Predicting the presence of uncommon elements in unknown biomolecules from isotope patterns

Marvin Meusel,[†] Franziska Hufsky,[†,¶] Fabian Panter,[‡] Daniel Krug,[‡] Rolf Müller,[‡] and Sebastian Böcker*[,†]

[†]*Chair for Bioinformatics, Friedrich Schiller University Jena, 07743 Jena, Germany*

[‡]*Department of Microbial Natural Products, Helmholtz-Institute for Pharmaceutical Research Saarland, Helmholtz Centre for Infection Research and Pharmaceutical Biotechnology, Saarland University, 66123 Saarbrücken, Germany*

[¶]*RNA Bioinformatics and High Throughput Analysis, Friedrich Schiller University Jena, 07743 Jena, Germany*

E-mail: sebastian.boecker@uni-jena.de

This is a preprint of: Marvin Meusel, Franziska Hufsky, Fabian Panter, Daniel Krug, Rolf Müller, and Sebastian Böcker. Predicting the Presence of Uncommon Elements in Unknown Biomolecules from Isotope Patterns. Analytical Chemistry 88(15), 7556–7566, 2016. DOI https://doi.org/10.1021/acs.analchem.6b01015. The final publication is available from ACS Publications.

Abstract

Motivation: The determination of the molecular formula is one of the earliest and most important steps when investigating the chemical nature of an unknown compound. Common approaches use the isotopic pattern of a compound measured using mass spectrometry. Computational methods to determine the molecular formula from this isotopic pattern require a fixed set of elements. Considering all possible elements

severely increases running times and more importantly the chance for false positive identifications as the number of candidate formulas for a given target mass rises significantly if the constituting elements are not pre-filtered. This negative effect grows stronger for compounds of higher molecular mass as the effect of a single atom on the overall isotopic pattern grows smaller. On the other hand, hand-selected restrictions on this set of elements may prevent the identification of the correct molecular formula. Thus, it is a crucial step to determine the set of elements most likely comprising the compound prior to the assignment of an elemental formula to an exact mass.

Results: In this paper, we present a method to determine the presence of certain elements (sulfur, chlorine, bromine, boron and selenium) in the compound from its (high mass accuracy) isotopic pattern. We limit ourselves to biomolecules, in the sense of products from nature or synthetic products with potential bioactivity. The classifiers developed here predict the presence of an element with a very high sensitivity and high specificity. We evaluate classifiers on three real-world datasets with 663 isotope patterns in total: 184 isotope patterns containing sulfur, 187 containing chlorine, 14 containing bromine, one containing boron, one containing selenium. In no case do we make a false negative prediction; for chlorine, bromine, boron, and selenium, we make ten false positive predictions in total. We also demonstrate the impact of our method on the identification of molecular formulas, in particular the number of considered candidates and running time.

Availability: The element prediction will be part of the next SIRIUS release, available from `https://bio.informatik.uni-jena.de/software/sirius/`. The 86 mass spectra from the *myxo* dataset will be made available upon publication.

Introduction

Hyphenated high-resolution mass spectrometry, mostly coupled to liquid chromatography (LC-MS) or gas chromatography (GC-MS), is the predominant experimental platform for untargeted metabolomics and also plays an important role in other analytical fields requiring high information content and increased sample throughput, such as natural products research. Due to the underlying study designs these applications regularly bring about high numbers of unidentified mass spectral features, leading to the analytical challenge to identify as many as possible of the corresponding unknown compounds. One of the decisive steps investigating unknown compounds is to determine its molecular formula, which can serve as a starting point for the structural elucidation. High-throughput molecular formula annotation workflows are required for the analysis of complex biological samples. The vast numbers of unknowns detected in mass spectral datasets acquired from these samples necessitate efficient methods for formula generation, since the process is computationally expensive and error-prone. As an example, a bacterial extract could contain 600-700 unknown substances[1], and studies of the metabolomes of higher organisms exceed these numbers: the Human serum metabolome contains more than 4000 metabolites visible by LC-MS analysis[2]. Especially for non-model organisms, an astounding number of metabolites to date remain uncharacterized with respect to their structure and function. Contrary to proteins and other bio-polymers, which are constructed from a well-defined set of building blocks, the structure of metabolites is much less defined and thus the structure elucidation process is labor-intensive and usually

requires additional techniques like NMR spectroscopy. The capability to efficiently compute high-confidence molecular formulas facilitates this objective.

In the following, we will limit ourselves to *biomolecules*, that is, molecules that are products of nature, or synthetic products with potential bioactivity. Many biomolecules, among them an overwhelming number of substances found ubiquitously across the domains of life, are composed of six elements, i.e. carbon, hydrogen, nitrogen, oxygen, phosphorus, and sulfur[3]. In contrast, secondary metabolites and other biomolecules, such as drugs or pesticides, occasionally contain less frequently occurring elements, which we refer to as *uncommon elements* throughout this paper. To name a few examples, marine organisms and some terrestrial bacteria produce halogenated compounds incorporating bromine or chlorine[4,5]; antibiotics containing boron have been reported[6]; metabolites of higher plants can contain selenium[7,8]; finally, fluorine and iodine have been detected in the metabolism of certain organisms[9,10]. Microbial secondary metabolites are the subject of natural product screening workflows forming the basis for the discovery of novel drug candidates, thus strengthening our motivation to improve computational tools for their identification[11–13]. However, it is understood that methods developed for biomolecule identification are likely to find application in other analytical areas too, since the underlying fundamental challenge of unknown characterization is common to many varieties of small-molecule investigations using mass spectrometry, ranging from in-vitro drug metabolism studies to pesticide screening.

Mass spectrometric instrumentation has seen significant improvement in terms of resolution and mass accuracy over the last two decades, however exact (monoisotopic) mass alone is insufficient to determine the molecular formula of a compound even for sub-ppm mass accuracy[14]. To this end, several approaches use the natural isotopic distributions of elements to improve molecular formula determination[15], assuming a mass accuracy of about 10 ppm or better. Certain approaches limit computation to molecular formulas present in a database[16]; but since many natural compounds are absent from any database, this restriction is unacceptable. Other approaches compute all candidate molecular formulas (considering a fixed set of elements) that are sufficiently close to the measured peak mass[17–19], simulate an isotope pattern for each candidate molecular formula[20–23], and compare it to the measured one[24–27]. Here, the isotope pattern may also contain accurate masses of isotope peaks. Some approaches additionally take the fragmentation pattern of compounds into account[28–31]. All of these approaches require the researcher to specify the set of elements to be considered.

For compounds above 400 Da, the number of molecular formulas increases rapidly[14,24,31], in particular when considering a set of elements beyond CHNOPS. This severely increases not only running time, but also chances for false identifications. On the other hand, manually adjusted restrictions on elements, or the maximum allowed number of atoms of a particular element, could exclude the correct molecular formula from the candidate set[32] and hence prevent the discovery of an interesting compound featuring an uncommon element. The presence of certain uncommon elements can be inferred by manual inspection of the isotope pattern in mass spectra, see Fig. 1 in ref.[23] for an instructive example. Transforming such observations into robust classifiers is nevertheless a challenging problem, as both masses and intensities are perturbed in real-world measurements. Previous efforts along these lines have, for example, been related to estimating the number of chlorine and bromine atoms from Electron Ionization mass spectra as part of the NIST Mass Spectral Search Program[33]. Here, we present a *fully automated method* to robustly determine the elements for the com-

pound under study from the isotope pattern. The set of elements determined by our method then serves as input for the next step of the analysis, where accurate masses, isotope pattern (and potentially other information) are used to determine the molecular formula of a compound under investigation. The four most abundant elements in living organisms — carbon, hydrogen, nitrogen, and oxygen — form our basic set of elements. We add phosphorus to this set, as it is relatively common in bio-compounds (e.g. nucleotides, ATP) and, in addition, has only a single stable isotope, thus cannot be predicted on the basis of isotope pattern (the latter applies also to iodine and fluorine). We present a method to predict the presence of the elements sulfur (S), chlorine (Cl), bromine (Br), boron (B) and selenium (Se) from isotope patterns.

Our method uses Machine Learning, that is, algorithms that can learn from data, and make predictions based on data. We employ Supervised Machine Learning where the computer is presented with example inputs and desired outputs (the training data), and the goal is to learn rules that map inputs to outputs. For each element, we use a binary classifier that classifies the data into two groups (contains the element vs. does not contain the element). For classification, we use random forests[34]. For each element, we present three classifiers based on the number of observed isotope peaks (three, four, and five or more peaks). We find that the more unique the isotopic distribution of an element is, the more precise our predictions are. Furthermore, availability of more isotopic peaks from the mass spectrum also improves prediction quality. Evaluating the classifiers on three real-world datasets of a total of 663 isotope patterns measured on three different instruments results in no false negative predictions, and few false positive predictions for all elements but sulfur. We show that our method significantly decreases the number of molecular formulas that have to be considered by subsequent analysis steps, resulting in a massive decrease of running time, while at the same time also slightly improving identification rates. The method presented here will be integrated in an upcoming release of the SIRIUS 3 software package[24,31].

Background

The most common method to determine a molecular formula is to simulate an isotope pattern for each candidate molecular formula, and compare it to the measured one[24–27]. This requires a fixed set of elements to be considered for the generation of candidate molecular formulas. Obviously, the number of candidates increases considerably with the size of the set of elements. For example, consider *rifampicin* with monoisotopic mass 822.405 Da and mass accuracy 6 ppm: we reach 2,358 candidates for the set of elements CHNOP, and 117,029 candidates for CHNOPSClBr, an almost 50-fold increase caused by only three uncommon elements. (The above numbers ignore molecular formula filters such as the "Senior rules"[35], but the effect is comparable.) On the theoretical side, the number of decompositions of a certain mass can be approximated with high accuracy using a polynomial[24]: Assuming a *relative error*, this polynomial has degree k for k elements.

Isotopes are variants of an element with different numbers of neutrons. Different isotopes of the same element have nearly identical chemical properties but different mass. Most elements in nature have more than one stable isotope, and each of these isotopes occurs in nature with a certain abundance[36–38]. The totality of isotopes is the *isotopic distribution* of an element.

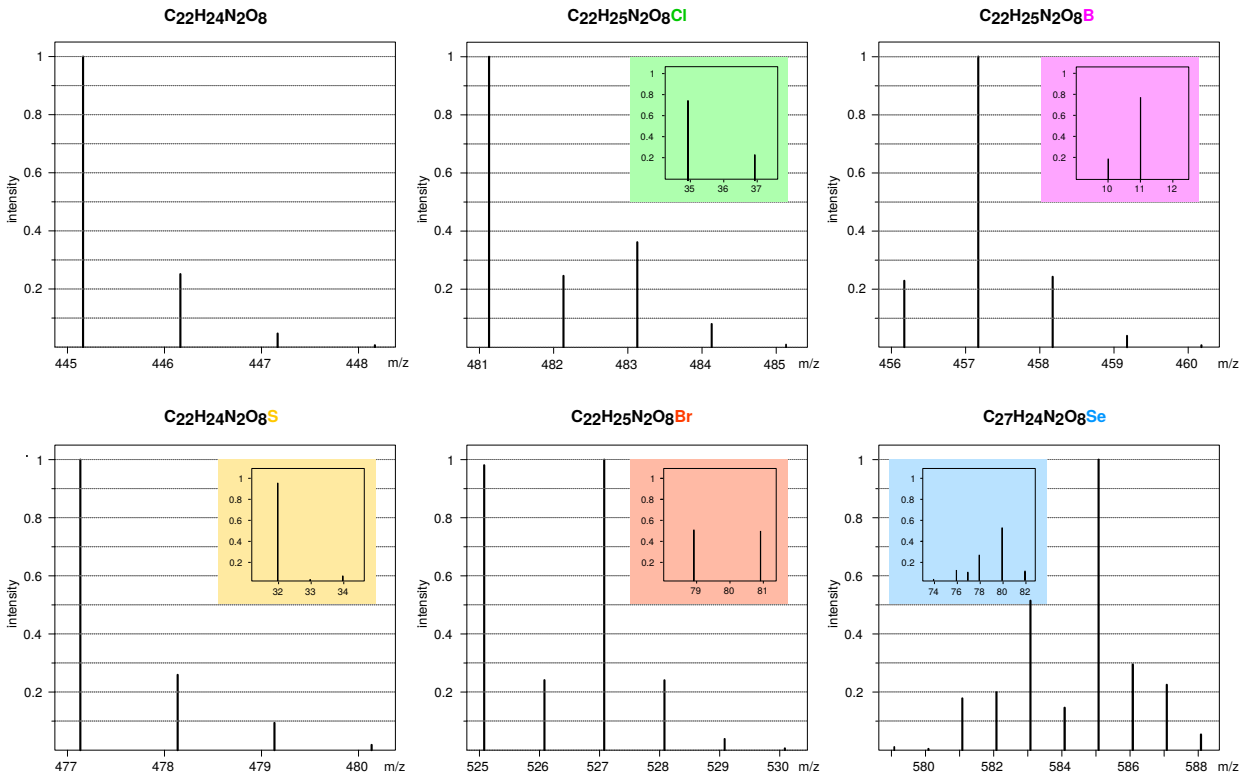

Figure 1: Effect of the uncommon elements S (yellow), Cl (green), Br (red), B (magenta), and Se (blue) on the isotope pattern. $[M+H]^+$ ionization is assumed. Individual isotopic distributions of the corresponding uncommon elements are shown in the colored boxes. All molecular formulas come from the molecular structure databases.

The mass difference of successive isotopes differs from element to element, despite the fact that in all cases, one or more neutrons are added: For example, the mass difference of ^{12}C and ^{13}C is 1.00335 Da (Dalton), whereas the mass difference of ^{10}B and ^{11}B is 0.99637 Da. The different isotope abundances and masses of the isotopes that are contained in a compound generate a characteristic set of peaks, called an isotope pattern. In comparison to C, H, N, O, and P (referred to as "CHNOP" in the following), the elements S, Cl, Br, B, and Se lead to more "distinctive" isotope patterns, as the most abundant isotope of these elements has relative abundance less than 95 %.

In this paper, we will not consider the isotopic fine structure (isotopologues) of a compound; instead, we limit ourselves to the *mean peak masses*[22] (or *accurate masses*[21]) of isotopic peaks, combining all isotopologues with identical nominal mass. Throughout this paper, the isotopologue where each atom is the isotope with the lowest nominal mass is referred to as *monoisotopic*, see for example ref.[39]. For certain elements such as boron or selenium, this is *not* the most abundant isotope. We refer to the peak at the monoisotopic mass as the *monoisotopic peak* or +0 peak, which is followed by the +1, +2, ... peaks. Referring to a peak "before" ("after") another peak, means it has a smaller (higher, respectively) mass.

The isotopic distribution of elements influences the isotope pattern of the compound in both mass differences and intensities of peaks. Small compounds containing only CHNOP

have an intense monoisotopic peak; the +1 peak has much lower intensity, and intensity of subsequent peaks decreases further. The more "distinctive" isotopic distributions of S, Cl, Br, B, and Se are reflected in the shape of the isotope pattern of a compound, see Figure 1 and Supplementary Section S.1.

Methods

We use a set of binary classifiers to predict the presence of uncommon elements from isotope patterns of compounds using supervised machine learning. We create one classifier for the presence of each of the uncommon elements Cl, Br, B, S and Se. Further, we create a classifier "CHNOPS" for compounds that contain only the elements C, H, N, O, P, and S; we use this classifier to demonstrate the discriminative power of the isotope pattern for identifying uncommon elements. The classifier cannot rule out the presence of fluorine or iodine in a compound, or any other element that has only a single stable isotope.
To learn and predict the presence of uncommon elements from isotope patterns, the patterns need to be transformed to a set of numerical features characterizing the data. Those features are based on the masses and intensities of the peaks. Measured isotope patterns vary in the number of detected peaks. In particular for small compounds, we get very short isotope patterns, as only a few peaks are intense enough to be detected as signal. Using classifiers for arbitrary peak number would result in missing (peak) features for smaller isotope patterns, making the training more complicated. Hence, we construct different classifiers for isotope patterns of size three, four and five. In total, we train 18 classifiers (five uncommon elements plus CHNOPS times three isotope pattern sizes).
Isotope patterns with less than three peaks do not contain enough information for uncommon element prediction. However, for isotope patterns with less than three peaks, the presence of an uncommon element is very unlikely since the +2 should be intense enough for detection if an uncommon element is present (see Figure 1). For isotope patterns with more than five peaks, we consider only the first five peaks. Classifiers for isotope patterns with more than five peaks did not improve the results significantly (data not shown).

Features

We use features based on the intensity and mass of the isotope peaks. The full list of features can be found in Supplementary Table S.1. The number of features varies according to the number of peaks in the isotope pattern: There are 21 (38, 39) intensity features and 4 (7, 11) mass features for three (four, five, respectively) isotope peaks.

Intensity features. We use the intensity of each peak, as well as the minimal, maximal and median peak intensity. Since the presence of some elements results in zigzag shaped isotope patterns, we further use the sums, minima and maxima of intensities for even and odd peaks, respectively. Furthermore, we use the index of the most intense, second most intense and third most intense peak. We use quotients and differences of all pairs of peak intensities for the first four peaks. Finally, we use all combinations of sums of peak intensities.

Mass features. We use monoisotopic mass, and all pairwise mass differences between isotope peaks as features. Mass is included, since the influence of uncommon elements on the shape of the isotope pattern is stronger for small compounds than for large ones. The mass difference between consecutive peaks of the isotope pattern is influenced by elements with more than one stable isotope: For example, if Cl is present, the mass difference between the monoisotopic peak and +2 peak decreases due to the mass difference of 1.99705 Da between ^{35}Cl and ^{37}Cl.

Random forests

For classification we use random forests, a supervised machine learning method[34]. A random forest consists of a set of unpruned decision trees. Each tree is trained on a random sample (with replacement) of the training data. This technique is called bagging[40]. For the determination of a split (a node in a decision tree), we use a subset of $m \approx \sqrt{M}$ features, where M is the total number of features[41]. We use the random forest implementation of the Mahout library* (version 0.9) for training.
We set $m = 5$ for isotope patterns of size three, and $m = 7$ for isotope patterns of size four and five. For each classifier, we generate a random forest consisting of 100 trees. For all other parameters we use the default values of the Mahout library. The classes `DecisionTreeBuilder` and `Bagging` are used to construct the trees. By default, the minimum set size for a split is 2 and the minimum variance proportion is 0.001. We use the `OptIgSplit` class to compute the best split.
For classification, the input is run through all of the decision trees in the random forest, and the final classification is done by voting: every decision tree votes for true or false. A threshold for the ratio of positive votes is used to determine the classification of the forest.

Datasets

For the training and evaluation of the classifiers, we want to use a very large set of isotope patterns. Unfortunately, there is only limited experimental data available for isotope patterns of biomolecules. In this paper, we use three measured datasets from different instruments for evaluation; in addition, we use simulated isotope patterns of molecular formulas from several compound databases (see Table 1).
The *myxo* dataset consists of 88 isotope patterns measured on a Bruker MaXis 2G qTOF spectrometer (Bremen, Germany). The corresponding compounds are secondary metabolites from *Myxobacteria*, with several exceptions mentioned below. Compounds range in mass from 192.009 to 2213.962 Da, with average mass 623.277 Da and median mass 591.307 Da. From these compounds, 24 contain sulfur (20 contain a single sulfur atom, 2 contain two, and 2 contain three sulfur atoms), 8 contain chlorine (6 with one chlorine atom, 2 with two chlorine atoms), and one compound contains a single boron atom. We added selenomethionine to this dataset as a positive example for a selenium-containing compound. 9 compounds have a measured isotope pattern of length three, 47 of length four, and 32 of length five or more (3 of length six, 5 of length seven, 1 of length nine).

*`https://mahout.apache.org/`

Table 1: Overview of the datasets used for training and evaluation. Total number of molecular formulas and numbers of molecular formulas that are positive examples for the different classifiers are given. In addition to the simulated *evaluation set*, the measured datasets are used exclusively for evaluation.

	simulated		measured		
	training	*evaluation*	*myxo*	*pesticide*	*CASMI*
all	1 128 059	51 097	88	43	532
CHNOPS	604 506	42 713	78	15	370
sulfur S	502 529	9 892	24	14	146
chlorine Cl	345 799	5 502	8	27	152
bromine Br	171 596	3 015	0	1	13
boron B	56 808	141	1	0	0
selenium Se	16 680	98	1	0	0

The *pesticide* dataset is taken from Stravs et al.[30]. The dataset consists of isotope patterns from different pesticides measured on a LTQ Orbitrap XL from Thermo Fisher Scientific (San José, USA) with electrospray ionization in positive and negative mode. Compounds range in mass from 198.056 to 443.125 Da, with average mass 275.353 Da and median mass 260.016 Da. From the 60 mass spectra, 43 show an isotope pattern with at least three peaks. For the 18 shorter isotope patterns, none of the compounds contains any of the elements SClBrBSe. For the remaining 43 isotope patterns, 14 isotope patterns contain a single sulfur atom, 27 chlorine (22 with a single chlorine atom, 4 with two, and 1 with three chlorine atoms), and one isotope pattern contains a single bromine atom. The dataset contains 18 isotope patterns with a measured isotope pattern of length three, 18 of length four, and 7 of length five or more (1 of length six).
The *CASMI* dataset was measured by Martin Krauss (Helmholtz Centre for Environmental Research, Leipzig, Germany), and processed by Emma Schymanski (Eawag, Dübendorf, Switzerland) as part of the Critical Assessment of Small Molecule Identification challenge 2016[†] using RMassBank[30]. Measurements were performed on a Q Exactive Plus Orbitrap (Thermo Scientific) with electrospray ionization in positive and negative mode. MS1 spectra were extracted for substances with [M+H]+ (positive) and [M-H]- (negative mode) ions. We removed one compound containing silicon. The remaining dataset contains 628 independent mass spectra from 512 compounds. Compounds range in mass from 67.042 to 776.687 Da, with average mass 259.500 Da and median mass 246.071 Da. From the 628 mass spectra, 532 show an isotope pattern with at least three peaks. For the 96 shorter isotope patterns, none of the compounds contain any of the elements SClBrBSe. For the remaining 532 isotope patterns, 146 patterns contain sulfur (128 with a single sulfur atom, 17 with two, and 1 with four sulfur atoms), 152 chlorine (84 with a single chlorine atom, 42 with two, 19 with three, 3 with four and 3 with six chlorine atoms) and 13 bromine (11 with a single bromine atom and 2 with two bromine atoms). The dataset contains 140 isotope patterns with a measured isotope pattern of length three, 220 of length four, and 172 of length five

[†]`http://www.casmi-contest.org/2016/`

or more (49 of length six, 2 of length seven and 1 of length 8).

For training and evaluation, we use simulated isotope patterns of molecular formulas of compounds from eleven compound databases: ChEBI (Chemical Entities of Biological Interest)[42], ChEMBL version 19[43], DrugBank version 4.2[44], HMDB (Human Metabolome Database) version 3.6[45], Indofine[‡], KEGG (Kyoto Encyclopedia of Genes and Genomes)[46], KNApSAcK[47], MolMall[§], PubChem[48], UNPD (Universal Natural Product Database)[49], and ZINC (ZINC Is Not Commercial) version 12[50], see Supplementary Section S.3 and Supplementary Table S.2 for details.
Some of the databases, in particular PubChem, contain many records which are uncommon for biomolecules. To remove such molecular formulas, we filtered all datasets using the following *ad hoc* rules: (1) the compound has a monoisotopic mass between 100 and 1500 Da; (2) the compound contains only elements from CHNOPSClBBrSe and contains at least one carbon and one hydrogen atom; (3) the compound contains at most five atoms of sulfur, chlorine, boron or bromine. For the training data only, we also discarded (4) charged or generic compounds or complexes. We stress that molecular formulas passing these restrictions do not necessarily correspond to biomolecules, and that biomolecule molecular formulas may fail one or more of these rules; these data are used to train and evaluate our method.
We split the databases into a large *training set* that is used to construct the classifiers, and a smaller *evaluation set*. As the *training set*, we use PubChem, ChEMBL and ZINC. PubChem is the largest dataset and contains reasonably many compounds even for Se and B. It is particularly useful to train cases such as compounds incorporating a high number of uncommon elements, or an unlikely combinations of uncommon elements. We observe that molecular formulas from ChEMBL and ZINC are almost completely covered by PubChem. As the *evaluation set*, we use HMDB, ChEBI, DrugBank, Indofine, KEGG, KNApSAcK, MolMall, and UNPD.
Several molecular formulas are contained in more than one database; clearly, we consider only a single occurrence of each. Further, the *training set* should be independent from the *evaluation set*. To this end, all molecular formulas from the *evaluation set* are removed from the *training set*. In total, the *training set* contains 1,128,059 molecular formulas, and the *evaluation set* contains 51,097 molecular formulas (see Table 1).

Isotope pattern simulation

For the simulation of isotope patterns, we use SIRIUS[24] to compute the exact isotope pattern. For the *training* and *evaluation set*, we add noise to the peak intensities to simulate measured isotope patterns. In Machine Learning, adding noise often improves classification results, as it forces the method to search for complex patterns and improves generalizability. It is not necessary that the added noise has the same characteristics as noise observed in the real-world data. For each peak, we calculate the simulated intensity $I = I_{\text{exact}} \cdot \mathcal{N}(1, \sigma_{\text{IR}}) + \mathcal{N}(0, \sigma_{\text{IA}})$, where $\mathcal{N}(1, \sigma_{\text{IR}})$ is a normally distributed relative noise with mean one and $\mathcal{N}(0, \sigma_{\text{IA}})$ is a normally distributed absolute noise with mean zero. After adding noise to all peaks of an isotope pattern, the peak intensities are normalized to 100 %.

[‡] http://indofinechemical.com/
[§] http://www.molmall.net/

In addition, we add noise to the peak masses. For each peak, we calculate the simulated mass $m = m_{\text{exact}} + \mathcal{N}(0, \sigma_{\text{M}})$, where $\mathcal{N}(0, \sigma_{\text{M}})$ is an additive, normally distributed absolute noise with mean zero. We are aware that mass deviation is not constant throughout the mass range, but rather depends on the mass of the compound. Nevertheless, we refrain from using a multiplicative mass noise: The features for the classifiers are build on *mass differences* between the peaks instead of peak masses. To our knowledge, the distribution of deviation of dependent mass differences has not been investigated. To make our classifiers robust, we prefer to consider the worst case scenario, which is an additive noise with the highest mass deviation we have observed.

Noise parameters of the evaluation set. To find realistic noise parameters, we investigate measured isotope patterns from a Bruker Maxis 2G qTOF mass spectrometer, measured at different concentration levels. Based on the accuracy of these patterns, we define a *standard noise profile* with standard deviations $\sigma_{\text{IA}} = 0.0015$, $\sigma_{\text{IR}} = 0.04$, and $\sigma_{\text{M}} = 0.0013\,\text{Da}$. In addition, we define a *high noise profile* with standard deviations $\sigma_{\text{IA}} = 0.006$, $\sigma_{\text{IR}} = 0.07$, and $\sigma_{\text{M}} = 0.0018\,\text{Da}$ to generate particularly hard cases. The reader is reminded that we measure *mass differences* between isotope peaks. Both noise profiles are used for generating isotope patterns for the *evaluation set*.

Noise parameters of the training set. Instead of using exact simulated isotope patterns, we found that adding noise to the training data improves classification results. We found that using relatively high noise during training, improves classification results even for the *standard noise profile*. Thus, for simulating the *training set*, we choose noise parameters that are higher than the *standard noise profile* of the *evaluation set*: namely, $\sigma_{\text{IA}} = 0.005$, $\sigma_{\text{IR}} = 0.05$ and $\sigma_{\text{M}} = 0.0015\,\text{Da}$.

Learning phase

For most classifiers, we have many more negative examples than positive examples in the *training set*. Due to the optimization function of random forests, unbalanced data leads to suboptimal classification results for the smaller class[51]. For each classifier, we generate a training subset by randomly drawing without replacement to reach the same number of positive and negative examples. The CHNOPS, S, Cl, Br and B classifiers are trained using a subset of 100,000 molecular formulas (50,000 positive and 50,000 negative examples), and Se with 25,000 molecular formulas (12,500 positive and 12,500 negative examples).
There exist improved downsampling techniques to balance the data, such as distance based methods[52]. Due to the large number of compounds and high dimensionality of the training data, computation would be very time consuming. We found that random downsampling delivers very good results, and refrained from testing improved techniques.

Results

For each element, we present three classifiers based on the number of observed isotope peaks (3, 4, and 5 or more peaks). Classifiers are trained using simulated isotope patterns of more

than a million compounds. The resulting classifiers give excellent separation, with area under curve (AUC) above 0.99 for all elements but sulfur. We then concentrate on classifiers with very high sensitivity, so that we almost never miss an occurrence of an uncommon element. In contrast, we put less focus on the specificity of our classifiers, resulting in more false positive predictions for certain elements. The reasoning for doing so, is that our classifiers are meant as a first step of an automated pipeline for isotope pattern analysis: After generating and scoring candidates, the downstream method can sort out cases where our classifiers made a false positive prediction. In contrast, if our classifiers make a false negative prediction, wrongly stating that some uncommon element is absent from a compound, the downstream analysis has no means to correct this.
Training of all 18 classifiers required 3,5 h on two Intel Sandy Bridge processors at 2.3 GHz with six cores and 64 GB RAM. We find that training requires more time and the decision trees are larger, the less unique the isotopic distribution of the element is (see Supplementary Table S.3 and S.4). For example, in the random forest of the four peaks sulfur classifier, the resulting trees have a median of 9,450 nodes (median tree depth 47). Construction of this random forest required 23 min. In contrast, for the four peaks boron classifier, trees contain only a median of 13 nodes (depth four) and learning required only 7 min. For an example of a decision tree, see Figure 2.

Prediction quality on simulated data

There exist two types of wrong classifications: On the one hand, a classifier may return false ("element not present") although the element is part of the compound (*false negative*); on the other hand, a classifier may return true ("element present") when an element is not part of the compound (*false positive*).
In Machine Learning, one is generally interested in classifiers with high *sensitivity* (true positive rate) and high *specificity* (1−false positive rate), where:

$$
\begin{aligned}
\textit{sensitivity} &= \text{true positive rate} = 1 - \text{false negative rate} \\
&= \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \\
\textit{specificity} &= \text{true negative rate} = 1 - \text{false positive rate} \\
&= \frac{\text{true negatives}}{\text{true negatives} + \text{false positives}}
\end{aligned}
$$

It must be understood that reaching both high sensitivity and specificity simultaneously is usually not possible, and that we might have to trade in sensitivity for specificity, or vice versa. For the downstream analysis we have in mind (identification of the molecular formula), false negative predictions are much worse than false positive predictions. Once an element that is actually part of the compound is not considered for molecular formula identification, it is difficult or impossible to counteract this failure in the subsequent analysis steps. In contrast, a false positive prediction will "only" increase the running time and the probability of identifying a wrong molecular formula. Thus, sensitivity needs to be very high, even at the cost of specificity. For the CHNOPS classifier, which is included to evaluate the overall power of the approach, we treat sensitivity and specificity as equally important.

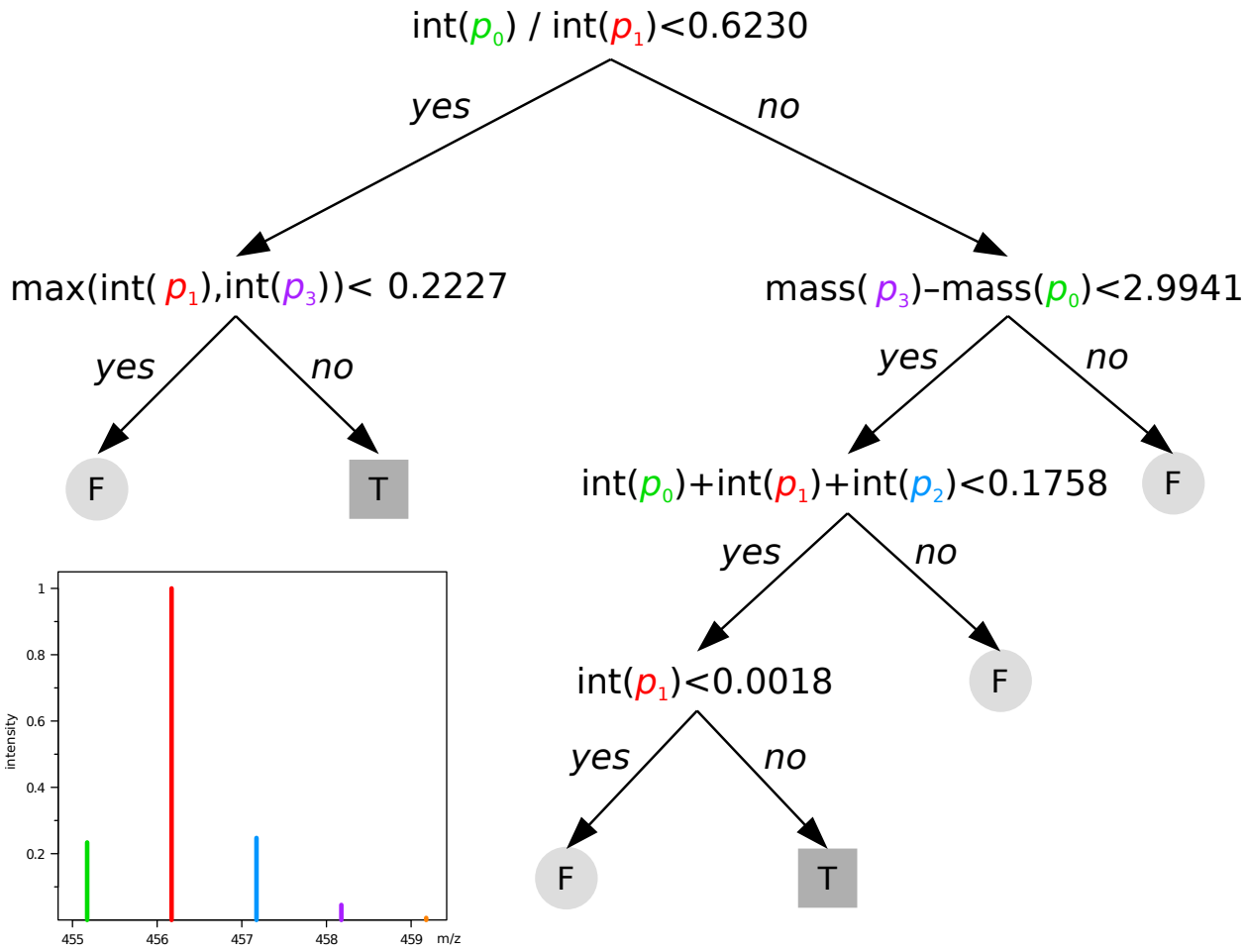

Figure 2: Example decision tree from the random forest of the boron classifier for isotope patterns with five peaks. This tree is combined with 99 other decision trees to reach the final decision whether or not boron is present in the unknown compound.

Receiver Operating Characteristic (ROC) and the area under the ROC curve (AUC) can be used to evaluate the ability of a classifier to distinguish between true and false examples. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at all possible threshold settings of the classifier. When decreasing the threshold, both the true positive rate and the false positive rate increase. A perfect classifier would result in an area under curve of 1, that is, there is a threshold where the classifier reaches TPR of 1 without predicting any false positives. Random classifiers have area under curve of 0.5.

ROC curves for the classifiers with three peaks are shown in Figure 3. Results for four and five peaks can be found in Supplementary Figure S.1 in the Appendix. For all classifiers, we achieve area under curve above 0.97 even for the *high noise profile*, see Table 2. The lowest area under curve was achieved for the S classifier on three peaks. Compared to the other elements to be detected, sulfur has the least distinctive isotopic distribution and is hard to distinguish from isotope patterns containing only CHNOP. The more unique the isotopic distribution of the elements get, the higher the area under curve of the corresponding

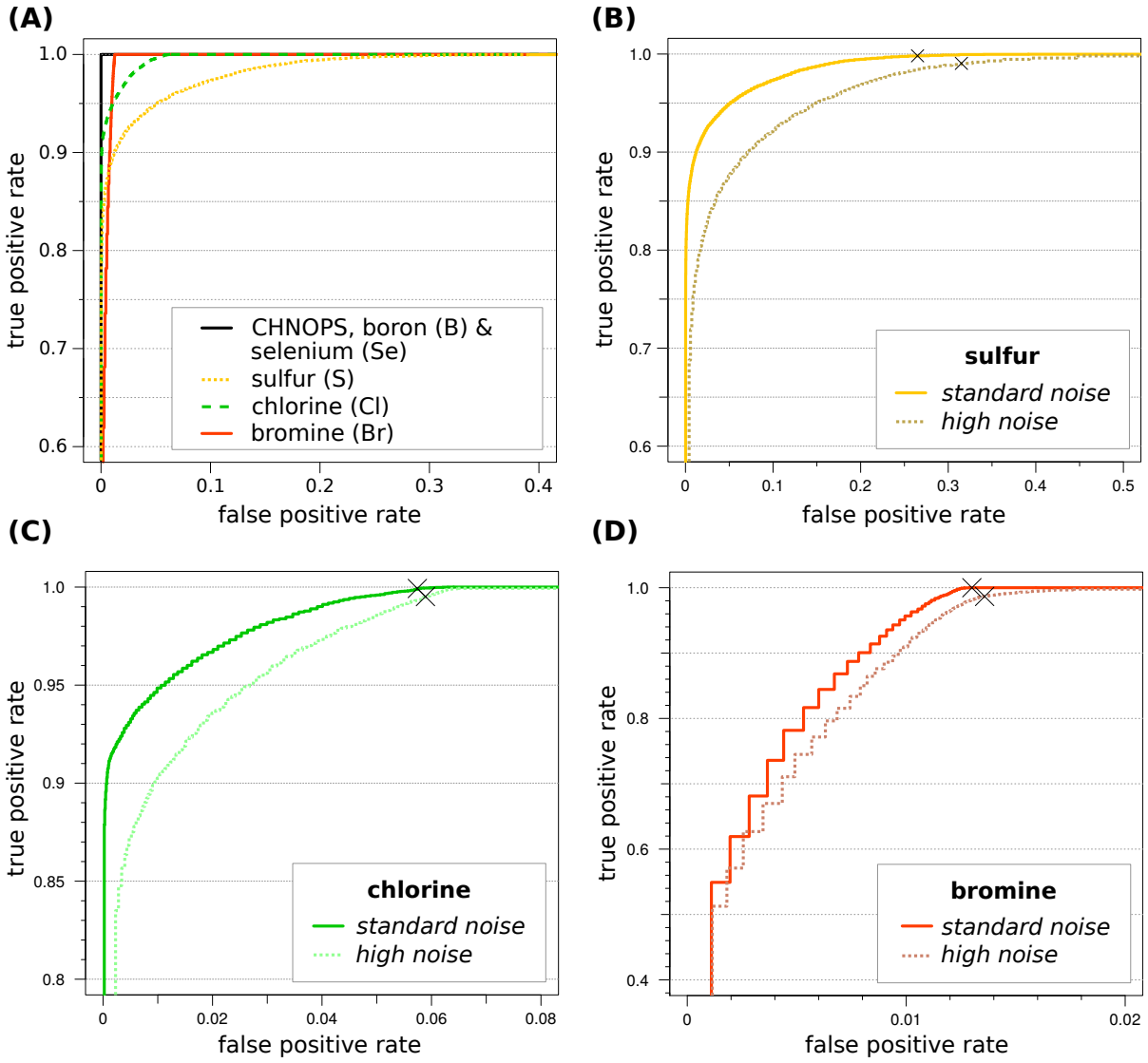

Figure 3: **(A)** ROC curves for all classifiers using three peaks at the *standard noise profile*. **(B)**–**(D)** ROC curves for S (yellow), Cl (green), and Br (red) using three peaks at the *standard noise profile* and the *high noise profile*. In comparison to **(A)**, ROC curves **(C)** and **(D)** are *zoomed in to the top left corner*. AUCs for all ROC curves can be found in Table 2. The chosen tradeoffs between FPR and TPR for the analysis on the measured datasets are marked (×).

Table 2: AUCs for the *standard noise profile* and the *high noise profile*. If decimals are given, the given values are rounded.

noise profile:	*standard*			*high*		
# peaks:	three	four	five	three	four	five
CHNOPS	1.0	1.0	1.0	1.0	1.0	1.0
S	0.992	0.996	0.999	0.974	0.985	0.996
Cl	0.998	0.999	1.0	0.994	0.997	0.999
Br	0.997	0.997	1.0	0.994	0.997	1.0
B	1	1	1	1	1	1
Se	1	1	1	1	1	1

classifier. For Se and B we get an area under curve of 1 in all cases.
We find that for all elements, area under curve increases for classifiers using isotope patterns with more peaks. This is not surprising, since isotope patterns with more peaks contain more information. In particular for the *high noise profile*, using more peaks improves the prediction.

Table 3: How many false positive predictions do we have to accept in order to reach a certain sensitivity? We report false positive rates (FPR) for all 18 classifiers at fixed true positive rates (TPR). TPR 0.998 corresponds to two missed positive predictions in 1000 examples; TPR 1 means missing not a single positive example. Results for the *standard noise profile*. FPR and TPR used for the analysis on the measured datasets in **bold**. 'N/A', classifier cannot reach the desired TPR.

		three peaks			four peaks			five peaks		
	desired TPR	0.998	0.999	1	0.998	0.999	1	0.998	0.999	1
resulting FPR	CHNOPS	0.0	**0.0**	0.005	0.0	**0.0**	0.001	0.0	**0.0**	0.001
	S	**0.265**	0.313	N/A	**0.162**	0.173	0.332	**0.055**	0.072	0.160
	Cl	0.056	**0.057**	0.063	0.054	**0.055**	0.060	0.004	**0.008**	0.027
	Br	0.012	0.013	**0.013**	0.012	**0.012**	0.013	0.0	0.0	**0.0**
	B	0.0	0.0	**0.0**	0.0	0.0	**0.0**	0.0	0.0	**0.0**
	Se	0.0	0.0	**0.0**	0.0	0.0	**0.0**	0.0	0.0	**0.0**

To further evaluate the tradeoff between sensitivity and specificity, we analyze the FPRs for all 18 classifiers at fixed TPRs for the *standard noise profile* (see Table 3; results for the *high noise profile* can be found in Supplementary Table S.5). For Cl, FPR is below 6.5 % in all cases, and for Br, FPR is below 1.5 % in all cases. As expected, FPR improves when using more isotope peaks. For S, we get higher FPRs than for the other classifiers. For three peaks, we do not reach a TPR of 100 %.
We additionally evaluated the B classifiers against isotope patterns with extremely high noise level ($\sigma_{\mathrm{IA}} = 0.008$, $\sigma_{\mathrm{IR}} = 0.1$, and $\sigma_{\mathrm{M}} = 0.002\,\mathrm{Da}$) to account for the range of natural variation in the isotopic distribution of B. The prediction quality remains the same.
For CHNOPS, we get a FPR 0.0 up to a TPR of 99.9 %. Even for 100 % true positives, the FPR is below 1 %.

We manually chose the tradeoff between FPR and TPR used for the analysis on the measured datasets, see Table 3. If several voting thresholds fall into the particular FPR/TPR tradeoff, we choose a reasonable one (see Supplementary Table S.6).
We have also tested the training of different classifiers for different mass ranges, since the mass of a compound influences the shape of its isotope pattern. However, using two or three classifiers for different mass ranges did not improve prediction quality (data not shown).

Prediction quality on measured data

In the second part of our evaluation, we apply our classifiers to three measured datasets. The measured isotope patterns have different numbers of peaks; for each pattern, we choose the appropriate classifier. For each classifier, we use the voting threshold with the best tradeoff between FPR and TPR (see Table 3 and Supplementary Table S.6). Note that there is only one positive example each for B and Se.
See Table 4 for prediction results. In *no case* do we observe a false negative prediction for elements S, Cl, Br, B, and Se. To this end, our classifiers have *perfect sensitivity* for the measured data. A false negative prediction would result in an element being excluded from further analysis, despite being present in the compound.
Recall that false positive prediction, in contrast, can be corrected in the downstream analysis of identifying the exact molecular formula. For Cl, we get seven false positive predictions: See Supplementary Fig. S.2 for the isotope patterns of bromazil ($C_9H_{13}BrN_2O_2$) and pinensin A ($C_{96}H_{139}N_{27}O_{30}S_2$). We note that pinensin A with mass 2213.962 Da is much larger than the mass range used for training our classifiers. For both compounds, the CHNOPS classifier reports a conflicting prediction. For Br, we have three false positive predictions in the *CASMI* dataset. For B and Se we do not get any false positive predictions.
For S, we get 40 false positive predictions and, thus, would use an unnecessary element for the molecular formula computation. Of these, 18 compounds are from the *myxo* dataset. Compounds in the *myxo* dataset are much larger than compounds in the other datasets; for larger compounds, the presence of sulfur is more veiled, in particular if only a single sulfur atom is contained. See Supplementary Fig. S.3 for an example.

Molecular Formula Identification

To show the use of our predictors in a pipeline for molecular formula identification, we integrated it into the SIRIUS pipeline for molecular formula identification using isotope patterns[24], see ref.[32,53] for details. Since biomolecules in the three datasets may potentially contain elements CHNOPSClBrBSe, we use this set of elements when running SIRIUS to compare against. For our predictors, we use elements CHNOP plus those from SClBrBSe that were predicted (including false positive predictions) to be present in the biomolecule. We use a mass deviation of 10 ppm.
For the *myxo* dataset, we excluded pinensin A ($C_{96}H_{139}N_{27}O_{30}S_2$, 2213.962 Da). Processing this compound using the alphabet CHNOPSClBrBSe, we ran into an out-of-memory exception. Processing it with the predictor-based version, 19 746 670 candidate molecular formulas were evaluated in 15 min, and the correct molecular formula was at rank 10414. For three

Table 4: Classification results on the three measured datasets. Number of real positives (P) and negatives (N), as well as number of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions. Recall that false positive predictions (orange) can be corrected in the downstream analysis of molecular formula identification; in contrast, false negative prediction (red) cannot.

		real		predicted			
		P	N	TP	TN	FP	FN
myxo	CHNOPS	78	10	78	10	0	0
	S	24	64	24	46	18	0
	Cl	8	80	8	79	1	0
	Br	0	88	0	88	0	0
	B	1	87	1	87	0	0
	Se	1	87	1	87	0	0
pesticide	CHNOPS	15	28	15	28	0	0
	S	14	29	14	27	2	0
	Cl	27	16	27	15	1	0
	Br	1	42	1	42	0	0
	B	0	43	0	43	0	0
	Se	0	43	0	43	0	0
CASMI	CHNOPS	368	164	368	164	0	0
	S	146	386	146	366	20	0
	Cl	152	380	152	375	5	0
	Br	13	519	13	516	3	0
	B	0	532	0	532	0	0
	Se	0	532	0	532	0	0

compounds in the *myxo* dataset, we had to extend the mass deviation to 30 ppm to include the correct molecular formula.

For the *CASMI* dataset, we excluded four compounds and mass spectra, as these were solvents and contain an exceptionally high number of fluorines. For the remaining compounds, we used an upper bound of 6 fluorines (the remaining compound with the highest number of fluorines was flufenoxuron, $C_{21}H_{11}ClF_6N_2O_3$, an insecticide). We assume for both approaches that we know whether iodine is present in the compound; as discussed below, the presence of iodine can be determined using the fragmentation pattern of the compound.

All running time measurements were performed on an Intel Sandy Bridge processor at 2.3 GHz with 128 GB RAM.

We observe a massive decrease in the number of candidate molecular formulas we have to consider (see Table 5): We observe a 32-fold to 261-fold decrease in candidate molecular formulas filtered with SENIOR rules[35], compared to the default method. As expected, this results in a similar decrease in running times: The total running time for processing all three datasets decreased from 18.6 min to 9.5 sec. We also observe a slight improvement in identification performance.

Table 5: Results of the molecular formula identification using SIRIUS. We use CHNOP-SClBrBSe when running the identification without element prediction (top). For our predictors, we use CHNOP plus those elements from SClBrBSe that were predicted to be present (bottom). For *CASMI*, we also added I to the set of elements for all compounds that contain iodine, and allow up to six fluorines for all compounds. We report correct identifications (top 1, top 3, top 10), average number of decompositions (before and after SENIOR filtering) and average running time for all three datasets.

	Dataset	*myxo*	*pesticide*	*CASMI*
	No. spectra	87	43	528
	Median mass (Da)	592.315	260.016	281.484
without element prediction	Correct identifications	12	13	287
	Top 3 identifications	20	36	360
	Top 10 identifications	30	42	433
	Average # decompositions			
	... before SENIOR filtering	2737085	2242.0	44007.4
	... after SENIOR filtering	571072.4	836.2	13052.5
	Average running time in ms	11277.2	20.1	256.9
with element prediction	Correct identifications	13	13	287
	Top 3 identifications	21	37	364
	Top 10 identifications	32	42	435
	Average # decompositions			
	... before SENIOR filtering	12708.8	73.4	597.6
	... after SENIOR filtering	2181.0	26.1	174.9
	Average running time in ms	64.5	1.0	7.2

Conclusion

We have presented classifiers to predict the presence of uncommon elements, based on the isotope pattern of the biomolecule. The thresholds have been chosen with regards to the setting discussed in the introduction, where an extremely high sensitivity is more important than a low false positive rate. Evaluating the classifiers on a real-world dataset, we found no false negative predictions and would, thus, never miss an element for molecular formula identification. Depending on the application, it is also possible to select thresholds based on other criteria, e.g. keeping the false positive rate low.

We have shown how our classifiers improve the subsequent steps of analyzing isotope patterns, in particular by significantly reducing the number of candidate molecular formulas we have to consider. Molecular formula identification based on isotope patterns alone is not adequate in practice (see Table 5), but results can be much improved by accompanying the isotope pattern data by fragmentation data: For example, for 56 of 60 compounds in the complete *pesticide* dataset, Böcker and Dührkop[31] inferred the correct molecular formula using both data types combined; in all cases, the correct answer was ranked in the top 5. Running time differences will be much larger if we also analyze tandem mass spectra, as processing

a single candidate may require several seconds: Böcker and Dührkop[31] report an instance where 3106 candidates were considered, requiring a total of 12.5 h for this compound.
We have used random forests for the prediction as they deliver high quality predictions and are not susceptible to noise, as has been shown in the noise parameter evaluation. Classifiers have been trained on simulated data with relatively high noise, and have not been tailored toward any particular instrumental platform. Training on instrument-specific data may further increase prediction quality.
From the uncommon elements we have considered in this paper, sulfur is the most abundant one (Table 1). It is also the element which is hardest to predict, since the effect of a single sulfur atom on the isotope pattern is less pronounced than for other uncommon elements. We have not taken into account the presence of the characteristic peak at +1.995796 Da from the monoisotopic peak in the isotopic fine structure, as detection of this peak is dependent on the resolution of the MS instrument, which is not covered in this paper. Assuming adequate resolution, combining the classifiers presented here with a classifier based on the +1.995796 Da peak will result in improved classification performance for sulfur.
Using the isotope pattern alone, it is not possible to predict the presence of iodine and fluorine. To a certain degree, it is possible to predict iodine using tandem mass spectral data, based on characteristic common losses and fragments[31]. In contrast, the presence of fluorine is hard to predict using tandem MS data, too; hence, it might be advisable to allow for a few fluorine atoms in the molecular formula candidates, if there is any reason to believe that fluorine might be present in the sample compounds[4,9]. Predicting the presence of other elements that have characteristic isotope distributions, such as silicon, is also possible by the methods presented here if required by the application.
Selenium-containing compounds result in a monoisotopic peak with small intensity, which may not be detected in the mass spectrum. In this case, we cannot proceed by decomposing the monoisotopic peak. Instead, we can decompose the average mass (that is, the molecular mass) of the compound[22]; the average mass can be estimated from the experimental spectrum by taking the weighted average of isotope peak masses.
Until now, the selection of elements to include for the calculation of molecular formulas has been a manual decision. This has not been seen as a critical bottleneck in the past, as molecular formula determination was mainly performed as a manual "low throughput" procedure in the course of structural elucidation. Today, methods for small molecule analysis are being applied at a large scale in numerous research fields, including diverse applications such as the investigation of industrially relevant plants and microbes, clinical pharmacology studies, and natural products discovery. The underlying mass spectrometry applications share a tendency to produce datasets from which hundreds to thousands of unknown molecules have to be analyzed. Generation of molecular formulas is among the most commonly performed first-pass analyses in unknown identification workflows, and several advanced methods rely on this upfront evaluation step[31,54]. Whereas efforts have been made to evaluate the fit of a measured isotope pattern against the theoretical isotope pattern of a molecular formula, as well as evaluation of the corresponding statistical significance and robustness of the method[55], the issue of predicting the presence or absence of elements has been widely disregarded. To this end, we believe that our method is an important step in the automated annotation of novel biomolecules from mass spectrometry data.

Supporting Information The Supporting Information is available free of charge via the Internet at `http://pubs.acs.org`. Additional information on methods and results as noted in the text (PDF).

Corresponding Author: Correspondence should be addressed to S.B. (`sebastian.boecker@uni-jena.de`).

Funding: M.M. funded by Deutsche Forschungsgemeinschaft (BO 1910/16). F.H. funded by Carl Zeiss Stiftung.

Competing Interests: The authors declare that they have no competing financial interests.

Acknowledgement

We thank Emma Schymanski and Michael Stravs for providing the *pesticide* data. We particularly thank Emma Schymanski and Martin Krauss for providing the *CASMI* dataset.

References

(1) Hoffmann, T.; Krug, D.; Hï¿½ttel, S.; Mï¿½ller, R. Improving natural products identification through targeted LC-MS/MS in an untargeted secondary metabolomics workflow. *Anal Chem* **2014**, *86*, 10780–10788.

(2) Psychogios, N. et al. The Human Serum Metabolome. *PLoS One* **2011**, *6*, e16957.

(3) Issaq, H. J.; Van, Q. N.; Waybright, T. J.; Muschik, G. M.; Veenstra, T. D. Analytical and statistical approaches to metabolomics research. *J Sep Sci* **2009**, *32*, 2183–2199.

(4) Fujimori, D. G.; Walsh, C. T. What's New in Enzymatic Halogenations. *Curr Opin Chem Biol* **2007**, *11*, 553–560.

(5) Pauletti, P. M.; Cintra, L. S.; Braguine, C. G.; da Silva Filho, A. A.; Andrade e Silva, M. L.; Cunha, W. R.; Januário, A. H. Halogenated indole alkaloids from marine invertebrates. *Mar Drugs* **2010**, *8*, 1526–1549.

(6) Elshahawi, S. I.; Trindade-Silva, A. E.; Hanora, A.; Han, A. W.; Flores, M. S.; Vizzoni, V.; Schrago, C. G.; Soares, C. A.; Concepcion, G. P.; Distel, D. L.; Schmidt, E. W.; Haygood, M. G. Boronated tartrolon antibiotic produced by symbiotic cellulose-degrading bacteria in shipworm gills. *Proc Natl Acad Sci U S A* **2013**, *110*, E295–E304.

(7) Çakır, Ö.; Turgut-Kara, N.; Arı, Ş. In *Advances in Selected Plant Physiology Aspects*; Montanaro, G., Dichio, B., Eds.; InTech, 2012; pp 209–232.

(8) Yuan, L.; Zhu, Y.; Lin, Z.-Q.; Banuelos, G.; Li, W.; Yin, X. A novel selenocystine-accumulating plant in selenium-mine drainage area in Enshi, China. *PLoS One* **2013**, *8*, e65615.

(9) O'Hagan, D.; Schaffrath, C.; Cobb, S. L.; Hamilton, J. T. G.; Murphy, C. D. Biochemistry: biosynthesis of an organofluorine molecule. *Nature* **2002**, *416*, 279.

(10) Dembitsky, V. Bromo-and Iodo-Containing Alkaloids from Marine Microorganisms and Sponges. *Russ J Bioorg Chem* **2002**, *28*, 170–182.

(11) Harvey, A. L. Natural products in drug discovery. *Drug Discov Today* **2008**, *13*, 894–901.

(12) Gerwick, W. H.; Moore, B. S. Lessons from the past and charting the future of marine natural products drug discovery and chemical biology. *Chem Biol* **2012**, *19*, 85–98.

(13) Cragg, G. M.; Newman, D. J. Natural products: a continuing source of novel drug leads. *Biochim Biophys Acta* **2013**, *1830*, 3670–3695.

(14) Kind, T.; Fiehn, O. Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics* **2006**, *7*, 234.

(15) Scheubert, K.; Hufsky, F.; Böcker, S. Computational Mass Spectrometry for Small Molecules. *J Cheminformatics* **2013**, *5*, 12.

(16) Matsuda, F.; Shinbo, Y.; Oikawa, A.; Hirai, M. Y.; Fiehn, O.; Kanaya, S.; Saito, K. Assessment of metabolome annotation quality: A method for evaluating the false discovery rate of elemental composition searches. *PLoS One* **2009**, *4*, e7490.

(17) Böcker, S.; Lipták, Zs. Efficient Mass Decomposition. Proc. of ACM Symposium on Applied Computing (ACM SAC 2005). 2005; pp 151–157.

(18) Böcker, S.; Lipták, Zs. A fast and simple algorithm for the Money Changing Problem. *Algorithmica* **2007**, *48*, 413–432.

(19) Dührkop, K.; Ludwig, M.; Meusel, M.; Böcker, S. Faster mass decomposition. Proc. of Workshop on Algorithms in Bioinformatics (WABI 2013). 2013; pp 45–58.

(20) Kubinyi, H. Calculation of isotope distributions in mass spectrometry: A trivial solution for a non-trivial problem. *Anal Chim Acta* **1991**, *247*, 107–119.

(21) Rockwood, A. L.; Haimi, P. Efficient calculation of accurate masses of isotopic peaks. *J Am Soc Mass Spectrom* **2006**, *17*, 415–419.

(22) Böcker, S.; Letzel, M.; Lipták, Zs.; Pervukhin, A. Decomposing metabolomic isotope patterns. Proc. of Workshop on Algorithms in Bioinformatics (WABI 2006). 2006; pp 12–23.

(23) Kind, T.; Fiehn, O. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics* **2007**, *8*, 105.

(24) Böcker, S.; Letzel, M.; Lipták, Zs.; Pervukhin, A. SIRIUS: Decomposing isotope patterns for metabolite identification. *Bioinformatics* **2009**, *25*, 218–224.

(25) 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*, 753–63.

(26) Valkenborg, D.; Mertens, I.; Lemière, F.; Witters, E.; Burzykowski, T. The isotopic distribution conundrum. *Mass Spectrom Rev* **2012**, *31*, 96–109.

(27) Loos, M.; Gerber, C.; Corona, F.; Hollender, J.; Singer, H. Accelerated isotope fine structure calculation using pruned transition trees. *Anal Chem* **2015**, *87*, 5738–5744.

(28) Rasche, F.; Svatoš, A.; Maddula, R. K.; Böttcher, C.; Böcker, S. Computing fragmentation trees from tandem mass spectrometry data. *Anal Chem* **2011**, *83*, 1243–1251.

(29) Pluskal, T.; Uehara, T.; Yanagida, M. Highly accurate chemical formula prediction tool utilizing high-resolution mass spectra, MS/MS fragmentation, heuristic rules, and isotope pattern matching. *Anal Chem* **2012**, *84*, 4396–4403.

(30) Stravs, M. A.; Schymanski, E. L.; Singer, H. P.; Hollender, J. Automatic recalibration and processing of tandem mass spectra using formula annotation. *J Mass Spectrom* **2013**, *48*, 89–99.

(31) Böcker, S.; Dührkop, K. Fragmentation trees reloaded. *J Cheminformatics* **2016**, *8*, 5.

(32) Dührkop, K.; Hufsky, F.; Böcker, S. Molecular Formula Identification Using Isotope Pattern Analysis and Calculation of Fragmentation Trees. *Mass Spectrom* **2014**, *3*, S0037.

(33) Mallard, W. G.; Sparkman, O. D. *NIST Standard Reference Database 1A: NIST/EPA/NIH Mass Spectral Library (NIST 14) and NIST Mass Spectral Search Program (Version 2.2) User's Guide*; U.S. Department of Commerce, 2014.

(34) Breiman, L. Random forests. *Mach Learn* **2001**, *45*, 5–32.

(35) Senior, J. Partitions and Their Representative Graphs. *Amer J Math* **1951**, *73*, 663–689.

(36) Audi, G.; Wapstra, A.; Thibault, C. The AME2003 atomic mass evaluation (II): Tables, graphs, and references. *Nucl Phys A* **2003**, *729*, 129–336.

(37) Wieser, M. E. Atomic weights of the elements 2005 (IUPAC Technical Report). *Pure Appl Chem* **2006**, *78*, 2051–2066.

(38) de Laeter, J. R.; Böhlke, J. K.; Biévre, P. D.; Hidaka, H.; Peiser, H. S.; Rosman, K. J. R.; Taylor, P. D. P. Atomic weights of the elements. Review 2000 (IUPAC Technical Report). *Pure Appl Chem* **2003**, *75*, 683–800.

(39) Dittwald, P.; Valkenborg, D.; Claesen, J.; Rockwood, A. L.; Gambin, A. On the Fine Isotopic Distribution and Limits to Resolution in Mass Spectrometry. *J Am Soc Mass Spectrom* **2015**, *26*, 1732–1745.

(40) Breiman, L. Bagging predictors. *Mach Learn* **1996**, *24*, 123–140.

(41) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*, 2nd ed.; Springer Series in Statistics; Springer-Verlag New York, 2009.

(42) Hastings, J.; de Matos, P.; Dekker, A.; Ennis, M.; Harsha, B.; Kale, N.; Muthukrishnan, V.; Owen, G.; Turner, S.; Williams, M.; Steinbeck, C. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res* **2013**, *41*, D456–D463.

(43) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL bioactivity database: an update. *Nucleic Acids Res* **2014**, *42*, D1083–D1090.

(44) Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* **2006**, *34*, D668–D672.

(45) Wishart, D. S. et al. HMDB 3.0: The Human Metabolome Database in 2013. *Nucleic Acids Res* **2013**, *41*, D801–D807.

(46) Kanehisa, M.; Sato, Y.; Kawashima, M.; Furumichi, M.; Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* **2016**, *44*, D457–D462.

(47) Afendi, F. M.; Okada, T.; Yamazaki, M.; Hirai-Morita, A.; Nakamura, Y.; Nakamura, K.; Ikeda, S.; Takahashi, H.; Altaf-Ul-Amin, M.; Darusman, L. K.; Saito, K.; Kanaya, S. KNApSAcK family databases: integrated metabolite-plant species databases for multifaceted plant research. *Plant Cell Physiol* **2012**, *53*, e1.

(48) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound databases. *Nucleic Acids Res* **2016**, *44*, D1202–D1213.

(49) Gu, J.; Gui, Y.; Chen, L.; Yuan, G.; Lu, H.-Z.; Xu, X. Use of natural products as chemical library for drug discovery and network pharmacology. *PLoS One* **2013**, *8*, 1–10.

(50) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: a free tool to discover chemistry for biology. *J Chem Inf Model* **2012**, *52*, 1757–1768.

(51) Chen, C.; Liaw, A.; Breiman, L. *Using Random Forest to Learn Imbalanced Data*; Technical Report 666, 2004.

(52) Bekkar, M.; Alitouche, T. A. Imbalanced Data Learning Approaches. *Int J Data Mining Knowl Manage Process* **2013**, *3*, 15–33.

(53) Dührkop, K.; Scheubert, K.; Böcker, S. Molecular Formula Identification with SIRIUS. *Metabolites* **2013**, *3*, 506–516.

(54) Dührkop, K.; Shen, H.; Meusel, M.; Rousu, J.; Böcker, S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc Natl Acad Sci U S A* **2015**, *112*, 12580–12585.

(55) Ipsen, A.; Want, E. J.; Ebbels, T. M. D. Construction of confidence regions for isotopic abundance patterns in LC/MS data sets for rigorous determination of molecular formulas. *Anal Chem* **2010**, *82*, 7319–7328.