ACS Partner Journal

Journal of the American Society for Mass Spectrometry

pubs.acs.org/jasms

Article

Automated Identification of Ions Observed in Mass Spectra of Inorganic Compounds Using Isotopic Distribution Brute Force: Methodology and Performance Measurements

Published as part of Journal of the American Society for Mass Spectrometry virtual special issue "Asilomar: Computational Mass Spectrometry".

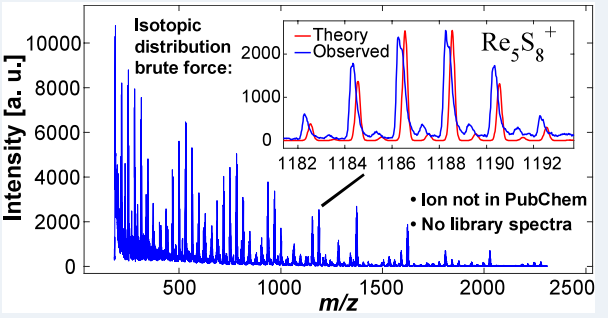
Viacheslav V. Lebedev,* Daniil I. Yarykin, and Aleksey K. Buryak

Cite This: *J. Am. Soc. Mass Spectrom.* 2024, 35, 1806−1817

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: This Article describes the method of isotopic distribution brute force, which can be used to identify ions registered in mass spectra of inorganic compounds in an automated manner when a library search cannot be conducted. A detailed description of the isotopic distribution brute force methodology is presented, including a discussion of computation-related difficulties. The ability of the proposed algorithm to identify various inorganic ions is tested on a small set of real-life low-resolution mass spectra of lead halides and copper halides. An evaluation of the isotopic distribution brute force performance is conducted using the low-resolution experimental mass spectra of natural rhenium sulfide and lead(II) chloride. Based on identification results and obtained performance measurements, we formulate the empirical restrictions on the input data, ensuring that the application of isotopic distribution brute force is feasible from the standpoints of search space reduction and identification time.



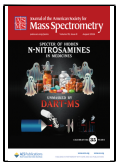
INTRODUCTION

Identification of registered ions is the final stage of mass spectrum processing.[1,2] Mass spectrometry is predominantly applied in the analysis of certain organic sample types (proteins, biologically active substances, drugs, etc.).[3] A significant amount of experimental data has been accumulated in this area of mass spectrometry application. As the result, the ions registered in mass spectra of such samples are usually identified using a library search.[4] Library search implies finding the reference mass spectrum that best matches the experimental one.[4] Signals of some of the experimentally registered ions may not be observed in the matching library mass spectrum, and various methods for the identification of these ions have been developed.[2] Many of such identification methods involve full-exhaustion (or "brute force") testing of possible ion compositions subject to certain constraints or a modification of this approach.

However, the areas of application of mass spectrometry are not limited to the analysis of organic samples. In particular, experiments involving mass spectrometry have been conducted to study inorganic construction materials,[5] rocks,[6] catalysts, etc. Nevertheless, the amount of data accumulated in those fields is relatively small when compared with the analysis of organic samples. Few libraries comprising mass spectra of inorganic compounds are known. Such libraries commonly have a thematic nature and contain a limited amount of data. Moreover, reference mass spectra of inorganic substances are often available only for certain ionization conditions (usually electron ionization and electrospray ionization). All of the above means that, generally, ions registered in the mass spectra of inorganic substances cannot be identified by means of a library search. The use of other identification methods, including those based on brute force, is required to identify such ions.

One of fundamentals of mass spectrometry, namely, the ability to detect various isotopes of elements present in the formed ion, plays an important role in identification of signals observed in mass spectra of inorganic compounds.[3] This fundamental provision implies that the formed ions are registered as a series of several peaks separated from each other by an approximately integer m/z value. Features that describe such a series of peaks for an ion with a given

Received: April 16, 2024
Revised: June 23, 2024
Accepted: July 16, 2024
Published: July 23, 2024



composition, i.e., the number of peaks, their position along the m/z axis, and relative intensities in series, are together referred to as the "isotopic distribution" of the ion.

Some of the existing approaches for the identification of organic ions make use of the isotopic distribution observed in mass spectra. Such approaches primarily include approximation methods, notably the well-developed Averagine model.[7,8] Approximation methods allow the estimation of the elemental composition of a given ion based on certain input data, which commonly includes the observed isotopic distribution of the precursor ion. These methods have demonstrated the ability to identify ions of certain organic compound classes rather accurately.[9] However, approximation methods are essentially based on pattern matching and can only be used to identify ions composed of similar structural units.[8] Because of this, identification techniques that approximate the elemental composition of organic ions based on observed isotopic distributions are predominantly applied within the framework of peptide studies.[9,10]

In general cases, however, the role of isotopic distribution is "often forgotten"[11] when fragments of organic samples are identified. This is largely a consequence of the fact that chemical elements often found in organic compounds either occur in the form of one most-abundant isotope with a molar fraction approaching 1, like C, H, N, and O, or are monoisotopic, like, e.g., F and P.[12] Thus, the isotopic distributions of two organic substances with similar masses but different compositions will differ only slightly. At the same time, it is known that the measured isotopic ratios are inevitably influenced by a number of equipment-specific factors, even when the real content of various isotopes in a sample is close to average values.[13] Combined with low abundances of heavy isotopes of C, N, O, etc., the dissimilarity of the observed and theoretical isotopic distributions that results from such factors often leads to organic fragment being misidentified.[13]

As a consequence of the above, organic ions observed in the mass spectrum cannot be reliably identified based on isotopic distribution alone. This led to the development of methods that do not use the data on observed isotopic distribution during the identification of organic ions. One commonly used example of such methods is the monoisotopic mass search, which allows the selection of the true composition of an observed ion from a list of compounds whose monoisotopic masses are the closest to the mass of the given experimentally observed peak.[14] Monoisotopic mass search has recently been used together with a technique called "deisotoping".[15] Deisotoping implies that each ion observed in the mass spectrum is represented by one peak for identification purposes. This peak usually corresponds to a combination of the most abundant isotopes of elements that comprise the ion. Peaks corresponding to other isotopologues are not involved in identification and are deliberately neglected.

Unlike the elements that are often present in organic analytes, many elements found in the studied inorganic substances have two or more stable isotopes with natural abundances as high as several dozen percent.[12] Note that this study focuses on only ions composed of elements with known stable isotopes; however, the provisions presented here can be extended to purely radioactive elements if the abundances of their isotopes are defined. Of 118 currently known chemical elements, 80 elements have at least one stable isotope.[12] Per our calculations presented later, the above number includes 46 chemical elements with "non-trivial" isotopic distributions, i.e., those elements for which the abundance of the second most common isotope amounts to at least several percent. In particular, several metals that are commonly found in the Earth's crust (such as magnesium, iron, copper, and zinc), most of the noble metals (e.g., silver and ruthenium), some heavy metals (like rhenium, molybdenum, etc.), and halides (notably bromine and chlorine) satisfy this condition.[12] The isotopic distribution of ions that contain such elements is valuable information, which should not be neglected. High abundances of various isotopes of elements contained in such inorganic ions imply that the result of identification is influenced to a much lower extent when compared to organic compounds.

Even in well-developed fields of mass spectrometry applications, the registered signals are often identified manually in cases where the observed ion is expected to contain an inorganic adduct with a "non-trivial" isotopic distribution.[16] Manual identification is excessively time-consuming since it requires repeating routine, monotonous actions multiple times. As mentioned earlier, the process of organic ion identification is automated using the library search and chemoinformatics. However, the automated identification of ions observed in mass spectra of inorganic compounds receives significantly less attention.

An obvious yet apparently rarely used approach to the automated identification of ions observed in mass spectra of inorganic compounds is the brute force of all possible composition options for a given isotopic peak series while accounting for the distribution of relative intensities in such a series. For convenience, this approach is later termed "isotopic distribution brute force". The underlying concept has a trivial nature, yet each step of the approach involves making nontrivial assumptions and the nontrivial selection of algorithms and their parameter values.

In this Article, we attempt to summarize the information on the automated identification of ions observed in the mass spectra of inorganic substances by means of isotopic distribution brute force. Since any application of brute force involves testing multiple options and performing a large number of calculations, we also conducted test identification runs and carried out performance measurements to formulate empirical restrictions on input data, subject to which the application of isotopic distribution brute force is suitable for the processing of experimental inorganic mass spectrometry data.

The conceptual limitations of the approach and difficulties that arise at its individual stages will be demonstrated using mass spectra of two different inorganic compounds, namely, lead(II) chloride and natural rhenium sulfide. Two mass spectra of the same compounds were used to conduct performance measurements. The ability of the proposed algorithm to identify various ions was tested on the set of seven experimental mass spectra of lead halides and copper halides, which were acquired during real-life studies of construction materials.[17] All of the mass spectra used in this Article were obtained using a low-resolution mass spectrometer under conditions of laser desorption-ionization. Graphical representations of all the used mass spectra are shown in Figures S1−S9. We failed to find publicly available libraries that contained reference mass spectra of lead halides, copper halides, and ReS_2 rhenium sulfide for the given ionization conditions. Furthermore, it is known that multiple cluster ions

https://doi.org/10.1021/jasms.4c00153
J. Am. Soc. Mass Spectrom. 2024, 35, 1806−1817

can be formed when laser desorption-ionization methods, e.g. matrix-assisted (MALDI) or surface-enhanced laser desorption-ionization (SELDI), are applied. Such clusters are commonly characterized by compositions that are not observed under traditional ionization conditions, i.e., electron ionization or electrospray ionization. All of the above did not allow the identification of the ions detected in the presented mass spectra by means of a library search or by monoisotopic mass brute force using data from, e.g., the PubChem database. This stimulated us to search for other approaches to automated identification.

THEORY

Problem Statement. Let B be a set of ion compositions that could hypothetically be registered in the mass spectrum (or in certain part of it) of an inorganic compound. Let us denote the number of molecular formulas in set B as M. For each hypothetical composition $b_m, m = 1, ..., M$ to be tested, there exists a theoretical isotopic distribution, i.e., the discrete distribution of mass peak intensities, denoted as $T^{(m)}$. The intensities of signals observed at m/z values corresponding to $T^{(m)}$ in the experimental mass spectrum are distributed as $X^{(m)}$.

The isotopic distribution brute force implies solving a hypothesis testing task for each tested composition, $b \in B$. The hypothesis testing aims to establish whether the theoretical and observed intensity distributions $T^{(m)}$ and $X^{(m)}$ are identical. Thus, the problem solved during isotopic distribution brute force can be written as follows:

$$\forall\, m = 1, ..., M:$$

$$H_0: X^{(m)} \overset{d}{=} T^{(m)}$$

$$H_1: X^{(m)} \overset{d}{\neq} T^{(m)}$$

If the null hypothesis H_0 is true, meaning that the theoretical and observed distributions are identical, then molecular formula b_m can be considered as a candidate composition for an ion that is represented by the given series of isotopic signals with a known most intense peak.

Identification Algorithm. Identification of ions by means of isotopic distribution brute force involves three steps for each tested composition $b_m, m = 1, ..., M$. First, the theoretical isotopic distribution $T^{(m)}$ is calculated. Next, each theoretical peak is matched with a peak observed in the experimental mass spectrum, meaning that the observed distribution $X^{(m)}$ is set. Finally, distributions $X^{(m)}$ and $T^{(m)}$ are tested for identity, and the decision is made.

Input Data. Three types of input data are required to conduct ion identification using isotopic distribution brute force, namely, the studied mass spectrum, the set of tested compositions B, and algorithm parameter values.

The list of algorithm parameters differs depending on which methods are used to perform each step of the identification. Generally, such a list includes the following: (1) a value of the stop criterion for the calculation of theoretical isotopic distribution, (2) peak detection settings, (3) the allowed mass tolerance for matching theoretical peaks with observed ones, and (4) a threshold value (or confidence level) for measure, which is used to determine whether distributions are identical. These parameters will be discussed in further paragraphs.

The only formal requirement of the input mass spectrum is that the mass spectrum must be presented as two vectors (or, in terms of informatics, arrays) of equal length. Theoretically, the studied mass spectrum may have any resolution and may be presented in either original (profile) or reduced (centroid) dimensions. However, the values of algorithm parameters need to be adjusted depending on the mentioned properties of the mass spectrum. In this study, we assume that the isotopic distribution brute force is conducted using low-resolution mass spectra of original dimensions as input data. The mass spectra are assumed to have undergone baseline correction.

The set of compositions B to be tested during brute force can be filled manually. However, this contradicts the general idea of automating the identification process. We believe that the generation of tested compositions using additional input data is a preferred option. A common technique, which is commonly used by equipment manufacturers,[18] implies generating the elemental compositions based on the set of allowed chemical elements $A = \{a_1, a_2, ..., a_L\}$ and vectors containing minimum and maximum numbers of atoms of each element, $C^{(min)} = (c_1^{(min)}, c_2^{(min)}, ..., c_L^{(min)})^T$ and $C^{(max)} = (c_1^{(max)}, c_2^{(max)}, ..., c_L^{(max)})^T$, respectively. The generation of compositions itself could be implemented using one of the algorithms that allow the formation of all the combinations of factor variable values.[19] Within the scope of task solved, the numbers of atoms of each element are factor variables, and their values lie in the $[c^{(min)}, c^{(max)}]$ range.

The set of allowed chemical elements can be formed based on various additional data. Some of the data sources are listed in the Supporting Information, section "Input data. Forming the set of compositions to be tested".[20] However, preliminary information about the sample and expert judgment of the researcher play a decisive role in the formation of a set of allowed chemical elements A.

The minimum and maximum numbers of atoms of each element in tested compositions can also be set manually or calculated automatically. In the performance assessment part of this study, respective numbers are calculated automatically based on user-defined lower and upped bounds (r_{min} and r_{max} parameters, respectively) of the m/z range where identification is carried out. The detailed description of calculation procedure is given in the Supporting Information (same section mentioned previously).[14] Note that at the end of generation, any compositions that satisfy or do not satisfy certain conditions can be deleted from the resulting set using array indexing options.

Calculation of Theoretical Isotopic Distribution. The check for the presence of an ion with certain composition in the mass spectrum starts with the generation of a theoretical isotopic distribution.

The vast majority of programs and software packages designed for calculating theoretical isotope distributions implement algorithms that are based on one of two approaches. The first approach involves the expansion of polynomials describing the isotope abundances of elements that form the ion. The second approach implies the usage of convolution, namely, the Fourier transform. A detailed overview of the mathematics and history of both approaches is given in ref 21.

The advantages and limitations of the two named approaches are well-known. Polynomial expansion allows the preservation of information about the exact isotopic composition of each peak. However, since calculations have

https://doi.org/10.1021/jasms.4c00153
J. Am. Soc. Mass Spectrom. 2024, 35, 1806−1817

to be performed sequentially, the expansion of polynomials is time- and memory-expensive for compositions containing a large number of atoms.[22,23] In contrast, Fourier transform allows the memory usage and calculation time to be minimized; however, the information on the isotopic composition of peaks is lost.

Since late 2000s, a new concept has been developed to calculate the isotopic distribution while preserving the exact isotopic compositions of peaks and reducing the computation time.[22] This concept utilizes a polynomial-like description of isotopic abundances and a "hierarchical" calculation scheme. Hierarchical calculations imply that the most probable isotopologues are computed first. This allows the calculation of isotopic distribution to be stopped once a user-defined cumulative relative intensity of peaks is reached and avoids the computation of "rare" isotopologues.

In this study, all the theoretical distributions were calculated using IsoSpec algorithm version 2 (IsoSpec2), which is based on the previously mentioned hierarchical concept.[23,24] The selection of IsoSpec2 in particular is based on the linear asymptotic complexity of this algorithm, which means that the calculation time grows at a rate no greater than the linear one as the total number of isotopologues for a given molecule increases. Furthermore, IsoSpec2 is implemented in the C++ programming language, which also allows us to expect a low absolute calculation time. Version 2 was selected due to the better reported performance compared to the original IsoSpec algorithm.[24]

Two difficulties of computational nature can be manifested at the theoretical distribution calculation step, and both of those difficulties can seriously impact identification results.

Handling the Rare Isotopologues. The first difficulty is related to the processing of isotopologues with a low probability of occurrence. As mentioned in previous studies,[13,14] the peaks corresponding to such isotopologues may have background intensities or may not be registered in the experimental mass spectrum at all for various reasons. Thus, retaining "too rare" isotopologues in the theoretical distribution is impractical.

In many implementations of polynomial and convolution algorithms, the above problem is solved by discarding theoretical peaks with a relative intensity lower than the threshold. However, as highlighted in ref 22, the choice of such a threshold value is not obvious. Furthermore, a simple removal of theoretical peaks with lower-than-threshold intensity does not allow for controlling the cumulative probability of occurrence of discarded isotopologues.

In contrast, hierarchical algorithms, where the most probable isotopologues are computed first, allow the *coverage* of the theoretical distribution to be controlled. Instead of the relative intensity threshold, hierarchical algorithms require the specification of the minimum cumulative relative intensity of the peaks in the resulting distribution. Such cumulative intensity can be interpreted as the cumulative probability of observing all the computed isotopologues in the experimental mass spectrum and, therefore, shows the share of the complete distribution covered by the calculation. Because of that, such a parameter is used as a calculation stop criterion in hierarchical algorithms.

The choice of an appropriate cumulative probability value is also not obvious. However, unlike the simple discarding of peaks with intensity lower than a threshold, the calculation of theoretical distribution until the specified cumulative probability is reached ensures that the coverage of the calculation is the same for all tested compositions. Additionally, such coverage does not depend on the number of atoms in the tested composition.

Formally, the minimum cumulative probability can be set equal to any value in the range [0; 1]. Confidence levels of 0.9, 0.95, 0.99, and, rarely, 0.999 are commonly used in mathematical statistics when hypothesis testing is conducted. The minimum cumulative probability for the isotopic distribution calculation can be set to one of these values as well.

Note that lower values of the calculation stop criterion imply greater loss of information regarding the theoretical distribution. This is especially relevant when the inorganic ion being identified contains many atoms of elements that have the one most abundant isotope and several additional isotopes with low (up to ~1%) abundances. The probability of observing the rare isotopologues in the experimental mass spectrum increases as the number of atoms of such elements grows. At a certain point, peaks corresponding to such rare isotopologues will become clearly visible in the experimental mass spectrum. However, those isotopologues may be missing from the theoretical distribution if the calculation is stopped early.

An illustration of relatively large information loss can be found in Figure 1, which depicts the theoretical isotopic distribution of $Re_5S_8^+$ ions calculated up to a cumulative probability of 0.9. As can be seen in Figure 1, isotopologues corresponding to four peaks that are clearly distinguishable in the experimental mass spectrum were not computed at such a calculation stop criterion value. These four isotopologues together accounted for 0.0409 probability of occurrence.

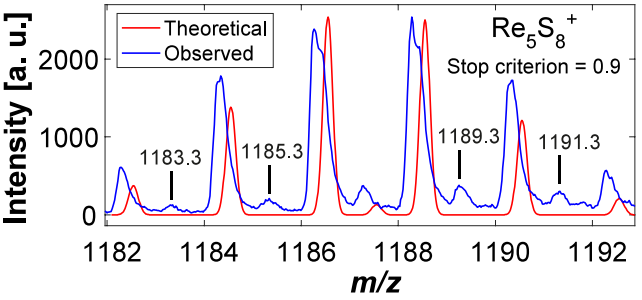


Figure 1. Comparison of observed (blue line) and aggregated theoretical (red line) isotopic distributions of the $Re_5S_8^+$ ion. Theoretical isotopic distribution was calculated up to a cumulative probability of 0.9. Note that information on four "odd" theoretical peaks is lost.

Further in this study, all theoretical isotopic distributions are calculated up to cumulative probability of 0.99, unless otherwise stated explicitly. Such a value was chosen to minimize the loss of information incurred due to stopping the calculation early and to reduce the effect of information loss on the comparison of distributions.

Adjusting the Resolution of the Theoretical Isotopic Distribution. The second difficulty that arises from the calculation of the theoretical isotopic distribution involves the adjustment of the resolution of theoretical distribution. Some of the calculated peaks can be separated from each other by very low m/z values, e.g., <0.01 Da. Peaks with such an m/z difference cannot be resolved if the experimental mass spectrum is acquired on low-resolution equipment. This

https://doi.org/10.1021/jasms.4c00153
J. Am. Soc. Mass Spectrom. 2024, 35, 1806−1817

phenomenon can be illustrated with the example of $PbCl_3^-$. The complete theoretical isotopic distribution of this ion contains 16 isotopologues (see Table S1 in the Supporting Information for the peak list). However, when the $PbCl_3^-$ ion was registered on the used mass spectrometer, 6 out of 16 theoretical peaks were not resolved, as depicted in Figure 2.

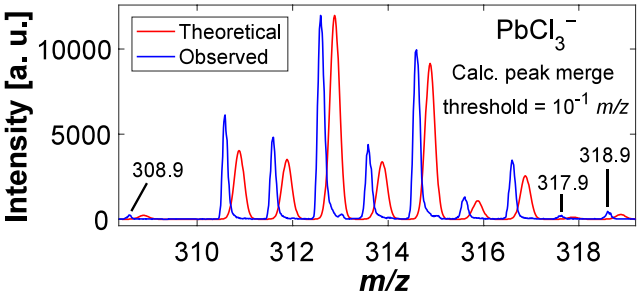


Figure 2. Comparison of observed (blue line) and aggregated theoretical (red line) isotopic distributions of the $PbCl_3^-$ ion. The threshold for merging neighboring theoretical peaks was 0.1 m/z. Note that resolutions of observed and theoretical distributions agree well following the merger of calculated peaks.

In order for theoretical and experimental isotopic distributions to be compared correctly, the "resolution" of the theoretical distribution must match the resolution of the experimental mass spectrum. This can be achieved by merging those calculated peaks that are separated by an m/z distance less than the specified threshold. In this study, the merger of the neighboring calculated peaks was conducted using the same approach as in IsoPro software,[25] which is an implementation of the J. Yergey polynomial algorithm.[26] Such an approach implies that all consecutive peaks separated by an m/z value less than a threshold are merged into one aggregated peak. The aggregated peak is assigned an m/z value of the most intense peak among the merged ones and a cumulative intensity of all merged peaks.

We failed to find recommendations on the selection of the m/z threshold for merging neighboring calculated peaks in the literature. In our opinion, the value of this parameter can be set to 10^{-k}, $k \in N$ for simplicity purposes. The value of the exponent k should be selected based on the characteristics of the equipment used and should reflect the last decimal place up to which the mass of the registered ion can be accurately measured using the given equipment. The merger of neighboring theoretical peaks inevitably undermines the performance of brute force. However, if the resolution of the theoretical distribution is not adjusted, meaningless identification results may be obtained. In this study, the k value was set to 1, i.e., the threshold distance for peak merge amounted to 0.1 m/z. As shown in Figure 2, merging peaks from the theoretical distribution of the $PbCl_3^-$ ion resulted in good agreement of observed and aggregated theoretical data when the threshold distance was set to 0.1 m/z. This indicates that the chosen threshold value for the m/z distance is sufficient for use under the described conditions.

Matching the Theoretical and Observed Peaks. Once the theoretical isotopic distribution is calculated, theoretical peaks have to be matched with signals observed in the experimental mass spectrum. A very simple concept is used to perform peak matching in various mass spectrum processing techniques that involve such matching. This concept implies that the search for observed peak that matches a theoretical peak with m/z value of t_i (I = 1, ..., n_m) is conducted in the range $[t_i - \varepsilon, t_i + \varepsilon]$ of the experimental mass spectrum, where ε is the used-defined mass tolerance.[27−29]

The value of mass tolerance ε can be selected in a manner similar to the one described earlier for the peak merge m/z threshold value, i.e., as 10^{-k}, $k \in N$. However, in practice, a multiplication of the mass tolerance by a scalar equation is commonly applied in order to account for small calibration errors. Thus, the mass tolerance can be set to $a \times 10^{-k}$, $k \in N$, $a > 0$. In particular, the value of $\varepsilon = 2 \times 10^{-2}$ Da was earlier proposed for high-resolution data.[30] Additionally, a peak binning method for the dynamic adjustment of the ε value was described in the literature.[27] However, we believe that the application of dynamic peak binning as part of isotopic distribution brute force is excessive and may influence identification results in an unexpected way.

The observed peaks that match theoretical ones can be selected either from the list of profile or centroid mass spectrum signals or from a list of peaks detected using some method. It is evident that the amount of arithmetic operations will decrease as the dimensions of the searched array decrease. Thus, the preliminary dimensionality reduction and peak detection will reduce the time required for execution of the isotopic distribution brute force.

Depending on the dimensionality of the array that is used to select the matching observed peaks, two special cases are possible. If the match is selected from a "short" peak list, there is a possibility that relatively low-intensity signals corresponding to rare isotopologues of the detected ion will not be included in the above list after the peak detection algorithm finishes its work. In this case, the search range $[t_i - \varepsilon, t_i + \varepsilon]$ may contain no experimental signals that could be matched with the given theoretical peak. The second case is directly opposite and implies that the search range $[t_i - \varepsilon, t_i + \varepsilon]$ contains several experimentally observed signals.

The introduction of fictitious zero-intensity peaks[27] and the application of highly sensitive peak detection algorithms based on the calculation of the signal-to-noise ratio[29,31] were previously proposed in the literature as the solutions to the first case problem. The problem arising from the second case apparently received little attention in previous studies. In order to minimize the influence of both mentioned difficulties, in this study each theoretical peak is matched with the experimental signal characterized by maximum intensity in the $[t_i - \varepsilon, t_i + \varepsilon]$ search range, and the searched array is taken from profile mass spectrum. A more detailed theoretical discussion of this aspect can be found in the Supporting Information, section "Matching the theoretical and observed peaks".

Comparison of Theoretical and Observed Isotopic Distributions: Goodness-of-Fit Criteria. In the course of isotopic distribution brute force, the evaluation of each tested composition is concluded by comparing the theoretical isotopic distribution with the corresponding observed distribution.

Earlier we formulated the task solved during the identification of ions by means of brute force as testing the hypothesis of whether the theoretical and observed isotopic distributions are identical. This type of task can technically be solved using the goodness-of-fit tests defined in mathematical statistics. Nevertheless, with few exceptions,[32] goodness-of-fit criteria are not used to test the identity of two isotopic distributions in practice.

https://doi.org/10.1021/jasms.4c00153
J. Am. Soc. Mass Spectrom. 2024, 35, 1806−1817

The calculation of goodness-of-fit test statistics involves operations with frequencies and sample size.[33−37] Within the scope of task solved, absolute intensities of observed peaks that match the peaks from the theoretical distribution can be naturally interpreted as frequencies. Respectively, the sample size is given by the cumulative intensity of peaks in the observed distribution. Such cumulative intensity may vary from several dozen to several hundred thousand absolute units for various ions. Meanwhile, at least the most commonly used goodness-of-fit criteria are known to have low statistical power, i.e., the ability to correctly reject the null hypothesis when the alternative is true, when sample size is either very small or very large.[33,38,39] Estimates of the sample sizes required can be found in the corresponding section of the Supporting Information.

In practice, all of the above leads to obtaining meaningless identification results when using goodness-of-fit criteria to compare theoretical and observed distributions in the course of isotopic distribution brute force. In such a case, multiple unrealistic elemental compositions would be assigned to low-intensity (yet distinguishable) ions. In the meantime, the goodness-of-fit tests would not pass when evaluating the isotopic distributions of true compositions for high-intensity ions. Figure S10 and Table S2 from the Supporting Information depict the example of this phenomenon for ions $Re_5S_8^+$, $Re_5MoS_8^+$, $PbCl_3^-$, and $Pb_2Cl_3O^-$.[40] Thus, goodness-of-fit tests cannot be used to compare isotopic distributions during ion identification by means of brute force.

Comparison of Theoretical and Observed Isotopic Distributions: Vector Similarity Measures. The task of testing the identity of theoretical and observed isotopic distributions can be seen in an alternative way. Once the theoretical distribution is calculated and each theoretical peak is matched to the experimental one, two relative intensity vectors of equal length are formed. Thus, the decision on the identity of distributions can be made using the vector similarity measures.

Such an approach to the comparison of mass spectra (and their parts) has been studied much better than the application of goodness-of-fit criteria and is particularly used in library search. A detailed review of similarity measures that are commonly used to compare intensity vectors of two mass spectra was presented in previous papers.[41,42] The theoretical and observed isotopic distributions of an individual ion can also be considered as separate mass spectra and thus can be analyzed using such vector similarity measures.

Cosine similarity and its numerous modifications are the vector similarity measures that are the most widely used in mass spectrometry. Recall that the original cosine similarity is calculated as

$$S_C(X^{(m)}, T^{(m)}) = \frac{\sum_{i=1}^{n_m} x_i^{(m)} t_i^{(m)}}{\sqrt{\sum_{i=1}^{n_m} (x_i^{(m)})^2} \sqrt{\sum_{i=1}^{n_m} (t_i^{(m)})^2}}$$

Given the non-negativity of intensities, the unmodified cosine similarity may take values in the range [0; 1]. Two modifications of cosine similarity that scale the resulting value to [0; 1000) range were listed in ref 42. One of those modifications, namely, the identity match factor, is known to be directly used in library search software.

Norms, i.e., geometric distance measures, are also used to quantify the similarity between two mass spectra. In particular, l_1 and l_2 norms (Manhattan and Euclidean distance, respectively) were applied in the practical part of study.[42] The use of the mentioned norms involves difficulties when selecting the threshold value for distributions to be considered identical, since the resulting calculated distance has no easily interpreted upper bound (e.g., 1). However, we believe that this inconvenience can be eliminated by applying l_1 and l_2 norms together with weights and indicator functions (see the Supporting Information, section "Using the weights and indicators together with geometric norms").

Pearson correlation coefficient (yields values in range [-1; 1]), Wasserstein distance, and partial correlation were also suggested as alternative measures of similarity of mass spectra in previous studies.[41,42] We shall note that Wasserstein distance has no fixed upper bound, like the l_1 and l_2 norms, and calculation of the partial correlation in the form proposed in ref 41 requires a large number of arithmetic operations. Thus, we believe that the aforementioned alternative metrics cannot be easily used as the measures of isotopic distribution similarity in the course of brute force.

Data presented in the literature show that the quality of comparison of mass spectra attained by application of traditional vector similarity measures may vary significantly depending on the set of input data used. For example, the usage of "modified" cosine similarity in ref 43 resulted in a maximum precision of over 0.8; meanwhile, a precision of only ∼0.5 was reported for the original cosine similarity.[41] However, it is commonly agreed that similarity measures based on cosine distance allow those reference mass spectra that evidently do not match the experimental data to be reliably excluded from consideration,[41−43] which is sufficient for isotopic distribution brute force. In the experimental part of this study, we use the unmodified cosine distance as the spectral similarity measure.

Conceptual Limitations of Ion Identification by Means of Isotopic Distribution Brute Force. The described approach to the automated identification of ions has several conceptual limitations. First of all, for obvious reasons, the isotopic distribution brute force cannot be used to identify ions composed only of monoisotopic elements. As mentioned earlier, only the elements that have stable isotopes are considered in this study for simplicity purposes. Of 80 such chemical elements, 26 are monoisotopic.[12] Thus, isotopic distribution brute force can technically be used to identify ions formed of $\frac{80-26}{80} = 71.25\%$ elements with known stable isotopes.

Next, among chemical elements that have two or more stable isotopes, there are at least eight elements (notably N and O) for which around 99% of natural abundance is attributable to just one isotope.[12] While the identification of ions composed only of such elements (and, additionally, any monoisotopic elements) is technically possible, meaningful results are unlikely to be obtained under such conditions in practice. As mentioned earlier, the peaks corresponding to rare isotopologues may not be registered in the mass spectrum in some cases, since the real abundances of isotopes in the sample may differ from the reported average values.

One special case is worth mentioning separately. Certain ions may simultaneously contain elements, the integer mass of one of which is a multiple of the integer mass of the other. For example, a series of clusters with the most intense peaks placed at distances of ∼16 Da was observed when the mass spectra of natural rhenium sulfide were acquired in negative ion mode

https://doi.org/10.1021/jasms.4c00153
J. Am. Soc. Mass Spectrom. 2024, 35, 1806−1817

during this study. Such a mass difference may indicate the presence of oxygen. The mass of ^{32}S, which is the most abundant sulfur isotope, amounts to ~31.9721, which is approximately equal to the mass of two atoms of the ^{16}O oxygen isotope ($2 \times 15.9949 \approx 31.9898$).

The mass difference of $31.9898 - 31.9721 = 0.0177$ Da can only be detected using high-resolution equipment. Assuming an ion mass of 1000 Da and a charge state of 1, the required resolution in this case is $\frac{1000}{0.0177} \approx 56500$ units. However, the application of high-resolution mass spectrometers still cannot be considered as common practice. Meanwhile, the "replacement" of one S atom by two O atoms does not alter the resulting theoretical isotopic distribution significantly, since isotopes ^{32}S and ^{16}O account for ~95% and ~99% of theoretical S and O abundance, respectively. As can be seen in Figure 3, the theoretical isotopic distributions of $Re_2S_5^-$ and

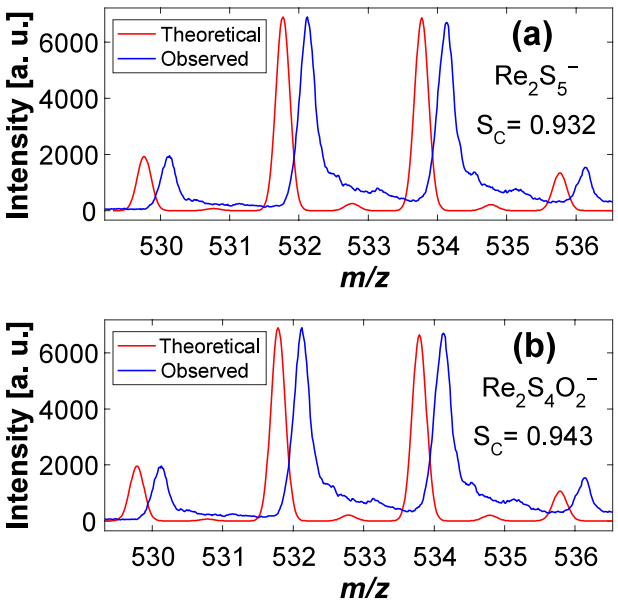


Figure 3. A comparison of theoretical (red line) and observed (blue line) isotopic distributions for (a) $Re_2S_5^-$ and (b) $Re_2S_4O_2^-$ ions. Note that both theoretical distributions are nearly identical.

$Re_2S_4O_2^-$ ions are almost identical (cosine similarity yields values of 0.932 and 0.943, respectively). Under conditions of low resolution, only a list of candidate compositions of the given ion can be obtained by using the isotopic distribution brute force in mentioned cases.

In addition to the discussed conceptual shortcomings of specific isotopic distribution brute force, every brute force technique is prone to low performance and reduced effectiveness when the amount of processed input data is high, i.e., the search space is large. In such cases, a performance drop is driven by the need to perform a large number of arithmetic operations, and lower effectiveness is a consequence of higher probability of random false-positive matches. However, such negative effects may not be pronounced until a certain number of objects, e.g., hypothetical compositions, as in the described case, are tested. In the next sections, we attempt to determine the constraints on the input data, subject to which the isotopic distribution brute force produces useful results both effectiveness-wise and performance-wise.

EXPERIMENTAL SECTION

As part of this study, dedicated software was developed in IPCE RAS in order to determine whether the isotopic distribution brute force can be used to identify inorganic ions in practice. The program was written in C++. The IsoSpec2 package (C++ API, GPL license) is used by the program internally to calculate theoretical isotopic distributions.[24] The graphical interface was created using the WxWidgets library (GPL license). All other functions were implemented using the standard library of C++. Additional technical details on program implementation can be found in the Supporting Information, section "Implementation of software for isotopic distribution brute force".

Two practical aspects of the proposed approach were evaluated during the experiments. The first evaluated aspect was the ability to correctly identify inorganic ions with various compositions and signal-to-noise ratios by means of the proposed algorithm. Test identification was conducted on a data set containing seven low-resolution laser-desorption ionization mass spectra of lead halides and copper halides. All the mass spectra in the data set were acquired during a real investigation of the surface of construction material according to the method described in ref 17. The data set notably included two mass spectra of lead(II) halides placed on a copper support. This allowed us to test the ability of the algorithm to establish the composition of heterometallic clusters based on the observed isotopic distribution. Graphical representations of all the spectra from the data set are shown in Figures S3–S9.

A total of 61 ions, including 51 unique species, were manually identified in the mass spectra from the tested data set. The latter number comprised 9 copper bromides, 6 copper chlorides, 3 mixed chlorine-bromine copper clusters, 5 lead chlorides, 10 oxide-chloride lead clusters, 4 lead oxide and lead hydroxide species, 4 lead clusters with potassium adducts, and individual element ions of lead and bromine. Full lists of ions for each of the mass spectra from the data set can be found in Tables S4–S10.

During the trial identification, different sets of allowed elements were used for each mass spectrum from the data set depending on the studied compound. Tested compounds were generated from the minimum and maximum numbers of atoms. The minimum number of atoms of each element was always set to 0. In order to imitate the assumed common scenario of program usage by the researcher, the maximum number of elements was equal to 5 in the majority of the cases. As an exception, a maximum of 1 atom of those elements that commonly form adducts under laser desorption–ionization conditions (K, H, and, in the case of the simultaneous presence of chlorine and bromine in sample, Br or Cl) was allowed. Complete sets of allowed elements and maximum allowed number of atoms are presented in Table S11 of the Supporting Information.

For the test identification purposes, the minimum value of cosine similarity for a tested composition to be considered a candidate for the true composition of the ion was set to 0.95. The calculation of the local signal-to-noise ratio for the most intensive peak in the observed distribution was additionally performed to filter out unwanted assignments of hypothetical compositions to background signals. Computation of the local signal-to-noise ratio was based on the method presented in ref 43 and is described in the Supporting Information, section

https://doi.org/10.1021/jasms.4c00153
J. Am. Soc. Mass Spectrom. 2024, 35, 1806–1817

"Calculation of local signal-to-noise ratio for validation of results of test identification" in detail.

Following the assessment of identification quality, an evaluation of performance was carried out to determine whether the isotopic distribution brute force was computationally viable. The assessment of isotopic distribution brute force runtime was conducted using two mass spectra as input data, namely, the mass spectrum of $PbCl_2$ (negative ion mode, Figure S8) from the data set used previously and the mass spectrum of natural rhenium sulfide (positive ion mode, Figure S1). During the performance assessment, the number of studied mass spectra was intentionally limited to 2, since the performance of the isotopic distribution brute force depends on the number of arithmetic operations performed. Thus, by increasing the number of tested compositions and making the compositions more complex, it is possible to obtain representative data on the performance of isotopic distribution brute force using a limited number of input data and study asymptotics in more detail.

During the performance assessment, tested elemental compositions were generated in an automated manner using data on identification range bounds, as described in the theoretical part (Input Data) and in the Supporting Information. The identification range was set to [0; 2000] Da, since most of nonbackground signals registered in the studied mass spectra were observed in the respective m/z range. The sets of allowed chemical elements comprised a maximum of four elements (Re, Mo, S, and O for rhenium sulfide and Pb, Cl, O, and K for $PbCl_2$). The presence of these elements in samples was either identified during manual processing or established based on preliminary information about the sample.

Performance measurements were conducted on a regular Acer Aspire A317-51KG office laptop (manufactured in 2019). Detailed specifications of the laptop are presented in Table S13. A total of 21 different identification scenarios were run. Each identification scenario was executed 50 times in order to reduce the influence of instant CPU load by other applications. The average total runtime and the average time of execution of individual brute force operations are presented in Table S14 of the Supporting Information for each calculation scenario.

Parameter values used at various stages of the brute force are listed in Table S3. Except for the peak matching search range, all parameters were set to the values that were previously discussed in theoretical part. Since some of the used mass spectra required nonlinear internal calibration, the search for matching peaks was conducted on an extended range of ±0.5 m/z so as not to introduce the effect of calibration parameters on identification result. As mentioned earlier, profile mass spectra were searched when peak matching was conducted. This allowed the elimination of the possible influence of peak detection or centroiding parameters on the observed identification quality. Furthermore, the used format of mass spectrum files (Bruker Flex/XMASS) implies that time-of-flight and, respectively, m/z values are recorded at a fixed discretization rate.[44,45] Thus, searching the profile mass spectrum for matching peaks also contributed to correct assessment of the asymptotics of peak matching execution time, since each search range contained roughly the same number of signals.

In addition to the parameters discussed in the theoretical part, the program accepts another input parameter, namely, the assumed charge state of ions to be identified. This enables the program to be directly used for the identification of ions with a user-specified charge state. For simplicity purposes, the charge state was assumed to be equal to 1 throughout this study.

RESULTS AND DISCUSSION

Test Identification. The aggregated and spectrum-wise results of the automated identification of lead and copper halide ions by means of isotopic distribution brute force are presented in Table S12 of the Supporting Information.

During the test practical application of isotopic distribution brute force, correct elemental compositions were successfully established for 58 ions, i.e., for 95% of all the ions identified manually. This notably included all nine occurrences of seven unique heterometallic Pb−Cu clusters, the theoretical and observed isotopic distributions of some of which are depicted in Figure S11. The ability to identify such clusters by isotopic distribution brute force may find use in, e.g., studies of alloys.

For 50 ions, or 86% of ions successfully identified by isotopic distribution brute force, the median rank of the true ion composition in the list of candidates ordered by cosine similarity value in decreasing order was equal to 1. Note that the median rank of the true composition in the list of candidates was equal to 1 when measured both for the entire data set and for each mass spectrum individually. The mean rank of true compositions amounted to 1.42 among all of the identified ions. In our opinion, such results confirm the robustness of the comparison of isotopic distributions by means of cosine similarity.

An important issue of all of the brute force techniques is the increasing number of false positive matches that arise from pure arithmetical reasons as the search space increases. This issue manifested during the test identification. A total of 281 elemental compositions were reported as "found" in the entire used data set of mass spectra. Of those, 136 compounds were the candidate compositions for ions identified manually, and the other 148 compositions presumably were false-positive matches. Such a proportion of around one candidate match to one false-positive match was maintained for each individual mass spectrum from the data set as well.

However, the numbers of matches was unevenly distributed across the data set. For identification scenarios with up to four elements (mass spectra no. 1−3, 6, and 7) and up to around 400 tested formulas, no more than 25 potentially matching compositions per mass spectrum were found by isotopic distribution brute force. In our opinion, the researcher should be able to extract useful information from identification results in such cases, especially given that some of the mismatches can be additionally filtered out by applying restrictions on the signal-to-noise ratio. On the other hand, identification scenarios where 5 and 6 elements were allowed (mass spectra 4 and 5; 2500 and 5000 compositions tested) resulted in over 50 and over 150 matches, respectively. Such results are less likely to be adequately processed by the user following the end of the identification. However, note that some of those matches could also have been avoided by a proper calibration of m/z values in mass spectra and by the use of more strict mass tolerance for peak matching.

Unwanted assignments of hypothetical compositions to low intensity peaks or background signals can be prevented by invalidating the result of distribution comparisons for compositions with a signal-to-noise ratio (SNR) of highest observed peak lower than the threshold. The ions that were successfully identified by isotopic distribution brute force were

https://doi.org/10.1021/jasms.4c00153
J. Am. Soc. Mass Spectrom. 2024, 35, 1806−1817

characterized by a double-digit median SNR of the most intense peak, which amounted to ∼34 for the entire data set and exceeded 18 for each individual mass spectrum. The minimum SNR of the highest peak among identified ions was roughly equal to 6. The corresponding values for assumed false-positive matches were several times lower. As it can be seen from Figure 4, discarding the identification results for

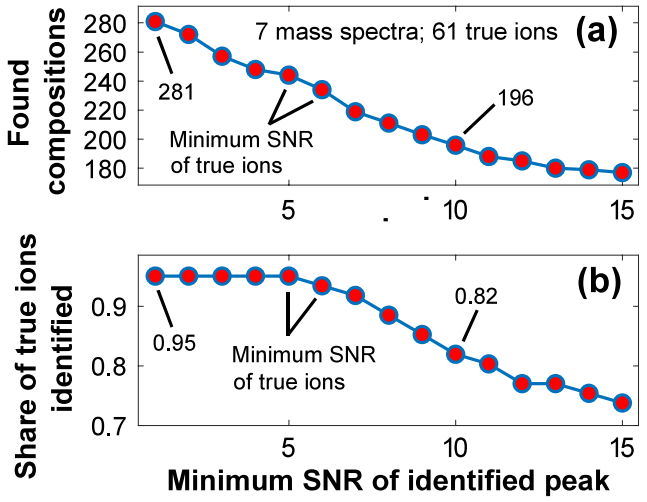


Figure 4. Influence of minimum signal-to-noise ratio restriction on (a) the total number of compounds allegedly found in studied mass spectra during isotopic distribution brute force and (b) the share of true ions that were correctly identified by isotopic distribution brute force.

signals with a SNR lower than 10 would reduce the total number of potentially found compositions by almost 30% to 196. This would come at a cost of failing to identify eight true ions during the "batch" search over the entire spectrum, with the share of identified ions dropping to 82%. Note that the unidentified ions would be distributed over mass spectra relatively evenly in such a case. However, filtered-out ions may be successfully identified during a "targeted" isotopic distribution brute force with higher restrictions on allowed composition and lower restrictions on SNR.

Based on the above, we may suggest the following practical recommendations for isotopic distribution brute force to be used effectively. When the processing of a new mass spectrum is started, the set of allowed elements should comprise only the main chemical elements that are expected to be present in the studied sample. Later, the amount of tested compositions should be increased gradually, e.g., by adding allowed elements. Once some of the ions are identified and the need to test more complex elemental compositions arises, the researcher should introduce additional restrictions, e.g., by limiting the allowed number of atoms, to effectively use the proposed approach for identification of the remaining ions of interest. The brute force results for signals with single-digit (less than 10) SNR values of the top peak should be evaluated carefully and may be discarded during the initial search runs. The user should remember that the observed peaks corresponding to rare isotopologues would have even lower SNR values. Conducting loosely constrained brute force with more than four allowed elements over the entire mass spectrum is not recommended. Targeted isotopic distribution brute force may later be conducted to identify selected ions that were not identified during the search over the entire mass spectrum.

Performance Assessment. The results of the performance measurements revealed no clear dependence of the brute force runtime on any single numeric characteristic of the tested compositions set. However, such dependencies were observed for some of the individual operations. The execution time of the theoretical isotopic distribution calculation also exhibited two competing dependencies on two parameters.

In all the considered identification scenarios, the calculation of theoretical isotopic distributions for tested compositions by means of IsoSpec2 algorithm constituted the largest (∼50−75%) part of the total brute force runtime. Per the authors of IsoSpec2, the time of isotopic distribution calculation should increase linearly along with the increase in the number of isotopologues for input composition.[23,24] However, in practice, two competing asymptotics are observed for identification scenarios, which involve testing the compositions with similar allowed elements. When the total number of isotopologues for all tested compositions was relatively low, the time of isotopic distribution calculation (and the total brute force runtime) increased linearly and was almost directly proportional to the number of tested compositions. This can be seen in Figure 5a,

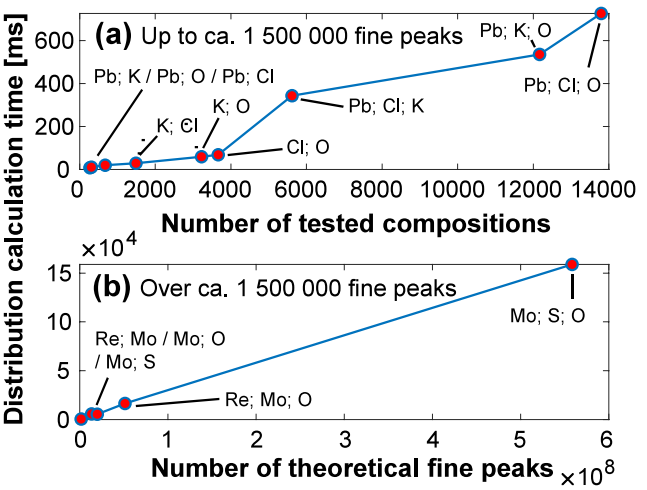


Figure 5. Runtime asymptotics of calculation of theoretical isotopic distributions (a) for calculation scenarios where up to ca. 1.5 million isotopologues were computed premerger and (b) for calculation scenarios where over 1.5 million peaks were computed premerger. Scenarios with similar allowed elements are shown.

where the distribution calculation time for scenarios with various combinations of Pb, Cl, O, and K allowed elements is plotted against the number of tested compositions. We believe that such a dependency is observed due to implementation peculiarities. Our implementation of brute force involves creating a new object of the IsoSpec::FixedEnvelope class, which is a wrapper over distribution generator, for the calculation of each theoretical isotopic distribution. It is likely that until a certain number of isotopologues (or theoretical fine structure peaks) is reached, the time for actually calculating the distribution is negligible compared to the time required to initialize the fields of the FixedEnvelope class. Once the threshold number of isotopologues is exceeded, the time of the distribution calculation increases linearly and proportionally to the total number of fine peaks (see the graph in Figure 5b for scenarios involving Re, Mo, S, and O elements). This agrees

https://doi.org/10.1021/jasms.4c00153
J. Am. Soc. Mass Spectrom. 2024, 35, 1806−1817

with the claimed linear asymptotics of the IsoSpec2 algorithm.[22,23] Based on obtained results, we believe that the change of the calculation time dependency parameter occurs when the total number of isotopologues exceeds ∼1.5 million, which can be seen when the performance results for {Pb, Cl, O}, {Pb, Cl, K} and {Pb, K, and O} scenarios are examined. Under the described conditions, the specified threshold was exceeded when approximately 15 000 compositions were tested.

The second largest contribution to the total runtime came from the sorting of peaks within theoretical isotopic distributions by the m/z value. This sorting is necessary to adjust the resolution of the theoretical distribution in further steps. The sorting time depends on the number of m/z array element swaps. In general, such a number of swaps cannot be estimated in advance. The number of swaps and, accordingly, the sorting time increase with the total number of isotopologues. However, such a dependency cannot be unambiguously described by any law. As the total number of calculated isotopologues grows, sorting constitutes an increasingly larger share of the execution time (an increase from ∼5% to 50% in the cases considered). Under the described conditions, sorting was completed within acceptable time (several dozens of milliseconds) when the number of tested compositions did not exceed 17 000, as in {Pb, K, O}, {Re, S, and O}, and {Pb, Cl, and O} scenarios. We believe that the peak sorting operation should be optimized first if the software implementing isotopic distribution brute force is developed further.

Merging neighboring calculated peaks during the adjustment of resolution of theoretical distribution made the third largest contribution (∼1−5%) to the total brute force runtime. Theoretically, the execution time of this operation should depend linearly on the number of merges. In practice, however, such a dependence was violated in some cases. We believe that those violations originate from peculiarities of implementation. In our implementation of peak merge, the indices of retained m/z values are inserted into the temporary vector without prior memory reservation. If a large number of peaks are retained in distributions following the peak merger, a significant amount of memory will have to be dynamically allocated, which implies the repeated execution of computationally expensive copying operations. Because of that, the time required for the completion of the peak merger operation in scenarios that involve retaining many peaks, e.g., {Re, O, S} and {Cl, K, and O}, may exceed the respective time for scenarios where many peaks are merged. Reserving the memory for temporary vector will likely reduce the peak merger time at the cost of increased memory usage in the course of brute force, which may be a limiting factor if operations are executed in a parallel manner. The merger of neighboring calculated peaks was accomplished within acceptable 100−200 ms for all realistic identification scenarios (up to 60 000 tested compositions, as in the {Cl, K, and O} scenario) when the current implementation of brute force was tested.

Calculating the theoretical isotopic distribution, sorting calculated peaks by m/z, and merging neighboring calculated peaks together accounted for ∼75−99% of the total identification time in all the considered cases. Thus, the three mentioned operations determine the performance of the isotopic distribution brute force.

The time required to execute two remaining steps, namely, the peak matching and comparison of distributions (including the calculation of cosine similarity), increased linearly with the number of aggregated peaks after the merger. Runtime asymptotics of peak matching are shown in Figure 6. Since

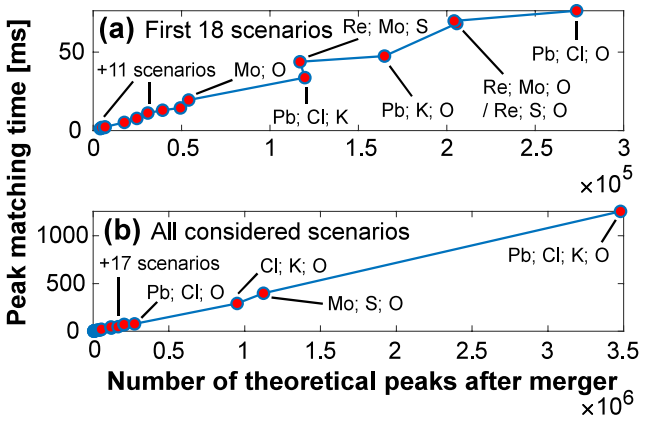


Figure 6. Runtime asymptotics of peak matching operation: (a) first 18 scenarios ordered by peak matching time and (b) all considered scenarios. Note the linear dependence of the peak matching execution time on the number of peaks after the merger.

the shapes of asymptotic graphs are very similar for both of these operations, the corresponding graph for distribution comparison is not shown here and is depicted in the Figure S12 instead.

Similar linear dependence of execution time on the number of merged peaks was observed for some technical operations, e.g., the collection of information to be displayed in identification results (asymptotics is shown Figure S13). The execution time of another technical operation, namely, the division of calculated m/z values by specified charge state, also behaved as expected and depended linearly on the total number of isotopologues before the merger, with an exception of the {Re, Mo, S} case (see Figure S14). All four mentioned operations together were accomplished in less than 0.5 s when all realistic identification scenarios were considered (up to 60 000 tested compositions and up to 1 million peaks after merger). Thus, we may conclude that peak matching, comparison of distributions, and technical operations do not limit the performance of isotopic distribution brute force.

Finally, we shall note that a significant reduction in performance of isotopic distribution brute force was observed when Mo was included into the set of allowed elements. The asymptotic dependences presented earlier were not violated; however, the absolute execution time was several times higher than that in scenarios that did not involve the testing of Mo-containing compositions. For example, evaluation of 124 compositions in the {Re, Mo} scenario was completed in ∼1000 ms versus ∼14 ms for the {Re, S} option where 371 formulas were tested.

The exact reason for such a sharp increase in execution time is unclear. However, we assume that this phenomenon is due to the internal logic of the IsoSpec2 algorithm. In the examples discussed above, Re and Mo have a total of nine isotopes, while Re and S together have 6 isotopes. Theoretically, each increase in the number of isotopes of allowed elements greatly increases the number of comparisons and other arithmetic operations required to obtain the smallest possible set of isotopologues. This is indirectly confirmed by the fact that the time of distributions calculation in the {Re, Mo} scenario is

https://doi.org/10.1021/jasms.4c00153
J. Am. Soc. Mass Spectrom. 2024, 35, 1806−1817

comparable to those in the {Re, S, O} and {Pb, K, O} cases (498 ms versus 536 and 580 ms, respectively; allowed elements have a total of 9, 8, and 8 isotopes). Thus, the number of isotopes of allowed elements also influences the performance of the isotopic distribution brute force.

CONCLUSIONS

The results of trial identification of ions observed in the studied mass spectra have shown that the isotopic distribution brute force is capable of correctly establishing the elemental composition of various inorganic ions, including those of rather complex nature, e.g., heterometallic clusters.

The effectiveness and the performance of the isotopic distribution brute force in the described implementation are influenced by several parameters that characterize the input data. Values of some of those parameters cannot be estimated before the identification is conducted. This does not allow us to formulate generalized numeric constraints on input data, subject to which the application of isotopic distribution brute force is feasible from the standpoints of results interpretability and identification time.

However, based on results of trial practical application, we suggest that the isotopic distribution brute force can be successfully used to conduct automated identification of inorganic ions composed of up to four chemical elements. In such a case, a brute-force search over the entire spectrum produces countable and small number of matches, commonly up to 20−30 formulas, with more than half of matches being true compositions of observed ions. In our opinion, such results can be adequately processed by the researcher. The correct compositions of ions that simultaneously contain more elements, e.g., 5 or 6, could also be determined by isotopic distribution brute force; however, additional filtering of results, e.g., by signal-to-noise ratio of highest peak, has to be applied to invalidate false-positive matches resulting from pure arithmetic reasons. In addition, we believe that there is a potential of further improvement of the effectiveness of the proposed identification approach by application of advanced preprocessing techniques. However, a detailed discussion of such techniques lies beyond the scope of this study.

The best performance of isotopic distribution brute force is achieved when the allowed elements together have up to 8 or 9 stable isotopes. For search lists composed of two or three elements satisfying such a condition, ca. 15 000 compositions of compounds with masses ranging from 0 to 2000 Da can be tested in less than 1 s using an ordinary office laptop. Under the described conditions, a brute force performance of 10 to 40 formulas per 1 ms was observed experimentally. Such performance is presumably lower than the performance of a library search, which recently showed the capability of comparing $\frac{3,009,902}{108.5 \times 60 \times 1000} \approx 4.62$ pairs of entire mass spectra per millisecond.[46,47] However, automated evaluation of several thousand compositions per second may still greatly reduce the time spent on ion identification if the library mass spectra are not available for the studied substance. Taking the constraints listed above on search space into account, the application of isotopic distribution brute force in its presented implementation is not limited by performance in all realistic usage scenarios.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jasms.4c00153.

Graphical representations of used mass spectra, additional theoretical considerations, technical details, detailed results of the test identification, and detailed results of conducted performance measurements (PDF)

AUTHOR INFORMATION

Corresponding Author

Viacheslav V. Lebedev − *A. N. Frumkin Institute of Physical Chemistry and Electrochemistry, Russian Academy of Sciences, Moscow 119071, Russian Federation;* orcid.org/0009-0003-2678-558X; Phone: +74959520462; Email: glory.leb@gmail.com

Authors

Daniil I. Yarykin − *A. N. Frumkin Institute of Physical Chemistry and Electrochemistry, Russian Academy of Sciences, Moscow 119071, Russian Federation*

Aleksey K. Buryak − *A. N. Frumkin Institute of Physical Chemistry and Electrochemistry, Russian Academy of Sciences, Moscow 119071, Russian Federation*

Complete contact information is available at: https://pubs.acs.org/10.1021/jasms.4c00153

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The work was supported by the Ministry of Science and Higher Education of the Russian Federation (grant agreement No. 075-15-2024-534).

REFERENCES

(1) Hilario, M.; Kalousis, A.; Pellegrini, C.; Müller, M. Processing and classification of protein mass spectra. *Mass Spectrom. Rev.* **2006**, *25* (3), 409−449.

(2) Milman, B. L.; Zhurkovich, I. K. New Trends in Chemical Identification Methodology. *J. Anal. Chem.* **2024**, *79* (2), 119−133.

(3) De Hoffmann, E.; Stroobant, V. *Mass Spectrometry: Principles and Applications*, 3rd ed.; John Wiley & Sons, 2007.

(4) Hansen, M. E.; Smedsgaard, J. A new matching algorithm for high resolution mass spectra. *J. Am. Soc. Mass Spectrom.* **2004**, *15* (8), 1173−1180.

(5) Pytskii, I. S.; Minenkova, I. V.; Kuznetsova, E. S.; Zalavutdinov, R. Kh.; Uleanov, A. V.; Buryak, A. K. Surface chemistry of structural materials subjected to corrosion. *Pure Appl. Chem.* **2020**, *92* (8), 1227−1237.

(6) Smith, C. B. Mass spectrometry in geochemistry: An established tool in the earth sciences. *Nucl. Instrum. Methods Phys. Res., Sect. B* **1988**, *35* (3−4), 364−369.

(7) Senko, M. W.; Beu, S. C.; McLafferty, F. W. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J. Am. Soc. Mass Spectrom.* **1995**, *6* (4), 229−233.

(8) Goldfarb, D.; Lafferty, M. J.; Herring, L. E.; Wang, W.; Major, M. B. Approximating Isotope Distributions of Biomolecule Fragments. *ACS Omega.* **2018**, *3* (9), 11383−11391.

(9) Valkenborg, D.; Jansen, I.; Burzykowski, T. A model-based method for the prediction of the isotopic distribution of peptides. *J. Am. Soc. Mass Spectrom.* **2008**, *19* (5), 703−712.
(10) Yao, X.; Diego, P.; Ramos, A. A.; Shi, Y. Averagine-Scaling Analysis and Fragment Ion Mass Defect Labeling in Peptide Mass Spectrometry. *Anal. Chem.* **2008**, *80* (19), 7383−7391.
(11) Claesen, J.; Dittwald, P.; Burzykowski, T.; Valkenborg, D. An Efficient Method to Calculate the Aggregated Isotopic Distribution and Exact Center-Masses. *J. Am. Soc. Mass Spectrom.* **2012**, *23* (4), 753−763.
(12) Holden, N. E.; Coplen, T. B.; Böhlke, J. K.; Tarbox, L. V.; Benefield, J.; de Laeter, J. R.; Mahaffy, P. G.; O'Connor, G.; Roth, E.; Tepper, D. H.; Walczyk, T.; Wieser, M. E.; Yoneda, S. *IUPAC Periodic Table of the Elements and Isotopes (IPTEI) for the Education Community —Update 2019*; IUPAC, 2019. https://iupac.org/wp-content/uploads/2015/02/IPTEI_postprint_20190301.pdf (accessed 2024-03-29).
(13) Claesen, J.; Rockwood, A.; Gorshkov, M.; Valkenborg, D. The isotope distribution: A rose with thorns. *Mass Spectrom. Rev.* **2023**, DOI: 10.1002/mas.21820.
(14) Cody, R. B.; Fouquet, T. Elemental Composition Determinations Using the Abundant Isotope. *J. Am. Soc. Mass Spectrom.* **2019**, *30*, 1321−1324.
(15) Boiko, D. A.; Kozlov, K. S.; Burykina, J. V.; Ilyushenkova, V. V.; Ananikov, V. P. Fully Automated Unconstrained Analysis of High-Resolution Mass Spectrometry Data with Machine Learning. *J. Am. Chem. Soc.* **2022**, *144* (32), 14590−14606.
(16) Long, D.; Eade, L.; Sullivan, M. P.; Dost, K.; Meier-Menches, S. M.; Goldstone, D. C.; Hartinger, C. G.; Wicker, J. S.; Taškova, K. AdductHunter: identifying protein-metal complex adducts in mass spectra. *J. Cheminf.* **2024**, *16*, 15.
(17) Goncharova, I. S.; Pytskii, I. S.; Buryak, A. K. Application of mass spectrometry with laser desorption/ionization for studies of lead clusters on surfaces of different types. *Prot. Met. Phys. Chem. Surf.* **2014**, *50*, 723−732.
(18) *MS: Elemental Composition Calculations and their Interpretation*. JEOL USA, 2006. https://www.jeolusa.com/LinkClick.aspx?fileticket=GaapMLKGCVk%3d&tabid=337&portalid=2&mid=5080 (accessed 2024-04-01).
(19) Knuth, D. E. *Fascicle 3: Generating All Combinations and Partitions*, 1st ed.; The Art of Computer Programming, Vol. *4*; Addison-Wesley Professional, 2005.
(20) Smith, R. M. Elemental Composition from Peak Intensities. In *Understanding Mass Spectra: A Basic Approach*, 2nd ed.; John Wiley & Sons, 2004; pp 56−98. DOI: 10.1002/0471479357.ch2.
(21) Valkenborg, D.; Mertens, I.; Lemière, F.; Witters, E.; Burzykowski, T. The isotopic distribution conundrum. *Mass Spectrom. Rev.* **2012**, *31* (1), 96−109.
(22) Li, L.; Kresh, J. A.; Karabacak, N. M.; Cobb, J. S.; Agar, J. N.; Hong, P. A Hierarchical Algorithm for Calculating the Isotopic Fine Structures of Molecules. *J. Am. Soc. Mass Spectrom.* **2008**, *19* (12), 1867−1874.
(23) Łącki, M. K.; Startek, M.; Valkenborg, D.; Gambin, A. IsoSpec: Hyperfast Fine Structure Calculator. *Anal. Chem.* **2017**, *89* (6), 3272−3277.
(24) Łącki, M. K.; Valkenborg, D.; Startek, M. IsoSpec2: Ultrafast Fine Structure Calculator. *Anal. Chem.* **2020**, *92* (14), 9472−9475.
(25) Senko, M. W. *IsoPro*, ver. 3.0; Cornell University: Ithaca, NY, 1998.
(26) Yergey, J. A. A general approach to calculating isotopic distributions for mass spectrometry. *Int. J. Mass Spectrom Ion Phys.* **1983**, *52* (2−3), 337−349.
(27) Feng, X.; Zhang, W.; Kuipers, F.; Kema, I.; Barcaru, A.; Horvatovich, P. Dynamic binning peak detection and assessment of various lipidomics liquid chromatography-mass spectrometry pre-processing platforms. *Anal. Chim. Acta* **2021**, *1173*, 338674.
(28) Kou, Q.; Wu, S.; Liu, X. A new scoring function for top-down spectral deconvolution. *BMC Genomics.* **2014**, *15*, 1140.
(29) Morris, J. S.; Coombes, K. R.; Koomen, J.; Baggerly, K. A.; Kobayashi, R. Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics.* **2005**, *21* (9), 1764−1775.
(30) Basharat, A. R.; Ning, X.; Liu, X. EnvCNN: A Convolutional Neural Network Model for Evaluating Isotopic Envelopes in Top-Down Mass-spectral Deconvolution. *Anal. Chem.* **2020**, *92* (11), 7778−7785.
(31) Yang, C.; He, Z.; Yu, W. Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis. *BMC Bioinf.* **2009**, *10*, 4.
(32) Zhu, P.; Bowden, P.; Tucholska, M.; Marshall, J. G. Chi-square comparison of tryptic peptide-to-protein distributions of tandem mass spectrometry from blood with those of random expectation. *Anal. Biochem.* **2011**, *409* (2), 189−194.
(33) Pearson, K. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *London, Edinburgh Dublin Philos. Mag. J. Sci.* **1900**, *50* (302), 157.
(34) Kolmogorov, A. N. Sulla Determinazione Empirica di Una Legge di Distribuzione. *Giornale dell'Istituto Italiano degli Attuari.* **1933**, *4*, 83−91.
(35) Smirnov, N. V. On the Estimation of Discrepancy between Empirical Curves of Distribution for Two Independent Samples. *Moscow Univ. Math. Bull.* **1939**, *2* (2), 3−11.
(36) Anderson, T. W.; Darling, D. A. A Test of Goodness of Fit. *J. Am. Stat. Assoc.* **1954**, *49* (268), 765−769.
(37) Gleser, L. J. Exact Power of Goodness-of-Fit Tests of Kolmogorov Type for Discontinuous Distributions. *J. Am. Stat. Assoc.* **1985**, *80* (392), 954−958.
(38) Guenther, W. C. Power and Sample Size for Approximate Chi-Square Tests. *American Statistician.* **1977**, *31* (2), 83−85.
(39) Dimitrova, D. S.; Kaishev, V. K.; Tan, S. Computing the Kolmogorov-Smirnov Distribution When the Underlying CDF is Purely Discrete, Mixed, or Continuous. *Journal of Statistical Software.* **2020**, *95* (10), 1−42.
(40) R Core Team. *R: A Language and Environment for Statistical Computing*, ver. 4.3.1; R Foundation for Statistical Computing: Vienna, Austria, 2023. https://www.r-project.org/. (accessed 2024-03-29).
(41) Kim, S.; Zhang, X. Comparative Analysis of Mass Spectral Similarity Measures on Peak Alignment for Comprehensive Two-Dimensional Gas Chromatography Mass Spectrometry. *Comput. Math. Methods Med.* **2013**, *2013*, 509761.
(42) Moorthy, A. S.; Kearsley, A. J. Pattern Similarity Measures Applied to Mass Spectra. In *Progress in Industrial Mathematics: Success Stories*; Cruz, M.; Parés, C.; Quintela, P., Eds; SEMA SIMAI Springer Series (ICIAM2019SSSS), Vol. *5*; Springer, 2020; pp 43−53. DOI: 10.1007/978-3-030-61844-5_4.
(43) Huber, F.; van der Burg, S.; van der Hooft, J. J. J.; Ridder, L. MS2DeepScore: a novel deep learning similarity measure to compare tandem mass spectra. *J. Cheminf.* **2021**, *13*, 84.
(44) Wells, G.; Prest, H.; Russ, C. W., IV *Signal, Noise and Detection Limits in Mass Spectrometry*; 5990−7651EN; Agilent Technologies, Inc., 2023. https://www.agilent.com/cs/library/technicaloverviews/public/5990-7651EN.pdf (accessed 2024-06-20).
(45) Titulaer, M. K.; Siccama, I.; Dekker, L. J.; van Rijswijk, A. L. C. T; Heeren, R. M. A.; Sillevis Smitt, P. A.; Luider, T. M. A database application for pre-processing, storage and comparison of mass spectra derived from patients and controls. *BMC Bioinf.* **2006**, *7*, 403.
(46) Gibb, S.; Strimmer, K. MALDIquant: a versatile R package for the analysis of mass spectrometry data. *Bioinformatics.* **2012**, *28* (17), 2270−2271.
(47) Bittremieux, W.; Meysman, P.; Noble, W. S.; Laukens, K. Fast Open Modification Spectral Library Searching through Approximate Nearest Neighbor Indexing. *J. Proteome Res.* **2018**, *17* (10), 3463−3474.