Manuscript with Tracked Changes

Click here to download Manuscript with Tracked Changes Manuscript_TrackedChanges.docx

IMPROVED METABOLITE IDENTIFICATION WITH MIDAS AND MAGMA THROUGH MS/MS SPECTRAL DATASET-DRIVEN PARAMETER OPTIMIZATION

Dries Verdegem[1,2], Diether Lambrechts[3], Peter Carmeliet[2] & Bart Ghesquière[1]

(1) Metabolomics Expertise Center, Vesalius Research Center (VRC), VIB, Leuven, B-3000, Belgium; (2) Laboratory of Angiogenesis and Neurovascular link, Vesalius Research Center (VRC), VIB, Department of Oncology, University of Leuven, Leuven, B-3000, Belgium; (3) Laboratory for Translational Genetics, Vesalius Research Center (VRC), VIB, Department of Oncology, University of Leuven, Leuven, B-3000, Belgium

Corresponding author: B. Ghesquière, Ph.D.
Metabolomics Core Facility
Vesalius Research Center
VIB, KU Leuven, Campus Gasthuisberg O&N4
Herestraat 49 - 912,
B-3000, Leuven, Belgium
tel: 32-16-32.27.33; fax: 32-16-37.25.85

e-mail: bart.ghesquiere@vib-kuleuven.be

ABSTRACT

INTRODUCTION: LC-MS/MS based untargeted metabolomics is evoking high interests in the metabolomics and broader biology community for its potential to uncover the contribution of unanticipated metabolic pathways to phenotypic observations. The major challenge for this methodology is making the computational metabolite identification as reliable as possible in order to reduce subsequent target candidate validation to a minimum. Metabolite library matching techniques based on precise masses and fragment mass patterns have become the *de facto* method in the field. However, in the literature the original methods are often under-validated, making it complicated to judge their intrinsic value.

OBJECTIVES: We aimed to demonstrate that large MS/MS metabolite spectral libraries can be used not only to validate and compare, but also to improve the methods.

METHODS: Several computational tools for metabolite identification (MAGMa, CFM-ID, MetFrag, MIDAS) were applied on a large MS/MS dataset derived from Metlin. Their performance was first compared and for the two best-performing tools (MAGMa and MIDAS), the performance was then improved by applying a parameter fine-tuning procedure.

RESULTS: We confirmed MIDAS and MAGMa as the state-of-the-art freely available tools for metabolite identification. Moreover, we were able to identify optimized working parameters, engendering an improvement in their performance. For MAGMa, dynamic, metabolite-dependent optimized parameters were obtained using machine learning techniques.

CONCLUSION: We were able to achieve an incremental increase in the identification accuracy of MIDAS and MAGMa. A wrapper script (MAGMa+) capable of calling MAGMa with tailored parameters is made available for download.

Keywords

Untargeted metabolomics, metabolite identification, MAGMa, method comparison, method optimization, machine learning

Introduction

Metabolomics is the study of the overall metabolic footprint of cells, tissues or body fluids. It involves the characterization of small molecules (typically below 2,000 Da) and their abundances. The technology of choice in metabolomics is often mass spectrometry coupled to chromatography (GC-MS, GCxGC-MS, CE-MS, LC-MS, LC-MS/MS, LC-MS^n) (Dunn et al. 2013). However, for unbiased (untargeted) metabolic profiling where no metabolite pre-selection is made and global metabolomics characterization and identification is desired, liquid chromatography – multistage mass spectrometry setups (LC-MS/MS and LC-MS^n) are of particular interest. On the chromatography side, although liquid chromatography (LC) and its relatively lower retention reproducibility impedes the use of retention times for metabolite identification, it is also characterized by straightforward sample processing and high sensitivity, which works in its favor in an untargeted setup. On the mass spectrometry side, multistage mass spectrometry (MS/MS and MS^n) provides besides the mass of the molecular parent ion also the masses of molecular substructures obtained after one or more steps of fragmentation. This multiplies the amount of information contained in the spectra as the set of masses of both fragments and parent ion is much more unique to a given molecule than its parent mass alone. When in addition high-resolution MS (HRMS) is applied, metabolite identification is further enhanced, thereby narrowing down the number of possible metabolite matches.

Not surprisingly, the most promising computational methods developed hitherto for metabolite identification (linking back mass spectrometry information to metabolite identity) mainly rely on tandem LC-MS/MS and, to a lesser extent, multi-stage LC-MS^n data (Vaniya and Fiehn 2015). However, reliable identification remains problematic and is typically considered as the major bottleneck for untargeted metabolomics (Dunn et al. 2013; Neumann and Bocker 2010). Compared to the field of proteomics where a multitude of well-established software tools is available for the identification of proteins and peptides (Haga and Wu 2014), the metabolomics field is seriously lagging behind. This is mainly due to the much larger structural variability in metabolites as compared to proteins and peptides. In recent years, a number of computational approaches have been suggested to solve the metabolite identification problem (Allen et al. 2014; Duhrkop et al. 2015; Heinonen et al. 2012; Ridder et al. 2012; Wang et al. 2014; Wolf et al. 2010), but none has become an established part of untargeted metabolomics pipelines.

One major approach that has gained remarkable momentum is to search for putative candidates in molecular structure databases using algorithms that match the spectral information to structure information. These algorithms have previously been classified in distinct categories (Hufsky et al. 2014), of which the “combinatorial fragmentation” category is particularly attractive because of the common applicability on ESI-MS/MS data and free availability of the corresponding software tools. In combinatorial fragmentation, it is attempted to explain the measured fragmentation peaks by molecular substructures generated *in silico* in a combinatorial manner. Examples of freely available tools based on this algorithmic solution are MetFrag (Wolf et al. 2010), Metabolite Identification via Database Searching (MIDAS) (Wang et al. 2014) and Mass Annotation based on in silico Generated Metabolites (MAGMa) (Ridder et al. 2012).

The original approach to metabolite identification, predating the computational strategies, consists of spectral matching between experimental and metabolite standard-based library MS/MS spectra (Degtyarenko et al. 2008; Horai et al. 2010; Smith et al. 2005; Tautenhahn et al. 2012; Wishart et al. 2013; Wishart et al. 2009). Although intrinsically powerful and still most commonly used, this approach inevitably suffers from the instrument and sample dependence of the spectra (Neumann and Bocker 2010) and the relatively small size of these libraries compared to the molecular structure databases. A different track to metabolite identification is the data-driven approach, in which existing MS/MS spectral libraries are not applied directly for metabolite identification, but indirectly as benchmark or training data holders to build or improve the purely computational methods mentioned above.

FingerID (Heinonen et al. 2012) was the first tool employing such a new paradigm. A support vector machine (SVM) model was trained using MS/MS spectra obtained from MassBank (Horai et al. 2010) for the prediction of molecular features, which are then used in a second step to match molecular structures. Another tool providing an adequate example of the power of the data-driven strategy is CFM-ID (Competitive Fragmentation Modeling based IDentification) (Allen et al. 2014). Using the MS/MS spectra of ±3,500 metabolites randomly selected from the Metlin database (Smith et al. 2005; Tautenhahn et al. 2012) as training data for machine learning based models, the authors built software capable of predicting the MS/MS spectra corresponding to metabolites and of the subsequent metabolite identification through spectral matching (Allen et al. 2014). In the latest Critical Assessment of Small Molecule Identification contest (CASMI 2015), created to follow up and stimulate the research on computational untargeted metabolomics assignment methods, this approach immediately provided CFM-ID a top ranking position.

Recently, the new machine learning based tool CSI:FingerID (Duhrkop et al. 2015) was described. Inspired by the original FingerID, its functioning is based on a comparison of molecular fingerprints predicted by the software for the experimental MS/MS data on the one hand and calculated from potential structure matches on the other hand. The authors observed a considerable performance improvement, however the method is currently not freely available for batch processing, preventing independent evaluation.

Here, we propose a different form of data-driven approach to method development, i.e. data-driven improvement of previously mentioned combinatorial fragmentation tools. MAGMa for instance is a rule-based software using a fairly simple yet highly parameterized formula to calculate a matching score between computationally generated molecule fragments of putative metabolite matches and MS^n peaks. Moreover, it also consistently obtains top rankings in the CASMI contest. Its strength lies in it being “model-free”, i.e. its complete independence of experimental setup since the matching score is not biased towards specific instrument-characteristic peak intensities. We investigated whether the performance of MAGMa (and another tool, MIDAS) can be further improved by optimizing its parameters using MS/MS data.

In all, this study aimed to serve two purposes. First, despite obviously useful efforts such as CASMI, researchers are prevented from straightforwardly choosing a preferred identification tool for their untargeted metabolomics pipeline because of the often limited method validation in the original papers of these tools. Moreover, even in CASMI, success is, besides by the intrinsic power of the identification algorithm itself, to a large extent determined by database pre-filtering based on provided meta-data, thereby hampering true validation. We therefore used MS/MS spectra obtained from Metlin to perform an extensive performance comparison on four different freely available computational untargeted metabolomics tools (MetFrag, MIDAS, CFM-ID and MAGMa) in identical conditions.

Second, more importantly, we show that parameterized combinatorial fragmentation tools can be generated that benefit not only from the intrinsic advantage of being model-free, but in addition also from domain knowledge as it is found in the spectral datasets, thereby taking profit of the best of two worlds. Indeed, by maximizing for average identification success rates when applied to large spectral datasets, a data-driven optimization (fine-tuning) of their working parameters and subsequent performance improvement becomes conceivable. To achieve this improvement, we used the MS/MS data obtained from Metlin as training set to determine the optimum points in the parameter spaces of the two tools, MIDAS and MAGMa, which are characterized by (1) the best performances in our method comparison and simultaneously have (2) the highest levels of parameterization. Since we identified for MAGMa multiple complementary parameter sets, each suitable for a subset of metabolites, we developed a random-forest (RF) classifier capable of detecting which parameters to apply to which identification problem.

Materials and Methods

Untargeted metabolomics identification tools

Four software tools were considered in this study: MAGMa, CFM-ID, MetFrag and MIDAS. FingerID and CSI:FingerID have not been included in our study. FingerID is available as a methodology only, not as a black box tool, as the user is still required to build the predictive models. Although this provides flexibility, it also limits the generalizability of the FingerID performance obtained in any particular test setup. CSI:FingerID in contrast is available as a blackbox tool, however, as mentioned, it does not support batch processing, which is required for large-scale evaluations.

Two of the four selected tools, CFM-ID and MetFrag were applied as such, without modifying the source code and no attempt was made to optimize their performance. For CFM-ID, the trained models params_metab_se_cfm and negative_se_params were chosen for negative and positive mode identifications, respectively. The param_config.txt files were left unchanged while as param_output.log files we chose param_output0.log in both ionization modes. MetFrag required no particular initial setup.

MAGMa uses a formula that determines the cost of matching a fragmentation tree with a molecular structure (equations 1 & 2). Smaller final costs, which are in essence a sum of weighted penalties dealing with both matched and unmatched peaks, imply better matches. The penalties are determined by parameter values (table 1) while the weights coupled to each penalty are determined by corresponding peak intensities. In the original paper, all seven parameter values are fixed based on educated assumptions, which however leaves room for the (data-driven) optimization reported here.

$$Substructure\ score\ (LSS) = \sum_{\substack{b\,\in\,disconn.\\ bonds}} h_b p_b \tag{1}$$

$$Candidate\ score = \frac{1}{\sum_p \sqrt{I_p}} \left(\sum_{\substack{p\,\in\,matched\\ peaks}} \sqrt{I_p} \times \min(LSS_p, MSP) + \sum_{\substack{p\,\in\,missed\\ peaks}} \sqrt{I_p} \times MSP \right) \tag{2}$$

MIDAS also has a rule-based approach to solve the identification problem. Instead of relying on penalties, it applies rewards to calculate a matching score between fragmentation spectrum and molecular structures (equations 3, 4 & 5), resulting in higher scores for better matches.

$$Candidate\ score = \sum_{j=1}^{n} \max_{1 \leq i \leq m} \left(S(F_j) \cdot I_i \cdot \lambda_{ij} \right) \tag{3}$$

$$S(F) = \begin{cases} 0.5^b S(P)\ for\ detected\ precursor\ (P) \\ 0.1^b S(P)\ for\ undetected\ precursor\ (P) \end{cases} \tag{4}$$

$$\lambda = \begin{cases} 2\ (1 - pnorm(\varepsilon, 0, \Delta/2))\ for\ matched\ fragments\ and\ peaks \\ \qquad 0\ for\ unmatched\ fragments\ and\ peaks \end{cases} \tag{5}$$

The summation happens over all n generated fragments and considers all m peaks. For each *in silico* fragment, the score increases if the mass match (λ, a function with Gaussian behavior defined over mass distance), the matching peak intensity (I) and the plausibility of generation of the fragment (S) increase. The plausibility score formula (equation 4) contains three optimizable parameters: the values of 0.5 and 0.1 (for ease of reference these will hereafter be referred to as dpp (detected precursor parameter) and upp (undetected precursor parameter)) and b (1 in case of linear bond break, 2 in case of ring pair bond break).

MS/MS SPECTRAL DATASET

As MS/MS spectral dataset for the evaluation and optimization procedure, we used a selection of the metabolic fragmentation data of the Metlin (Smith et al. 2005; Tautenhahn et al. 2012) database, obtained on a 6510 Q-TOF (Agilent Technologies). At the time of retrieval (November 2014), Metlin contained the MS/MS spectra of 11,736 metabolites. For 490 of those, the corresponding molecular structure could not be straightforwardly obtained and were therefore removed from the dataset. Furthermore, Metlin contains a large proportion of peptides (dipeptides, tripeptides and tetrapeptides). 5,692 of the remaining 11,246 metabolites with MS/MS spectra correspond to these peptides and were equally removed from our selection to prevent the results from being biased towards this particular metabolite class. We did retain the amino acids in the dataset. Hence, the final selection database contained ESI-MS/MS spectra of 5,554 metabolites recorded either in positive or negative mode or both.

To ascertain the diversity the remaining dataset, we linked as many metabolites as possible to the Human Metabolome Database (HMDB) (Wishart et al. 2013) using an in-house metabolite ID converter. The HMDB metabolite classification indicated that the 2,200 metabolites, which could in this way be linked to a HMDB ID, belong to as many as 25 different super classes (supplementary Fig. 1), 198 different classes and 519 different parent types. We also checked the mass distribution of the selected metabolites, which shows a representative distribution (supplementary Fig 2.). These numbers demonstrate the broad coverage of the selection database.

Metlin spectra are typically recorded at different collision energies (0V, 10V, 20V and 40V). Therefore, and partially in line with earlier evaluation setups (Wang et al. 2014; Wolf et al. 2010), the collected data were used to generate four distinct data subsets: positive_composite, positive_select, negative_composite and negative_select. “Positive” and “negative” refer to the ionization mode. “Composite” implies that the spectra at all four collision energies were merged into one spectrum. If the same peak (within 10 ppm or 0.002 Da distance) appeared at different collision energies, only the highest intensity one was retained. “Select” refers to the single lowest collision energy spectrum for which at least 75% of the molecule is fragmented. If for any fragmentation spectrum thus obtained, almost the entirety (95%) of the total intensity was divided over only one or two peaks, they were excluded from the dataset. The resulting (generated) dataset contained 4,736 positive_composite, 4,376 positive_select, 3,480 negative_composite and 2,950 negative_select spectra.

MOLECULAR STRUCTURE DATABASE

We compiled a molecular structure database using structures collected from several available metabolite databases (Kegg, HMDB, LMID, ChEBI, Metlin) or general small molecule databases (PubChem). A total of 67,780 metabolites were included in our database. It was assured that the MS/MS spectral dataset metabolites were completely covered in the molecular structure database such that the identification software was in principle always capable of assigning the correct molecule. Several variants of the database were generated, each compatible with one of the identification tools. All database variants contained for each metabolite at least the InChI code and one identifier, in accordance with the MS/MS spectral dataset, which allows making the link between both when evaluating the results.

EVALUATION PROCEDURE

The four metabolites identification software packages were compared using identical, relatively wide window settings (relative m/z precision = 20.0 ppm, absolute m/z precision = 0.005 Da). These windows were chosen so as to increase the probability that the parent-mass-based metabolite pre-selection from the structure database contained the correct metabolite. Simultaneously, the larger number of candidate metabolites request more selective power from the scoring algorithms, which can therefore be more thoroughly tested. All described software packages return a list of putative metabolites, each with a score reflecting the plausibility of kinship between metabolite and spectrum. As evaluation criterion we used the total number of correctly identified MS/MS spectra, i.e. spectra for which the software provides the best score for the correct metabolite. Stereochemical differences were ignored upon metabolite comparison.

Although some of the identification tools are capable of dealing with variable adducts and of determining the correct neutral molecule mass if adduct information is passed to it, most do not. For conformity, we did not provide the precursor masses as they occur in the spectral dataset. Instead, since for all MS/MS spectra, the corresponding metabolite is known, we provided the calculated $[M+H]^+$ or $[M-H]^-$ ion mass as precursor mass to the tools, together with the setting that the H-adduct was to be expected. Any observed fragment peaks corresponding to an m/z value larger than $mz_p - 1.5 \times \mathrm{mass_H}$ (where mz_p is the precursor mass, calculated as described above and $mass_H$ is the hydrogen mass), were excluded from the input fragmentation spectra.

For both CFM-ID and MetFrag we made sure that, for each identification request, a minimal subset of the entire molecular structure database was presented to these tools. These subsets contained only structures with potentially matching parent masses (determined using the window settings indicated above) and resulted in much improved run times. For MIDAS and MAGMa, no such procedure was required.

OPTIMIZATION PROCEDURE

The windows and adduct settings of the evaluation procedure were also applied during the optimization of MAGMa and MIDAS. Here, the larger mass window settings are again useful since the larger resulting molecule candidate set makes a tougher and more selective optimization environment and the tools therefore increasingly benefit from better parameter settings.

The MAGMa pars.py source code file contains the different parameter settings, which we adapted during the optimization procedure (see table 1 for the list of parameters allowed to vary). Since the values of the different parameters only make sense relative to each other, one parameter, the single bond break penalty, was kept fixed during the optimization. The triple bond break penalty was not optimized since very little known metabolites have triple bonds and consequently, this parameter was expected to have little influence on the overall performance in our setup. The h-parameters (bond strengths of carbon-carbon bonds versus bonds involving non-carbon atoms are interesting optimization subjects but were not optimized here for combinatorial explosion reasons. The parameters that were optimized are: double bond break cost (DBBC), aromatic bond break cost (ABBC) and missing substructure penalty

(MSP). This latter corresponds to the penalty given when an observed fragmentation peak cannot be explained by any of the generated substructures.

The parameter settings of MIDAS are provided in the weightedscore.py source code file. The b parameter corresponds to the C_self variable in the function TreeLikeBreakBondsDepthFirst, while the dpp and upp variables (0.1 and 0.5) are defined in the function ExactBondsInfo. Again, since only relative changes to the parameter values make sense, b for single bond breaks was left untouched during optimization.

To search for global optimum parameter values we performed a straightforward systematic exploration of parameter combinations, where one parameter was allowed to vary while the others were kept constant. Although metaheuristic search procedures exist to speed up optimization problems, we did not employ them in our study. The highly intractable nature of the optimization problem results in unpredictable space characteristics, while in fact some knowledge of the optimization space is required to prevent such metaheuristic searches from getting stuck in local minima. However, for time reasons, our manual search procedure was performed akin to a simulated annealing strategy, with more interesting regions being defined in greater detail.

As evaluation criterion for the parameter settings during optimization, the number of correct first position assignments was used. If the identification tool returned besides the correct metabolite, other metabolites with the same score, this was not considered a correct assignment. Instead, the worst-case ranking (determined by the size of the pool of metabolites with identical scores) was given to the correct metabolite match. Stereochemistry was again ignored.

Since organic covalent bonds can be expected to have different tendencies to break in positive and negative ionization mode, separate optimal parameters were searched for the both ionization modes, both in the case of MAGMa and MIDAS. To obtain the optimal positive (resp. negative) mode parameters, the results on the positive_composite and positive_select (resp. negative_composite and negative_select) datasets were pooled. In doing so, the optimized parameters were ensured to be compatible with the presence of both large and small numbers of fragmentation peaks, again making them less machine and/or setup dependent.

DYNAMIC PARAMETER SELECTION

For MAGMa, differential and complementary performance was observed for different parameter settings. Indeed, even though the optimized parameters resulted in the best overall performance, different parameter settings were observed to provide correct identifications for some spectra on which the optimized parameters failed. To exploit this phenomenon, we searched for a series of additional parameter combinations according to:

$$\begin{aligned}&\{dbbc_a^{max}, abbc_a^{max}, msp_a^{max}\}\\&= \arg\max(a \times |f_{MAGMa}(dbbc, abbc, msp) - f_{MAGMa}(dbbc_0^{max}, abbc_0^{max}, msp_0^{max})| + (1 - a)\\&\times |f_{MAGMa}(dbbc, abbc, msp)|, \{dbbc, abbc, msp\})\end{aligned} \tag{6}$$

In this equation, $\{dbbc, abbc, msp\}$ stands for the set of double bond break cost, aromatic bond break cost and missing substructure penalty respectively, f_{MAGMa} describes the set of correct identifications obtained by MAGMa for a specific set of parameters while a can take any value in the interval $[0,1]$. The set $\{dbbc_0^{max}, abbc_0^{max} mpp_0^{max}\}$ are the original optimized parameters obtained using:

$$\{dbbc_0^{max}, abbc_0^{max} msp_0^{max}\} = \arg\max(|f_{MAGMa}(dbbc, abbc, msp)|, \{dbbc, abbc, msp\}) \tag{7}$$

The 0-index refers to $a = 0$ as equation (6) effectively reduces to equation (7) for the smallest possible value of a. Hence, in going from $a = 0$ to $a = 1$, one gradually goes from a parameter optimization with a focus on the total number of correctly assigned MS/MS spectra, towards an optimization focusing on the amount of differentially correctly assigned spectra (as compared to the $a = 0$ optimized parameter set). The described series was determined for $a \in \{0.1, 0.2, \dots, 0.9, 1.0\}$, with the intention of retaining one second set of MAGMa parameters which is maximally compliant with the first set ($a = 0$) in a manner described hereafter. We observed however that some subsequent a-values resulted in the same parameter set values, reducing the number of singular parameter sets of interest (to 6 instead of 10). The described procedure is again performed separately for both positive and negative ionization mode.

If two separate parameter sets are considered, each capable of correctly assigning a subset of MS/MS spectra where the other fails, one can attempt to determine whether there is a molecular basis underlying this performance difference. We therefore selected the metabolites that were uniquely correctly identified using the singular parameter sets obtained with $a = 0$ and each $a \in \{0.1, 0.2, \dots, 0.9, 1.0\}$ respectively, to end up with a series of pairs of distinct metabolite sets. For each two sets, we trained a two-class random forest classifier, potentially able to predict the membership of a random metabolite to any of such two groups (if a molecular structure discrepancy does exist) using scikit-learn (Pedregosa et al. 2011). Initially, a total of 5,809 molecular fingerprints were used as classification features, comprising all CDK (Steinbeck et al. 2003), FP3 (as defined in Open Babel (O'Boyle et al. 2011)), Klekota-Roth (Klekota and Roth 2008), MACCS (Durant et al. 2002) (as defined in RDKit; www.rdkit.org) and CACTVS (Ihlenfeldt et al. 2002) (as defined in PubChem; https://pubchem.ncbi.nlm.nih.gov/) fingerprints available in SMARTS (Smiles

Arbitrary Target Specification) format. These fingerprints encode for predefined functional and sub-structural groups. The calculation of their absence/presence in the metabolites was done using an adapted version of RDKit.

The number of features was subsequently reduced to a minimal set of significant fingerprints using recursive feature elimination with cross-validation (RFECV) (Guyon et al. 2002). This procedure was performed in a 10-fold cross-validation setup and in each recursive step the ten features with the lowest importance were removed. The number of estimators for all random forest classifiers during the RFECV step was set to 10,000. The two final minimal sets (one for each ionization mode) were obtained by merging the RFECV-obtained feature sets of each unique $a = 0$ versus $a \in \{0.1, 0.2, \ldots, 0.9, 1.0\}$ opposition. The reduced feature sets contained 376 and 689 features for positive and negative ionization mode, respectively. The final, production-level classifiers, trained on the same datasets but applying the minimal instead of the full feature sets and used to get a sense of the degree of molecular difference between sets of molecules also contained 10,000 estimators. The feature elimination was indeed observed to result in improved classification accuracies.

To determine whether, and if so which, performance improvements exist upon combined use of a second set of MAGMa parameters ($a \in \{0.1, 0.2, \ldots, 0.9, 1.0\}$), maximally compliant with the first set ($a = 0$), two independent types of performance profiles were established. First, a theoretical performance profile was calculated using the MAGMa output when applied to our Metlin-derived spectral dataset and the binary classification accuracy obtained for each of the a-values through 10-fold cross validation as follows:

$$\begin{aligned}
&\mathrm{P}(a) \\
&= |f_{MAGMa}(dbbc_0^{max}, abbc_0^{max}, msp_0^{max}) \cap f_{MAGMa}(dbbc_a^{max}, abbc_a^{max}, msp_a^{max})| \\
&+ accuracy^{RF}(a) \\
&\times (|f_{MAGMa}(dbbc_0^{max}, abbc_0^{max}, msp_0^{max}) - f_{MAGMa}(dbbc_a^{max}, abbc_a^{max}, msp_a^{max})| \\
&+ |f_{MAGMa}(dbbc_a^{max}, abbc_a^{max}, msp_a^{max}) - f_{MAGMa}(dbbc_0^{max}, abbc_0^{max}, msp_0^{max})|)
\end{aligned} \tag{8}$$

By applying a theoretical profile employing cross-validation accuracies, we were allowed to draw conclusions from the training/discovery (Metlin) set, while still avoiding undesired overfitting pitfalls. Second, an empirical performance profile was determined for two unseen validation MS/MS datasets (see below) as a function of the a-values. Therefore, for each unseen spectrum, the metabolite identification was performed using both parameter settings in each unique combination ($a = 0$ vs. $a \in \{0.1, 0.2, \ldots, 0.9, 1.0\}$) and the molecular structure class membership was predicted for the top three identifications of both parameter settings (more than three if multiple putative metabolites with identical scores need to be taken into account and less than three if less putative metabolites are returned by MAGMa). From these (more or less) six predictions, a consensus class membership was determined and the parameter set corresponding to the consensus score was used for identification. In case of a tied score, preference was given to the $a = 0$ parameter set. Finally, the ideal combination of parameter sets is extracted from both the theoretical and empirical performance profiles.

VALIDATION

The parameter improvements and dynamic parameter selection proposed for MAGMa were validated on two independent MS/MS dataset, generated by using the MoNA (MassBank of North America) repository. The two datasets ("MoNA TOF" and "MoNA FT") were produced by filtering out the MS/MS spectra obtained on LC-ESI-ITTOF and LC-ESI-QTOF (MoNA TOF) and LC-ESI-ITFT and LC-ESI-QFT (MoNA FT) instrument types, respectively. Of both dataset, only MS/MS spectra where 95% of the total intensity is divided over at least three peaks are kept for identification (see above). Also, spectra of which the corresponding metabolite is not present in our molecular structure database (see above) were discarded. In case of duplicate spectra (same metabolite recorded in same condition), only one spectrum (with the number of MS/MS closest to 15; a reasonable number for metabolite identification) was kept for the final dataset. The resulting MoNA TOF dataset contained 308 positive mode spectra and 148 negative mode spectra, whereas the MoNA FT.dataset contained 513 positive mode spectra and 176 negative mode spectra.

MAGMA+

We developed a python wrapper script with identical command line arguments as the original MAGMa program. The wrapper script calls MAGMa with two complementary parameter sets. Subsequently, only one of both resulting rankings is retained based on the class memberships of the top three ranked metabolites for both parameter set as explained above. A schematic representation of the MAGMa+ workflow and its utility in an untargeted metabolomics pipeline can be found in Fig 1. The script, together with the positive and negative mode feature files and classifiers it requires is available for download at: https://github.com/savantas/MAGMa-plus. It can easily be incorporated in existing untargeted metabolomics pipelines using MAGMa.

Results and Discussion

Comparison of identification methods

For our evaluation setup, we determined for each tool the amount of correct identifications (the correct metabolite is ranked first). We identified MAGMa (57.9% correct identifications) as the best performing identification tool, closely followed by MIDAS (57.8%). MetFrag (53.4%) and CFM-ID (45.3%) provided less correct identifications (Fig. 2a). Several findings in the evaluations are noteworthy. Of all tools evaluated, MAGMa is unique in dealing with true multistage (MS^n, with n>2) fragmentation data. Even though one might have expected an underperformance when restricting the input to MS2 data, we nonetheless observed that even in situations of only a single level of MS fragmentation, it outperformed the other methods. However, none of the other tools was completely redundant when compared to the MAGMa performance. Indeed, Fig. 2b shows the distribution of correct assignments by the different tools, thereby giving insight in how different a tool's performance is compared to the others. We observed that every tool was capable of correctly assigning a unique set of hundreds of spectra, whereas the other tools failed on this set. A similar behavior was observed recently (Duhrkop et al. 2015).

Our evaluation identified a suboptimal performance for CFM-ID. This was surprising, especially when considering its top ranking in CASMI. There might be several testbed-related explanations for this observation. First, CFM-ID is currently not capable of processing charged metabolites, resulting in a number of mismatches in our validation setup. Second, CFM-ID is based on spectral comparison, which potentially works well if the *in silico* spectra it generates are accurate and discriminative. This requirement can be obtained through broad coverage during the machine learning training phase. Indeed, larger training sets generally yield models with stronger generalization performance. CFM-ID was trained on a dataset of 3,476 metabolites from Metlin, while in our validation setup, a total of 5,554 Metlin metabolites were evaluated. Hence, in the present study, CFM-ID inevitably deals with such generalization issues whereas for the much smaller CASMI challenge, these issues might have been less pronounced. Third, for its participation to CASMI, CFM-ID was allowed to search in several databases, which were expected to contain the contest metabolites. A consensus ranking was subsequently defined based on the combined identification results (www.casmi-contest.org/2014/results-cat2.shtml). Hence, the CFM-ID CASMI achievement was likely in part attributable to additional filtering, besides to the algorithmic power of spectral prediction itself.

Furthermore, CFM-ID and MAGMa might be complementary with regard to the structure database size. CFM-ID has a performance comparable to MAGMa when using the large PubChem database (over 50 million compounds)

(Duhrkop et al. 2015). However, as the applied structure database decreases in size to a more reasonable number of 300,000 compounds, MAGMa starts outperforming CFM-ID (Duhrkop et al. 2015). Because in our study the molecular structure database was still smaller (67,780 metabolites), MAGMa performed visibly better than CFM-ID. Molecular structure databases should in theory be as big and complete as possible. It is indeed being increasingly suggested that databases such as KEGG and HMDB contain only a small fraction of all possible human metabolites. Metabolites typically included are either highly abundant in cells or involved in well-studied pathways, whereas those resulting from undocumented side-reactions or enzymatic mutations typically linked to disease tend to be lacking (Jeffryes et al. 2015). Yet, in practice, searching in larger, “all-covering” databases has an adverse effect on the identification quality, up to the point where the current software typically has correct metabolite identification rates of only 10-20% (Duhrkop et al. 2015). As computational metabolite identification becomes useful only at much higher correct identification rates, database size reduction through pre-filtering based on the biological context should probably still be the standard procedure in untargeted metabolomics setups in order to boost the performance. In this context, it seems worthwhile to perform tool evaluations with smaller structure databases, as is the case here.

Unfortunately, we were not able to include CSI:FingerID (Duhrkop et al. 2015) (a recently developed machine learning based tool) for independent comparison in our study as it does not exist as a stand-alone tool or cannot currently be accessed through an application programming interface (API). Its reported performance improvement compared to CFM-ID is likely due to the fact that it does not compare spectra but molecular fingerprints deduced from both the MS/MS spectrum to be identified and from the putative metabolite structures, since such fingerprints can be expected to be less dependent on experimental setup (Duhrkop et al. 2015).

Another noteworthy observation is the relatively good performance of MIDAS as compared to its performance in an earlier report (Duhrkop et al. 2015). The precise reason for this discrepancy is not entirely clear, although the difference in structure database sizes between both studies is a possible explanation. A further possibility is that the MIDAS performance is relatively more sensitive to the instrumental setup and the richness of the MS/MS spectra. Indeed, although MAGMa outcompetes MIDAS overall in our evaluation setup, MIDAS is superior when only considering the "composite" spectra (see above), in which more MS/MS peaks are present (Supplementary Table1). In this respect, it is also important to notice that the observations reported here were obtained using a dataset consisting of QTOF data only. We therefore also applied our tool comparison procedure on the MoNA FT dataset introduced above, even though it is smaller in size and therefore more difficult to draw conclusion from. Here too, MIDAS starts outperforming MAGMa (Fig. 2c), which might be, besides a platform effect, attributable to the rather consistent high abundance of MS/MS peaks in this latter dataset. Further investigation is clearly needed to completely apprehend the behavior of different tools in different setups. However, in all our comparisons, MAGMa and MIDAS invariably outperform the other considered metabolite identification tools.

MIDAS OPTIMIZATION

Interestingly, the two best-performing tools of our evaluation study are those characterized by high levels of parameterization. We therefore performed a parameter optimization using our Metlin derived datasets for positive and negative ionization mode separately. As explained in the Methods, MIDAS has three optimizable parameters that account for differential plausibility of fragment structures (*dpp* and *upp*) and differential bond vulnerability (*b*). In the original MIDAS implementation, the three parameters *dpp*, *upp* and *b* take the values of 0.5, 0.1 and 2.0 respectively. After having evaluated 181 different parameter combinations, we obtained (0.4, 0.1, 1.5) and (0.9, 0.1, 1.5) as improved parameters for positive and negative ionization mode respectively (Fig. 3).

The shift of the *b* parameter from 2.0 to 1.5 for both positive and negative mode in the optimized situation indicates that the cost of breaking a pair of ring bonds to cleave open a ring structure is actually smaller than twice the cost of breaking a single bond, which was originally anticipated. The break-up of the *dpp* parameter into two separate values for positive (0.4) and negative (0.9) ionization mode, suggests that for positive mode spectra, it is less exceptional to observe a specific fragment in the MS/MS spectrum without also observing its supposed immediate precursor than it is for negative mode spectra. These optimized parameters resulted in a performance increase from 57.8% to 58.6% of correct assignments in the Metlin dataset, thereby also surpassing the original MAGMa performance (Fig. 6). Rsearchers willing to use MIDAS with the improved parameters are thus encouraged to update their values in the corresponding source code file (see Materials and Methods).

MAGMA OPTIMIZATION

Inspired by the potential of the procedure, we performed a similar parameter optimization for MAGMa. To obtain a confident representation of the optimization space, we evaluated over 23,450 different parameter settings. The parameters for single, double, triple and aromatic bond break cost and unexplained MSn peak cost in the publicly available version of MAGMa were set to (1.0, 2.0, 3.0, 3.0 and 10.0) respectively. These bond break costs were chosen to more or less reflect the corresponding bond dissociation energies (a double bond is twice as strong as a single bond, and a triple or aromatic bond were anticipated to be three-fold as strong). Our calculations indicate that when applying equation (7) to pursue maximum average correct match rank, these parameter values converged towards the value set (1.0, 1.7, 3.0, 1.2, 7.7) for positive mode spectra and (1.0, 7.0, 3.0, 1.9, 10.5) for negative mode spectra (Fig. 4). Using the original parameter set, MAGMa has an average correct identification score of 57.9%. Applying the optimized parameter sets, this score increased to 59.6%. These values correspond to an increase in the total number of correctly assigned spectra in our setup of 272 (from 8,993 to 9,265) and a performance increase of 3.1%, thereby already strengthening the leading position of MAGMa.

Next, we wanted to verify whether MAGMa with optimized parameters shows better performance due to a mere increase in the number of correctly identified metabolites or whether there was also a performance shift involved. We observed that such a shift was indeed present, since the original and optimized MAGMa had only 8,738 correct identifications in common. Hence, the original MAGMa was still capable of correctly assigning 255 MS/MS spectra while MAGMa with improved parameters no longer had the correct metabolite ranked first for these spectra. To further investigate and exploit this behavior, we continued searching the MAGMa parameter space for other parameter sets (both in positive and negative mode) that gradually maximized this amount of uniquely correctly assigned MS/MS spectra as compared to the first optimized sets (following the formalization of equation 6). In the visual representation of these series of parameter sets (Fig. 4), the points "A" correspond to the first sets of optimized parameters mentioned above ((1.0, 1.7, 3.0, 1.2, 7.7) and (1.0, 7.0, 3.0, 1.9, 10.5)), which correctly assign 5,429 and 3,836 spectra in positive and negative modes, respectively. The last points on the white lines ("G") correspond to parameter sets with a smaller total number of correctly assigned MS/MS spectra (4,970 positive mode and 3,312 negative mode MS/MS spectra respectively), but with the highest amount of uniquely correctly identified spectra compared to the corresponding "A" parameter sets (362 and 304 for positive and negative mode respectively). They thereby constituted the most complementary parameter sets. Parameter sets in between those end points have intermediate behavior (Table 2).

One anticipated observation from these data series was the overestimation of the original parameter value (points "Z") for aromatic bond break cost. An aromatic bonds is typically weaker than a triple and even than a double bond, hence the original ABBC value of 3.0 seemed overrated. The optimal values for the aromatic bond break cost in the first optimized ("A") parameter sets take the smaller values of 1.2 (pos. mode) and 1.9 (neg. mode). Even over the entire series of optimized parameters, ABBC varied relatively little (within the [0.1,1.9] range), and thus stay well under the original value of 3.0.

Since the differential performance of parameter sets (points "A" versus subsequent points “B”, “C”, etc., shortly called "x" hereafter) naturally also manifested itself in uniquely correctly identified metabolites by the opposed sets (Table 2), it became legitimate to determine whether there is a latent molecular ground governing the performance differences of different parameter sets. In other words, we wanted to determine whether the different parameter sets performed in the observed complementary manner because they were each optimized for explaining the fragmentation behavior of different molecule types. To achieve this, we trained two-class random forest classifiers based on selected molecular fingerprint features (see Materials and Methods) for each of the "A" versus "x" uniquely correctly identified metabolites. The prediction accuracies obtained using ten fold cross-validation ranged between 58.1 ± 11.1% and 73.9 ± 4.9% for positive ion mode and between 62.7 ± 5.2% and 76.6 ± 2.5% for negative ion mode, while stratified random sampling in most cases resulted in significantly lower prediction accuracies (Table 2). The obtained values thus indicate that there is indeed a latent but predictable molecular difference between the molecules correctly identified by most considered parameter set pairs.

A direct consequence of this observation is that point "A" parameters can be combined with a second "x" parameter set resulting in a further improved MAGMa performance if the parameter set to be applied is chosen on the fly for a given task based on the random forest molecular classifier outcome. In this way, the method is allowed to combine the performance of two complementary parameter sets. To determine which "x" parameter set ("B", "C", etc.) can best be combined with "A", we carried out three performance profiles. These included a discovery set (Metlin) based theoretical performance profile calculated using equation (8) and two empirical performance profile based on the MoNA derived validation spectral dataset using the MAGMa+ script that dynamically selects between two parameter sets based on the determined metabolite class. The performance profiles pointed out an optimal performance for the "A-D" parameter set combination in positive mode and for the "A-C" parameter set combination in negative mode (Fig. 5).

Despite the complete independence and different instrumental origin of our validation datasets, we observed mostly consistent behavior between both the theoretical and empirical performance profiles. For both positive and negative ionization mode, intermediate parameter set combinations provided the largest performance improvements. The lesser performance of preceding combinations (e.g. "A-B") can be explained by corresponding weak (often non-significant) molecular classifier accuracies (little molecular difference) for these combinations. Later parameter set combinations (e.g. "A-G") demonstrate weaker performance since they are characterized by larger differences in total correct assignment numbers (e.g. between the sets "A" and "G"), which would have to be supported by a more accurate molecular classifier to outperform the optimal combinations.

Importantly, the performance of the withheld parameter set combinations, "A-D" (positive mode) and "A-C" (negative mode), which corresponds to the performance of the MAGMa+ script, is further improved relative to the initial optimized "A" parameter sets (Fig. 5 & 6). The theoretical MAGMa+ performance determined on the Metlin dataset is 60.6%, which is slightly higher than the 59.6% obtained using the optimized "A" parameter sets alone and incrementally higher than the 57.9% obtained with the original MAGMa parameters (Fig. 6a). These MAGMa+ results correspond to 243 more positive mode spectra getting correctly identified (4.6% improvement) and 184 more negative mode spectra getting correctly identified (5.0% improvement). On average, the theoretical performance improvement of MAGMa+ is 4.7% over the original MAGMa performance. For the validation datasets, performances increase from 37,7% to 39.9% for the MoNA TOF dataset and from 73.9% to 75.5% for the MoNA FT dataset with MAGMa+ as compared to the original MAGMa (Fig. 6b&c). These increases are attributable to performance improvements of 3.8% (MoNA TOF) and 2.6% (MoNA FT) for positive ion mode and of 11.9% (MoNA TOF) and 0.8% (MoNA FT) for negative ion mode. The improvements obtained on these validation datasets are more variable and less conclusive due to their smaller size. Nonetheless, of importance is their consistent positive sign, indicating the validity of the parameter optimization approach. Averaging theoretical and empirical performance improvement resulted in the winning values of 3.7% (pos. mode) and 5.9% (neg. mode) reported in Fig. 5.

The higher identification scores accomplished on the MoNA FT dataset are not related to a superior performance of MAGMa/MAGMa+ on MS/MS spectra generated on Orbitrap platforms, but due to technical reasons. The MoNA FT dataset indeed contained relatively more metabolites with more unique parent masses for which the search in the structure database return only one or few putative candidates.

CONCLUDING REMARKS

Even though parameter optimizations are a fairly basic intervention, our study proves their usefulness and more generally the usefulness of data-driven approaches for future method development. Considering their different underlying principles and complementary behavior with respect to correct assignments (Fig. 2), future efforts to invigorate the untargeted metabolomics capabilities should be further divided over all (types of) tools. The next generation tools will have to combine the best ideas of all existing algorithms and, with the big data wave hitting the biological fields (and gradually also metabolomics) these days, the real game-changers in the field will emerge from their combination with large datasets. Both the purely machine learning based methods (e.g. CSI:FingerID) as MAGMa and MIDAS, for which a data-driven incremental optimization has been reported here, have demonstrated the validity of this strategy. However, for now, in the case of high-throughput untargeted metabolomics, MAGMa+ still has advantages over CSI:FingerID. The latter can currently not be included in metabolomics identification pipelines and cannot be integrated with customized molecular structure databases. MAGMa+ is therefore better suited as the workhorse identification tool in practical environments. Furthermore, as more laboratories add MS^n-generating mass spectrometers to their toolkit, MAGMa/MAGMa+, being unique in their ability to deal with these data, will undoubtedly continue to increasingly demonstrate their true potential.

Acknowledgements

The authors wish to thank Marco Saerens and Pascal Francq (UCLouvain) and Yorick Poels, Matthieu Moisse and Bram Boeckx (VIB – KU Leuven) for providing computing power and technical support for this research.

Compliance with Ethical Standards

Funding

This study was supported by a Federal Government Belgium grant (IUAP P7/03), long-term structural Methusalem funding by the Flemish Government, grants from the Research Foundation Flanders (FWO), the Foundation Leducq Transatlantic Network (ARTEMIS), Foundation against Cancer, an ERC Advanced Research Grant (EU-ERC269073), an ERC Consolidator Grant (RCN:191995), an AXA Research Fund, and by VIB.

Conflict of interest

The authors declare that they have no conflict of interest.

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

References

Allen, F., Greiner, R., & Wishart, D. (2014). Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics*, 1-13.

CASMI (2015). Critical Assessment of Small Molecule Identification. http://www.casmi-contest.org2015.

Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., et al. (2008). ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res, 36*(Database issue), D344-350, doi:10.1093/nar/gkm791.

Duhrkop, K., Shen, H., Meusel, M., Rousu, J., & Bocker, S. (2015). Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc Natl Acad Sci U S A*, doi:10.1073/pnas.1509788112.

Dunn, W. B., Erban, A., Weber, R. J. M., Creek, D. J., Brown, M., Breitling, R., et al. (2013). Mass appeal: metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics, 9*(1), S44-S66, doi:DOI 10.1007/s11306-012-0434-4.

Durant, J. L., Leland, B. A., Henry, D. R., & Nourse, J. G. (2002). Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Comput Sci, 42*(6), 1273-1280.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning, 46*(1-3), 389-422, doi:Doi 10.1023/A:1012487302797.

Haga, S. W., & Wu, H. F. (2014). Overview of software options for processing, analysis and interpretation of mass spectrometric proteomic data. *J Mass Spectrom, 49*(10), 959-969, doi:10.1002/jms.3414.

Heinonen, M., Shen, H., Zamboni, N., & Rousu, J. (2012). Metabolite identification and molecular fingerprint prediction through machine learning. *Bioinformatics, 28*(18), 2333-2341, doi:10.1093/bioinformatics/bts437.

Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., et al. (2010). MassBank: a public repository for sharing mass spectral data for life sciences. *J Mass Spectrom, 45*(7), 703-714, doi:10.1002/jms.1777.

Hufsky, F., Scheubert, K., & Böcker, S. (2014). Computational mass spectrometry for small-molecule fragmentation. *TrAC Trends in Analytical Chemistry, 53*, 41-48.

Ihlenfeldt, W. D., Voigt, J. H., Bienfait, B., Oellien, F., & Nicklaus, M. C. (2002). Enhanced CACTVS browser of the Open NCI Database. *J Chem Inf Comput Sci, 42*(1), 46-57.

Jeffryes, J. G., Colastani, R. L., Elbadawi-Sidhu, M., Kind, T., Niehaus, T. D., Broadbelt, L. J., et al. (2015). MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. *J Cheminform, 7*, 44, doi:10.1186/s13321-015-0087-1.

Klekota, J., & Roth, F. P. (2008). Chemical substructures that enrich for biological activity. *Bioinformatics, 24*(21), 2518-2525, doi:10.1093/bioinformatics/btn479.

Neumann, S., & Bocker, S. (2010). Computational mass spectrometry for metabolomics: identification of metabolites and small molecules. *Anal Bioanal Chem, 398*(7-8), 2779-2788, doi:10.1007/s00216-010-4142-5.

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., & Hutchison, G. R. (2011). Open Babel: An open chemical toolbox. *J Cheminform, 3*, 33, doi:10.1186/1758-2946-3-33.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.

Ridder, L., van der Hooft, J. J., Verhoeven, S., de Vos, R. C., van Schaik, R., & Vervoort, J. (2012). Substructure-based annotation of high-resolution multistage MS(n) spectral trees. *Rapid Commun Mass Spectrom, 26*(20), 2461-2471, doi:10.1002/rcm.6364.

Smith, C. A., O'Maille, G., Want, E. J., Qin, C., Trauger, S. A., Brandon, T. R., et al. (2005). METLIN: a metabolite mass spectral database. *Ther Drug Monit, 27*(6), 747-751.

Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E., & Willighagen, E. (2003). The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *J Chem Inf Comput Sci, 43*(2), 493-500, doi:10.1021/ci025584y.

Sud, M., Fahy, E., Cotter, D., Brown, A., Dennis, E. A., Glass, C. K., et al. (2007). LMSD: LIPID MAPS structure database. *Nucleic Acids Res, 35*(Database issue), D527-532, doi:10.1093/nar/gkl838.

Tautenhahn, R., Cho, K., Uritboonthai, W., Zhu, Z., Patti, G. J., & Siuzdak, G. (2012). An accelerated workflow for untargeted metabolomics using the METLIN database. *Nat Biotechnol, 30*(9), 826-828, doi:10.1038/nbt.2348.

Vaniya, A., & Fiehn, O. (2015). Using fragmentation trees and mass spectral trees for identifying unknown compounds in metabolomics. *Trends in analytical chemistry : TRAC, 69*, 52-61, doi:10.1016/j.trac.2015.04.002.

Wang, Y., Kora, G., Bowen, B. P., & Pan, C. (2014). MIDAS: a database-searching algorithm for metabolite identification in metabolomics. *Anal Chem, 86*(19), 9496-9503, doi:10.1021/ac5014783.

Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y., et al. (2013). HMDB 3.0--The Human Metabolome Database in 2013. *Nucleic Acids Res, 41*(Database issue), D801-807, doi:10.1093/nar/gks1065.

Wishart, D. S., Knox, C., Guo, A. C., Eisner, R., Young, N., Gautam, B., et al. (2009). HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res, 37*(Database issue), D603-610, doi:10.1093/nar/gkn810.

Wolf, S., Schmidt, S., Muller-Hannemann, M., & Neumann, S. (2010). In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics, 11*, 148, doi:10.1186/1471-2105-11-148.

LEGENDS TO THE FIGURES AND TABLES

FIGURE 1: OVERVIEW OF THE MAGMA+ APPLICABILITY AND OPERATION PRINCIPLE.

a. A typical untargeted metabolomics software pipeline consists of several modules, each dealing with different aspects of the data analysis (e.g. data preprocessing, statistical analysis, metabolite identification, biological interpretation). One of the most challenging aspects of this pipeline is the reliable identification of metabolites. MAGMa+, described in this study, is an optimized version of MAGMa providing this functionality. **b.** MAGMa+ is a wrapper script for the original MAGMa tool. Upon execution, it calls MAGMa twice with different parameters. The metabolite class of individual metabolites in the output list of both runs is determined using a carefully selected and fit random forest (RF) model. Out of both MAGMa outcomes, MAGMa+ then selects the one obtained with the parameters corresponding to the most represented metabolite class.

FIGURE 2: PERFORMANCE COMPARISON BETWEEN METFRAG, MIDAS, CFM-ID AND MAGMA.

a. Cumulative correct assignment percentages among the top k rank positions (k = 1 to 20) provided by the four tools when applied to our evaluation setup. These scores correspond to the combined performance on the four datasets (positive_composite, positive_select, negative_composite and negative_select). **b.** Venn diagram representations of the tool performances on the four evaluation datasets combined are shown. Numbers represent correct spectral assignment counts. The numbers within each tool set sum up to the numbers underneath the tool names. To indicate the original setup of the tools, the applied parameters have been specified when applicable. For MAGMa and MIDAS, (2.0, 3.0, 10.0) and (0.5,0.1,2.0) correspond to the original parameter settings of (DBBC, ABBC, MSP) and (dpp, upp, b) respectively. MIDAS and MAGMa generally performed better although all tools were capable of uniquely correctly assigning metabolites to MS/MS fragmentation spectra. **c.** Cumulative correct assignment percentages obtained with the four tools under comparison when applied to the MoNA FT validation dataset. Here too, MIDAS and MAGMa outperform CFM-ID and MetFrag.

FIGURE 3: REPRESENTATION OF THE MIDAS PARAMETER OPTIMIZATION SPACE

Panels **a** and **b** (resp. **d** and **e**) represent two orthogonal planes in the positive (resp. negative) ionization mode parameter optimization. Browner colors imply better performance. White squares belong to parameter set combinations that have not been evaluated. The color bar values correspond to the number of correct spectral assignments. Grey and white frames indicate original and optimized parameter, respectively. The optimized (dpp, upp, b) parameters are (0.4, 0.1, 1.5) for positive mode and (0.9, 0.1, 1.5) for negative mode spectra. Panel **c** (resp. **f**) contains the same plane as panel **b** (resp. **e**), however, with the additional upp parameter value of 0.0 also evaluated. From these panels it is clear that 0.1 is indeed the optimal value for upp, and reducing it further results in a severe drop in assignment quality.

FIGURE 4: REPRESENTATION OF THE MAGMA PARAMETER OPTIMIZATION SPACE

The double bond break cost (DBBC) - aromatic bond break cost (ABBC) - missing substructure penalty (MSP) space for positive (**a**) and negative (**b**) ionization mode. A total of 23,458 parameter combinations (small dots) have been evaluated. The parameter sets corresponding with the best performance in both cases are labeled with "A". The original parameter setting (2.0, 3.0, 10.0) (identical for both ionization modes) has the "Z"-label. The color code of a parameter set represents the corresponding performance with respect to "A". Blue represents the worse performance, red the best. Color bar values represent the number of correct spectral assignments. The second end points, "G" for both ionization modes, correspond to the parameter sets of which the performance is most differential from "A". Intermediate points result from the gradual transition between both end points and are defined by equation (6).

FIGURE 5: PERFORMANCE PROFILES OF MAGMA PARAMETER SET COMBINATIONS

The theoretical (turquoise) and empirical (orange and blue) performance profiles obtained on the different MAGMa parameter set combinations for positive (**a**) and negative (**b**) ionization mode spectra are presented. Theoretical values are derived from the Metlin-based MS/MS dataset using equation (8). Empirical values are derived from the MoNA TOF and MoNA FT MS/MS dataset using the different MAGMa parameter set combinations. Right-hand-side y-axis values correspond to amounts of correct spectral assignments. The dashed black line represents the average (theoretical-empirical-empirical) performance improvement as compared to the original parameter settings ("Z") and was used to identify the optimal parameter set combinations. The positions where the average performance improvement is highest are indicated with semi-transparent red lines. Average performance improvements of 3.7% ("A-D") and 5.9% ("A-C") were obtained for positive and negative ionization mode, respectively.

Figure 6: Performance comparison between original and optimized MIDAS and MAGMa

a. Cumulative correct assignment percentages among the top k rank positions (k = 1 to 20) obtained in five setups on the Metlin-derived MS/MS dataset. For two of the setups, the MIDAS tool was applied (both with original and optimized parameters). The specified values correspond to the MIDAS (dpp, upp, b) parameter set. The remaining scoring outcomes are obtained with MAGMa in three different setups: (1) original parameters, (2) optimized "A" parameters and (3) the MAGMA+ combination of "A-D" (positive mode) and "A-C" (negative mode) parameters. Since the latter scoring arose from a theoretical evaluation (equation 8), only the rank-1 score is available. Legend values specified for MAGMa correspond to the (DBBC, ABBC, MSP) parameters. The inset shows a zoom of the rank-1 region. **b** and **c.** Same as (**a**) but obtained on the MoNA TOF and MoNA FT MS/MS datasets, respectively. Only the MAGMa outcomes were determined.

Table 1: The list of all parameters determining the MAGMa performance with their corresponding original value.

The "p" parameter describes bond break penalties of which the exact value depends on the type of bond. Stronger bonds break less readily and were therefore given higher p-values. The "h" parameter value depends on the presence of heteroatoms. Since a carbon-carbon bond is stronger than a bond involving heteroatoms, the corresponding h-value of the former was given a higher number. The MSP parameter is the penalty factored in for fragmentation peaks which could not be matched to any substructure. Not all parameters in the list are subject to optimization.

TABLE 2: MAGMA PERFORMANCE DETAILS FOR THE PARAMETER SETS OF INTEREST.

The information provided for each unique parameter set of interest in order of appearance is: (1) its label given in Figure 3; (2) the corresponding parameter values; (3) the a-values resulting in the parameter set when substituted in equation (6); (4) the number of correctly assigned MS/MS spectra from our Metlin-based dataset it provides (decreasing with increasing label symbol); (5) the number of uniquely correctly assigned MS/MS spectra by the "A" parameter set on the one hand (left side) and the parameter set under consideration on the other hand (right side); (6) the number of uniquely correctly identified metabolites by the "A" parameter set on the one hand (left side) and the parameter set under consideration on the other hand (right side); (7) the classification accuracy of the random forest (RF) molecular classifier trained on the uniquely correctly identified metabolites of parameter set "A" versus the parameter set under consideration as determined by 10-fold cross validation; (8) the classification accuracy of stratified random sampling (SRS) from the uniquely correctly identified metabolites of parameter set "A" versus the parameter set under consideration as determined by 10-fold cross validation and (9) the Student's T-test hypothesis test that the RF classifier outperforms the stratified random classification and thus that molecular differences are present between the opposed sets of metabolites. Two RF classifiers, the "A-B" and "A-C" classifiers for positive ionization mode, did not provide significantly better classification than stratified random sampling, implying there is little molecular difference between those groups.

Figure 1

Click here to download Figure Figure1.pdf

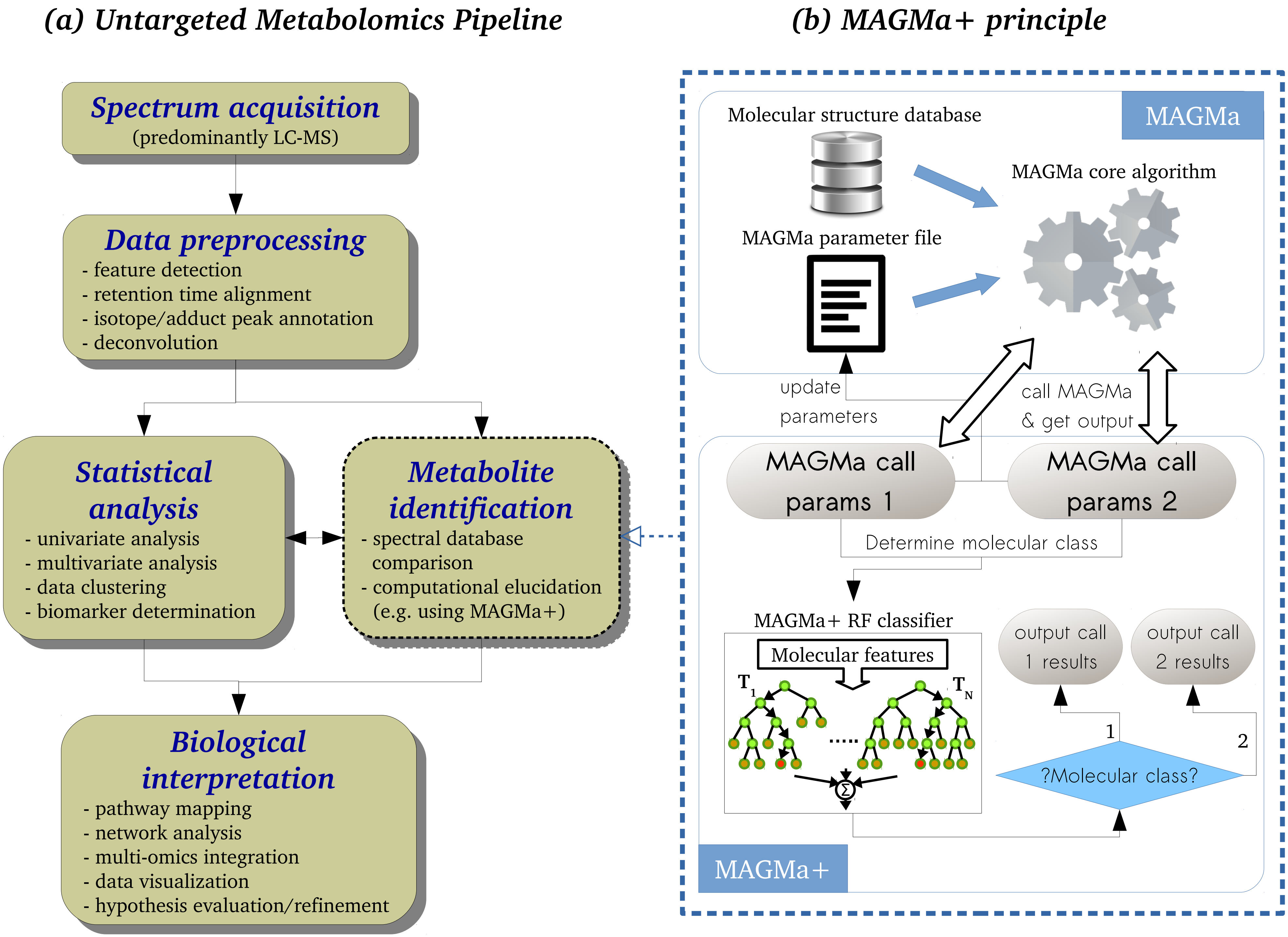


Figure 2

Click here to download Figure Figure2.eps

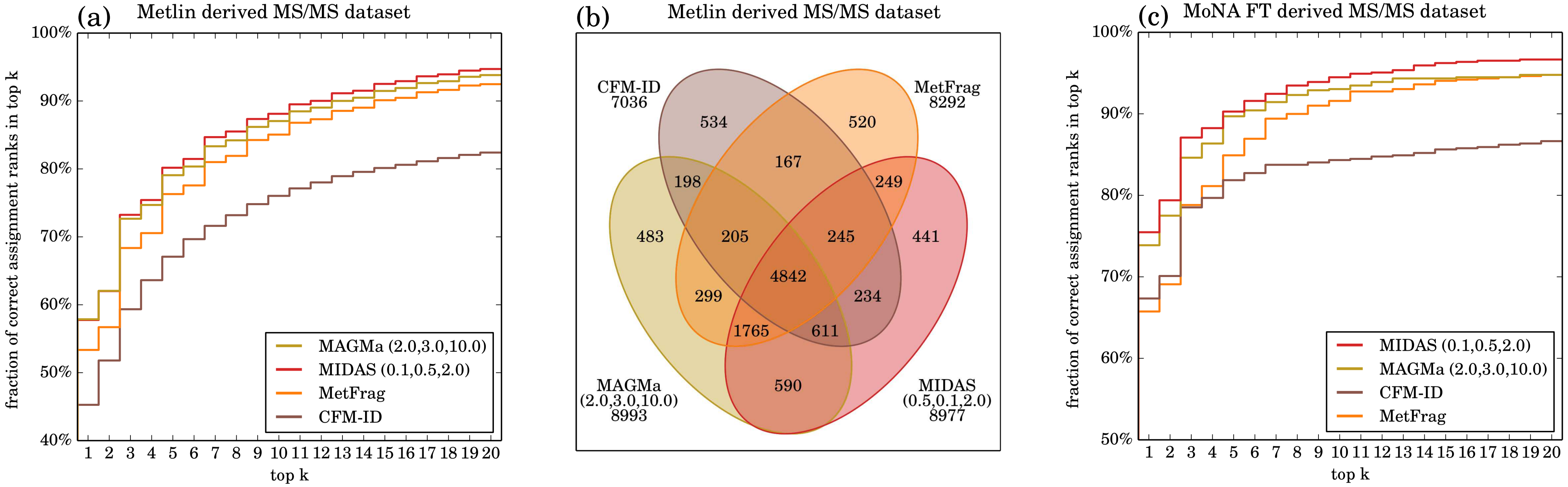


Figure 3

Click here to download Figure Figure3.pdf

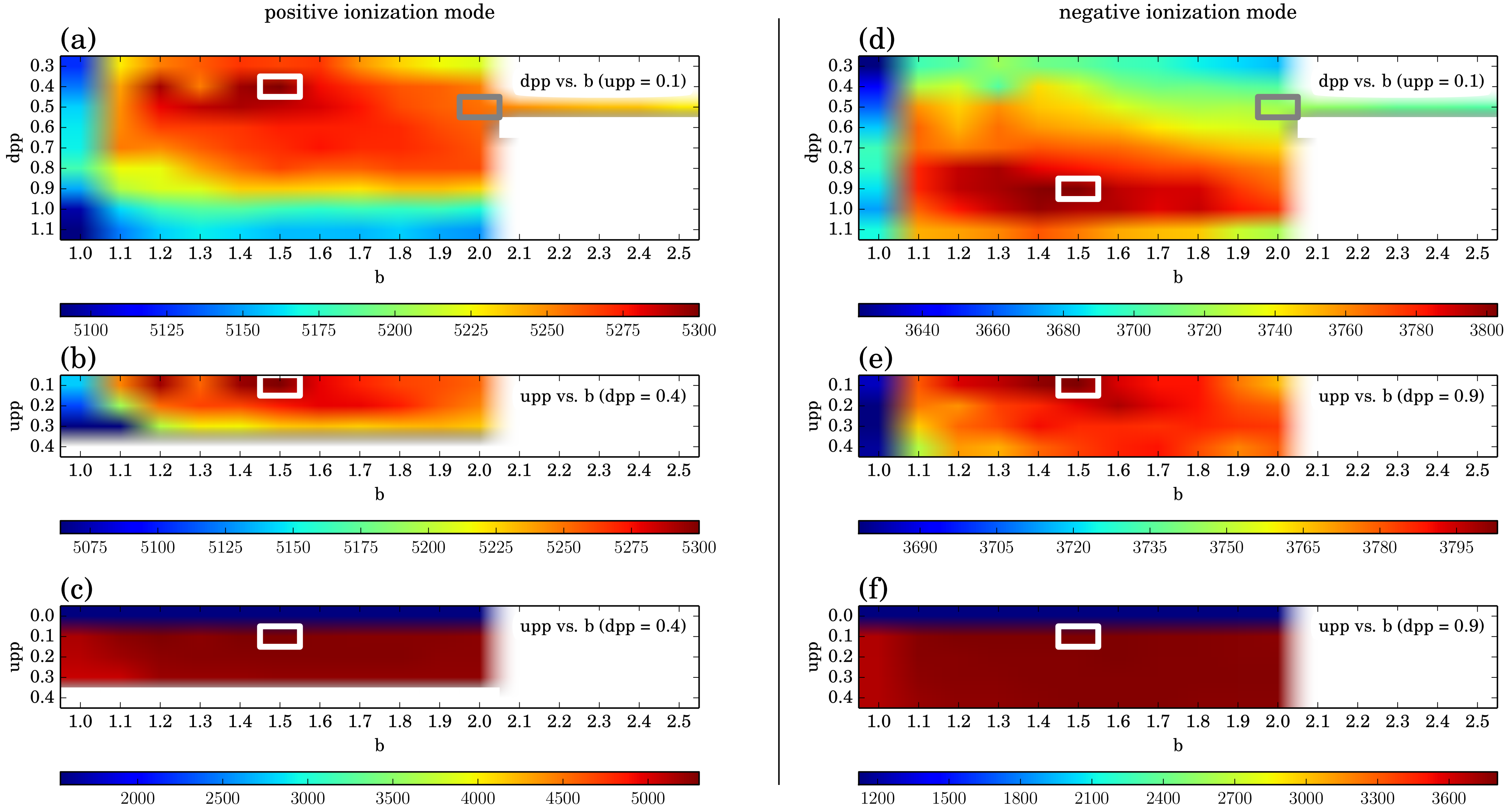


Figure 4

Click here to download Figure Figure4.tiff

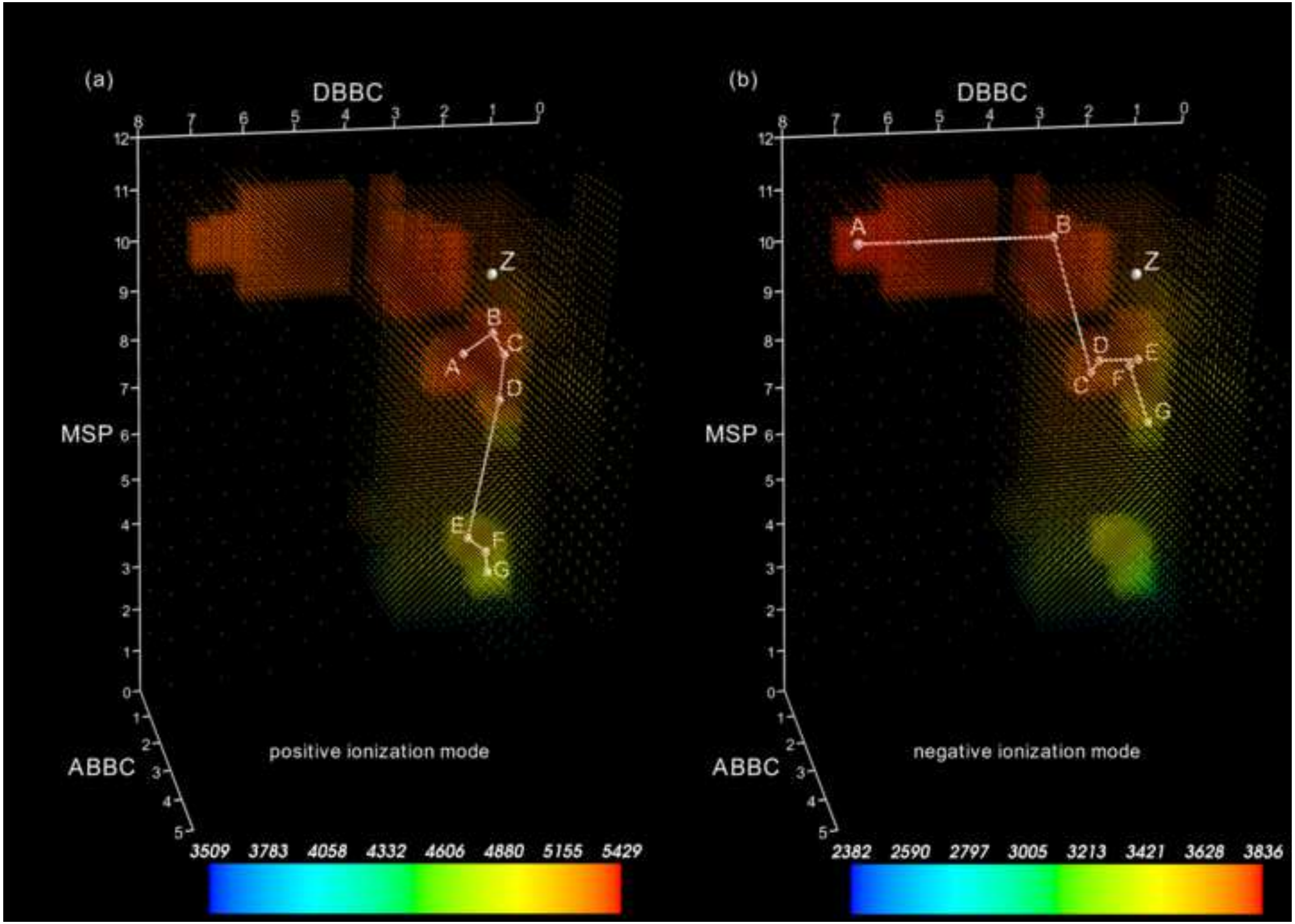


Figure 5

Click here to download Figure Figure5.eps

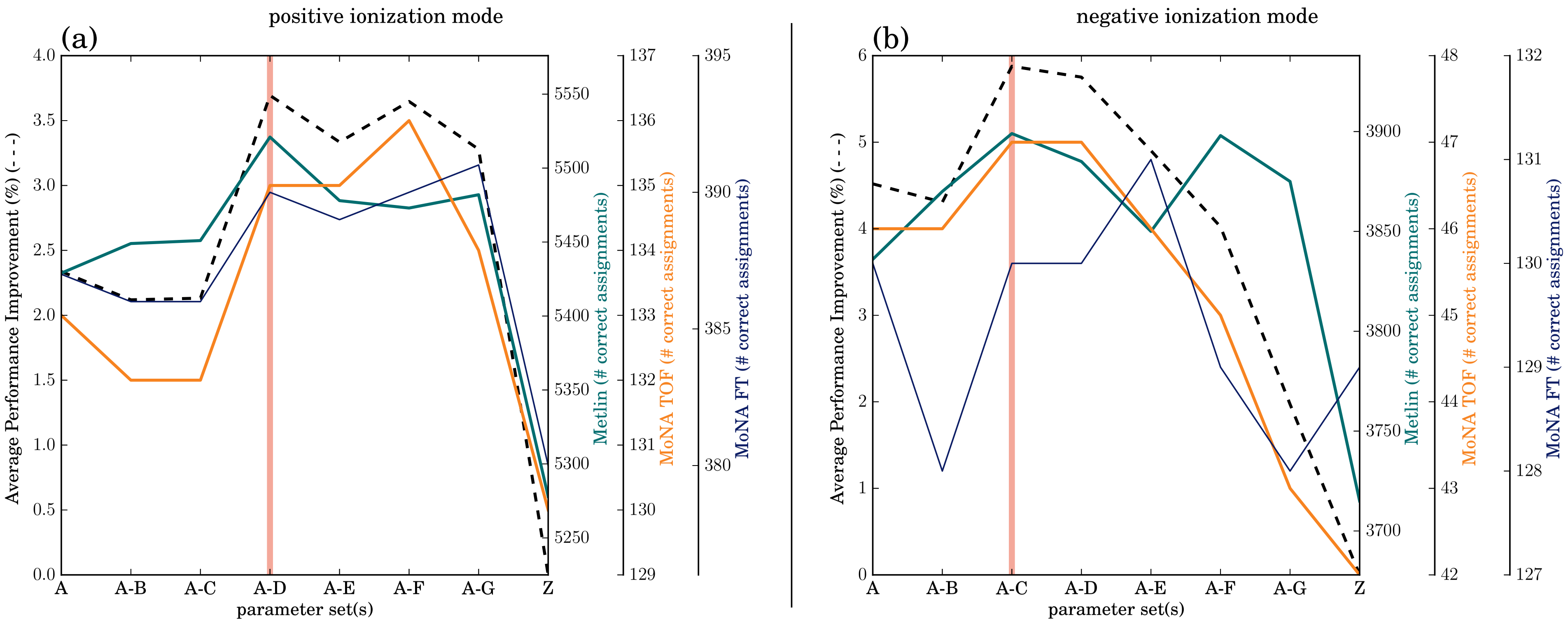
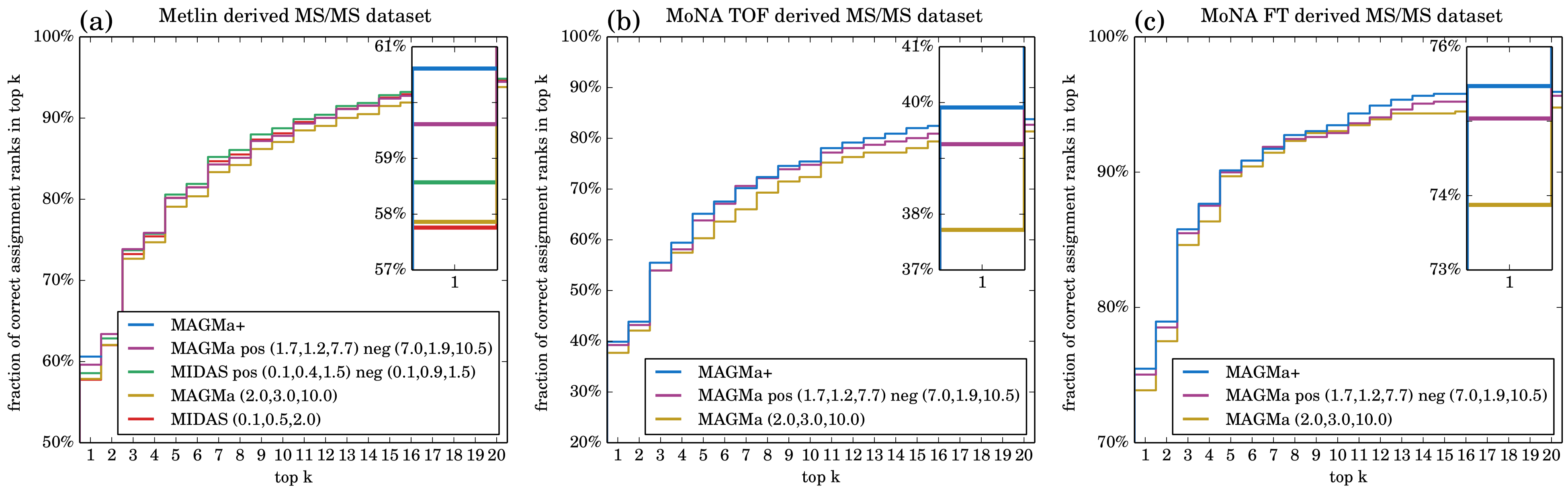


Figure 6

Click here to download Figure Figure6.eps



Parameter	Optimized	Original Value
Single Bond Break Cost	NO	p = 1.0
Double Bond Break Cost	YES	p = 2.0
Triple Bond Break Cost	NO	p = 3.0
Aromatic Bond Break Cost	YES	p = 3.0
Carbon-Carbon Bond Break Cost	NO	h = 2.0
Carbon-Heteroatom Bond Break Cost	NO	h = 1.0
Missing Substructure Penalty	YES	MSP = 10.0

Table 2

Click here to download Table Table2.docx

Positive Ion Mode

point label	parameter set	*a*-values	# correct assignments	# unique correct assignments ("A" vs. "x")	# unique 1st rank metabolites ("A" vs. "x")	RF molecular classifier accuracy (%)	SRS classifier accuracy (%)	T-test p-value
A	(1.7, 1.2, 7.7)	0.0	5429	0 - 0	0 - 0			
B	(1.1, 1.2, 8.1)	0.1	5427	100 - 98	54 - 59	60.6±12.4	53.5±20.1	0.3779
C	(0.9, 1.4, 7.7)	0.2	5419	169 - 159	89 - 97	58.1±11.1	50.8±8.0	0.1252
D	(0.7, 0.9, 6.5)	0.3-0.6	5392	260 - 223	134 - 127	72.9±8.7	49.3±8.3	1.4e-5
E	(1.1, 0.8, 3.3)	0.7	5189	572 - 332	294 - 169	68.7±6.3	54.9±8.0	0.0007
F	(0.8, 1.1, 3.1)	0.8-0.9	5111	676 - 358	360 - 182	69.6±5.8	54.6±5.8	3.3e-5
G	(0.8, 1.3, 2.7)	1.0	4970	821 - 362	445 - 183	73.9±4.6	58.5±5.5	4.5e-6
Z	(2.0, 3.0, 10.0)		5278					

Negative Ion Mode

point label	parameter set	*a*-values	# correct assignments	# unique correct assignments ("A" vs. "x")	# unique 1st rank metabolites ("A" vs. "x ")	RF molecular classifier accuracy (%)	SRS classifier accuracy (%)	T-test p-value
A	(7.0, 1.9, 10.5)	0.0	3836	0 - 0	0 - 0			
B	(3.0, 1.2, 10.2)	0.1-0.3	3830	111 - 105	57 - 58	67.0±10.3	50.0±17.8	0.0234
C	(1.9, 0.9, 7.2)	0.4-0.5	3798	206 - 168	106 - 93	71.9±11.5	47.8±12.4	0.0005
D	(1.7, 0.8, 7.4)	0.6	3777	244 - 185	131 - 106	68.4±8.0	53.5±8.0	0.0009
E	(0.9, 0.9, 7.4)	0.7	3732	312 - 208	160 - 124	62.7±5.2	51.4±8.8	0.0039
F	(0.9, 0.5, 7.1)	0.8	3632	447 - 243	239 - 143	73.8±5.6	51.8±4.1	1.9e-8
G	(0.2, 0.1, 5.6)	0.9-1.0	3312	828 - 304	432 - 179	76.6±2.5	56.1±6.6	8.0e-8
Z	(2.0, 3.0, 10.0)		3715					

Supplementary Material

Click here to access/download
Supplementary Material
SupplementaryMaterial.docx