	3.26	2.65	1.70	**4.41**	4.22	1.43	0
	min	0	0	0	**1**	0	0	0
	max	**21**	10	8	14	16	5	0
stereo centers	mean	**0.75**	0.66	0.38	0.33	0.59	0.29	0
	min	**0**	**0**	**0**	**0**	**0**	**0**	**0**
	max	**9**	6	3	4	8	6	0
electronegative atoms	mean	4.74	1.78	1.32	2.40	**5.94**	1.76	5.49
	min	0	0	0	1	2	0	**4**
	max	**14**	8	5	6	12	8	8

Molecular properties computed using DataWarrior [22]. Any molecules failing computation were excluded from the comparison. Structure databases have been generated from the respective publications and additional chemical identifier were collected using MSMetaEnhancer [23].

mass spectra for a previously published set of environmental compounds [1]. Furthermore, we then investigate the applicability of this methodology on our chemically diverse dataset by comparing the spectra to the accompanying high-resolution EI mass spectral library [25] and outline aspects influencing prediction accuracy.

Methods

Workflow The developed semi-automated workflow for predicting mass spectra integrates various functions to streamline the process, significantly enhancing researchers' capacity to conduct high-throughput mass spectra simulations. To the best of our knowledge, this addresses a deficiency in available open-source end-to-end processing solutions. It includes

approaches such as managing input file preparation through templates and a global parameter file for every stage of the workflow on a high-performance computing (HPC) cluster as well as accompanying scripts for batch job submission, progress monitoring, result collection and analysis.

Starting with the structure data file (SDF) input, the workflow progresses through these steps: (i) Creating and Structuring Files: Arranging files for molecular geometry optimization with GAMESS[26], handling molecule reading, directory creation, 3D conformation generation, input file writing, and optimization job script setup, (ii) Molecular Optimization: Dispatches optimization tasks to the HPC cluster and gathers optimized molecular inputs for spectral simulations, (iii) QCxMS Spectral Prediction: Submits neutral MD, prepares production runs, and executes them on the HPC cluster using batch job processing, (iv) MSPs Generation and Analysis: Retrieves spectral results and formats them into `msp` file.

Additionally, the workflow includes post-processing tools for summarizing simulations and removing unnecessary files. It offers visualization and analysis tools for enhanced usability. A high-level overview of this workflow is illustrated in Figure 1. Implemented with Bash and Python, the software repository, including statistical analysis code, is archived on Zenodo[27]. Computations were conducted on the metacentrum HPC cluster (MetaCloud; `https://www.metacentrum.cz/en/cloud/`).

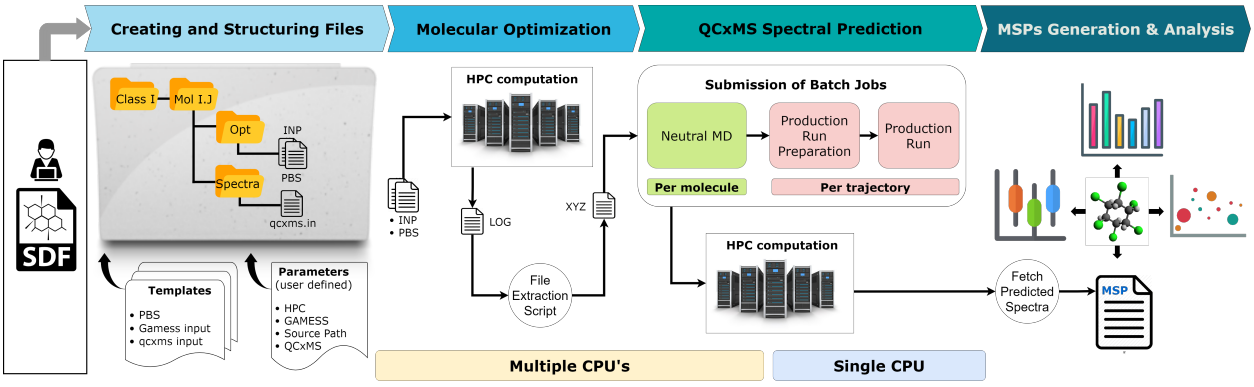


Figure 1: Semi-automatic EI spectra prediction workflow. From an SDF input, the workflow creates all input files required for batch spectral prediction.

Dataset The RECETOX Exposome HR-[EI+]-MS library[25] was used as data source. The library contains 56 distinct chemical classes, covering molecules ranging from 12 to 80 atoms, representing a diverse range of compounds in terms of both size and elemental composition[1] (Table 2). Isotopically labeled analytes were removed from the library as those are not handled by QCxMS, leaving 367 structures.

3D Conformer generation Molecular structures and other metadata, e.g., ChemOnt chemical class[28], were extracted from the RECETOX Exposome HR-[EI+]-MS library[25] in SDF using the RDKIT package[29]. Subsequently, three-dimensional conformers were created from simplified molecular input line entry system (SMILES) descriptors using a customized variant of the rdconf code (see `https://github.com/dkoes/rdkit-scripts/blob/master/rdconf.py`), which leverages RDKIT functions while employing the universal force field (UFF) method. Molecular structures were optimized using GAMESS at the 6-31G level of theory and gradient convergence tolerance of 5×10^{-4} prior to QCxMS analysis.

Spectra prediction using QCxMS QCxMS calculations were performed using default parameters (electron energy: 70 eV; excess energy: 0.6 eV/atom; initial temperature: 500 K; simulation time: 10 ps). The ground state trajectories and production runs were performed using the SQM GFN1-xTB method. After fragmentation, the ionization energies (IEs) were calculated using GFN1-xTB, and the fragment with the lowest IE obtained the positive charge. Default number of trajectories were computed (25x the number of atoms in the molecule) for every molecule.

Post-processing Results were collected and converted into a library in the `msp` format and the predicted spectra underwent further filtering using the matchms[30] package. Filtering steps were tailored to ensure reliable comparison between predicted spectra and reference spectra: (i) the m/z values were restricted to the range of 70 to 700, (ii) peak intensities were normalized to the peak of maximum intensity, (iii) peaks with intensities below 1% of the

maximum peak intensity were removed. A second dataset containing the top 5 intensity ions from both predicted and experimental spectra was created because in practice, only minimal spectral information is often available for annotation of compounds in experimental datasets. Chemical properties related to atomic composition and molecular structure were computed using DataWarrior[22] (see Table 2). Statistical analysis was performed using scipy[31] and pandas[32].

Spectral matching Spectral matching was performed using matchms. The *CosineHungarian* score was used with a tolerance of 0.0035 Da (i.e., 5 ppm at 700 m/z), intensity power of 1 and m/z power of 0. In spectral similarity computation, only peaks that fall within the specified tolerance are taken into account. After this filtering process, spectrum vectorization can be understood as the intersection between the compared spectra. As matchms does not retain entries with 0 matching ions in the outputs, these were systematically integrated into the score tables, assigning scores and match values of zero for pairs missing in the output files. The workflow for spectral matching and related datasets are available online[33,34].

Results

Geometry optimization A subset of 48 molecules did not converge using the previously described geometry optimization parameters[27]. This is most likely due to the generated conformer resulting in an unfavorable initial configuration for geometry optimization. The computational framework allows users to adjust key parameters such as basis sets, gradient convergence tolerance, and the maximum number of steps to address convergence challenges. However, our primary focus at this stage is to enable the generation of molecular structures for spectral predictions, rather than achieving high-level molecular optimization. Therefore, users also have the flexibility to employ alternative tools for refining these molecular structures. In this context, we leveraged the xTB code[35] to address the molecular optimization of the aforementioned molecules.

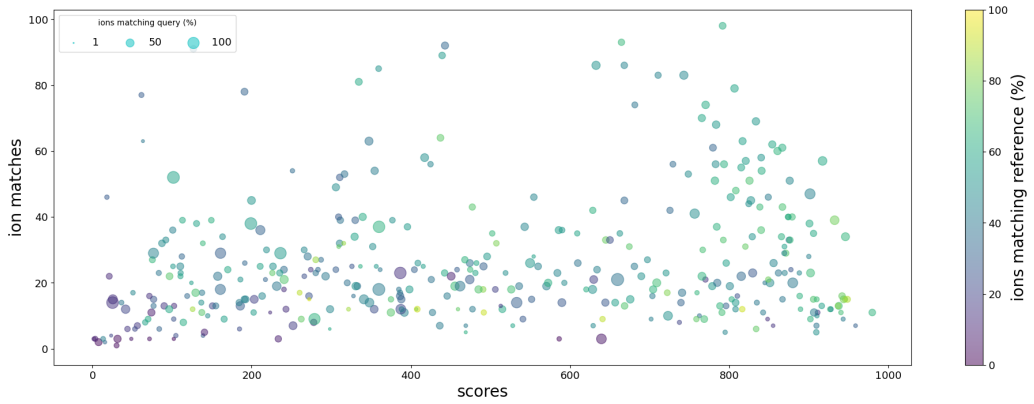


Figure 2: Cosine similarity scores (x-axis) versus the absolute number of matched ions (y-axis) for predicted and experimental spectra. The color scale and size represent ion matches in % normalized by the number of peaks contained in the experimental and predicted spectrum, respectively.

Accuracy of predicted spectra A comparison between the predicted and experimentally acquired spectra, based on intensity ratio similarity (cosine score) and the number of matching ions, does not show any direct trend indicating the failure of the method to achieve meaningful predictions, with the scores and ion matches heterogeneously distributed (Figure 2).

Physicochemical properties Investigating the correlations of spectral matching scores and computed chemical properties reveals a weak positive correlation between cosine similarity score and number of ion matches. Furthermore, there is a negative correlation between scores and molecular flexibility, the number of rotatable bonds, the number of atoms and number of electronegative atoms. These properties are co-correlated. The number of matching ions follows a similar trend. However, it shows weaker correlation with the number of atoms. Notably, molecular complexity does not correlate with the cosine similarity score, nor the number of matched ions (Figure 3).

Chemical class Comparisons were performed based on ChemOnt[28] chemical class hierarchies to investigate prediction accuracy amongst structurally-related groups (Figure 4).

At the superclass level, only the benzenoids and organoheterocyclic compounds have a

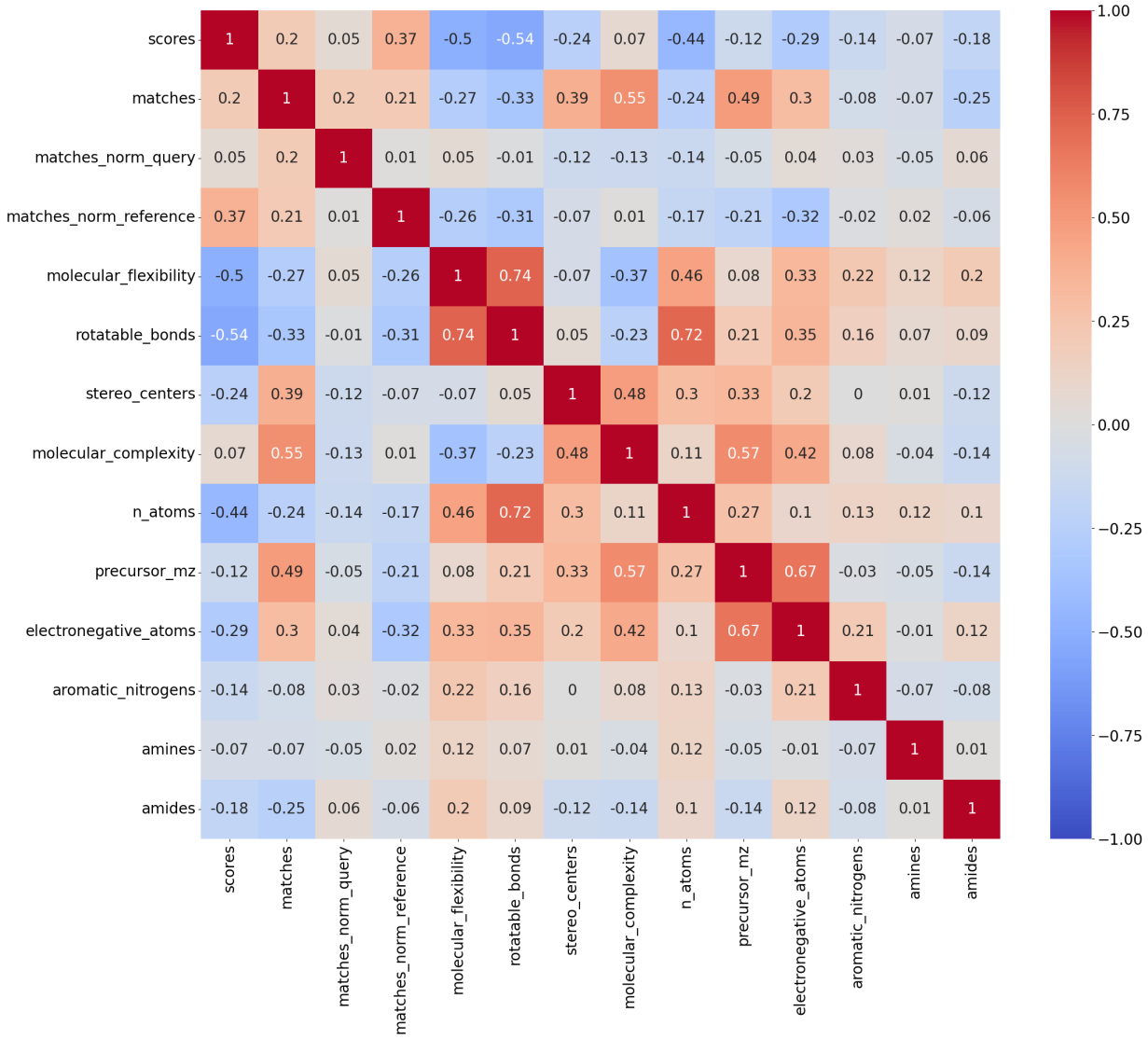
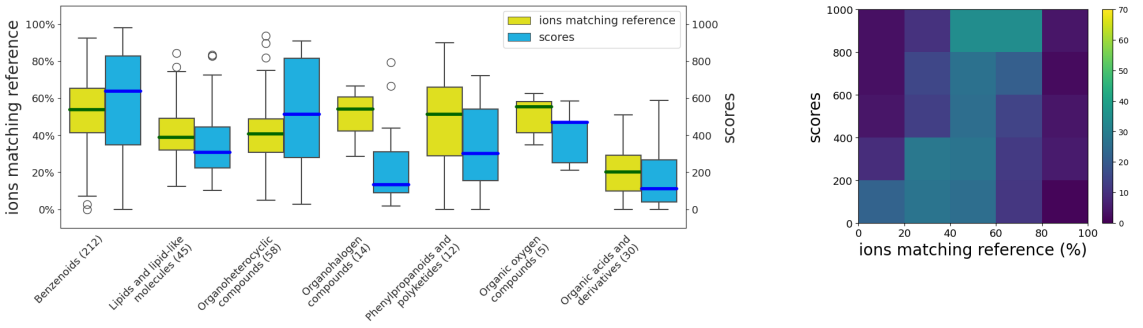
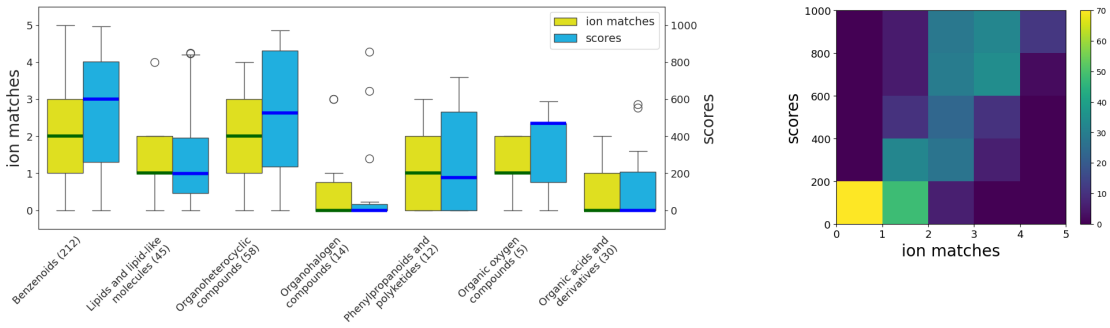


Figure 3: Pearson's correlations of chemical properties and spectral matching results. Chemical properties computed using DataWarrior. Molecular complexity is used as an aggregate measure[36]. The number of ion matches is given as absolute number as well as normalized by the number of peaks in the predicted and reference spectra.



(a) Predicted spectra containing all peaks. Ion matches are normalized against the number of peaks contained in the experimentally acquired spectra.



(b) Predicted spectra filtered for top 5 ions

Figure 4: Spectral matching of predicted to experimental spectra at the superclass level. The color scale for counts has been aligned across histograms.

median spectral cosine score greater than 500 and an interquartile range (IQR) exceeding 800. However, the prediction accuracy for lipid and lipid-like molecules, organohalogen compounds, phenylpropanols and polyketides as well as for organic acids and derivatives, is comparatively lower, with all having median scores below 400. The lipid and lipid-like molecules are characterized by a high number of rotatable bonds (on average ≥ 7) and contain additional atoms beyond C, O and H such as P and S. The organic acids superclass shows the lowest median cosine similarity score of 110.5, encompassing most S and P containing molecules in the dataset. Low spectral similarity scores have previously been demonstrated for ODTs[18]. Particularly for those structures containing S and P, the SQM methods show poor accuracy in simulating fragmentation pathways.

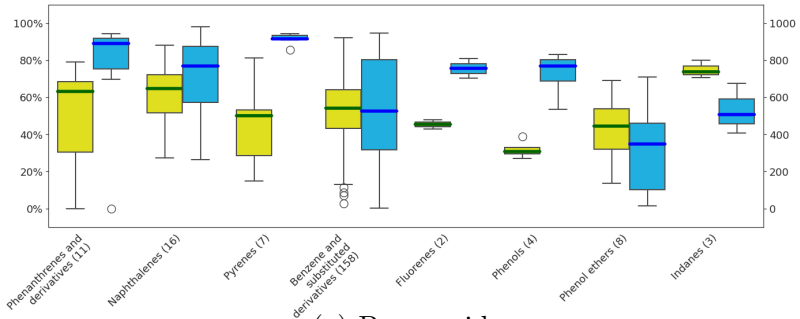
Overall, when considering only the top 5 intensity ions the spectral match between pre-

dicted and experimental spectra is even lower (Figure 4b). For 71 spectra, not a single of the top 5 experimentally measured ions is predicted at the correct m/z. In practice, three or more ions are often used to corroborate annotation, yet three or more ions are only predicted correctly for 95 of the 367 molecules.

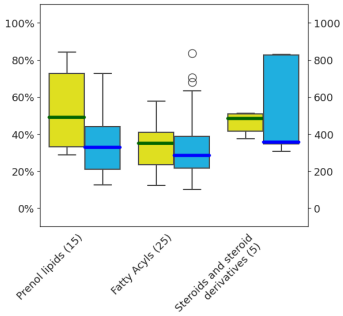
On the class level (see Figure 5), two groups stand out for their high prediction accuracy: (i) phenanthrenes and pyrenes (benzenoids), and (ii) benzofurans, benzodioxins and benzimidazoles (organoheterocyclic compounds), totaling 34 molecules. These molecules, characterized by their relatively planar geometry (molecular flexibility ≤ 0.4), were reliably predicted with median scores ≥ 800. However, certain individual classes showed lower prediction performance, indicated by median scores ≤ 350. These include (i) phenol ethers (benzenoids), (ii) alkyl halides (organohalogen compounds) and (iii) azoles, benzothiazoles and dizianaphthalenes (organoheterocyclic compounds).

In comparison to other benzenoids, the phenol ethers in the dataset consistently exhibit at least 3 rotatable bonds, with some also containing aromatic nitrogens or being halogenated. Similarly, the organoheterocyclic classes possess a higher count of aromatic nitrogens and rotatable bonds (both ≥ 2), along with electronegative atoms (≥ 5). Additionally, Benzothiazoles, which contain S, are not effectively characterized by the SQM methods.

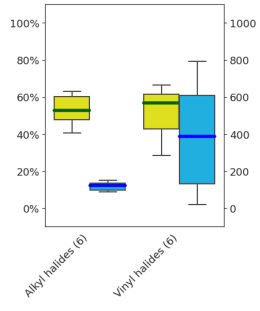
As benzenes and substituted derivatives consist of 158 molecules, we further inspected those at the subclass level, as depicted in Figure 6. Biphenyls and derivatives constitute the largest subclass, with 41 molecules, and are most accurately predicted, with a median cosine similarity above 800. The outliers such as bitertanol (3,3-dimethyl-1-(4-phenylphenoxy)-1-(1,2,4-triazol-1-yl)butan-2-ol) and bifenazate (propan-2-yl N-(2-methoxy-5-phenylanilino)carbamate), each having 6 rotatable bonds, are observed within this subclass. Accordingly, subclasses containing molecules with greater structural flexibility, such as phenyl methylcarbamates, phenyl methylcarbamic acids and diphenylethers are predicted less accurately.



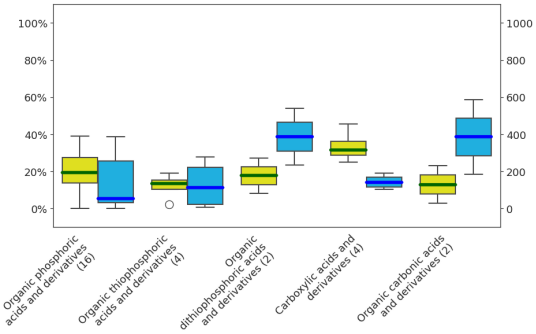
(a) Benzenoids



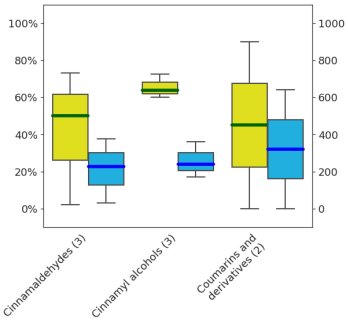
(b) Lipids and lipid-like molecules



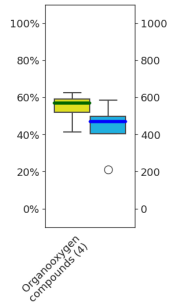
(c) Organohalogen compounds



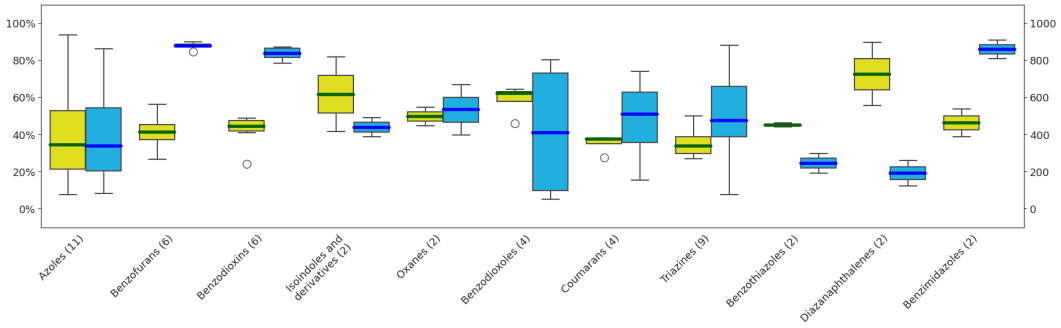
(d) Organic acids and derivatives



(e) Phenylpropanoids and polyketides



(f) Organic oxygen compounds



(g) Organoheterocyclic compounds

Figure 5: Boxplot of spectral comparisons between predicted and reference EI mass spectra per chemical class within each superclass. The bar denotes the median while outliers are depicted as circles with spectral similarity score (blue, right y-axis) and absolute number of matched ions (yellow, left y-axis) displayed. Superclasses and classes represented only by a single molecule are excluded.

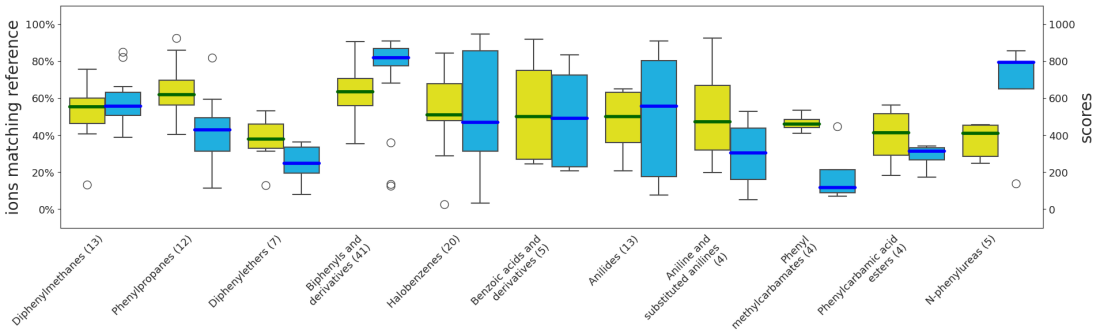


Figure 6: Boxplots for subclasses of benzene and substituted derivatives. The bar denotes the median while outliers are depicted as circles. Subclasses with less than 3 molecules have been removed from the figure.

Elemental composition The class-based analysis reveals a trend that spectra prediction of molecules containing certain atom types, especially aromatic nitrogens present in compounds like azoles, as well as the presence of P, is less accurate. Therefore, in addition to the class-based analysis, we also analyzed the spectra matching results regarding the elemental composition of each molecule.

To isolate the influence resulting from the presence of N in the molecule, we directly compared molecules containing the same atom types but with and without N (Figure 7). Median scores and ion match rates are lower for every group of molecules, except for those containing the chemical compositions (i) Br,C,H,(N),O; (ii) C,H,(N),O,S; and (iii) C,Cl,H,(N),O,S, when N is part of the chemical composition.

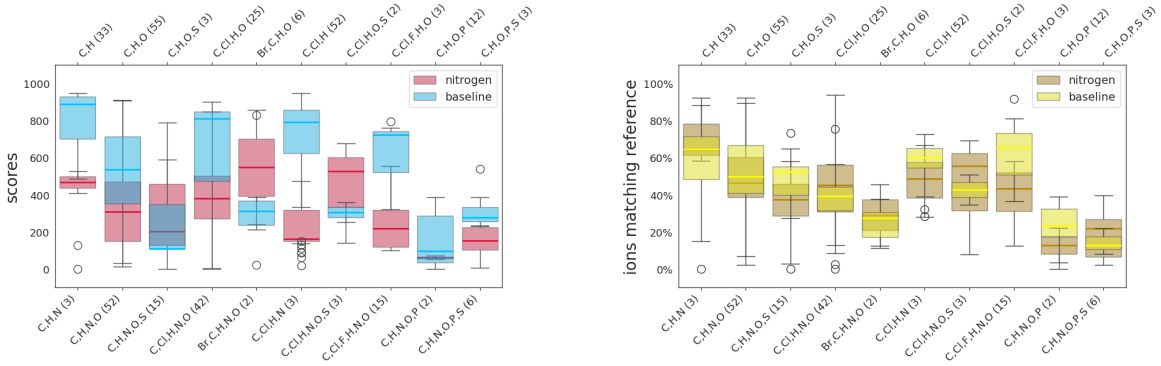


Figure 7: Comparison of spectra prediction accuracy for molecules of selected chemical compositions based on the presence of N. The selected groups molecules contain all of the form X →X + N present in the dataset.

Similarly, the presence of P negatively influences prediction accuracy using our chosen methodology (Figure 8). P often serves as a central atom, resulting in a more flexible structure. Within our dataset, P containing molecules have an average number of rotatable bonds of 8.4 over 2.9 for all other molecules respectively. This is further supported by the negative correlation of the rotatable bond count and cosine similarity and number of matching ions (Figure 3).

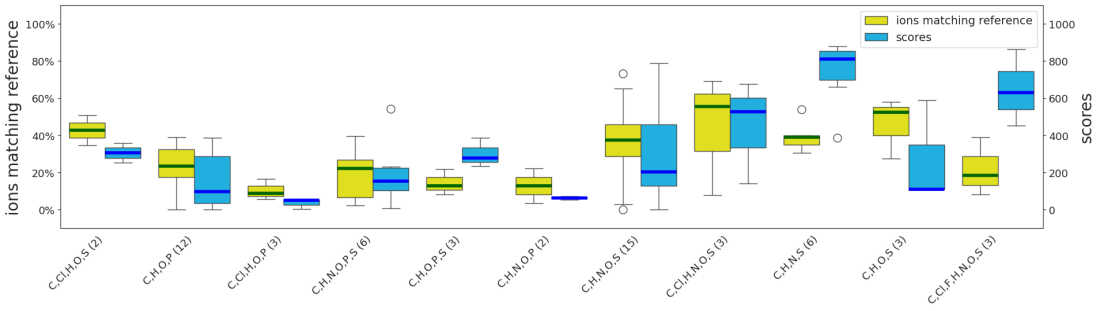


Figure 8: Results for molecule groups containing P and/or S atoms. Groups consisting only of a single molecule have been removed from the boxplot.

Discussion

Our results show that performance of the QCxMS and xTB methods vary significantly across the diverse dataset. Only few chemical groups are predicted with sufficient accuracy (e.g., pyrenes, phenanthrenes, benzofuranes, benzodioxins and benzimidazoles and biphenyls and derivatives, totaling 75 molecules) for direct use in spectral matching based annotation workflows i.e. average spectral similarity scores above 800. In particular, for the vast majority of analytes, even the most abundant ions were not accurately predicted i.e. wrong m/z, limiting the current potential to incorporate predicted spectra into suspect annotation workflows. The presence of atoms other than C, O and H leads to less accurate spectral prediction, as can be seen on the results of P and/or S containing molecules depicted in Figure 8 and the comparison of N containing groups and their respective N-free counterparts in Figure 7. Additionally, the extensive presence of electronegative atoms such as halogens (e.g., in alkyl

and vinyl halides, see Figure 5c) negatively affect the simulations, as halogen bonds typically require atom pairwise corrections[24]. We repeated predictions of the alkyl halides using the GFN2-xTB method that models anisotropic electrostatic interactions and does not employ specific element pair corrections. Though it reportedly improves predictions of molecules containing bonds between molecules with large differences in electronegativity[37], we did not observe any improvement in spectra quality for our subset[38]. Beyond chemical composition, the related three dimensional molecular structure is the main determinant. While planar geometries are predominantly predicted with high accuracy, central atoms with many rotatable bonds (e.g., P in organic acids, see Figure 5d) and other stereo centers leads to lower prediction accuracy for the QCxMS method.

The observed limitations with regard to the chemical space for which accurate predictions are possible hinders application for many parent environmental chemicals which are enriched with halogens and/or S compared to endogenous analytes. In addition, the spectral prediction for many metabolites is likely to also be poor due to their higher proportion of P and rotatable bonds[39]. However, previous investigations to predict spectra of endogenous analytes rarely included P (Table 1). Simply tuning simulation parameters may not significantly enhance predictions, instead, it is proposed that accurate calculation of potential energy surfaces or the incorporation of excited-state MD may enhance accuracy within the semi-empirical framework[40]. Recently, Wang et al.[17] incorporated excited states into the MD steps of EI spectra using the binary-encounter-Bethe (BEB) model, achieving higher accuracy, although limitations still exist, particularly for non-organic molecules.

The semi-empirical xTB method enables systematic processing of larger sets of molecules, though true *high throughput* processing would require a speedup in several orders of magnitude. Including initial testing, we scheduled $524,334$ compute jobs with a total usage of $43,201$ central processing unit (CPU) days on our HPC cluster. The files containing logs, structures and computed trajectories require $\sim$ 2TB storage space. Considering that such infrastructure and capacity might not be generally available to researchers further highlights

the need for more flexible, accessible and efficient computational methods[41,42]. For similar reasons, current works are limited to the use of SQM over DFT based methods. Future work should consider the application of such *ab initio* MD for molecules not well characterized by the SQM based methods.

Conclusions

We present an open-source workflow for prediction of high-resolution EI mass spectra and performance assessment based on the QCxMS, matchms, RDKIT, xTB and GAMESS software packages. Additionally, we provide the largest set of predicted mass spectra of environmental chemicals so far and perform an unbiased analysis based on structural taxonomy classifications and chemical element composition. Our results show that further developments of SQM-based MD to improve prediction accuracy for molecules containing electronegative atoms (e.g., halogens, nitrogen) with high molecular flexibility (e.g., multiple rotatable bonds and stereo centers) are crucial. Current methods are insufficient for practical suspect screening applications which require at least 3 characteristic (i.e., highly abundant) ions for correct identification, as these conditions were not met for the vast majority of predicted spectra when compared with experimental high-resolution spectra. Furthermore, improvements with regard to computational efficiency and accessibility are essential to advance the field of QC based *in silico* mass spectra prediction. This especially holds true considering the need for additional comprehensive large scale studies which are required to characterize the applicable chemical space and potential pitfalls of these methodologies.

Associated content

Data Availability statement

All data and scripts used in this work are hosted and archived on Zenodo[27]. The code repository containing the workflow as well as the scripts used to generate tables and figures is publicly available at `https://github.com/RECETOX/ei_spectra_predictions`. Note that this repository is subject to further development, so please refer to the Zenodo archive for information specifically related to this publication.

Author Contributions

H.H. - Data curation; Formal analysis; Software; Visualization; Writing – original draft; Writing – review and editing. W.Y.R. - Data curation; Formal analysis; Investigation; Software; Writing – original draft; Writing – review and editing. Z.A. - Data curation; Software; Visualization; Writing – review and editing. A.K. - Resources; Writing – review and editing. J.K. - Funding acquisition; Writing – review and editing. E.J.P. - Conceptualization; Formal analysis; Resources; Supervision; Writing – original draft; Writing – review and editing.

Acknowledgement

H.H., W.Y.R., Z.A., J.K. and E.J.P thank the RECETOX Research Infrastructure (LM2023069) financed by the Ministry of Education, Youth and Sports, and the Operational Programme Research, Development and Education (the CETOCOEN EXCELLENCE project No. CZ.02.1.01/0.0/0.0/17 _043/0009632) for supportive background. This work was supported from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857560. Computational resources were provided by the project e-INFRA CZ (LM2018140). This presentation reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains. The authors acknowledge the support of the Freiburg Galaxy Team and Björn Grüning, Bioinformatics, University of Freiburg (Germany) funded by the Collaborative Research Centre 992 Medical Epigenetics (DFG grant SFB 992/1 2012) and the German Federal Ministry of Education and Research BMBF grant 031 A538A de.NBI-RBC.

Acronyms

BEB binary-encounter-Bethe. 16

CID collision induced dissociation. 3

CPU central processing unit. 16

Da dalton. 4, 8

DFT density functional theory. 3, 17

EI electron ionization. 2–6, 16, 17

GC-MS gas chromatography - mass spectrometry. 2

GNN graph neural network. 2

HPC high-performance computing. 6, 16

IE ionization energy. 7

IQR interquartile range. 11

MD molecular dynamics. 3, 6, 16, 17

ML machine learning. 2, 3

MLP multi-layer perceptron. 2

NEIMS neural electron-ionization mass spectrometry. 2

ODT organophosphorus flame retardant. 4, 11

QC quantum chemistry. 2, 3, 17

QCEIMS quantum chemical electron ionization mass spectrometry. 3

SDF structure data file. 6, 7

SMILES simplified molecular input line entry system. 7

SQM semiempirical quantum mechanical. 3, 4, 7, 11, 12, 17

TOF time-of-flight. 2

UFF universal force field. 7

References

(1) Price, E. J.; Palát, J.; Coufaliková, K.; Kukučka, P.; Codling, G.; Vitale, C. M.; Koudelka, Š.; Klánová, J. Open, High-Resolution EI+ Spectral Library of Anthropogenic Compounds. *Frontiers in Public Health* **2021**, *9*.

(2) Stettin, D.; Poulin, R. X.; Pohnert, G. Metabolomics Benefits from Orbitrap GC–MS—Comparison of Low- and High-Resolution GC–MS. *Metabolites* **2020**, *10*, 143.

(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) Krettler, C. A.; Thallinger, G. G. A map of mass spectrometry-based in silico fragmentation prediction and compound identification in metabolomics. *Briefings in Bioinformatics* **2021**, *22*, 1–25.

(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*, 700–708.

(6) Zhu, H.; Liu, L.; Hassoun, S. Using Graph Neural Networks for Mass Spectrometry Prediction. *arXiv* **2020**,

(7) Young, A.; Wang, B.; Röst, H. MassFormer: Tandem Mass Spectrum Prediction for Small Molecules using Graph Transformers. **2021**, 1–14.

(8) Murphy, M.; Jegelka, S.; Fraenkel, E.; Kind, T.; Healey, D.; Butler, T. Efficiently predicting high resolution mass spectra with graph neural networks. **2023**, 1–18.

(9) Goldman, S.; Bradshaw, J.; Xin, J.; Coley, C. W. Prefix-tree Decoding for Predicting Mass Spectra from Molecules. **2023**,

(10) Zhu, R. L.; Jonas, E. Rapid Approximate Subset-Based Spectra Prediction for Electron Ionization-Mass Spectrometry. *Analytical Chemistry* **2023**, *95*, 2653–2663.

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

(12) 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*, 1735–1751.

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

(14) Bauer, C. A.; Grimme, S. First principles calculation of electron ionization mass spectra for selected organic drug molecules. *Org. Biomol. Chem.* **2014**, *12*, 8737–8744.

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

(16) 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*, 4879–4895.

(17) 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*, 4403–4410.

(18) Schreckenbach, S. A.; Anderson, J. S.; 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. *Journal of the American Society for Mass Spectrometry* **2021**, *32*, 1508–1518.

(19) Lee, J.; Kind, T.; Tantillo, D. J.; Wang, L.-P.; Fiehn, O. Evaluating the Accuracy of the QCEIMS Approach for Computational Prediction of Electron Ionization Mass Spectra of Purines and Pyrimidines. *Metabolites* **2022**, *12*.

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

(21) 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*, 1559–1566.

(22) Sander, T.; Freyss, J.; von Korff, M.; Rufener, C. DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis. *Journal of Chemical Information and Modeling* **2015**, *55*, 460–473.

(23) Troják, M.; Hecht, H.; Čech, M.; Price, E. J. MSMetaEnhancer: A Python package for mass spectra metadata annotation. *Journal of Open Source Software* **2022**, *7*, 4494.

(24) 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*, 1989–2009.

(25) Price, E. J.; Palát, J.; Coufaliková, K.; Kukučka, P.; Codling, G.; Vitale, C. M.; Koudelka, Š.; Klánová, J. RECETOX Exposome HR-[EI+]-MS library. 2021; `https://doi.org/10.5281/zenodo.4471217`.

(26) Barca, G. M. J. et al. Recent developments in the general atomic and molecular electronic structure system. *The Journal of Chemical Physics* **2020**, *152*, 154102.

(27) Rojas, W. Y.; Hecht, H.; Ahmad, Z. RECETOX/ei_spectra_predictions: v0.3. 2024; `https://doi.org/10.5281/zenodo.10853686`.

(28) 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*, 61.

(29) Landrum, G. et al. rdkit/rdkit: 2023_09_3 (Q3 2023) Release. 2023; `https://doi.org/10.5281/zenodo.10275225`.

(30) Huber, F.; Verhoeven, S.; Meijer, C.; Spreeuw, H.; Castilla, E.; Geng, C.; van der Hooft, J.; Rogers, S.; Belloum, A.; Diblen, F.; Spaaks, J. matchms - processing and similarity evaluation of mass spectrometry data. *Journal of Open Source Software* **2020**, *5*, 2411.

(31) Virtanen, P. et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **2020**, *17*, 261–272.

(32) The pandas development team pandas-dev/pandas: Pandas. `https://github.com/pandas-dev/pandas`.

(33) Y. Rojas, W. RECETOX Spectral Similarity Top 5 Peaks Galaxy Workflow and History. 2024; `https://doi.org/10.5281/zenodo.10842560`.

(34) Y. Rojas, W. RECETOX Spectral Similarity All Peaks Galaxy Workflow and History. 2024; `https://doi.org/10.5281/zenodo.10842462`.

(35) Bannwarth, C.; Caldeweyher, E.; Ehlert, S.; Hansen, A.; Pracht, P.; Seibert, J.; Spicher, S.; Grimme, S. Extended tight-binding quantum chemistry methods. *WIREs Computational Molecular Science* **2021**, *11*, 1–49.

(36) von Korff, M.; Sander, T. About Complexity and Self-Similarity of Chemical Structures in Drug Discovery. Chaos and Complex Systems. Berlin, Heidelberg, 2013; pp 301–306.

(37) 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. *Journal of Chemical Theory and Computation* **2019**, *15*, 1652–1671.

(38) Hecht, H. QCxMS prediction of alkyl halides comparison of GFN1-xTB and GFN2-xTB. 2024; `https://doi.org/10.5281/zenodo.10839047`.

(39) Khanna, V.; Ranganathan, S. Physiochemical property space distribution among human metabolites, drugs and toxins. *BMC Bioinformatics* **2009**, *10*, 1–18.

(40) 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*, 2215–2287.

(41) Sarojini, D.; Burrows-Schilling, C.; Thomas, K.; Mizumoto, C. Towards Developing a Guide to Choosing National High-Performance Computing Resources. Practice and Experience in Advanced Research Computing. New York, NY, USA, 2023; pp 382–385.

(42) IDC *High Performance Computing in the EU: Progress on the Implementation of the European HPC Strategy*; 2015; p 137.