Summer project:

Improving Tandem Mass Spectrometry based Metabolite Identification using Collision Cross Section Measurements

Buğra Aker Yılmaz
Supervisor: Juho Rousu
Instructor: Eric Bach

August 28, 2019

Contents

1 Introduction

The identification of *unknown* metabolites contained in biological samples, e.g. blood, urine or plant-extracts, is an important and at the same time challenging task in the field of *untargeted Metabolomics*. Metabolites are small molecules[1] involved in biological processes. The term "unknown" means, that the molecular structure is not known. Typically a biological sample, analyzed in an untargeted Metabolomics study, contains a large number of unknown molecules, e.g. ranging from hundreds to thousands.

1.1 Metabolite Identification using Mass Spectrometry

Due to its ability to perform high-throughput screening, its high sensitivity and its applicability to a wide range of molecules, *mass spectrometry* is one of most widely used analysis method in Metabolomics. A mass spectrometer can separate molecules based on their mass[2] and measures their abundance. However, only the mass and abundance of an unknown molecule, is not sufficient for its identification. Therefore, typically two mass spectrometers are put in series with a molecular fragmenter in between. This setup is called *tandem mass spectrometer* and the resulting measurements of the unknown metabolites are called *tandem mass spectrum* (MS/MS spectrum, compare Figure 1).

A typical workflow of the metabolite identification using MS/MS spectra is shown in Figure 1. After the the tandem mass (MS/MS) spectra (in the figure only one is illustrated) are measured, some kind of similarity measure between the MS/MS spectrum and a set of so called molecular candidates is calculated. Those candidates are ranked and the highest ranked molecular candidates structure is considered as identification. Current state-of-the-art approaches for the identification of metabolites using MS/MS spectra are utilizing kernel based machine learning (Dührkop et al., 2015; Brouard et al., 2016; Dührkop et al., 2019) and reach roughly 40% correct identifications.

1.2 Collision Cross-Section Values from Ion Mobility Mass Spectrometry

Tandem mass spectrometry (MS/MS) delivers very rich information about the unknown molecules in biological samples and high identification rates can be achieved. Nevertheless, additional (orthogonal) information can be gathered, when mass spectrometry is combined with other measuring methods such as *ion mobility spectrometry (IMS)* (Kanu et al., 2008).

[1] Weight less than 1000-1500Da

[2] More accurately: Mass-to-Charge ration, [m/z].

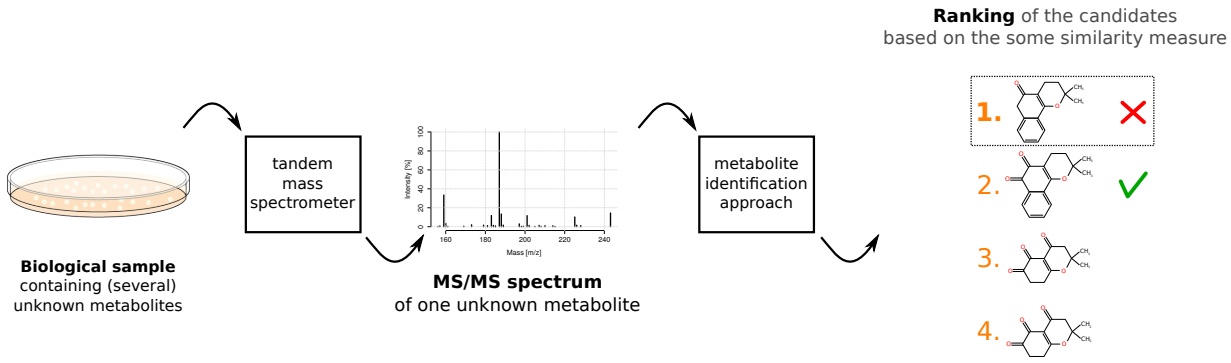


Figure 1: General workflow of the metabolite identification using tandem mass (MS/MS) spectra. The correct molecular candidates structure is at ranked at the second position in this example. Figure by Eric Bach.

IMS is a method to separate the molecules in biological samples, which could not be separated by the solely relying on mass spectrometry. In this way, additional information about the molecules can be gained to improve the metabolite identification performance (Zhou et al., 2016; Tejada-Casado et al., 2018).

The typical workflow in an IMS-MS setup is as follows: First the biological samples is injected into the ion mobility spectrometer. There the molecules fly through a gas and depending on their size and shape they separate and leave the system at different so called drift times. Due to this separation, isomers, isobars and conformers can be distinguished (Kanu et al., 2008). Subsequently, the separated molecules enter the mass spectrometer (compare Section 1.1) and will be separated by mass. The separation of for example isomers would not be possible only solely based on the mass of the molecules, so that the IMS adds orthogonal information.

As mentioned in the previous Section 1.1, the output of the mass spectrometer is the MS/MS spectrum of a molecule. On the other hand, the output of the IMS, based on the drift times, are the so called *collision cross-section (CCS)* values for each molecule. That means, the IMS-MS workflow provides a set of (MS/MS spectrum, CCS value)-tuples, that can be used for the identification of the unknown compounds.

1.3 Related work

One of the most common approaches to use additional (or orthogonal) information for the metabolite identification is additional filtering of candidate metabolites. There have been two studies to apply this with experimental CCS (Paglia et al., 2014) and with predicted CCS values (Zhou et al., 2016) to further restrict the candidate sets that are obtained by mass (m/z) filtering. In other works, the efficiency of using CCS values to decrease the

number of identification candidates from a set obtained by a m/z query to a metabolite database is investigated in these studies.

Paglia et al. (2014) and Nichols et al. (2018) demonstrate only one case in which the use of CCS in addition to m/z proved to be effective, yet the studies are bound by the small number of compounds of which CCS is measured. Zhou et al. (2016), on the other hand, illustrates a decrease in the number of identification candidates for 441 metabolites, with 15% over-filtering i.e. filtering of 15% of true candidates. However, the study does not report the rate of filtered candidates with respect to the total number of candidates, which makes it hard to compare with the over-filtering rate. Therefore, whether CCS measurements or predictions in addition to m/z measurements can be effectively used to identify metabolites in a large scale is still a valid question.

Tejada-Casado et al. (2018) provides an analysis of CCS measurements of 173 ions. They report a high correlation between m/z and the CCS values. They also conclude that different families of compounds might have different tendencies in terms of CCS and the difference between CCS measurements in different ionization modes can be informative.

Nichols et al. (2018) inspects the separation of isomers (molecules sharing the same chemical formula) via CCS measurements in metabolite identification. In the work, it is argued that MS/MS and high precursor mass is not enough to identify most metabolite isomers. It is demonstrated that CCS measurements in some ion-modes can contain orthogonal information and concluded that nearly half of the isomeric compounds have more than 2.0% difference in CCS of 500 metabolites of which CCS are measured.

2 KEPACO Group

This project was carried out in Kernel Methods, Pattern Analysis and Computational Metabolomics (KEPACO)[3] group, which is a research group located in Department of Computer Science, Aalto University, Espoo, Finland.[4] It also has structural connection to Helsinki Institute for Information Technology.[5] As the name implies, the group's work is focused on machine learning tasks on biological molecules called metabolites and the learning method is mainly on Kernel Methods.

The group has an informal weekly gathering called KEPACOffee, during which everyone presents weekly challenges and progresses on their projects. Thus, everyone has a chance to follow the research or teaching activity in the group and get feedback on theirs. In addition, there is a presentation of a research project or a paper from the literature.

[3] https://research.cs.aalto.fi/kepaco/
[4] https://www.aalto.fi/en/department-of-computer-science
[5] https://www.hiit.fi/

2.1 People

Professor Juho Rousu is head of the group. Assistant Professor Rohit Babbar is also a part of the group. There are three post-doc researchers and three doctoral candidates. For, up-to-date information, group website can be visited. The group consists of people from a wide range of different countries, making it easier for other international researchers to adopt.

2.2 Research

The current and past research projects of the group are

- TensorBiomed - Tensor Learning for Biomedicine, 2018-2019
- MACOME - Machine Learning for Computational Metabolomics, 2017-2021
- FCHealth - Foundations of Computational Health
- D4Health - Data-Driven Decision Support for Digital Health, 2016-2018
- LiF - Living Factories, 2014-2017
- MIDAS - Metabolite Identification through Algorithms and Statistical Learning, 2013-2017.

The publication forums include high-quality journals such as *Nature Methods*, *PNAS*, *Bioinformatics*, *PLOS Computational Biology*, *Machine Learning* as well as leading conferences such as *ISMB*, *ICML*, *ICDM* and *AISTATS*. More information on publications, researches and activities are present on the group website.

3 Methods

This section introduces the IOKR metabolite identification framework developed by Brouard et al. (2016) and the collision cross-section prediction model developed by Plante et al. (2019). It furthermore explains how the CCS measurements and predictions could be used to improve the IOKR metabolite identification.

3.1 Input Output Kernel Regression (IOKR)

This approach considers an input kernel $\kappa_x : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ measuring the similarity between MS/MS spectra, as well as an output kernel $\kappa_y : \mathcal{M} \times \mathcal{M} \mapsto \mathbb{R}$ measuring the similarity between molecular structures. The input kernel is associated with a feature map $\phi : \mathcal{X} \mapsto \mathcal{F}_x$ and the output kernel is associated with a feature map $\psi : \mathcal{M} \mapsto \mathcal{F}_y$. The metabolite identification problem is decomposed in two tasks (Brouard et al., 2016).

In the **first task**, the output feature map ψ is approximated by learning a mapping $h : \mathcal{X} \mapsto \mathcal{F}_y$, i.e. between the set of MS/MS spectra and the output feature space. Given a set of training examples $\{(x_i, y_i) \in \mathcal{X} \times \mathcal{M}\}_{i=1}^{\ell}$, the mapping is learned by solving the following (regression) optimization problem:

$$\min_{h \in \mathcal{H}} \sum_{i=1}^{\ell} \|h(x_i) - \psi(y_i)\|_{\mathcal{F}_y}^2 + \lambda \|h\|_{\mathcal{H}}^2, \tag{1}$$

where $\lambda > 0 \in \mathbb{R}$ is the regularization parameter, and $\mathcal{H}$ the hypothesis space, i.e. set of possible functions h. The mapping h writes as $h(x') = \sum_{i=1}^{\ell} \mathbf{c}_i \kappa_x(x_i, x')$, with $\mathbf{c}_i \in \mathcal{F}_y$ being the model parameters.

In the **second task**, the molecule corresponding to a (in practice new) MS/MS spectrum x' is predicted by comparing the predicted feature vector $h(x')$ with a set of molecular candidates $\mathcal{M}(x')$. These candidate molecules are extracted from a large molecular structure database based on the (estimated) molecular formula of x'. The predicted molecular structure y' is found by solving the following optimization problem[6]:

$$y' = f(x') = \underset{y \in \mathcal{M}(x')}{\arg\max} \langle h(x'), \psi(y) \rangle_{\mathcal{F}_y}, \tag{2}$$

where $f : \mathcal{X} \mapsto \mathcal{Y}$. In other words: The molecular candidate y with the highest score[7] (inner product) is considered to be the metabolite identification of x'.

3.2 Collision Cross-Section Prediction

The neural network model provided by Plante et al. (2019) is used to predict the CCS values of the molecular candidates. A molecule's isomeric SMILES and ion configuration are used to predict its CCS value. The model's architecture is given in Figure 2.

The authors provide a pre-trained model, where the training of the model is achieved in two steps, which is known as transfer learning. First, the convolutional layers of the

[6] Practically this is nothing else than an exhaustive search in $\mathcal{M}(x')$.

[7] Or: The lowest rank

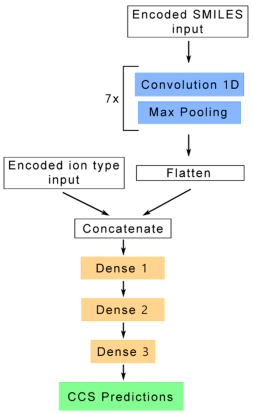


Figure 2

Figure 3: The Architecture of the Deep Neural Network for CCS prediction provided by Plante et al. (2019).

model is trained with multi-task learning to predict polar surface area, logS, refractivity, polarizability, logP (from ALOGPS database), and logP (from Chemaxon database) of roughly 70 thousand compounds in order to learn a valid representation of a molecule as output of the convolutional layers. Then, fixing the parameters of convolutional layers, the 3 dense layers of the network is trained to predict CCS of the compounds.

3.3 Learning with Multiple Input Kernels

The collision cross-section values, can be considered as a second input representation (besides the MS/MS). For that, a second input kernel $\kappa_z : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ with an associated feature map $\varphi : \mathbb{R} \mapsto \mathcal{F}_z$ is considered.

The **first task** of the IOKR framework (Section 3.1) can be changed in the following way. Given a set of training examples $\{(x_i, z_i, y_i) \in \mathcal{X} \times \mathbb{R} \times \mathcal{M}\}_{i=1}^{\ell}$, where $z_i \in \mathbb{R}$ are measured (or predicted) CCS values, the function h (here $h : \mathcal{X} \times \mathbb{R} \mapsto \mathcal{F}_y$) can be learned by solving:

$$\min_{h \in \mathcal{H}} \sum_{i=1}^{\ell} \|h(x_i, z_i) - \psi(y_i)\|_{\mathcal{F}_y}^2 + \lambda \|h\|_{\mathcal{H}}^2. \tag{3}$$

where the optimal mapping functions writes as:

- $$h(x', z') = \sum_{i=1}^{\ell} \mathbf{c}_i \left(\kappa_x(x_i, x') + \kappa_z(z_i, z')\right) \tag{4}$$

  or alternatively

- $h(x', z') = \sum_{i=1}^{\ell} \mathbf{c}_i \left(\kappa_x(x_i, x') \cdot \kappa_z(z_i, z')\right)$ (5)

Important to note is, that this approach assumes the knowledge of measured CCS values alongside the MS/MS spectra *during model training*. In addition, it does not utilizes the ionization modes of the input compounds in which the CCS is measured.

3.3.1 Addition of CCS Kernel(s) to MS/MS Spectra Kernel

When combining input kernels by addition (See Equation 4), in practice one would give a weighting to each kernel. That means, a weighted some if kernel is calculated instead. Different strategies to determine the weights have been introduced in the literature. In this project, we focused on two types of additive kernel combination:

- **UNIMKL:** This method proposes a uniform combination of input kernels. If w is the n dimensional weight vector with weights $w_1, w_2, ..., w_n$ for input kernels $\kappa_1(x, x')$, $\kappa_2(x, x')$, $...\kappa_n(x, x')$, then, the weight vector is given by $w_i = 1/n$ for $i \in n$.
- **ALIGNF:** This method is proposed by Cortes et al. (2012). The method finds the weight vector w that maximizes centered alignment between the combined input kernels and the output kernel, with $\|w\|_2 = 1$.

3.3.2 Multiplication of CCS Kernel with MS/MS Spectra Kernel

The optimal mapping function for this method is given in Equation 5. We investigate three different centering and normalization pipelines for the multiplicative kernel combination, each of which results in a different feature space. Let $\omega : \mathcal{Z} \mapsto \mathcal{F}_z$ be the feature mapping associated with CCS kernel and $\chi : \mathcal{X}, \mathcal{Z} \mapsto \mathcal{F}_{x,z}$ be the feature mapping associated with combined input kernel.

- **Centering and Normalization Before Multiplication:** All input kernels are centered and normalized separately before multiplication. The MS/MS kernels are combined before the multiplication. The corresponding feature space is

$$\chi_{BM}(x', z') = \frac{\phi(x') - \frac{1}{l}\sum_l \phi(x_i)}{\|\phi(x') - \frac{1}{l}\sum_l \phi(x_i)\|} \otimes \frac{\omega(z') - \frac{1}{l}\sum_l \omega(z_i)}{\|\omega(z') - \frac{1}{l}\sum_l \omega(z_i)\|}$$

- **Centering and Normalization Before & After Multiplication:** Resulting kernel from the "Centering and Normalization Before Multiplication" method is centered

and normalized in addition to all kernels being centered and normalized separately before multiplication. The MS/MS kernels are combined before the multiplication. The corresponding feature space is

$$\chi_{BAM}(x', z') = \frac{\chi_{BM}(x', z') - \frac{1}{l}\sum_l \chi_{BM}(x_i, z_i)}{\|\chi_{BM}(x', z') - \frac{1}{l}\sum_l \chi_{BM}(x_i, z_i)\|}$$

- **Centering and Normalization After Multiplication:** Kernel-matrices of MS/MS and CCS are multiplied and the resulting matrix is centered and normalized. The MS/MS kernels are first multiplied by CCS kernel individually, before linear combination. The corresponding feature space is

$$\chi_{AM}(x', z') = \frac{\phi(x') \otimes \omega(z') - \frac{1}{l}\sum_l \phi(x_i) \otimes \omega(z_i)}{\|\phi(x') \otimes \omega(z') - \frac{1}{l}\sum_l \phi(x_i) \otimes \omega(z_i)\|}$$

Multiplying each MS/MS kernel by CCS kernel individually, before the linear combination is mathematically equivalent to linearly combining MS/MS kernels, before multiplication with CCS kernel, even though the results may practically differ. Therefore, only one of the two orders is implemented for each method.

3.4 Late Fusion of IOKR with Predicted CCS Values

An approach to incorporate the CCS predictions is the filtering of the candidate sets used in the **second task** of the IOKR framework (Section 3.1), so that candidates with additional CCS predictions to the measured CCS of the input molecule are removed from the candidate sets based on a filtering threshold. Ideally, CCS predictions of false candidates are further from the true CCS of the input compound than the CCS prediction of the true candidate. This provides reduced number of false candidates, which may improve the performance of the metabolite identification.

Another approach to incorporate the CCS predictions is the modifications of the **second task** of the IOKR framework (Section 3.1), so that the scores of candidates with closer CCS prediction to the measured CCS of the input molecule are up-weighted. This approach enables us to combine the CCS predictions in a *softer* manner than the filtering, such that candidates with further CCS prediction to the measured CCS of the input molecule are not removed from the candidate set, rather down-weighted.

Important to note is that the approaches in this section does *not* assume the knowledge of measured CCS values alongside the MS/MS spectra during model training, but only during the scoring phase.

3.4.1 Candidate Filtering Using Predicted CCS values of Candidates

We investigate two types of filtering strategies:

- **Thresholding on True CCS:** Filtering based on a relative threshold on true CCS value of the input compound. Specifically, removing candidates whose predicted CCS values are not in the range $[z' - t * z', z' + t * z']$, t being the threshold and z' being the CCS measurement of an input compound.
- **Thresholding on Candidate Set Size:** Filtering based on a relative threshold on the candidate set size. Specifically, removing $t\%$ of the candidates whose predicted CCS values are furthest from the true CCS of the input compound, where t is the threshold.

3.4.2 Molecular Candidate Re-ranking

Let $g : \mathcal{M} \mapsto \mathbb{R}$ be a CCS prediction function, then to each molecular candidate $y \in \mathcal{M}(x')$ for a given MS/MS spectrum $x' \in \mathcal{X}$ (with measured z' a CCS value), $g(y) = \tilde{z} \in \mathbb{R}$ can be predicted and used to re-score (re-rank) the molecular candidates:

- **Re-ranking by Addition:**

$$y' = f(x', z') = \underset{y \in \mathcal{M}(x')}{\arg\max} \left(\langle h(x'), \psi(y) \rangle_{\mathcal{F}_y} + \kappa_z(z', g(y)) \right)$$

- **Re-ranking by Multiplication:**

$$y' = f(x', z') = \underset{y \in \mathcal{M}(x')}{\arg\max} \left(m(\langle h(x'), \psi(y) \rangle_{\mathcal{F}_y}) \cdot \kappa_z(z', g(y)) \right),$$

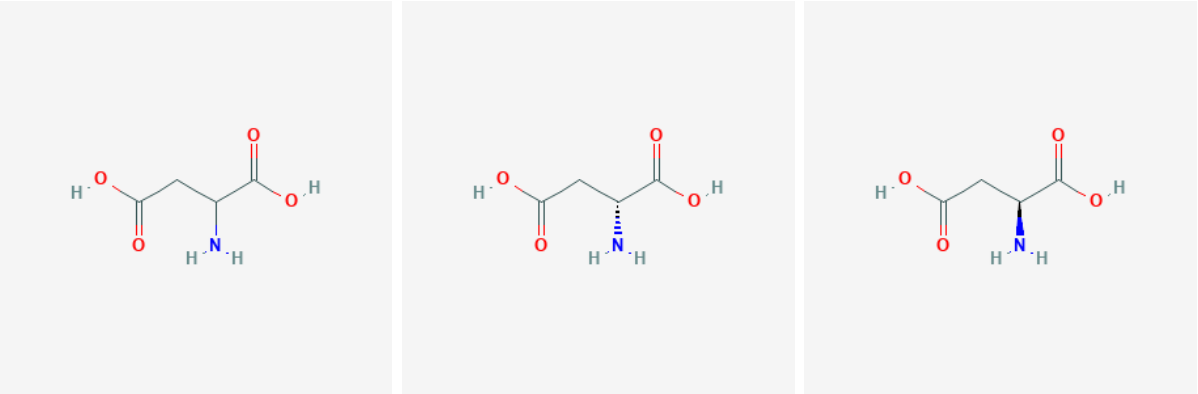
  where $m(x)$ is a monotonically increasing mapping function used to map the values in range $(-\infty, \infty)$ to a non-negative range. Mapping of IOKR scores is needed, since $\langle h(x'), \psi(y) \rangle_{\mathcal{F}_y}$ may be negative, which would result in down-weighting instead of up-weighting the candidate when $\kappa_z(z', g(y))$ is relatively high. Moreover, we assume that $\kappa_z(z', g(y))$ is always non-negative, otherwise another mapping function with the same characteristics of $m(x)$ is needed.

4 Datasets

This section introduces two merged datasets we worked with in this project, as well as definitions and analyses related to them. The term merged here refers to the fact that we are finding (MS/MS, CCS)-tuples, by merging independent MS/MS and CCS databases.

(a) InChIKey examples

InChIKey			Stereo Chemistry	Structure
Level 1	**Level 2**	**Level 3**		
CKLJMWTZIZZHCS	UHFFFAOYSA	N	No stereo information	Figure 4b
CKLJMWTZIZZHCS	UWTATZPHSA	N	Stereo isomer 1	Figure 4c
CKLJMWTZIZZHCS	REOHCLBHSA	N	Stereo isomer 2	Figure 4d



(b) DL-Aspartic acid (c) D-Aspartic acid (d) Aspartic acid

Figure 4: Different molecules with the same main InChi layer. DL-Aspartic acid (b) bares no stereo configuration, while D-Aspartic acid (c) and Aspartic acid (d) are stereo-isomers. Images retrieved from PubChem.

We used an MS/MS database containing 14968 MS/MS spectra (13497 unique molecular structures) and a set of 1640 CCS measurements (1141 unique (2D) molecular structures) to construct our merged datasets. Even though, the CCS dataset provides information about the stereo-chemistry and also different measurements for some stereo-isomers, the MS/MS dataset does not provide this information. The later fact is a design-decision[8] tandem mass spectrometry cannot distinguish stereo-chemistry. Table 1 gives an example for these datasets.

4.1 2D and 3D Molecular Structures

In this project, we have used the "2D" and "3D" molecular structure concepts as follows:

- **2D Molecular Structure:** The level of information carried by only InChiKey level-1 which is hash of the main layer of InChi. The main layer of the InChi contains

[8]Partly a propagated one from publicly available MS/MS databases.

information regarding the connectivity of a molecule, but does not contain charge, stereo-chemical and isotopic information.

- **3D Molecular Structure:** The level of information carried by the main and stereo-chemical layer of InChi.

Figure 4 illustrates these concepts. All of the molecules share the same 2D molecular structure given in Figure 4b, while other two carries stereo-chemistry information and have different 3D molecular structures. It is important to note that in general, different InChiKeys might also be due to different isotopic patterns.

Table 1: CCS and MS/MS Spectra Dataset Examples

(a) CCS Dataset Example

InChIKey	**Ion Mode**	**CCS (Å²)**
A-X-N	M+H	132.5
A-Y-N	M+H	132.6
A-Z-N	M+H	134.8
D-W-N	M+Na	122.3

(b) MS/MS Spectra Dataset Example

InChIKey L1	**Ion Mode**	**MS/MS ID**
A	M+H	1188546
D	M+Na	1188706

Table 2: Merged Dataset Examples (CCS Values given in $Å^2$)

(a) Merged 2D Dataset

ID	**Ion Mode**	**CCS**	**MS/MS ID**
A	M+H	133.3	1188546
D	M+Na	122.3	1188706

(b) Merged 3D Dataset

ID	**Ion Mode**	**CCS**	**MS/MS ID**
A-X-N	M+H	132.5	1188546
A-Y-N	M+H	132.6	1188546
A-Z-N	M+H	134.8	1188546
D-W-N	M+Na	122.3	1188706

4.2 2D Merged (MS/MS, CCS)-Dataset: Positive Ionization Mode

Table 2a illustrates the merging. A particular MS/MS is matched to a CCS value, if their *2D molecular structures* and *ionization mode*, i.e. adduct configuration, match. CCS measurements from different stereo-isomers are averaged before the merge. Candidate sets are constructed based on the molecular formula of the metabolites, in other words, candidate sets are molecules which have the same molecular formula. Moreover, only metabolites with different 2D structures are included in the candidate sets. Fingerprints (molecular descriptors) are obtained by the main layer of InChi, thus calculated by 2D

molecular structure. Therefore, this configuration assumes metabolite identification on the level of 2D molecular structure.

The merged dataset consist of **779 (MS/MS, CCS)-tuples** (625 unique molecular structures). The dataset contains molecules only in positive ionization modes and has the following ionization mode distribution: 715 "M+H", 63 "M+Na" and 1 "M". To each MS/MS spectra, a set of molecular candidates is associated with a total number of 1.5 Million structures that will be used to determine the metabolite identification performance.

4.3 3D Merged (MS/MS, CCS)-Dataset: Positive Ionization Mode

Table 2b illustrates the merging. Instead of averaging CCS measurements from different stereo-isomers, we matched *every* CCS measurement with an MS/MS measurement if they have the same 2D structure. This means that we have an MS/MS measurement coupled with different CCS measurements. As a consequence, the input molecule now does not only have 2D information but also 3D information, since CCS values are based on measurements of molecules having 3D structure information. That enabled us to construct candidate sets, which contain stereo-isomers. For every (MS/MS, CCS) tuple, we considered a candidate as the true one if they have the same 3D structure with the related molecule's CCS measurement, even though MS/MS data is based on only 2D structure. Fingerprints (molecular descriptors) are obtained by the isomeric SMILES, thus calculated by 3D molecular structure. Therefore, this configuration assumes metabolite identification on the level of 3D molecular structure.

We constructed two candidate sets:

- **Molecular Mass:** A set of molecular candidates whose exact mass, determined from the molecular formula, is within a +/- 5ppm window around the ground-truth exact molecular mass.
- **Molecular Formula:** A set of molecular candidates only containing molecules with the same molecular formula as the ground-truth structure and whose masses are within a +/- 5ppm window around the ground-truth exact molecular mass.

The candidate set based on molecular formula is a proper subset of that based on molecular mass.

The merged dataset consists of **1060 (MS/MS, CCS)-tuples** with 779 unique MS/MS spectra, 647 unique 3D structures (full InChiKeys) and 615 unique 2D structures (InChiKeys level-1s). The dataset contains molecules only in positive ionization modes and has the following ionization mode distribution: 949 "M+H", 110 "M+Na" and 1 "M". Including

Table 3: IOKR Metabolite identification (10 Fold-CV) for the 2D Merged Dataset. Best results for each column are in bold. Abbreviations BA, BAM and AM in parenthesis for multiplication methods stands for Centering and Normalization Before Multiplication, Centering and Normalization After Multiplication, Centering and Normalization Before & After Multiplication, respectively.

CCS Kernel(s)	**MKL**	**Top-1**	**Top-5**	**Top-10**	**Top-20**
None	UNIMKL	32.0%	**60.3%**	69.5%	77.5%
None	ALIGNF	31.1%	57.1%	66.7%	73.9%
Gaussian	UNIMKL	**32.3%**	**60.3%**	**70.3%**	**77.9%**
10 Gaussian and 1 GRK	ALIGNF	31.2%	57.2%	66.7%	73.9%
Gaussian	Multiplication (BM)	28.0%	54.5%	62.7%	70.2%
Gaussian	Multiplication (BAM)	27.6%	54.2%	63.0%	69.9%
Gaussian	Multiplication (AM)	30.8%	57.8%	66.8%	74.7%

repeated candidate sets for the input molecules having the same molecular formula, total number of molecular candidates is nearly 4.7 million in the candidate set constructed based on molecular mass and is slightly above 3 million in candidate set based on true molecular formula. The number of stereo-isomers of an input compound in its candidate set based on molecular mass and molecular formula are equal and in average, is 13.48 stereo-isomers.

5 Experiments

This section illustrates the experiments done in the project. The following subsections include those conducted on the related datasets.

5.1 2D Merged Dataset: Positive Ionization Mode

Table 3 shows the IOKR metabolite identification performance using the 2D Merged Dataset of 779 (MS/MS, CCS)-tuples. The details of the dataset is given in the Section 4.2 and the details of the experiments are provided in the following subsections.

The same 10-fold Cross Validation structure is used for all experiments, where all (MS/MS, CCS)-tuples belonging to the same molecular structure (based on their InChI) are in the same fold. For all experiments, the output kernel is Gaussian Kernel on Tanimoto kernel and 14 MS/MS spectra (and fragmentation tree) kernels have been used as input.

5.1.1 Addition of CCS Kernel(s) to MS/MS Spectra Kernel

Table 3 contains the results discussed in this section. The CCS information is used in addition to MS/MS spectra kernels. Details of the method is given in Section 3.3, where the objective function is obtained by addition of CCS and MS/MS kernel functions.

In the UNIMKL method, only one kernel on CCS values are used, whose parameter is selected by maximizing the entropy of the training CCS kernel matrix. On the other hand, in the ALIGNF method, 11 kernels on CCS values are used, in order to let the ALIGNF algorithm perform the parameter selection. 10 of 11 CCS kernels were Gaussian Kernels with different parameters and 1 was a Gaussian Response Kernel (Cichonska et al., 2018). Results show only a very slight increase in the metabolite identification performance. Since there is almost no difference between the results of two experiments with ALIGNF method, weights are further investigated. It seems that all of the weights for CCS kernels were below 0.01, which indicates that CCS kernels provide only little information to increase the centered alignment with the output kernel

5.1.2 Multiplication of CCS Kernel with MS/MS Spectra Kernel

Table 3 contains the results of this section. The CCS information is used by multiplication with MS/MS spectra kernels. That is, the kernel matrix of CCS kernel is multiplied element-wise with the kernel matrix of combined MS/MS spectra kernels. Details of the method is given in Section 3.3, where the objective function is obtained by multiplication of CCS and MS/MS kernel functions.

The parameter of CCS kernel is selected by maximizing the entropy of the training kernel matrix and UNIMKL is used as additive combination method of MS/MS spectra kernels, for all experiments. The centering and normalization pipeline "Centering and Normalization After Multiplication" performs the best, however the multiplicative kernel combination stays below the baseline performance.

5.2 3D Merged Dataset: Positive Ionization Mode

Table 4 shows the IOKR metabolite identification performance using the 3D Merged Dataset of 1060 positive MS/MS-spectra with the CCS information. The details of the dataset is given in the Section 4.2 and the details of the experiments are provided in the following subsections.

The results are calculated using 10-fold CV, where all spectra belonging to the same molecular structure (based on their InChI) are in the same fold. The same MS/MS kernels with

Table 4: IOKR Metabolite identification (10 Fold-CV) for the 3D Merged Dataset with predicted CCS values for candidates. Best results for each column and candidate set type are in bold.

(a) Multiple Kernel Learning

Candidate Set	**CCS Kernel**	**MKL**	**Top-1**	**Top-5**	**Top-10**	**Top-20**
Molecular Formula	None	UNIMKL	25.3%	**49.3%**	**60.3%**	**69.3%**
Molecular Formula	Gaussian	UNIMKL	**25.4%**	**49.3%**	60.2%	**69.3%**
Molecular Formula	Gaussian	ALIGNF	21.8%	46.3%	57.1%	66.6%
Molecular Mass	None	UNIMKL	25.1%	**49.4%**	**60.3%**	**69.0%**
Molecular Mass	Gaussian	UNIMKL	**25.3%**	**49.4%**	60.2%	**69.0%**
Molecular Mass	Gaussian	ALIGNF	22.0%	46.1%	57.0%	66.2%

(b) Re-ranking the IOKR scores based only on MS/MS using predicted CCS values

Candidate Set	**Re-ranking Method**	**Top-1**	**Top-5**	**Top-10**	**Top-20**
Molecular Formula	Multiplication	20.7%	45.6%	56.3%	67.6%
Molecular Formula	Addition	**25.9%**	**49.0%**	**59.9%**	**69.9%**
Molecular Mass	Multiplication	20.8%	45.1%	56.3%	67.5%
Molecular Mass	Addition	**25.8%**	**48.4%**	**59.8%**	**69.6%**

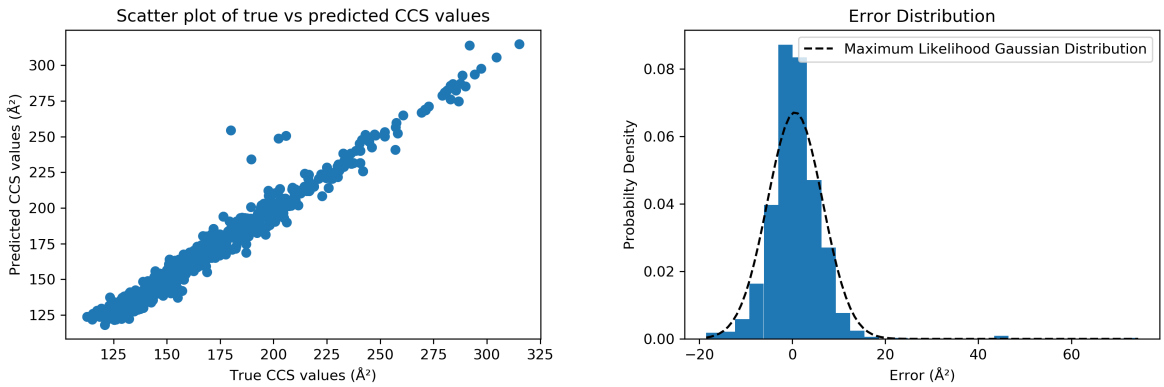
the previous experiments are used. Output kernel is Gaussian kernel on Tanimoto kernel for all experiments. The experiments are repeated for candidate sets based on true molecular formula and molecular mass.

5.2.1 Addition of CCS Kernel(s) to MS/MS Spectra Kernel

Table 4a contains the results of this section. The CCS information is used in addition to MS/MS spectra kernels. Details of the method is given in Section 3.3, where the objective function is obtained by addition of CCS and MS/MS kernel functions. For both ALIGNF experiments, the weights for CCS kernels are investigated and were below 10^{-6}. Uniform kernel combination (UNIMKL), even though performing better than ALIGNF, could not outperform the baseline.

5.2.2 Prediction of CCS Values of Molecular Candidates

The details of the model we used to predict CCS values are given in Section 3.2. We trained 10 different models for 10 different cross validation folds, so that, during training,



(a) (b)

Figure 5: (a) Scatter plot of predicted vs. true CCS values and (b) error histogram and fitted Gaussian Distribution ($Normal(0.60, 5.94)$) by maximizing the likelihood for 1045 input metabolites.

no model has seen the true 2D structures that are present in the related fold. We only trained the dense layers, keeping the convolutional layers unchanged. Models are trained using 2359 (molecule, CCS) tuples at average.

We used the 3D molecular structure (isomeric SMILES) of the candidates and the ground-truth ionization mode of the corresponding MS/MS measurement to predict their CCS values. Figure 5a shows the scatter plot for predictions and Figure 5b shows the error distribution. Due to unseen SMILES structure during the training of convolutional network, the model was unable to predict for 26324 candidates in the candidate sets based on molecular mass and 362 candidates in the candidate sets based on molecular formula. These number corresponds to 0.56% and 0.01% of all candidates, respectively. Also, prediction of CCS values for candidates was only possible for Ionization Modes with adduct configuration M+H, M+Na due to insufficient number of training samples for other adduct configurations. The predicted CCS value for candidates are always on the same ionization mode as the related input compound's. Mean Relative Error is 2.40% and relative error is less than 12% except for 4 outliers. R^2 is 0.97 and Spearman correlation is 0.98.

5.2.3 Analysis of Predicted Candidate CCS Values

Figure 6 shows the box plots of predicted CCS values of 25 randomly selected candidate sets based on true molecular formula and molecular mass. It can be seen that even though CCS predictions range from roughly 120 to 300, the individual candidate sets are concentrated on much smaller windows. In other words, the predicted CCS values of candidates which has the same molecular formula or the mass are more concentrated. Moreover, predicted

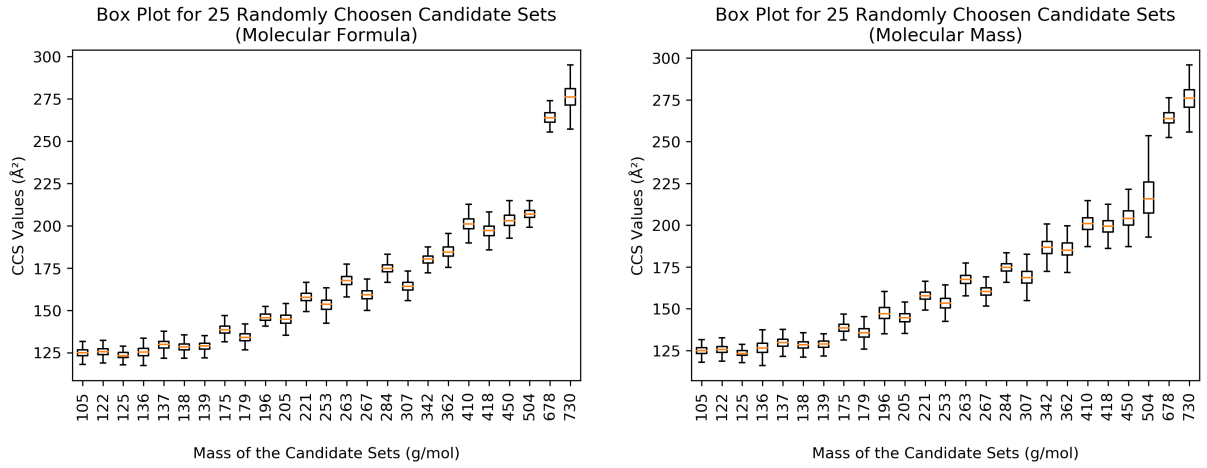


Figure 6: Box plot for predicted CCS values of 25 randomly selected candidate sets. The candidate sets on both plots are for the *same* input compounds. Candidate sets are sorted by their mass which is calculated from the molecular formula of the corresponding input compound.

CCS values seem to be correlated with mass and the variation of predicted CCS values for candidates sets based on molecular mass appear to be slightly larger than those based on molecular formula.

Assuming that the estimator has the *unbiased* error distribution of fitted distribution on errors (see Figure 5b, $Normal(0, 5.94)$), we have calculated the Z (standard) scores of all candidates. More specifically, $g(y)$ being the CCS prediction of a molecular candidate $y \in \mathcal{M}(x')$, z' being the CCS measurement of related input compound and σ_{error} being the standard deviation of the assumed error distribution, the Z score of a candidate is calculated using the following formula:

$$Z_{score}(y) = \frac{g(y) - z'}{\sigma_{error}} \tag{6}$$

Under this error distribution assumption, the predicted CCS value of a true candidate is a random variable with distribution $Normal(z', \sigma_{error})$. The same operation given in above equation corresponds to standardization of the distribution of this random variable. Then, all true candidates would have predicted CCS values with distribution $Normal(0, 1)$, which enables us to compare the distribution of predicted CCS values of true candidates with the distribution of predicted CCS values of all candidates

The distribution for Z (standard) scores for all candidates are given in Figure 7 together with a standard normal distribution. From the figure, we see that the Z-scores of candidates

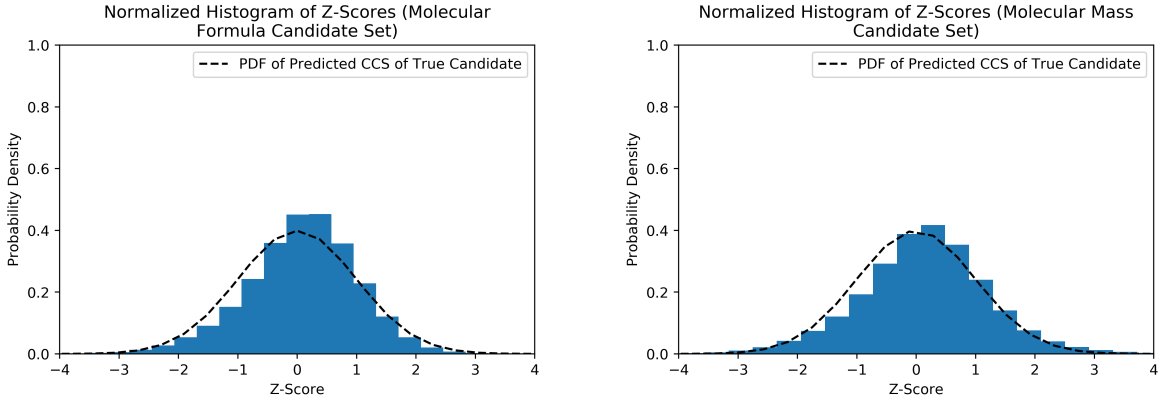


Figure 7: Normalized histograms of Z (standard) scores of predicted CCS values of candidates (See Equation 6) with probability density function of standard normal distribution. Z-scores are calculated using the standard deviation of the error distribution given in Figure 5b. Only 4.10% of the candidates based on true molecular formula are out of the $95,5\%$ confidence interval and 7.37% of the candidates based on molecular mass.

are almost distributed according to standard normal distribution. Thus, we can say that in general the CCS predictions of candidates *almost* follow the same distribution as the CCS prediction of true candidates. This makes it difficult to distinguish candidates based on their predicted CCS values. In addition, the candidates based on molecular mass have higher variation in Z-scores compared to candidates based on molecular formula, showing that molecular formula is a stronger restriction than molecular mass and predicted CCS values has more orthogonal information to molecular mass than molecular formula.

We investigated the predicted CCS distribution of stereo-isomers of input compounds. Here, by stereo-isomer we mean molecules sharing the same 2D structure, but with different 3D structures (see Section 4.1). Figure 8 shows the histogram of relative distance in predicted CCS values of stereo-isomers of input compounds and all candidates based on molecular mass. Relative distance in predicted CCS value of a candidate is computed as follows:

$$Relative(y) = 100 \cdot \frac{|g(y) - z'|}{z'} \tag{7}$$

From the figure, we can see that the distribution of relative distances of all candidates appear to have slightly higher variation than that of stereo-isomers. In addition, stereo-isomers in a small mass interval have predicted CCS values closer to input compounds than an average molecule in both candidate sets.

We have compared random identification of stereo-isomers with CCS prediction based

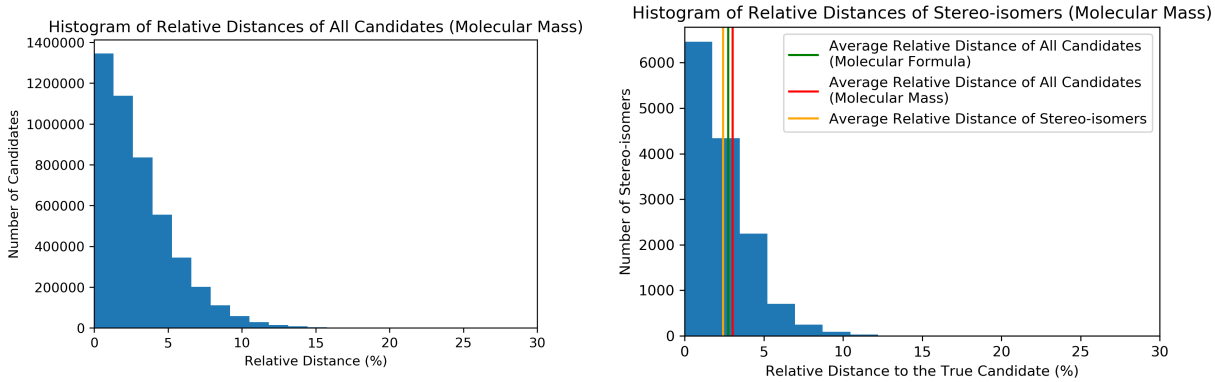


Figure 8: Histograms of relative distance (See Equation 7) of predicted CCS value of all candidates based on molecular mass and stereo-isomers of input compounds.

Table 5: Random vs. predicted CCS based identification of stereo-isomers of input compounds. The equation for CCS based identification is provided in Equation 8 Best results for each column are in bold.

Method	**Top-1**	**Top-5**	**Top-10**	**Top-20**
Random	43.9%	74.93%	84.18%	90.85%
CCS based	**44.0%**	**76.93%**	**86.27%**	**92.56%**

identification. Since MS/MS spectra cannot distinguish stereo-isomers, we assumed they would be randomly scored. Thus, this experiments gives a better insight if the CCS predictions can help to distinguish stereo-isomers and provide additional information to MS/MS spectra. CCS based identification is computed using the following formula:

$$y' = \underset{y \in \mathcal{M}(x')}{\arg\min} |g(y) - z'| \tag{8}$$

Table 5 provides the results for this experiment. We can see from the results that with predicted CCS information, identification is slightly improved.

5.2.4 Candidate Filtering Using Predicted CCS values of Candidates

We have calculated True Positive Rate (TPR) and False Positive Rate (FPR) for all candidate sets with different common thresholds, considering true candidates as positive instances (total of 1045) and false candidates as negative instances (nearly 4.7 million for candidates sets obtained by mass filtering). Expected behaviour of the filtering is reducing the number of false positives (low FPR), while keeping the number of true positives as high as possible (high TPR). The ROC curves for filtering based on a threshold on true

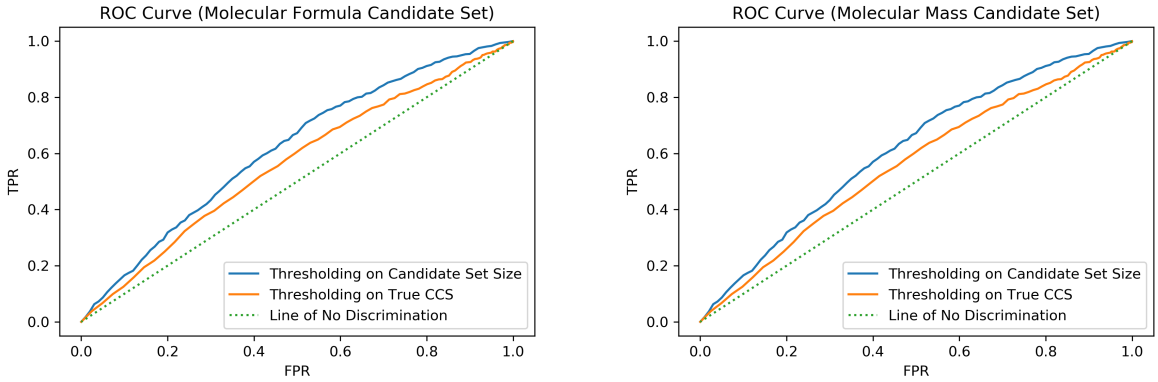


Figure 9: ROC curves for different thresholds on true CCS value of the input compound in the range $[0\%, 40\%]$ and for different thresholds on relative threshold on candidate set size in the range $[0\%, 100\%]$.

CCS and candidate set size (See Section 3.4.1) are given in Figure 9. Candidates with unpredicted CCS values (See Section 5.2.2) are always kept inside the filtered candidate sets for thresholding on true CCS and removed before the filtering, i.e. they are not included in the calculation of FPR for thresholding on candidate set size.

The linear line on the ROC curves, known as "line of no-discrimination", is the expected behaviour of a random filtering. Thresholding on the candidate set size performing slightly better, both approaches are above the line of no-discrimination, for both candidate sets. Thus, we can infer that we have orthogonal information to both true molecular formula and molecular mass. However, the method eliminates too many true candidates in exchange for filtering false ones to effectively restrict the candidate sets before IOKR identification.

5.2.5 Molecular Candidate Re-ranking

Table 4b summarizes the results of this section. We have used Gaussian Kernel on CCS values, and the scores of the baseline results, i.e. scores based only on MS/MS spectra, in both methods. Candidates with unpredicted CCS values (See Section 5.2.2) are removed.

For the re-scoring by multiplication method, we have used the logistic function as the mapping function (See Section 3.4.2), which makes the equation of the second task of IOKR as follows:

$$y' = f(x', z') = \underset{y \in \mathcal{M}(x')}{\arg\max} \frac{1}{1 + e^{-\langle h(x'), \psi(y) \rangle_{\mathcal{F}_y}}} \cdot e^{-\gamma(g(y) - z')^2} \tag{9}$$

The choice of the steepness of the logistic function is arbitrary and 0 is the theoreti-

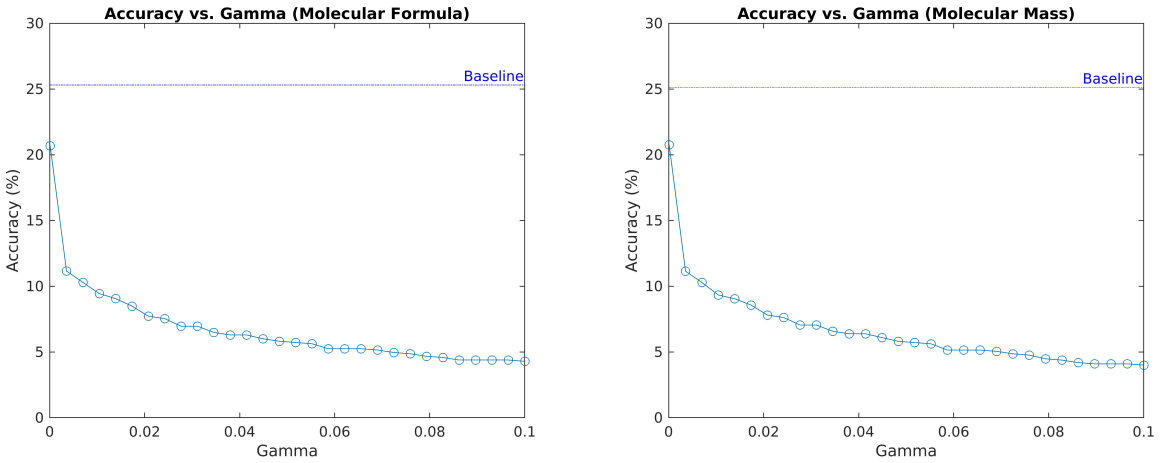


Figure 10: Accuracy as a function of γ parameter of Gaussian Kernel on CCS values used in Re-ranking by Multiplication method (See Equation 9). Circles correspond to calculated points.

cal midpoint of the scores; thus, used as midpoint of the function. The γ parameter of the Gaussian kernel is chosen by grid search from 30 linearly separated values in range $[10^{-4}, 0.1]$, by maximizing the number of correctly identified compounds on complete data. Chosen gamma parameter is 10^{-4} for both candidate sets.

The accuracy as a function of the parameter is given in Figure 10. The γ parameter of a Gaussian Kernel can be interpreted as the *sensitivity* of the similarity measurement. The lesser the γ parameter of a Gaussian kernel, the less important the difference between the feature vectors. For example in our case, Gaussian Kernel with $\gamma = 10^{-4}$ was almost always 1 for different candidates, reducing the influence of predicted CCS values. We see that the higher the sensitivity of the CCS kernel, the less accurate predictions become.

The equation of the second task of IOKR for the re-scoring by addition method is:

$$y' = f(x', z') = \underset{y \in \mathcal{M}(x')}{\arg\max} \left((1 - w) \cdot \langle h(x'), \psi(y) \rangle_{\mathcal{F}_y} + w \cdot e^{-\gamma (g(y) - z')^2} \right) \tag{10}$$

The γ parameter of the Gaussian kernel is chosen by grid search from 30 linearly separated values in range $[10^{-4}, 0.1]$, by maximizing the number of correctly identified compounds on complete data. The weight w for the scores are chosen by grid search as well, from 20 linearly separated values in range $[0, 0.5]$. The chosen γ is 10^{-4} and the chosen weights w are below 0.25 for both candidate sets. Re-scoring by addition method performing better, we did not observe a consistent increase in the results.

Figure 11 gives accuracy as a function of γ parameter and weight of Gaussian Kernel. We

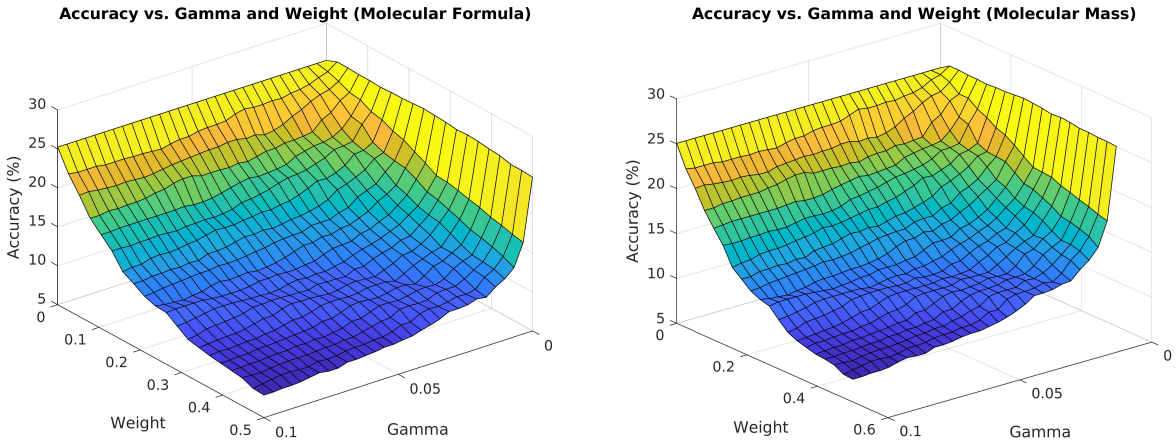


Figure 11: Accuracy as a function of γ parameter of Gaussian Kernel on CCS values and weight of the CCS kernel used in Re-ranking by Addition method (See Equation 10).

see a decrease in accuracy with an increase on the weight and gamma parameter, which confirms the conclusions of re-ranking by multiplication method. The slight increase in the performance might be related to over-fitting, since the grid search of the parameters are directly done on complete-data.

6 Conclusion

In this study, we inspected the use of CCS measurements and predictions to improve metabolite identification in addition to molecular formula, molecular mass and MS/MS spectra. Even though there is orthogonal information in the CCS values to both molecular formula and mass, we were not able to effectively utilize this information to improve our identification method based on MS/MS spectra. This might be due three reasons:

- There might be limited amount of orthogonal information to MS/MS spectra in constructed datasets.
- The CCS prediction might not be sufficiently accurate.
- The combination methods for MS/MS spectra and CCS values might not be "good" enough to discover the orthogonal information.

However, it should be noted that the MS/MS spectra is a very rich source of information for a molecule. It contains information regarding the fragmentation pattern of a

molecule which is much more informative than only mass, molecular formula or CCS of the molecule. Therefore, CCS measurements of molecular fragments may still provide additional information to MS/MS spectra. Moreover, wider CCS datasets and more accurate CCS predictions could provide deeper insight regarding the question of whether MS/MS spectra based metabolite identification can be improved using collision cross section values or not.

Glossary

high-throughput screening Measuring a large amount of different molecules from the same biological in within one experiment.. 3

molecular candidates Potential molecular structures, that *could* be, i.e. "are candidates", the correct molecular structure of an unknown molecule. The set of molecular candidates, can be defined from large molecular structure databases such as PubChem.. 3, 4, 7, 14

molecular formula Description of a molecule in terms of the amount of different atoms in it. For example: Water has the molecular formula H2O, i.e. two hydrogen and one oxygen atom.. 7

molecular structure Description of a molecule in terms of the geometrical arrangement of its atoms, their connections and their distribution in space. The molecular structure can be used to visualize the molecule.. 3, 7, 12–14

References

Brouard, C., Shen, H., Dührkop, K., d'Alché Buc, F., Böcker, S., and Rousu, J. (2016). Fast metabolite identification with input output kernel regression. *Bioinformatics*, 32(12):i28–i36.

Cichonska, A., Pahikkala, T., Szedmak, S., Julkunen, H., Airola, A., Heinonen, M., Aittokallio, T., and Rousu, J. (2018). Learning with multiple pairwise kernels for drug bioactivity prediction. *Bioinformatics*, 34(13):i509–i518.

Cortes, C., Mohri, M., and Rostamizadeh, A. (2012). Algortihms for learning kernels based on centered alignment. *Journal of Machine Learning Research*, 13:795–828.

Dührkop, K., Fleischauer, M., Ludwig, M., Aksenov, A. A., Melnik, A. V., Meusel, M., Dorrestein, P. C., Rousu, J., and Böcker, S. (2019). Sirius 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat Methods*, 16:299–302. Doi 10.1038/s41592-019-0344-8.

Dührkop, K., Shen, H., Meusel, M., Rousu, J., and Böcker, S. (2015). Searching molecular structure databases with tandem mass spectra using csi:fingerid. *Proceedings of the National Academy of Sciences (PNAS)*, 13:12580–12585.

Kanu, A. B., Dwivedi, P., Tam, M., Matz, L., and Hill Jr., H. H. (2008). Ion mobility–mass spectrometry. *Journal of Mass Spectrometry*, 43(1):1–22.

Nichols, C. M., Dodds, J. N., Rose, B. S., Picache, J. A., Morris, C. B., Codreanu, S. G., May, J. C., Sherrod, S. D., and McLean, J. A. (2018). Untargeted molecular discovery in primary metabolism: Collision cross section as a molecular descriptor in ion mobility-mass spectrometry. *Analytical Chemistry*, 90(24):14484–14492.

Paglia, G., Williams, J. P., Menikarachchi, L., Thompson, J. W., Tyldesley-Worster, R., Halldórsson, S., Rolfsson, O., Moseley, A., Grant, D., Langridge, J., Palsson, B. O., and Astarita, G. (2014). Ion mobility derived collision cross sections to support metabolamics applications. *Analytical Chemistry*, 86(8):3985–3993.

Plante, P.-L., Francovic-Fontaine, E., May, J. C., McLean, J. A., Baker, E. S., Laviolette, F., Marchand, M., and Corbeil, J. (2019). Predicting ion mobility collision cross-sections using a deep neural network: Deepccs. *Analytical Chemistry*, 91(8):5191–5199.

Tejada-Casado, C., Hernández-Mesa, M., Monteau, F., Lara, F. J., del Olmo-Iruela, M., García-Campaña, A. M., Bizec, B. L., and Dervilly-Pinel, G. (2018). Collision cross section (ccs) as a complementary parameter to characterize human and veterinary drugs. *Analytica Chimica Acta*, 1043:52 – 63.

Zhou, Z., Shen, X., Tu, J., and Zhu, Z.-J. (2016). Large-scale prediction of collision cross-section values for metabolites in ion mobility-mass spectrometry. *Analytical Chemistry*, 88(22):11084–11091. PMID: 27768289.