GC-MS characterisation of novel artichoke (*Cynara scolymus*) pectic-oligosaccharides mixtures by the application of machine learning algorithms and competitive fragmentation modelling

Carlos Sabater, Agustín Olano, Nieves Corzo*, Antonia Montilla

Instituto de Investigación en Ciencias de la Alimentación CIAL, (CSIC-UAM) CEI (UAM + CSIC), C/ Nicolás Cabrera, 9, E-28049 Madrid (Spain).

*Author to whom correspondence should be addressed:
C/ Nicolás Cabrera 9,
E-28049 Madrid (Spain).
Tel: +34 910017900
Fax: +34 910017905
E-mail: nieves.corzo@csic.es

ABSTRACT

Novel artichoke pectic-oligosaccharides (POS) mixtures have been obtained by enzymatic hydrolysis using four commercial enzyme preparations: Glucanex®200G, Pentopan®Mono-BG, Pectinex®Ultra-Olio and Cellulase from *Aspergillus niger*. Analysis by HPAEC-PAD showed that Cellulase from *A. niger* produced the greatest amount of POS (310.6 mg g^{-1} pectin), while the lowest amount was produced by Pentopan®Mono-BG (45.7 mg g^{-1} pectin). To determine structural differences depending on the origin of the enzyme, GC-MS spectra of di- and trisaccharides have been studied employing three machine learning algorithms: multilayer perceptron, random forest and boosted logistic regression. Machine learning models allowed characteristic *m/z* ions patterns to be established for each enzyme based on their GC-MS spectra with high prediction rates (above 95% on the test set). Possible chemical structures were given for some *m/z* ions having a decisive influence on these classifications. Finally, it was observed that several ions could be formed from specific POS structures.

Keywords: artichoke pectin; enzymatic hydrolysis, pectic-oligosaccharides; neural network; *in silico* fragmentation

1. Introduction

Pectin, one of the most structurally complex families of polysaccharides in nature, is mainly composed of linear chains of α-1,4-D-galacturonic acid (GalA) called homogalacturonan (HG), which comprise approximately 70% of total pectin. The other two domains present ramified chains as rhamnogalacturonan type I (RG-I), a polymer with alternate sequences of GalA and α-(1, 2) linked L-rhamnosyl residues which may be substituted at *O*-4 with linear or branched oligosaccharides, and rhamnogalacturonan II (RG-II), a most complex structure consisting of 12 different types of sugars in over 20 different linkages (Mohnen, 2008; Gullón et al., 2013). Pectins present interesting properties which are directly influenced by its structural characteristics (monomeric composition, presence and distribution of side chains, degree of methyl-esterification and acetylation, molar mass, and charge distribution along their backbone) (Herbstreith & Fox, 2018; Maric et al., 2018).

The most commonly used pectins in the food industry are extracted from citrus peel, apple and sugar-beet by-products. However, the recent optimization of a procedure of pectin extraction from artichoke by-products could provide a new source with interesting techno-functional properties (Sabater, Corzo, Olano & Montilla, 2018a).

By the partial depolymerization of pectins through chemical or enzymatic methods pectic-oligosaccharides (POS) can be obtained and it is known that POS are one of the most promising candidates to be recognised as prebiotics (Gullón et al., 2013). Enzymatic hydrolysis of pectins shows several advantages, such as high regio- and stereoselectivity, obtaining structurally different POS mixtures (Babbar, Dejonghe, Gatti, Sforza, & Elst, 2016; da Moura, Macagnan, & da Silva, 2015). Most food-grade POS studied are produced by commercial pectinases with three different enzymatic activities polygalacturonase (PG), pectin lyase (PL) and pectin esterase (PE) involving reactions of hydrolysis, β-elimination

and removal of methyl or acetyl groups, respectively. Commercial pectinases are usually produced by *Aspergillus* spp. (Combo, Aguedo, Goffin, Wathelet, & Paquot, 2012).

POS mixtures can be formed of various types of oligomers and these structures are often characterised using chromatographic techniques such as HPAEC-PAD (Babbar, Dejonghe, Sforza, & Elst, 2017; Combo et al., 2013; Gómez, Gullón, Yáñez, Parajó, & Alonso, 2013) and mass spectrometry (MS) (Leijdekkers, Huang, Bakx, Gruppen, & Schols, 2015; Ognyanov et al., 2016). MS allows structural patterns of complex molecules to be determined and has become the analytical method of choice in metabolomics research. However, it generates large amounts of intricate data that needs to be interpreted to extract chemical information and ensure that they are of valuable (Boccard & Rudaz, 2014, Yi et al., 2016). The identification of unknown compounds is the main bottleneck, so computational tools assisting structure elucidation and *de novo* identification of small molecules have been developed including *in silico* fragmentation methods (Allard et al., 2016; Ruttkies, Schymanski, Wolf, Hollender, & Neumann, 2016). These methods help in the deduction of possible structures of metabolites in spectra interpretation. *Competitive fragmentation modelling* (CFM) has been proposed as an *in silico* fragmentation method suitable for common MS techniques such as GC-EI and significantly outperforms existing state-of-the-art computational methods (Allen, Greiner, & Wishart, 2015; Allen, Pon, Greiner, & Wishart, 2016). It has been validated and tested on the NIST database, and it could be used as an alternative when no reference standard is available for measurement.

Other computational tools are focused on data modelling to find reproducible patterns and discover valuable information on biological events and chemical/structural properties (Käll, Canterbury, Weston, Noble, & MacCoss, 2007; Yi et al., 2016). A number of machine learning classification methods have been applied in the analysis of MS spectra. Supervised and semi-supervised classification methods support *a priori* known data structures to train

patterns and rules to predict new data. When the relationship between MS data and chemical structures is complex, simple classifiers may not be efficient. Today there exists a multitude of algorithms to determine classification patterns among samples, including support vector machines (SVM) and random forests (RF), which can also be applied to the field of metabolomics (Lin et al., 2014; Uarrota et al., 2014) and also food science (Lim et al., 2017; Sabater et al., 2018b). Tree-based models have been used for proteomic mass spectra classification (Geurts et al., 2005) and recent works reported high prediction rates based on MALDI-TOF data (Rossel & Arbizu, 2018). Artificial neural networks (ANN) and boosted models are other machine learning algorithms used to classify spectral data of small molecules (Gosav, Praisler, & Birsa, 2011), which recognize chemical substructures from MS data (Varmuza, He, & Fang, 2003), feature selection in MS-based proteomic profiling (Gertheiss & Tutz, 2009) or protein biomarker discovery (Yasui et al., 2003). It has been reported that machine learning has greatly improved performance relative to the bond-breaking approaches and even CFM (Schymanski et al., 2017).

Therefore, the aim of this study was characterised by GC-MS novel artichoke POS obtained by enzymatic hydrolysis using commercial enzyme preparations with different main activities. In order to accomplish this, structural characteristics of di- and tri-POS were determined by mass spectral mining using three machine learning algorithms: multilayer perceptron (MLP), random forest (RF) and boosted logistic regression (BLR).

2. Materials and methods

2.1. Standards and reagents

Analytical reference substances such as sucrose, D-arabinose, D-xylose, L-rhamnose, D-galactose, D-mannose, D-glucose, galacturonic acid (GalA), digalacturonic acid (Di-GalA), kestose, nystose and β-phenyl glucoside were purchased from Sigma Aldrich (Steinheim,

Germany). Trigalacturonic acid (Tri-GalA) was from Carbosynth (Compton, UK). Four commercial enzyme preparations were studied (**Table 1**). Cellulase from *Aspergillus niger* was acquired from Sigma Aldrich (Steinheim, Germany). The rest of the enzyme preparations were a generous gift from Novozymes (Bagsvaerd, Denmark).

Artichoke pectin was previously extracted in our laboratory using a commercial cellulase, Celluclast®1.5L (artichoke by-product powder concentration 6.5%, enzyme dose 10.1 U g^{-1}, extraction time 48 h). This pectin has GalA content of 69.5%, degree of methylation of 19.5% and molecular mass (Mw) ranging from 4.8 to 660 kDa (Sabater et al., 2018a).).

2.2. Enzyme characterisation

Enzyme characterisation assays were carried out following the method described by Martínez, Gullón, Yánez, Alonso, and Parajó (2009) with modifications. A solution of 2% (w/v) of polygalacturonic acid, chosen as HG standard (Sigma, purity>85%, GalA content greater than 96%), was dissolved in 0.05 M acetate buffer (pH 5.0). Pectinase activity was measured by quantifying the amount of reducing sugar groups liberated after incubation of polygalacturonic acid solutions with 5 mg mL^{-1} or 5 µL mL^{-1} of enzyme at 50 °C for 5 min, using the method of DNS and GalA as standard. One unit of pectinase activity was defined as the amount of enzyme required to release 1 µmol of GalA per min at 50 °C.

2.3. Formation of pectic-oligosaccharides (POS)

Pectic-oligosaccharides (POS) were obtained by enzymatic hydrolysis of 2% (w/v) artichoke pectin solutions dissolved in 0.05 M acetate buffer (pH 5.0) using 0.54-6.75 U mL^{-1} of enzyme (Table 1) following the method of Gómez, Yanez, Parajó and Alonso (2016) with some modifications.

144 Enzymatic hydrolysis was performed in a final volume of 1 mL incubated in an orbital
145 shaker at 50 ºC and 750 rpm. Aliquots were withdrawn from the reaction mixture at the
146 different times (0.5, 1 and 4 h) and immediately immersed in boiling water for 5 min to
147 inactivate the enzyme. Samples were stored at −18 ºC for subsequent analysis. Enzymatic
148 reactions were carried out in duplicate and analyses were performed twice for each enzymatic
149 treatment. In addition, four hydrolysis replicates were prepared for reactions incubated at 4 h
150 for their GC-MS characterization.

Table 1. Determination of commercial enzyme preparation activities (U g^{-1} and U mL^{-1}) using polygalacturonic acid as substrate.

Enzyme preparation	Microorganism	Declared activity	Hydrolase activity (U g^{-1})	Enzyme dose** (U mL^{-1})
Pectinex®Ultra-Olio	*Aspergillus aculeatus/ Aspergillus niger*	Pectin-lyase	202.6 ± 7.4[a,b]*	6.75
Glucanex®200G	*Trichoderma harzianum*	Glucanase	189.2 ± 28.5[b]	0.63
Pentopan®Mono-BG	*Thermomyces lanuginosus*	1,4-endoxylanase	161.5 ± 3.6[b]	0.54
Cellulase from *A. niger*	*Aspergillus niger*	Cellulase	263.7 ± 18.5[a]	0.88

*Enzyme preparation in liquid form: U mL^{-1}.

** Enzyme dose used for POS production

[a,b] Statistically significant differences between enzymes

156 *2.4. Analytical techniques*

157 Monosaccharides were quantified by GC-FID as trimethyl silylated oximes (TMSO)
158 (Sabater et al., 2018a). In addition, samples containing di- and tri-POS formed were analysed
159 by GC-MS on an Agilent Technologies 7890A gas chromatograph coupled to a 5975CMSD
160 quadrupole mass detector (Agilent Technologies) to characterise low M_w POS obtained.
161 Separations were carried out using a fused silica capillary column HP-5MS (5% phenyl
162 methylsilicone, 30 m × 0.25 mm × 0.25 µm thickness; J&W Scientific, Folsom, CA, USA).
163 Helium was used as carrier gas at a flow rate of 0.8 mL min^{-1}. Injector temperature was 200
164 °C. The oven temperature was initially 200 ºC and held for 5 min, then increased at a rate of
165 3 ºC min^{-1} to 250 ºC and held for 1 min, then increased at a rate of 10 ºC min^{-1} to 320 ºC and
166 held for 70 min. Injections were made in the split mode (1:5). The mass spectrometer was
167 operated in electrospray ionisation mode at 70 eV. Mass spectra were acquired using Agilent
168 ChemStation MSD software. Internal standard (β-phenyl glucoside) was added to the
169 samples. Identification of trimethylsilyl oxime derivatives of carbohydrates was carried out
170 by comparison of their relative retention times and mass spectra with those of standard
171 compounds (GalA, Di-GalA and Tri-GalA).

172 Average M_w distribution of POS formed were determined by HPSEC-ELSD
173 following the methods described by Sabater et al. (2018a). For HPSEC-ELSD analysis,
174 samples (0.65 mg mL^{-1}) were dissolved in water filtered using a 0.45 µm syringe filter
175 (Symta, Spain) and analysed in an Agilent Technologies 1220 Infinity LC System
176 (Böblingen, Germany) The separation of carbohydrates was carried out on a TSK-GEL
177 G5000PW_{XL} column (300 mm x 7.8 mm, 10 µm particle size) and TSK-GEL G2500PW_{XL}
178 column (300 mm x 7.8 mm, 6 µm particle size) in combination with a TSK-GEL PW_{XL} guard
179 column (40 mm x 6 mm, 12 µm particle size) (Tosoh Bioscience, Montgomeryville, PA,
180 USA) using 0.01 M NH_4Ac, as mobile phase and elution in isocratic mode at a flow rate of

0.5 mL min^{-1} for 50 min. Molecular weight of carbohydrates was calculated by the external calibration method using solutions of commercial pullulan standards (M_w 0.342-788 kDa) (Fluka Analytical) in the range 10–2250 mg L^{-1}.

POS obtained with different commercial enzyme preparations were quantified by HPAEC-PAD using a DIONEX ICS2500 system (Dionex Corp., Sunnyvale, CA, USA) incorporating a GP50 gradient pump and an ED50 electrochemical detector using a gold and Ag/AgCl as working and reference electrodes, respectively. Separations were carried out at room temperature in a CarboPac PA-1 column (2 x 450 mm) and a CarboPac PA 1 guard column (4 x 50 mm) as a flow rate of 1 mL min^{-1}. The mobile phases were (A) 0.1 M NaOH and (B) 1 M NaOAc in 0.1 M NaOH). The elution profile was as follows: 0-15 min, 0-5% B; 15-60 min, 5-70% B; 60-65 min, 70-100% B; 65-70 min, 100% B; 70-70.1 min, 100-0% B; and finally column re-equilibration by 0% B from 70.1 to 85 min. Pectin neutral monosaccharides (arabinose, xylose, rhamnose, galactose), neutral di- tri- and tetrasaccharides (sucrose, kestose and nystose) as well as GalA and its derivatives (Di- GalA and Tri-GalA) standards were used for identification. Acquisition and processing of data were achieved with Chromeleon 6.7 software (Dionex Corp. CA, USA).

2.5. Data analysis

ANOVA tests and Tukey´s test for $p < 0.05$ were applied to all data generated.

Structural characteristics of POS obtained with Pentopan®Mono-BG, Glucanex®200G, Cellulase from *A. niger* and Pectinex®Ultra-Olio were determined by GC-MS spectral mining. In order to find reproducible patterns which could be applied on new samples, three machine learning models were evaluated in a supervised classification task: multilayer perceptron (MLP), random forest (RF) and boosted logistic regression (BLR). To ensure these algorithms were trained with valuable information, *m/z* fragments whose abundances

were statistically different among groups ($p < 0.05$) corresponding to known POS ruptures were selected.

Signal processing. To increase model performance, discrete wavelet transform (DWT) was applied to GC-MS data (Xia, Wu, & Yuan, 2007; Li, Li, & Yao, 2007). In this study, a 16 level DWT was computed (indicating the depth of the decomposition) with a "la8" decomposition filter (Daubechies orthonormal compactly supported wavelet of length = 8). In the threshold step, DWT coefficients under percentile 15% were removed to denoise the signal. Once the signal was reconstructed, all variables were scaled and centered.

Classification of MS spectra. 462 mass spectra of known pectic sugars and unknown POS released during enzymatic hydrolysis were classified using MLP, RF and BLR.

MLP is the most common kind of artificial neural network (ANN), a family of broadly used models that allow modeling complex and highly non-linear processes. MLP is formed by an input layer (i.e. pre-processed GC-MS spectra), an output layer (i.e. enzyme used to obtain POS) and several neurons or nodes organised in hidden layers, where each neuron in a layer is connected with each neuron in the next layer through a weighted connection. In this case, an MLP was built with 1 hidden layer consisting in 25 neurons. The activation function (a transformation applied to the input signal to determine whether the information that the neuron is receiving is relevant or not) was logistic.

In RF, a multitude of decision trees are constructed, outputting the different classes. Each node is split using the best among a subset of predictors randomly chosen. In this case, RF model was built with 500 trees and 50 variables tried at each split.

On the other hand, BLR is considered as an ensemble method that uses a weighted average of predictions of individual classifiers (Geurts et al., 2005). The iterations specify the maximal number of trees to be fitted. In this work, a BLR consisting in 11 iterations was chosen.

All the models were trained with 70% of the data, 10-fold cross-validated and then tested with 30% of data from each group (corresponding to new samples). Variable importance analysis was carried out to determine the most influential *m/z* ions in each model. In MLP, influent *m/z* ions were determined by the sum of the product of raw input-hidden and hidden-output connection weights while a permutation of the out of-bag-error was chosen for RF (an estimation of the classification error as trees are added to the forest). For BLR, importance coefficients were determined by calculating the area under the ROC (Receiver Operating Characteristic) curve.

All statistical analyses were computed on R v3.5.0. DWT was performed using wavelets package (Aldrich, 2013). MLP was computed using the RSNNS package (Bergmeir & Benitez, 2012) and RF classification was performed with random Forest package (Liaw & Wiener, 2002). For BLR, caTools package was employed (Tuszynski, 2014).

In silico fragmentation. After determining the most important *m/z* ions in di- and tri-POS classification, chemical structures of some fragments have been proposed employing the competitive fragmentation modeling source code (CFM-ID) developed by Allen et al. (2016). GC-EI fragmentation patterns of POS structures described in the bibliography (Atmodjo, Hao, & Mohnen, 2013) were determined and compared to those of enzymatically obtained POS.

3. Results and discussion

The polygalacturonase activity of four commercial enzyme preparations using polygalacturonic acid as a substrate was studied. As can be seen in **Table 1,** Cellulase from *A. niger*, Pectinex Olio, Glucanex and Pentopan showed high activities (263.7, 202.6, 189.2 and 161.5 U mL^{-1}, respectively). Moreover, these enzymes have different declared enzymatic activities including cellulase, pectin-lyase,, exo-β-D-galactofuranosidase and 1, 4-

endoxylanase, respectively and, therefore, different hydrolysis patterns could be expected. The complementary activities could be of great importance, so Cellulase from *A. niger* presented higher polygalacturonase activity than cellulase activity using polygalacturonic acid and carboxymethyl cellulose as substrates (Sabater et al. 2018a).

3.1. Pectic-oligosaccharides (POS) obtainment from artichoke pectin hydrolysis

Enzymatic obtainment of POS derived from artichoke pectin using the four tested enzymes was studied. In **Table 1** the dose of each enzyme used to hydrolyse artichoke pectin is shown. Monosaccharides released during enzymatic hydrolysis were quantified by GC-FID. **Figure 1** shows the amounts found of each compound, GalA, neutral monosaccharides derived from pectins, such as xylose, arabinose, rhamnose and galactose as well as other unidentified monosaccharides present in artichoke hydrolysates. Pectinex Olio released significantly higher amounts of GalA (14.8 mg 100 mg^{-1} pectin) and neutral sugars (11.5 mg 100 mg^{-1} pectin), as expected, due to the high amount of added enzyme (6.75 U mL^{-1}), to attempt obtaining a large amount of low M_w POS. This elevated GalA release may also be due to their declared activity, pectin-lyase. In the other enzymatic preparations, although the enzyme dose was similar at 0.54, 0.63 and 0.88 U mL^{-1} of reaction mixture, the hydrolysis patterns of artichoke pectin were very different, probably because of their different main activities. Therefore, Cellulase from *A. niger* released a significantly major amount of GalA and Glucanex released mainly arabinose (in amounts significantly higher than Cellulase from *A. niger* and Pentopan but still lower than the ones obtained with Pectinex Olio). Lower hydrolysis rates were expected for Glucanex considering its main activity (glucanase) which may not produce significant ruptures in the pectin backbone. Results obtained with Cellulase from *A. niger* indicate significant pectinase secondary activity, even higher than its declared activity. On the contrary, the hydrolytic activity of Pentopan (α-1,4-endoxylanase) was very

low showing a significantly lower monosaccharide release, probably due to a lower presence of xylose in artichoke pectin ramified chains. Interestingly, Cellulase from *A. niger* and Pectinex Olio also released 0.2-0.5 mg 100 mg^{-1} pectin of unidentified monosaccharides probably due to other secondary enzymatic activities. Neutral monosaccharides released from artichoke pectin hydrolysis, mainly arabinose and galactose, depict the relevance of a side chain of RG-I structures, possibly arabinans, galactans and arabinogalactans. As previously reported in Sabater et al. (2018a), artichoke pectin contains a higher percentage of neutral monosaccharides compared to pectin extracted from apple and citrus pectin (Bonnin, Garnier, & Ralet, 2014). Considering these differences, artichoke POS would be expected to show different structures to those derived from other pectin sources.

M_w distribution of enzymatic hydrolysates was determined by HPSEC-ELSD. Pectin from artichokes showed three main fragments with M_w of 660, 105 and 4.8 kDa, as has been previously reported (Sabater et al., 2018a) and after enzymatic treatments, differences in the chromatographic profiles were observed depending on the enzyme selected (**Figure 2).** As can be seen, Pectinex Olio and Cellulase from *A. niger* (**Figure 2b**) showed the most different patterns of POS average M_w distribution compared to the original product (**Figure 2a**), producing the release of a wider variety of fragments and higher amounts of POS with M_w between 50 and 3 kDa (fractions with M_w < 3 kDa were not quantified due to coelution with other compounds present in enzymatic preparations). In contrast, chromatographic profiles of Pentopan hydrolysates showed hardly any modification with respect to artichoke pectin while Glucanex produced an important reduction de M_w producing modified pectins (M_w ~ 50 kDa) (**Figure 2a** and **c**).

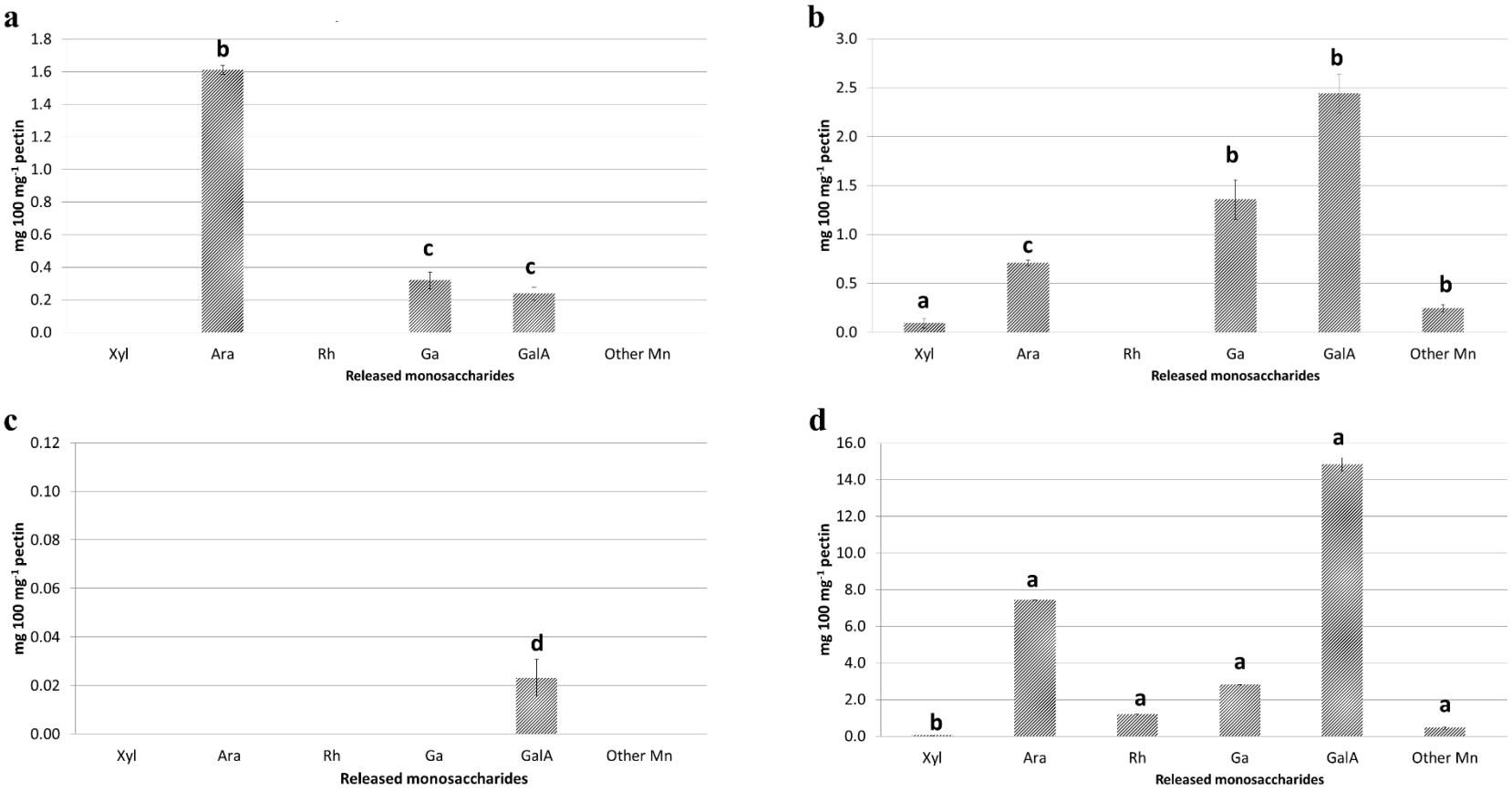


Figure 1. Monosaccharides (mg 100 mg^{-1} pectin) found in enzymatic hydrolysates of artichoke pectin after 4 h of reaction using: **a)** Glucanex® 200G (0.63 U mL^{-1}); **b)** Cellulase from *Aspergilus niger* (0.88 U mL^{-1}); **c)** Pentopan® Mono-BG (0.54 U mL^{-1}); **d)** Pectinex® Ultra-Olio (6.75 U mL^{-1}).Xyl: xylose, Ara: arabinose, Rh: rhamnose, Ga: galactose, GalA: galacturonic acid, Other Mn: unidentified monosaccharides. [a,b,c,d] Statistically significant differences between enzymes.

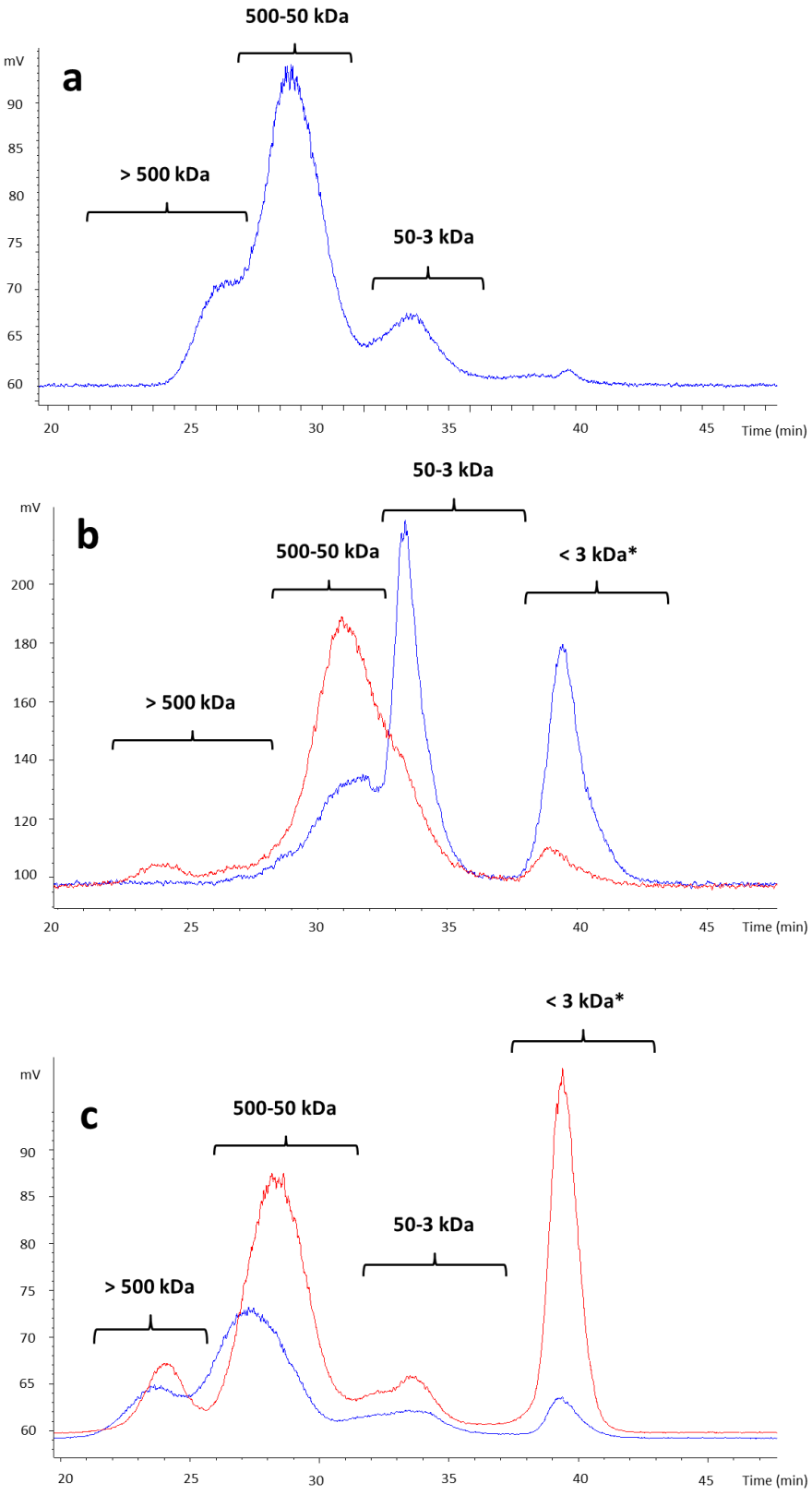
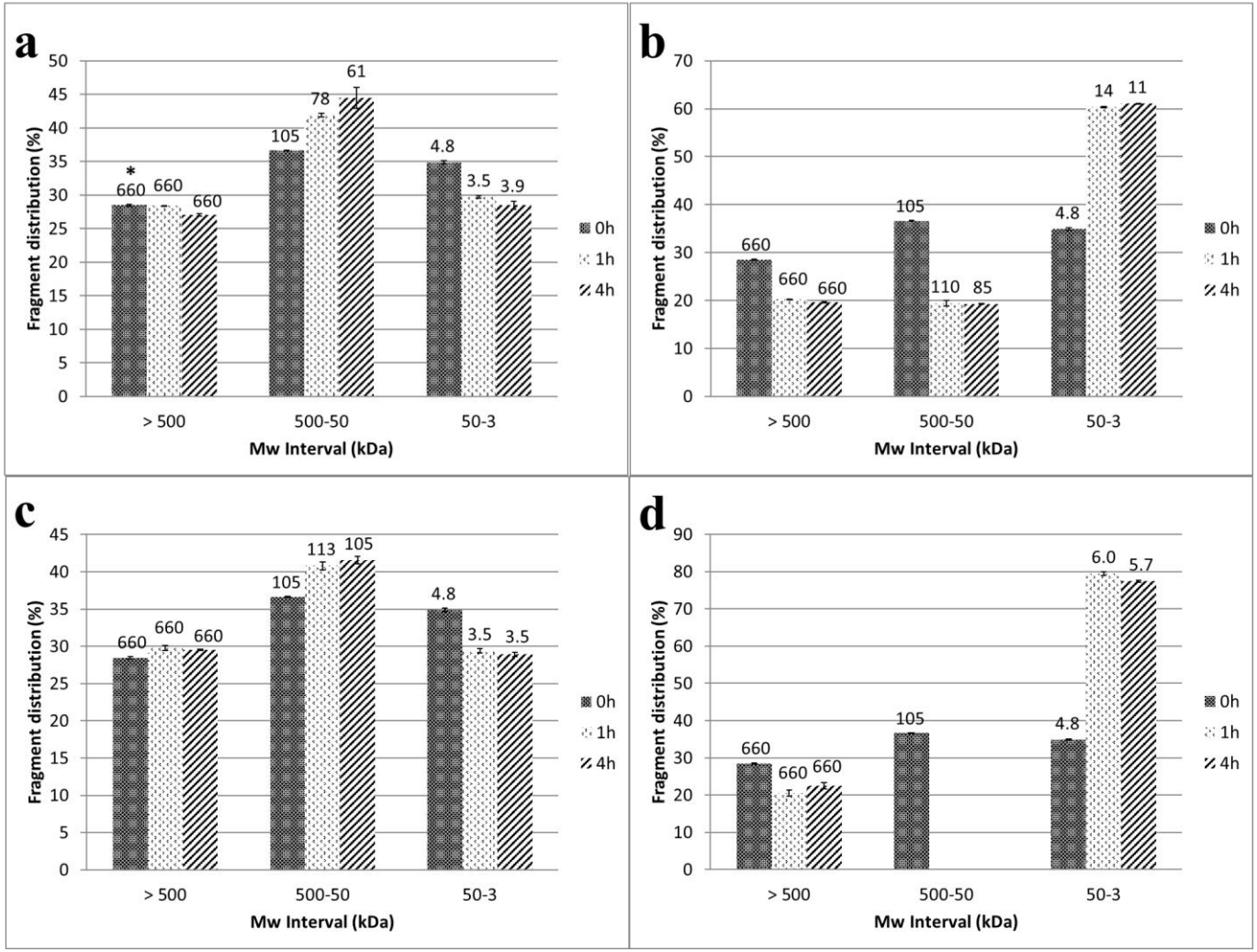


Figure 2. HPSEC-ELSD profiles and molecular weight ranges of a) artichoke pectin; and hydrolysates obtained from artichoke pectin after 4 h of reaction using b) Cellulase from *A. niger* (red) and Pectinex®Ultra-Olio (blue), c) Pentopan®Mono-BG (blue) and Glucanex®200G (red). *Compounds not quantified due to coelution with other present in enzymatic preparations.

315 In general, average M_w distribution of POS produced with these preparations varied from
316 78 to 3.5 kDa (**Figure 3**). Enzymatic hydrolysis with Cellulase from *A. niger* produced, after
317 1 h of reaction, POS of M_w around 14 kDa (60.3% of fragments determined) and the M_w
318 decreased slightly after 4 h up to 11 kDa, while Pectinex Olio formed high amounts of POS
319 with lower M_w (5.7-6.0 kDa, 77-80% of fragments determined). As has been stated, no
320 changes were observed in the HPSEC-ELSD profiles of the other of hydrolysates (Figure 2c).
321 These results highlight the importance of enzymatic activity in POS structural characteristics
322 and several studies reported the influence of enzyme and reaction conditions over POS
323 formation (Combo et al., 2012; Babbar et al., 2016).

324 Differences in hydrolysis patterns were also observed in the HPAEC-PAD analysis.
325 Figure 4 shows a chromatographic profile of carbohydrates found in enzymatic hydrolysate
326 of artichoke pectin after 4 h of reaction with enzyme preparations. Neutral sugars were eluted
327 before acid sugars, GalA (peak 1), Di-GalA (peak 2), Tri-GalA (peak 3), and unknown acid
328 POS with degree of polymerization (DP) of 2, 3 and > 3 were found. In comparison to
329 retention times with commercial standards, these compounds could correspond to
330 oligosaccharides with one/two molecules of GalA and one or more molecules of neutral
331 sugars. Total acid POS formed with the four enzymes used were quantified (**Table 2**) and
332 ranged from 33.4 to 310.6 mg g^{-1} pectin. Pectinex Olio released significantly higher amounts
333 of GalA, Di-GalA and Tri-GalA. Interestingly, Cellulase from *A. niger* formed significant
334 higher amounts of larger oligosaccharides. Maximum formation of POS was reached at 4 h of
335 enzymatic hydrolysis, with a POS yield of 45.7, 63.7, 128.5 and 310.6 mg g^{-1} pectin for
336 Pentopan, Glucanex, Pectinex Olio and Cellulase from *A. niger*, respectively.



337

338 **Figure 3**. Molecular mass distribution of fragments with M_w > 3 kDa determined by HPSEC-ELSD of different hydrolysates of
339 artichoke pectin after 1 and 4 h of reaction using: **a)** Glucanex® 200G (0.63 U mL^{-1}); **b)** Cellulase from *Aspergilus niger* (0.88 U mL^{-1});
340 **c)** Pentopan® Mono-BG (0.54 U mL^{-1}); **d)** Pectinex® Ultra-Olio (6.75 UmL^{-1}). *Average M_w of the fragments within M_w interval.

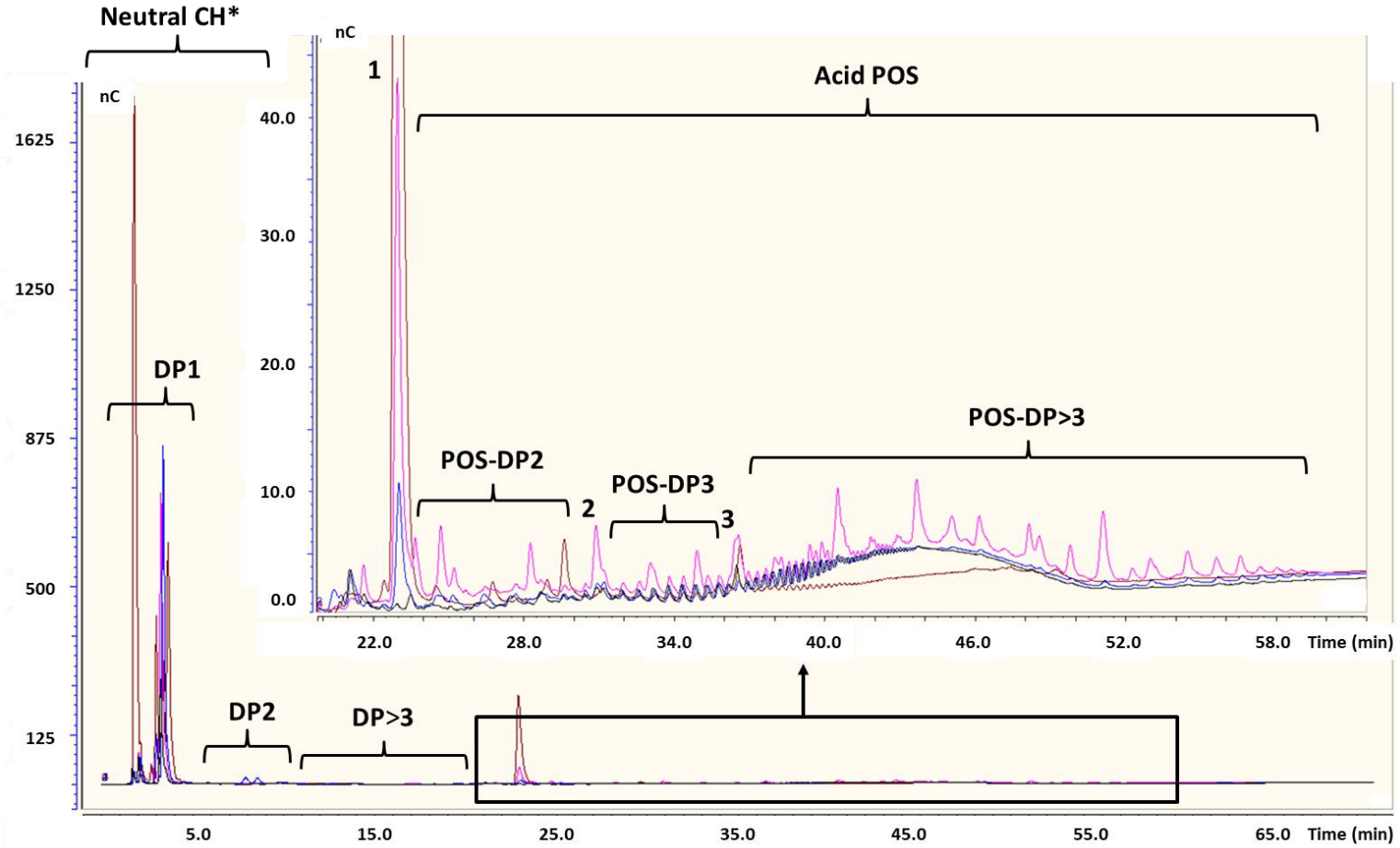


Figure 4. HPAEC-PAD profiles of carbohydrates found in artichoke pectin hydrolysates using commercial enzyme preparation Cellulase from *A. niger* (pink), Pectinex Olio (brown), Glucanex (blue) and Pentopan (black) (50 ºC, pH =5.0, 4h, 0.54-6.75 U mL^{-1} enzyme). CH: carbohydrates, POS: pectic-oligosaccharides. Peaks: DP1, DP2 and DP3 neutral mono-, di- and trisaccharides DP > 3, neutral oligosaccharides; (1) Galacturonic acid; (2) Digalacturonic acid; (3) Trigalacturonic acid; POS-DP2, acid disaccharides; POS-DP3, acid trisaccharides; POS-DP>3, others acid oligosaccharides. (*) Compounds not quantified.

Table 2. POS formation (HPAEC-PAD) and galacturonic acid (GalA) release (mg g^{-1} pectin) during enzymatic hydrolysis (0.5, 1 and 4h) of artichoke pectin using the four studied enzymes. Di-GalA: digalacturonic acid, Tri-GalA: trigalacturonic acid, POS: pectic-oligosaccharides.

Hydrolysis	Time (h)	GalA (peak 1)	Acid POS (mg g^{-1} pectin)					
			POS-DP2	Di-GalA (peak 2)	POS-DP3	Tri-GalA (peak 3)	POS-DP > 3	Total POS*
Pentopan®Mono-BG	0.5	3.0 ± 0.2 ^{d}	12.2 ± 1.0 ^{d}	7.8 ± 0.6 ^{b}	6.4 ± 0.2 $^{c, d, e}$	7.0 ± 0.5 $^{a, b}$	-	33.4 ± 2.8 ^{e}
	1	6.5 ± 0.6 ^{d}	12.6 ± 0.9 ^{d}	8.0 ± 0.6 ^{b}	7.7 ± 0.5 $^{c, d}$	7.4 ± 0.2 $^{a, b}$	-	35.7 ± 2.5 ^{e}
	4	10.0 ± 0.7 ^{d}	12.8 ± 1.1 ^{d}	8.2 ± 0.6 ^{b}	7.3 ± 0.2 $^{c, d}$	7.4 ± 0.5 $^{a, b}$	-	45.7 ± 3.9 ^{e}
Glucanex®200G	0.5	6.5 ± 0.6 ^{d}	17.9 ± 1.3 $^{b, c, d}$	7.3 ± 0.5 ^{b}	0.1 ± 0.0 ^{e}	6.2 ± 0.2 $^{b, c}$	9.6 ± 0.8 ^{c}	41.1 ± 2.9 ^{e}
	1	11.0 ± 0.8 ^{d}	20.2 ± 1.4 $^{b, c}$	7.3 ± 0.5 ^{b}	4.6 ± 0.1 $^{d, e}$	7.7 ± 0.5 $^{a, b}$	22.1 ± 1.6 ^{c}	61.9 ± 5.3 $^{d, e}$
	4	11.4 ± 1.0 ^{d}	23.3 ± 1.6 $^{a, b}$	7.7 ± 0.5 ^{b}	6.9 ± 0.5 $^{c, d, e}$	7.1 ± 0.2 $^{a, b}$	18.7 ± 1.6 ^{c}	63.7 ± 4.5 $^{d, e}$
Cellulase from *A. niger*	0.5	15.0 ± 1.1 ^{d}	17.9 ± 1.3 $^{b, c, d}$	7.1 ± 0.5 ^{b}	11.7 ± 0.3 ^{c}	7.2 ± 0.5 $^{a, b}$	184.5 ± 13.0 ^{b}	228.4 ± 19.4 ^{b}
	1	22.6 ± 1.9 ^{d}	20.2 ± 1.4 $^{b, c}$	7.3 ± 0.5 ^{b}	9.1 ± 0.6 $^{c, d}$	7.6 ± 0.2 $^{a, b}$	196.2 ± 16.6 $^{a, b}$	240.4 ± 17.0 ^{b}
	4	43.0 ± 3.0 ^{d}	23.3 ± 2.0 $^{a, b}$	7.4 ± 0.5 ^{b}	47.3 ± 1.3 ^{b}	7.4 ± 0.5 $^{a, b}$	225.2 ± 15.9 ^{a}	310.6 ± 26.4 ^{a}
Pectinex® Ultra-Olio	0.5	247.3 ± 21.0 ^{c}	16.0 ± 1.1 $^{c, d}$	5.8 ± 0.4 ^{b}	53.5 ± 3.8 $^{a, b}$	8.4 ± 0.2 ^{a}	13.4 ± 1.1 ^{c}	97.1 ± 6.9 $^{c, d}$
	1	382.1 ± 27.0 ^{a}	22.1 ± 1.9 ^{b}	13.8 ± 1.0 ^{a}	55.6 ± 1.6 ^{a}	8.3 ± 0.6 ^{a}	15.2 ± 1.1 ^{c}	115.0 ± 9.8 ^{c}
	4	303.8 ± 25.8 ^{b}	29.0 ± 2.1 ^{a}	16.1 ± 1.1 ^{a}	55.2 ± 3.9 ^{a}	5.0 ± 0.1 ^{c}	23.2 ± 1.6 ^{c}	128.5 ± 9.1 ^{c}

*__Total POS__: Σ POS-DP2, Di-GalA, POS-DP3, Tri-GalA and POS-DP > 3.

^{a,b,c,d,e} Statistically significant differences between groups.

3.2. GC-MS characterisation of POS-DP2 and POS-DP3 obtainment from artichoke pectin

To delve into structural characteristics of novel artichoke POS mixtures released during enzymatic hydrolysis, a GC-MS study of these oligosaccharides was carried out. In **Figure 5** a CG-MS profile of hydrolysate of artichoke pectin using Cellulase from *A. niger* is shown. Di-GalA and Tri-GalA as well as unknown POS-DP2 and POS-DP 3 were detected. Similar GC-MS profiles have been obtained for the other enzymes used.

Interestingly, specific *m/z* ions 277, 321, 333 and 423, derived from β-cleavage fragmentation of uronic acids, as well as GalA characteristic ions *m/z* 332 and 540 were detected in Di-GalA MS spectra (Petersson, 1974; Füzfai, Kovács, & Molnár-Perl 2004). Taking into account the scarce information found in the bibliography about the GC-MS spectra of this type of compound a mass spectral study employing three machine learning algorithms (multilayer perceptron, MLP; random forest, RF; and boosted logistic regression, BLR) was carried out. Therefore, full mass spectra of all disaccharides and trisaccharides and unknown POS found in the enzymatic hydrolysates were classified. The number of spectra included in this study is shown in **Table 3**. A total of 462 MS spectra were used for data analysis and classified according to the enzyme used, Glucanex-unknown POS (n = 104); Cellulase from *A. niger*-unknown POS (n =116); Pentopan-unknown POS (n = 44); Pectinex Olio-unknown POS (n = 128) and different known pectic sugars (arabinose, rhamnose, xylose, galactose, GalA and Di- and Tri-GalA, n = 70).

For data analysis, 151 fragments in the range of *m/z* 61-546, which were statistically different among groups ($p < 0.05$) and might correspond to known POS ruptures, assessed by CFM-ID (Allen et al., 2016), were selected. Then, each MS spectra was decomposed and reconstructed using the DWT. Unknown POS were classified according to the enzyme used (Glucanex, Pentopan, Cellulase from *A. niger* and Pectinex Olio), and were

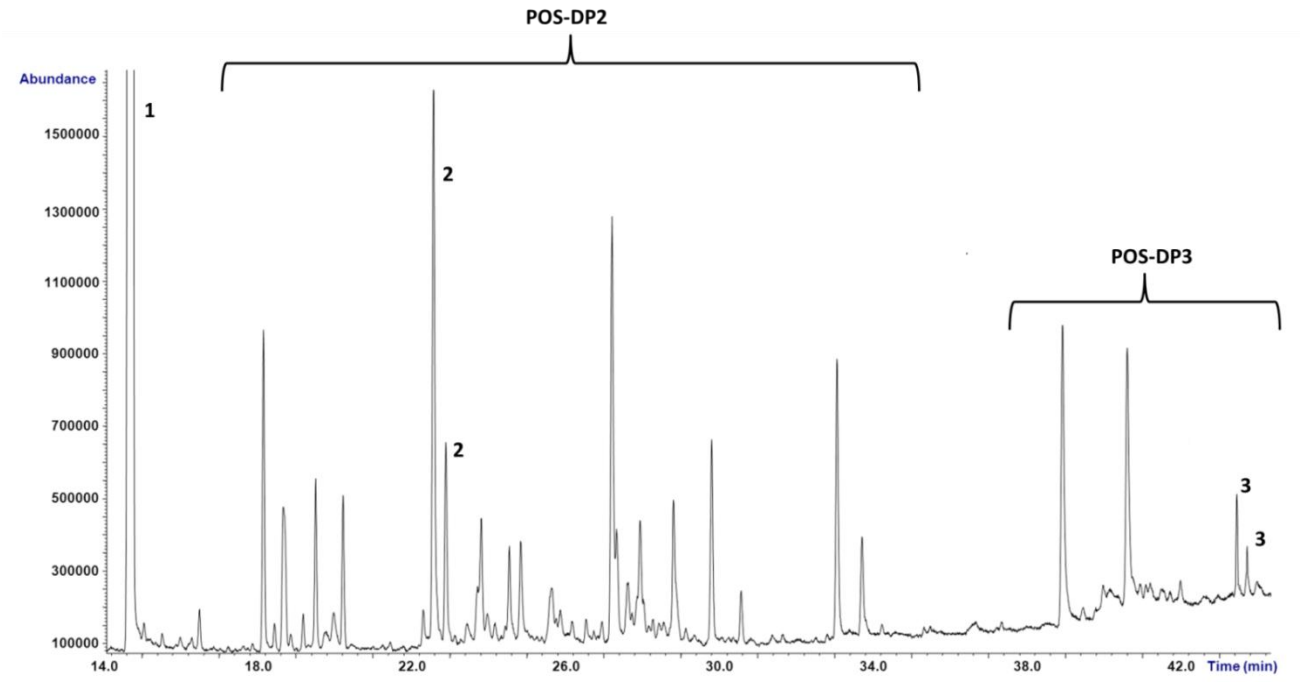


Figure 5. GC-MS profile of hydrolysate of artichoke pectin obtained by incubation with cellulase from *A. niger*. Peaks: (1) Internal standard, (2) Digalacturonic acid (Di GalA), (3) Trigalacturonic acid (Tri-GalA). POS-DP2: unknown pectic disaccharides, POS-DP3: unknown pectic trisaccharides.

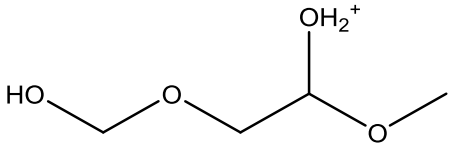
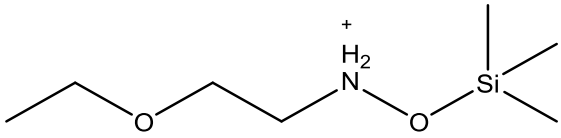
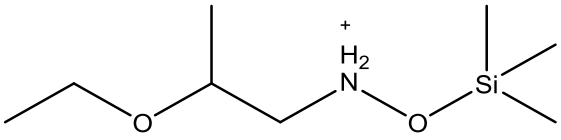
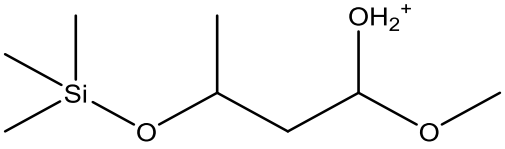
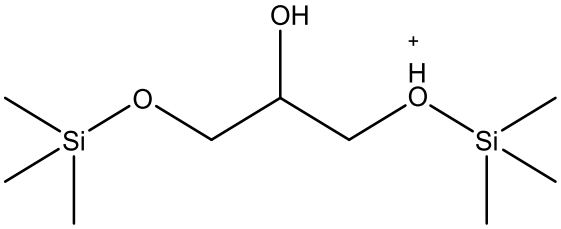
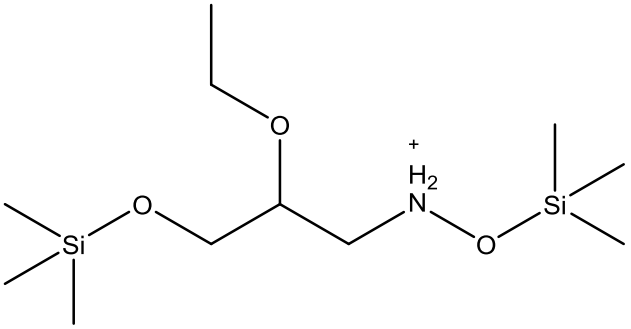
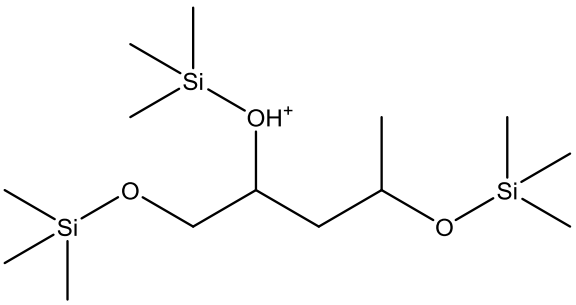
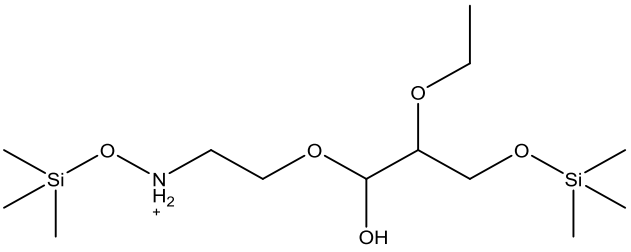
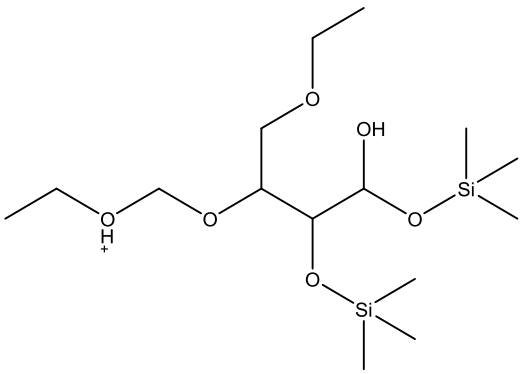
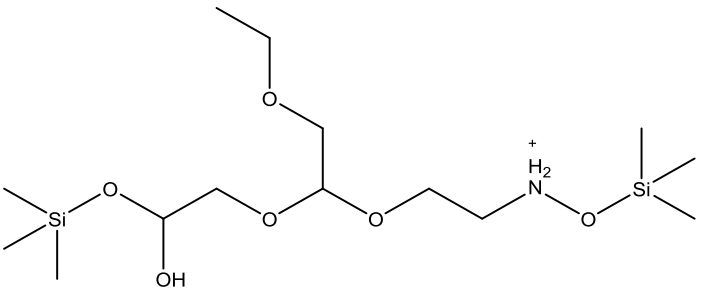
382 **Table 3.** Number of GC-MS spectra (n = 462) of oligosaccharides obtained from GC-MS-EI analysis of hydrolysates from artichoke pectins
383 included in machine learning study. GalA: galacturonic acid, Di-GalA: digalacturonic acid, Tri-GalA: trigalacturonic acid, POS: pectic-
384 oligosaccharides.

Artichoke pectin + enzyme	**Commercial standards**	**Number of hydrolysates**	**Number of GC-MS spectra**					
			Known pectic monosaccharides	**Di-GalA**	**Unknown POS-DP2**	**Tri-GalA**	**Unknown POS-DP3**	**Total**
Glucanex®200G	-	4	-	8 (2 per hydrolysate)	60 (15 per hydrolysate)	8 (2 per hydrolysate)	44 (11 per hydrolysate)	16 known POS 104 unknown POS
Cellulase from *A. niger*	-	4	-	8 (2 per hydrolysate)	84 (21 per hydrolysate)	8 (2 per hydrolysate)	32 (8 per hydrolysate)	16 known POS 116 unknown POS
Pentopan®Mono-BG	-	4	-	8 (2 per hydrolysate)	24 (6 per hydrolysate)	8 (2 per hydrolysate)	20 (5 per hydrolysate)	16 known POS 44 unknown POS
Pectinex®Ultra-Olio	-	4	-	8 (2 per hydrolysate)	112 (28 per hydrolysate)	-	16 (4 per hydrolysate)	8 known POS 128 unknown POS
-	Standard	-	10[a]	2[b]	-	2[c]	-	14

385 [a] Known monosaccharides including galacturonic acid, arabinose, rhamnose, xylose and galactose. [b] Digalacturonic acid standards. [c] Trigalacturonic acid standards.

386 also differentiated from known pectic sugars. These models were validated and tested on 30%
387 of new samples. The training, 10-fold cross-validation and test rates were:
388 a) MLP: 100, 91.1 and 95.7%, respectively.
389 b) RF: 100, 97.8 and 100%, respectively.
390 c) BLR: 100, 98.1 and 100%, respectively.
391 As can be observed, machine learning algorithms found a highly reproducible
392 classification pattern that could be applied on new samples. DWT pre-processing allowed
393 model prediction rates to increase up to 95-100%. When these models were tested on new
394 samples, MLP showed false negative rates of 0.72, 1.45 and 2.17% for POS obtained with
395 Glucanex, Pentopan and Pectinex Olio preparations, respectively. The false positive rates
396 were 0.72% for known pectic sugars, also for POS obtained with cellulase from *A. niger*,
397 Pentopan and Pectinex Olio were 0.72% and 1.60% for POS from Glucanex preparations.
398 Although MLP showed high accuracy in its predictions (above 95%), RF and BLR correctly
399 classified 100% of test samples. Therefore, RF and BLR showed 100% sensitivity, specificity
400 and balanced accuracy values for each class. In contrast, MLP showed test sensitivity values
401 between 85.7-96.7% for Pentopan, Pectinex Olio and Glucanex POS and 100% for Cellulase
402 from *A. niger* POS and known sugars. MLP specificity values ranged from 98.2 to 99.2% for
403 all the classes and balanced accuracy between sensitivity and specificity was 92.5, 95.7, 97.4,
404 99.5 and 99.6% for Pentopan, Pectinex Olio, Glucanex and Cellulase from *A. niger* POS and
405 known sugars, respectively. Interestingly, MLP possessed higher specificity for POS
406 classification, obtaining the lowest accuracy for Pentopan POS. However, all these values
407 were above 90%. Higher accuracy for known sugars might be due to structural similarities
408 among POS mixtures. The most influential *m/z* ions and their importance in each model were
409 determined (**Supplementary material Tables S1, S2 and S3**).
410

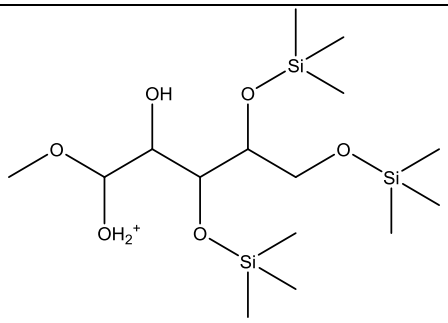
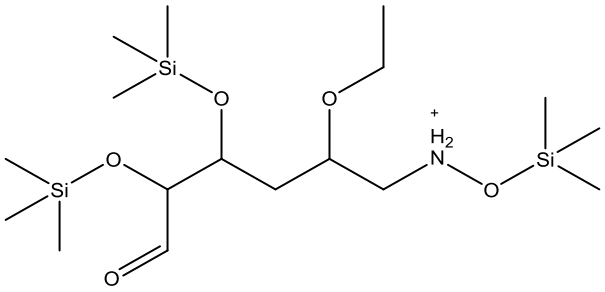
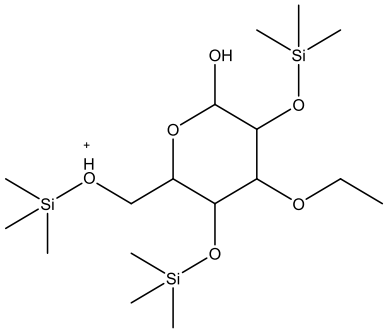
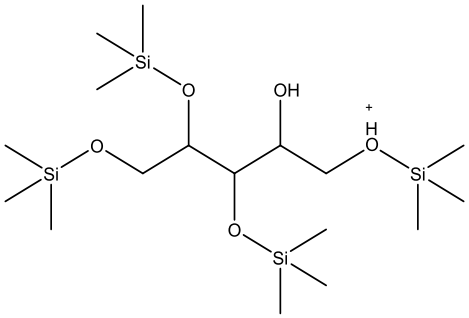
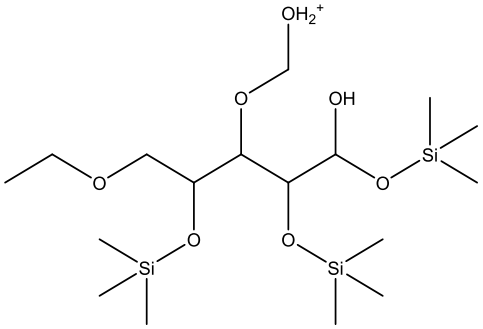
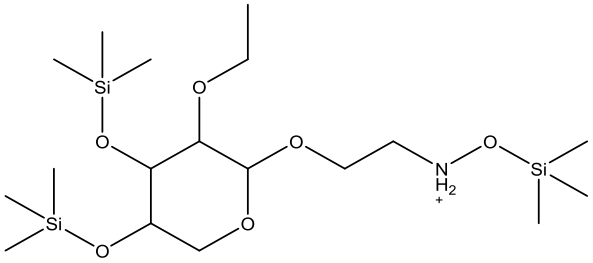
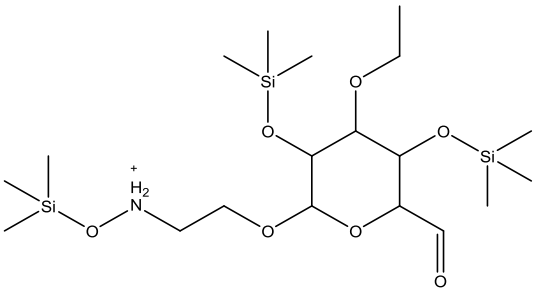
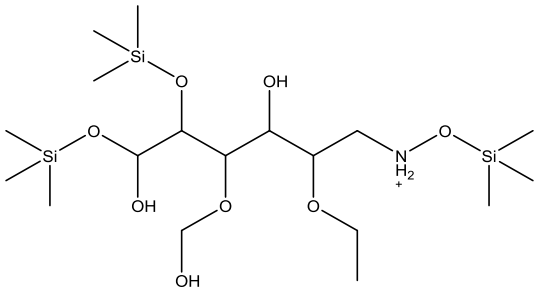
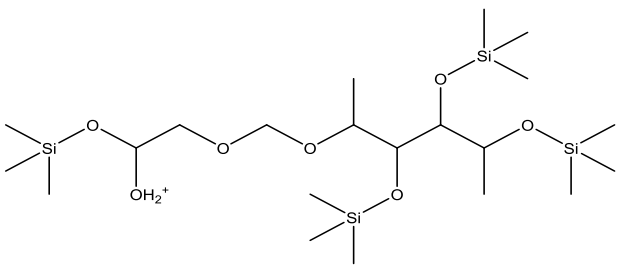
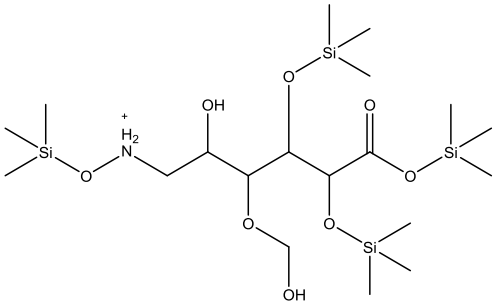
Table 4. Possible chemical structures of the most relevant *m/z* ions in POS classification given by machine learning algorithms, determined by CFM-ID. These ions correspond to TMSO fragments from pectic di- or trisaccharides.

m/z	Structure	*m/z*	Structure
123		**178**	
192		**193**	
237		**280**	
337		**340**	
369		**370**	

414

415

416 **Table 4**. Cont.

m/z	Structure	*m/z*	Structure
399		**424**	
425		**441**	
443		**454**	
482		**488**	
529		**530**	

In addition, overall accuracy and kappa values obtained with each model via their resampling distributions were compared (a comparative account is shown in Supplementary material Figure S1). The accuracy indicates the number of instances that were classified correctly, while kappa is a more robust measure that takes into account the possibility of a correct classification by chance. Both values were similar in the three models. BLR and RF showed mean accuracy and kappa values of 0.97-0.98, significantly higher than those of MLP (0.91 and 0.89, respectively). In all three models, these metrics were high, indicating high predictive power.

To deepen into POS structural differences established by these models, a total 50 *m/z* values of the most relevant fragments were selected and probable chemical structures of these fragments were suggested employing CFM-ID code (Allen et al., 2016). Proposed structures of some interesting *m/z* ions (n=20) are shown in **Table 4,** while the other structures of *m/z* ions (n=30) are shown in **Supplementary material Table S4**

First, complete feasible fragments for candidate molecules (i.e. monosaccharides, POS-DP2 and DP3) were calculated by systematically breaking bonds within the molecule (Allen et al., 2015; Allen et al., 2016) in order to obtain a POS *in silico* fragmentation library. For monosaccharides, pectic neutral sugars (arabinose, rhamnose, xylose and galactose) and GalA (methylated/acetylated or not) were considered. The POS-DP2 included, taking into account more possible structures present in HG and RG-I, were Di-GalA (methylated or not), xylose-α(1,3)-GalA (methylated/acetylated or not), GalA-α(1,2)-rhamnose (acetylated or not), rhamnose-α(1,4)-GalA (acetylated or not), galactose-α(1,4)-rhamnose and arabinose-α(1,4)-rhamnose (Atmodjo et al., 2013; Mohnen, 2008). Tri-POS library was generated considering combinations of all POS-DP2 structures and pectic monomers, following the main criteria reported by Mohnen (2008) and Atmodjo et al. (2013). Then, we proposed a chemical structure for each *m/z* ion by looking for these specific ruptures in our library for

studying structural characteristics of known pectic sugars and unknown POS released during enzymatic hydrolysis. The abundances of several of the most relevant ions in machine learning classification were significantly higher ($p < 0.05$) in known pectic sugars released by the four enzymes, so fragments *m/z* 207 and 309 could be originated from all compounds studied. Ions *m/z* 311, 382, 414, 486, 498, 500, 501, 502, 513, 528, 532, 544, 545 and 546 were also higher in known pectic sugars and could also be derived from GalA-containing POS or GalA oligomers bonded by α(1,4) glycosidic linkages.

Then, statistically significant structural differences among unknown POS-DP2 and POS-DP3 released with the four preparations were studied:

Pentopan. Some *m/z* ions were significantly more abundant in unknown POS present in these hydrolysates: ion *m/z* 237 which could be derived from all pectic sugars and POS and *m/z* 193, 425 and 530 originated from xylose and POS ruptures. Other more specific ions higher in these POS were *m/z* 488 originated from POS containing rhamnose-α(1,4)-GalA (acetylated); ion *m/z* 399 is formed from GalA-α(1,2)-rhamnose while *m/z* 529 is formed from POS containing rhamnose-α(1,4)-GalA (acetylated), GalA-α(1,2)-rhamnose (acetylated) or xylose-α(1,3)-GalA (methylated/acetylated or not); and ion *m/z* 454 which could also be formed from POS containing arabinose-α(1,4)-rhamnose.

Glucanex. Some fragments with higher abundance in unknown POS from Glucanex, *m/z* 280 derived from rhamnose-α(1,4)-GalA (acetylated), *m/z* 370 derived from GalA-α(1,2)-rhamnose (acetylated), and *m/z* 369 which could be derived from acetylated dimers of rhamnose-α(1,4)-GalA or GalA-α(1,2)-rhamnose as well as GalA-containing POS bonded by α(1,4) glycosidic linkages. Other important ions in these POS were *m/z* 178 and 192 which could also be originated from the rupture of xylose-α(1,3)-GalA (acetylated). Interestingly, *m/z* 123 derived from POS containing GalA (methylated) was significantly higher in

Glucanex POS and lower in those obtained with cellulase from *A. niger* perhaps this enzyme produced slight demethylation.

Pectinex Olio. The MS study of unknown POS obtained with Pectinex Olio revealed that the abundances of *m/z* 337, which could be originated from all pectic sugars and all POS, and *m/z* 340, 424 and 443, specifically originated from rhamnose-α(1,4)-GalA, GalA-α(1,2)-rhamnose or xylose-α(1,3)-GalA (methylated/acetylated or not), were significantly higher in these hydrolysates.

Cellulase from A. niger. MS spectra of unknown POS formed with Cellulase from *A. niger* showed higher abundances of ions *m/z* 441, originated from galactose and all POS, and *m/z* 482, specifically derived from GalA-α(1,2)-rhamnose (acetylated) and xylose-α(1,3)-GalA (acetylated).

Other *m/z* ions which were relevant in more than one group were *m/z* 531 present in all POS containing GalA units and *m/z* 483 present in galactose-α(1,4)-rhamnose or GalA-α(1,2)-rhamnose. Similarly, abundances of *m/z* 353, 354, 457, 458, 471 and 472 were higher in unknown POS and could be derived from the rupture of most POS, with special importance for those containing rhamnose-α(1,4)-GalA, GalA-α(1,2)-rhamnose or xylose-α(1,3)-GalA. More specifically, *m/z* 225, 444, 469, 484 and 542, formed from the rupture of GalA and molecules containing GalA-α(1,4)-GalA dimers, were relevant in known pectic sugars and also in unknown POS from Pentopan and Glucanex preparations. Finally, *m/z* 439 formed from rhamnose-α(1,4)-GalA, GalA-α(1,2)-rhamnose and galactose-α(1,4)-rhamnose, was relevant in both hydrolysates with Pentopan or with Cellulase from *A. niger* preparations. At last, the dimer structures above referred may be also present in trisaccharides, but no trisaccharides specific m/z ions were relevant for POS classification.

The assignment of these glycosidic linkages was possible due to the most frequent structures found in pectins (Mohnen, 2008 and Atmodjo et al., 2013).

As expected, *m/z* fragments possibly formed from the rupture of dimers and trimers containing neutral sugars and GalA (methylated/acetylated or not) units were relevant in unknown POS. In general terms, these results suggest that Glucanex and Pectinex Olio preparations produce POS that may contain rhamnose, xylose and GalA (methylated/acetylated or not). In addition, arabinose may also be present in POS structures produced by Pentopan. Moreover, Cellulase from *A. niger* may produce POS that contain neutral sugars such as galactose, rhamnose and xylose, and also acetylated GalA units.

A GC-MS spectral study might be considered as a first approach to determine structural characteristics of POS enzymatically obtained using commercial preparations with different activities. MS data modelling may lead to a better understanding of differences observed in the chromatographic profiles of these samples. It has been demonstrated that is possible to classify complex oligosaccharides according to the enzyme used for their obtainment, by extracting chemically relevant information from their full spectra. These models could be applied on new reaction mixtures containing novel oligosaccharides to evaluate the activity of different enzyme preparations considering their high prediction rates and allowing consistent structural information to be obtained. The ability to determine structural similarities and differences among novel POS can be of great importance to establish structure-function relationships which could be very useful to extrapolate results from biological assays.

CONCLUSIONS

Novel artichoke POS mixtures have been enzymatically obtained and characterised. Differences in the chromatographic profiles of POS were observed according to the enzyme used suggesting different structural properties, confirmed by several chromatographic techniques. Cellulase from *A. niger* and Pectinex® Ultra-Olio formed large amounts of POS, with M_w around 14 and 6.0 kDa, respectively while Glucanex produced high M_w POS and

516 modified pectins, and Pentopan-showed chromatographic profiles similar to those of original
517 artichoke pectin. POS were analysed by HPAEC-PAD, showing that Cellulase from *A. niger*
518 preparation produced the highest amount (310.6 mg g^{-1} pectin). Mass spectra of unknown
519 POS-DP2 and -DP3 have been studied and classified using machine learning algorithms
520 (multilayer perceptron, random forest and boosted logistic regression) obtaining high
521 prediction rates on the test set. These models confirm structural differences observed in the
522 hydrolysis profiles of commercial preparations with various enzymatic activities and could be
523 used to establish structure-function relationships.
524

525 **Acknowledgments**

526 This work has been funded by MICINN of Spain, Projects AGL2014-53445-R and
527 AGL2017-84614-C2-1-R and by the Spanish Danone Institute. Carlos Sabater thanks his
528 FPU Predoc contract from Spanish MECD (FPU14/03619). Authors are also thankful to
529 Riberebro (La Rioja, Spain) for kindly providing the artichoke by-products studied in this
530 paper and Ramiro Martinez (Novozyme) for enzyme supply.
531

532 **References**

533 Aldrich, E. (2013). Wavelets: A package of functions for computing wavelet filters, wavelet
534 transforms and multiresolution analyses. R package version 0.3-0. URL
535 https://CRAN.R-project.org/package=wavelets. Last accessed: 30/07/18
536 Allard, P. M., Péresse, T., Bisson, J., Gindro, K., Marcourt, L., Pham, V. C, Roussi, F.,
537 Litaudon, M., & Wolfender, J. L. (2016). Integration of molecular networking and in-
538 silico MS/MS fragmentation for natural products dereplication. *Analytical*
539 *chemistry*, *88*, 3317-3323.

Allen, F., Greiner, R., & Wishart, D. (2015). Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics*, *11*, 98-110.

Allen, F., Pon, A., Greiner, R., & Wishart, D. (2016). Computational prediction of electron ionization mass spectra to assist in GC/MS compound identification. *Analytical Chemistry*, *88*, 7689-7697.

Atmodjo, M. A., Hao, Z. Y., & Mohnen, D. (2013). Evolving Views of Pectin Biosynthesis. Ed. S. S. Merchant. *Annual Review of Plant Biology*, *64*, 747-779.

Babbar, N., Dejonghe, W., Gatti, M., Sforza, S., & Elst, K. (2016). Pectic oligosaccharides from agricultural by-products: production, characterization and health benefits. *Critical reviews in biotechnology*, *36*, 594-606.

Babbar, N., Dejonghe, W., Sforza, S., & Elst, K. (2017). Enzymatic pectic oligosaccharides (POS) production from sugar beet pulp using response surface methodology. *Journal of food science and technology*, *54*, 3707-3715.

Bergmeir, C., & Benitez J. M. (2012). Neural Networks in R Using the Stuttgart Neural Network Simulator: RSNNS. Journal of Statistical Software, 46, 1-26. URL http://www.jstatsoft.org/v46/i07/. Last accessed: 30/07/18

Boccard, J., & Rudaz, S. (2014). Harnessing the complexity of metabolomic data with chemometrics. *Journal of Chemometrics*, *28*, 1-9.

Bonnin, E., Garnier, C., & Ralet, M. C. (2014). Pectin-modifying enzymes and pectin derived materials: applications and impacts. *Applied Microbiology and Biotechnology*, 98(2), 519-532.

Combo, A. M. M., Aguedo, M., Goffin, D., Wathelet, B., & Paquot, M. (2012). Enzymatic production of pectic oligosaccharides from polygalacturonic acid with commercial pectinase preparations. *Food and bioproducts processing*, *90*, 588-596.

564 Combo, A. M. M., Aguedo, M., Quiévy, N., Danthine, S., Goffin, D., Jacquet, N., Blecker,
565 C., Devaux, J., & Paquot, M. (2013). Characterization of sugar beet pectic-derived
566 oligosaccharides obtained by enzymatic hydrolysis. *International journal of*
567 *biological macromolecules*, *52*, 148-156.
568 da Moura, F. A., Macagnan, F. T., & da Silva, L. P. (2015). Oligosaccharide production by
569 hydrolysis of polysaccharides: a review. *International Journal of Food Science &*
570 *Technology*, *50*, 275-281.
571 Füzfai, Z., Kovács, E., & Molnár-Perl, I. (2004). Identification and quantitation of the main
572 constituents of sour cherries: Simultaneously, as their trimethylsilyl derivatives, by
573 gas chromatography-mass spectrometry. *Chromatographia*, *60*, S143-S151.
574 Gan, C. Y., & Latiff, A. A. (2011). Extraction of antioxidant pectic-polysaccharide from
575 mangosteen (*Garcinia mangostana*) rind: Optimization using response surface
576 methodology. *Carbohydrate Polymers*, *83*, 600-607.
577 Gertheiss, J., & Tutz, G. (2009). Supervised feature selection in mass spectrometry-based
578 proteomic profiling by blockwise boosting. *Bioinformatics*, *25*, 1076-1077.
579 Geurts, P., Fillet, M., De Seny, D., Meuwis, M. A., Malaise, M., Merville, M. P., &
580 Wehenkel, L. (2005). Proteomic mass spectra classification using decision tree based
581 ensemble methods. *Bioinformatics*, *21*, 3138-3145.
582 Gómez, B., Gullón, B., Yáñez, R., Parajó, J. C., & Alonso, J. L. (2013). Pectic
583 oligosacharides from lemon peel wastes: Production, purification, and chemical
584 characterization. *Journal of Agricultural and Food Chemistry*, *61*(42), 10043-10053.
585 Gómez, B., Yañez, R., Parajo, J. C., & Alonso, J. L. (2016). Production of pectin-derived
586 oligosaccharides from lemon peels by extraction, enzymatic hydrolysis and membrane
587 filtration. *Journal of Chemical Technology and Biotechnology*. *91*, 234-247.

588 Gosav, S., Praisler, M., & Birsa, M. L. (2011). Principal component analysis coupled with
589 artificial neural networks—A combined technique classifying small molecular
590 structures using a concatenated spectral database. *International Journal of Molecular*
591 *Sciences*, *12*, 6668-6684.
592 Gullón, B., Gómez, B., Martínez-Sabajanes, M., Yáñez, R., Parajó, J. C., & Alonso, J. L.
593 (2013). Pectic oligosaccharides: Manufacture and functional properties. *Trends in*
594 *Food Science & Technology*, *30*, 153-161.
595 Herbstreith & Fox. (2018). The Specialists for Pectin. URL http://www.herbstreith-
596 fox.de/index.php?id=54&L=1. Last accessed: 30/07/18
597 Käll, L., Canterbury, J. D., Weston, J., Noble, W. S., & MacCoss, M. J. (2007). Semi-
598 supervised learning for peptide identification from shotgun proteomics
599 datasets. *Nature methods*, *4*, 923-925.
600 Leijdekkers, A. G., Huang, J. H., Bakx, E. J., Gruppen, H., & Schols, H. A. (2015).
601 Identification of novel isomeric pectic oligosaccharides using hydrophilic interaction
602 chromatography coupled to traveling-wave ion mobility mass spectrometry.
603 *Carbohydrate Research*, *404*, 1-8.
604 Li, X., Li, J., & Yao, X. (2007). A wavelet-based data pre-processing analysis approach in
605 mass spectrometry. *Computers in Biology and Medicine*, *37*, 509-516.
606 Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*,
607 18-22.
608 Lim, D. K., Long, N. P., Mo, C., Dong, Z., Cui, L., Kim, G., & Kwon, S.W. (2017).
609 Combination of mass spectrometry-based targeted lipidomics and supervised machine
610 learning algorithms in detecting adulterated admixtures of white rice. *Food Research*
611 *International*, *100*, 814-821.

612 Lin, Z., Gonçalves, C. M. V., Dai, L., Lu, H. M., Huang, J. H., Ji, H., Wang, D., Yi, L., &
613 Liang, Y. (2014). Exploring metabolic syndrome serum profiling based on gas
614 chromatography mass spectrometry and random forest models. *Analytica Chimica*
615 *Acta*, *827*, 22-27.
616 Maric, M., Grassino, A. N., Zhu, Z. Z., Barba, F. J., Brncic, M., & Brncic, S. R. (2018). An
617 overview of the traditional and innovative approaches for pectin extraction from plant
618 food wastes and by-products: Ultrasound-, microwaves-, and enzyme-assisted
619 extraction. *Trends in Food Science & Technology*, *76*, 28-37.
620 Martínez, M., Gullón, B., Yanez, R., Alonso, J. L., & Parajó, J. C. (2009). Direct Enzymatic
621 Production of Oligosaccharide Mixtures from Sugar Beet Pulp: Experimental
622 Evaluation and Mathematical Modeling. *Journal of Agricultural and Food Chemistry*,
623 *57*, 5510-5517.
624 Mohnen, D. (2008). Pectin structure and biosynthesis. *Current opinion in plant biology*, *11*,
625 266-277.
626 Ognyanov, M., Remoroza, C., Schols, H. A., Georgiev, Y., Kratchanova, M., & Kratchanov,
627 C. (2016). Isolation and structure elucidation of pectic polysaccharide from rose hip
628 fruits (*Rosa canina* L.). *Carbohydrate Polymers*, *151*, 803-811.
629 Petersson, G. (1974). Gas-chromatographic analysis of sugars and related hydroxy acids as
630 acyclic oxime and ester trimethylsilyl derivatives. *Carbohydrate Research*, *33*, 47-61.
631 Rossel, S., & Martínez Arbizu, P. (2018). Automatic specimen identification of Harpacticoids
632 (Crustacea: Copepoda) using Random Forest and MALDI-TOF mass spectra,
633 including a post hoc test for false positive discovery. *Methods in Ecology and*
634 *Evolution*, *9* 1421-1434.

635 Ruttkies, C., Schymanski, E. L., Wolf, S., Hollender, J., & Neumann, S. (2016). MetFrag
636 relaunched: incorporating strategies beyond *in silico* fragmentation. *Journal of*
637 *Cheminformatics*, *8*, 3.
638 Sabater, C., Corzo, N., Olano, A., & Montilla, A. (2018a). Enzymatic extraction of pectin
639 from artichoke (*Cynara scolymus L.*) byproducts using Celluclast®1.5L.
640 *Carbohydrate Polymers, 190,* 43–49.
641 Sabater, C., Montilla, A., Ovejero, A., Prodanov, M., Olano, A., & Corzo, N. (2018b).
642 Furosine and HMF determination in prebiotic-supplemented infant formula from
643 Spanish market. *Journal of Food Composition and Analysis, 66,* 65–73.
644 Schymanski, E. L., Ruttkies, C., Krauss, M., Brouard, C., Kind, T., Dührkop, K., Allen, F.,
645 Vaniya, A., Verdegem, D., Böcker, S., Rousu, J., Shen, H., Tsugawa, H., Sajed, T.,
646 Fiehn, O., Ghesquière, B., & Neumann, S. (2017). Critical Assessment of Small
647 Molecule Identification 2016: automated methods: *Journal of Cheminformatics*, 9, 1-
648 21.
649 Tuszynski, J. (2014). caTools: Tools: moving window statistics, GIF, Base64, ROC AUC,
650 etc.. R package version 1.17.1. https://CRAN.R-project.org/package=caTools. Last
651 accessed: 30/07/18
652 Uarrota, V. G., Moresco, R., Coelho, B., da Costa Nunes, E., Peruch, L. A. M., de Oliveira
653 Neubert, E., Rocha, M., & Maraschin, M. (2014). Metabolomics combined with
654 chemometric tools (PCA, HCA, PLS-DA and SVM) for screening cassava (Manihot
655 esculenta Crantz) roots during postharvest physiological deterioration. *Food*
656 *Chemistry*, *161*, 67-78.
657 Varmuza, K., He, P., & Fang, K. T. (2003). Boosting applied to classification of mass
658 spectral data. *Journal of Data Science*, *1*, 391-404.

659 Xia, J. M., Wu, X. J., & Yuan, Y. J. (2007). Integration of wavelet transform with PCA and
660 ANN for metabolomics data-mining. *Metabolomics*, *3*, 531-537.
661 Yasui, Y., Pepe, M., Thompson, M. L., Adam, B. L., Wright Jr, G. L., Qu, Y., Potter, J. D.,
662 Winget, M., Thornquist, M., & Feng, Z. (2003). A data-analytic strategy for protein
663 biomarker discovery: profiling of high-dimensional proteomic data for cancer
664 detection. *Biostatistics*, *4*, 449-463.
665 Yi, L., Dong, N., Yun, Y., Deng, B., Ren, D., Liu, S., & Liang, Y. (2016). Chemometric
666 methods in data processing of mass spectrometry-based metabolomics: a
667 review. *Analytica Chimica Acta*, *914*, 17-34.

Supplementary data

Click here to download Supplementary data: New Supplementary Material.docx