MESSAR: Automated recommendation of metabolite substructures from tandem mass spectra

Youzhong Liu[1,2‡], Aida Mrzic[1,2‡], Pieter Meysman[1,2], Thomas De Vijlder[3], Edwin P. Romijn[3], Dirk Valkenborg[4], Wout Bittremieux[1,2,5], Kris Laukens[1,2*]

1 Department of Mathematics and Computer Science, University of Antwerp, Antwerp, Belgium
2 Biomedical Informatics Network Antwerpen (biomina), University of Antwerp, Antwerp, Belgium
3 Pharmaceutical Development & Manufacturing Sciences (PDMS), Janssen Research & Development, Beerse, Belgium
4 Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Hasselt University, Diepenbeek, Belgium;
5 Department of Genome Sciences, University of Washington, Seattle WA 98195, USA.

‡ Authors contributed equally and should be considered co-first authors.
* kris.laukens@uantwerpen.be

Abstract

Despite the increasing importance of non-targeted metabolomics to answer various life science questions, extracting biochemically relevant information from metabolomics spectral data is still an incompletely solved problem. Most computational tools to identify tandem mass spectra focus on a limited set of molecules of interest. However, such tools are typically constrained by the availability of reference spectra or molecular databases, limiting their applicability to identify unknown metabolites. In contrast, recent advances in the field illustrate the possibility to expose the underlying biochemistry without relying on metabolite identification, in particular via substructure prediction. We describe an automated method for substructure recommendation motivated by association rule mining. Our framework captures potential relationships between spectral features and substructures learned from public spectral libraries. These associations are used to recommend substructures for any unknown mass spectrum. Our method does not require any predefined structural motifs or features, and therefore it can be used for the partial identification of unknown unknowns. The method is called MESSAR (MEtabolite SubStructure Auto-Recommender) and is implemented in a free online web service available at `messar.biodatamining.be`.

Introduction

Metabolomics is an emerging "omics" science involving the high-throughput analysis of metabolites or small biomolecules, with highly relevant applications in drug and biomarker discovery [1, 2]. One standard method for metabolite analysis is mass spectrometry (MS), preceded by a separation technique, such as gas chromatography (GC) or liquid chromatography (LC). Advances in MS instrumentation enable the simultaneous detection and quantification of thousands of metabolites in a biological sample. Chemical identification of these metabolites is a key step towards biochemical interpretation of studied samples. To obtain structural information, tandem MS (MS/MS) is applied

to record the fragment m/z of targeted molecules. Structure elucidation from MS/MS data has always been a challenging and time-consuming task with a vast number of potentially interesting metabolites that are still unknowns. The main reason is that current MS/MS databases (spectral libraries) only contain a limited number of historical spectra, far below the number of metabolites in reality [3, 4].

Advances in computational tools have led to a considerable extension of the search space that can be examined and have resulted in an improvement of the identification accuracy by using massive molecular databases (for example, PubChem currently contains over 100 million compounds [5]). These tools start by filtering the molecular database using the precursor m/z of the unknown spectra, yielding up to thousands of structure candidates.

To subsequently score and rank these candidates two categories of algorithms have been proposed. First, *in silico* fragmentation tools simulate theoretical spectra for each candidate metabolite and compare those with the query spectrum [6–9]. Second, machine learning (ML) methods learn intermediate representations, such as molecular fingerprints [10] from historical spectrum–structure relationships. These representations are then used to score spectrum–candidate matches. A typical example hereof is the CSI:FingerID tool [11]. However, both types of algorithms still have certain limitations. For example, *in silico* fragmentation tools only cover simple fragmentations and cannot accurately simulate complex rearrangement reactions [12]. Additionally, none of these tools can identify "unknown unknowns", i.e. compounds that have not been structurally described yet and are therefore not present in any molecular database. As a result typically only a fraction of compounds can be identified correctly [13].

Recently a third category of computational tools has been introduced in non-targeted metabolomics. With a focus on the *de novo* identification of unknowns, these tools aim to predict substructures rather than the full metabolite structure. The basic concept for this strategy is that metabolites often share substructures, resulting in similar patterns in their MS/MS spectra. Typical spectral features are product ions, neutral losses, or mass differences [14–17].

One important tool to explore spectral similarity is the Global Natural Products Social Molecular Networking (GNPS) resource [18]. GNPS consists of a large metabolite network where metabolites with similar MS/MS spectra are connected so that structurally annotated metabolites can be used for the identification of their neighbors. However, such a network-based approach may fail to connect metabolite pairs if they have a low spectral similarities despite sharing important substructures such as small functional groups. Additionally, the annotation of neighboring nodes still requires manual intervention.

MS2Analyzer is an alternative approach which enables user-defined searches of MS/MS features in thousands of spectra to retrieve substructure information [19]. The drawback of this method, however, is that meaningful MS/MS features can be overlooked if they are not explicitly specified by users.

MS2LDA is a recent framework proposed by van der Hooft et al. [20–22]. It decomposes unlabeled MS/MS spectra into patterns of co-occurring fragments and losses, referred to as "Mass2Motifs", which are indicative of biological substructures. These Mass2Motifs patterns are automatically extracted from complex MS/MS spectra using unsupervised text mining techniques. This approach combines the notion of "overall similarity" from molecular networking and the "important MS/MS features" from MS2Analyzer. However, like MS2Analyzer, the extracted motifs have to be structurally annotated based on expert knowledge, which requires extensive domain expertise and is time-consuming.

Here we introduce a new method for the recommendation of substructures for MS/MS spectra, working independent of molecular databases. Our tool, called MESSAR (MEtabolite SubStructure Auto-Recommender), is inspired on the concept of association

rule mining (ARM). ARM has been designed to discover interesting relations based on frequently co-occurring items, and it has previously been used to find relations between unassigned mass spectra [23–25].

We use a collection of labeled spectra (reference spectra with a known corresponding molecular structure from a spectral library) to mine co-occurring patterns of "MS/MS features + substructures", yielding "MS/MS feature(s) → substructure" rules. These rules capture recurring patterns found in mass spectra and assign them potential substructures. As such, they can be used to partially replace expert-driven annotations in tools such as MS2Analyzer and MS2LDA.

Similar to CSI:FingerID, MESSAR is a supervised tool to predict metabolite substructures. However, unlike CSI:FingerID, the model training (rule generation) of MESSAR is conducted in an automated fashion. Whereas CSI:FingerID relies on predefined molecular fingerprints, MESSAR directly generates substructures for each training metabolite by breaking its chemical bonds. Moreover, MESSAR rules are filtered automatically through comparison with "decoy rules" (random rules derived from the same training data). All statistically sound rules are used for substructure prediction.

During prediction, MESSAR matches spectral features from unknown MS/MS spectra against the rule database and retrieves matching substructure candidates. We rank the candidates for users according to an "interestingness" score. MESSAR is currently designed for positive ion mode LC-MS/MS data. It is available as a free online web service at `messar.biodatamining.be`.

Materials and Methods

Training spectral libraries

MESSAR generates rules from target and decoy spectral libraries built by Scheubert et al. According to the data descriptions [26], the target library consists of 4138 positive ion high-quality labeled spectra acquired on Q-TOF instruments. For each spectrum, Scheubert et al. have computed fragmentation tree that annotates a subset of fragments with molecular formulas and removed non-annotated peaks that usually represent isotopic peaks, chemical noise... [11]. Meanwhile, neutral exact masses were assigned to formula annotated fragments. The decoy library was a randomized version of the target library. The decoy process kept the library labels (molecular structures) and randomized the corresponding mass spectra through re-rooted fragmentation trees (Fig 1A). Such process mimics very "noisy" experimental spectra, therefore extracted patterns can probably reflect spurious feature-substructure relations in the target library. For both target and decoy libraries, mass spectra representing the same structure were combined, and duplicated fragments were removed. The final training data consists of 3146 target and decoy spectra (S1 Data).

Spectral feature extraction

Both fragment masses and distances between fragments (mass differences) were interesting spectral features for model training. However, if a training spectrum has 20 mass peaks, extracted spectral features consist of $\binom{20}{2} = 190$ mass differences. Rule generation can be computationally expensive and highly dominated by mass differences with no structural information. Therefore, we only used edges of fragmentation trees as target spectral features since they represent losses due to fragmentation reactions and probably holds structural information (Fig 1B). For the decoy spectrum of the same molecule, we extracted a random subset of mass differences that had the same size of the corresponding

target spectrum (Fig 1D). In this way, similar amount of spectral features were used to generate target and decoy rules.

Substructure generation

For every metabolite in the spectral library a set of substructures is created using the "breaking of retrosynthetically interesting chemical substructures" (BRICS) algorithm [27] (RDKit in Python). We disconnect one or two chemical bonds in the metabolite and collect the resulting two or three substructures obtained from every iteration (Fig 1C). The following trivial substructures are discarded: (i) CHO substructures that contain less than six carbon and oxygen atoms, and (ii) CHON substructures with less than four non-H atoms. Substructures are stored as SMILES codes in the rule database.

Rule mining and statistical evaluation

After combining substructures with extracted spectral features, association rules were mined separately for target and decoy libraries (Fig 1EF). Details about ARM can be found in S1 Text. A MESSAR rule with shape $X \Rightarrow Y$ describes the potential dependency of a substructure (Y) on a spectral feature pattern (X can contain up to three co-occurring masses and mass differences). Therefore, each rule can be considered as a binary classifier that diagnoses whether a metabolite contains a substructure according to the corresponding spectral feature presence/absence. To evaluate the statistical performance of MESSAR rules, we created confusion matrices using their corresponding training library (e.g. the confusion matrix of a target rule was computed based on the entire target library):

Table 1. Confusion matrix for the MESSAR rule $X \rightarrow Y$.

	X	$\neg X$
Y	TP	FP
$\neg Y$	FN	TN

$\neg X = FP + TN$, number of metabolites whose spectra do not contain X
$\neg Y = FN + TN$, number of metabolites for which substructure Y is absent
TP, true positives; FP, false positives; FN, false negatives; TN, true negatives

We use three evaluation measures for binary classifiers, namely *lift*, *mcc* and F_1:

$$\boldsymbol{lift}(X \rightarrow Y) = \frac{TP \times N}{(TP + FN)(FP + TN)}$$

$$\boldsymbol{mcc}(X \rightarrow Y) = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

$$\boldsymbol{F_1}(X \rightarrow Y) = \frac{2TP}{2TP + FP + FN}$$

Here $N = TP + FP + FN + TN$. The *lift* measures how much more often X and Y occur together than expected if they were statistically independent [28], with a lift of 1 that indicates independence. It is able to detect interesting rules that contain rare spectral patterns or substructures. The *mcc* (Matthews correlation coefficient) is a balanced measure of TP, FP, FN, and TN, and indicates the agreement or disagreement between prediction and observation [29]: a coefficient of +1 represents a perfect prediction, 0 no better than random prediction, and −1 indicates total disagreement. F_1 (or *F-score*) is

an harmonic mean of precision and recall of a binary classifier. *F-score* reaches its best value at 1 (perfect precision and recall).

Rule filtering

According to the definition of above metrics, only rules with a *lift* higher than 1 and a positive *mcc* are statistically sound (better than random). We applied these basic filters on both target and decoy rules. Target rules were further filtered by defining a *F-score* threshold for 1% FDR. Based on the target-decoy method, the simple FDR associated with a particular *F-score* threshold is the ratio between number of accepted decoy rules (above the threshold) and target rules [30]. The simple FDR was estimated using the *prozor* package in R.

Web tool for substructure recommendation

The MESSAR web tool recommends substructures for an input spectrum based on the pretrained rule database (Fig 1). MESSAR expects as input a list of MS/MS peaks and the precursor m/z. MESSAR then uses the fragments and the computed mass differences to query its database for matching rules. Additionally, multiple rules that suggest the same substructure are aggregated. The resulting MESSAR output is a list of substructure recommendations, which can be sorted according to the sum of *lift*, *mcc* or *F-scores*. The client interface of the web tool was developed using the R Shiny framework.

Test data

To evaluate the quality of substructure prediction, we used labeled spectra derived from two spectral libraries (S2 Data): (i) "MASSBANK" was downloaded from GNPS website(`https://gnps.ucsd.edu/ProteoSAFe/libraries.jsp`), the library contains all types of compounds analyzed on Q-TOF instruments, (ii) "DRUG" consists of standards (mostly drugs) acquired on Q Exactive and TripleTOF instruments (`http://mona.fiehnlab.ucdavis.edu/`), complemented by other drug-related compounds provided by Janssen Pharmaceutica, all acquired on a Q Exactive instrument. If multiple spectra were available for the same compound, the spectrum with the highest total ion intensity was kept. Both datasets were filtered so that there was no overlap in molecular structures with the spectral library used for MESSAR training. After all, "MASSBANK" and "DRUG" contained 2759 and 1005 test spectra, respectively.

Fisher's exact test

Fisher's exact test (F test) is used to evaluate the statistical significance of substructures that have been correctly predicted by MESSAR. For each test spectrum the F test is performed on a contingency table (same shape as Table 1) examining relationships between "true/false" and "predicted/left-behind" substructures. The total number of rules $N = TP+FP+FN+TN$ remains 1740, which represents all potential substructure recommendations in the rule database. A substructure is correctly predicted if it is present in the reference molecule corresponding to the test spectrum. TP represents the number of true substructures that are correctly predicted by MESSAR, while FP signifies predicted substructures that do not occur in the reference molecule. The one-sided F test concludes whether MESSAR predicts more correct substructures than false positives.

Substructure recommendation by CSI:FingerID and MS2LDA

CSI:FingerID program (Windows GUI SIRIUS-4.0.1) was downloaded from `https://bio.informatik.uni-jena.de/software/sirius/`. The "DRUG" data was submitted to the GUI as a mgf format file. The precursor formula was not available for fragmentation tree building, and all elements were considered (classical CHNOPS along with Br, Cl, F and I since they might be present in drug compounds). The annotation tolerance was set at 20 ppm. CSI:FingerID predicted substructures as five types of molecular fingerprints, namely CDK, PubChem CACTVS, Klekota-Roth, FP3 and MACCS. All fingerprints are coded as SMARTS strings. Structure annotation was achieved by searching PubChem, and the top 10 candidates were saved for evaluation.

The same mgf file was submitted to `http://ms2lda.org/` for M2M searching and annotation. The minimum intensity of MS2 peaks was set at 1, and the width of MS2 bins at 0.005 Da. The motifs found were further inferred using predefined 500 GNPS motifs.

Results and Discussion

First, we briefly describe the model training (rule generation) procedure and statistical properties of generated rules. Second, the biochemical relevance of the rules is revealed through a comparison with MS2LDA patterns. Third, the substructure annotation performance is evaluated based on two independent sets of test spectra, and the statistical significance of the predictions is evaluated. Finally, the usefulness of the MESSAR output is compared to CSI:FingerID and MS2LDA.

Outline of MESSAR rule generation

MESSAR predicts molecular substructures from a MS/MS spectrum based on significant "MS/MS feature(s) $\rightarrow$ substructure"-like patterns (rules) in existing labeled spectra. To seek such patterns, the classical ARM algorithm was applied on a subset of GNPS spectral library. The algorithm discovers rules among labeled spectra based on two parameters, support and confidence (S1 Text). An overview of the rule generation procedure from the spectral library is depicted in Fig 1 that includes following steps described in the Materials and Methods section: (A) retrieving training data: target and decoy spectral libraries from passatutto software [26], (B) extracting spectral features from the target library, (C) generating molecular substructures of training compounds, (D) extracting spectral features from the decoy library, (E) ARM on target features and substructures followed by statistical evaluations, (F) generating random rules from the decoy database and scoring, (G) filtering target rules for statistical soundness and by FDR estimation with the Target-Decoy method. All target rules above the score threshold chosen were saved in the MESSAR rule database.

Statistical properties of MESSAR rules

Using the parameters defined in S1 Text, we have generated 10550 and 10597 rules from target and decoy spectral libraries, respectively (before the step G). In the target database we can observe several obvious substructure recommendations. For example, rules 1–3 in Table 2 consist of simple substructure recommendations based on exact mass difference matches. These rules usually capture the direct link between mass differences and loss of substructures. However, most rules describe potential latent associations between low-molecular-weight spectral features and complex substructures (Table 2, Rules 4–7) and cannot be explained directly.

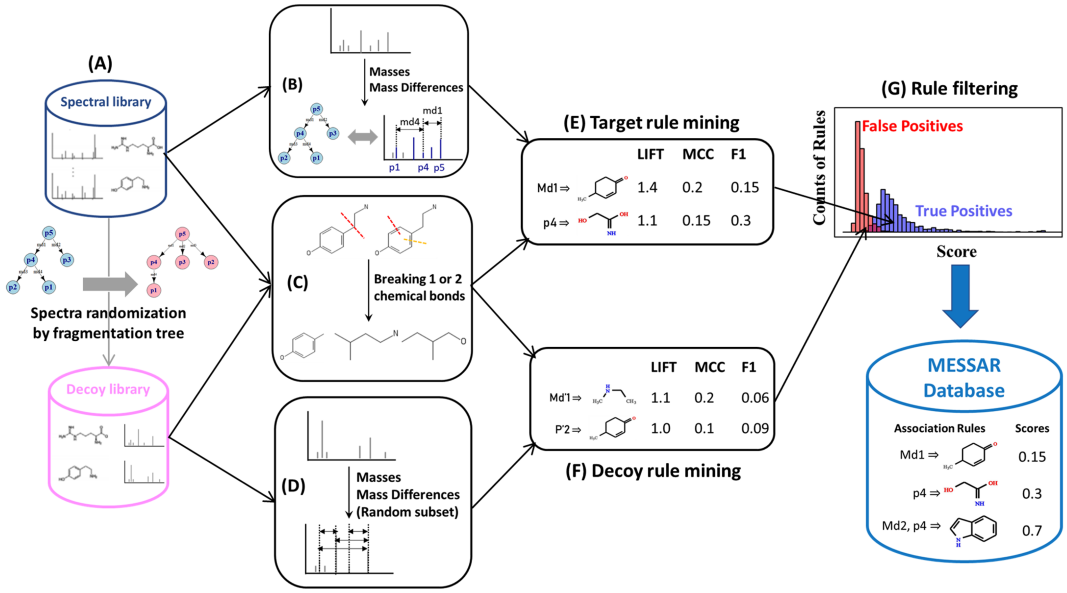

Fig 1. Workflow for target and decoy rule generation. Based on training data from an LC-MS/MS spectral library substructure recommendations are generated. Our method consists of (A) retrieving training MS/MS spectra, (B) extracting "calibrated" spectral features from the target library, (C) generating molecular substructures, (D) extracting spectral features from the decoy library, (E) rule mining using the target spectral features and substructures followed by statistical evaluations, (F) rule mining using the decoy database and statistical evaluations, (G) filtering target rules and saving valid rules to the final database.

Table 2. Examples of MESSAR rules from the target database.

ID	Type	Body (X)	Head (Y)	lift	mcc	F-score
1	MDiff.	78.047	c1ccccc1	1.2	0.1	0.25
2	MDiff.	35.9767	Cl	10.3	0.85	0.85
3	MDiff.	57.0215	CN=CO	2.4	0.2	0.28
4	Mass	132.0808	C(Cc1c[nH]c2ccccc12)N=	12.6	0.32	0.22
5	MDiff, MDiff	{58.0055, 46.0055}	OCC(O)=NCC(=O)O	8.0	0.25	0.24
6	Mass, Mass	{117.0699, 103.0542}	N1CCC(CC1)c1ccccc1	9.5	0.17	0.11
7	Mass, MDiff	{105.0699, 52.0313}	C1CCC2=CC(=O)CCC2(C)C1	10.1	0.2	0.16

During model training some spurious target spectral features–substructure associations will be created purely by chance. Therefore, it is important to discriminate between meaningful and spurious rules. Since a direct biochemical evaluation is not feasible, we verified whether the statistical measurements, *lift*, *mcc* and *F-score*, can be used to select meaningful rules (higher scores indicate more meaningful rules). The decoy rules served as negative controls and were evaluated in the same way as the target rules (Fig 1EF) Firstly, there was no exact agreement between the three metrics – *lift* and *F1* were almost independent from each other (Fig 2F), while a weak correlation was observed between *mcc* and *lift* (Fig 2D). Such inconsistency indicated that one metric can be more adapted than another depending on the dataset [31]. To make a selection, we compared the score distributions of target and decoy rules for each metric. It can be seen that the *lift* of most decoy rules lay between 0 and 5 (Fig 2A). However, a significant amount of decoy rules held an extremely high *lift*, probably because this measurement overestimated the importance of rare rules. *Mcc* showed better discrimination between target and decoy

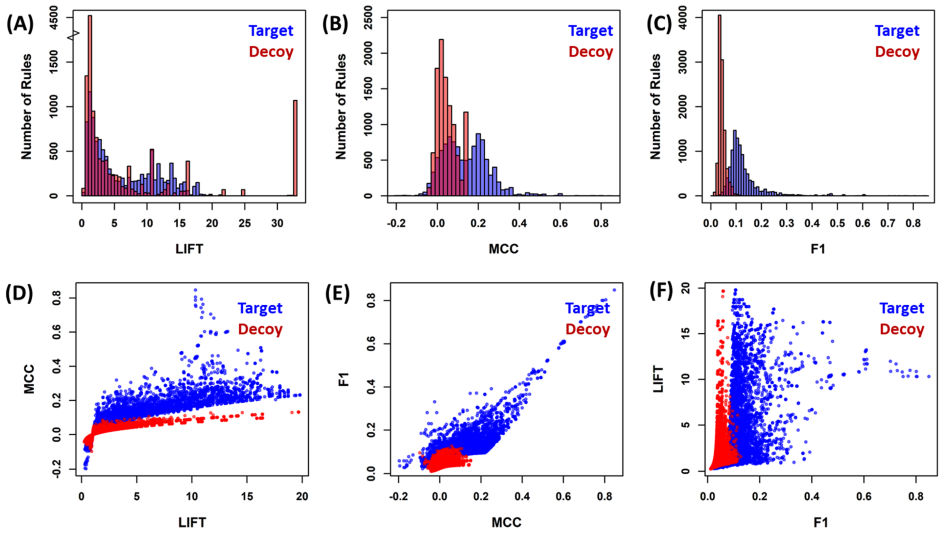

Fig 2. Statistical evaluation of MESSAR rules (A) mcc, (B) lift and (C) F-score distribution of target and decoy rules, scatter plot between (D) mcc and lift, (E) F1 and mcc, (F) lift and F1.

rules – all decoy rules held an *mcc* below 0.2, while target rules can have an *mcc* up to 0.8 (Fig 2B). Nevertheless, this statistical metric also classified more than 70% target rules below 0.2. The best target-decoy discrimination was observed for *F-score* – decoy rules centered around 0.03 while target rules displayed a skewed Gaussian-like distribution around 0.11 with a long right tail (Fig 2C). There was little overlap between two score distributions. Compared to *mcc*, *F-score* allowed a clear target-decoy separation for "weak" rules, i.e. rules with *mcc* below 0.2 (Fig 2E), meanwhile it was well correlated with *mcc* for "stronger" rules ($mcc > 0.2$). In summary, *F-score* appeared to be the best metric to distinguish meaningful rules from spurious patterns. It showed agreement with another metric while allowing better discrimination for rules with low *mcc*. Besides, no dependency was observed between *F-score* and parameters used for rule generation (S1 Fig). It will be the only metric for FDR control, rule ranking and model testing (substructure recommendation) in the manuscript. We note that *lift* and *mcc* can also be used for substructure recommendation, available through our web tool. After filtering rules for statistical soundness and for 1% FDR (threshold defined based on the shape of curve in S2 Fig), target rules with a *F-score* higher than 0.092 were kept. The final MESSAR database consists of 7785 association rules (S1 File). There are 1740 unique substructures in the database, indicating that several spectral features might suggest the same substructure.

Comparison between MESSAR rules and MS2LDA patterns

MS2LDA [20, 22] is an unsupervised tool that discovers patterns across fragmentation spectra. As it operates within a similar scope as MESSAR, a detailed comparison is warranted. The major difference between MESSAR and MS2LDA is that MS2LDA requires frequent patterns of spectral features (fragments and neutral losses) to be manually annotated by MS experts after extraction from unlabeled spectra (without structural information). These two steps result in a set of "Mass2Motifs" (M2Ms) that couple spectral features to molecular sub-structures, similar to MESSAR rules. Moreover,

spectral feature–substructure associations in M2Ms are highly confident, and as such they can be used as a ground truth to assess the biochemical relevance of MESSAR rules.

We compared the 7785 MESSAR rules with the 500 positive ion mode M2Ms derived from GNPS [20]. Each M2M consists of up to 200 motif features (fragments and losses), sorted by their probabilities. The twenty most probable features were searched against the MESSAR target rules using a 20 ppm mass window around the MESSAR spectral features, and feature types were required to agree (examples given in Fig 3A and 3B). When a single motif feature matched multiple MESSAR rules, *F-score* was used as tiebreaker to select the most important rule.

Overall, 159 out of 500 GNPS M2Ms can be partially explained by MESSAR rules. Among them 41 motifs have received manual annotations by MS experts (Fig 3A and 3C, S1 Table). Interestingly, the selected MESSAR rules showed a striking similarity to expert knowledge in terms of substructure prediction, ranging from simple (ethyl phenol, piperazine) to complex (indole, cinchonan-9-ol, cytisine) substructures. According to experts' knowledge, structural annotations of 17 motifs were identical or very similar to the substructure recommendation of at least one matched MESSAR rule. 19 motifs showed partial structural matches with MESSAR rules (with a few atoms of difference or belong to the same chemical family). Matched rules can even capture the structural similarity between motifs. For example, most rules matching with motifs 1, 32, 39, 50, and 274, which are all sterone related, correctly recommend the steroid core.

On the other hand, after manual annotation of the M2Ms there are still 400 unknown motifs, of which 118 receive substructure recommendations based on MESSAR rules (illustrated in Fig 3B and 3C). The examples in S2 Table further show the ability of MESSAR as an automated substructure recommender to support structural annotation of M2Ms.

The biochemical relevance of data-derived MESSAR rules has been validated through a comparison with the ground truth, expert-driven substructure annotation of the GNPS spectral patterns. These two approaches have been shown to be complementary, with about 30% of M2Ms that can be partially explained by MESSAR rules. Furthermore, an important strength of MESSAR compared to MS2LDA is that MESSAR requires no manual curation, while it is still able to derive meaningful substructure recommendations from spectral patterns.

Validating MESSAR rules using authentic standards

For a given unknown spectrum MESSAR can predict potential substructures based on the presence of certain spectral feature(s). This functionality is available to users in the MESSAR web tool. Here we illustrate the intermediate steps of the substructure prediction procedure based on an example of a reference compound in the "Drug" dataset (Cinnarizine in Fig 4). We queried the rule database using the peaks and mass differences from the test spectrum. With a 20 ppm mass window, the query resulted in 85 matched rules. There were only 60 individual substructures predicted as some recommendations include the same substructure (S3 Fig). As the agreement between rules predicting the same substructure increases the likelihood of that substructure being present, the final recommendation score is summed over all rules that predict the same substructure. In this way, substructure candidates were scored and ranked, and the top 20 candidates were used for performance evaluation. Meanwhile, Cinnarizine can be decomposed into 141 unique theoretical substructures (see Method section). However, only 18 of them were covered by our 1740-substructure rule database. By taking the top 20 candidates, MESSAR predicted 10 substructures correctly (out of 18), translating to a smaller-than-0.001 *p-value* (see Fisher's exact test in Method section).

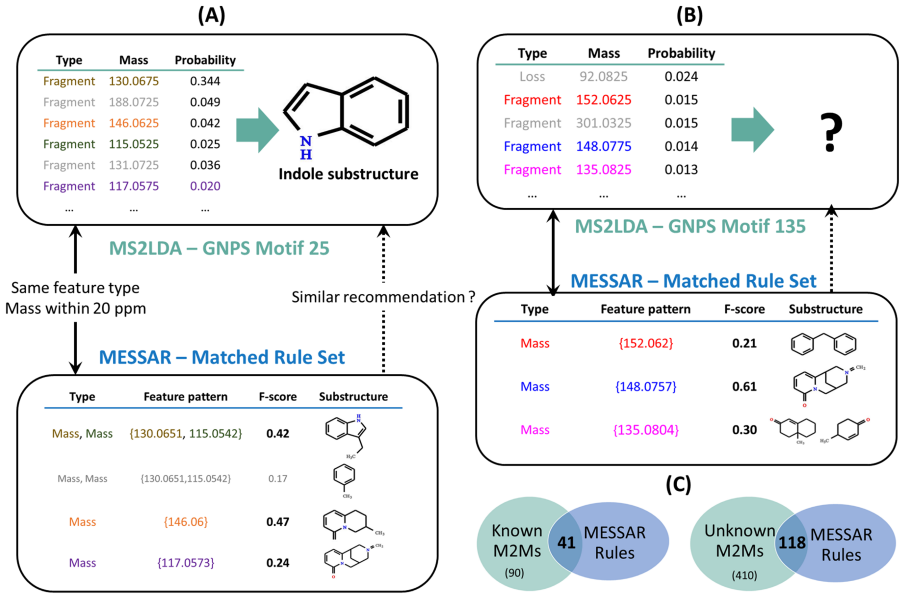

Fig 3. Example comparison between Mass2Motifs and MESSAR rules. A) The twenty most probable features of GNPS motif 25 (fragments or losses) were searched against the MESSAR rule database. A 20 ppm mass window was used and feature types were required to agree ("Fragment" = "Mass", "Loss" = "MDiff"). Only the most significant rule (with highest F-score) was reported if several MESSAR rules matched with a single motif feature. The MESSAR substructure recommendations for motif 25 were compared with the annotation by MS experts. B) GNPS motif 135 has not been annotated by experts yet. Here several potential substructures were assigned by searching in the MESSAR rule database. C) MESSAR rules matched with 41 annotated GNPS M2Ms and can explain 118 unknown M2Ms.

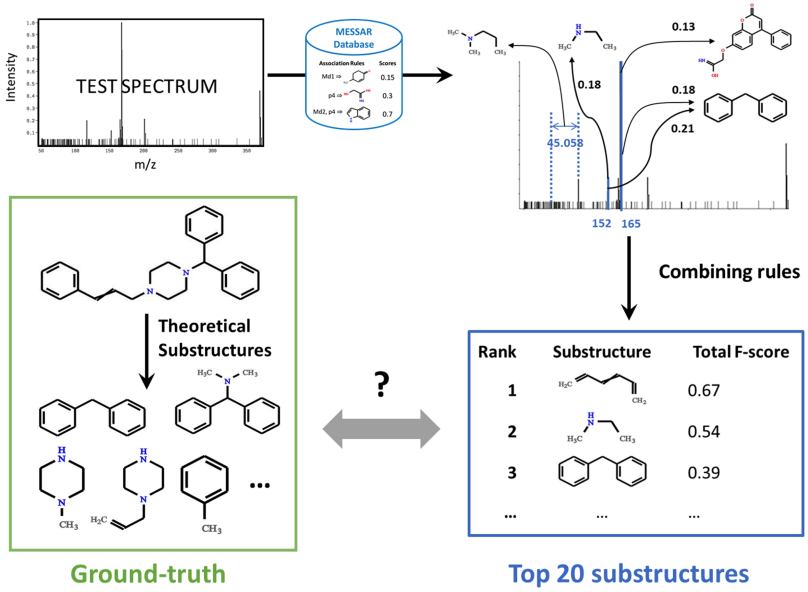

Fig 4. Substructure recommendation and validation for an "unknown" spectrum. Spectral features of the test spectrum are searched against the rule database. The matched rules are aggregated if several rules suggest the same substructure. The sum of F-scores is used to rank predicted substructures. Top 20 candidates are compared with actual substructures of the reference compound.

With the same testing parameters (20 ppm mass window, top 20 substructure candidates), substructure recommendation of MESSAR was validated on 3764 high-resolution positive ion mode MS/MS spectra. The performance was evaluated separately for 2759 spectra from "MASSBANK" and 1005 spectra from "Drug". Before evaluation, we noticed that theoretical substructures of many reference compounds were almost absent in the MESSAR database, which can be explained by the lack of training data or by a huge structural difference between training and test compounds. Since low sample size can lead to low statistical power, we present here only results from compounds for which MESSAR database covers at least 5 theoretical substructures (1455 and 520 spectra from "MASSBANK" and "DRUG"). For each spectrum we evaluated the statistical significance of the correctly assigned substructures using a one-sided Fisher's exact test. The performance of MESSAR on a test spectrum is validated if the *p-value* is smaller than 0.001, which indicates a significant amount of correct predictions. In this way, 454 test spectra were validated from 'MASSBANK" and 252 from "DRUG" (Fig 5). A better performance was clearly observed with the "DRUG" dataset. Such difference can be related to the quality of spectra or the type of compounds.

Usefulness of substructure recommendation tools

Similar to other substructure discovery tools the MESSAR output is a ranked list of substructure recommendations. In practice experts use the top-ranked candidates to facilitate a *de novo* identification or approximate characterization of unknowns. Therefore, ideally the top candidates of a good substructure prediction tool should contain rich, useful, and reliable structural knowledge. On this basis, we evaluated MESSAR, MS2LDA and CSI:FingerID substructure predictions on the "DRUG" datasets (details in Method section).

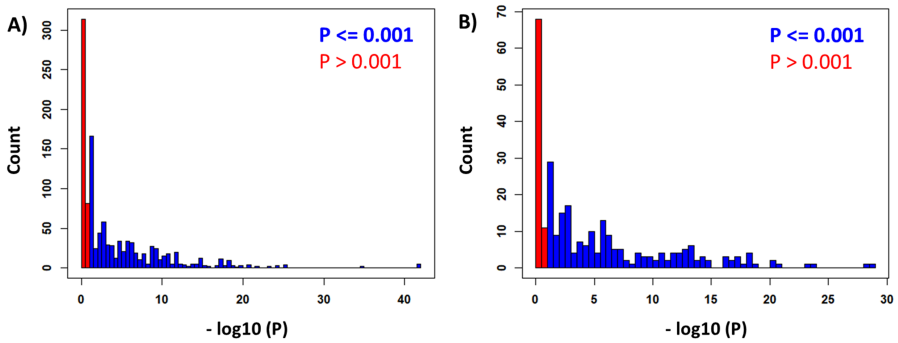

Fig 5. P-value distribution of the Fisher's exact test. We evaluated the statistical significance of the substructure prediction for reference compounds in "MASSBANK" (A) and "DRUG" (B). The blue color denotes test spectra for which a statistically significant number of correct substructures were predicted. The *p-value* and the confusion matrix obtained for each spectrum are provided as S2 File.

As shown in S3 Table, we provided top eight substructure candidates of MESSAR and CSI:FingerID (ranked by posterior probability), as well as MS2LDA motif annotations to a MS expert. Without knowing the ground-truth, the expert selected maximum three most informative substructures/annotations from each software. 'Informative" means that the candidate indicates an important chemical group or is likely a combination of smaller substructures. Such candidates should be useful for structure elucidation or biochemical interpretations. In contrast, common annotations such as "carboxylic group" or "acetyl loss" were not included. We presented here examples (S3 Table) of "unsolvable" unknowns, i.e., CSI:FingerID fails to correctly identify their true structure by searching molecular databases (based on its top 10 ranked structures).

Although it is not possible to completely solve the structures of these spectra, meaningful substructures were recommended by all three tools. Compared to the ground truth, MESSAR and/or CSI:FingerID were able to predict chlorobenzenes (Drug 119, 128, 130, 142 and 540) and indole substructures (drugs 25, 124, 130 and 324). MESSAR and MS2LDA could capture the steroid core for drugs 138, 167 and 201. By comparing, each software has its own advantages. CSI:FingerID predicts well small functional groups – it is the only tool that reveals the presence of sulfur-containing groups in drugs 119, 432 and 720. However, we tend to find similar candidates at the top of the list probably because of the usage of predefined, fixed molecular fingerprints by CSI:FingerID. MS2LDA provides more complex and diverse substructure annotations by assigning multiple motifs to the spectrum, although not all motif annotations correspond to the ground-truth. MESSAR recommends from simple to complex substructures and is useful for different types of compounds. Additionally, the combined interpretation of outputs can significantly aid in the substructure discovery for unknown metabolites. For instance, combining MESSAR and CSI:FingerID reveals the presence of both diphenyl substructure and sulfur group in drug 432.

Conclusions

MESSAR was inspired by the idea that the presence of spectral features in a MS/MS spectrum is linked to substructures of the metabolite. Such relations can be revealed through association rule mining in public spectral libraries. The rules obtained are used to predict the presence or absence of substructures for any unlabeled spectrum, providing

end users a ranked list of substructure recommendations. The MESSAR web tool can be used to assist *de novo* annotation of unknown metabolites, for example, to identify functional classes of unknown spectra that share substructures, to corroborate results from other chemical identification tools, to support NMR-based structure elucidation, etc.

An important advantage of MESSAR is that multiple substructures can be associated to a single mass spectrum, after which users can choose which recommendations to accept. MESSAR is inherently complementary to recently-published software MS2LDA as MS2LDA extracts co-occurring spectral features while MESSAR provides an automated structural annotation of these features. We envisage an improvement in substructure discovery by the combined use of expert-driven M2M annotations and data-driven MESSAR rules. Similarly, we anticipate that MESSAR will be useful for the functional analysis of complex biological matrices as it can quickly recognize substructure patterns.

Supporting information

S1 Text. Supporting method.

S1 Fig. Additional scatter plots between ARM parameters and F-score. A) Support. B) Confidence.

S2 Fig. Relation between F-score threshold and estimated FDR. The red vertical line indicates the F-score threshold estimated for 1% FDR.

S3 Fig. Examples of several MESSAR rules suggesting the same substructure. Four spectra from the DRUG and MASSBANK datasets were annotated by MESSAR. The mass fragments and mass differences (arrows) suggesting the recommended substructure are highlighted in red. Fragments not related to the substructure are in gray.

S1 Table. 41 structurally annotated positive ion mode M2Ms that matched with MESSAR rules.

S2 Table. Examples of unexplained positive ion mode M2Ms annotated by MESSAR rules.

S3 Table. Direct comparison of substructures recommended by MESSAR, CSI:FingerID, and MS2LDA.

S1 Data. MESSAR training data. GNPS spectral library and its decoy version in mgf format. The datasets are derived from `https://bio.informatik.uni-jena.de/passatutto`.

S2 Data. MESSAR testing data Two sets of labeled spectra for MESSAR testing

S1 File. MESSAR rule database

S2 File. Confusion matrices and p-values showing the performance of MESSAR on each test spectrum

Acknowledgments

Special thanks goes to Danh Bui-Thi and Charlie Beirnaert for insightful discussions. Many thanks to reviewers who provided invaluable suggestions to improve our manuscript.

References

1. Wishart DS. Emerging applications of metabolomics in drug discovery and precision medicine. Nature Reviews Drug Discovery. 2016;15(7):473–484.

2. Armitage EG, Barbas C. Metabolomics in cancer biomarker discovery: Current trends and future perspectives. Journal of Pharmaceutical and Biomedical Analysis. 2014;87:1 – 11.

3. Patti GJ, Yanes O, Siuzdak G. Innovation: Metabolomics: the apogee of the omics trilogy. Nature Reviews Molecular Cell Biology. 2012;13(4):263–269.

4. Nguyen DH, Nguyen CH, Mamitsuka H. Recent advances and prospects of computational methods for metabolite identification: a review with emphasis on machine learning approaches. Briefings in Bioinformatics;doi:10.1093/bib/bby066.

5. Kim S, Thiessen PA, Bolton E, Chen J, Fu G, Gindulyte A, et al. PubChem Substance and Compound databases. Nucleic Acids Res. 2016;44(D1):D1202–D1213.

6. Wolf S, Schmidt S, Müller-Hannemann M, Neumann S. In silico fragmentation for computer assisted identification of metabolite mass spectra. BMC Bioinformatics 2010. 2010;148:11.

7. Ridder L, van der Hooft JJJ, Verhoeven S, de Vos RC, van Schaik R, Vervoort J. Substructure-based annotation of high-resolution multistage MS(n) spectral trees. Rapid Communication in Mass Spectrometry. 2012;26(20):2461–2471.

8. Tsugawa H, Kind T, Nakabayashi R, Yukihira D, Tanaka W, Cajka T, et al. Hydrogen Rearrangement Rules: Computational MS/MS Fragmentation and Structure Elucidation Using MS-FINDER Software. Analytical Chemistry. 2016;88(16):7946–7958. doi:10.1021/acs.analchem.6b00770.

9. Allen F, Greiner R, Wishart D. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. Metabolomics. 2015;11(1):98–110.

10. Cereto-Massagué A, Ojeda MJ, Valls C, Mulero M, Garcia-Vallvé S, Pujadas G. Molecular fingerprint similarity search in virtual screening. Methods. 2015;71:58–63. doi:10.1016/j.ymeth.2014.08.005.

11. Dührkop K, Shen H, Meusel M, Rousu J, Böcker S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. Proceedings of the National Academy of Sciences of the United States of America. 2015;112(41):12580–12585.

12. Demarque DP, Crotti AEM, Vessecchi R, Lopes JLC, Lopes NP. Fragmentation reactions using electrospray ionization mass spectrometry: an important tool for the structural elucidation and characterization of synthetic and natural products. Natural Product Reports. 2016;33(3):432–455. doi:10.1039/c5np00073d.

13. Blaženović I, Kind T, Torbašinović H, Obrenović S, Mehta SS, Tsugawa H, et al. Comprehensive comparison of in silico MS/MS fragmentation tools of the CASMI contest: database boosting is needed to achieve 93% accuracy. Journal of Cheminformatics. 2017;9. doi:10.1186/s13321-017-0219-x.

14. Yang JY, Sanchez LM, Rath CM, Liu X, Boudreau PD, Bruns N, et al. Molecular Networking as a Dereplication Strategy. Journal of Natural Products. 2013;76(9):1686–1699.

15. Aguilar-Mogas A, Sales-Pardo M, Navarro M, Guimerá R, Yanes O. iMet: A Network-Based Computational Tool To Assist in the Annotation of Metabolites from Tandem Mass Spectra. Analytical Chemistry. 2017;86(6):3474–3482.

16. Mrzic A, Lermyte F, Vu TN, Valkenborg D, Laukens K. InSourcerer: a high-throughput method to search for unknown metabolite modifications by mass spectrometry. Rapid Communications in Mass Spectrometry. 2017;31(17):1396–1404.

17. Mahieu NG, Spalding JL, Gelman SJ, Patti GJ. Defining and Detecting Complex Peak Relationships in Mass Spectral Data: The Mz.unity Algorithm. Analytical Chemistry. 2016;88(18):9037–9046.

18. Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, et al. Sharing and community curation of mass spectrometry data with GNPS. Nature biotechnology. 2016;34(8):828–837. doi:10.1038/nbt.3597.

19. Ma Y, Kind T, Yang D, Leon C, Fiehn O. MS2Analyzer: A Software for Small Molecule Substructure Annotations from Accurate Tandem Mass Spectra. Analytical Chemistry. 2014;86(21):10724–10731.

20. van der Hooft JJJ, Wandy J, Barrett MP, Burgess KEV, Rogers S. Topic modeling for untargeted substructure exploration in metabolomics. Proceedings of the National Academy of Sciences of the United States of America. 2016;113(48):13738–13743.

21. van der Hooft JJJ, Wandy J, Young F, Padmanabhan S, Gerasimidis K, Burgess KEV, et al. Unsupervised discovery and comparison of structural families across multiple samples in untargeted metabolomics. Analytical Chemistry;in press.

22. Wandy J, Zhu Y, van der Hooft JJJ, Daly R, Barrett MP, Rogers S. Ms2lda.org: web-based topic modelling for substructure discovery in mass spectrometry. Bioinformatics;34(2):317–318.

23. Naulaerts S, P M, Bittremieux W, Vu TN, Vanden Berghe W, Goethals B, et al. A primer to frequent itemset mining for bioinformatics. Briefings in Bioinformatics. 2015;16(2):216–231.

24. Vu TN, Bittremieux W, Valkenborg D, Goethals B, Lemière F, Laukens K. Efficient Reduction of Candidate Matches in Peptide Spectrum Library Searching Using the Top k Most Intense Peaks. Journal of Proteome Research. 2014;13(9):4175–4183. doi:10.1021/pr401269z.

25. Vu TN, Mrzic A, Valkenborg D, Maes E, Lemière F, Goethals B, et al. Unravelling associations between unassigned mass spectrometry peaks with frequent itemset mining techniques. Proteome science. 2014;12(1):54.

26. Scheubert K, Hufsky F, Petras D, Wang M, Nothias LF, Dührkop K, et al. Significance estimation for large scale metabolomics annotations by spectral matching. Nature Communications. 2017;8(1):1494. doi:10.1038/s41467-017-01318-5.

27. Degen J, Wegscheid Gerlach C, Zaliani A, Rarey M. On the Art of Compiling and Using 'Drug-Like' Chemical Fragment Spaces. ChemMedChem. 2008;3(10):1503–1507. doi:10.1002/cmdc.200800178.

28. Brin S, Motwani R, Ullman JD, Tsur S. Dynamic Itemset Counting and Implication Rules for Market Basket Data. In: Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data. SIGMOD '97. New York, NY, USA: ACM; 1997. p. 255–264. Available from: `http://doi.acm.org/10.1145/253260.253325`.

29. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure. 1975;405(2):442–451. doi:10.1016/0005-2795(75)90109-9.

30. Käll L, Storey JD, MacCoss MJ, Noble WS. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. Journal of Proteome Research. 2008;7(1):29–34. doi:10.1021/pr700600n.

31. Bekkar M, Djemaa HK, Alitouche TA. Evaluation Measures for Models Assessment over Imbalanced Data Sets. Journal of Information Engineering and Applications. 2013;3(10):27–38–38.