

Article

Not peer-reviewed version

Simplistic Software for Analyzing Mass Spectra and Mixed Experimental-Theoretical Database for Identifying Poisonous and Explosive Substances

[Denis S. Tikhonov](#)^{*}, [Mikhail A. Kalinin](#), Alexander A. Maryewski, Aleksandr A. Avdoshin, Olgert Dallakyan, [Nikita A. Vasilev](#), Egor A. Eliseev, Mandy Koch, [Vladimir V. Rybkin](#)^{*}, [Denis G. Artiukhin](#)^{*}

Posted Date: 10 February 2025

doi: 10.20944/preprints202502.0623.v1

Keywords: mass spectra; substance identification; molecular dynamics; database; metric



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Article

Simplistic Software for Analyzing Mass Spectra and Mixed Experimental-Theoretical Database for Identifying Poisonous and Explosive Substances

Denis S. Tikhonov ^{1,2,*} , Mikhail A. Kalinin ³ , Alexander A. Maryewski ⁴ , Aleksandr A. Avdoshin ⁵ , Olger Dallakyan ⁵ , Nikita A. Vasilev ⁶ , Egor A. Eliseev ⁶ , Mandy Koch ³, Vladimir V. Rybkin ^{7,*}  and Denis G. Artiukhin ^{8,*} 

¹ Deutsches Elektronen-Synchrotron DESY, Notkestr. 85, 22607 Hamburg, Germany

² Free Moscow University, www.freemoscw.university

³ Institute of Chemistry, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany

⁴ Institute of Physical Chemistry (IPC), Karlsruhe Institute of Technology, Kaiserstrasse 12, 76131 Karlsruhe, Germany

⁵ Institute of Catalysis Research and Technology (IKFT), Karlsruhe Institute of Technology, Eggenstein- Leopoldshafen, 76344, Germany.

⁶ Independent researcher

⁷ HQS Quantum Simulations GmbH, Rintheimer Straße 23, 76131 Karlsruhe, Germany

⁸ Institut für Chemie und Biochemie, Freie Universität Berlin, Arnimallee 22, 14195 Berlin, Germany

* Correspondence: denis.tikhonov@desy.de; vladimir.rybkin@quantumsimulations.de; denis.artiukhin@fu-berlin.de

Abstract: A recent increase in targeted attacks using chemical warfare agents by dictators and authoritarian regimes against politicians, journalists, and other civilians is a major concern. To aid the civil investigators in identifying poisonous substances in such cases, we developed an algorithm and a lightweight and simple-to-use software, ToxicMassSceptic, with a database of 394 mass spectra entries, which include many poisonous and explosive agents. The identification relies on a window-based reduction of the experimental spectra and four statistical metrics that are combined into a single metametric. The software also features automatic spectral background removal. Furthermore, we provide the workflow for increasing the size of this database by performing theoretical calculations of mass spectra with a molecular dynamics-based approach. The accuracy of both the theoretical prediction workflow and ToxicMassSceptic is validated on the experimental spectra. Our results demonstrate that the proposed software package can aid in the preliminary identification of traces of poisonous and explosive substances.

Keywords: mass spectra; substance identification; molecular dynamics; database; metric

1. Introduction

The Chemical Weapons Convention [1], which entered into force in 1997, marked a breakthrough in a long-standing effort to end the production, storage, and eventual deployment of poisoning agents in a military setting. Despite its nearly universal adoption, multiple large-scale assaults involving chemical weapons have occurred in the decades after the adoption, most notably in Syria (before [2] and after [3] its accession to the convention) and Iraq [4]. In a concerning development, nerve combat agents, originally designed for indiscriminate large area use, have been employed in attempts on lives of individuals in urban environments. The most known case is the Tokyo subway sarin attack, performed in 1995 by the Aum Shinrikyo cult, that killed 13 and injured more than 6000 people [5,6]. In recent years, authoritarian regimes in Russia and North Korea [7,8] have made targeted attempts at using various poisons to assassinate dissidents and critics [9–11]. Thus, Russian democratic opposition leader Alexei Navalny [12,13] and a former Russian spy and double agent for British intelligence Sergei Skripal [14] were notoriously poisoned with the Novichok nerve agent, Ukrainian president Victor Yushchenko was poisoned during his presidential campaign of 2004 by the TCDD agent [15], and an exiled relative of North Korea's supreme leader Kim Jong Un, Kim Jong-nam [16], was killed using

VX nerve agent. Months after the attempt on Skripal, an unrelated British couple was poisoned with Novichok [17], apparently as a collateral of a Russian attack.

Although in the aforementioned high-profile cases the specific nerve agents were reliably identified, investigations into other apparent poisonings did not produce conclusive results on the nature of chemical agents used. In cases of Russian regime critics Pyotr Verzilov [18], Dmitry Bykov [19], Vladimir Kara-Murza Jr. [20], the latter being poisoned on two separate occasions, or in a recent chain of poisonings of dissident Russian journalists and activists after the outbreak of Russian aggression against Ukraine [21], the used substances were not definitively established, which might be due to delays in samples collection and their analysis.

A range of methods exists to identify the presence of chemical warfare agents in the laboratory or in the field. The most sensitive and informative of these are non-portable techniques: mass spectrometry (MS), nuclear magnetic resonance (NMR), and chromatographic methods, such as gas chromatography (GC) or high-pressure liquid chromatography (HPLC), coupled to MS [22–26]. In their review on the detection and destruction of chemical warfare agents, Kim et al. [27] provide numerous examples of MS techniques being used to identify organophosphorus nerve agents and other toxins at very low concentrations, in some cases *in vivo*. In most MS techniques, the molecules present in the sample undergo fragmentation upon ionization, which makes interpretation of mass spectra a cumbersome task even when dealing with a clean individual substance, increasing the likelihood of failure to identify a compound in the probe. In real-world forensic samples, often heavily contaminated and containing only traces of compounds, reliable identification becomes an exceedingly difficult task. Thus, a method to automatically identify poisons or other dangerous chemical compounds in mass spectra of impure samples is of great interest to a broad community of forensic experts, medical professionals, as well as independent sleuths. Since investigations are often conducted by individuals and teams with no technical education and at their own risk, we also note that a software piece to implement this method must be easy to install and operate without MS specialist knowledge.

Focusing on MS as the prime method to identify various species in experimental mixtures, we find ourselves with a wide selection of program tools for analyzing mass spectra. First of all, many producers of MS equipment provide accompanying software to be used with it. The MassHunter code by Agilent [28] is one such example. Secondly, the analysis software developed by the National Institute of Standards and Technology (NIST), such as the AMDIS (Automated Mass Spectrometry Deconvolution and Identification System) and MS Search [29–33] are commonly used. The drawback of these programs is that they are proprietary. As an alternative, there are also open-source software, such as the ProteoWizard [34], matchms [35,36], OpenMS/pyOpenMS [37,38], and FastEI [39]. However, most of these packages require both advanced user experience and proficiency in MS. Therefore, these software packages can be hard to use for non-experts.

Finding the reference spectra in the existing literature might also present a challenging problem. In the publicly accessible databases, such as those by The NIST Chemistry WebBook [40] or National Institute of Advanced Industrial Science and Technology (AIST) [41], experimental data for many substances are not present, for instance for the compounds described in the book by V.S. Mirzayanov [42]. One possible solution to this problem is to predict spectra from theory. In recent years, an algorithm to compute mass spectra by means of molecular dynamics (MD) simulations was proposed by S. Grimme [43]. This algorithm was used to predict the MS spectra, among others, of Tabun [44] and Novichok [45], experimental work therewith being greatly hindered by the inherent danger.

To address the outlined difficulties, we present a simple-to-use software package, ToxicMassSceptic, for the analysis of mass spectra, together with a database compiled from both MS experiments and theoretical computations, as well as the workflow for producing the theoretical mass spectra. The article has the following structure. First, in Sec. 2, we introduce the methodology: the structure and sources of the database, the digital formats of the data, and algorithms and workflows to compute and assign mass spectra, including the spectral similarity metrics. Secondly, we discuss the theoretical

computation of mass spectra and demonstrate applications of the methodology in Sec. 3. Finally, conclusions are outlined in Sec. 4.

2. Methods

2.1. Mass-Spectroscopic Database

2.1.1. Database Structure and File Formats

Our database has to be easy to extend even by inexperienced users. Therefore, we store it as a set of nested directories with the structure shown in Figure 1. The top-level directory ("database") contains the subdirectories that name the class of substances ("class #1", "class #2", etc.). Each of the subdirectories ("substance #1", "substance #2", etc.) contains folders with data on the specific substance. The recommended naming of these folders is "[Brutto chemical formula in the Hill notation]_[common name of the substance]". For every substance, the "ref.ms" file is required, which contains the reference mass spectrum of the given compound. It is optional but strongly suggested to supplement an entry with a file "INFO.txt" that contains information about the substance, e.g., common names, molar mass, links to substance Wikipedia and/or PubChem webpage, etc.

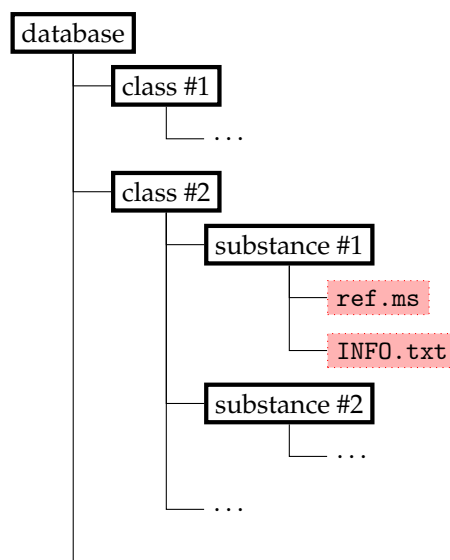


Figure 1. Schematic structure of the database with reference MS. The symbol "..." denotes similarly repeated structure.

The classes of substances in the presented database and the number of entries in each are shown in Table 1. While the classification of substances is almost always self-explanatory, assuming their separation into different chemical weapon agent types [46], environmental pollutants (such as polycyclic aromatic hydrocarbons (PAHs) [47,48] and dioxines [49,50]), a separate category (miscellaneous) had to be made to store different substances that did not fit into this arguably rigid framework.

The reference spectra of the molecules in the database (files "ref.ms") are formatted as two-column text files with pairs of numbers (x, y) in rows, where x is the integer mass-over-charge (m/z) position of the ion and y is the normalized intensity of the given ionic fragment in the MS; this format is usually denoted with an .xy file extension. The spectra in the "ref.ms" files have different normalization and are to be treated as not normalized, while normalization happens during the runtime. For a molecule with a spectrum of N fragment ions $\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$, the intensities are normalized such that

$$\sum_{i=1}^N y_i = 100\% . \quad (1)$$

Table 1. Classes of substances present in the database. In total, the database contains 394 entries, a few of which represent the same substances but different spectra.

Class of substances	N_{sub}
Acid Contaminants	9
Blister Agents	15
Blood Agents	6
Chlorophenols	7
Choking Agents	9
Dioxines	15
Explosives	59
Herbicides	7
Lachrymators	5
Nerve Agents	43
PAHs	16
Pesticides	31
Miscellaneous	172

2.1.2. Sources of Experimental Mass Spectra

Our database of molecular species borrowed mainly from the following sources: The NIST Chemistry WebBook [40], Spectral Database for Organic Compounds SDBS [41] organized by the AIST, Japan, and University of Rhode Island Explosives Database [51]. Since the Chemistry WebBook removed the option to download numerical MS data, most of the information from this database was extracted through manual digitizing the graphs (for details of this procedure, see ESI). The spectra for the two Novichok species, A-230 and A-232, were digitized from Ref. [52] using WebPlotDigitizer software [53].

2.1.3. Sources of Theoretical Mass Spectra

Theoretical mass spectra were computed using the workflow shown in Figure 2. All quantum chemical calculations, including conformational search and the MS calculation, were done with the GFN2-xTB method [54] as implemented in the xTB software [55], version 6.6.1. First, the initial molecular structure, obtained either from the NIST Chemistry WebBook, PubChem, or drawn in Jmol [56], was optimized with the xTB software. Then, conformational search was performed for this structure using CREST (version 2.12) [57,58], except for conformationally-rigid molecules. Subsequently, two augmented Born–Oppenheimer molecular dynamics (aBOMD) program packages were applied to calculate theoretical mass spectrum of the lowest energy conformer: QCxMS (version 5.2.1) [43,59,60], an original approach by S. Grimme, and DissMD, a software [61–63] based on the same idea. A detailed comparison of those approaches can be found in Sec. 3. Finally, the spectra obtained by the two theoretical approaches described above were combined as arithmetic means.

In QCxMS, the default settings were applied. The molecules were ionized by electron ionization (EI) with kinetic energy of electrons equal to 70 eV. The spectra were then collected by PlotMS (version 6.1.0). Since DissMD only simulates laser ionization, the ionization of molecules was modeled with an extreme ultraviolet (XUV) photon of 70 eV energy. In both QCxMS and DissMD calculations, the GFN2-xTB method was used to provide the potential energy surfaces for the aBOMD simulations, as this method was shown to be sufficiently accurate and computationally feasible for the mass spectra prediction [59,64].

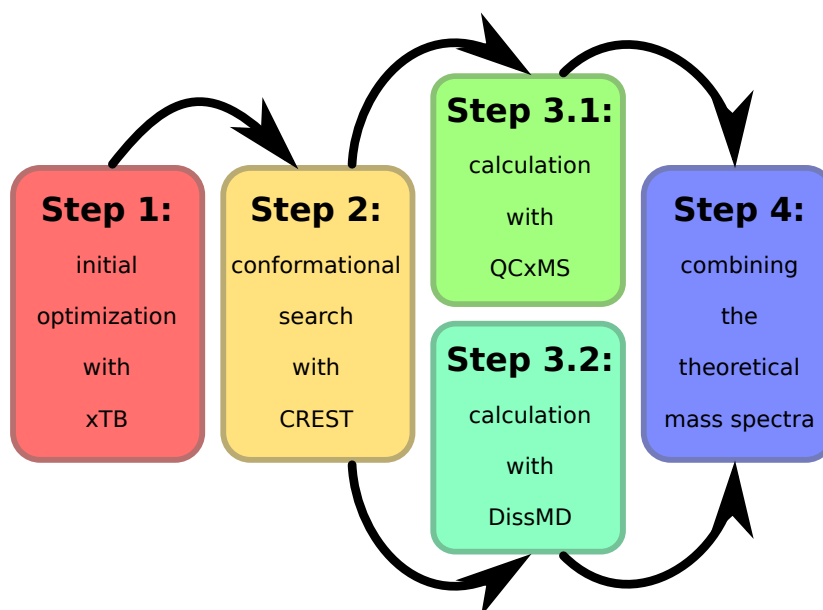


Figure 2. A general workflow scheme applied for the theoretical MS prediction for a given molecule.

2.2. Mass-Spectra Assigning Algorithm

2.2.1. Window-Function Based Assignment

The assignment was based on the assumption that there might be more than one species in the MS, which can be the case if the mixtures were not properly separated by chromatography or an alternative technique applied before the MS analysis. Therefore, the procedure involves finding only the relevant peaks in the tested spectrum to be compared with the reference database. For this, the window-based metrics were employed as described in more detail in the following.

Let us assume that we are interested in the possibility of species A with known reference spectrum of $N(A)$ peaks $\{(x_1, y_1), (x_2, y_2), \dots, (x_{N(A)}, y_{N(A)})\}$ to be present in the mixture. Intensities y_i can be represented as an $N(A)$ -dimensional vector $\mathbf{y}(A) = (y_1, y_2, \dots, y_{N(A)})$. Note that we require all intensities to be positive ($y_i > 0$ for $i = 1, \dots, N(A)$) and normalized to 100% as seen from Eq. (1). To make the comparison, we need to reduce the experimental dataset to an analogous $N(A)$ -dimensional vector of experimental intensities $\tilde{\mathbf{y}}(B|A) = (\tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_{N(A)})$, where $\tilde{y}_i$ is the spectral intensity around $x_i = m_i/z_i$ in experimentally-measured MS $I(x)$ of unknown species or mixture B . To that end, we integrate the raw experimental MS $I(x)$ with a window function $w(x|x_i)$ for a given position $x_i = m_i/z_i$ and obtain non-normalized intensities $(Y_1, Y_2, \dots, Y_{N(A)})$ as

$$Y_i = \int_0^{+\infty} I(x) \cdot w(x|x_i) dx, \quad (2)$$

where $w(x|x_i)$ is nonzero only in the vicinity of x_i . This mathematical operation essentially sums up the spectral intensity near an expected position x_i into a single value. Applying this transformation to every peak i in the reference spectrum $I(x)$ and subsequently normalizing resulting values Y_i such that

$$\tilde{y}_i = 100\% \times \frac{Y_i}{\sum_{j=1}^{N(A)} Y_j}, \quad (3)$$

we obtain experimental intensities $\tilde{y}_i$ at the discretized positions $x_i = m_i/z_i$ of the reference dataset.

Alternatively, if the experimental MS is presented in the form of discrete peaks, the integration procedure is replaced by the summation, namely

$$Y_i = \sum_{k=1}^M I_k \cdot w(x_k|x_i) dx, \quad (4)$$

where index k runs over all M peaks with intensities $I(x_k) = I_k$ identified in the experimental MS by the spectrometer's software.

In our program code, we implemented two types of window functions $w(x|x_i)$: A rectangular window,

$$w(x|x_i) = \begin{cases} 1 & , |x - x_i| \leq \sigma/2, \\ 0 & , |x - x_i| > \sigma/2, \end{cases} \quad (5)$$

and Gaussian window

$$w(x|x_i) = \exp\left(-\frac{(x - x_i)^2}{2\sigma^2}\right), \quad (6)$$

where σ is the width of the given window in m/z units. By default, the Gaussian window with $\sigma = 1/2$ is employed.

2.2.2. Assignment Metric

After defining the window-based reduction scheme of experimental data, we can discuss the route to identifying chemical species in our spectrum. To that end, we rely on a metametric, which is composed of several deterministic metrics. Thus, the simplest metric $N(B|A)$ that can be defined for a given reference spectrum A is the number of lines present in both A and B . It reads

$$N(B|A) = \sum_{i=1}^{N(A)} \theta(\tilde{y}_i - c), \quad (7)$$

where $c > 0$ is a small threshold (in our case, $c = 10^{-15}$) for numerical comparison of real numbers and $\theta(x)$ is the Heaviside step function of the form

$$\theta(x) = \begin{cases} 1, & x > 0, \\ 0, & x \leq 0 \end{cases}. \quad (8)$$

The expression in Eq. (7) can be normalized by the total number of lines in the reference spectrum $N(A)$ to produce the relative number of lines, i.e.,

$$P(B|A) = \frac{N(B|A)}{N(A)}. \quad (9)$$

More sophisticated metrics should also account for the distribution of fragment intensities. For this purpose, two sets of normalized values $\mathbf{y}(A)$ and $\tilde{\mathbf{y}}(B|A)$ can be treated as probability distributions. Thus, standard statistical distances for probability distributions can be employed. We chose four such measures: Kullback–Leibler divergence (D_{KL}) [65], Bhattacharyya distance (D_B) [66], Hellinger distance (D_H) [67], and cosine distance (D_C). In our case of two spectra, A and B , these four measures are given as [68,69]

$$D_{KL}(B|A) = \sum_{i=1}^{N(A)} \tilde{y}_i \cdot \ln\left(\frac{\tilde{y}_i}{y_i}\right), \quad (10)$$

$$D_B(B|A) = -\ln(BC(B|A)), \quad (11)$$

$$D_H(B|A) = \sqrt{1 - BC(B|A)}, \quad (12)$$

$$D_C(B|A) = 1 - \frac{\sum_{i=1}^{N(A)} y_i \tilde{y}_i}{\sqrt{\left(\sum_{i=1}^{N(A)} \tilde{y}_i^2\right) \cdot \left(\sum_{i=1}^{N(A)} y_i^2\right)}}, \quad (13)$$

respectively. In Eqs. (11) and (12), BC is the so-called Bhattacharyya dimensionless coefficient [66,70] given by

$$BC(B|A) = \frac{1}{100\%} \sum_{i=1}^{N(A)} \sqrt{\tilde{y}_i \cdot y_i}. \quad (14)$$

Here, the division by 100% is motivated by the fact of BC being defined for probability distributions normalized to 1. The three chosen measures of similarities for probability distributions from Eqs. (9)–(12) require that components of the vector $\tilde{\mathbf{y}}(B|A)$ are non-negative. Note that Eqs. (10)–(12) are undefined for $N(B|A) = 0$, which corresponds to the case of the species not being present in the spectrum.

The combined metametric is then constructed from Eqs. (9)–(13) such that

$$D_{\text{meta}}(B|A) = \frac{1}{P(B|A) \times \sum_{j=1}^{N(A)} Y_j} \times \left(\frac{D_{\text{KL}}(B|A)}{\varsigma_{\text{KL}}} + \frac{D_{\text{B}}(B|A)}{\varsigma_{\text{B}}} + \frac{D_{\text{H}}(B|A)}{\varsigma_{\text{H}}} + \frac{D_{\text{C}}(B|A)}{\varsigma_{\text{C}}} \right), \quad (15)$$

where Y_j is the non-normalized experimental intensity given by Eqs. (2) or (4) and ς_X is the standard deviation of the given metric $X = \text{KL}, \text{B}, \text{H}, \text{and C}$, computed over the whole available dataset as

$$\varsigma_X = \sqrt{\langle D_X^2 \rangle - \langle D_X \rangle^2} = \sqrt{\frac{1}{N_d} \sum_a D_X^2(B|A) - \left(\frac{1}{N_d} \sum_a D_X(B|A) \right)^2}, \quad (16)$$

where index a runs over all spectra in the database and N_d is the number of such spectra. The value of $D_{\text{meta}}(B|A)$ from Eq. (15) tends to zero if the two spectra A and B are similar and increases with the growing dissimilarity of the experimental spectrum from the reference. Although Bhattacharyya and Hellinger distances provide the same relative ranking of substances, it can be advantageous to use both in the metametric, as they might have different sensitivity at different values of the Bhattacharyya dimensionless coefficient BC .

2.2.3. Background Removal Algorithm

Experimentally measured spectra can contain signals from the background. This may result in empty areas of a spectrum producing negative intensities when using Eqs. (2) and (4). To avoid that, basic filtering of the experimental MS signal $I(x)$ can be performed. The simplest and most robust approach is probably a visual determination of the noise threshold level I_{thr} , and setting all the values $I(x) \leq I_{\text{thr}}$ to zero. However, a crude automatic routine can also be designed (for example, see Ref. [71]) assuming that non-zero peaks occupy only a minor part of the spectrum in all available m/z ranges and that the baseline signal is $I = 0$. To that end, we represent a spectrum in a discretized form with lines $I_1, I_2, \dots, I_M$. Then, the following procedure can be employed.

1. Calculate the standard deviation of $I(x)$ from baseline ($I = 0$) as $\text{SD}_0 = \sqrt{\frac{1}{M} \sum_{k=1}^M I_k^2}$.
2. Consider only values $I_k < q \cdot \text{SD}_0$, with $q \geq 1$ being an arbitrary selectivity coefficient, forming a new set $I_1^{(1)}, I_2^{(1)}, \dots, I_{M_1}^{(1)}$, where the upper index “(1)” indicates the iteration number and $M_1 \leq M$ is the number of elements in the new set.
3. Calculate the new standard deviation as $\text{SD}_1 = \sqrt{\frac{1}{M_1} \sum_{k=1}^{M_1} (I_k^{(1)})^2}$.
4. Repeat steps 2 and 3 until the number of elements in the set remains constant or a maximum number of iterations p is reached.

- Set values of the original mass spectrum below the final threshold $q \cdot SD_p$ to zero.

This automatic background removal procedure is implemented in our program code, with the default number of steps $p = 3$ and selectivity coefficient $q = 1.5$.

2.3. Software

The program code is written in Python version 3.8 for the Linux, MacOS, and MS Windows operational systems, distributed under an open source Apache License version 2.0 [72], and is managed using the version control system GIT [73] by the provider GitLab [74]. The source code is available under Ref. [75]. The list of program requirements includes Python packages such as NUMPY [76] and MATPLOTLIB [77]. The code has a clear version number and is accompanied by two types of documentation: (i) a README file in the Markdown format outlining external dependencies, package structure as well as the installation procedure and (ii) an automatically generated Doxygen [78] code documentation describing all constituting objects and functions. The package-management system PIP3 [79] governs the installation procedure. The code is aimed to be fully unit-tested. To that end, the package UNITTEST [80] is employed. The current code design enables the use of our program as an external Python library as well as through a command line interface.

2.4. Statistical Analysis of Results

Let us assume that the user is interested in testing N_{trials} number of different mixtures B . Each such i th mixture B_i contains a compound A_{true}^i , which is also present in the database. Furthermore, we assume that for each sample B_i , the top- K matching candidates $\mathbf{A}^i = \{A_1^i, A_2^i, \dots, A_K^i\}$ are suggested by our algorithm based on the metrics introduced above in Sec. 2.2.2. Here, each set $\mathbf{A}^i$ is sorted in descending order such that its first element is the most probable match. Therefore, index j denotes the rank of compound A_j^i , i.e., $j = R(A_j^i)$, with lower ranks being preferable. Then, the following scores can be introduced to assess the performance of our algorithm.

- Top-K accuracy* (also known as *Hit rate at rank K*), which is equal to the number of trials with the correctly identified compound being present in top-best K candidates $N_{\text{in top-K}}$ divided by the total number of trials N_{trials} and multiplied by 100%, i.e.,

$$\text{top-K accuracy} = \frac{N_{\text{in top-K}}}{N_{\text{trials}}} \times 100\% . \quad (17)$$

- Mean reciprocal rank (MRR)*, defined as

$$\text{MRR} = \frac{1}{N_{\text{trials}}} \sum_{i=1}^{N_{\text{trials}}} \frac{1}{R(A_{\text{true}}^i)} \times 100\% , \quad (18)$$

where $R(A_{\text{true}}^i)$ is the rank of the correctly identified compound A_{true}^i in trial i .

- Mean rank (MR)*, defined as

$$\text{MR} = \frac{1}{N_{\text{trials}}} \sum_{i=1}^{N_{\text{trials}}} R(A_{\text{true}}^i) . \quad (19)$$

The top- K score from Eq. (17) shows how often the correctly identified compound was present in the K most probable candidates predicted by the program code, whereas MRR from Eq. (18) evaluates the ability of the code to assign low ranks to relevant chemical compounds. In the case of an ideal assignment, when correct compounds always occupy the very top of the suggestion list, both scores are equal to 100%. The MR score from Eq. (19) is closely related to MRR, but is equal to or greater than 1.0 and tends to 1.0 for better performing recommendation systems.

3. Results and Discussion

3.1. Mass-Spectra Prediction Workflow

Predicted mass spectra presented in this work were computed using either QCxMS or DissMD. The latter is a part of the PyRAMD package [61,81,82]. Both algorithms employ Born–Oppenheimer molecular dynamics (BOMD), as proposed by S. Grimme in his seminal paper [43]. Before discussing our results, we first compare those two approaches.

A graphical representation of an aBOMD-based theoretical workflow for an MS spectrum prediction is depicted in Figure 3. First, multiple molecular geometries are generated, representing the gaseous ensemble of molecules in the spectrometer. Those structures are then used as initial points to start BOMD dynamics for ions. To include the electronic excitation effects, the BOMD dynamics are perturbed (or augmented) by the kinetic energy influx from an external energy reservoir, producing aBOMD trajectory. This energy, referred to as the internal excess energy (IEE), and the ion charge are ascribed according to the ionization procedure. If, upon the aBOMD trajectory propagation, a dissociation of the molecule is detected, the parent ion trajectory is stopped, and new aBOMD trajectories for the products are initiated by sharing the charge and IEE of the parent ion between fragments. Then, these trajectories of the daughter ions are propagated further. Finally, the mass spectra are computed from the ensemble of MD trajectories by counting the final products.

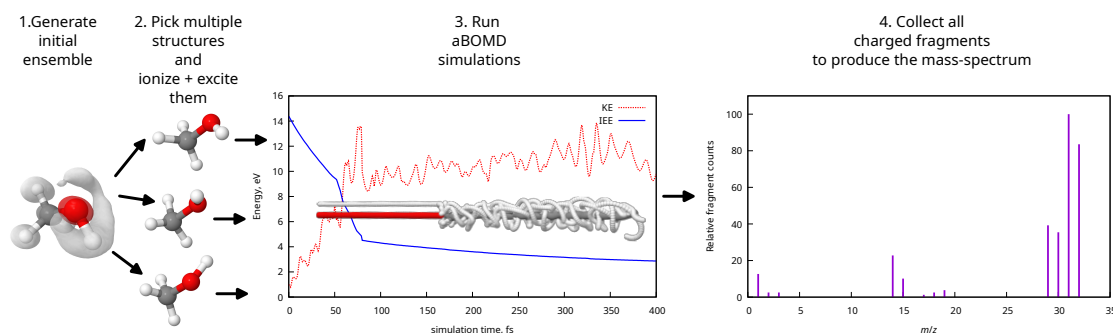


Figure 3. Graphical representation of a mass spectra simulation using aBOMD approach.

Despite this scheme's general simplicity, a few crucial components in the algorithm define the simulation behavior. The QCxMS and DissMD use two completely different approaches to generate initial conditions. In the QCxMS, the thermostated MD of the neutral molecule is performed to sample the initial structures and their velocities. In the DissMD, the simplified Wigner sampling [82,83] approach from a user-provided geometry is used, which, in principle, can include some of the nuclear quantum effects [84] for the lighter nuclei such as hydrogens. Furthermore, these two approaches also differ greatly in the ionization procedure and the assignment of the IEE. In QCxMS, an arbitrary Poisson-like distribution is employed [59,85]

$$P(\text{IEE}) = \frac{\exp(c \cdot \text{IEE} \cdot (1 + \ln(b/(c \cdot \text{IEE}))) - b)}{a \cdot \text{IEE} + 1}, \quad (20)$$

where $P(\text{IEE})$ is the probability of the ion to have the value of IEE upon ionization, whereas $a = 0.2$ eV, $b = 1$ eV, and $c = 1/N_{\text{ve}}$ are pre-defined parameters with N_{ve} being the number of valence electrons in the system. In the DissMD, however, an approach based on the electronic density of the states is used. Upon applying the maximum entropy principle and energy conservation to molecular ionization, one arrives at the following distribution [62]:

$$P(\text{IEE}) = \text{DoS}(\text{IEE}) \cdot (E_i - \text{IP} - \text{IEE})^{\frac{N_f}{2}-1}, \quad (21)$$

where $\text{DoS}(\text{IEE})$ is the electronic density of states of the ion, E_i is the total energy of the ionization event, IP is the sum of ionization potentials to reach a given ionization state, and N_f is the number of

degrees of freedom for the leaving particles. For the photoionization, which is the only available case in DissMD, $E_i = m h \nu$ and $N_f = 3 \cdot N_{re}$. In these expressions, m is the number of absorbed photons, $h = 6.626 \times 10^{-34}$ J·s is the Planck constant, ν is the photons' frequency, and N_{re} is the number of electrons removed upon ionization ($N_f = 3$ for single ionization, $N_f = 6$ for double ionization, etc.). Note, however, that Eq. (21) can still be applied for the electron impact ionization. In this case, E_i is the kinetic energy of the electrons and N_f is set to $3 \cdot (N_{re} + 1)$ to account for the leaving ionizing particle's degree of freedom. Unlike in the first version of the software, in which the explicitly computed excited states were used to obtain the electronic density of states [81], the current version of the DissMD uses a simplified heuristic model based on the Van-der-Waals volume and surface to approximate $\text{DoS(IEE)} \propto \text{IEE}^n$ as a power function with a single parameter n . In this case, Eq. (21) reduces to a beta-distribution [62].

The third crucial component of the simulation is the rate of internal conversion (IC), showing how fast the IEE decays into nuclear motions. For this purpose, the QCxMS uses the energy-gap law in the form [59]

$$k_{IC}^{-1} = \sum_{j>i}^M \frac{k_h}{N_{ve}} \exp(\alpha(\varepsilon_i - \varepsilon_j)) , \quad (22)$$

where $k_h = 2$ ps and $\alpha = 0.5 \text{ eV}^{-1}$ are constraints, ε_i is the energy of an i -th orbital, and M is the total number of orbitals. Contrary to that, in the DissMD a classical model of hot electrons with kinetic energy of IEE colliding with motionless nuclei is employed. In the DissMD prototype, a similar algorithm, based on an idea of electron–nuclear collision-induced IC, was used to compute the IC rates using the atomic electronic densities through the plasma frequency estimated from atomic charges [62]. However, in the newer code, it was replaced with a simplified model for the rate of such collisions is given as [63]

$$k_{IC} = \kappa \frac{\sqrt{m_e \text{IEE}}}{m_{amu}(L_0 + L_{mol})} N_n N_e , \quad (23)$$

where N_e and N_n are the total number of electrons and nuclei in the ion, respectively, m_e is the electron mass, m_{amu} is the atomic mass unit (dalton), L_{mol} is the molecular length (atomic-charge-product-weighted sum of all chemical bonds, determined from the covalent radii of atoms), $L_0 = 5 \text{ \AA}$ is the regularizing parameter, and $\kappa \approx 1.28$ is the fitted parameter based on the available experimental data [63].

When the dissociation is detected, the QCxMS and DissMD again proceed in a different fashion. The DissMD follows a direct route: Upon the detection of dissociation of ion M^{q+} into fragments A and B it calculates the energies of several channels



that satisfy the charge conservation $q_A + q_B = q$. Upon dissociation, the channels with non-negative kinetic energy release (KER) are assigned a probability proportional to this KER value. Subsequently, one of these channels is chosen according to those probabilities. This leads to a speedup in the calculation, as the neutral fragments are not propagated. However, this approach requires a larger number of trajectories to be computed. In the QCxMS, a concept of statistical charge, or statistical weighing, is used. In this approach, the MD is carried out for all fragments, but their associated intensities depend on the weight, which is determined as [59]

$$C_i = \frac{\exp\left(-\frac{IP_i}{k_B T}\right)}{\sum_j \exp\left(-\frac{IP_j}{k_B T}\right)} \quad (25)$$

with indices i and j running over the number of fragments, IP_j being the ionization potential of a given fragment, k_B being Boltzmann constant, and $T = \text{KE}/(3k_B N_n)$ being the instant temperature of nuclei, as computed from their kinetic energy (KE). With these fragment weights, it is also possible to directly

apply the isotopic distribution in the post-analysis, while the DissMD requires running simulations with different isotopes.

To demonstrate the predictive capabilities of the aBOMD-based approach for computing the mass spectra, we took four molecules, for which we had the available spectra: methanol (CH_3OH), novichok A-230 ($\text{C}_7\text{H}_{16}\text{FN}_2\text{OP}$), o-chlorophenoxyacetic acid ($\text{C}_8\text{H}_7\text{ClO}_3$), and vinclozolin ($\text{C}_{12}\text{H}_9\text{Cl}_2\text{NO}_3$). Structures of the most stable conformers of these molecules, according to CREST, can be found in Figure 4. As a metric to judge the similarity between spectra, we chose the number of peaks from the reference spectrum from Eq. (7), the Kullback–Leibler divergence given in Eq. (10), and the Bhattacharyya distance from Eq. (11).

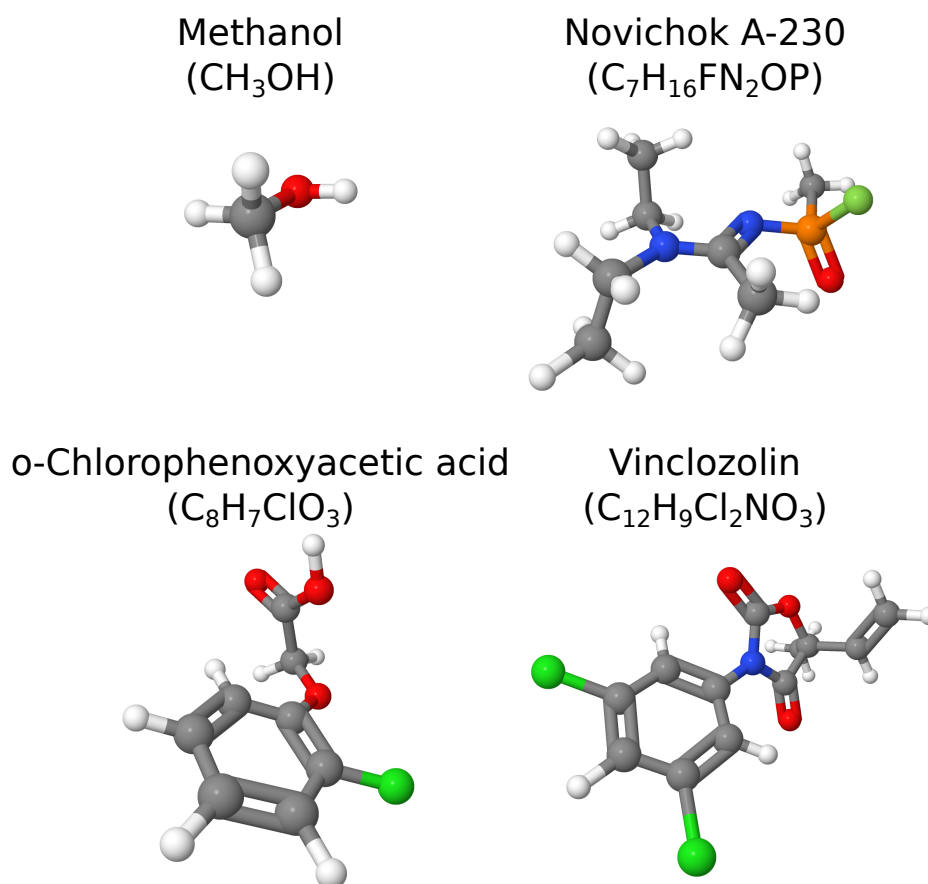


Figure 4. The most stable conformers of four test molecules used in the theoretical mass-spectra prediction.

The results of our comparison are demonstrated in Figures 5 and 6, and in Table 4. It is clear that the QCxMS, as the software specifically designed for EIMS predictions, outperforms DissMD. Nevertheless, in three out of four cases, DissMD provided extra fragments, which were missing in the QCxMS predictions. In all cases, the combination of the both methods allowed us to cover more than 80% of lines from experimental spectra. However, the relative intensities of the peaks are not always perfect, which can be a result of wrong ionization conditions in the simulations. Nevertheless, we can confirm the conclusions from previous studies in Refs. [44,45], stating that it is possible to use theoretically predicted mass spectra for the assignment of species with absent experimental reference spectra.

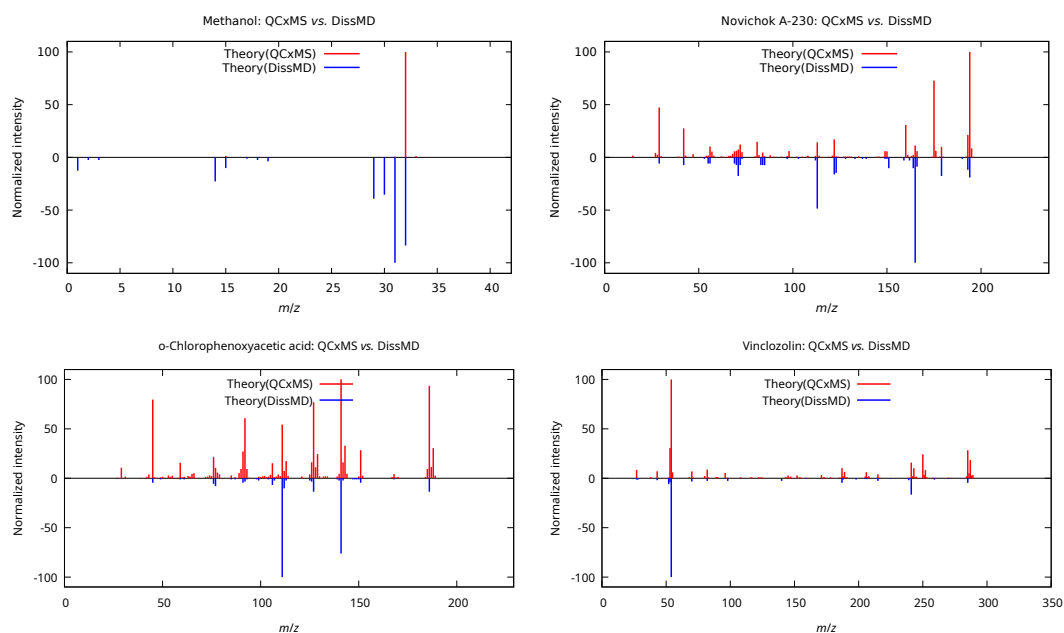


Figure 5. Comparison of the two theoretical mass spectra computed with QCxMS or DissMD software, for four test molecules (methanol, novichok A-230, o-chlorophenoxyacetic acid, and vinclozolin from Fig. 4).

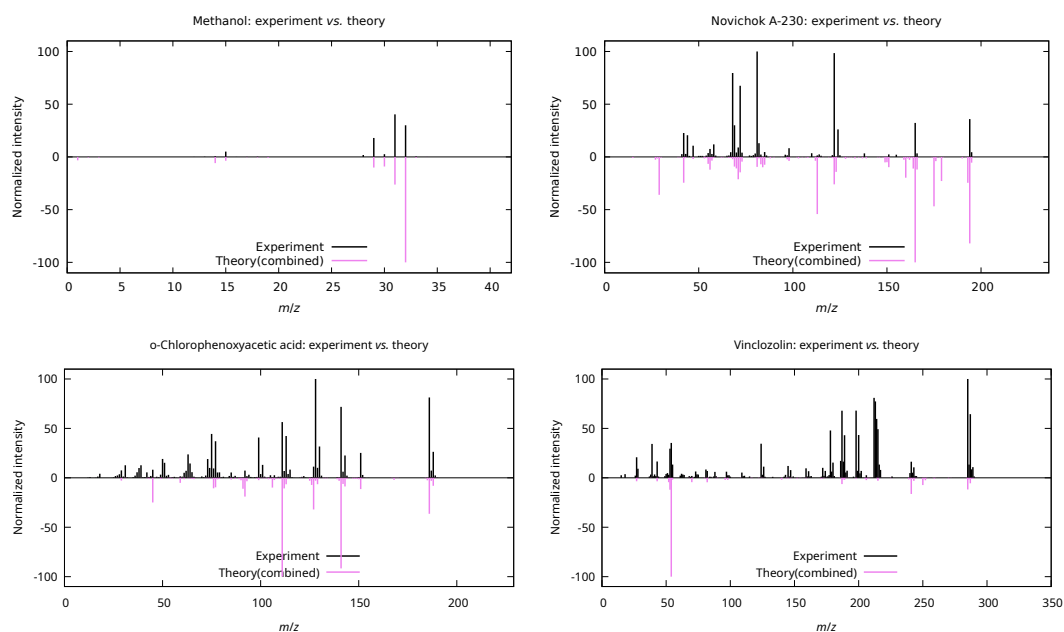


Figure 6. Comparison of the combined theoretical mass spectrum with the experimental one from the database for four test molecules (methanol, novichok A-230, o-chlorophenoxyacetic acid, and vinclozolin from Fig. 4).

Table 2. Comparison of theoretically predicted mass spectra with their experimental reference counterparts from the database. The number of lines N_{lines} is calculated via Eq. (7), while N_{ref} is the total number of peaks in the reference spectrum. The metrics D_{KL} and D_{B} are those given in Eqs. (10) and (11).

Spectrum	$N_{\text{lines}}/N_{\text{ref}}$	$P, \%$	$D_{\text{KL}}, \%$	D_{B}
Methanol (CH_3OH)				
QCxMS	9/16	56.2	109.90	0.47
DissMD	10/16	62.5	29.02	0.06
Combined	13/16	81.2	36.98	0.10
Novichok A-230 ($\text{C}_7\text{H}_{16}\text{FN}_2\text{OP}$)				
QCxMS	46/52	88.5	90.34	0.24
DissMD	17/52	32.7	180.01	0.60
Combined	46/52	88.5	117.12	0.32
o-Chlorophenoxyacetic acid ($\text{C}_8\text{H}_7\text{ClO}_3$)				
QCxMS	118/129	91.5	104.03	0.28
DissMD	30/129	23.3	137.05	0.56
Combined	120/129	93.0	98.48	0.30
Vinclozolin ($\text{C}_{12}\text{H}_9\text{Cl}_2\text{NO}_3$)				
QCxMS	84/105	80.0	122.95	0.42
DissMD	19/105	18.1	235.75	0.85
Combined	86/105	81.9	166.31	0.50

However, we would also claim that new software is probably due to development that would take the best algorithmic solutions from the QCxMS and DissMD. For the ionization stage, it makes more sense to assign the IEE from a physically sound model from Eq. (21). For computing the IC rate, one might use a better model of the electron-phonon coupling. One such possibility is demonstrated in Refs. [86,87], where the rate is calculated based on the Fermi–Dirac distribution and orbital overlaps for the two consecutive MD steps. For the treatment of dissociation, the QCxMS approach appears more suitable. However, instead of using the heuristically defined weights from Eq. (25), it would make more sense to use a modified version of the model introduced in Ref. [88], as it takes into account not only the ionization energies of fragments, but also the electron affinities, and the dissociation energies.

3.2. Performance Tests with Simulated Data

The ToxicMassSceptic features are subject to unit tests, ensuring the code works as expected. One of the production tests trials the performance of the code in the presence of noise and additional substances. We model the species with Gaussian-shaped peaks with randomly chosen standard deviation, that is, in the range between 0.05 and 0.1 m/z . We take the mass spectrum of a randomly chosen species from the database and generate a spectrum in the m/z range from 0 to 500 with 2000 points. Then, we add a background that consists of two components. First, the signal of the substance is mixed with a spectrum composed of signals from benzene (C_6H_6), oxygen (O_2), nitrogen (N_2), carbon dioxide (CO_2), and farnesene ($\text{C}_{15}\text{H}_{24}$), one of sesquiterpenes. The relative amounts of the background species are randomly chosen between 0.1 and 0.2. Then, a random uniformly distributed noise is added on top of that with a signal-to-noise (S/N) level randomly chosen from the interval between $S/N=100$ and $S/N=1000$. Then, this generated spectrum is passed through our assignment algorithm, including the background removal and the rating of the actual compound, which is stored. The mean rating of the spectra upon multiple trials should not exceed an MR (Eq. 19) threshold, which, in our case, is set to 5. The current version of the software routinely passes this test.

To further demonstrate the performance of our code and compare different metrics, we carried out assignments of 500 randomly generated spectra. To that end, we modified the settings described above by lowering the allowed signal-to-noise level to $5 \leq S/N \leq 100$, and additionally allowing peak intensities to vary by $\pm 50\%$ and their positions to be shifted by $\pm 0.2 m/z$. The assignment was repeated 48 times, leading to 24,000 trials in total and allowing us to compute the mean values and standard deviations for statistical parameters from Eqs. (17)–(19). The results of this analysis are shown in Table 3. As can be seen, the worse top-1 result is obtained using the cosine distance D_{C} , reaching an

accuracy level of only about 30%. The performance of other metrics is much higher and varies from about 55 to 91%. Similar trends are observed for the MRR and MR scores. The use of the proposed metametric D_{meta} was found to produce results of the highest quality in all cases.

Table 3. Performance of ToxicMassSceptic for simulated data. Results for the top-K from Eq. (17) and MRR from Eq. (18) scores are given in %. The MR is given according to Eq. (19).

	Top-1	Top-3	Top-5	Top-10	MRR	MR
D_{meta}	91 ± 1	98.6 ± 0.6	98.8 ± 0.4	99.0 ± 0.3	94.9 ± 0.8	1.8 ± 0.4
D_{KL}	55 ± 2	87 ± 2	94 ± 1	96.9 ± 0.9	72 ± 2	3.3 ± 0.7
D_{B}	61 ± 2	92 ± 1	97.0 ± 0.9	98.7 ± 0.7	77 ± 1	2.1 ± 0.3
D_{H}	61 ± 2	92 ± 1	97.0 ± 0.9	98.7 ± 0.6	77 ± 1	2.1 ± 0.3
D_{C}	30 ± 2	69 ± 2	84 ± 2	95 ± 1	53 ± 2	4.1 ± 0.5

3.3. Performance Test with Experimental Noisy Dataset

As an example of the mass spectra with noisy background, we took the strong-field-induced mass spectra of a tree-ring PAH fluorene ($\text{C}_{13}\text{H}_{10}$), which are openly available from Ref. [89]. Since fluorene is in the database, and the laser-induced fragmentation patterns look similar to those obtained with EI, we simply tested the identification of the species with the mass spectra obtained using different laser peak powers (from 1.5×10^{13} to 6.8×10^{13} W/cm²). In all of the cases, the automatic background removal was applied.

The background removal results are shown in Figure 7. As one can see, the background is indeed removed quite efficiently, leaving only the signals from the ion fragments. The cleaning in the range of higher masses is somewhat less effective, which is due to the overall background level increase, as clearly seen in a logarithmic plot. Nevertheless, such background removal was sufficient to identify fluorene in the case of all experimental spectra considered in this work. The results for the highest peak power spectrum are shown in Figure 8.

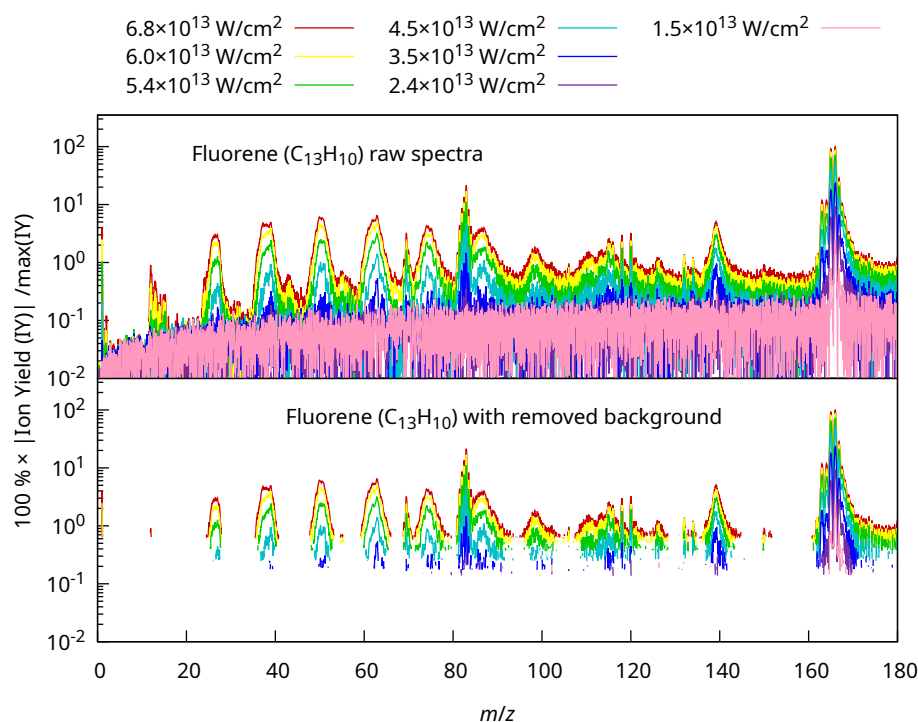


Figure 7. Experimental mass spectra of fluorene ($\text{C}_{13}\text{H}_{10}$) obtained by strong-field ionization with ultrashort laser pulses of varied peak intensity. The top figure shows raw experimental spectra, while the bottom one is after the background removal. Note that the logarithmic scale on the absolute intensity is used for y -axis, and the curve disappearance in the bottom figure means that the signal is zero.

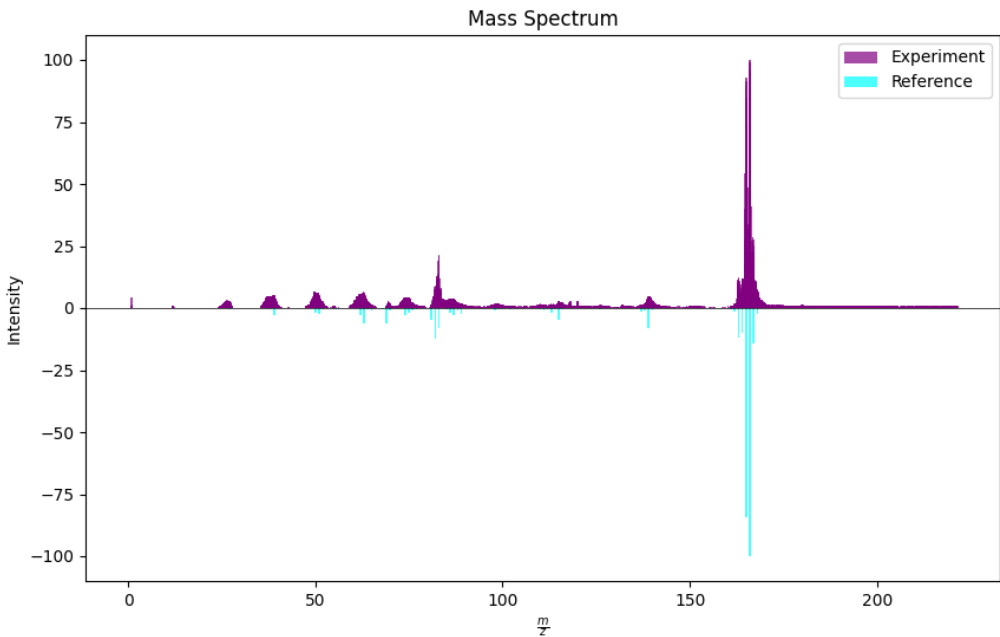


Figure 8. Comparison of the experimental and reference spectrum of fluorene for the highest (6.8×10^{13} W/cm²) peak power mass spectra.

3.4. Performance Tests with an Experimental Dataset of Cleaned Spectra

The mass spectra of 64 substances were recorded using GC (HP6890, Agilent Technologies) coupled to a single quadrupolar MS (HP5972A or HP5973, Agilent Technologies) or with GC (Trace 1310) coupled to MS (TSQ Duo Triple Quadrupole, Thermo Scientific). Helium was used as a carrier gas, and the spectra were measured in the range of 50–500 m/z . The EI was used to ionize species with an electron KE of 70 eV. More details on the measurement parameters are available in ESI.

The experimental dataset consists of several classes of substances: Acid contaminants, chlorophenols, dioxins, PAHs, pesticides, and herbicides. For each of the compounds from this dataset, the reference spectrum was added to the database, and then ToxicMassSceptic was tested to provide the assignment results. We ranked the performance in each dataset using six scores: Top-1, top-3, top-5, and top-10 accuracies from Eq. (17), MRR from Eq. (18), and MR from Eq. (19). The results of the test are given in Table 4. As one can see, most of the species were correctly identified in the top-3 best-matched substances, and the correct compound was the best-matched one 60% of the time, on average. With that, we conclude that the current performance allows the identification of species in unknown samples.

Table 4. Performance of the ToxicMassSceptic assignment algorithm on the experimental datasets of various classes of substances.

Substance class	N_{subst}	Top-1	Top-3	Top-5	Top-10	MRR	MR
Acid contaminants	9	44.4	55.6	55.6	66.7	52.8	17.0
Dioxins	4	75.0	100.0	100.0	100.0	87.5	1.2
PAHs	16	43.8	100.0	100.0	100.0	68.8	1.8
Pesticides	29	82.8	100.0	100.0	100.0	90.8	1.2
Herbicides	6	50.0	83.3	83.3	83.3	66.8	19.2

3.5. Testing Theoretical Reference Against Cleaned Experimental Data

In the dataset used in Sec. 3.4, there were three dioxines: 1,2-Dichlorodibenzo-p-dioxin, 1,3-Dichlorodibenzo-p-dioxin, and 1,4-Dichlorodibenzo-p-dioxin. These compounds are suitable for testing the assignment of experimental spectra against theoretically predicted mass spectra. For

that reason, we computed the theoretical mass spectra of these three structural isomers using the workflow shown in Fig. 2. In addition to that, in Sec. 3.1, we calculated theoretical mass spectra for o-chlorophenoxyacetic acid and vinclozolin, which were also present in the same database.

Thus, we took these five substances to test their identification with the ToxicMassSceptic software. The resulting ranking of these theoretical spectra (R) against their experimental counterparts is given in Table 5 in columns Threshold = 0%. As one can see, the results are acceptable. However, upon examination of the theoretical spectra, one can see that the number of reference lines (N_{lines}) is much larger than usually available for experimental spectra taken from various databases (which is typically of the order of a few tens of data points). Therefore, we have tried to remove some of the fragments with lower intensities from the theoretical spectra to see the effect on the identification of substances. In particular, we removed every lower-intensity peak by setting a relative threshold with respect to the most intensive one. We tried two settings: thresholds of 1% and 5%, which drastically reduced the number of lines and had an effect on the prediction performance (see Table 5). With 1% threshold, the MR value for this set of five spectra was slightly lower than the 0% and 5% settings, which indicates that there is an optimal amount of lines to represent a species in the database, as too many or too little may lead to misidentification of the species. Therefore, we recommend removing the weak intensity fragments when using ToxicMassSceptic for predicting theoretical mass spectra, as this improves the identification probability $P(B|A)$ (Eq. 9). The importance of the latter can be seen from the definition of the metametric from Eq. (15).

Table 5. Ranking (R) of and the number of lines (N_{lines}) in the theoretically predicted mass spectra of five substances against the experimental data. Different threshold values denote the removal of the weak intensity peaks from the reference dataset. 1,2-DpD, 1,3-DpD, 1,4-DpD, and o-CA denote 1,2-Dichlorodibenzo-p-dioxin, 1,3-Dichlorodibenzo-p-dioxin, 1,4-Dichlorodibenzo-p-dioxin, and o-chlorophenoxyacetic, respectively. The last row is the MR [see Eq. (19)] values for the dataset of these five molecules at a given threshold.

Substance	Threshold = 0%		Threshold = 1%		Threshold = 5%	
	R	N_{lines}	R	N_{lines}	R	N_{lines}
1,2-DpD	21	154	13	64	11	15
1,3-DpD	37	135	20	64	31	15
1,4-DpD	10	152	3	69	11	17
o-CA	5	140	3	45	3	19
Vinclozolin	2	201	14	26	36	7
MR	15.0		10.6		18.4	

4. Conclusions

In this paper, we have presented an algorithm and a computer program for identifying toxic and combat compounds using mass spectrometry, ToxicMassSceptic, that is easy to operate for nonprofessionals. An essential part of it is the database of substances, assembled from multiple different sources, most prominently from databases like the NIST Chemistry WebBook and the SDBS of AIST, as well as from quantum chemical modeling. The use of theoretically predicted mass spectra allowed us to obtain reference data for poisonous substances for which no publicly accessible data exist. According to our tests against simulated and experimental datasets, ToxicMassSceptic with the database can facilitate preliminary identification of possible traces of poisonous and explosive substances. However, the identification results can be biased toward the available database. The final conclusions regarding substance identification should always be based on expert opinion and validated with other experimental methods, such as NMR or rotational spectroscopy [90].

Author Contributions: Conceptualization, D.S.T.; methodology, D.S.T., D.G.A., and V.V.R.; software, D.S.T., A.A.M., A.A.A., O.D., V.V.R., D.G.A.; validation, D.S.T., D.G.A., V.V.R.; formal analysis, D.S.T., A.A.M., V.V.R., D.G.A.; investigation, D.S.T., M.A.K., A.A.M., A.A.A., O.D., N.A.V., E.A.A., M.K.; data curation, D.S.T., M.A.K., A.A.A., N.A.V., E.A.A., M.K., V.V.R.; writing—original draft preparation, D.S.T.; writing—review and editing,

A.A.M., D.G.A., V.V.R.; visualization, D.S.T., A.A.A., O.D.; supervision, D.S.T., D.G.A., V.V.R.; project administration, D.S.T., D.G.A., V.V.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: The latest version of the software can be obtained from the GitLab repository <https://gitlab.com/madschumacher/toxicmasssceptic/>. A stable version of the software and database is also available in the ESI. Besides the software, the ESI information on the experimental conditions for GC-MS measurements and the measurements themselves, procedures for manual digitizing of mass spectra from NIST Chemistry WebBook, the results of the statistical testing of ToxicMassSceptic on the generated dataset, and the simulated mass spectra. The full simulations of the mass spectra used here can be obtained from the Zenodo repository: <https://dx.doi.org/10.5281/zenodo.14831652>.

Acknowledgments: D.S.T. acknowledges DESY (Hamburg, Germany), a member of the Helmholtz Association HGF. In particular, D.S.T.'s calculations were enabled through the Maxwell computational resources operated at DESY. D.G.A. acknowledges funding provided by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) — project number 545861628. D.S.T. also acknowledges Dr. Andrei Benediktovitch, Dr. Vladimir Lipp, Dr. Andrey Zayakin, and Prof. Melanie Schnell for valuable discussions and support. D.G.A. acknowledges Prof. Benedikt Kaufer for valuable discussions.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AIST	National Institute of Advanced Industrial Science and Technology
EI	electron ionization
GC	gas chromatography
HPLC	high-pressure liquid chromatography
IC	internal conversion
IEE	internal excess energy
KE	kinetic energy
KER	kinetic energy release
MD	molecular dynamics
MR	mean rank
MRR	mean reciprocal rank
MS	mass-spectrometry
NIST	National Institute of Standards and Technology
NMR	nuclear magnetic resonance
PAHs	polycyclic aromatic hydrocarbons
XUV	extreme ultraviolet

References

1. Organisation for the Prohibition of Chemical Weapons. Convention on the Prohibition of the Development, Production, Stockpiling and Use of Chemical Weapons and on Their Destruction.
2. UN Secretary-General.; UN Mission to Investigate Allegations of the Use of Chemical Weapons in the Syrian Arab Republic (2013). Report of the United Nations Mission to Investigate Allegations of the Use of Chemical Weapons in the Syrian Arab Republic on the Alleged Use of Chemical Weapons in the Ghouta Area of Damascus on 21 August 2013 :: Note /: By the Secretary-General 2013.
3. United Nations. 'Reasonable Grounds to Believe' Syrian Government Used Chlorine Gas on Douma Residents in 2018, Head of Chemical Weapons Monitoring Organization Tells Security Council, 2023.
4. United Press International. 1988 Kurdish Massacre Labeled Genocide. https://www.upi.com/Top_News/Special/2010/03/08/1988-Kurdish-massacre-labeled-genocide/93471268062566/.

5. Ogawa, Y.; Yamamura, Y.; Ando, H.; Kadokura, M.; Agata, T.; Fukumoto, M.; Satake, T.; Machida, K.; Sakai, O.; Miyata, Y.; et al., An Attack with Sarin Nerve Gas on the Tokyo Subway System and Its Effects on Victims. In *Natural and Selected Synthetic Toxins*; chapter 22, pp. 333–355, [<https://pubs.acs.org/doi/pdf/10.1021/bk-2000-0745.ch022>]. <https://doi.org/10.1021/bk-2000-0745.ch022>.
6. Sugiyama, A.; Matsuoka, T.; Sakamune, K.; Akita, T.; Makita, R.; Kimura, S.; Kuroiwa, Y.; Nagao, M.; Tanaka, J. The Tokyo subway sarin attack has long-term effects on survivors: A 10-year study started 5 years after the terrorist incident. *PLOS ONE* **2020**, *15*, 1–12. <https://doi.org/10.1371/journal.pone.0234967>.
7. GURIEV, S.; TREISMAN, D., FEAR AND SPIN. In *Spin Dictators: The Changing Face of Tyranny in the 21st Century*; Princeton University Press, 2022; pp. 3–30.
8. Schulmann, E. The Russian political system in transition: Scenarios for power transfer. *NUPI Working Paper* **2018**, 883.
9. Brunka, Z.; Ryl, J.; Brushtulli, P.; Gromala, D.; Walczak, G.; Zięba, S.; Pieśniak, D.; Sein Anand, J.; Wiergowski, M. Selected Political Criminal Poisonings in the Years 1978–2020: Detection and Treatment. *Toxics* **2022**, *10*. <https://doi.org/10.3390/toxics10080468>.
10. Brunka, Z.; Ryl, J.; Brushtulli, P.; Gromala, D.; Walczak, G.; Zięba, S.; Pieśniak, D.; Sein Anand, J.; Wiergowski, M. Selected Political Criminal Poisonings in the Years 1978–2020: Detection and Treatment. *Toxics* **2022**, *10*, 468. <https://doi.org/10.3390/toxics10080468>.
11. Dewey, K. Poisonous Affairs: Russia's Evolving Use of Poison in Covert Operations. *The Nonproliferation Review* **2022**, *29*, 155–176. <https://doi.org/10.1080/10736700.2023.2229691>.
12. Bellingcat Investigation Team. FSB Team of Chemical Weapon Experts Implicated in Alexey Navalny Novichok Poisoning, 2020.
13. Steindl, D.; Boehmerle, W.; Körner, R.; Praeger, D.; Haug, M.; Nee, J.; Schreiber, A.; Scheibe, F.; Demin, K.; Jacoby, P.; et al. Novichok nerve agent poisoning. *The Lancet* **2021**, *397*, 249–252. [https://doi.org/10.1016/S0140-6736\(20\)32644-1](https://doi.org/10.1016/S0140-6736(20)32644-1).
14. May, T.; Prime Minister's Office, 10 Downing Street. PM Commons Statement on Salisbury Incident: 12 March 2018, 2018.
15. Sorg, O.; Zennegg, M.; Schmid, P.; Fedosyuk, R.; Valikhnovskyi, R.; Gaide, O.; Kniazevych, V.; Saurat, J.H. 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) poisoning in Victor Yushchenko: identification and measurement of TCDD metabolites. *The Lancet* **2009**, *374*, 1179–1185. [https://doi.org/10.1016/S0140-6736\(09\)60912-0](https://doi.org/10.1016/S0140-6736(09)60912-0).
16. Ng, E. Post-Mortem: VX Poison Killed Brother of North Korean Leader. <https://apnews.com/general-news-90e425dbaf1e44d1ba77e2eea890fc67>, 2017.
17. Amesbury Novichok Poisoning: Couple Exposed to Nerve Agent **2018**.
18. Charité – Universitätsmedizin Berlin. Pyotr Verzilov Receiving Treatment at Charité, 2018.
19. Bellingcat Investigation Team. Russian Poet Dmitry Bykov Targeted by Navalny Poisoners, 2021.
20. Countersanctions. How FSB Officers Tried to Poison Vladimir Kara-Murza. <https://theins.ru/en/politics/253146>.
21. Weiss, M. Blood Simple. Several Russian Journalists and Activists Were Poisoned in Europe. <https://theins.ru/en/politics/264280>, 2023.
22. Amend, N.; Niessen, K.V.; Seeger, T.; Wille, T.; Worek, F.; Thiermann, H. Diagnostics and Treatment of Nerve Agent Poisoning—Current Status and Future Developments. *Annals of the New York Academy of Sciences* **2020**, *1479*, 13–28. <https://doi.org/10.1111/nyas.14336>.
23. Rybal'chenko, I.V.; Baigil'diev, T.M.; Rodin, I.A. Chromatography–Mass Spectrometry Analysis for the Determination of the Markers and Biomarkers of Chemical Warfare Agents. *Journal of Analytical Chemistry* **2021**, *76*, 26–40. <https://doi.org/10.1134/S1061934821010111>.
24. Baygildiev, T.; Vokuev, M.; Braun, A.; Rybalchenko, I.; Rodin, I. Monitoring of hydrolysis products of mustard gas, some sesqui- and oxy-mustards and other chemical warfare agents in a plant material by HPLC-MS/MS. *Journal of Chromatography B* **2021**, *1162*, 122452. <https://doi.org/https://doi.org/10.1016/j.jchromb.2020.122452>.
25. Vokuev, M.F.; Baygildiev, T.M.; Plyushchenko, I.V.; Ikhalaynen, Y.A.; Ogorodnikov, R.L.; Solontsov, I.K.; Braun, A.V.; Savelieva, E.I.; Rybalchenko, I.V.; Rodin, I.A. Untargeted and targeted analysis of sarin poisoning biomarkers in rat urine by liquid chromatography and tandem mass spectrometry. *Analytical and Bioanalytical Chemistry* **2021**, *413*, 6973–6985. <https://doi.org/10.1007/s00216-021-03655-3>.
26. Vokuev, M.; Baygildiev, T.; Braun, A.; Frolova, A.; Rybalchenko, I.; Rodin, I. Monitoring of hydrolysis products of organophosphorus nerve agents in plant material and soil by liquid chromatography-tandem

- mass spectrometry. *Journal of Chromatography A* **2022**, 1685, 463604. <https://doi.org/https://doi.org/10.1016/j.chroma.2022.463604>.
27. Kim, K.; Tsay, O.G.; Atwood, D.A.; Churchill, D.G. Destruction and Detection of Chemical Warfare Agents. *Chemical Reviews* **2011**, 111, 5345–5403. <https://doi.org/10.1021/cr100193y>.
 28. Agilent Masshunter Quantitative Analysis software (RRID:SCR_015040).
 29. 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, 859–866, [[https://doi.org/10.1016/1044-0305\(94\)87009-8](https://doi.org/10.1016/1044-0305(94)87009-8)]. PMID: 24222034, [https://doi.org/10.1016/1044-0305\(94\)87009-8](https://doi.org/10.1016/1044-0305(94)87009-8).
 30. Stein, S.E. An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *Journal of the American Society for Mass Spectrometry* **1999**, 10, 770–781, [[https://doi.org/10.1016/S1044-0305\(99\)00047-1](https://doi.org/10.1016/S1044-0305(99)00047-1)]. [https://doi.org/10.1016/S1044-0305\(99\)00047-1](https://doi.org/10.1016/S1044-0305(99)00047-1).
 31. Place, B.J. Development of a Data Analysis Tool to Determine the Measurement Variability of Consensus Mass Spectra. *Journal of the American Society for Mass Spectrometry* **2021**, 32, 707–715, [<https://doi.org/10.1021/jasms.0c00423>]. PMID: 33591742, <https://doi.org/10.1021/jasms.0c00423>.
 32. Wallace, W.E.; Moorthy, A.S. NIST Mass Spectrometry Data Center standard reference libraries and software tools: Application to seized drug analysis. *Journal of Forensic Sciences* **2023**, 68, 1484–1493, [<https://onlinelibrary.wiley.com/doi/pdf/10.1111/1556-4029.15284>]. <https://doi.org/https://doi.org/10.1111/1556-4029.15284>.
 33. Stein, S.E. Estimating probabilities of correct identification from results of mass spectral library searches. *Journal of the American Society for Mass Spectrometry* **1994**, 5, 316–323, [[https://doi.org/10.1016/1044-0305\(94\)85022-4](https://doi.org/10.1016/1044-0305(94)85022-4)]. PMID: 24222569, [https://doi.org/10.1016/1044-0305\(94\)85022-4](https://doi.org/10.1016/1044-0305(94)85022-4).
 34. Chambers, M.C.; Maclean, B.; Burke, R.; Amodi, D.; Ruderman, D.L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egerton, J.; et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology* **2012**, 30, 918–920. <https://doi.org/10.1038/nbt.2377>.
 35. Huber, F.; Verhoeven, S.; Meijer, C.; Spreeuw, H.; Castilla, E.M.V.; Geng, C.; j. van der Hooft, J.J.; Rogers, S.; Belloum, A.; Diblen, F.; et al. matchms - processing and similarity evaluation of mass spectrometry data. *Journal of Open Source Software* **2020**, 5, 2411. <https://doi.org/10.21105/joss.02411>.
 36. de Jonge, N.F.; Hecht, H.; Strobel, M.; Wang, M.; van der Hooft, J.J.; Huber, F. Reproducible MS/MS library cleaning pipeline in matchms. *Journal of Cheminformatics* **2024**, 16, 88. <https://doi.org/10.1186/s13321-024-00878-1>.
 37. Röst, H.L.; Sachsenberg, T.; Aiche, S.; Bielow, C.; Weissner, H.; Aicheler, F.; Andreotti, S.; Ehrlich, H.C.; Gutenbrunner, P.; Kenar, E.; et al. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nature methods* **2016**, 13, 741–748.
 38. Röst, H.L.; Schmitt, U.; Aebersold, R.; Malmström, L. pyOpenMS: A Python-based interface to the OpenMS mass-spectrometry algorithm library. *PROTEOMICS* **2014**, 14, 74–77, <https://doi.org/https://doi.org/10.1002/pmic.201300246>.
 39. Yang, Q.; Ji, H.; Xu, Z.; Li, Y.; Wang, P.; Sun, J.; Fan, X.; Zhang, H.; Lu, H.; Zhang, Z. Ultra-fast and accurate electron ionization mass spectrum matching for compound identification with million-scale in-silico library. *Nature Communications* **2023**, 14, 3722. <https://doi.org/10.1038/s41467-023-39279-7>.
 40. "Mass Spectra" by NIST Mass Spectrometry Data Center, William E. Wallace, director in NIST Chemistry WebBook, NIST Standard Reference Database Number 69, Eds. P.J. Linstrom and W.G. Mallard, National Institute of Standards and Technology, Gaithersburg MD, 20899, <https://doi.org/10.18434/T4D303>, (retrieved January 11, 2025).
 41. sdbms : <https://sdbms.db.aist.go.jp/Disclaimer.aspx> (National Institute of Advanced Industrial Science and Technology, 05.01.2025).
 42. Mirzayanov, V.S. *State Secrets: An Insider's Chronicle of the Russian Chemical Weapons Program*; Outskirts Press, Inc.: Denver, Colorado, 2008.
 43. Grimme, S. Towards First Principles Calculation of Electron Impact Mass Spectra of Molecules. *Angewandte Chemie International Edition* **2013**, 52, 6306–6312, [<https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.201300158>]. <https://doi.org/https://doi.org/10.1002/anie.201300158>.
 44. Chernicharo, F.C.; Modesto-Costa, L.; Borges Jr, I. Molecular dynamics simulation of the electron ionization mass spectrum of tabun. *Journal of Mass Spectrometry* **2020**, 55, e4513, e4513 JMS-19-0159.R2, <https://doi.org/https://doi.org/10.1002/jms.4513>.

45. Chernicharo, F.C.S.; Modesto-Costa, L.; Borges Jr., I. Simulation of the electron ionization mass spectra of the Novichok nerve agent. *Journal of Mass Spectrometry* **2021**, *56*, e4779, <https://doi.org/https://doi.org/10.1002/jms.4779>.
46. Chauhan, S.; Chauhan, S.; D'Cruz, R.; Faruqi, S.; Singh, K.; Varma, S.; Singh, M.; Karthik, V. Chemical warfare agents. *Environmental Toxicology and Pharmacology* **2008**, *26*, 113–122. <https://doi.org/https://doi.org/10.1016/j.etap.2008.03.003>.
47. Srogi, K. Monitoring of environmental exposure to polycyclic aromatic hydrocarbons: a review. *Environmental Chemistry Letters* **2007**, *5*, 169–195. <https://doi.org/10.1007/s10311-007-0095-0>.
48. Låg, M.; Øvrevik, J.; Refsnes, M.; Holme, J.A. Potential role of polycyclic aromatic hydrocarbons in air pollution-induced non-malignant respiratory diseases. *Respiratory Research* **2020**, *21*, 299. <https://doi.org/10.1186/s12931-020-01563-1>.
49. Hites, R.A. Dioxins: An Overview and History. *Environmental Science & Technology* **2011**, *45*, 16–20. <https://doi.org/10.1021/es1013664>.
50. Kirkok, S.K.; Kibet, J.K.; Kinyanjui, T.K.; Okanga, F.I. A review of persistent organic pollutants: dioxins, furans, and their associated nitrogenated analogues. *SN Applied Sciences* **2020**, *2*, 1729. <https://doi.org/10.1007/s42452-020-03551-y>.
51. University of Rhode Island Explosives Database : <http://expdb.chm.uri.edu> (accessed on 01.11.2024).
52. Eskandari, M.; Faraz, S.M.; Hosseini, S.E.; Moradi, S.; Saeidian, H. Fragmentation pathways of chemical weapons convention-related organophosphorus Novichok agents: The electron ionization and electrospray ionization tandem mass spectroscopy and DFT calculation studies. *International Journal of Mass Spectrometry* **2022**, *473*, 116794. <https://doi.org/https://doi.org/10.1016/j.ijms.2021.116794>.
53. Rohatgi, A. Webplotdigitizer: Version 4.6, 2022.
54. 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, [<https://doi.org/10.1021/acs.jctc.8b01176>]. PMID: 30741547, <https://doi.org/10.1021/acs.jctc.8b01176>.
55. 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*, e1493, [<https://wires.onlinelibrary.wiley.com/doi/pdf/10.1002/wcms.1493>]. <https://doi.org/https://doi.org/10.1002/wcms.1493>.
56. Jmol: an open-source Java viewer for chemical structures in 3D.
57. Pracht, P.; Bohle, F.; Grimme, S. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169–7192. <https://doi.org/10.1039/C9CP06869D>.
58. Pracht, P.; Grimme, S.; Bannwarth, C.; Bohle, F.; Ehlert, S.; Feldmann, G.; Gorges, J.; Müller, M.; Neudecker, T.; Plett, C.; et al. CREST—A program for the exploration of low-energy molecular chemical space. *The Journal of Chemical Physics* **2024**, *160*, 114110, [https://pubs.aip.org/aip/jcp/article-pdf/doi/10.1063/5.0197592/19837939/114110_1.5.0197592.pdf]. <https://doi.org/10.1063/5.0197592>.
59. Ásgeirsson, V.; Bauer, C.A.; Grimme, S. Quantum chemical calculation of electron ionization mass spectra for general organic and inorganic molecules. *Chem. Sci.* **2017**, *8*, 4879–4895. <https://doi.org/10.1039/C7SC00601B>.
60. 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, [<https://doi.org/10.1021/acs.jpca.6b02907>]. PMID: 27139033, <https://doi.org/10.1021/acs.jpca.6b02907>.
61. Tikhonov, D.S. PyRAMD. <https://gitlab.desy.de/denis.tikhonov/pyramd>, 2024.
62. Tikhonov, D.S.; Datta, A.; Chopra, P.; Steber, A.L.; Manschwetus, B.; Schnell, M. Approaching black-box calculations of pump-probe fragmentation dynamics of polyatomic molecules. *Zeitschrift für Physikalische Chemie* **2020**, *234*, 1507–1531. <https://doi.org/doi:10.1515/zpch-2020-0009>.
63. Lee, J.W.L.; Tikhonov, D.S.; Allum, F.; Boll, R.; Chopra, P.; Erk, B.; Gruet, S.; He, L.; Heathcote, D.; Kazemi, M.M.; et al. The kinetic energy of PAH dication and trication dissociation determined by recoil-frame covariance map imaging. *Phys. Chem. Chem. Phys.* **2022**, *24*, 23096–23105. <https://doi.org/10.1039/D2CP02252D>.
64. Koopman, J.; Grimme, S. Calculation of Electron Ionization Mass Spectra with Semiempirical GFNN-xTB Methods. *ACS Omega* **2019**, *4*, 15120–15133, [<https://doi.org/10.1021/acsomega.9b02011>]. PMID: 31552357, <https://doi.org/10.1021/acsomega.9b02011>.

65. Kullback, S.; Leibler, R.A. On Information and Sufficiency. *The Annals of Mathematical Statistics* **1951**, *22*, 79–86.
66. Bhattacharyya, A. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.* **1943**, *p.* 99–109.
67. Hellinger, E. Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. *Journal für die reine und angewandte Mathematik* **1909**, *1909*, 210–271. <https://doi.org/doi:10.1515/crll.1909.136.210>.
68. Nilsson, N.; Håkansson, B.; Ortiz-Catalan, M. Classification complexity in myoelectric pattern recognition. *Journal of NeuroEngineering and Rehabilitation* **2017**, *14*, 68. <https://doi.org/10.1186/s12984-017-0283-5>.
69. Kailath, T. The Divergence and Bhattacharyya Distance Measures in Signal Selection. *IEEE Transactions on Communication Technology* **1967**, *15*, 52–60. <https://doi.org/10.1109/TCOM.1967.1089532>.
70. Bhattacharyya, A. On a Measure of Divergence between Two Multinomial Populations. *Sankhyā: The Indian Journal of Statistics (1933-1960)* **1946**, *7*, 401–406.
71. Potemkin, A.A.; Proskurnin, M.A.; Volkov, D.S. Noise Filtering Algorithm Using Gaussian Mixture Models for High-Resolution Mass Spectra of Natural Organic Matter. *Analytical Chemistry* **2024**, *96*, 5455–5461, [<https://doi.org/10.1021/acs.analchem.3c05453>]. PMID: 38530650, <https://doi.org/10.1021/acs.analchem.3c05453>.
72. Apache License Version 2.0, January 2004. <https://www.apache.org/licenses/LICENSE-2.0>, Access date: January 5, 2025.
73. Git Source Control Management Tool. <https://git-scm.com>, Access date: January 5, 2025.
74. GitLab. <https://about.gitlab.com>, Access date: January 5, 2025.
75. Toxic Mass Sceptic (TMS): Release 0.0.1, 2025. <https://gitlab.com/madschumacher/toxicmasssceptic>, Access date: January 5, 2025.
76. Harris, C.R.; Millman, K.J.; van der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; et al. Array programming with NumPy. *Nature* **2020**, *585*, 357–362. <https://doi.org/10.1038/s41586-020-2649-2>.
77. Hunter, J.D. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* **2007**, *9*, 90–95. <https://doi.org/10.1109/MCSE.2007.55>.
78. van Heesch, D. Doxygen. <http://www.doxygen.org>, Access date: January 5, 2025.
79. Package installer for Python PIP. <https://pypi.org/project/pip/>, Access date: January 5, 2025.
80. Unittest — Unit testing framework. <https://docs.python.org/3/library/unittest.html>, Access date: January 5, 2025.
81. Tikhonov, D.S. Metadynamics simulations with Bohmian-style bias potential. *Journal of Computational Chemistry* **2023**, *44*, 1771–1775, [<https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.27125>]. <https://doi.org/https://doi.org/10.1002/jcc.27125>.
82. Tikhonov, D.S. PyRAMD Scheme: A Protocol for Computing the Infrared Spectra of Polyatomic Molecules Using ab Initio Molecular Dynamics. *Spectroscopy Journal* **2024**, *2*, 171–187. <https://doi.org/10.3390/spectroscj2030012>.
83. Tikhonov, D.S.; Vishnevskiy, Y.V. Describing nuclear quantum effects in vibrational properties using molecular dynamics with Wigner sampling. *Phys. Chem. Chem. Phys.* **2023**, *25*, 18406–18423. <https://doi.org/10.1039/D3CP01007D>.
84. Markland, T.E.; Ceriotti, M. Nuclear quantum effects enter the mainstream. *Nature Reviews Chemistry* **2018**, *2*, 0109. <https://doi.org/10.1038/s41570-017-0109>.
85. Bauer, C.A.; Grimme, S. Automated Quantum Chemistry Based Molecular Dynamics Simulations of Electron Ionization Induced Fragmentations of the Nucleobases Uracil, Thymine, Cytosine, and Guanine. *European Journal of Mass Spectrometry* **2015**, *21*, 125–140, [<https://doi.org/10.1255/ejms.1313>]. PMID: 26307693, <https://doi.org/10.1255/ejms.1313>.
86. Medvedev, N.; Li, Z.; Tkachenko, V.; Ziaja, B. Electron-ion coupling in semiconductors beyond Fermi's golden rule. *Phys. Rev. B* **2017**, *95*, 014309. <https://doi.org/10.1103/PhysRevB.95.014309>.
87. Medvedev, N.; Milov, I. Electron-phonon coupling in metals at high electronic temperatures. *Phys. Rev. B* **2020**, *102*, 064302. <https://doi.org/10.1103/PhysRevB.102.064302>.
88. Tikhonov, D.S.; Lee, J.W.L.; Schnell, M. On the thermodynamic stability of polycations. *The Journal of Chemical Physics* **2024**, *160*, 244110, <https://doi.org/10.1063/5.0207526>.

89. Garg, D.; Chopra, P.; Lee, J.W.L.; Tikhonov, D.S.; Kumar, S.; Akcaalan, O.; Allum, F.; Boll, R.; Butler, A.A.; Erk, B.; et al. Ultrafast dynamics of fluorene initiated by highly intense laser fields. *Phys. Chem. Chem. Phys.* **2024**, *26*, 20261–20272. <https://doi.org/10.1039/D3CP05063G>.
90. Tikhonov, D.S.; Sueyoshi, C.J.; Sun, W.; Xie, F.; Khon, M.; Gougoula, E.; Li, J.; Berggötz, F.; Singh, H.; Tonauer, C.M.; et al. Scaling of Rotational Constants. *Molecules* **2024**, *29*. <https://doi.org/10.3390/molecules29245874>.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.