

Systems biology

Probabilistic framework for integration of mass spectrum and retention time information in small molecule identification

Eric Bach^{1,*}, Simon Rogers², John Williamson² and Juho Rousu¹

¹Department of Computer Science, School of Science, Aalto University, Espoo, Finland and ²School of Computing Science, University of Glasgow, Glasgow, UK

*To whom correspondence should be addressed.

Associate Editor: Pier Luigi Martelli

Received on August 18, 2020; revised on October 27, 2020; editorial decision on November 14, 2020; accepted on November 17, 2020

Abstract

Motivation: Identification of small molecules in a biological sample remains a major bottleneck in molecular biology, despite a decade of rapid development of computational approaches for predicting molecular structures using mass spectrometry (MS) data. Recently, there has been increasing interest in utilizing other information sources, such as liquid chromatography (LC) retention time (RT), to improve identifications solely based on MS information, such as precursor mass-per-charge and tandem mass spectrometry (MS²).

Results: We put forward a probabilistic modelling framework to integrate MS and RT data of multiple features in an LC-MS experiment. We model the MS measurements and all pairwise retention order information as a Markov random field and use efficient approximate inference for scoring and ranking potential molecular structures. Our experiments show improved identification accuracy by combining MS² data and retention orders using our approach, thereby outperforming state-of-the-art methods. Furthermore, we demonstrate the benefit of our model when only a subset of LC-MS features has MS² measurements available besides MS¹.

Availability and implementation: Software and data are freely available at https://github.com/aalto-ics-kepaco/msms_rt_score_integration.

Contact: eric.bach@aalto.fi

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

The identification of small molecules, such as metabolites or drugs, in biological samples is a challenging task posing a bottleneck in various research fields, such as biomedicine, biotechnology, environmental chemistry and drug discovery. In untargeted metabolomics studies, the samples typically contain thousands of different molecules, the vast majority of which remain unidentified (Aksenov *et al.*, 2017; da Silva *et al.*, 2015). Liquid chromatography (LC) coupled with tandem mass spectrometry (MS²) is arguably the most important measurement platform in metabolomics (Blaženović *et al.*, 2018), due to its suitability to high-throughput screening, its high sensitivity and applicability to a wide range of molecules. Briefly explained, LC separates molecules by their differential physicochemical interaction between the stationary and mobile phase, which results in retention time (RT) differences and MS separates molecular ions by their mass per charge (MS¹). Subsequently, MS² can be used to fragment molecules in a narrow mass window and to record the fragment intensities (MS²-spectrum). In an untargeted

metabolomics experiment, large sets of MS features (MS¹ and RT, plus optionally MS²), are observed, corresponding to the different molecules in the sample. Metabolite identification concerns then the structural annotation of the observed MS features.

In recent years, numerous powerful approaches (Nguyen *et al.*, 2018a; Schymanski *et al.*, 2017) to predict molecular structure annotations for MS² spectra have been developed (Allen *et al.*, 2014; Brouard *et al.*, 2016; Dührkop *et al.*, 2015, 2019; Nguyen *et al.*, 2018b, 2019; Ruttkies *et al.*, 2016, 2019). Typically, these methods output a ranked list of molecular structure candidates, that can be shown to human experts, or further post-processed, e.g. by using *additional* information available for the analysed sample. Sources of additional information include, e.g. RT (Bach *et al.*, 2018; Ruttkies *et al.*, 2016; Samaraweera *et al.*, 2018), collision cross-section (Plante *et al.*, 2019) or prior knowledge on the data generating process, such as the source organism's metabolic characteristics (Rutz *et al.*, 2019).

RT, i.e. the time that a molecule takes to elute from the LC column, is readily available in all LC-MS pipelines, and is frequently

used in aiding annotation (Stanstrup et al., 2015). A basic technique is to use the difference between the observed and predicted RT (Domingo-Almenara et al., 2019; Samaraweera et al., 2018) to prune the list of candidate molecular structures. A major challenge for utilizing RT information, however, is that the RT of the same molecule can vary significantly across different LC systems and configurations, necessitating system specific candidate RT reference databases and RT predictors. Different approaches have been proposed to tackle this challenge, such as using physicochemical properties (e.g. partition coefficient, LogP) as RT proxies (Hu et al., 2018; Ruttkies et al., 2016), RT mapping across LC systems (Stanstrup et al., 2015) or predicting retention orders, which are largely preserved within a family of LC systems (e.g. reversed phase) (Bach et al., 2018; Liu et al., 2019). Using LogP as an RT proxy is simple to implement, but only models the hydrophobic separation effects of the LC system. RT mapping, on the other hand, is limited to pairs of LC systems in which the same molecules have been measured. Retention order prediction can overcome those drawbacks, by learning the LC system's separation directly from RT data of multiple systems (Bach et al., 2018).

This study proposes a probabilistic framework to integrate MS¹ or MS²-based annotations with predicted retention order for improved small molecule identification given a set of MS features measured within one LC-MS run by building on the work by Bach et al. (2018) and Del Carratore et al. (2019). The latter proposed a *probabilistic* approach for integrating different types of additional information to MS¹ data, including RT information. We too define a probabilistic approach, but differ in how RT is handled. Where Del Carratore et al. (2019) use absolute RT information, we follow Bach et al. (2018) and use pairwise retention order predictions for molecules eluting within the same LC-MS run. In contrast to the work done by Bach et al. (2018), our model makes use of pairwise retention order information between all MS features rather than only the ones adjacent in terms of their RTs, resulting in more accurate annotations. Furthermore, our model allows to rank all candidate lists, instead of just returning the most likely candidate assignment for each MS feature, as done by Bach et al. (2018).

Our framework models the score integration as an inference problem on a graphical model, where the edges correspond to retention order predictions, the nodes correspond to MS features and the node labels correspond to candidate molecular structures, scored by a MS² based predictor, such as CSI: FingerID (Dührkop et al., 2015), MetFrag (Ruttkies et al., 2016) or IOKR (Brouard et al., 2016), or in the absence of MS² information, MS¹ precursor mass deviation. This graph is fully connected, which makes exact inference an NP-hard problem. To solve this challenge, we resort to efficient approximate inference, in particular spanning tree approximations (Marchand et al., 2014; Pletscher et al., 2009; Su and Rousu, 2015; Wainwright et al., 2005).

2 Materials and methods

2.1 Overall workflow

We assume data arising from a typical LC-MS-based experimental workflow (including chromatographic peak picking, and alignment): MS features consisting MS¹ measurement and the associated RT. A subset of these will include an MS² spectrum. In the following, we present our score-integration model in the most general form in which it is provided with MS features and a set of possible candidate molecular structures. The candidate list can be generated, e.g. by querying molecular structures from a structure database, such as ChemSpider (Pence and Williams, 2010), that have the same mass as the observed MS feature. In addition, we assume that to each candidate structure a score is assigned by either an MS²-based predictor, or, if no MS²-spectrum is available, a score based on the mass deviation of the candidates from the MS mass. For all molecular candidate pairs, associated with the different MS features, the retention order is predicted. Here, we use the Ranking Support Vector Machine (RankSVM)-based predictor by Bach et al. (2018). The candidate structure scores and predicted retention orders are

integrated through a probabilistic graphical model (described in the following). This allows us to rank the molecular candidate structures by their inferred marginal probabilities, given both the MS and RT information.

More formally, the output of an LC-MS experiment is given as a tuple set $\mathcal{D} = \{(x_i, t_i, C_i)\}$, with $x_i \in \mathcal{X}$ being the spectrum of feature i (either an MS² or a spectrum containing only a single peak at the mass of the precursor ion if no MS² information is available), $t_i \in \mathbb{R}_{\geq 0}$ being its RT, and $C_i = \{m_{i1}, \dots, m_{in_i}\} \subseteq \mathcal{M}$ being the associated molecular candidates. Here, $m_{ir} \in \mathcal{M}$ represents a molecular candidate structure and n_i is the number of molecular candidates for the i th MS feature. Figure 1 shows an overview of our workflow.

2.2 Probabilistic model

Let $G = (V, E)$ be an undirected graph, in which each node, $i \in V$ represents one observed MS feature, and with an edge for all MS feature pairs $E = \{(i, j) | i, j \in V, i \neq j\}$. The edge-set E does not contain any parallel edges. The number of MS features is denoted with N , i.e. $|V| = N$. We associate each node i in the vertex set with a discrete random variable z_i that takes values from the space $\mathcal{Z}_i = \{1, \dots, n_i\}$. Intuitively, z_i defines which candidate has been assigned to the i th MS feature. The full vector $\mathbf{z} = \{z_i | i \in V\}$ corresponds to the molecular structure assignment to each MS feature in the LC-MS experiment, and it takes values from the set $\mathcal{Z} = \mathcal{Z}_0 \times \dots \times \mathcal{Z}_N$. In this work, we consider \mathcal{Z} to be fixed and finite for a given a set of MS features, due to our definition of the molecular candidates sets, which assumes that we can restrict the putative annotation for a given MS feature.

2.2.1 Markov random field

The probability distribution of \mathbf{z} is given as a *pairwise Markov Random Field* (MRF) (MacKay, 2005):

$$p(\mathbf{z}) = \frac{1}{Z} \prod_{i \in V} \psi_i(z_i) \prod_{(i,j) \in E} \psi_{ij}(z_i, z_j), \quad (1)$$

composed of node ψ_i and edge ψ_{ij} potential functions, and omits higher-order cliques (hence the term pairwise). Above, $\psi_i : \mathcal{Z}_i \rightarrow \mathbb{R}_{>0}$ is the potential function of node i measuring how well the i 's candidates matches the measured MS information, and $\psi_{ij} : \mathcal{Z}_i \times \mathcal{Z}_j \rightarrow \mathbb{R}_{>0}$ encodes the consistency of the observed retention orders for MS feature i and j and the *predicted* retention order of their candidates z_i and z_j and $Z = \sum_{\mathbf{z} \in \mathcal{Z}} \left(\prod_{i \in V} \psi_i(z_i) \prod_{(i,j) \in E} \psi_{ij}(z_i, z_j) \right)$ is the partition function (MacKay, 2005).

2.2.2 Node potential function ψ_i

For each candidate $m_{ir}, r \in \mathcal{Z}_i$, we predict a matching score $\theta_{ir} = f(x_i, m_{ir}) \in \mathbb{R}$ expressing how well it matches the observed MS¹ or MS² spectrum x_i . For that, we assume a pre-trained model, such as CSI: FingerID (Dührkop et al., 2015), MetFrag (Ruttkies et al., 2016) or IOKR (Brouard et al., 2016). We use the latter two in our experiments as representative MS² scoring methods (Section 3.3). MetFrag performs an *in silico* fragmentation of m_{ir} , compares these fragments peaks with the observed ones in x_i and outputs a matching score. IOKR, on the other hand, can be used to directly predict a matching score $f(x, m)$ for any (MS² feature, molecular structure)-tuple. All matching scores θ_{ir} are normalized to the range $[0, 1]$. Finally, we express the potential of a molecular candidate m_r given the spectrum x_i as follows:

$$\psi_i(z_i = r) = \max(\theta_{ir}, c),$$

where $c > 0$ is a constant used to avoid zero potentials. In our experiments, we select c such that it is 10 times smaller than the minimum of all non-zero scores across all candidate sets.

2.2.3 Edge potential function ψ_{ij}

For each candidate pair $(r, s) \in \mathcal{Z}_i \times \mathcal{Z}_j$ associated with the MS pair (i, j) , we compute how well the candidates' predicted retention order

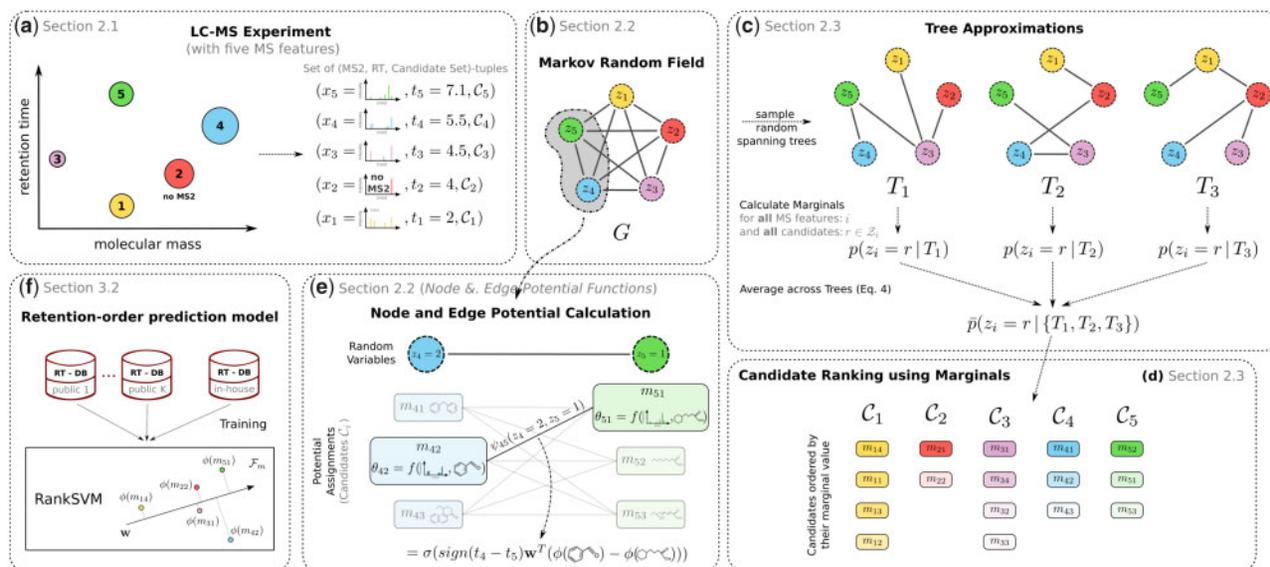


Fig. 1. Workflow of our framework and its main components. (a) Data acquisition in an LC-MS experiment resulting in a set of (MS^2, RT) -tuples of unknown molecules. (b) Illustration of the underlying graphical model. (c) Ensemble of spanning trees to approximate the MRF and their integration using averaged marginals. (d) Output to the user: ranked molecular candidate lists based on the approximated marginals. (e) Incorporation of the predicted retention orders for a particular assignment for z via the edge potential function. (f) Illustration of the RankSVM model

is aligned with the observed one defined by the RTs t_i and t_j . To this end, we apply the framework for retention order prediction developed by Bach et al. (2018). The edge potential $\psi_{ij}(z_i = r, z_j = s)$ is defined as follows:

$$\psi_{ij}(z_i = r, z_j = s) = \sigma(\text{sign}(t_i - t_j) \cdot \mathbf{w}^T(\phi_{ir} - \phi_{js})),$$

where $\mathbf{w} \in \mathbb{R}^{|\mathcal{F}_m|}$ is the RankSVM's parameter vector, and $\phi_{ir}, \phi_{js} \in \mathcal{F}_m$ are the feature vectors of the candidates' molecular structures, and $\sigma: \mathbb{R} \rightarrow (0, 1]$ is a monotonic function mapping the predicted preference value difference to a value between zero and one. In our experiments, we consider two mapping functions:

$$\begin{aligned} \text{Sigmoid: } & \sigma_{\text{sigmoid}}(x) = \frac{1}{1 + \exp(-kx)} \\ \text{Step - Function: } & \sigma_{\text{step}}(x) = \begin{cases} \epsilon, & x < 0 \\ 1, & x \geq 0 \end{cases}, \epsilon = 10^{-10}. \end{aligned}$$

The different functions can be interpreted as follows. The *sigmoid* makes full use of the information from the RankSVM margin, i.e. the score of each candidate pair depends on the preference score difference. In this work, we consider k as a hyper-parameter of our method that needs to be estimated from data (Section 3.5). The *step-function*, on the other hand, only differentiates between aligned and not aligned pairs.

2.2.4 Weighting of information sources

To control the contribution of each information source, i.e. MS information and retention orders, we introduce a modification on the potential functions:

$$p(\mathbf{z}) = \frac{1}{Z} \prod_{i \in V} \psi_i(z_i)^{1-D} \prod_{(i,j) \in E} \psi_{ij}(z_i, z_j)^D$$

with $D \in [0, 1]$. A D value close to one, e.g. will result in a score mainly based on the observed retention orders. In our experiments, we explain how this hyper-parameter can be estimated in practice (Section 3.5).

2.3 Ranking candidates through approximated marginals

We rank the molecular candidates using the marginals of the MRF (1). The marginal for the candidate r of MS feature i is given as:

$$p(z_i = r) = \sum_{\{z' \in Z \mid z'_i = r\}} p(z'). \quad (2)$$

In practice, the calculation of (2) is intractable due to the size of the domain Z of \mathbf{z} , which grows exponentially with the number of MS features, thus we will resort to approximate inference methods.

2.3.1 Tree approximation of G

To enable feasible inference of (2), we approximate the MRF (1) using spanning trees of the original graphical model G (Marchand et al., 2014; Pletscher et al., 2009; Su and Rousu, 2015; Wainwright et al., 2005). In the following let T be a spanning tree of G with the same nodes, but an edge-set $E(T) \subseteq E$, with $|E(T)| = N - 1$, that ensures T being a cycle-free single connected component. The probability distribution of the graphical model induced by T is given as:

$$p(\mathbf{z}|T) = \frac{1}{Z(T)} \prod_{i \in V} \psi_i(z_i) \prod_{(i,j) \in E(T)} \psi_{ij}(z_i, z_j). \quad (3)$$

As the graphical model associated with (3) is a tree, we can exactly infer its marginals through the *sum-product* algorithm (MacKay, 2005). The sum-product algorithm is a message-passing algorithm using dynamic programming that has linear time complexity in the number of MS features. See, e.g. MacKay (2005) for further details on the algorithm.

The output of the sum-product algorithm are the unnormalized marginals $\mu(z_i = r|T)$ for all $i \in V$ and $r \in \mathcal{Z}_i$. We calculate the normalized marginals as follows (MacKay, 2005):

$$p(z_i = r|T) = \frac{\mu(z_i = r|T)}{\sum_{r'=1}^{M_i} \mu(z_i = r'|T)}.$$

2.3.2 Random spanning trees sampling

We compare two approaches to retrieve spanning trees from G . The first approach is to randomly sample spanning trees from G (c.f.

Pletscher *et al.*, 2009; Su and Rousu, 2015; Wainwright *et al.*, 2005). We sample the trees by applying the minimum weighted spanning tree algorithm to a random adjacency matrix. If for an MS feature pair (i, j) both RTs are equal, i.e. $t_i=t_j$, then their corresponding edge is not sampled. This is justified by the observation, that MS features with a RT difference equal zero, do not impose constraints on the retention order of their corresponding candidates. We will refer to a sampled spanning tree as T_i . The second approach was implicitly used by Bach *et al.* (2018) and corresponds to a linear Markov chain where edges connect adjacent MS features ordered by increasing RT, which can be seen as a degenerate spanning tree. In the remaining text, we refer to this tree as T_{chain} .

2.3.3 Averaged marginal over a random spanning tree ensemble

Using tree-like graphical models for the inference is motivated by the exact and fast inference it enables us to do. However, a single tree, such as T_{chain} or a sampled T_i , will most likely be only a rough approximation of the original probability distribution (1). Therefore, the use of random spanning tree ensembles $\mathbf{T} = \{T_i\}_{i=1}^L$ has been proposed. In particular, Wainwright *et al.* (2005) show that an expectation over trees can be used to obtain an upper bound on the maximum *a posteriori* (MAP) estimate of the original graph, and showed that this approximation can be tight if the underlying trees agree about the MAP configuration. More recently, Marchand *et al.* (2014) demonstrated generalization bounds for joint learning and inference using tree ensembles. More applied work in using tree-based approximation can be found in Pletscher *et al.* (2009), who use majority voting and Su and Rousu (2015), who empirically study several aggregation schemes in multilabel classification.

Motivated by the mentioned literature, we opted to average the marginals of a random spanning tree ensemble \mathbf{T} , where for each tree T_i , we independently retrieve the marginals using the sum-product:

$$\bar{p}(z_i = r|\mathbf{T}) = \frac{1}{L} \sum_{i=1}^L p(z_i = r|T_i). \quad (4)$$

2.3.4 Max-marginals

The exact inference on trees allows us to use the max-marginal, as an alternative to the sum-marginal shown in Equation (2). The max-marginal is closely related to the MAP estimate. For a single tree T it is given as:

$$p_{\max}(z_i = r|T) = \max_{\{z' \in \mathcal{Z} | z'_i = r\}} p(z'|T).$$

The interpretation of the two marginals (sum and max) differs slightly. Whereas the sum-marginal expresses the *sum* of the probabilities of all configurations z' with $z'_i = r$, the max-marginal is the maximum probability that a configuration with the constraint $z'_i = r$ can reach. In our experiments, we compare the performance of both marginal types (Section 4.1). The *max-product* algorithm is used to calculate the unnormalized max-marginals $\mu_{\max}(z_i = r|T)$, which is a modification of the sum-product algorithm, in which summations are replaced by maximization. The normalized marginal can be calculated as (MacKay, 2005):

$$p_{\max}(z_i = r|T) = \frac{\mu_{\max}(z_i = r|T)}{\max_{r' \in \{1, \dots, n_i\}} \mu_{\max}(z_i = r'|T)}. \quad (5)$$

By plugging Equation (5) into (4) instead of the sum-marginal, we obtain the averaged max-marginal \bar{p}_{\max} .

2.3.5 Run-time complexity

Calculating the marginals for an individual tree and all MS features i has run-time complexity $\mathcal{O}(N \cdot n_{\max}^2)$. Here, N is the total number of features and $n_{\max} = \max_{i \in V} |Z_i|$ the maximum number of molecular candidates for a feature.

3 Material and experiments

3.1 Evaluation datasets

To evaluate our score-integration approach, we use two publicly available datasets. These are described in this section and summarized in Table 1.

3.1.1 CASMI 2016

The Critical Assessment of Small Molecule Identification (CASMI) challenge is a contest organized for the computational spectrometry community (Schymanski *et al.*, 2017). For its implementation in 2016, a dataset of 208 (MS², retention-time)-tuples was released. The dataset contains 81 negative and 127 positive ionization mode MS² spectra. The molecular candidate structure sets were extracted from ChemSpider, using a ± 5 ppm window around the monoisotopic exact mass of the correct candidate, by the challenge organizers.

3.1.2 EA (Massbank)

Massbank (Horai *et al.*, 2010) is a publicly available repository for MS data. For the development of MetFrag 2.2, Ruttkies *et al.* (2016) extracted 473 (MS², retention-time)-tuples of 359 unique molecular structures from Massbank (EA dataset). The dataset is split into 154 negative and 319 positive ionization mode MS² spectra. We used the molecular candidates provided by Ruttkies *et al.* (2016) extracted from ChemSpider using the molecular formula (MF) of the correct candidate.

For each dataset and ionization mode, we repeatedly subsample training and test (MS², RT)-tuple sets: CASMI (negative) 50-times $N_{\text{train}} = 31, N_{\text{test}} = 50$; CASMI (positive) 50-times $N_{\text{train}} = 52, N_{\text{test}} = 75$; EA (negative) 50-times $N_{\text{train}} = 45, N_{\text{test}} = 65$; and EA (positive) 100-times $N_{\text{train}} = 50, N_{\text{test}} = 100$. No molecular structure, determined by its InChI representation, appears simultaneously in test and training. The training set is used for the hyper-parameter selection (Section 3.5) and the test sets are used to assess the average identification performance of our score-integration framework (Section 3.4).

3.2 Training setup for the retention order predictor

To calculate the edge potentials of our MRF model (1), we use the RankSVM retention order prediction approach by Bach *et al.* (2018). The RankSVM model is trained using seven publicly available RT datasets. Six were published by Stanstrup *et al.* (2015) along with their RT mapping tool PredRet: UFZ_Phenomenex, FEM_long, FEM_orbitrap_plasma, FEM_orbitrap_urine, FEM_short and Eawag_XBridgeC18. The seventh dataset contains examples for which RTs were published as part of the training dataset for the CASMI 2016 challenge (Schymanski *et al.*, 2017). The joint dataset covered four different chromatographic columns all using H₂O → MeOH (with 0.1% formic acid as additive) as eluent. In total, the dataset contained 1248 (molecule, RT)-tuples of 890 unique molecular structures, after the same pre-processing as in Bach *et al.* (2018) was applied. We represent the molecular structures using Substructure counting fingerprints calculated with rcdk and CDK 2.2 (Willighagen *et al.*, 2017). We use the MinMax-kernel (Ralaivola *et al.*, 2005) to calculate the similarity between the fingerprints. For our experiments, we build an individual RankSVM model for each (MS, RT)-tuple subsample (Section 3.1), ensuring no molecular structure in the subsample is used for the RankSVM training.

3.3 MS²-based match scores from MetFrag and IOKR

We apply MetFrag (Ruttkies *et al.*, 2016) and IOKR (Brouard *et al.*, 2016) as representative methods to obtain MS² matching scores for the molecular structures in the candidate list of each MS² spectrum.

3.3.1 MetFrag

We use the latest MetFrag version 2.4.5 (<http://msbi.ipb-halle.de/cruttkie/metfrag/MetFrag2.4.5-CL.jar>) and utilize it as described in Ruttkies *et al.* (2016). The MS² matching scores are calculated using the FragmenterScore feature of MetFrag.

Table 1. Summary of the datasets used for the evaluation of our score-integration framework

Dataset	Ionization	Mass spectra info. MS ¹ info.	Molecular candidates ^a		Chromatography		Eluent
			#MS ²	Tot. #Cand.	Median #Cand.	Column	
CASMI 2016	Negative	Precursor m/z	81	74 589	420	Phenomenex Kinetex EVO C18	H2O → MeOH (both 0.1% formic acid)
CASMI 2016	Positive	Precursor m/z	127	183 633	919	Phenomenex Kinetex EVO C18	H2O → MeOH (both 0.1% formic acid)
EA (Massbank)	Negative	Precursor m/z	154	75 107	119.5	Waters XBridge C18	H2O → MeOH (both 0.1% formic acid)
EA (Massbank)	Positive	Precursor m/z	319	215 893	246	Waters XBridge C18	H2O → MeOH (both 0.1% formic acid)

^aExtracted from ChemSpider. CASMI: ± 5 ppm window around monoisotopic exact mass of correct candidate. EA: MF of correct candidate.

3.3.2 IOKR

Two IOKR models are trained, for negative and positive mode MS² spectra, respectively. The training (MS², molecular structure)-tuples are extracted from GNPS (Wang et al., 2016), Massbank and the CASMI 2016 training data. We remove training molecular structures that appear in our evaluation datasets (Section 3.1). This results in 3255 negative and 6773 positive mode training examples. We use a uniform combination of 16 MS² spectra and fragmentation tree (FT) kernels as input kernel (Supplementary Section S4). On the output side, we use the same molecular fingerprint definitions as Dührkop et al. (2019) as feature representation and a Gaussian kernel those distances are derived from the Tanimoto kernel (Brouard et al., 2019) as output kernel. For all MS² spectra in our evaluation datasets, we calculate the FTs using SIRIUS 4.0.1 (Dührkop et al., 2019) and keep the highest scoring tree for each spectra to calculate the MS² and FT kernels used by the IOKR.

3.4 Performance evaluation

In our experiments, we use the top- k accuracy to determine the metabolite identification performance, i.e. the percentage of correctly ranked molecular candidates at rank k or less. Different approaches can be used to determine the rank of the correct structure. We follow the protocol used by Schymanski et al. (2017). If multiple stereo-isomers were present in the candidate list, only the one with the highest MS²-score was retained. The correct molecular structure was found by comparing the InChIs containing no stereo information. The top- k accuracies are calculated the test sets.

3.5 Hyper-parameter estimation

The training set of each individual subsample is used to determine optimal weighting D between MS and retention order information. For that, we run the score-integration framework for a different D values, and calculate the area under-the-ranking curve up to rank 20: $\text{top20AUC} = \frac{1}{20} \sum_{i=1}^{\text{top}(i)} \frac{i}{N}$, where $\text{top}(i)$ is the number of correct structures up to rank i and N is the number of MS features. Subsequently, we select the retention order weight with the highest top20AUC (Supplementary Section S2). The optimal sigmoid parameter k is estimated using Platt's method (Lin et al., 2007; Platt, 2000) calibrated using RankSVM's training data (Section 3.2).

3.6 Experiments

3.6.1 Full MS² information available

We compare our approach for combining MS² and RT information for metabolite identification against the baseline, which only uses MS² information for the candidate ranking. This allowed us to quantify the performance gain by using RTs. Furthermore, we applied two recently published methods for the integration of MS² and RT scores and compared them to our approach. The first one is MetFrag 2.2 (Ruttkies et al., 2016), which exploits the RT information by establishing a linear relationship between the candidates' predicted LogP values with the observed RTs. Each molecular candidate receives an additional score by comparing its predicted RT against that observed for the corresponding MS² spectra. We use the

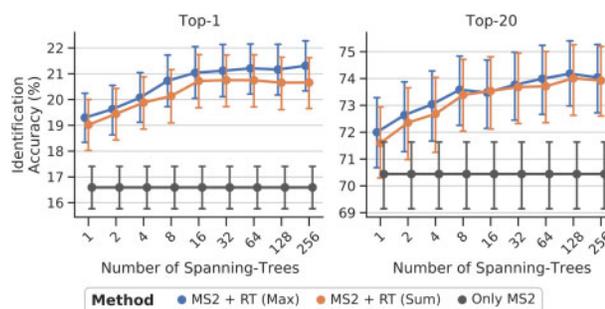


Fig. 2. Top- k accuracies, averaged across all datasets and ionizations, plotted against the number of random spanning trees (L) used for the approximation. The baseline using only MS² information is plotted in black. The sigmoid function is used in the score integration. The differences between the sum- and max-marginals' average performance for the L values is significant ($P < 0.001$ for Top-1 and 20, Two-sided Wilcoxon Signed-Rank test)

CDK (Willighagen et al., 2017) XLogP predictions, which are automatically calculated by the MetFrag software. The weight of the RT feature RetentionTimeScore is determined as described in Section 3.5. Our second comparison is the approach by Bach et al. (2018), which uses predicted retention order and dynamic programming over a chain-graph connecting adjacent MS features. Bach et al. (2018) focussed in extracting the most likely assignment z [Equation (3)] given the chain-graph using dynamic programming. Here, we use our generalized framework to also compute the marginals of all candidates given the chain-graph T_{chain} . The approach by Bach et al. (2018) implicitly used a hinge-sigmoid to compute edge potentials: $\sigma_{\text{hinge}}(x) = \min\left(\frac{2}{1+\exp(-kx)}, 1.0\right)$. Its parameter k is determined as described in the Supplementary Section S2. We refer to this method as Chain-graph.

3.6.2 Missing MS²

In a second experiment, we simulated the common application scenario, in which during an LC-MS experiment, a set of MS features (MS and RT) have been measured, but MS² spectra have only been acquired for a subset of the features. There can be multiple reasons for this, such as limited measuring time when using, e.g. data-dependent acquisition (Xiao et al., 2012) protocols, bad fragmentation quality or inability to deconvolute all spectra when using data-independent acquisition. In this case, besides the RT, only MS¹ related information is available for some features, which includes the mass of the ion (precursor m/z) and its isotope pattern. We use our proposed score-integration framework to perform structural identification when the proportion of MS features that have an MS² spectrum varies. We vary the percentage of available MS²-spectra from 0% to 100% and investigate how the joint use of MS¹, MS² and RT information can improve over the baseline solely relying on only MS information. For the candidates r of an MS feature i , simulated to be without MS² spectrum, we assign the following candidate score (Del Carratore et al., 2019):

Table 2. Identification accuracies (top- k) for the different datasets and ionization modes

Dataset	Method	Negative				Positive			
		Top-1	Top-5	Top-10	Top-20	Top-1	Top-5	Top-10	Top-20
CASMI 2016	MS ² + RT (<i>our</i>)	15.2 (***)	47.2 (***)	57.0 (**)	70.1 (***)	14.0 (***)	40.7 (***)	52.2 (***)	62.8 (***)
	MS ² + RT (Chain-graph)	13.2 (***)	49.4 (***)	61.0 (***)	69.4 (***)	11.9	36.5	50.2 (***)	60.7 (***)
	MS ² + RT (MetFrag 2.2)	14.0 (***)	42.0	55.5	71.2 (***)	13.7 (***)	36.2	46.2	57.5
	Only MS ²	11.1	44.2	55.3	68.0	11.8	37.3	47.0	58.3
EA Massbank	MS ² + RT (<i>our</i>)	28.7 (***)	61.9 (***)	73.8 (***)	83.6 (***)	27.3 (***)	61.6 (***)	72.9 (***)	80.7 (***)
	MS ² + RT (Chain-graph)	27.2 (***)	59.5 (***)	72.4 (***)	81.8 (***)	23.9 (***)	59.2	70.1	79.1 (***)
	MS ² + RT (MetFrag 2.2)	30.2 (***)	59.2 (***)	73.6 (***)	84.4 (***)	24.0 (***)	59.0	69.5	77.1
	Only MS ²	22.8	57.6	69.5	78.5	21.2	59.0	69.7	77.6

Note: Compares our score-integration framework (MS² + RT (*our*)), against the baseline (Only MS²), MetFrag 2.2 with predicted RT and the Chain-graph model. The best performance for each dataset and ionization is indicated by bold-font. The stars (*) represent the significant improvement over the baseline calculated using a one-sided Wilcoxon signed-rank test on the sample top- k accuracies (P < 0.05 (*), P < 0.01 (**), and P < 0.001 (***)).

Table 3. Pairwise test for significant improvement of the MS² + RT score-integration methods: *Our*, MetFrag 2.2 and Chain-graph

Method (MS ² + RT)	Top-1		Top-20	
	Chain-graph	MetFrag 2.2	Chain-graph	MetFrag 2.2
<i>Our</i>	8.7 · 10 ⁻²⁴	6.1 · 10 ⁻⁸	2.1 · 10 ⁻¹³	1.5 · 10 ⁻¹⁴
Chain-graph	—	n.s.	—	8.1 · 10 ⁻⁰⁴
MetFrag 2.2	4.3 · 10 ⁻⁰⁸	—	n.s.	—

Note: We show the P -values for testing the improvement of the row over the column method using a one-sided Wilcoxon signed-rank test. The test is performed over all top- k accuracy samples (datasets and ionization). MetFrag 2.2 and Chain-graph could not significantly outperform our framework. P -values ≥ 0.05 are marked with 'n.s.'.

$$\theta_{ir} = \frac{\mathcal{N}(m_i - m_{ir}|0, \sigma)}{\max_{s \in \{1, \dots, n_i\}} \mathcal{N}(m_i - m_{is}|0, \sigma)},$$

where m_i is the neutral exact mass of the measured ion calculated from the precursor m/z using the ground truth adduct, here either $[M + H]^+$ (positive) or $[M - H]^-$ (negative), and m_{ir} is the exact mass of candidate r associated with MS feature i , and $\sigma = \frac{\text{ppm} \cdot m_i}{2 \cdot 10^6}$ variance of Gaussian noise model. ppm expresses the MS-device accuracy, which we set to ppm=5.

4 Results

4.1 Parameters of our framework

This section investigates the influence of different settings for framework, such as number of random spanning trees or the marginal type.

4.1.1 Number of random spanning-trees and marginal type

Figure 2 shows the top- k accuracy as a function of the number of random spanning trees L averaged across the datasets and ionizations. The identification performance increases for larger L , whereby the improvement per tree decreases. For the top-1 performance remains similar for $L \geq 16$ trees. However, for top-20, we observe improvements till $L = 128$. Figure 2 also shows that the max-marginal approach performs slightly better than the sum-margin. An explanation could be that max-marginal is more robust against candidate configurations \mathbf{z} with very low probability. The sum-marginal averages over such cases, whereas the max-marginal only includes the one with maximum probability.

Table 4. Top- k accuracies averaged across all datasets for two MS²-scorers

MS ² -scorers	Method	Top-1	Top-5	Top-10	Top-20
MetFrag	MS ² + RT	21.3	52.9	64.0	74.3
	Only MS ²	16.7	49.5	60.4	70.6
IOKR	MS ² + RT	26.7	52.1	62.5	70.3
	Only MS ²	25.1	49.5	60.3	67.6

4.1.2 Comparison of the edge potential functions

The average metabolite identification performance does not differ much between two edge potential functions (see Supplementary Table S1). This is interesting specifically for the Step-function, which uses the predicted retention orders in a binary fashion only. However, the Sigmoid function still can significantly outperform the Step-function for top-1 and top-5 accuracy.

4.2 Performance of our score integration framework

Here, we compare our score-integration framework with other methods and evaluate it under different data setups. We use $L = 128$ with max-marginals and the Sigmoid as edge potential function for the experiments.

4.2.1 Comparison to other approaches

In Table 2, we compare the performance of our score-integration framework with other approaches from the literature that utilize RT information for metabolite identification. It can be seen that our framework performs well across all datasets and ionization modes and we reach significant improvements over the baseline (Only MS²). Especially for the positive mode spectra, our method seems to have an advantage, as both competing approaches, cannot consistently improve the identification by including RT information. The least performance improvement of our approach can be observed for the negative CASMI dataset, which might be due to the small training set. The other approaches, MetFrag 2.2 and Chain-graph, can consistently (top-1, 5, 10 and 20) improve the results only on particular (dataset, ionization mode) combinations. However, they almost always increase top-1 performance. The results in Table 3 show that our framework significantly outperforms MetFrag 2.2 and Chain-graph in terms of identification performance.

4.2.2 Influence of MS² scoring method

Table 4 shows the performance using of our score-integration framework for two difference MS² scoring methods, MetFrag and IOKR (Section 3.3). Retention order information (MS² + RT) can

Table 5. Top-*k* accuracies averaged on the CASMI data (pos. & neg.) using either MetFrag or IOKR as MS²-scorer for two different candidate sets: ‘All’ molecules queried using a mass window; only those with ‘correct molecular formula’

Candidate Set	Method	MetFrag				IOKR			
		Top-1	Top-5	Top-10	Top-20	Top-1	Top-5	Top-10	Top-20
All	MS ² + RT	14.6 (***)	44.0 (***)	54.6 (***)	66.5 (***)	26.0 (***)	48.0 (***)	60.0 (***)	69.1 (***)
	Only MS ²	11.4	40.7	51.2	63.2	24.4	46.0	58.4	65.5
Correct MF	MS ² + RT	17.7 (***)	48.4 (***)	59.8 (***)	71.0 (***)	30.6	52.3	66.2 (***)	75.1 (*)
	Only MS ²	13.1	46.0	56.9	68.7	30.6	53.9	65.3	74.8

Note: The stars (*) represent the significant improvement over the Only MS² (see Table 2 for details on the significance test).

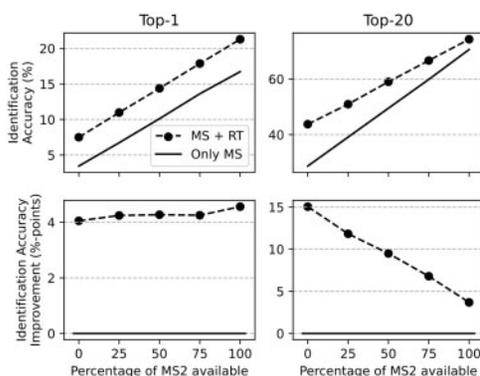


Fig. 3. Top-*k* accuracies and improvements averaged across all datasets. Plots for different percentages of available MS² spectra: 0% (only MS¹ and RT) to 100% of MS² spectra (previous experiments)

improve the identification performance in both cases; however the improvement with IOKR scores is lower.

4.2.3 Influence of the candidate set

Here, we study the effect of two commonly used ways of defining the candidate lists of molecular structures: first approach (‘All’) includes all molecules with a matching exact mass to the list, and the second approach (‘Correct MF’) only includes molecular structures matching the pre-determined MF [e.g. SIRIUS (Dührkop et al., 2019) uses this approach]. To determine the effect of the candidate set definition on our framework, we modify the CASMI dataset, such that for a spectrum *i* only candidates are used that have the same MFs as the correct structures. This leads to significantly reduced candidate sets: For the positive mode spectra, the median number of candidates decreases from 919 to 238 and for the negative ones from 420 to 58. For the Massbank data, we cannot do this modification, as the candidates are already restricted to structures with the correct MF. Table 5 shows that the baseline performance using MetFrag MS² scores (Only MS²) improves after filtering of the candidates. Further improvement can be reached by using retention order information (MS² + RT) even though the absolute improvement is slightly lower than without candidate filtering. For IOKR, we notice that RT information significantly improves the top-*k* accuracies when all matching exact mass candidates are used, whereas when the candidate sets only contain molecules with the correct, i.e. ground truth, MF, using RT information can only improve top-10 and top-20 accuracies.

4.3 Missing MS²

In Figure 3, we show the identification accuracy using our score-integration framework compared to the baseline (Only MS) when only some percentage of the MS features has an MS² spectrum. The features without spectra only use the precursor mass as MS information (Section 3.6). We vary the percentage from 0% to 100% with 25%-point steps. The retention order weight *D* was optimized using the 100% setting. At 0%, the score-integration framework only uses the mass of the candidates and their predicted retention order for

the ranking. In the absence of MS² information, we observe a high performance gain for top-20. The more MS² information we add, the smaller the gain in top-20 accuracy using the retention orders. The fact that RT is a weaker information than MS² could explain this observation. The more MS² are available, the less additional information RT can add. For the top-1, there is constant improvement for all MS²-%s.

5 Discussion

In this article, we have put forward a rigorous probabilistic framework for the integration of MS-based candidate structure and retention order predictions. Our framework allows the use of any of the popular models, such as CSI: FingerID, IOKR or MetFrag for scoring candidate structures on MS data.

Our method takes into account the retention orders of all candidate structure pairs in distinct candidate lists through an approximated fully connected MRF model. It generally achieves higher quality structural annotations of observed MS features than using a single Markov chain as implied in the Bach et al. (2018) model. It also improves on the method of Ruttkies et al. (2016), which uses predicted RTs, in three out of four experiments. For the latter approach, we believe using the RankSVM scores instead of the predicted LogP values could improve the performance. Both measures are proxies for retention behaviour and our results show that the RankSVM predicts the retention order more accurately than the LogP values (see Supplementary Table S2). We also demonstrate that our framework improves the identifications, if only a subset of the MS features come with an MS² spectrum. The framework is computationally efficient, e.g. ranking the candidates for a set of *N* = 75 MS features takes <4 min (see Section S.5), and can be trained using modest-sized datasets.

The amount of improvement using RT information was shown to depend on the dataset and MS² scorer (here MetFrag or IOKR). This indicates that RT information rather fine tunes the ranking given by the MS² scorer, e.g. by better tie-breaking. The underlying factors could be ambiguities in the candidate sets that can be only be resolved by RT or molecular features that cannot be predicted by MS. Stereochemistry is an obvious factor, but annotations of stereochemistry are not always provided for the RT databases limiting the use of this information for training better retention order prediction models. Thus improved modelling of stereochemistry features is an important open problem (Witting and Böcker, 2020). A further research direction could be to include the LC system’s configuration, e.g. column or eluent, into the retention order prediction. As LC systems can be configured to separate certain molecular classes, this could provide additional information to certain molecular candidates. Also, using the LC peak shape to train a model directly predicting the retention order probabilities could be more accurate, e.g. by incorporating RT variance. However, such data are currently not part of the public RT databases.

Acknowledgements

E.B. likes to thank Emma Schymanski for answering numerous questions on mass spectrometry and Sandor Szedmak for his helpful comments on the

computational framework. All authors thank Justin J.J. van der Hoof for critical reading and corrections.

Funding

This work was supported by the Academy of Finland [grant 310107 (MACOME)]; and the Aalto Science-IT infrastructure. S.R. and J.H.W. were supported by the Engineering and Physical Sciences Research Council [EP/R018634/1]. This work started during a visit by J.R. to S.R., funded through the Scottish Informatics and Computing Science Alliance (SICSA) distinguished visiting fellow scheme.

Conflict of Interest: none declared.

References

- Aksenov, A.A. *et al.* (2017) Global chemical analysis of biology by mass spectrometry. *Nat. Rev. Chem.*, **1**, 0054.
- Allen, F. *et al.* (2014) CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. *Nucleic Acids Res.*, **42**, W94–W99.
- Bach, E. *et al.* (2018) Liquid-chromatography retention order prediction for metabolite identification. *Bioinformatics*, **34**, i875–i883.
- Blaženović, I. *et al.* (2018) Software tools and approaches for compound identification of LC-MS/MS data in metabolomics. *Metabolites*, **8**, 31.
- Brouard, C. *et al.* (2016) Fast metabolite identification with Input Output Kernel Regression. *Bioinformatics*, **32**, i28–i36.
- Brouard, C. *et al.* (2019) Improved small molecule identification through learning combinations of kernel regression models. *Metabolites*, **9**, 160.
- da Silva, R.R. *et al.* (2015) Illuminating the dark matter in metabolomics. *Proc. Natl. Acad. Sci. USA*, **112**, 12549–12550.
- Del Carratore, F. *et al.* (2019) Integrated probabilistic annotation (IPA): a Bayesian-based annotation method for metabolomic profiles integrating biochemical connections, isotope patterns and adduct relationships. *Anal. Chem.*, **91**, 12799–12807.
- Domingo-Almenara, X. *et al.* (2019) The METLIN small molecule dataset for machine learning-based retention time prediction. *Nat. Commun.*, **10**, 1–9.
- Dührkop, K. *et al.* (2015) Searching molecular structure databases with tandem mass spectra using CSI: FingerID. *Proc. Natl. Acad. Sci. USA*, **112**, 12580–12585.
- Dührkop, K. *et al.* (2019) SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods*, **16**, 299–302.
- Horai, H. *et al.* (2010) MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.*, **45**, 703–714.
- Hu, M. *et al.* (2018) Performance of combined fragmentation and retention prediction for the identification of organic micropollutants by LC-HRMS. *Anal. Bioanal. Chem.*, **410**, 1931–1941.
- Lin, H.-T. *et al.* (2007) A note on Platt's probabilistic outputs for support vector machines. *Mach. Learn.*, **68**, 267–276.
- Liu, J.J. *et al.* (2019) Quantitative structure–retention relationships with non-linear programming for prediction of chromatographic elution order. *Int. J. Mol. Sci.*, **20**, 3443.
- MacKay, D.J. (2005) *Information Theory, Inference and Learning Algorithms*. Cambridge University Press. Cambridge.
- Marchand, M. *et al.* (2014) Multilabel structured output learning with random spanning trees of max-margin Markov networks. In: *NIPS*, Palais des Congrès de Montréal, Montréal Canada, pp. 873–881.
- Nguyen, D.H. *et al.* (2018a) Recent advances and prospects of computational methods for metabolite identification: a review with emphasis on machine learning approaches. *Brief. Bioinform.*, **20**, 2028–2043.
- Nguyen, D.H. *et al.* (2018b) Simple: sparse interaction model over peaks of molecules for fast, interpretable metabolite identification from tandem mass spectra. *Bioinformatics*, **34**, i323–i332.
- Nguyen, D.H. *et al.* (2019) ADAPTIVE: leArning DAta-dePendentT, concIse molecular Vectors for fast, accurate metabolite identification from tandem mass spectra. *Bioinformatics*, **35**, i164–i172.
- Pence, H. and Williams, A. (2010) ChemSpider: an online chemical information resource. *J. Chem. Educ.*, **87**, 1123–1124.
- Plante, P.-L. *et al.* (2019) Predicting ion mobility collision cross-sections using a deep neural network: DeepCCS. *Anal. Chem.*, **91**, 5191–5199.
- Platt, J. (2000) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Alexander, J.S. (eds) *Advances in Large Margin Classifiers*. MIT Press, pp. 61–74.
- Pletscher, P. *et al.* (2009) Spanning tree approximations for conditional random fields. *PMLR*, **5**, 408–415.
- Ralaivola, L. *et al.* (2005) Graph kernels for chemical informatics. *Neural Netw.*, **18**, 1093–1110.
- Ruttkies, C. *et al.* (2016) MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J. Cheminform.*, **8**, 3.
- Ruttkies, C. *et al.* (2019) Improving MetFrag with statistical learning of fragment annotations. *BMC Bioinformatics*, **20**, 376.
- Rutz, A. *et al.* (2019) Taxonomically informed scoring enhances confidence in natural products annotation. *Front. Plant Sci.*, **10**, 1329.
- Samaraweera, M.A. *et al.* (2018) Evaluation of an artificial neural network retention index model for chemical structure identification in nontargeted metabolomics. *Anal. Chem.*, **90**, 12752–12760.
- Schymanski, E.L. *et al.* (2017) Critical assessment of small molecule identification 2016: automated methods. *J. Cheminform.*, **9**, 22.
- Stanstrup, J. *et al.* (2015) PredRet: prediction of retention time by direct mapping between multiple chromatographic systems. *Anal. Chem.*, **87**, 9421–9428.
- Su, H. and Rousu, J. (2015) Multilabel classification through random graph ensembles. *Mach. Learn.*, **99**, 231–256.
- Wainwright, M.J. *et al.* (2005) Map estimation via agreement on trees: message-passing and linear programming. *IEEE Trans. Inf. Theory*, **51**, 3697–3717.
- Wang, M. *et al.* (2016) Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nat. Biotechnol.*, **34**, 828–837.
- Willighagen, E.L. *et al.* (2017) The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J. Cheminform.*, **9**, 33.
- Witting, M. and Böcker, S. (2020) Current status of retention time prediction in metabolite identification. *J. Sep. Sci.*, **43**, 1746–1754.
- Xiao, J.F. *et al.* (2012) Metabolite identification and quantitation in LC-MS/MS-based metabolomics. *Trends Analyt. Chem.*, **32**, 1–14.