analytical chemistry

Article

pubs.acs.org/ac

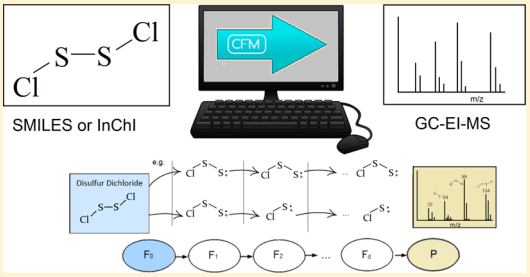
Computational Prediction of Electron Ionization Mass Spectra to Assist in GC/MS Compound Identification

Felicity Allen,* Allison Pon, Russ Greiner, and David Wishart

Department of Computing Science, University of Alberta, Edmonton T6G 2E8, Canada

Supporting Information

ABSTRACT: We describe a tool, competitive fragmentation modeling for electron ionization (CFM-EI) that, given a chemical structure (e.g., in SMILES or InChI format), computationally predicts an electron ionization mass spectrum (EI-MS) (i.e., the type of mass spectrum commonly generated by gas chromatography mass spectrometry). The predicted spectra produced by this tool can be used for putative compound identification, complementing measured spectra in reference databases by expanding the range of compounds able to be considered when availability of measured spectra is limited. The tool extends CFM-ESI, a recently developed method for computational prediction of electrospray tandem mass spectra (ESI-MS/MS), but unlike CFM-ESI, CFM-EI can handle odd-electron ions and isotopes and incorporates an artificial neural network. Tests on EI-MS data from the NIST database demonstrate that CFM-EI is able to model fragmentation likelihoods in low-resolution EI-MS data, producing predicted spectra whose dot product scores are significantly better than full enumeration "bar-code" spectra. CFM-EI also outperformed previously reported results for MetFrag, MOLGEN-MS, and Mass Frontier on one compound identification task. It also outperformed MetFrag in a range of other compound identification tasks involving a much larger data set, containing both derivatized and nonderivatized compounds. While replicate EI-MS measurements of chemical standards are still a more accurate point of comparison, CFM-EI's predictions provide a much-needed alternative when no reference standard is available for measurement. CFM-EI is available at https://sourceforge.net/projects/cfm-id/ for download and http://cfmid.wishartlab.com as a web service.



Gas chromatography/mass spectrometry (GC/MS) is widely used in analytical chemistry to first separate compounds using gas chromatography then fragment and identify their components via electron ionization-mass spectrometry (EI-MS). Putative identification of compounds using EI-MS commonly involves searching for a closely matching mass spectrum within a database of previously collected EI-MS reference spectra,[1−3] often the Wiley Registry of Mass Spectral Data[4] or the NIST/NIH/EPA MS database.[5] In cases where the query molecule is contained within the reference database, the resulting accuracy levels are quite good.[2] However, the main drawback to this approach is that the reference database often does not contain a reference spectrum for the target compound. This is particularly problematic in the field of metabolomics, where most of the compounds of interest are not in any existing spectral database. For example, only 4023 out of >40000 compounds in the Human Metabalome Database (HMDB)[6] have spectra of any kind recorded in either HMDB or in the NIST/NIH/EPA MS database. Indeed, even the NIST database, with >240000 compounds, has surprisingly few EI-MS spectra for natural products or chemicals of biological interest. Only 12525 compounds in the NIST database have a matching entry in either HMDB, the Chemical Entities of Biological Interest Database (ChEBI),[7] or our own private list of more than 190000 plant-derived compounds. Consequently, finding alternative means for identifying metabolites for which no measured reference spectra are available is particularly important in metabolomics, where GC/MS (EI-MS) techniques are frequently used. The importance of addressing this challenge has been highlighted in a number of recent reviews, including those by Kind and Fiehn[8] and Scheubert et al.[9]

Predicting EI-MS spectra from structure is not a new idea. In fact, it was one of the first problems to be tackled by the emerging field of artificial intelligence in the 1960's. Investigators working on the Dendral project[10] separated the overall problem into three main steps, which they labeled "plan", "generate", and "test". The "plan" step involved narrowing the chemical search space by extracting structural information directly from the target mass spectrum. A range of machine-learning methods were proposed to address this step, most of which were aimed at identifying likely substructures of the target molecule.[11−18] This is the approach that is routinely applied as part of the NIST 2014/EPA/NIH MS Search.[17] The "generate" step generates candidate chemical structures from within that refined search space. Algorithms for exhaustively generating structural isomers subject to various constraints[19−21] largely solved this problem. However, such exhaustive enumeration often results in more candidate compounds than

Received: April 25, 2016
Accepted: July 6, 2016
Published: July 6, 2016

DOI: 10.1021/acs.analchem.6b01622
Anal. Chem. 2016, 88, 7689−7697

can be processed with current compute resources. So these algorithms are often now replaced by queries to large public chemical databases (e.g., HMDB[6] and PubChem[22]), which return a smaller subset of candidates. The subsequent "test" step, in which the obtained candidates are ranked according to whether they would be expected to produce the target spectrum, is the focus of this work. This involves generating fragmentation events that can explain the peaks in the target spectrum, and also, importantly, determining a score for each candidate based on the likelihood or quality of those explanations.

Dendral[10] proposed an "expert system" to specify how a molecule would fragment, which involved the collation of many user-defined rules. For instance, if the molecule was identified as a ketone then it would be subject to a McLafferty rearrangement.[23] Several commercial packages now exist that use a similar rule-based approach. These include Mass Frontier (Thermo Scientific, www.thermoscientific.com) and MS Fragmenter (ACD Laboratories, www.acdlabs.com), which contain thousands of manually curated rules to predict fragmentations. MOLGEN-MS[24] is another commercial program that also applies rule-based fragmentations. These programs all produce so-called "bar-code" spectra, in which all predicted peaks are of equal height.

As more rules have been added to these methods, they have been able to predict more fragmentations for any given molecule. This allows such rule-based systems to achieve near-perfect recall (i.e., they can provide an explanation for almost every peak in a target spectrum). In general, these systems achieve improved precision (i.e., a higher percentage of the predicted fragments actually occur) by leaving out some of the rules.[25] The difficulty comes when deciding which rules to leave out.

Rather than relying on a large library of fragmentation rules, another class of algorithms has emerged that apply combinatorial fragmentation procedures. Mainly developed for ESI-MS/MS, where rule-based methods are less established, these algorithms enumerate all possible fragments of a chemical structure by systematically and recursively breaking all bonds[26−29] or by enumerating all connected substructures of the input molecule.[30,31] Like the rule-based methods, these methods are capable of generating large numbers of fragments and often achieve near-perfect recall. In an effort to combat the associated precision problem, they typically employ various heuristics in their scoring protocols. For example, MetFrag[28] uses an estimate of the energy of each broken bond, combined with a bonus if the neutral loss formed is one of a common subset. While these heuristics certainly help to alleviate the precision problem, this paper shows that there are ways to improve on this.

Several projects have attempted to estimate the likelihood of a given fragmentation event from data. As far back as the 1990s, Gasteiger et al. used logistic regression[32] and neural networks[33] to predict fragmentation probabilities for α-cleavages from hand-labeled EI-MS data. However, no implementation appears to have survived.

More recently, Kangas et al.[34] proposed a machine learning approach to obtain bond dissociation energies for lipids. Their method uses a neural net within a kinetic Monte Carlo simulation, trained using a genetic algorithm on ESI-MS/MS data. However, this method has not yet been applied to general classes of metabolites, besides lipids, nor to EI-MS data.

Quantum chemical and molecular dynamics methods have also been applied to this problem with some success.[35−37] However, the computational demands of these methods is exceedingly high (several thousand CPU hours per molecule).

Competitive Fragmentation Modeling (CFM)[38] is a method for mass spectrum prediction that was recently developed in the context of ESI-MS/MS, in an attempt to improve the precision of combinatorial methods. From here onward, we denote the original ESI-MS/MS version of this method as CFM-ESI. It uses a probabilistic, generative model to predict both the mass and intensity values of peaks in the spectrum of a given molecule. The method was shown to be effective in modeling the relative likelihoods of fragmentation events, producing spectra with significantly improved Jaccard scores over full enumeration bar-code spectra. It was also shown to translate well to the problem of metabolite identification, outperforming existing methods MetFrag and FingerID,[18] at the time of testing. Another method, CSI:FingerID, has since been reported to achieve better performance than CFM-ESI on a different ESI-MS/MS identification task;[39] however, it is not applicable to EI-MS.

In this paper, we propose several modifications to CFM-ESI to make it applicable to the EI-MS spectra typically generated by GC/MS instruments and report the results of extensive empirical testing of the method on EI-MS data.

EXPERIMENTAL SECTION

In this section, we first provide a brief outline of the CFM-ESI method, and then we describe the proposed modifications to make it applicable to the EI-MS spectra, and finally we provide details of the empirical testing we carried out.

Competitive Fragmentation Modeling (CFM-ESI). CFM-ESI uses a probabilistic, generative model for the fragmentation processes occurring within a mass spectrometer. The model is a fixed-length, stochastic Markov process of transitions between discrete fragment states F_0, F_1, ..., F_d (see Figure 1) that each take values from the set containing all possible fragments. The state space of possible fragments is enumerated using a combinatorial approach based on systematic bond disconnection.[26] This involves breaking every bond in the molecule, and every pair of bonds in each ring, in turn, and considering all hydrogen rearrangements within each pair of resulting fragments.

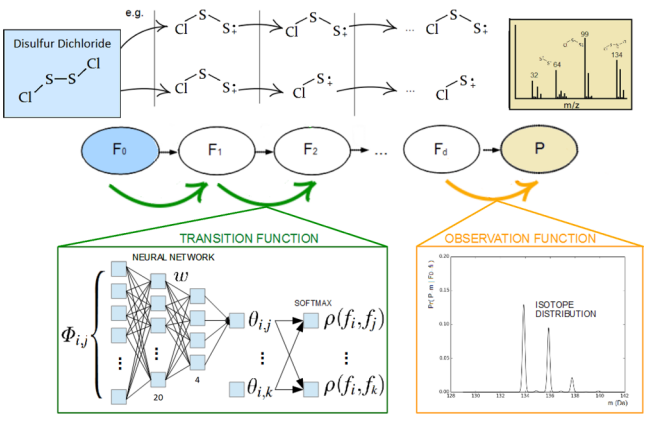


Figure 1. Schematic of CFM-EI showing possible sequences of fragments leading to a spectral peak. A neural network is included within the transition function. Extensions to handle isotopes are included in the observation function.

DOI: 10.1021/acs.analchem.6b01622
Anal. Chem. 2016, 88, 7689−7697

The transition from any fragment f_i to a possible child fragment f_j is assigned a break tendency value $\theta_{i,j} \in \mathbb{R}$. This is defined as $\theta_{i,j} := w^T \Phi_{i,j}$, where $\Phi_{i,j}$ is a vector of chemical features describing the possible transition (e.g., the atoms that exist on either side of the broken bond) and w is a vector of parameters corresponding to those features. The probability that f_i transitions to f_j at a single time step is then defined as

$$\Pr(f_j|f_i) \approx \rho(f_i, f_j) = \begin{cases} \frac{\exp \theta_{i,j}}{1 + \sum_k \exp \theta_{i,k}} : f_i \neq f_j \\ \frac{1}{1 + \sum_k \exp \theta_{i,k}} : f_i = f_j \end{cases} \quad (1)$$

where this softmax function is used to model the competition between different fragmentation events originating from the same parent ion f_i and ensures that all probabilities sum to one. The second part of eq 1 indicates that self-transitions are allowed and are effectively assigned a break tendency value of $\theta_{i,i} = 0$.

The probability of producing a peak at mass m is modeled by a real-valued, random variable P. The conditional $P = m|F_d$ is modeled as a Gaussian with variance σ determined by the mass tolerance of the instrument and mean given by the mass of F_d:

$$\Pr(P = m|F_d) \approx g(m, F_d; \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{m - \text{mass}(F_d)}{\sigma}\right)^2\right\} \quad (2)$$

The predicted spectrum is computed as the marginal of P conditioned on the input molecule F_0 (see Figure 1) (i.e.,

$$\Pr(P = m|F_0; w) \approx \sum_{F_1} \cdots \sum_{F_d} \rho(F_0, F_1) \cdots \rho(F_{d-1}, F_d) g(m, F_d; \sigma) \quad (3)$$

Parameters for the complete model are estimated from training data using a maximum likelihood approach. Full details can be found in Allen et al.[38] and Allen.[40]

From CFM-ESI to CFM-EI. In order to make CFM applicable to EI-MS, modifications were required to account for (1) odd-electron ions and (2) isotopes, as detailed in the following sections. We also explored (3) the incorporation of an artificial neural network within the CFM transition function. This latter modification was not specific to EI-MS but was tested within that context. A schematic of CFM-EI, showing the isotope and neural network modifications is provided in Figure 1.

Odd Electron Ions. The even-electron rule[23,41] is a rule-of-thumb that applies to the vast majority of ions fragmenting in a mass spectrometer. It states that even-electron ions can only produce even-electron fragments, whereas odd-electron ions can produce either odd- or even-electron fragments. In ESI-MS/MS the precursor is even, so only even-electron ions can occur. However, as the precursor in EI-MS is odd, both odd and even-electron ions occur.

In CFM-ESI, when enumerating the fragment state space, two possibilities are generated for each break (and any associated hydrogen rearrangements), by assigning the charge to either of the two resulting fragments (i.e., to determine which becomes the ion, and which the neutral loss). In CFM-EI, when fragmenting odd-electron ions, rather than generating just two possibilities, we generate four possibilities (i.e., all combinations of which side includes the charge and which side the radical).

Isotopes. Isotope peaks are often absent in ESI-MS/MS due to the use of narrow isolation widths around the precursor mass. In contrast, isotope peaks are common in EI-MS and can be useful in deciding between alternative explanations for the same peak. This is because some fragments may have the same monoisotopic mass but different expected isotopic distributions. MOLGEN-MS[24] makes use of isotope information in its scoring function.

Isotopic peaks can be incorporated quite naturally within the CFM observation model. While CFM-ESI[38] used a Gaussian observation function (see eq 2), CFM-EI instead uses a weighted sum of Gaussians corresponding to the peaks in the fragment's expected isotope spectrum. Denoting the expected isotope spectrum for fragment f_i as $\mathbb{E}(S_{f_i})$, and defining this as a set of mass and intensity pairs $\{(m',h')\}$, normalized such that $\sum_{(m',h') \in \mathbb{E}(S_{f_i})} h' = 1$, the new observation function (replacing eq 2) becomes

$$g(m, F_d; \sigma) = \sum_{(m',h') \in \mathbb{E}(S_{F_d})} \frac{h'}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{m - m'}{\sigma}\right)^2\right\} \quad (4)$$

Various algorithms[42−45] have been proposed to compute the expected isotope spectrum for a given molecular formula. We use the program emass[42] for this purpose, thresholding the result to include only isotopic peaks with normalized intensity above 0.01.

Eq 4 can be substituted into eq 3 to compute the predicted spectrum as before but now including the isotopic peaks at their expected intensities. During parameter training, the computation of the expected marginal probabilities of the fragment states becomes slightly more complicated. To address this, the different isotopes for the same fragment are effectively considered as different fragments, but their marginal probabilities are then accumulated to give the marginal of the multi-isotopic fragment. This causes the method to favor those explanatory fragments for which all isotopic peaks are present in their expected proportions. However, it also allows the method to accept such explanations when some of those peaks are missing (including the monoisotopic peak), if no better explanations are available. Further details and a simple example of this computation are provided in the Supporting Information.

Neural Network. As already noted above, CFM-ESI modeled the break tendency of a particular bond within a moelcule as $\theta_{i,j} := w^T \Phi_{i,j}$ (i.e., a linear function of the chemical features describing that break). A natural extension is to replace this linear function with a more complex function using an artificial neural network. Toward that end, we let $\theta_{i,j}$ be the output of a multilayer perceptron, for which the inputs are given by the feature vector $\Phi_{i,j}$. The parameters of the model are again denoted by w.

In order to estimate these parameters, we use Expectation Maximization (EM) as before, which maximizes the expected log-likelihood Q, but employ a modified form of the backpropagation algorithm to compute the partial gradients $\frac{\partial Q}{\partial w^l}$ in the maximization (M) step, as described in the Supporting Information. The expectation (E) step proceeds as before but uses the neural network to compute the $\theta_{i,j}$ values on each iteration.

Data Sets. We used the following three EI-MS data sets from the NIST/EPA/NIH Mass Spectral Library.[5] All data were measured at integer mass accuracy using a single energy of

DOI: 10.1021/acs.analchem.6b01622
Anal. Chem. 2016, 88, 7689−7697

70 eV. (1) Small Molecule Set (17324 molecules): this set was designed to allow rapid comparison of various CFM-EI model and parameter configurations. Molecules were randomly selected from those compounds within the main NIST library that had good spectrum quality [weighted recall of full enumeration (see Models for Comparison) spectra above 50%] and low CFM-EI spectrum prediction compute times [Fragmentation graph computation in less than 10 s (see the Supporting Information)], and no overlap with the two validation sets below. We identified 987 molecules in this set as being derivatized, due to the presence of methoxy, O-silyl, N-silyl, or S-silyl groups. The full set was randomly divided into 5 groups for use within a 5-fold cross validation framework, ensuring duplicates (determined by matching the c,h and i sections of the InChI string) were allocated to the same cross-fold group. A model was then trained on the full set, for validation with the two sets described next, with which there were no overlapping molecules. (2) Kerber Set (100 molecules): Kerber et al.[24] and Schymanski et al.[46] report the results of applying MOLGEN-MS, Mass Frontier, and MetFrag on this small data set, which was extracted from the NIST library. We include this set to compare the performance of CFM-EI against those previous results. (3) Replicate Set (20588 molecules): this set contained entries from the NIST replicate set. The initial set had 33782 molecules. We removed 296 molecules because they were not computable by CFM-EI (see Supporting Information) and another 12898 because they were duplicates (e.g., stereoisomers) of another molecule in the set. We identified 992 molecules within this set as being derivatized (D), again using the presence of methoxy and silyl groups.

Model Configuration. The CFM-EI model was configured with a fragmentation and model depth of 2. We tested both the original transition function and the extension to include a neural network. The neural network included two hidden layers, one with 20 nodes and the other with 4. Each hidden node used a rectified linear unit (reLU) activation function, with half the units assigned a negative activation function, as recommended by a number of previous publications.[47,48] The final output node was a linear unit (see Figure 1). This basic network configuration was selected with the expectation that it would provide moderate modeling ability but use an order of magnitude less parameters than the linear model.

Details of the 996 binary chemical features used to describe each break are provided in the Supporting Information. On average, only 23 of these features are "on" for any given break (as measured on the Small Molecule Set). When the original linear transition function was used, all quadratic combinations of the features were also included. When the neural network was used, the quadratic features were not included, as it was expected that the neural network would capture this information more efficiently. Features never encountered in the training data (the small molecule set) were removed. Table 1 shows the resulting number of parameters in each model. Although the models have a large number of parameters, the sparsity of the feature vector means that only a small proportion of these parameters are used for any given break, as shown in the final column of Table 1.

Table 1. Number of Parameters for Each Model

model	# parameters	av. # parameters used per break
linear (with quadratic features)	160787	288
neural net	18509	549

Spectrum Prediction. A trained CFM-EI model was used to predict a spectrum for each molecule in the training (within a cross-validation framework) and validation sets. Since the data was collected with integer mass accuracy, we combined the resulting peaks into integer bins by rounding each mass to the nearest integer and summing the intensity values of peaks within the same bin. Unlike in CFM-ESI, no further postprocessing was applied to remove low intensity peaks. This was because the NIST EI-MS spectra generally had more peaks than the ESI-MS/MS spectra used in Allen et al.[38] The use of integer mass accuracy in the NIST data also means that there are fewer possible peak locations, so fewer peaks to be potentially discarded. Instead we use the Dot Product metrics (described next) to account for the size of the peaks when scoring a spectral match.

Metrics. We used the following metrics (see Supporting Information for equations) to assess the quality of each predicted spectrum, when compared to a target reference spectrum. We used the weighted dot product metric defined by Stein and Scott,[2] in which intensities are raised to the power 0.6 and the masses to the power 3 (Stein Dot Product), as is recommended for searching against the NIST database. We also used our own reweighted version of this metric, in which both the intensities and masses are raised to the power 0.5 (Dot Product), since we were concerned that Stein's weighting may overemphasize the higher mass peaks at the expense of information contained in the lower mass peaks. We also calculated Weighted Recall and Weighted Precision scores, as defined in Allen et al.[38] (and in the Supporting Information).

Models for Comparison. CFM-EI is one of the only computational methods to predict the intensities of peaks as well as their m/z values. The other exception is the quantum chemistry-based method QCEIMS.[36] However, QCEIMS is too computationally demanding to allow signficant comparison on the data sets used here, whereas CFM-EI can typically predict an EI-MS spectrum for molecules of molecular mass less than 800 Da in less than 10 min on a single CPU (see Supporting Information). All other currently available EI-MS and ESI-MS/MS tools can be viewed as generating bar code spectra in which all (nonisotopic) peaks are the same height. Of those methods, we do not have access to commercial methods MassFrontier and MOLGEN-MS[24] but were able to use previously reported results to compare against them in a compound identification task discussed later. MetFrag[28] is open source; however, it does not produce a computationally accessible spectrum but rather focuses on computing only those peaks that match a peak in a target spectrum (i.e., in a compound identification setting). We provide comparisons with MetFrag in various metabolite identification tasks discussed later.

To assess CFM-EI's spectrum prediction performance on the Small Molecule Set, we compare against our own bar code spectra, our previous CFM-ESI model, and replicate measured spectra as follows: (1) Full Enumeration {Enum} x {Iso, −}: the predicted spectrum includes all possible fragments that could be produced from the starting molecule, both with (Iso) and without (−) isotopes, all with uniform intensity values (including the isotope peaks). (2) Measured: (replicate set only) this model uses the measured spectrum from the main NIST library for the corresponding molecule in place of the predicted spectrum. Since these spectra are measured rather than predicted, this provides an upper limit for the best possible

DOI: 10.1021/acs.analchem.6b01622
Anal. Chem. 2016, 88, 7689−7697

predictions. Remeasurement variability should be the main reason that these spectra are not a perfect match for the target spectrum. (3) CFM EI Models {NN, Lin} x {Iso, −}: we consider four configurations of the CFM-EI model. These are all combinations of the following: with (Iso) versus without (−) isotopes; and with (NN) versus without (Lin) the neural network extensions. (4) Old CFM ESI Models {ESI} x {Iso, −}: Our CFM-ESI model, trained on ESI-MS/MS data as described in Allen et al.,[38] but modified to allow for odd-electron ionization, and applied with (Iso) and without (−) isotopes. The resulting predicted spectra (at three energy levels) were quantified at unit mass resolution and added together. For the tests on the Replicate Set, we applied only 3 models: the NN-Iso CFM-EI model, since it performed best of all CFM models on the Small Molecule Set (see Results and Discussion); the Enum-Iso model, to show the equivalent "bar-coded" result; and the measured spectra.

Metabolite Identification. Next we applied our EI-MS spectrum predictions to a series of metabolite identification tasks. For each target molecule in each test set, we first obtained a set of possible candidate molecules for the target (as detailed next) and then generated the predicted MS spectrum for each candidate. We then ranked the candidates based on how closely their predicted spectrum matched the target spectrum. We considered all four spectrum prediction metrics used above to compare the target and predicted spectra. In GC/MS studies of small molecules, the molecules may be either derivatized (with trimethylsilane) or may be left underivatized. Consequently, we obtained or prepared both derivatized and underivatized sets of candidate chemical structures.

Candidate Selection. The candidate sets were produced using the methods listed below. In all cases, we removed molecules that were uncomputable by CFM-EI (see Supporting Information) or could not be computed by CFM-EI in reasonable time (less than 10 min per compound for candidate sets of size greater than 100, or less than 1 h per compound for candidate sets of size less than 100, to keep total compute time down) (see Supporting Information). For easier comparison with MetFrag, we also removed those compounds that MetFrag was unable to process. Both CFM-EI and MetFrag were evaluated only on the subset of candidates that could be processed by both programs. Further details of the numbers of molecules for which the query included the correct molecule (since some were not found in the databases), and the median numbers of candidates in each set, are provided with the results in Table 2. (1) HMDB: we queried HMDB (Human Metabalome Database)[6] for all molecules within 0.5 Da of the known molecular mass of the target compound. This simulates the case where the molecule is thought to be a naturally occurring metabolite, but there is some uncertainty in the target mass range. (2) PubChem: we queried the PubChem compound database[22] for all molecules within 10 ppm of the known molecule mass of the target compound. This simulates the case where little is known about the candidate compound, but the parent ion mass is known with good accuracy (via high mass accuracy MS). (3) dHMDB and dPubChem: we used the derivitization tool,[31] provided as part of MetFrag, to produce derivatized variations for all entitites in HMDB that could be derivatized. We allowed replacement of up to 8 carbonyl, amino, and thiol groups, and set the maximum mass limit to 800 Da. This resulted in a total of 37403 derivatized entities. The derivatized version of PubChem, produced for Ruttkies et al.,[31] was made available to us by the authors. The two derivatized databases were queried for molecules within 0.5 Da (for dHMDB) and 10 ppm (for dPubChem) of the known molecular mass. This simulates the case where derivatization has been carried out, and so it makes sense to search among only derivatized compounds. (4) MOLGEN: for the Kerber Set, to compare with previously published results in Kerber et al.,[24] we used candidate sets of all possible isomers for each molecule as generated by MOLGEN and made available in the supporting information of Schymanski et al.[25] Using all structural isomers like this is a very extreme test case, and since the number of structural isomers grows at least exponentially with molecular size, it is only possible for test molecules such as these with low molecular masses. (5) NIST: for comparison with the case where you have a reference database of measured spectra (rather than computationally predicted spectra), we used the entire main library of the NIST EI-MS database as a candidate set.

Methods for Comparison. Using the Small Molecule Set, we compared the ranking performance of two CFM-EI models (NN-Iso and Lin-Iso) and one CFM-ESI model (ESI-Iso) when querying PubChem, to see whether better prediction performance translated to better identification performance. We also assessed the differences in identification performance obtained using each of the four spectrum prediction metrics to rank candidates.

On the other two validation sets, we compared the ranking performance of the best performing CFM-EI model (NN-Iso with Dot Product), against that of MetFrag,[28] and where possible, MOLGEN-MS[24] and MassFrontier (using the results in Schymanski et al.[46]).

MetFrag was run using the recent update MetFrag2.2 CL[49] using FragmenterScore only (i.e., no use of patent or reference counts). Both CFM-EI and MetFrag used an absolute mass tolerance of 0.5 Da to determine matching peaks. Example configuration files used for both programs are available in the Supporting Information.

We also compared CFM-EI's performance to that achievable when measured spectra are available for all candidate compounds, by querying the Replicate Set against the NIST candidate set. For the measured spectra, we used Stein's Dot Product to compare spectra, and thus rank candidates, as recommended by Stein and Scott.[2] For CFM-EI, we report results using both Stein's Dot Product and our own Dot Product.

Metrics. We considered both absolute rankings and relative rankings within each candidate set. In the case of the former, we dealt with tied scores by taking the expected average ranking given a uniform distribution over those candidates with equal scores. For the latter, we employed the RRP score used in Kerber et al.[24] and Schymanski et al.,[25] which is defined as

$$\mathrm{RRP} = \frac{1}{2}\left(1 + \frac{\mathrm{BC} - \mathrm{WC}}{\mathrm{TC} - 1}\right) \quad (5)$$

where BC is the number of candidates with better scores, WC is the number of candidates with worse scores, and TC is the total number of candidates. A value of 0.0 indicates perfect identification, whereas 0.5 indicates that performance is no better than random.

DOI: 10.1021/acs.analchem.6b01622
Anal. Chem. 2016, 88, 7689−7697

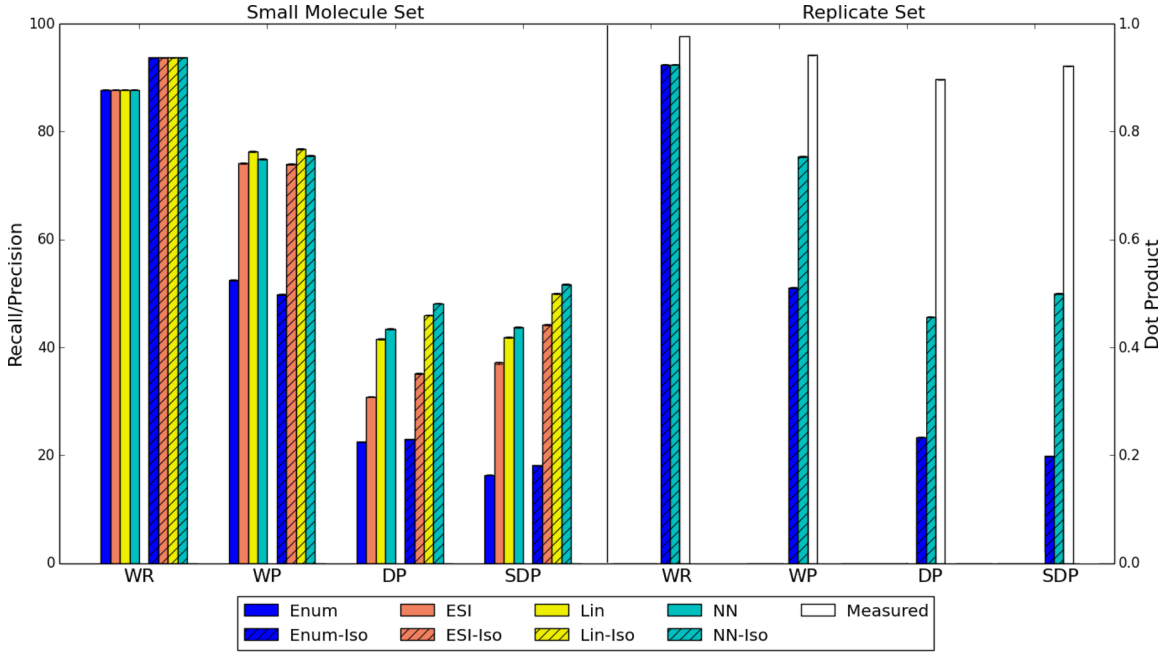


Figure 2. Spectrum prediction results for the Small Molecule Set (left) and Replicate Set (right). The *x* axis shows the four metrics: Weighted Recall (WR), Weighted Precision (WP), Dot Product (DP), and Stein Dot Product (SDP). Enum and Enum-Iso use full enumeration bar code spectra; ESI and ESI-Iso are CFM-ESI models; Lin, Lin-Iso, NN, and NN-Iso are all CFM-EI models; and measured uses replicate measured spectra. Bars display means ± standard error (too small to see). For all metrics, larger values are better.

RESULTS AND DISCUSSION

In this section, we present the results for the tests of spectrum prediction performance, followed by the metabolite identification results.

Spectrum Prediction. The spectrum prediction results are presented in Figure 2. The values obtained are very similar for the two molecule sets, suggesting that we have not overfitted to the Small Molecule Set, nor lost much accuracy in selecting less computationally intensive molecules for training the models.

For both molecule sets, the high Weighted Recall values suggest that most of the peaks in each spectrum can be explained by a fragmentation event generated by CFM-EI. There are a small proportion of possible fragmentation events that CFM-EI cannot explain, as can be seen in the higher Weighted Recall scores for the measured spectra on the Replicate Set. These may include non-hydrogen rearrangements and fragmentation events requiring depths greater than 2. Note that the Weighted Recall metric is independent of the predicted intensity values, and so since there is no postprocessing to remove low intensity peaks, this metric always assigns the same scores to CFM-EI, CFM-ESI, and the full enumeration.

The other three metrics show that, when taking the predicted intensities into account, CFM-EI significantly outperforms the full enumeration models. This demonstrates that it is able to differentiate between likely and unlikely fragmentations.

Since the Dot Product scores incorporate the intensities of both the measured and predicted spectra, they are good metrics for how well each model predicts the spectrum. Using either the Stein Dot Product or the Dot Product metric and looking at the cross-validation results on the Small Molecule Set, we see that the best performing CFM-EI model uses both the isotope and neural network extensions. Although the performance of the neural network model is only a little better than the linear model, it achieves this using far fewer model parameters and so is a more efficient representation. The new models, specifically trained on EI-MS data also outperform the old CFM-ESI model, which was trained on ESI-MS/MS data.

On the Replicate Set, the comparison with the remeasured spectra shows that CFM-EI still falls short of providing a spectrum that is as reliable as those produced by physically measuring the spectrum. This is not unexpected, and shows that computational methods still have room for improvement.

Metabolite Identification. The results for the Small Molecule Set, when querying PubChem for candidates, are shown in Figure 3. When ranking candidates using the weighted recall scores, we see that the performance is no better than random. This is equivalent to using a full enumeration spectrum for matching and is similar to the match value scoring used in Kerber et al.[24] that Schymanski et al.[25] showed was not effective. The main difference here is the details of the full enumeration.

The best result (RRP = 0.0882) was achieved when ranking candidates using the Dot Product metric, demonstrating that our predicted intensity values help rank candidates correctly. The performance using the NN-Iso model was better than that obtained using either the Lin-Iso or ESI-Iso model, showing that in these cases at least, better prediction performance translated to better identification performance.

The RRP results for validation testing with the Kerber and Replicate sets are presented in Table 2. Standard error values were all less than 0.01 for tests on the Kerber data set and less than 0.001 for tests on the other data sets. On the Kerber Set, CFM-EI outperforms MassFrontier, MOLGEN-MS and MetFrag. The RPP score achieved is 0.199, which means that nearly 20% of candidates score better than the correct candidate. However, one should note that this is a very extreme test case, in which the comparison is between a large number of very similar molecules, and this result is substantially better than any previously reported on this set.[24,25,46]

DOI: 10.1021/acs.analchem.6b01622
Anal. Chem. 2016, 88, 7689−7697

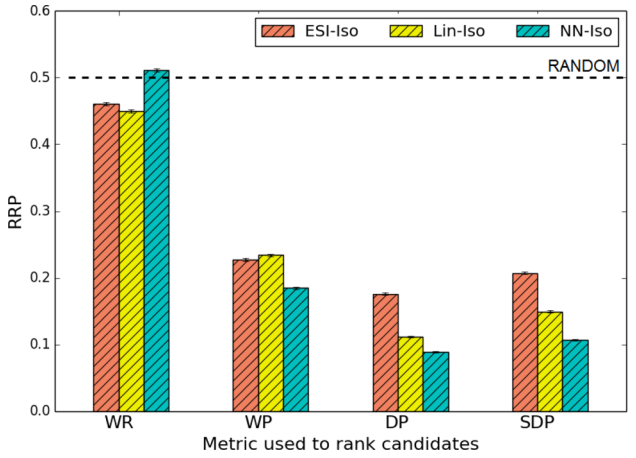


Figure 3. CFM-EI (NN-Iso and Lin-Iso) and CFM-ESI (ESI-Iso) identification performance on the Small Molecule Set when querying PubChem (median number of candidates = 1015). The *x* axis shows the metrics used to rank candidates: Weighted Recall (WR), Weighted Precision (WP), Dot Product (DP), and Stein Dot Product (SDP). Bars display mean relative ranking performance (RRP) scores. Error bars are too small to be seen. Note than an RRP of 0.0 is perfect and an RRP of 0.5 is no better than random.

CFM-EI's performance on the replicate set when querying HMDB and PubChem is better, and it substantially outperforms MetFrag. Both programs achieved RRP scores when querying HMDB that are very similar to those achieved when querying PubChem. This suggests that the characteristics of a molecule that make it more likely to be found in HMDB are independent of those characteristics that make it identifiable from its mass spectrum.

The results for derivatized compounds are a little worse than those obtained for nonderivatized compounds but not substantially so. The fact that only 987 from 17324 molecules in the training data are derivatized may be a factor. Although given that MetFrag also performs worse on the derivatized set, it may also be that this is a harder test, due to the similarity of the derivatized candidates. For example, if multiple locations on the molecule are feasible for derivatization, this provides multiple distinct but very similar candidates, only one of which is considered to be the correct one in this testing. It may be possible to improve these results by combining information from multiple derivatizations (e.g., 1 TMS, 2 TMS, 3 TMS, etc.) of the same compound. We have not yet attempted this.

Since average RRP scores can be unduly affected by outliers, we also compared per-molecule RRP scores between CFM-EI and MetFrag in the final column of Table 2. Average RRP differences seem to translate well to whom-beats-whom statistics on these tests, such that CFM still outperforms MetFrag in all tests.

Absolute ranking results for these same tests are shown in the left four axes of Figure 4. The lower number of candidates retrieved from HMDB for each molecule means that the similar RRPs translate to much better ranking performance than for PubChem. When querying HMDB, the target molecule was correctly identified in 45% of the cases and ranked in the top 10 in 86% of cases. When querying PubChem, the target molecule was correctly identified in 13% of cases and ranked in the top 10 in 45% of the cases. For the derivatized versions of these databases, the correct molecule was identified in 33% and 8% of cases, respectively, and ranked in the top 10 in 80% and 39%, respectively.

We reiterate that, while PubChem provides an interesting test case for our algorithms, it is generally a poor database choice for anyone wishing to do EI-MS- or GC/MS-based studies of metabolites, natural products, or environmental contaminants. With less than 1% of Pubchem's molecules having a biological or natural product origin, one is already dealing with a significant challenge of how to eliminate a 100:1 excess of false positives. So we would regard the results from the PubChem assessment as a "worst-case" scenario, and the results from the HMDB assessment as a more typical experimental scenario, in which we know something about the target compound of interest. It is likely that introducing further information about the compound of interest, as done in Ruttkies et al.,[49] would further increase the identification rates obtained. This might include retention indices, species of origin, sample type, abundance, or the likelihood of it appearing in publications or patents (a proxy measure for the relative abundance or likelihood of being detected).

The rightmost subfigure of Figure 4 shows the results obtained when querying the Replicate Set against the NIST candidates.

When querying against the measured reference spectra in the NIST database, the correct candidate was retrieved at rank 1 in 77% of cases. This is consistent with the results reported in Stein and Scott[2] and suggests that the combined effects of measurement variability, spectrum quality, and the information content in mass spectra (or lack thereof), mean that even actual measured spectra do not allow for perfect identification performance. In this same test, when CFM-EI was tasked with searching NIST, it was able to retrieve the correct candidate at rank 1 in 10% of cases. Given that there are more than 200000 candidates, this result is not bad. When restricted to consider compounds with the correct molecular formula, the rate of correct identifications increases to 42.6%. This scenario is often enabled by follow-up analysis with high mass accuracy MS.[43] Even when the molecular formula is not uniquely

Table 2. Average RRP of MassFrontier (MFrt), MOLGEN-MS (M-MS), MetFrag (MFrag), and CFM-EI (NN-Iso)[a]

Data Set	Query	N[b]	M[c]	MFrt	M-MS	MFrag	CFM	CFM v MFrag
Kerber	MOLGEN	100	802	0.268	0.273	0.354	0.199[d]	65 (69)
replicate	HMDB	3071	53	–	–	0.314	0.096[d]	77 (87)
replicate	PubChem	20133	1070	–	–	0.335	0.097[d]	84 (86)
replicate (D)[e]	dHMDB	540	43	–	–	0.411	0.128[d]	79 (85)
replicate (D)[e]	dPubChem	700	641	–	–	0.424	0.104[d]	82 (84)

[a]The final column is the percent molecules for which CFM-EI achieves a better (better or equal) RRP than MetFrag. Results for MassFrontier and MOLGEN-MS were taken from Schymanski et al.[46] [b]N: The number of molecules for which the correct molecule was in the candidate list. [c]M: The median number of candidates for those N molecules. [d]The best results of each condition. [e](D) indicates derivatized compounds only.

DOI: 10.1021/acs.analchem.6b01622
Anal. Chem. 2016, 88, 7689−7697

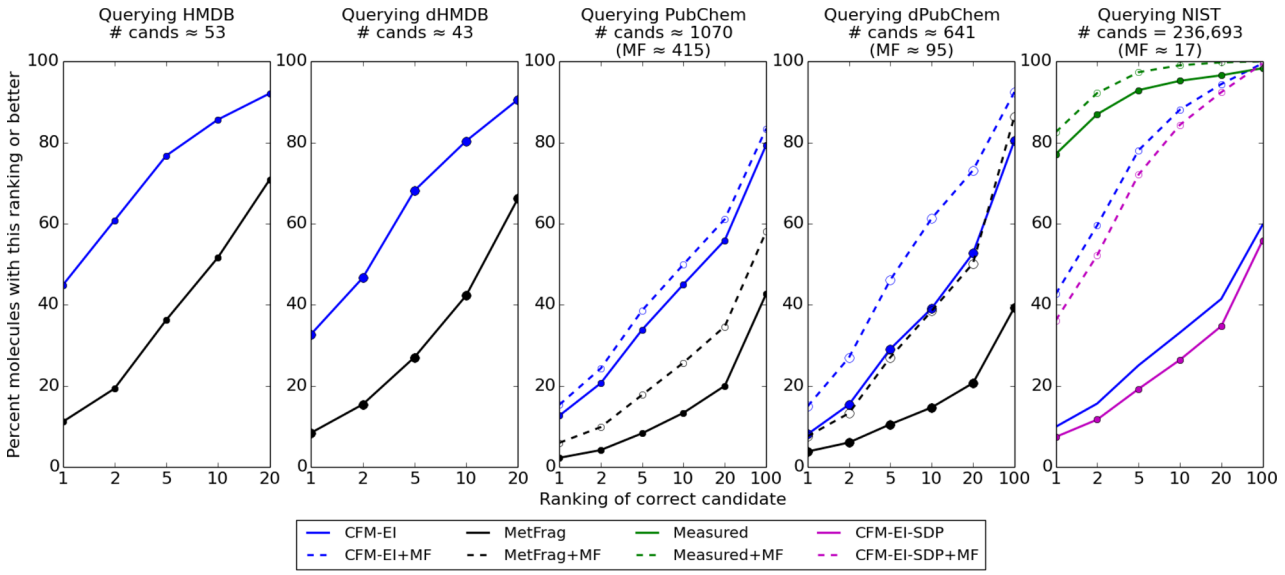


Figure 4. Absolute ranking results obtained using the replicate set, querying (from left to right) HMDB, dHMDB, PubChem, DPubchem, and NIST for candidate molecules. Solid lines indicate rankings achieved using the full set of candidates. Dashed lines indicate rankings achieved when narrowing the set of candidates to include only those with the correct molecular formula (MF). CFM-EI-SDP (in magenta) indicates that CFM-EI was run using Stein's Dot Product metric to compare spectra. All other CFM-EI results (in blue) use our Dot Product metric. # cands ≈ N: The median number of candidates is N. MF ≈ N: the median number of candidates with the correct MF is N.

identifiable, knowing an approximate mass substantially reduces the range of candidates to be considered. However, in some cases there is sufficient uncertainty surrounding the mass of the compound (e.g., if the molecular ion peak is absent in the mass spectrum), that a search of the entire database is required.

CONCLUSION

The CFM-EI tool provides an effective extension of the CFM method for use with EI-MS spectra (i.e., the spectra typically generated by GC-MS experiments). The spectrum prediction performance of CFM-EI has been benchmarked in cross-validation testing on the NIST database. It provides substantial improvements over the so-called bar code spectra commonly used for metabolite identification purposes. The method has also been extensively validated on multiple metabolite identification tasks. Head-to-head comparisons under multiple query conditions show that the CFM-EI significantly outperforms existing state-of-the-art computational methods.

These results also demonstrate that a gap still remains between identification performance obtainable when using computationally predicted spectra vs using real measured spectra. This confirms the view of Sumner et al.,[50] that metabolite identifications should ultimately be confirmed using comparisons with real measurements of reference standards.

Despite this apparent shortcoming, collecting reference spectra of chemical standards can be expensive, time-consuming, and is often infeasible, whereas computational methods offer a rapid, cost-effective alternative. It may be expected that computational methods will continue to be used as they are now: to narrow the chemical search space and hence reduce the experimental work load. Since CFM-EI outperforms other computational methods, it is an important contribution in this area and should help to reduce the time and cost of metabolite identifications.

Windows executables, cross-platform source code, and the trained models used in the results presented here are freely available at https://sourceforge.net/projects/cfm-id/. Test molecule lists, configuration files and per-molecule results can also be found there. A web server interface is also provided at http://cfmid.wishartlab.com/, which provides access to the trained CFM-EI model used here, along with examples of predicted spectra. Predicted spectra for all compounds in both HMDB and dHMDB (the derivatized version of HMDB used in these experiments) are also made available both on the sourceforge site and through the web server interface.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.6b01622.

Training CFM-EI with the isotope extensions, the backpropagation equations used in the neural network extensions, chemical features, test metrics used, computability criteria (with lists of example uncomputable compounds), and compute times (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: felicity.allen@ualberta.ca. Tel: +1 (780) 492-2285. Fax: 1 (780) 492-6393.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank Tom Wenseleers for his help with the NIST data, Christoph Ruttkies for providing the derivatized version of PubChem, Emma Schymanski for providing details of previous results, and Yannick Djoumbou for providing recommendations for functional groups. The authors were funded by NSERC, AICML, AIHS, Genome Alberta, and CIHR. This work was carried out using the Compute Canada Westgrid facility.

DOI: 10.1021/acs.analchem.6b01622
Anal. Chem. 2016, 88, 7689−7697

REFERENCES

(1) Stein, S. *Anal. Chem.* **2012**, *84*, 7274−7282.
(2) Stein, S. E.; Scott, D. R. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 859−866.
(3) Vinaixa, M.; Schymanski, E. L.; Neumann, S.; Navarro, M.; Salek, R. M.; Yanes, O. *TrAC, Trends Anal. Chem.* **2016**, *78*, 23−35.
(4) McLafferty, F. W. *Wiley Registry of Mass Spectral Data*, 11th ed.; John Wiley and Sons, 2016.
(5) Stein, S. *Standard Reference Library 1*; National Institute of Standards and Technology: Gaithersburg, MD, 2014.
(6) Wishart, D. S.; et al. *Nucleic Acids Res.* **2013**, *41*, D801−D807.
(7) Hastings, J.; de Matos, P.; Dekker, A.; Ennis, M.; Harsha, B.; Kale, N.; Muthukrishnan, V.; Owen, G.; Turner, S.; Williams, M.; Steinbeck, C. *Nucleic Acids Res.* **2013**, *41*, D456−63.
(8) Kind, T.; Fiehn, O. *Bioanal. Rev.* **2010**, *2*, 23−60.
(9) Scheubert, K.; Hufsky, F.; Böcker, S. *J. Cheminf.* **2013**, *5*, 12.
(10) Lindsay, R. K.; Buchanan, B. G.; Feigenbaum, E. A.; Lederberg, J. *The DENDRAL Project*; McGraw-Hill, 1980.
(11) Curry, B.; Rumelhart, D. E. *Tetrahedron Comput. Methodol.* **1990**, *3*, 213−237.
(12) Varmuza, K.; Werther, W. *J. Chem. Inf. Model.* **1996**, *36*, 323−333.
(13) Klawun, C.; Wilkins, C. L. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 249−257.
(14) Eghbaldar, a.; Forrest, T.; Cabrol-Bass, D. *Anal. Chim. Acta* **1998**, *359*, 283−301.
(15) Kwok, K.-S.; Venkataraghavan, R.; McLafferty, F. W. *J. Am. Chem. Soc.* **1973**, *95*, 4185−4194.
(16) Lowry, S. R.; Isenhour, T. L.; Justice, J. B.; McLafferty, F. W.; Dayringer, H. E.; Venkataraghavan, R. *Anal. Chem.* **1977**, *49*, 1720−1722.
(17) Stein, S. E. *J. Am. Soc. Mass Spectrom.* **1995**, *6*, 644−655.
(18) Heinonen, M.; Shen, H.; Zamboni, N.; Rousu, J. *Bioinformatics* **2012**, *28*, 2333−41.
(19) Brown, H.; Masinter, L. *Discrete Mathematics* **1974**, *8*, 227.
(20) Benecke, C.; Grüner, T.; Kerber, a.; Laue, R.; Wieland, T. *Fresenius' J. Anal. Chem.* **1997**, *359*, 23−32.
(21) Wieland, T.; Kerber, a.; Laue, R. *J. Chem. Inf. Model.* **1996**, *36*, 413−419.
(22) Bolton, E.; Wang, Y.; Thiessen, P.; Bryant, S. *Chapeter 12 in Annual Reports in Computational Chemistry*; American Chemical Society: Washington, D.C., 2008; Vol. *4*.
(23) McLafferty, F. W.; Turecek, F. *Interpretation of Mass Spectra*, 4th ed.; University Science Books, 1993.
(24) Kerber, A.; Meringer, M.; Rücker, C. *Croat. Chem. Acta* **2006**, *79*, 449−464.
(25) Schymanski, E. L.; Meringer, M.; Brack, W. *Anal. Chem.* **2009**, *81*, 3608−3617.
(26) Hill, A. W.; Mortishire-Smith, R. J. *Rapid Commun. Mass Spectrom.* **2005**, *19*, 3111−3118.
(27) Heinonen, M.; Rantanen, A.; Mielikainen, T.; Kokkonen, J.; Kiuru, J.; Ketola, R.; Rousu, J. *Rapid Commun. Mass Spectrom.* **2008**, *22*, 3043−3052.
(28) Wolf, S.; Schmidt, S.; Müller-Hannemann, M.; Neumann, S. *BMC Bioinf.* **2010**, *11*, 148.
(29) Wang, Y.; Kora, G.; Bowen, B. P.; Pan, C. *Anal. Chem.* **2014**, *86*, 9496−9503.
(30) Ridder, L.; van der Hooft, J. J. J.; Verhoeven, S.; de Vos, R. C. H.; van Schaik, R.; Vervoort, J. *Rapid Commun. Mass Spectrom.* **2012**, *26*, 2461−71.
(31) Ruttkies, C.; Strehmel, N.; Scheel, D.; Neumann, S. *Rapid Commun. Mass Spectrom.* **2015**, *29*, 1521−1529.
(32) Gasteiger, J.; Hanebeck, W.; Schulz, K.-P. *J. Chem. Inf. Model.* **1992**, *32*, 264−271.
(33) Gasteiger, J.; Li, X.; Simon, V.; Novič, M.; Zupan, J. *J. Mol. Struct.* **1993**, *292*, 141−160.
(34) Kangas, L. J.; Metz, T. O.; Isaac, G.; Schrom, B. T.; Ginovska-Pangovska, B.; Wang, L.; Tan, L.; Lewis, R. R.; Miller, J. H. *Bioinformatics* **2012**, *28*, 1705−13.
(35) Grimme, S. *Angew. Chem., Int. Ed.* **2013**, *52*, 6306−6312.
(36) Bauer, C. A.; Grimme, S. *Org. Biomol. Chem.* **2014**, *12*, 8737−44.
(37) Bauer, C. A.; Grimme, S. *J. Phys. Chem. A* **2016**, *120*, 3755−3766.
(38) Allen, F.; Greiner, R.; Wishart, D. *Metabolomics* **2015**, *11*, 98−110.
(39) Dührkop, K.; Shen, H.; Meusel, M.; Rousu, J.; Böcker, S. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 12580−12585.
(40) Allen, F. *Competitive Fragmentation Modeling of Mass Spectra for Metabolite Identification*. Ph.D. Thesis, University of Alberta, 2016.
(41) Karni, M.; Mandelbaum, A. *Org. Mass Spectrom.* **1980**, *15*, 53−64.
(42) Rockwood, A. L.; Haimi, P. *J. Am. Soc. Mass Spectrom.* **2006**, *17*, 415−9.
(43) Böcker, S.; Letzel, M. C.; Lipták, Z.; Pervukhin, A. *Bioinformatics* **2009**, *25*, 218−224.
(44) Fernandez-De-Cossio Diaz, J.; Fernandez-De-Cossio, J. *Anal. Chem.* **2012**, *84*, 7052−7056.
(45) Claesen, J.; Dittwald, P.; Burzykowski, T.; Valkenborg, D. *J. Am. Soc. Mass Spectrom.* **2012**, *23*, 753−763.
(46) Schymanski, E. L.; Gallampois, C. M. J.; Krauss, M.; Meringer, M.; Neumann, S.; Schulze, T.; Wolf, S.; Brack, W. *Anal. Chem.* **2012**, *84*, 3287−3295.
(47) Glorot, X.; Bordes, A.; Bengio, Y. *AISTATS* **2011**, *15*, 315−323.
(48) Krizhevsky, A.; Sutskever, I.; Hinton, G. *NIPS* **2012**, 1097−1105.
(49) Ruttkies, C.; Schymanski, E. L.; Wolf, S.; Hollender, J.; Neumann, S. *J. Cheminf.* **2016**, *8*, 3.
(50) Sumner, L. W.; et al. *Metabolomics* **2007**, *3*, 211−221.

DOI: 10.1021/acs.analchem.6b01622
Anal. Chem. 2016, 88, 7689−7697