Automated classification of diglucosides glycosidic linkage with ion mobility spectrometry data by machine learning approaches

Michal Radecki

A thesis in the field of machine learning and biochemistry

for the Masters by Research degree in Medical Sciences

Lancaster University

September 2020

Abstract

Unambiguous characterization of carbohydrate products remains a challenging endeavour. The current state-of-the-art techniques are NMR-based approaches, which require large amounts of purified sample and are challenging to annotate. Recently, a strategy has been developed combining gas-phase ion-mobility spectrometry with tandem mass spectrometry to separate and characterize isomeric product ions. Crucially, this can provide information about the stereochemistry that MS alone is often “blind” to because certain monosaccharides have the same *m/z* (i.e. they are isomeric) and are therefore indistinguishable. Given the amount of data this approach produces, there is a need for a method to rapidly annotate the produced data. The initial strategy involves developing a method that can discern the number of peaks within an IMS spectrum, where it has been shown that product ions derived from α-glucosides produced similar features (2 peaks) whereas β-glucosides only produced a single peak. It was reported that an IMS signal can be approximated as a sum of spectral line shapes (such as Gaussian, Lorentzian or Voigt). Current results show that the approximation method allows analysis of the signal in terms of peaks description. Using this approach, we built a feature matrix from IMS data for different diglucosides, which was then used to train a machine learning classifier able to distinguish between α- and β-glucosides. The performance of the classifier proved that automated classification of glucosides by their bonding type is achievable.

Table of Contents

1. Introduction **5**

1.1. Background 5

1.1.1. Carbohydrates overview 5

1.1.2. Mass spectrometry 6

1.1.3. Ion mobility spectrometry 10

1.1.4. Travelling wave ion mobility spectrometry 11

1.2. Motivation of the study 11

1.3. Thesis outline 14

2. Literature overview **15**

2.1 Catalogue bases approaches 15

2.2 Brute force approaches 16

2.3 *De novo* sequencing approaches 16

2.4 Classifying a subgroup of glycans 18

2.5 Conclusion 18

3. Research objectives and formal questions **19**

4. Methodology **20**

4.1. Model 21

4.2. Threshold selection 22

4.3. Least-squares 23

4.4. Minimisation using Levenberg-Marquardt algorithm 24

4.4.1. Gradient descent 24

4.4.2. Gauss-Newton method 25

4.4.3. Levenberg-Marquardt algorithm 25

4.5. Initial parameter estimation 26

4.6. Feature extraction algorithm 28

4.7. Classification using Gaussian naïve Bayes classifier 29

5. Results 30

5.1. Example fits 30

5.2. Feature matrix 32

5.3. K-fold validation 32

5.4. Accuracy 33

5.5. AUC-ROC curve 34

6. Conclusions 35

6.1. Research objectives 35

6.2. Formal requirements 36

6.2. Future work 36

7. Literature overview 37

Appendix 41

1. Introduction

Carbohydrates and their analogues represent one of the largest groups of biomolecules found within nature. Methods have been developed capable of providing structural information of proteins [1] or nucleotides [2] in an autonomous, robust and efficient manner in a various area of applications. However, due to the complexity of carbohydrates structure and lack of technology providing sufficiently informative data, carbohydrate sequencing is still an open research question. In this study, it is shown that the latest findings related to ion mobility spectrometry have a great potential to make contribution to the carbohydrates sequencing challenge. The first section introduces the background knowledge necessary to comprehend the study, research questions and objectives.

1.1 Background

This subsection provides information regarding general knowledge on carbohydrates and experiments used to generate data throughout the project – mass spectrometry (MS), tandem mass spectrometry (MSMS) and ion mobility mass spectrometry (IMS).

1.1.1 Carbohydrates overview

Carbohydrates are ubiquitous biological polymers that are important in a broad range of biological processes. On the whole, carbohydrates exist in a wide variety of sizes, from simple monosaccharides, through oligosaccharides (typically less than 20 building blocks), to polysaccharides (even more than 20 basic units). Glycans (a compound consisting of a large number of monosaccharides linked glycosidically [3]) can be described by its composition, connectivity, and configuration (Fig. 1.). Monosaccharide can also differ by the number of atoms that they contain: triose (3), tetrose (4), pentose (5), hexose (6), heptose (7), etc. Carbon atoms are numbered starting from 1 along the backbone. Simple linear monosaccharides with unbranched skeleton have one hydroxyl group (OH) and one carbonyl group (O=H) attached to one of the carbon atoms. When the carbonyl group is attached to the carbon atom at position one (aldehyde), the monosaccharide is defined as aldose. It might be the case that the carbonyl group is attached between two carbons (ketose) and the monosaccharide of this type is a ketose – usually they have the carbonyl group at position 2. Following this pattern, one can define more classifications of monosaccharides resulting in names such as 'aldohexose', 'ketoriose', etc.

Many simple monosaccharide building blocks often differ by the stereochemistry at a single position, as it is in the case of glucose and galactose. A pair of carbohydrates, which possess this characteristic, are defined as stereoisomers. The connectivity of glycans is defined by the point of attachment of their building blocks. Each monosaccharide hydroxyl group can be a potential location of connection to the next monosaccharide. It must be stressed out, that carbohydrates, unlike proteins or lipids, can be considered nonlinear on account of this attribute. It means that they are more like branched structures with diverse regiochemistry rather than linearly connected elements. Furthermore, two monosaccharides can be connected in two different configurations– α- or β-linkage, which is a further stereochemical difference. Potentially, glycans can also be attached to other biopolymers, such as lipids, proteins or other organic molecules, in the process of glycosylation. In this process, at least three (the most abundant) classes of glycans can be produced: N-linked glycans [4], O-linked glycans [5], phosphoglycans [6]. Moreover, glycans can vary by their charge. They can be extremely negatively charged (e.g. glycosaminoglycans) or neutral.

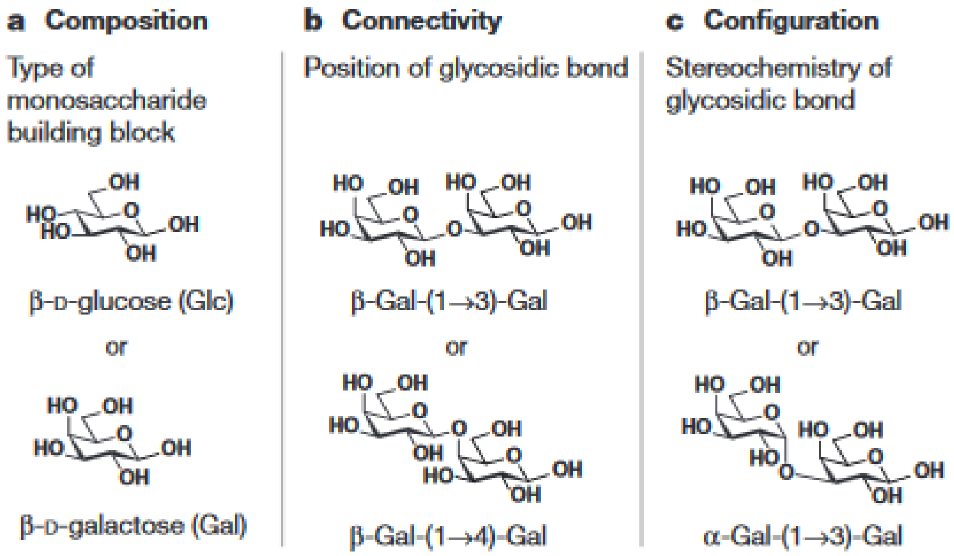


Figure 1. Structural characteristics of glycans. a) Composition is defined by a single building block, which can differ in a position of single carbon atom. b) Hydroxyl group, which is the connection point between building blocks defines the connectivity of a glycan. c) Each connection between monosaccharide, i.e. glycosidic bond, occur in two different configurations - α- or β-configuration.

All of the stated characteristics make carbohydrates potentially highly information-rich molecules, *i.e.* a carbohydrate can be defined by a high number of complex features, not only quantitative but also spatial. Despite the fact that the current technologies allow extracting a variety of data describing glycans, a manual annotation is still a challenging task, particularly defining all glycan characteristics at once. Current approaches [literature] relies on tandem mass spectrometry data (MS) (Fig. 2.b.), which provides information on mass/charge (*m/z*) and abundance of ions of a given sugar, which can be characterised by a reference to standard depending on an analysed compound. Since MS provides a numerous number of signals strongly correlated with the composition of a glycan, it is hard to retrieve the part of signal defining a possible stereochemistry of a compound. Therefore, MS is found to be inefficient (nevertheless, potentially useful) in distinguishing between α-linkage and β-linkage, thus the connectivity. However, a recently introduced approach which combines MS and ion mobility spectrometry (IMS) [7] has been reported to have a great potential to solve this problem. IMS discriminates ions based on their mobility which is inversely proportional to the rotationally averaged collision cross section area (CCS)-to-charge ratio, an intrinsic property of an ion’s structure. [8] This data has been showed to contain information on the stereochemistry of glycosidic bonds between monosaccharide building blocks, which is the main motivation of the study. More information on MS and IMS will be provided in the following subsections.

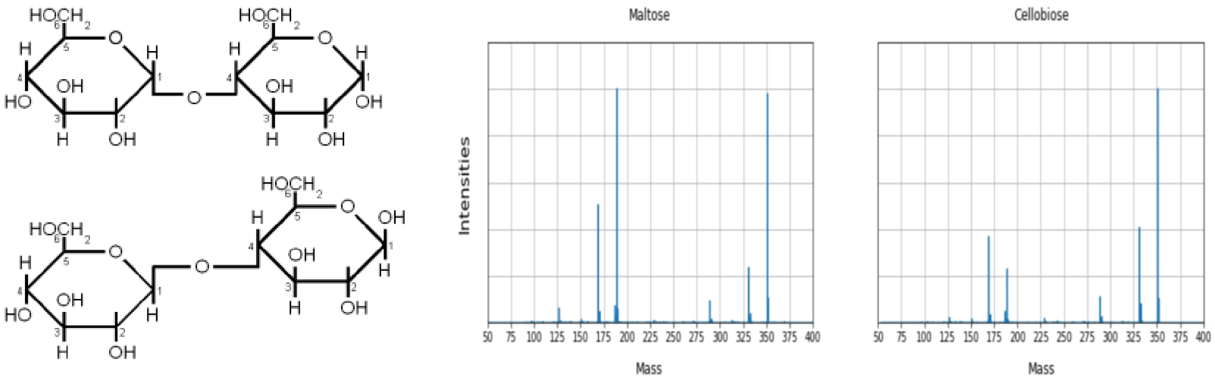


Figure 2. A) Two examples of disaccharides with different types of glycosidic bonds: 1. maltose - α-linkage, cellobiose – β-linkage. Despite the fact that their connectivity and composition is identical, they are still stereoisomers due to the different bonding type. B) Example of mass spectrometry data, which is a traditional approach to carbohydrate sequencing. Since it provides information only of the mass aspect of a given test sample, it produces an extremely similar output for stereoisomers, e.g. maltose and cellobiose.

1.1.2. Mass spectrometry

Mass spectrometry (MS) is one of the most popularized and the most universal analytical technics. It is used in a wide area of fields (both academic and commercial), *e.g.* pharmaceutical, clinical, environmental, geological or biotechnical. The major advantage of MS is its sensitivity, i.e. the small changes in an analysed compound produce a significant difference in the produced spectrum, which makes the small difference detectable. Moreover, MS is considered to be relatively rapid. Furthermore, the fact that it might be coupled with other techniques such as liquid chromatography (LC-IMS) [9], ion mobility spectrometry (IMMS) [10] or MS itself – tandem mass spectrometry [11]. It also possesses an an ability of dealing with mixtures (in some cases).

MS is used to define a composition of a given sample. There are many types of mass spectrometers, however, all of them can be generalized to the following steps:

1. Ionization of analyte into the gas-phase – the sample is vaporized by a heater and bombarded by high-energy electrons. Some of the electrons in the atoms of the sample are knocked out, which produces positively charged ions. The ions are accelerated so that they all have the same kinetic energy. Then, they are deflected by a magnetic field according to their masses. Ions with lower masses are deflected more and the ions with higher masses are deflected less. There are soft ionization techniques like ESI and MALDI, which allow the entire molecule to enter the gas phase with no dissociation and hard ionization such as EI, which dissociate the analyte during ionization.
2. Separation based on m/z and sorting of the ions – done by the mass analyser (*e.g.* a ToF, quadrupole and quadrupole ion traps - quadrupole can be very useful as they can be used as a mass filter only letting fragments with a particular *m/z*).
3. Detection of the presence of ions by a detector and abundance of ions within a packet.
4. The electronics calculate the *m/z* and abundance of the detected packet. Mass-to-charge ratio is proportional to the deflection of an ion. The abundance of a packet is the number of ions with particular *m/z* divided by the overall number of ions within a packet. The data is plotted and displayed to the user as intensity of ions *vs m/z* (*e.g.* Figure 2. B)).

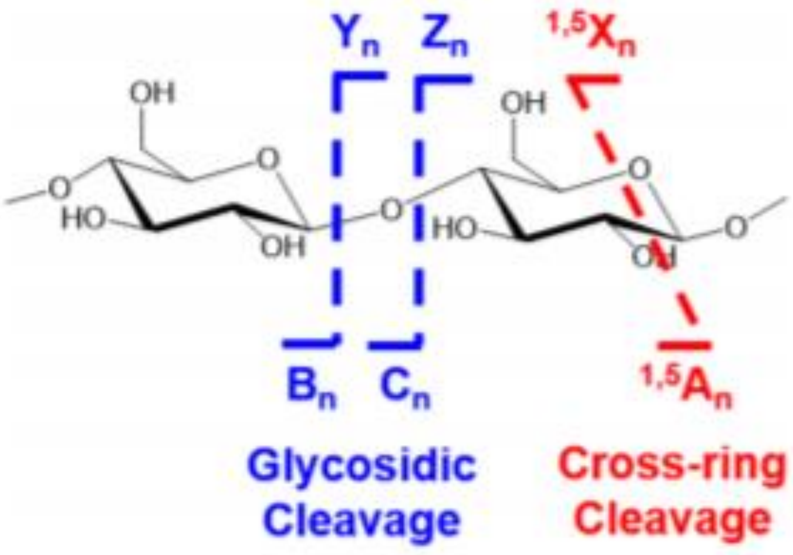


Figure 3. Domon-Costello nomenclature [12] for common fragments generated by dissociation techniques. This kind of dissociation might be achieved using tandem mass spectrometry. The fragmentation showed in the further subsections (Figure 4) is derived from this nomenclature.

Combined with presumed knowledge of on the analysed compound, MS profiles of carbohydrates can show their composition. Tandem mass spectrometry (MS/MS) is particularly useful to dissociate glycans into smaller fragments that can reveal structural motifs directly. Glycans dissociate either across the glycosidic bond (forming B-, Z-, C-, and Y-ions) or across the monosaccharide ring (A- and X-ions) (Fig. 3). This kind of experiment reveals broad sequence and composition information, but it does not directly hold any information on stereochemical assignment of the monosaccharide building block. If only disaccharides (or basically glycans of relatively small size) were taken into consideration that would be somehow possible to classify them by their linkage. A spectrum could be compared to a reference molecule of a known linkage fragmented under the same conditions – this would leave a spectral 'fingerprint' of stereochemistry of the linkage. However, spectral matching will fail with larger compounds. Only three monomeric units: mannose, galactose and glucose can produce 144 possible disaccharides – in the case of trisaccharides this will lead to significantly larger number, counted in 10's of thousands. This fact makes the spectral matching inefficient, which will be discussed in the next chapters. However, the conclusion is that identifying linkage type using MS is achievable, but only for small molecules. In contrast, the quality of features obtained from IMS data is not depended on the molecule size.

1.1.3. Ion mobility spectrometry

Ion mobility spectrometry (IMS) is an atmospheric pressure technique for trace analysis of gas-phase analytes. During the last decade IMS has come of age as an analytical method with over 50000 stand-alone ion mobility spectrometers currently employed throughout the world [7]. The technique has a great potential to be coupled with MS due to the fact it reveals the stereochemistry of a compound, which provides orthogonal information to MS data - thus it is introduced in the study. Traditional ‘drift-time’ ion mobility spectrometry (IMS) measures the time that it takes an ion to migrate through a buffer gas in the presence of a low electric field. The primary condition of this low electric field is that the thermal energy supplied from the collisions of the buffer gas is greater than the energy the ions obtain from the electric field. Thus, the ions have energies similar to that of the buffer gas and diffusion processes are dominant. Ion mobility under low-field conditions can be thought of as ‘directed diffusion.’ Under these low-field conditions, the velocity of the ion is directly proportional to the electric field. This proportionality constant is called the ion mobility constant and is related to the ion’s collision cross-section (CCS), which allows distinguishing structural features of ions in gas phase [8]. CCS can be obtained using Mason-Schamps equation [13] and it defined as the area around a particle in which the centre of another particle must be in order for a collision to occur. Literally speaking, this measure be compared to a resistance – the bigger CCS a molecule has, the drift time is lower, like it was harder for it to get through the gas.

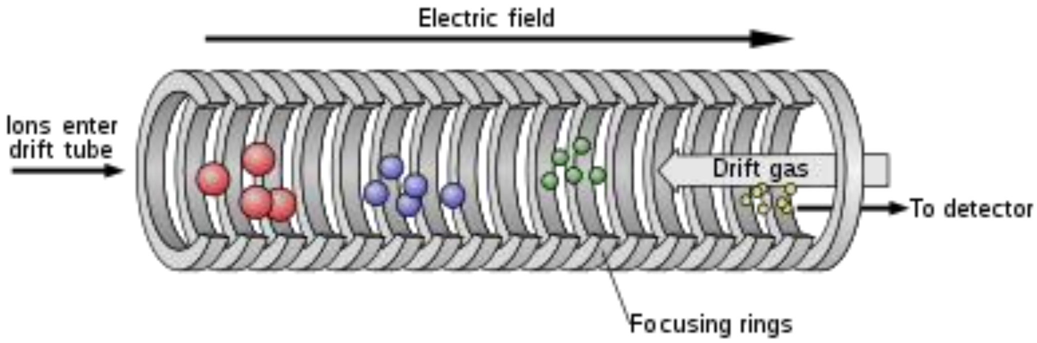


Figure 4. A general example of IMS mechanism. Ions with different mobilities have different drift times, which can be recorded by the detector.

1.1.4. Traveling wave ion mobility spectrometry

Traveling wave ion mobility spectrometry (TWIMS) [14] uses a non-uniform drift field. In this a sequence of symmetric potential waves continually propagating through a tube propels ions along with velocity different species transit the tube in unequal times. As the waves pass along the device, ions can ‘surf’ on the wave front for a period of time before being overtaken by the wave. Ions are separated as they travel through the device as higher mobility ions undergo less ‘roll over’ events on the waves than the lower mobility ions. TWIMS generally allows mobility separation of a similar resolution to occur in a shorter device than DTIMS, resulting in an instrument with a smaller footprint. Whilst ion motion in a TWIMS is relatively complex, the resulting IMS separation can be calibrated to allow the easy, quick measurement of CCS for all of the ions present.

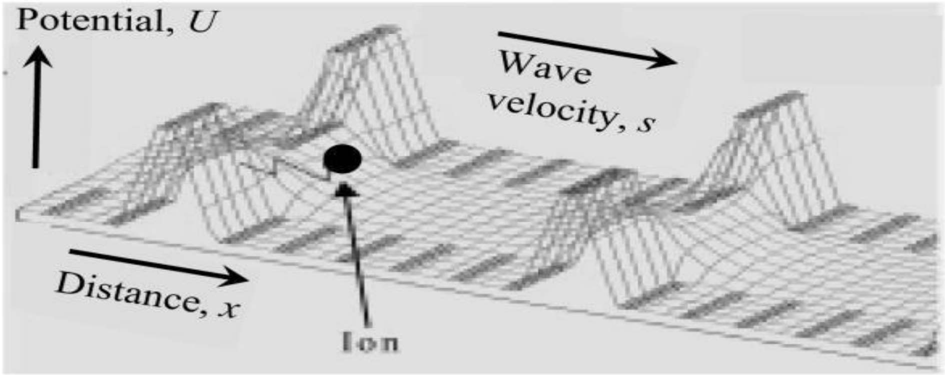


Figure 5. Ion propelling through a drift tube.

1.2. Motivation of the study

The experiment of extracting necessary data was conducted outside of the study [15, 16, 17]. A series of reducing glucoside standards (Glcα/β1-2/3/4/6Glc) **1−8** and different reducing sugars (Glca1-1aGlc and Glca1-1bGlc) **9-10**, which differ only in the regio- and stereochemistry of the glycosidic bond, were analysed with a series of experiments that combined tandem mass spectrometry and traveling wave ion mobility separation (Fig. 2.a. and Fig. 2a.). The instrumentation used in the experiment was SYNAPT G2-S Mass spectrometer (Waters, Manchester, UK). The synapt G2S is based on used nESI ionization [18], quadrupole mass selection (mass analyser) and dissociation-IMS-ToF [19]. Thus, as a generalisation one can define it as tandem mass spectrometry coupled with travelling ion mobility spectrometry.

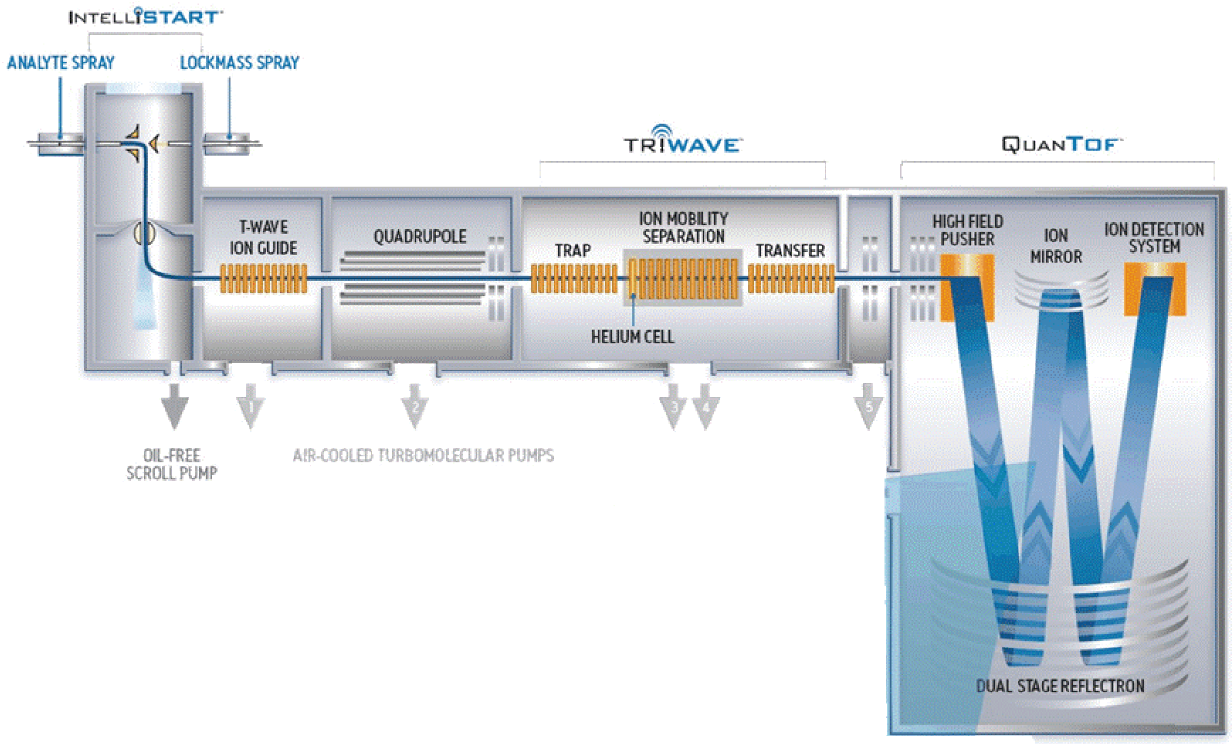


Figure 6. A diagram showing the outline of SYNAPT G2-S.

MS data might have extremely similar features for two epimers (Fig 4.a.), therefore it is relatively hard to distinguish those using MS data on its own. However, one can observe an interesting pattern occurring in IM spectrometry data produced by previously separated ions. The provided example of 12 diglucosides indicates that α-glucose terminating precursor usually produces more than one peak in comparison with its β-glucose analogue. This pattern slightly differs for reducing naturally occurring sugars (Glcα1-4GlcNAc and Glcβ1-4GlcNAc) at Y ion. However, in pattern recognition there are usually biases in the pattern and they are considered as outliers. How those inconsistencies are treated will be discussed in the further sections. In the study, Glcβ1-1βGlc and Glcα1-1βGlc will be omitted. The fingerprint left in IMS data by those molecules is effectively a mixture of the two linkages in a single molecule

– those are co-fragmented and non-separable. Literally, this means that IMS contains mixed spectra of potentially two different glycosidic linkages.

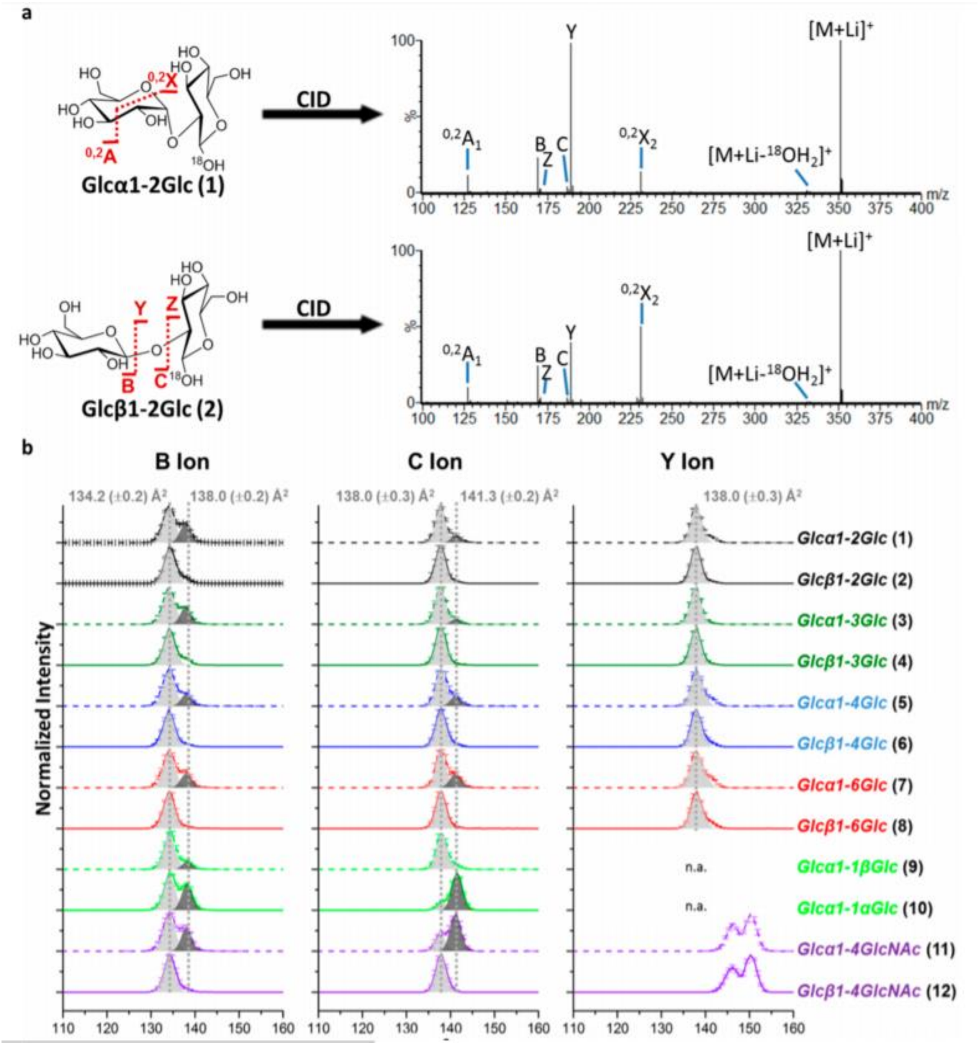


Figure 7. a) An example of diglucosides regioisomers, which MS data is identical. b) IMS data of 12 sugars indicating that IMS holds a footprint of stereochemistry of the glycosidic linkage.

Observing the occurring pattern, one can conclude that the information on the number of peaks in IM spectrometry might allow creating a sufficiently informative dataset for developing a statistical method to classification. The aim of this work is to build a machine learning classifier able to distinguish any provided diglucoside by its bonding type. It has to be mentioned that IMS will provide information on stereochemistry for any group of glycans, not only diglucosides as reported in this study. However, it is also important to stress that this current approach is only applicable for lithiated disaccharides – it has an extremely low intensity for sodiated disaccharides, which makes it hard to analyse them. In the case of different class of glycans the data will be different - the peaks will have different characteristics. They might be more separated or more overlapped, the peaks will differ in height. Nevertheless, the observed pattern will occur, which gives an idea to capture and describe it using analytical methods.

1.4. Thesis outline

The thesis is divided into following sections:

2. Literature overview – analysis and comparison of existing approaches to automated carbohydrates classification.
3. Research questions – after studying the current literature the research questions are formulated.
4. Methodology – a concept of extracting information from IMS data and using it for glycosidic bond classification is studied in this section. More particularly, the main ideas behind it are threshold selection, signal approximation by parameter optimisation, feature selection from the optimised parameters, training and evaluating a machine learning classifier.
5. Results – the results of the curve fitting, feature extraction and classification are provided, *i.e.* plots of fitted functions, feature matrix, accuracy and AUC curve. The results are discussed.
6. Conclusion – the research objectives are reviewed and matched with the research outcomes. Moreover, the potential possibilities of extending the model are defined.
7. Appendix – plots of all IMS signal function approximations.

2. Literature overview

Relatively few attempts have been made in order to automate the process of carbohydrate sequencing. One particular challenge is classification of the stereochemistry of the glycosidic linkage. Most of the introduced approaches only have an ability to successfully classify composition (*i.e.* classify monosaccharides defining the building blocks) of the carbohydrates, providing little to no information on connectivity. In this section the existing endeavours to automated glycans classification will be introduced, analysed and compared. The first three subsections describe algorithms capable of classifying all of the possible glycans. The fourth subsection compiles the approaches, which aim to label a specific subgroup of carbohydrates. The conclusions made as the summary of the section along with the assumptions defined in the introduction will allow formulating the research questions.

2.1. Catalogue based approaches

The first methods [20-22] of glycans classification used a predefined database of MS spectra. Those techniques rely on a simple search made on the database. Given an unknown MS spectrum, the algorithm explores the whole storage space and compares the input with each entry. The one with the highest score, which can be a simple distance metric (such as Euclidean distance between each point in the spectra), is considered to be the result. Preferably, the algorithm output is a list of scores of a fixed number of potential candidates with the highest results. Those approaches have been found to be efficient and robust since they reach up to 100% accuracy in terms of the composition of the glycans. However, the need of the previously defined library of glycans is the main limitation of these methods. The number of possibilities is generally too high to experimentally capture all of them and store in a database. Also, similarities between reference MS spectra might lead to an increased number of false positives in the classification. There were multiple attempts made to create a glycans database: Glycan Mass Spectral DataBase [23], EuroCarbDB [24] and many others. However, none of them tries to describe all of the existing carbohydrates and they rather specialize in a certain group. This fact makes the sequencing of undefined samples impossible since it is not known, which database should be chosen as the reference in matching. Moreover, as a consequence of the usage of MS data these techniques are unable to classify the type of glycosidic linkages between the building blocks, but only the composition. It would be possible to incorporate IMS data to this type of technique in order to attempt to classify bonding type. However, due to the limitation resulting from the requirement of a database, this approach is still considered to be inefficient since it needs a significant human effort to be made. In conclusion, there is a need to create an automated method, which brings IMS and MS data together, and do not use any reference system – the sequencing should be performed *de novo*.

2.2. Brute force approaches

An exhaustive comparison of a given experimental mass spectrum to all theoretical structures is considered to be a possible brute force approach, *e.g.* STAT [25]. In greater detail, the algorithm generates all of the possible structures of glycans *de novo* as it is running and compares each of them with the input spectrum. This kind of approach has a significant limitation since it is computationally inconvenient to generate a large number of candidate spectra due to the potential complexity of glycans structure. Hence, the user is prompt to enter a series of likely information to be contained in the sample, such as monosaccharides presented in the sample or precursor ion. Without prompting the user to input the data the searching time might be potentially high and/or it might require significant computational sources. Unfortunately, the need of the user input puts an end to the idea of the full automation in this problem, since it makes the approach only partially automated. This drawback certainly limits the technique's applications, where a rapid annotation of newly approached molecules is crucial. For example, in a possible scenario of testing a substance on a production line on the go (rather than testing it by statistical sampling) and without the knowledge of the potential glycans contained the algorithm would not carry its task.

It has to be stressed out that the method might be efficient, but that would be only for relatively small glycans due to the short simulation time of potential candidates. Above all, the algorithm is not able to correctly characterise glycosidic linkage due to MS data limitation. In this study is considered to be main drawback of this approach.

2.3. *De novo* sequencing approaches

In summary to the previously stated methods, one can conclude that there is a need to create a *de novo* sequencing platform, *i.e.* an approach, which does not rely on previously stored (or simulated) data and classify newly approached samples without comparing them with any database. This kind of approaches have been successfully implemented, however, with particular limitations described in this subsection and addressed in this research project.

GlycoDeNovo [26] is a fully automated algorithm, which creates a topology of a glycan given its tandem mass spectrum. GlycoDeNovo builds an interpretation-graph from each glycan tandem mass spectrum by identifying all potential glycosidic fragments (sequence ions) that can be interpreted as a combination of a monosaccharide (root) and one or more previously interpreted peaks (branches) which lead to the assignment of the precursor peak. The algorithm has been proved to be highly efficient in terms of composition classification of glycans. It has achieved a nearly perfect accuracy, oscillating around 95%, which is considered to be distinctly robust performance. Nevertheless, the performance of the algorithm in classifying bonding type is much less than 50% on average, which potentially means that it assigns the glycosidic linkage type in random manner. This phenomenon might have a theoretical proof in the statement that MS/MS data is often 'blind' to information on the bonding type. It needs to be mentioned that this approach performs well for relatively large oligosaccharides – in testing phase a saccharide consisting of 11 building blocks was used. The authors do not imply that the model should be limited by the larger molecules.

GLYCH [27] is an algorithm for glycan characterisation using MS/MS data. The approach consists of three steps. First, the potential bond linkages between monosaccharaides are identified. This is achieved by analysing the appearance pattern of cross-ring ions. GLYCH is the first algorithm mentioned in the work, which address this problem by separating it and providing it an individual methodology. The next step is to use a dynamic programming [28] algorithm to define the most probable composition of a glycan given its mass spectrum. Finally, oligosaccharide structures are re-evaluated, taking into account the double fragmentation ions. The proper evaluation of the algorithm is missing in the study since the experiment involves only 6 different oligosaccharides. Furthermore, the algorithm strongly prefers linear glycans rather than the ones with branching structure. For the test group of glycans, GLYCH has successfully annotated both composition and linkages. In conclusion, the algorithm outperforms all of the previously stated solutions. However, in terms of linkage type classification, the performance limited to a certain subgroup of glycans – linear oligosaccharides. It needs to be stressed that this approach perform well for relatively large oligosaccharides – in testing phase a hexasachharide was used and the authors do not imply that the model should be limited by the larger molecules.

LODES [29] is another technique of glycans sequencing using tandem mass spectrometry. This technique can successfully classify linkages between building blocks. However, it is suited only for relatively small glycans (trisaccharides are evaluated).

One can conclude that there is a need to build a classifier able to distinguish linkage type of glycans without a limitation on their size and with an ability to classify them by their linkage.

Clearly, the first steps are to prove that IMS data can be a starting point of the idea of a full automation in glycan classification. Concluding, in this work the focus will be made on regiochemistry of glycans rather than their composition.

2.4. Classifying a subgroup of glycans

While the previously indicated approaches aimed to classify any given glycan, in this subsection the focus is made on classifying a specific predefined subgroup of carbohydrates. Since the pattern introduced in the first section was observed only for specific diglucosides, it might be important to bring closer the other research works which objective is to annotate a subset of glycans.

Discrimination of disease phenotypes in complex biological samples, such a serum using MS continues to be an active area of research. As an example [30], analysis of single N-linked glycans can lead to discriminate disease phenotypes associated with esophageal adenocarcinoma [31]. In the mentioned work, the samples of glycans were extracted from sera of healthy controls and patients diagnosed with that Barrett's Esophagus (BE), high grade dysplasia (HGD), and esophageal adenocarcinoma (EAC). Then, the samples were analysed with MALDI-IMS-MS. The goal of the research was to train a statistical model, which is able to identify a specific N-glycans in the serum responsible for stated diseases. The results of the algorithm are considered to be highly robust since they reached 100% accuracy for the training set and the model correctly classified 23 out of 26 patients in the prediction set.

In conclusion, there are potential applications of classifying only a certain subgroup of glycans, when these glycans can be connected with a specific phenomenon to be classified. This has been mentioned because in the study only a subgroup is classified (however, the technique might is applicable for variety of glycans as stated in the introduction).

2.5 Conclusion

All of the stated approaches (apart from LODES, which misses an ability to classify larger glycans) are missing an ability of classifying the bonding type between building blocks of oligosaccharides – connectivity. Hence, the main goal of the study is to develop an approach, which utilise IMS data and focus on classifying the linkage type. This should be possible because of the characteristic of IMS spectra described in the introduction section.

3. Research objectives and formal requirements

Per the conclusion of the previous sections, one can form the following research objectives, which are studied in the next sections.

> Research objective 1. Perform feature extraction on the reducing diglucoside standards (Glcα/β1-2/3/4/6Glc) and two other reducing sugars (Glcα1-4GlcNAc and Glcβ1-4GlcNAc) by building a dataset giving information on the number of peaks in IM spectrometry data.
>
> Research objective 2. Train a machine learning classifier able to distinguish a glycosidic bondage type amongst reducing diglucoside standards (Glcα/β1-2/3/4/6Glc) and two other reducing sugars (Glcα1-4GlcNAc and Glcβ1-4GlcNAc). Moreover, evaluate the performance of the classifier.

Furthermore, formal requirements of the implemented model/software are defined.

> I. The user effort of using the classifier should be minimised, *i.e.* ideally, the system shall be fully autonomous – the user should only provide a raw data input and receive a prediction.
>
> II. The system must be robust, *i.e.* it cannot fail when approaching newly discovered types of samples. It is not specified that it shall classify them, but it should be able to discard them.
>
> III. The model should be self-explainable, *i.e.* its parameters shall have an understandable meaning and constraints should be also comprehensible. This will lead to increase the extensibility of the system, as it will be easily interpretable.

4. Methodology

In this section, a possible approach to feature extraction from IM spectrometry data is presented. IMS signal can be approximated using a sum of spectral line shapes like Gaussian, Lorentzian or Voigt. In this work, a Gaussian function is selected to be a model of the signal (Fig, 4). For the purpose of this research problem, there is no important distinction between the chosen model out of those three. The Gaussian function is the simplest and the most popular among the three stated functions. Most importantly, its parameters will allow extracting the information necessary to perform the further classification. More specifically, the aim is to gain knowledge about the difference between the heights of peaks in the signal. This will allow analysing the pattern explained in the introduction section, i.e. α -diglucosides produce a double peak and β-diglucosides produce a single peak when it comes to their IMS spectra.

4.1. Model

The first step of the feature extraction phase is to fit a sum of two Gaussian functions, each corresponding to one component, i.e. a peak in the signal. To understand how fitting this curve translates into the feature extraction process, there is a need to describe the assumed model and its parameters. Normally, a Gaussian function (Eq. 2.) describes a probability distribution, so its scaling factor is set in the way that all values of the function sum up to one, which is a basic law of probability.

$$f(x) = \frac{1}{\sigma * \sqrt{2 * \pi}} * e^{\frac{-(x-\mu)^2}{(2\sigma^2)}}$$

However, for IMS signal approximation purpose (rather than estimating probability distribution), the values are not summing to one since the domain of the signal might be any random interval in the real numbers space. Hence, the scaling factor of the usual Gaussian function is replaced by a parameter h, which defines the height of a peak. The μ parameter, which is originally the mean of a distribution is replaced by x_0, which states the centre of a peak. This replacement has only a symbolic meaning since the parameter has the same definition in case of the Gaussian distribution. Finally, the σ parameter, formerly defining a standard deviation of a distribution, is changed to w, which is half-width of a peak. A half-width of a peak defines a width of the peak in the middle of its height. Summarizing, the following function f_g (Eq. 1) defines a single component in IMS signal.

$$f_g(x; h, x_o, w) = h * 2^{\frac{-(x-x_o)^2}{w^2}}$$

However, the signal might consist of more than one component. Specifically, studying this phenomenon is the goal of the research, i.e. distinguishing between a single peaked signal and double peaked signal. For this purpose, a mixture model F_g is introduced. In statistics and signal approximation, a mixture model is defined as a sum of any number of different functions. In this case, the model F_g can be described as a sum of two previously defined Gaussian functions f_g with parameters h_i, x_{o_i}, w_i, where i is the number of a component.

$$F_g\left(x; h_1, x_{o_1}, w_1, h_2, x_{o_2}, w_2\right) = f_g\left(x; h_1, x_{o_1}, w_1\right) + f_g\left(x, h_2, x_{o_2}, w_2\right)$$

Summarising, the defined model has six parameters $h_1, \sigma_1, w_1, h_2, \sigma_2, w_2$. To extract the previously stated information from IMS signal, the h_1, h_2 parameters need to be approximated to correctly describe the height of a single component. More particularly, a difference between these two parameters can capture the previously stated phenomenon, i.e. if the differences are significantly high one can say that there's a single peak in IMS data or if the difference is low, it is assumed that there are two peaks in the signal.

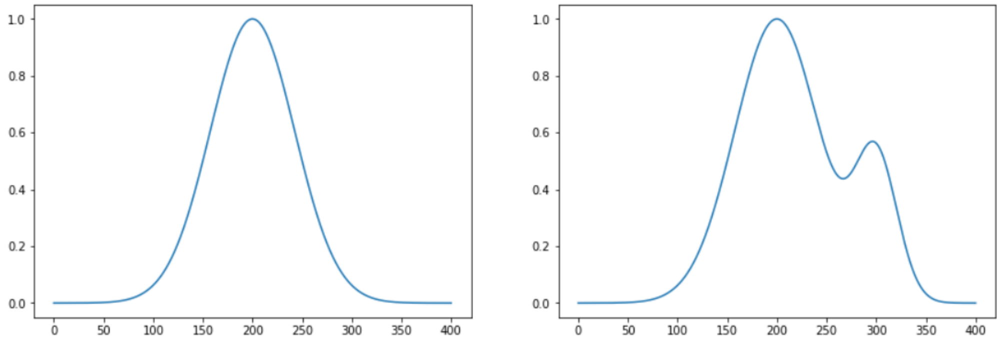


Figure. 8. An example plots of a Gaussian function with manually set parameters. The first plot shows a curve with one component, i.e. $h_2 = 0$. *The second curve is an example of* h_2 *being a positive number.*

4.2. Threshold selection

One of the signal characteristics is that it consists of the true signal produced by a specific physical phenomenon and the noise. In the case of the peak interpretation process, the noise component can be a cause of detection of unwanted peaks. Besides from a significantly lower height than true peaks, the random fluctuations produced by the noise component can visually resemble the peak-shapes. Therefore, a minimum intensity considered in curve fitting problem must be found (Fig. 5) before performing the curve fitting. This is achieved by a heuristic process, which relies on the standard deviation of intensity. For each spectrum i a threshold (Eq. 3) is calculated separately using the following equation.

$$\tau_i = \sigma_{Intensity_i} * k$$

The method was optimised experimentally and the k parameter was decided to be set to 0.2. As previously stated, this is a heuristic process without any explainable meaning and the value of the constant was set basing on observations made during the exploration the analysed samples. The calculation of τ allows rescaling the data in a way that the data points are shifted in the direction of the x-axis by its value.

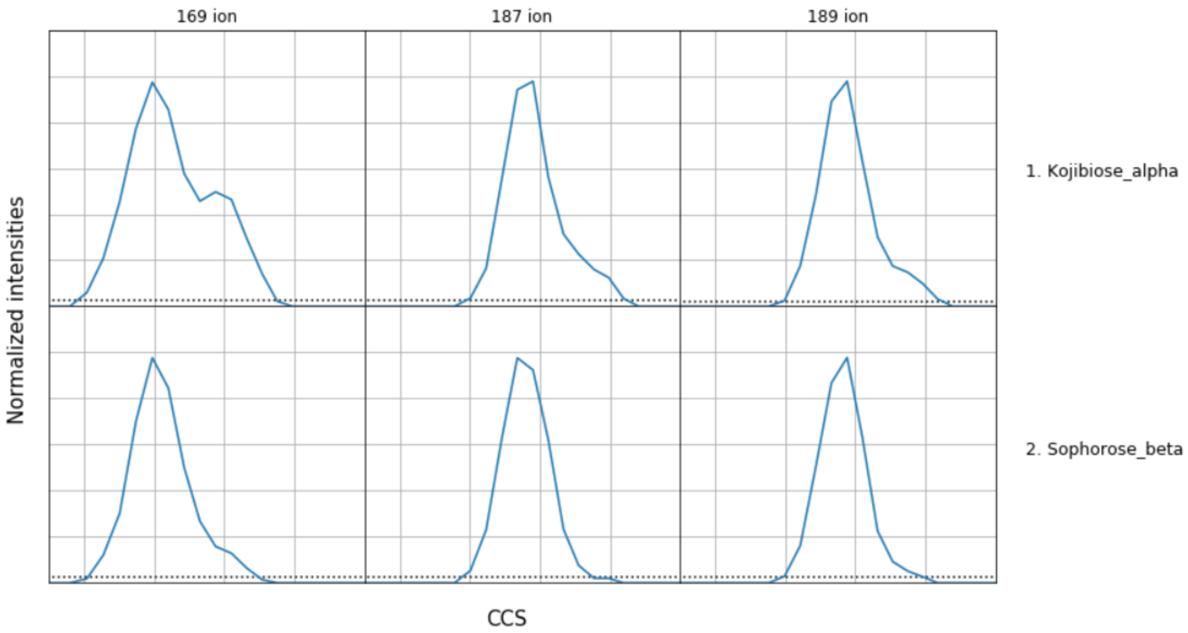


Figure. 9. An example of thresholds (red dotted lines) calculated for x, y and z ions of Kojibiose. The zoomed spaces show how the minimal fluctuations in the data reassemble the actual peaks, which is the reason why they have to be removed or omitted in the further analysis.

It must be mentioned that this approach cannot be considered robust and reliable. This is due to the fact that the k parameter has not been fully explored and it was set manually. It is not a fact that the k parameter set to 0.2 will define the threshold correctly for newly approached samples. Prospectively, the automated optimisation of k parameter or another method of threshold detection needs to be introduced. Alternatively, setting the k parameter will be left to the final user of the proposed methodology. However, this would be considered as a great drawback of the introduced technique, since the goal is to implement a robust, autonomous method, which does not rely on the user input.

4.3. Least-squares

One of the possible approaches to curve fitting problem is the least-squares approach [32]. The least-squares involves solving an optimization problem. In particular, its goal is to optimise parameters of a given function so the function recreates the shape of a given set of points as closely as possible. This is done by solving a minimization problem for an expression defining an overall difference between any discrete point of the signal and its equivalent in the continuous space of the function values, i.e. least squares. To introduce this technique for the research problem, a least-squares expression for the Gaussian mixture model is defined (Eq. 5.).

$$SSR(\phi, x) = \sum_{i=1}^{n} \left|\left|y_i - F_g(\phi, x_i)\right|\right|^2$$

Where ϕ, x, n, y_i are defined as a vector of parameters ($h_1, \sigma_1, w_1, h_2, \sigma_2, w_2.$), CCS of IMS data, number of points, the intensity of $i - th$ point respectively. As previously stated, in order to find parameters of the F_g function, which would produce the best fit to x and y datapoints, the SSR expression must be minimised (Fig. 5) with respect to the parameter vector ϕ. More precisely, the value of this vector must be set in a way that the SSR is as close to zero as it is possible. This might be achieved in the following way:

I. Parameters are initialised in order to give the algorithm an initial guess and a starting position in the parameters space.

I. Partial derivative with respect to each parameter is calculated for this point. This step might be described as the search for the direction in the parameter space. A partial derivative calculated with respected to a given parameter holds information about the steepest ascent of the function in the direction of the axis defined by this parameter. With this knowledge, a gradient vector is built, i.e. a vector holding all of the possible partial derivatives. Technically, this vector shows a direction of the steepest ascent of the function, i.e. the direction in which the function increases its values the most rapidly.

II. Basing on the learning step (which defines how much parameters values can be changed in a single iteration) the initially set ‘position’, i.e. value of the parameters, is changed. The choice of the learning step depends of the minimization technique used, which will be explained in the further section.

III. The algorithm terminates when specific convergence criteria are met. Those might vary depending on the chosen algorithm and will be introduced in the further section.

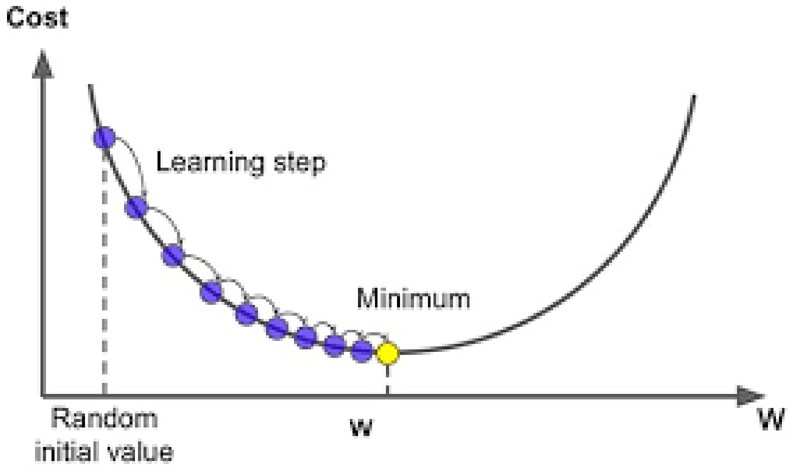


Figure. 10. An example of minimising SSR function with respect to one parameter w. The parameter is set to an initial value and then changed gradually along with the decrease of SRR. It has to be mentioned that in the research problem, a multidimensional space is considered. Since it is visually impossible to plot the parameter space, showing 2-D case is reasonable example, which enables a proper understanding of the minimisation process.

Since the number of observations k is greater than the number of parameters and the defined model is nonlinear, the problem is considered being nonlinear. This is the reason why the outline of the algorithm has an iterative flow i.e. the solution is being find in the number of steps rather than a single computation. It must be mentioned that this approach main disadvantage is that it might be stopping on a local minimum of SSR. Possibly, the solution might be found very quickly but it can be a solution, which SSR is very distant from the potentially smallest SSR. In this case, the algorithm will stop, as it does not distinguish a local and a global maximum Therefore a proper parameter initialisation might significantly increase the efficiency of this technique. Furthermore, a proper 'guess' of initial parameters might significantly speed up the algorithm since the local (or global) minimum will be simply reached faster. There are several possible techniques to minimise SSR, i.e. 'move the ball' towards the optimal solution. In this work, the Levenberg-Marquardt algorithm is employed to solve this problem.

4.4. Minimisation using Levenberg-Marquardt algorithm

The Levenberg-Marquardt (LM) [33] can be described as a combination of two widely used minimization methods: the gradient descent [34] and the Gauss-Newton method [35]. Broadly speaking, the algorithm acts more like the gradient descent when the current solution is far from the optimal value and like the Gauss-Newton method when the solution is close to its optimal value. Therefore, in order to get in-depth understanding of the LM algorithm, it is reasonable to define those two algorithms first. It has to be mentioned that the LM algorithm was chosen in this study for particular reasons. Since the analysis is performed on a very limited number of samples, there is a need to make the most out of them. Inaccurate peak fitting method would result in losing a relatively significant amount of information. The algorithms have been experimentally compared resulting in the choice of LM algorithm. The gradient descent and Gauss-Newton method were proved to incorrectly fit the curve in even 50% cases. However, the LM method is closely related to the other algorithms, so they will be introduced in the further subsections.

4.4.1 Gradient descent

The gradient descent is the simplest and the most common technique in minimisation problems. It updates parameter values in the 'downhill' direction with a predefined learning step, i.e. the parameter η. The learning step has to be set manually and it does not change during the search for an optimal solution. The gradient descent can be expressed as the following equation

$$\phi_{n+1} = \phi_n + \eta \nabla_{f_n}$$

Where ∇_{f_n} is the gradient vector in the point n. The main drawback of this approach is the choice of the learning step. If the learning step is too small the convergence of the algorithm might be achieved relatively slow. In case of too high learning step the optimal solution might be missed by the algorithm.

4.4.2. Gauss-Newton method

The Gauss-Newton method is another method for minimizing a sum of squares objective function. It presumes that the objective function is approximately quadratic in the parameters near the optimal solution. As previously mentioned, in each iteration step the parameter vector ϕ is replaced by a solution $\phi + \delta$. For a small $|\delta| \vee$, a local Taylor series expansion leads to the linear approximation

$$F_g(\phi + \delta) \approx F_g(\phi) + J\delta,$$

Where J is the Jacobian matrix of $F_g(\phi)$, i.e. a matrix holding scalar values defining all possible first-order partial derivatives of this function. As previously stated, the algorithm operates iteratively. At each step, it is required to find δ that minimised the following quantity.

$$\left|y - F_g(\phi + \delta)\right| \approx \left|y - F_g(\phi) - J\delta\right| = SSR - J\delta \vee$$

The minimum is obtained when $J\delta - SSR$ is orthogonal to the column space of J, which leads to $J^T(J\delta - SSR) = 0$. Hence, δ is a solution of the normal equations [36]:

$$J^T J\delta = J^T SSR$$

Where $J^T J$ is the Hessian matrix, i.e. a matrix holding all possible second-order partial derivatives. The iteration step derived from the normal equations is defined as follows:

$$\phi + \delta = (\phi J^T J)^{-1} J^T SSR$$

4.4.3. Levenberg-Marquardt algorithm

In comparison to the Gauss-Newton method the Levenberg-Marquardt algorithm solves a slightly different of normal equations, i.e. augmented normal equations:

$$N\delta = J^T SSR$$

Where all of the non-diagonal elements of N are the same as corresponding elements of $J^T J$ and the diagonal elements are defined as $N_{ii} = \mu + [J^T J]_{ii}, \mu > 0$. The process of adjusting the diagonal of N is called *damping* and μ is called *a damping term*. If damping leads to a reduction in SSR, one iteration step is completed and the damping term is decreased. Otherwise, the damping term is increased and the augmented normal equations are solved again until SSR is decreased and one iteration step is done. This explains why the LM method can be considered as a combination of the gradient descent and the Gauss-Newton algorithm – for a small damping term the algorithm will behave more like the gradient descent and for a large damping term, it will be more like the Gauss-Newton algorithm.

4.5. Initial parameter estimation

To accurately initialise the parameters for each IMS spectrum, the data needs to be normalised. Parameters such as height h and width w have the same initial values for each spectrum, hence, the range of values of variables x and y shall be identical all samples. Normalisation to range $(0,1)$ is chosen in this work (Eq. 6).

$$x = \frac{x - x_{min}}{x_{max} - x_{min}}; \; y = \frac{y - y_{min}}{y_{max} - y_{min}}$$

With pre-processed data, one can start initialising parameters. In order to initialise h and x_o parameter, an initial peak detection method needs to be introduced. It must be stressed out that the introduced method will not be accurate and can be used only for initialisation rather the feature extraction in general. The approach relies on the second derivative $f''(x)$ and its characteristics, which implies that the local minima of $f''(x)$ reflects local maxima of an underlying function and its inflection points (Figure X). In case of IMS data, the peaks are usually strongly overlapping. The first derivative would not detect such peaks since mathematically they are very similar to the concept of inflection points. Thus, the second derivative is calculated for each spectrum and then its local minima are found by comparing each point to its neighbours.

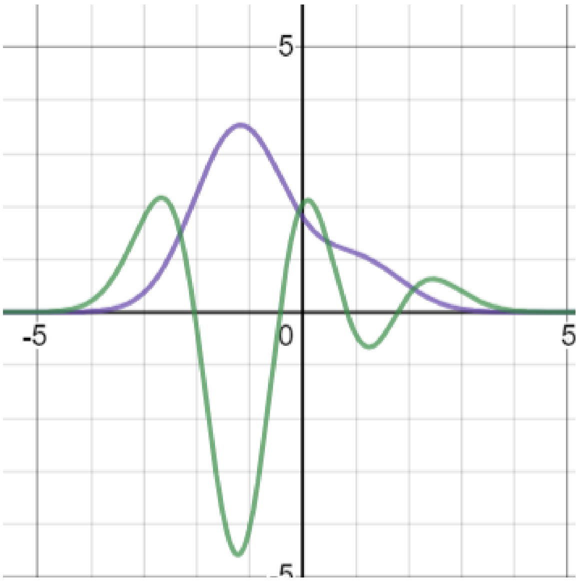


Figure 11. The plot shows how the second derivative f'' (green) can be used to obtain a local maxima and inflection point of a function f (blue). As might be noticed an unwanted peak has been detected in point of the inflection point probably resulting from the noise.

The derivatives (Fig. 5) are calculated in a discrete domain, hence finite differencing technique is employed (Eq. 7). The second derivative is derived by taking the derivative of the first derivative using the same method.

$$f'(x) = \frac{f(x+h) - f(x)}{h}$$

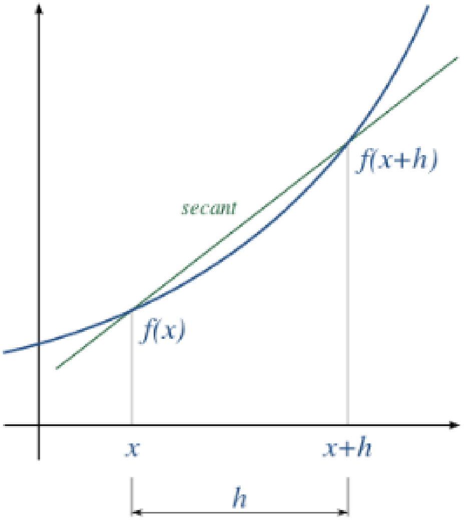


Figure. 12. An example of how a derivative can be calculated using discrete points rather than continuous. In continuous domain h tends to be an infinitesimal number, i.e. the number infinitely close to zero. However, since it is impossible to derive an infinitely small number on a machine h is simply a distance between two neighbouring points.

It was observed that this technique produces the number of peaks significantly higher than the ground truth implies, therefore a set of correct peaks must be chosen from the whole initial set. One can conclude that this phenomenon occurs due to the fact that random noise in the signal can produce inflection points 'placed on the peaks' (Figure X). The introduced thresholding method does not remove this kind of noise from the signal, thus there is a need to filter the undesired peaks. The feature selection process will be described in the next subsection. The remaining parameter w is set using empirical observations rather than calculations and it takes the value of 0.1.

In addition to the initialisation, there is also a need to put constraints on the parameters in order to achieve the best fit. More particularly, both of height parameters are set to have a minimum value equal to the initial value. The x_o parameters are constrained to be placed in a distance of 0.01on the x-axis in both ways. The width parameter shall be always greater than zero. The constraints were defined by experimental observations and it is not certain if they will give a foundation to a fully robust solution, but it will be proved that they allow a successful automation of the process.

Taking all into consideration, the set of rules for the parameters initialisation (on the left) and constraints (on the right) are defined as follows:

$$h = h_{init} \qquad h > h_{init}$$

$$w = 0.01 \qquad w > 0$$

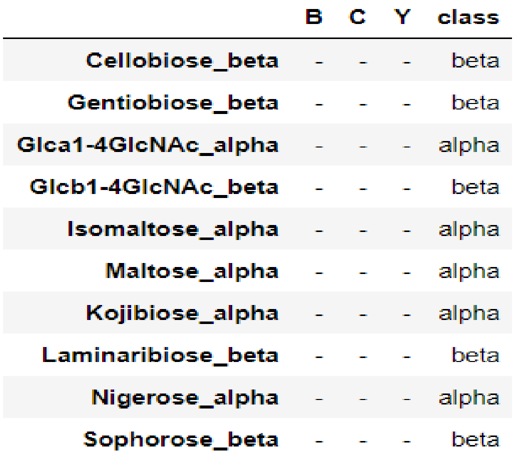
$$x_o = x_{init} \qquad x_o \in (x_{init} - 0.01, x_{init} + 0.01)$$

4.6. Feature extraction algorithm

As previously mentioned, in order to find true peak positions a proper filtration of the initially found peaks needs to be performed. Thus, a simple iterative method is proposed to pick the correct initial values of x and h. Set Z containing the result of permutation of the initially found peaks is created, i.e. a set of 2-tuples such as (h_i, h_j), where $i \neq j$. More precisely, the set Z contains all of the possible pairs generated from the set of initially found peaks and a pair cannot contain instances of the same peak. Then, the previously described fitting algorithm is applied for each candidate peaks. SSR is calculated for each fit and the tuple of candidate peaks with the lowest score is chosen. Summarising, the algorithm is defined as follows:

1. The initial search for peaks is performed using the second derivative method.
2. Set Z is generated for the initially detected peaks.
3. The curve fitting algorithm is used for each pair in set Z.
4. The pair (h_1, h_2) with the lowest SSR is chosen.

The tuple (h_1, h_2) contains valuable information about the difference between heights of the peaks. The overall goal of the feature extraction phase was to gain this particular information from the data. Hence, a difference $|h_1 - h_2|$ is calculated to finally form a feature matrix used in the further diglucosides classification. A structure of the matrix can be visualised as the following table. It can be visually and structurally compared with Figure 2. since it holds the same information but in the numerical form. Each cell of the table corresponds to a distance between peaks of a given diglucoside and ion separated in the mass spectrometry phase. With this data, one can attempt to build a machine learning model able to distinguish between α- and β -diglucosides, which is the overall goal of the study.



	B	C	Y	class
Cellobiose_beta	-	-	-	beta
Gentiobiose_beta	-	-	-	beta
Glca1-4GlcNAc_alpha	-	-	-	alpha
Glcb1-4GlcNAc_beta	-	-	-	beta
Isomaltose_alpha	-	-	-	alpha
Maltose_alpha	-	-	-	alpha
Kojibiose_alpha	-	-	-	alpha
Laminaribiose_beta	-	-	-	beta
Nigerose_alpha	-	-	-	alpha
Sophorose_beta	-	-	-	beta

Table 1. Structure of the feature matrix extracted from the IMS signals of reducing diglucosides.

4.7. Classification using Gaussian naïve Bayes classifier

There are plenty of different classifiers used in machine learning and statistics, e.g. support vector machines, deep neural networks, linear classifiers or decision trees. Due to the uncomplicated nature of data, i.e. a relatively low number of dimensions (three, each dimension corresponding to one ion) and small number of samples, there is no need to employ highly advanced techniques in the study.

The problem of classifying samples given a training set with predefined labels is considered to be supervised learning. It builds a model that is able to make predictions given evidence. When the model is trained it might take any input of the same type (dimensions and data type of each variable) and decide which of predefined labels should be assigned to this sample.

In order to classify the given series of diglucosides by their bondage type, a Gaussian naïve Bayes classifier [37] is employed. The model is based on Bayes' theorem and makes a 'naïve' assumption of the variables being independent. It means that a single variable does not have any impact on the other variables in the dataset. It involves calculating prior and posterior probabilities. The prior probability is defined as follows:

$$P(c) = \frac{N_C}{N}$$

where N_c is the number of samples of a given class and N is the number of samples. The posterior probability is defined as:

$$P(x|c) = P(x_1|c) * P(x_2|c) * \dots * P(x_n|c)$$

In case of usual naïve Bayes classifier, one deals with discrete data. However, the feature matrix obtained in the feature extraction phase contains only continuous variables. $P(x_i|c)$ for continuous features is defined as:

$$P(x_i|c) = \frac{1}{\sqrt{2 * \pi * \sigma_c^2}} * e^{\frac{-\left(x_i - \mu_{x_i},c\right)^2}{2 * \sigma_{x_i,c}^2}}$$

Where $\sigma_{x_i,c}^2$ and $\mu_{x_i,c}^2$ are the standard deviation and the mean of x_i feature within c class. Finally, the probability of a class c defining a newly approached sample x is given as follows:

$$P(c_i|x) = \frac{P(x|c_i) * P(c_i)}{\sum_j P\left(x|c_j\right) * P\left(c_j\right)}$$

In the classification problem, the probability is calculated for each class and the class with the highest probability is considered to be the predicted result. Precisely, a probability of a given IMS spectra of being α-glucoside and β-glucoside are calculated. The higher probability is considered to be the result. It is possible to take a different approach in choosing the predicted class, which is introduced in the result section.

5. Results

In this section, the results of the feature extraction and classification are presented. More specifically, a table of fits is shown (Figure 6.), the extracted features are displayed (Table 2.), the evaluation routine is described, the AUC-ROC curve is plotted (Figure 7) and the accuracy of the model is calculated. The results are described and analysed.

5.1 Example fits

First, the example fits are presented (Figure 6). As it might be observed, the fitting algorithm performs satisfactorily in several scenarios. The Y ion of nigerose is an example of clearly separated peaks, which technically should be the less difficult to describe by a function. The fitted function curves are also separated creating a specific 'valley' between them, as should be expected. The B ion extracted from nigerose and the Y ion of laminaribiose show the second possible case of IMS signal structure. The peaks are highly overlapped and there is no clear gap between them, which is considered to be the hardest type of IMS data to be approximated by a function. Nonetheless, the algorithm performed decently even in this case. It created an inflection point in the place of the second peak, which is a perfect mapping of the actual IMS signal. The C ion derived from Laminaribiose is an example of the single peaked signal. One possible difficulty in fitting the curve to this type of data might be creating two very similar peaks, which heights are roughly the half of the maximum point. Potentially, sum of those functions would recreate the true shape of the signal but properties of each component would not be consistent with the actual components. The fitting algorithm performed well in this case. Only one component tends to describe the whole signal and the second component is minimized so technically it can be treated as it would not exist. The Y ion derived from Glcα1-4GlcNAC is an example of distanced peaks, where a strong 'valley' is formed between them. The algorithm also performs well in this situation. Generally speaking, the algorithm performed well in all of the situations introduced in the study. It could be easily modified to more advanced scenarios, where we deal with larger number of peaks by simply increasing the number of components of Gaussian functions and extending a tuple of initially defined peaks. In the case the number of components is known, these quantities can stay fixed. However, if they are not, there is a need to specify the maximum number of components, which might be found in the spectrum. Then, another exhaustive search could be performed. The main disadvantage of this approach is its complexity. In the case when the number of peaks is known the complexity is $O(N)$ (linear) and when the number of peaks is not known and two brute force searches are performed, it is increased to $O(N^2)$ (quadratic). Since in the study the number of peaks is low and known, this approach has been employed. The main idea behind the study is to prove that automated classification of linkages of diglucosides (so other lihiated disaccharides) is achievable. As long as the desired effect of parameterizing the signal is obtained and the pattern is recognized, the feature extraction process is considered to be successful. However, in the case of more complex glycans it would be reasonable to employ more advanced and rapid methods [38]. The curve fitting problem for MS/IMS spectra has been an open research question and improving the current state-of-the-art approaches is considered to be a problem outside of this study.

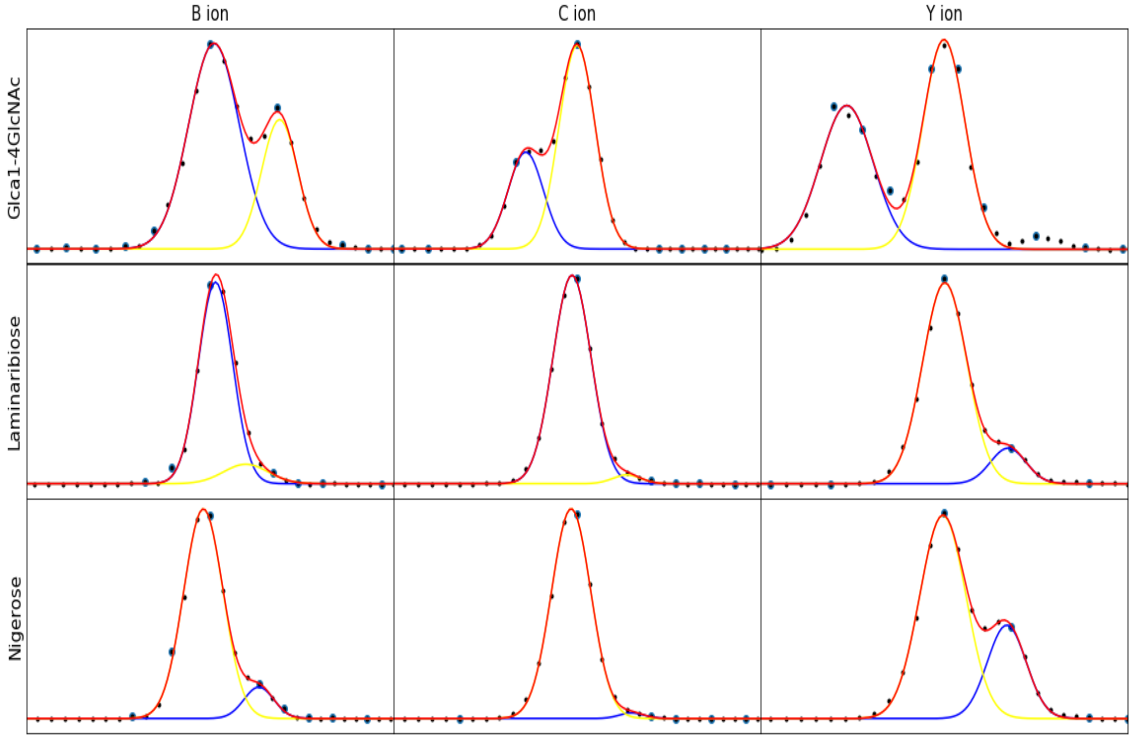
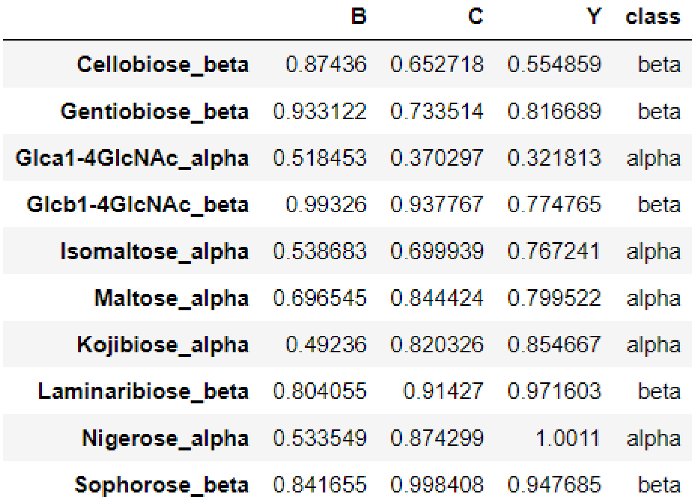


Fig. 13. The table of example fits performed in the study. The table contains all possible scenarios for two peaks: highly overlapped peaks as in the case of Laminibriose – Y ion, clearly separated but still close peaks – Glcα1-4GlcNAC – C, and clearly distinctive, distant peaks as in the case of Glcα1-4GlcNAC – Y. All the approximations obtained in the experiment can be found in the appendix.

5.2. Feature matrix

The following feature matrix has been built by comparing heights of components of the approximated functions. The feature matrix faithfully reflects the pattern for α- and β-diglucosides, which is a consequence of correctly parameterized curves, as showed in the previous subsection. As deduced in the introduction section, one outlier has been produced – Y ion for Glcβ1-4GlcNAC. It has to be stressed that outliers are something, which is usually expected in data analysis. Nevertheless, the value is replaced by a mean of the feature to avoid skewing the classification results. In such a limited space even one outlying point might make a significant difference in the classification.



	B	C	Y	class
Cellobiose_beta	0.87436	0.652718	0.554859	beta
Gentiobiose_beta	0.933122	0.733514	0.816689	beta
Glca1-4GlcNAc_alpha	0.518453	0.370297	0.321813	alpha
Glcb1-4GlcNAc_beta	0.99326	0.937767	0.774765	beta
Isomaltose_alpha	0.538683	0.699939	0.767241	alpha
Maltose_alpha	0.696545	0.844424	0.799522	alpha
Kojibiose_alpha	0.49236	0.820326	0.854667	alpha
Laminaribiose_beta	0.804055	0.91427	0.971603	beta
Nigerose_alpha	0.533549	0.874299	1.0011	alpha
Sophorose_beta	0.841655	0.998408	0.947685	beta

Table 2. Features extracted from IMS data, which are used to classify diglucosides by a bonding type. As might be observed, the values of alpha samples are usually smaller compared to the beta samples. This pattern allows classifying the diglucosides. The classification is performed by training a Naïve-Bayes classifier with the matrix as explained in the methodology section.

5.3. K-fold validation

In order to conduct this experiment with only 10 samples, the abbreviation of k-fold cross validation is employed. The k-fold validation testing is used when the number of samples is limited. In machine learning the usual testing approach involves splitting the dataset into two groups – a training set, which is used to construct a model and test set, which is used for the evaluation purposes. However, in this study, the number of samples is relatively small so casual splitting would produce too few samples for the training purposes since it would provide not enough information to capture the pattern observed in the data. The k-fold validation procedure is defined as follows:

1. Choose one sample from the dataset as a test sample. The remaining seven samples are considered to be the training set.
2. Perform a training phase, i.e. calculate prior and posterior probabilities, using the training set.
3. Make a prediction for the test sample.
4. Repeat the first three steps until all of the samples are used as a test set once.

5.4. Accuracy

In order to evaluate the k-fold validation procedure a performance metric must be employed. For purpose of this study, accuracy is chosen. This basic metric is defined as follows:

$$A = \frac{P}{N}$$

Where P is the number of correctly classified samples and N is the count of classification attempts.

The accuracy obtained in the experiment, when two classes were treated as equally important equals 0.7, which means that the classifier made three mistakes in 8 trials. However, if it is decided to drop y ion from the feature space, the accuracy increases to 0.9 and the mistake is made only for Cellobiose – a sample, which was not accurately parameterized by the fitting algorithm for C and Y ions. Furthermore, if the c ion is dropped the accuracy is increased to 1, which means that all of the samples in the k-validation procedure were classified successfully. The increased accuracy leads to the conclusion that ion is the most important feature in the built dataset, *i.e.* it brings the most information to the classification outcome.

Nevertheless, it is hard to accurately approximate the performance of the classification due to the limited number of samples of each class. However, this fact is considered to be an initial limitation of the research problem, since the number of given saccharides standards abundant in nature is limited. It has to be stressed out that the dataset could be extended to different carbohydrates than diglucosides. The pattern, which allows classification, does not occur only for this certain type of sugars as stated in the introduction.

5.5. AUC-ROC curve

As previously stated, in the classification routine each of the class is treated as equally important. However, this is not always the case. It might be possible to place emphasis on only one class. For example, one might want to specifically classify only alpha- over beta-diglucosides and vice versa. In this case, a classification threshold needs to be defined. When two classes are considered to be of equal weight the threshold is set to 0.5. In other words, the class with higher probability is considered to be the result. However, when the preference is put on one of the classes the threshold for this class should be higher. For example, when it is chosen to classify alpha linkages the threshold might be set to 0.7. This indicates that the model classify alpha-diglucosides only when there is 70% probability that given sample represents this class.

In order to evaluate the model for different thresholds AUC-ROC curve is analysed. The AUC-ROC curve shows how the true-positive rate and false positive rate changes when the threshold of classification is changed. It provides information on the performance of the model in the whole space of the threshold parameter. From the obtained AUC-ROC curve it might be deduced that the classification algorithm would perform well for different thresholds. However, a number of samples decrease an evaluation power of this metric. This is due to the fact that it takes a slightly discretized form and does not cover more possibilities in the TPR and FPR.

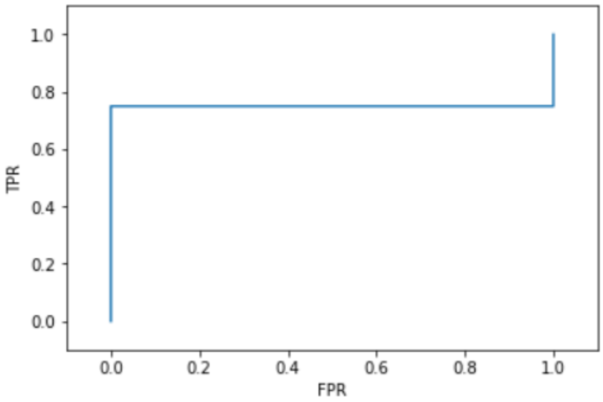


Fig. 14. AUC-ROC curve obtained in the diglucosides classification experiment. The rectangular shape implies a limited number of samples in the dataset.

6. Conclusions

In this section the research objectives and formal requirements of the approach are revisited. All of the findings related to the implemented algorithm are summarized and confronted with the biochemical knowledge on glycans and experiments, which provided data for the analysis.

6.1. Research objectives

In summary to the description of curve fitting approach in the results section, the algorithm performed well in all of the testing scenarios. However, it has a high complexity and it is not known how it deals with more complex samples. From the theoretical point of view, the algorithm should be able to fit a gaussian component for each true positive in the initial detection, i.e. using the second derivative method. This means that it can be assumed that it will perform well with more signals with higher number of modes – even extremely overlapped peaks will produce an inflection point in the signal, which will be initially detected and will allow correct approximation – this is a brute force approach so given solution space containing solutions must converge. However, this would require a further hyperparameter tuning, which is not considered to be the desired. As previously mentioned, it would be beneficial to employ more rapid and non-parametrised techniques for larger molecules, when the focus needs to be made on the fast annotation, which does not assume *priori* configuration of the algorithm.

The current approach provides a feature matrix suitable only for signal consisting from two components. It could be extended to more complex cases. For example, for three mode signal the features can be defined as:

$$(x_{h_1-h_2}, x_{h_2-h_3}, x_{h_3-h_1}, x^1_{h_1-h_2}, x^1_{h_2-h_3}, x^1_{h_3-h_1}, \dots, x^n_{h_1-h_2}, x^n_{h_2-h_3}, x^n_{h_3-h_1})$$

Where $x^1, x^2, \dots, x^n$ are ions, which are derived in the MSMS fragmentation and analysed by IMS. The difference $|h_j - h_i|$ represents all possible combinations between peaks in term of the difference in their heights.

However, it has to be reminded as this technique will be useful only in classifying not co-fragmented. As mentioned in the introduction section the glycans containing isomers connected with each other by more than one glycosidic linkage might be problematic for this technique. However, if glycans with monomeric building blocks with no co-fragmented glycosidic linkages were assumed, it might be possible to extend the technique. The output of such a classifier will have multiple components, where each correspond to one linkage in the glycan. This kind of problem is considered to be *multi-label* classification []. In terms of extending the feature extraction phase for glycans different than diglucosides, the B, Y, Z ions need to be specified. In the case of this research, a knowledge about B, Y, Z ions were presumed, and those ions have been chosen from all the available IMS spectra. The necessity of specifying those ions is considered to be a disadvantage of this technique – unless the target group is only one standard of glycans (a group built from the same building blocks).

The classifier exploited the bias in data, i.e. the pattern of one peak for b-glucosides and two peaks for a-glucosides. It is hard to justify this phenomenon since it was not theoretically explained. One interesting conclusion is that B ion was the most important in the classification. Practically, this means that the difference between a-glucosides and b-glucosides were the most significant within this feature, i.e. double-peaked signals produced a small difference between the peaks and single-peaked signals produced a high difference between the peaks. However, since this phenomenon has not been described yet, it should be assumed that this kind of overfitting is not desirable. Newly approached samples could differ in ion B signal intensities in comparison to the analysed samples, which will fail the classification.

6.2. Formal requirements

The classifier does not require any user input in terms of the training phase and prediction phase. The user provides a raw input spectrum of a diglucoside and receives a glycoside linkage classification with probability of belonging to each class provided.

The system might fail when approaching an unknow sample. As previously stated, masses of B, Y, Z ions are presumed and newly approached sample with a different mass of those ions will be not successfully recognized. In order to classify a sample built from a different monosaccharides than glucose, the classifier needs to be retrained using data from that group. The classifier will not classify a newly approached sample as unknow either. Classifying a sample as unrecognized is defined as *open set* classification and can be considered in the further work.

The model used for approximation is considered to be easily understandable. Its parameters correspond to visually observable characteristics of the curve such as height or width. The Gaussian-Naïve classifier is also considered to be interpretable since it can justify its choices by providing *prior* and *posterior* probabilities.

6.3. Future work

The main outcome of the project is exploiting bias observed in IMS data for diglucosides and showing that it might lead to automated sequencing of glycans in terms of the glycosidic linkage type. The *de novo* sequencing approaches introduced in the literature overview were missing this ability. They are able to provide a sequential structure of glycans by determining their composition (GlycoDeNovo, GLYCH) or glycosidic linkages for smaller ions (LODES). Those approaches are fully autonomous and nonparameterized, so they overcome the drawbacks of the approach introduced in the study.

It might be concluded that extending similar techniques by IMS data, so adding a single dimension to the input, might solve the defined problems. Developing a *de novo* sequencing method, which utilizes tandem MS data complemented by orthogonal information provided by IMS, will be certainly a huge breakthrough in a field of glycomics.

References

[1] Yuting Xu, Deeptak Verma, Robert P. Sheridan, Andy Liaw, Junshui Ma, Nicholas M. Marshall, John McIntosh, Edward C. Sherer, Vladimir Svetnik, and Jennifer M. Johnston. Deep Dive into Machine Learning Models for Protein Engineering. Journal of Chemical Information and Modeling 2020 60 (6), 2773-2790

[2] Zheng, X., Xu, S., Zhang, Y., Xinxiang Huang. Nucleotide-level Convolutional Neural Networks for Pre-miRNA Classification. Sci Rep 9, 628 (2019)

[3] Mulloy B, Hart GW, Stanley P. In: Varki A, Cummings RD, Esko JD, et al., editors. Structural Analysis of Glycans. Essentials of Glycobiology. 2nd edition. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2009. Chapter 47.

[4] Stanley P, Schachter H, Taniguchi N. N-Glycans In: Varki A, Cummings RD, Esko JD, et al., editors. Essentials of Glycobiology. 2nd edition. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2009. Chapter 8.

[5] Brockhausen I, Schachter H, Stanley P. O-GalNAc Glycans. In: Varki A, Cummings RD, Esko JD, et al., editors. Essentials of Glycobiology. 2nd edition. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2009. Chapter 9.

[6] G.B. Ogden, P.C. Melby, Leishmania, In: Moselio Schaechte editors. Encyclopedia of Microbiology (Third Edition), Academic Press, 2009, Pages 663-673.

[7] Abu B. Kanu, Prabha Dwivedi, Maggie Tam, Laura Matz and Herbert H. Hill Jr., Ion mobility–mass spectrometry, Journal of mass spectrometry 2008; Volume 43, Pages 1–22

[8] Pukala, Tara, Importance of collision cross section measurements by ion mobility mass spectrometry in structural biology, Special Issue: Perspectives on the future of Mass Spectrometry, Volume33, Issue S3, 2019, Pages 72-82.

[9] Karpievitch, Y. V., Polpitiya, A. D., Anderson, G. A., Smith, R. D., & Dabney, A. R. (2010). Liquid Chromatography Mass Spectrometry-Based Proteomics: Biological and Technological Aspects. The annals of applied statistics, 4(4), 1797–1823.

[10] Kanu, Abu B., Dwivedi, Prabha, Tam, Maggie, Matz, Laura, Hill Jr., Herbert H. Ion mobility–mass spectrometry, Journal of Mass Spectrometry, Volume 43, 2008, Pages 1-22.

[11] Mittal, Rama Devi. Tandem mass spectroscopy in diagnosis and clinical research. Indian journal of clinical biochemistry. Volume 30, 2015, Pages 121-123.

[12] Bruno Domon & Catherine E Costello, A systematic nomenclature for carbohydrate fragmentations in FAB-MS/MS spectra of glycoconjugates. Glycoconjugate Journal volume 5, 1988, pages 397–409.

[13] Arif Ahmed, Yun Ju Cho, Myoung-han No, Jaesuk Koh, Nicholas Tomczyk, Kevin Giles, Jong Shin Yoo, and Sunghwan Kim, Application of the Mason−Schamp Equation and Ion Mobility Mass Spectrometry To Identify Structurally Related Compounds in Crude Oil, Analytical Chemistry, 2011, Volume 83, Pages 77-83

[14] Lanucara, F., Holman, S., Gray, C. et al. The power of ion mobility-mass spectrometry for structural characterization and the study of conformational dynamics. Nature Chem, 2014, Volume 6, pages 281–294.

[15] Christopher J. Gray, Baptiste Schindler, Lukasz G. Migas, Martina Pičmanová, Abdul R. Allouche, Anthony P. Green, Santanu Mandal, Mohammed S. Motawia, Raquel Sánchez-Pérez, Nanna Bjarnholt, Birger L. Møller, Anouk M. Rijs, Perdita E. Barran, Isabelle Compagnon, Claire E. Eyers, and Sabine L. Flitsch, Bottom-Up Elucidation of Glycosidic Bond Stereochemistry, Analytical Chemistry 2017, Volume 89 (8), Pages 4540-4549.

[16] Both P, Green AP, Gray CJ, Sardzík R, Voglmeir J, Fontana C, Austeri M, Rejzek M, Richardson D, Field RA, Widmalm G, Flitsch SL, Eyers CE. Discrimination of epimeric glycans and glycopeptides using IM-MS and its potential for carbohydrate sequencing. Nat Chem. 2014, Volume 6, Pages 65-74.

[17] Pičmanová M, Neilson EH, Motawia MS, Olsen CE, Agerbirk N, Gray CJ, Flitsch S, Meier S, Silvestro D, Jørgensen K, Sánchez-Pérez R, Møller BL, Bjarnholt N. A recycling pathway for cyanogenic glycosides evidenced by the comparative metabolic profiling in three cyanogenic plant species. Biochem J. 2015, Volume 469, Pages 375-89

[18] Peetz O, Hellwig N, Henrich E, Mezhyrova J, Dötsch V, Bernhard F, Morgner N. LILBID and nESI: Different Native Mass Spectrometry Techniques as Tools in Structural Biology. J Am Soc Mass Spectrom. 2019, Pages 181-191.

[19] Ibrahim, Yehia M et al. “Characterization of an Ion Mobility-Multiplexed Collision Induced Dissociation-Tandem Time-of-Flight Mass Spectrometry Approach.” International journal of mass spectrometry, Volume 293, 2010, Pages 34-44.

[20] Tseng K, Hedrick JL, Lebrilla CB. Catalog-library approach for the rapid and sensitive structural elucidation of oligosaccharides. Anal Chem, 1999, Pages, 3747-54

[21] Joshi HJ, Harrison MJ, Schulz BL, Cooper CA, Packer NH, Karlsson NG. Development of a mass fingerprinting tool for automated interpretation of oligosaccharide fragmentation data. Proteomics. 2004, 1650-64.

[22] Klaus Karl Lohmann, Claus-W. von der Lieth, GlycoFragment and GlycoSearchMS: web tools to support the interpretation of mass spectra of complex carbohydrates, Nucleic Acids Research, Volume 32, 2004, Pages 261–266

[23] Kameyama A, Kikuchi N, Nakaya S, Ito H, Sato T, Shikanai T, Takahashi Y, Takahashi K, Narimatsu H. A strategy for identification of oligosaccharide structures using observational multistage mass spectral library. Anal Chem. 2005, Pages 4719-25.

[24] von der Lieth, C. W., Freire, A. A., Blank, D., Campbell, M. P., Ceroni, A., Damerell, D. R., Dell, A., Dwek, R. A., Ernst, B., Fogh, R., Frank, M., Geyer, H., Geyer, R., Harrison, M. J., Henrick, K., Herget, S., Hull, W. E., Ionides, J., Joshi, H. J., Kamerling, J. P., … Haslam, S. M. (2011). EUROCarbDB: An open-access platform for glycoinformatics. Glycobiology, Volume 21, Pages 493–502.

[25] Sara P. Gaucher, Jeff Morrow, and Julie A. Leary STAT: A Saccharide Topology Analysis Tool Used in Combination with Tandem Mass Spectrometry Analytical Chemistry 2000, pages 2331-2336

[26] Tang, H., Mechref, Y., & Novotny, M. V. (2005). Automated interpretation of MS/MS spectra of oligosaccharides. Bioinformatics (Oxford, England), 21, 431–439.

[27] Hong, P., Sun, H., Sha, L., Pu, Y., Khatri, K., Yu, X., Tang, Y., & Lin, C. GlycoDeNovo - an Efficient Algorithm for Accurate de novo Glycan Topology Reconstruction from Tandem Mass Spectra. Journal of the American Society for Mass Spectrometry, 2017, Volume 28, Pages 2288–2301

[28] Richard Bellman, Dynamic Programming, Scienc, 1966, Pages 34-37

[29] Tsai, Shang-Ting, Liew, Chia Yen, Hsu, Chen, Huang, Shih-Pei, Weng, Wei-Chien, Kuo, Yu-Hsiang, Ni, Chi-Kung. Automatic Full Glycan Structural Determination through Logically Derived Sequence Tandem Mass Spectrometry, ChemBioChem, 2019, Pages 2351-2359

[30] B.K. Lavine, C.G. White, T. Ding, M.M. Gaye, D.E. Clemmer, Wavelet based classification of MALDI-IMS-MS spectra of serum N-Linked glycans from normal controls and patients diagnosed with Barrett's esophagus, high grade dysplasia, and esophageal adenocarcinoma. Chemometrics and Intelligent Laboratory Systems, Volume 176, 2018, Pages 74-81.

[31] Manuel Pera, Carlos Manterola, Oscar Vidal Luis Grande, Epidemiology of esophageal adenocarcinoma, Special Issue: Seminars in Surgical Oncology: Recent Advances in the Management of Esophageal Adenocarcinoma, 2005, Pages 151-159

[32] “The Method of Least Squares”, Steven J. Miller, Mathematics Department, Brown University, 2006.

[33] “The Levenberg-Marquardt algorithm for nonlinear least squares curve-fitting problems”, Henri P. Gavin, Department of Civil and Environmental Engineering Duke University September, 2020

[34] “An overview of gradient descent optimization Algorithms”, Sebastian Ruder, Insight Centre for Data Analytics, NUI Galway, 2017

[35] Gratton, Serge & Lawless, Amos & Nichols, Nancy. Approximate Gauss–Newton Methods for Nonlinear Least Squares Problems. SIAM Journal on Optimization, 2007, Volume 18, Pages 106-132.

[36] Roncat, Andreas. Visualizing Normal Equations in Least-Squares Adjustment. Conference: Proceedings of the 16th International Conference on Geometry and Graphics, 2014.

[37] Watson, T, An empirical study of the naive Bayes classifier, 2001.

[38] Hauschild, A. C., Kopczynski, D., D'Addario, M., Baumbach, J. I., Rahmann, S., & Baumbach, J. Peak detection method evaluation for ion mobility spectrometry by using machine learning approaches. Metabolites, 2013, Volume 3, Pages 277–293.

7. Appendix

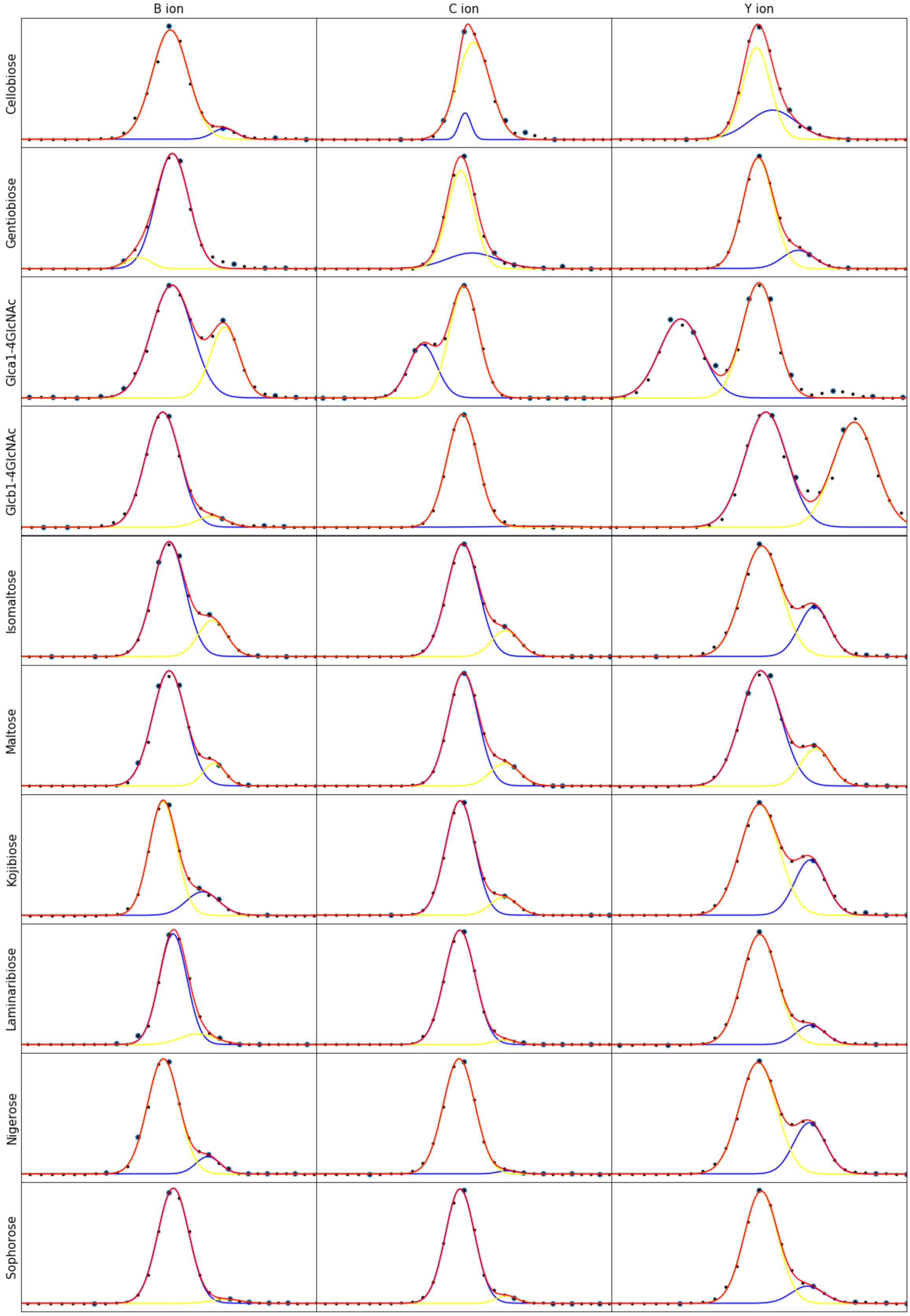


Figure 15. Curve fitting results