



University of Southern Denmark

Evaluation of spectral libraries and sample preparation for DIA-LC-MS analysis of host cell proteins

A case study of a bacterially expressed recombinant biopharmaceutical protein

Heissel, Søren; Bunkenborg, Jakob; Kristiansen, Max Per; Holmbjerg, Anne Fich; Grimstrup, Marie; Mørtz, Ejvind; Kofoed, Thomas; Højrup, Peter

Published in:

Protein Expression and Purification

DOI:

10.1016/j.pep.2018.03.002

Publication date:

2018

Document version:

Accepted manuscript

Document license:

CC BY-NC-ND

Citation for published version (APA):

Heissel, S., Bunkenborg, J., Kristiansen, M. P., Holmbjerg, A. F., Grimstrup, M., Mørtz, E., Kofoed, T., & Højrup, P. (2018). Evaluation of spectral libraries and sample preparation for DIA-LC-MS analysis of host cell proteins: A case study of a bacterially expressed recombinant biopharmaceutical protein. *Protein Expression and Purification*, 147, 69-77. <https://doi.org/10.1016/j.pep.2018.03.002>

Go to publication entry in University of Southern Denmark's Research Portal

Terms of use

This work is brought to you by the University of Southern Denmark.

Unless otherwise specified it has been shared according to the terms for self-archiving.

If no other license is stated, these terms apply:

- You may download this work for personal use only.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying this open access version

If you believe that this document breaches copyright please contact us providing details and we will investigate your claim. Please direct all enquiries to puresupport@bib.sdu.dk

1 **Title**

2 SWATH-MS based absolute quantification of host cell proteins associated with a biopharmaceutical protein.

3 **Authors**

4 Søren Heissel 1), Jakob Bunkenborg 2), Max Per Kristiansen 3), Anne Fich Holmbjerg 3), Marie Grimstrup 2),
5 Ejvind Mørtz 2), Thomas Kofoed 2), Peter Højrup* 1)

6 *Corresponding author: php@bmb.sdu.dk

7 1) Department of Biochemistry and Molecular Biology, University of Southern Denmark, 5230 Odense M,
8 Denmark

9 2) Alphalyse A/S, 5220 Odense SØ, Denmark

10 3) Statens Seruminstitut, 2300 Copenhagen S, Denmark

11 **Abstract**

12 Recombinantly expressed biopharmaceutical proteins often undergo a series of purification steps with the aim of
13 removing contaminating material. Depending on the application of the protein, there are various requirements for the
14 degree of purity, but host cell proteins (HCP) will in general remain in small amounts. We describe a method using
15 microflow LC-MS/MS with extensive, multidimensional spectral libraries followed by 1D SWATH-acquisition for fast
16 and robust absolute quantification across several orders of magnitude of the HCPs associated with a
17 biopharmaceutical protein. The HCP content of this BP has never been studied by LC-MS, and the calculated purity
18 was found to be in agreement with ELISA data. A set of evaluated reference proteins was spiked into the sample
19 before and after digestion and evaluated for absolute label-free quantification through a Top3 quantification strategy.
20 The label-free quantification was further evaluated using stable isotope standard (SIS) peptides from selected HCPs.
21 The methodology provides a general, fast and cost-efficient identification and label-free absolute quantification
22 strategy, which can easily be transferred to different samples for HCP characterization and purity assessment. The
23 dynamic range of the HCPs was established, and it was shown that 50 % of the molar HCP amount originated from less
24 than 10 proteins. The developed strategy can be deployed during development processes or quality control for other
25 biopharmaceutical proteins to evaluate the amounts of distinct HCPs and to identify possible risk factors and antigenic
26 protein contaminants. The speed of the analysis (a few days) along with the easy setup allow for evaluation of the
27 purification process in parallel to process development.

28 **Introduction**

29 Recombinant biopharmaceutical proteins are often obtained as a crude mixture containing culture
30 medium, DNA, lipids, proteins and other biological components¹ and are therefore subjected to a series of
31 purification steps, ultimately yielding the purified biopharmaceutical protein. The protein impurities which
32 almost always remain, called host cell proteins (HCPs), are of special interest, since these may cause
33 unwanted immunogenic responses or possess proteolytic activity^{2,3}. Manufacturers are required to monitor
34 and document the HCP content throughout production; however, there are no regulations on the analytical
35 methods applied. HCP content is usually studied using immunogenic methods such as ELISA assays⁴⁻⁶ where
36 polyclonal antibodies are raised through inoculation using proteins produced by the null cell line, i.e. the
37 host cell-line which does not produce the biopharmaceutical protein.

38 In recent years, mass spectrometry (MS) has increasingly been applied for HCP-based workflows^{3,5,7-9},
39 offering identification of individual proteins which may aid in identifying possible risk factors and evaluate
40 the purification process. The main hurdles when analyzing HCPs in biopharmaceutical protein samples by
41 MS are the complexity and the high dynamic range, as the drug protein is often >100 times more abundant
42 than the most abundant HCP. Due to the many challenges, a robust strategy for quantification of HCPs is
43 needed. The biopharmaceutical protein used in this study (henceforth referred to as BP) is produced by
44 Statens Serum Institut (DK) for diagnostic testing in the field of tuberculosis (TB) infections and was chosen,
45 as it has been routinely studied by ELISA assays, providing a benchmark for the results obtained from the
46 LC-MS/MS analyses. The BP is purified by tangential flow filtration (TFF) which removes contaminants
47 below ~10 kDa. Due to remaining non-proteinous buffer components, which may interfere with the
48 analyses, extensive sample clean-up is required before introducing the sample to the mass spectrometer.
49 The most common workflow consists of protein purification, digestion in a urea-containing buffer and
50 clean-up by solid phase extraction. Another strategy, which has shown great potential is the suspension
51 trapping strategy, published by Zougman and coworkers¹⁰, where proteins are precipitated onto a quartz
52 filter trap and desalted before digestion directly on-trap. For absolute quantification of many proteins in a
53 solution, several label-free strategies exist, which are based on spiking the sample with reference proteins
54 in known amounts, providing cost-efficient and sufficiently accurate strategies for absolute protein
55 quantification¹¹⁻¹⁴. A number of studies by Doneanu and coworkers show how multidimensional LC-MS
56 workflows can provide accurate absolute quantitative results of HCPs down to very low amounts^{15,16},
57 however, by applying multidimensional workflows, analysis time is significantly increased. Data dependent
58 acquisition (DDA) was used to identify HCP peptides with high confidence and to select which transitions to
59 monitor, thus functioning as spectral libraries. Quantitative data were recorded using the data independent

60 acquisition (DIA) strategy *Sequential Window Acquisition of all Theoretical Mass Spectra* (SWATH)¹⁷, where
61 ions are fragmented untargeted based on large, predefined selection windows. DIA has been reported to
62 allow for accurate MS2 quantification of several thousands of peptides across a large dynamic range¹⁸.

63 In this study we characterized the HCP content of a BP using DDA and SWATH-MS. Tryptic digests of the
64 raw harvest media (henceforth referred to as P10), and the final commercial product (henceforth referred
65 to as P15) were evaluated for generating the most comprehensive spectral library. P10 and P15 were
66 analyzed both with and without pre-fractionation of the tryptic peptides by high-pH RP-HPLC, which has
67 previously proven a feasible option for increasing the chromatographic power in proteomic workflows^{19,20}.
68 The generated spectral libraries can be used for analyses several months apart, making this method highly
69 suitable for monitoring BPs during production, requiring only quantitative measurements. All quantitative
70 analyses were recorded on a 1-dimensional separation of P15. Absolute HCP quantification was performed
71 by spiking with a set of reference proteins in known amounts and evaluated using stable isotope standard
72 (SIS) peptides originating from a set of selected HCPs. We evaluated acetone precipitation along with MCX
73 purification and suspension trapping for establishing a sample preparation strategy which produces
74 accurate quantitative results. Non-specific cleavage searches of the tryptic digested samples were
75 performed in order to identify low-abundant non-specific products of the biopharmaceutical protein, which
76 may be misinterpreted as HCP peptides and thereby introduce false positives. The HCPs associated with
77 this BP have not previously been characterized by LC-MS, and this study shows that a catalog of potential
78 immunogenic targets can be created fast and cost-efficient. The general setup developed in this study
79 proved to be sensitive, quantitative and robust. Compared to ELISA assays, quantitative data for individual
80 proteins were obtained. This additional information can speed up the development process and
81 improve/simplify the down-stream process. The unique information regarding individual HCPs can improve
82 the regulatory approval process, and in the end increase the possibility for bringing new safe drugs and
83 vaccines to the market.

84 **Experimental**

85 All chemicals, proteins and reagents were purchased from Sigma-Aldrich, St. Louis, MO, USA, unless
86 otherwise specified. The BP was provided by Statens Serum Institut (Copenhagen, Denmark).

87 **Amino acid analysis**

88 For determination of protein amounts, Amino acid analysis (AAA) was applied. Aliquots of protein- or
89 peptide-containing samples were lyophilized by vacuum centrifugation and hydrolyzed using 6M HCl under
90 vacuum at 110°C for 20 hours²¹. The samples were then analyzed by a Biochrom 30+ analyzer (Biochrom

91 Ltd., Cambridge, UK), using post-column derivatization by ninhydrin. The protein concentration in the BP
92 sample was determined to be 0.336 µg/µL upon delivery.

93 **Acetone precipitation strategy**

94 Proteins from 300 µL (100.8 µg) of BP were precipitated by the addition of 4*volumes of ice-cold acetone
95 followed by incubation overnight at -18°C. After incubation, the sample was centrifugated at 14000*g using
96 a Minispin table-top centrifuge (Eppendorf) followed by removal of the supernatant. The pellet was washed
97 with an ice-cold mixture of acetone/ethanol/water (2:2:1) followed by another centrifugation at 14000*g.
98 The washing step was performed twice. The sample was lyophilized and redissolved in 7M urea, 2M
99 thiourea in 50 mM ammonium bicarbonate (AMBIC). Cysteines were reduced using 10 mM dithiothreitol
100 (DTT) at RT for 45 minutes and subsequently alkylated using 23 mM iodoacetamide (IAA) for 45 minutes at
101 RT in the dark. An additional aliquot of DTT was added to quench excess IAA. The sample was diluted to a
102 final concentration of 4.5 M urea, and proteins were digested by the addition of 44 mAU Lysyl
103 Endopeptidase (Wako Chemicals, Japan) followed by incubation for 3 hours at RT. The sample was diluted
104 to 9 times the starting volume using 50 mM AMBIC and modified trypsin (Promega) was added at 4 % w/w
105 and incubated at 37°C for 20 hours. Digestion was stopped by the addition of trifluoroacetic acid (TFA) to a
106 final concentration of 1 %. Peptides were purified by MCX mixed-bed cartridges (Waters, MA, USA)
107 according to manufacturer specifications.

108 **Suspension trap purification**

109 50 µg of the BP sample was lyophilized and resolubilized in 2% SDS, 50 mM AMBIC. Cysteines were reduced
110 and alkylated as previously described, with the exception of being reduced at 57°C. Proteins were purified
111 as described by Zougman, Selby and Banks¹⁰, with minor modifications. Briefly; the quartz traps were
112 constructed using micro-quartz fiber filter paper (Ahlstrom Munktell, Sweden) stomped into a D200 pipette
113 tip (Gilson) using a syringe. No C18 trap was included. A solution of 90 % methanol, 50mM AMBIC was used
114 to precipitate the proteins. Tryptic digestion was done first on the filter at 47°C for 60 minutes using 1 %
115 w/w modified trypsin (Promega), followed by elution to a low-binding Eppendorf tube using 50 mM AMBIC,
116 and further digestion overnight with 2% w/w non-modified trypsin (Sigma) at 37°C.

117 **High-pH fractionation**

118 30 µg of tryptic digested P10 and P15 were dissolved in 50 µL 0.1 % triethylamine (TEA) and separated
119 using a 1260 Infinity HPLC (Agilent Technologies, Germany) equipped with a 2.1*150mm Kinetex Evo C18
120 column (Phenomenex, USA) operating at 45°C and 200 µL/min. Solvent A consisted of 0.1 % TEA in water
121 and solvent B consisted of 90 % ACN. 0.1 % TEA in water. Solvent B levels were kept at 2 % for 2 minutes,

122 and then increased to 40 % over 17 minutes, followed by a sharp increase to 90 % over 6 minutes, where it
123 was kept for 5 minutes. 10 fractions were collected in 2-minute intervals and concatenated to yield 5
124 fractions.

125 **Preparation of protein quantification standards**

126 The reference-protein mix for quantification was constructed as a ladder, consisting of several proteins
127 with varying MW in various amounts. All proteins were purchased from Sigma Aldrich. The ladder consisted
128 of horse myoglobin, bovine carbonic anhydrase II, human serotransferrin and bovine alcohol
129 dehydrogenase. Proteins were dissolved individually in 6 M urea, 50 mM TEAB (Fluka Analytical, St. Louis,
130 MO, USA). Reduction and alkylation was carried out as described above. The protein ladders were
131 constructed using two different strategies; one where the appropriate amounts of reference proteins
132 (spanning several orders of magnitudes) were added to the BP samples, before any other sample
133 preparation followed by protein purification, digestion and clean-up. This strategy was used for calculating
134 the absolute amounts of the HCPs. The other strategy was to digest the reference proteins and purify the
135 peptides individually, determine the amounts by AAA, mix the ladder and spike it into the sample
136 immediately before injection into the MS. This strategy was used to evaluate the accuracy of the
137 quantification strategy. The individual reference proteins were digested as described in the “acetone-MCX”
138 section. All protein digests were purified using Oasis HLB cartridges (Waters, MA, USA) according to the
139 manufacturer’s specifications. Another mixture was created containing peptides from all reference proteins
140 in near-equal amounts, and used for spectral library generation. The Hi3 quantification standard was
141 purchased from Waters (Product number 186006011) and consisted of the 6 best ionizing peptides from
142 the tryptically digested rabbit Phosphorylase B²². The lyophilized peptides were redissolved in UHQ H₂O,
143 aliquoted and frozen for subsequent experiments.

144 **Preparation of stable isotope standard peptides**

145 A total of 21 tryptic peptides from 6 selected HCPs (A2RHZ5, A2RKI9, A2RMQ7, POA3J1, A2RMV9 and
146 A2RL56) were synthesized with a heavy-isotope lysine or arginine (JPT peptides, Berlin, Germany) and
147 delivered non-purified and non-quantified. The 6 HCPs were selected based on their individual estimated
148 quantity in order to span the entire dynamic range of the HCPs. All peptides were purified by RP-HPLC and
149 quantified by AAA. See Supporting Information for procedure and AAA results.

150 **LC-MS/MS analyses**

151 The peptides were separated using an Eksigent 425 microflow HPLC system (Sciex, Warrington, UK)
152 equipped with a reversed phase CSH C18 column (0.3mm, 100mm, 1.9µm, Waters) and analyzed using a

153 TripleTOF 6600 Q-TOF mass spectrometer (Sciex, Warrington, UK) run in high-sensitivity mode. The flow
154 rate was 5 μ L/min and the column temperature was kept at 55°C. Solvent A consisted of 0.1% formic acid
155 (FA) in water. Solvent B consisted of 0.1% FA in acetonitrile. During the 80 minute gradient the solvent B
156 level went from 2 to 29 % in 60 minutes, from 29 to 55 % in 5 minutes, from 55 to 75 % in 2 minutes and
157 remained at 75 % for 5 minutes and then returned to 2 % over 2 minutes and stayed at 2 % for the
158 remaining 6 minutes. DDA analyses were carried out by a full MS1 spectrum with an accumulation time of
159 250 ms, followed by fragmentation of the top 25 most intense multiply charged ions. MS2 spectra were
160 acquired with an accumulation time of 50 ms. The SWATH analyses consisted of a parent spectrum,
161 followed by 52 MS/MS acquisitions with variable window sizes from m/z 349.5-1700. The window
162 distribution consisted of a single window of 21 m/z from 349.5-370.5, followed by a single window of 15
163 m/z from 369.5-384.5, followed by 33 windows of 13 m/z stretching from 383.5-780.5 m/z. This was
164 followed by 9 windows of 26 m/z from 779.5-1005.5 and 6 windows with a size of 51 m/z from 1004.5-
165 1305.5, two windows with a size of 101 m/z from 1304.5-1505.5 and a single window of 195.5 m/z from
166 1504.5-1700. An accumulation time of 33 ms was utilized and the entire cycle time was approximately 1.9
167 seconds. The default charge state was set to 2+. All samples were analyzed in triplicates.

168 **Database searching**

169 The DDA files were recalibrated and converted to mgf-format using Protein Pilot version 5.0.1.0, 4895
170 (Sciex, Warrington, UK) and searched using the search engine MASCOT (Version 2.2.04, Matrix Science Ltd.,
171 London, UK)²³. A mass tolerance of 10 ppm on the MS1 level and 0.1 Da on the MS2 level was applied.
172 Oxidation of methionine was included as variable modification. Carbamidomethylation of cysteine was
173 applied as a static modification. For database searching, the complete proteome of *Lactococcus lactis* strain
174 MG1363 was retrieved from www.uniprot.org (2383 sequences), and the drug protein sequence was added
175 manually. For experiments including protein spikes, the necessary proteins were added manually. All search
176 results had a threshold 1 % FDR on the PSM level and were exported as MASCOT .DAT files for further
177 analysis. For non-specific searches, Andromeda was utilized through MaxQuant²⁴ (v. 1.5.3.30). Peptide
178 lengths of 7-25 amino acids were allowed. The data was searched with a 1 % FDR on PSM and protein level.

179 **Label-free absolute quantification**

180 The MASCOT search results were imported into Skyline version 3.5.0.9319²⁵ and used as spectral
181 libraries. Only unique peptides were processed for quantification. The same mass deviations and
182 parameters as the MASCOT search were applied. Quantification was performed on the P15 sample,
183 spiked with reference proteins, recorded using DIA. The peak areas for the top 5 transitions (γ - and
184 b-ions) per peptide were exported. Subsequent data analysis was performed using R version 3.3.1²⁶
185 and Microsoft Excel version 14.0.7173.5000. Entries with a library dot product (dotp) < 0.7 and an
186 isotope dot product (idotp) < 0.85 were discarded in order to filter out low-quality entries. Peak
187 areas for all transitions were accumulated for each individual peptide to yield a peptide response. All
188 proteins identified with less than 3 peptides were not included in the quantitative analyses. A mean
189 intensity (summed peptide intensity / number of peptides, MeanInt) and Top3 intensity (summed
190 intensity for the 3 most intense peptides) were exported. The signal and amount of the reference
191 proteins were used for deriving absolute HCP amounts. For the comparison with heavy peptide-
192 based quantification, a digest of P15 was purified by acetone-MCX and spiked with purified peptides.

193 **Absolute quantification using heavy peptides**

194 The heavy peptide standards were mixed in equimolar amounts and spiked into 2 μg of BP digest
195 (acetone precipitated and MCX purified) in amounts ranging from 5 fmol to 1 pmol. All dilution steps
196 were analyzed in triplicates using SWATH and quantified using Skyline. The intensity of the three
197 most intense γ -ions were summed and used for quantification. The observed responses from the
198 native peptides were normalized against the heavy responses. Peptides which produced poor
199 linearity throughout the dilution steps were discarded.

200

201 **Results and discussion**

202

203 **Identified HCPs**

204 In order to achieve a complete HCP characterization, it is important to have extensive identification
205 of the peptides belonging to each HCP. Tryptic digests of P10 and P15 (1D and 2D) were analyzed by
206 LC-MS. Using the 1-dimensional strategy, 608 HCPs were identified from P10, while 256 HCPs were
207 identified from P15. P10, being the raw harvest media, had not been subjected to any purification
208 and it is therefore expected that many HCPs are present at higher amounts compared to P15. The
209 high-pH fractionations lead to a significant increase in the number of identified HCPs (1175 HCPs in
210 P10 and 441 HCPs in P15). As seen from Figure 1A, a large subset of proteins were only identified in
211 P10, representing HCPs which are removed to such a high degree during purification that they
212 remain undetectable in the final product.

213 **Utilizing different spectral libraries**

214 The search results were used for generation of spectral libraries and applied to a 1D separated P15
215 digest, recorded using DIA in order to study the effects on quantification. The number of quantifiable
216 proteins was significantly higher when using the 2D-separated samples as libraries. By combining all
217 four strategies in the spectral library, additional peptides were identified, thereby allowing
218 quantification of additional HCPs, which were not quantifiable using the individual libraries. The
219 abundances of these additional proteins were at least two orders of magnitude below the most
220 abundant HCP (as determined by their Top3 values) and therefore these HCPs did not contribute
221 significantly to the combined HCP amount. These proteins do, however, still represent possible
222 immunogenic targets, which are essential to monitor. Figure 1B shows the distribution of
223 quantifiable proteins. The benefits of an extensive spectral library have previously been
224 described^{27,28} and the presented results suggest that a combination of the various libraries provides
225 the most extensive library, which may be used for comprehensive quantification of HCPs in future
226 batches of the BP, requiring only a 1-D separation of a tryptic digest, thereby saving analysis time.

227 **Reference proteins**

228 The PhosB standard was spiked into a BP P15 digest in amounts ranging from 500 amol to 50 pmol
229 per injection, and the Top3 signal of PhosB was normalized against the Top3 signal of a constant HCP
230 in order to establish the linear range. The multi-point calibration curve was compared to a single-
231 point calibration, based on extrapolation of the signal obtained with 500 fmol PhosB. The multi-point
232 calibration curve was identical to the single-point calibration curve between 5 fmol and 5 pmol on-

233 column, which was determined as the linear range. The calibration plots are presented in Supporting
234 Information Figure S-1.

235 A number of diverse protein candidates were evaluated in order to comprise a representative spike-
236 in ladder. Linearity of the protein ladders was investigated using the Top3 method¹⁴ and the MeanInt
237 method¹³. The final ladder consisted of human serotransferrin, bovine carbonic anhydrase 2, horse
238 alcohol dehydrogenase and horse myoglobin. These proteins were selected based on their varying
239 MW, which in turn resulted in a varying number of peptide candidates for quantification. Other
240 proteins were discarded either due to too few peptide candidates or low purity (e.g. β -casein). The
241 quantification strategies were evaluated using the purified peptides spiked into the sample
242 immediately before injection. The spiked proteins spanned 20 – 600 fmol on-column per injection.
243 Calibration curves were constructed using both Top3 and MeanInt. The Top3 method showed the
244 highest degree of linearity across the entire range ($r^2=0.999$) and was therefore chosen as the
245 method of quantification. When using MeanInt, the amount of large proteins tended to be
246 underestimated while overestimating the amounts of small proteins. The Top3 strategy may involve
247 a “one-point calibration” strategy using the signal and amount of a single protein, followed by
248 extrapolation. However, if this protein is not sufficiently digested, or some of the peptides are
249 partially lost during processing, the signal and thereby the calibration curve is greatly affected. By
250 including several proteins in the calibration curve, variations between proteins and in the sample
251 preparation process may be compensated for. Excellent linearity was observed both when spiking
252 with purified peptides and intact proteins. Spiking with intact proteins produced an r^2 of 0.996 for
253 the acetone-MCX experiment and 0.976 for the suspension trap experiment with amounts ranging
254 from 40 fmol to 2.4 pmol per 4 μ g of sample. Supporting Figure S-1 shows the linearity of the
255 different strategies.

256 Signal-to-amount ratios (Top3 signal/fmol) were calculated for all individual reference proteins in
257 order to investigate possible variations between the reference proteins. When spiking with purified
258 peptides, the ratios of all four proteins were consistent throughout the entire range, not varying
259 greatly from each other nor from the ratio obtained using all proteins. When spiking with the intact
260 reference proteins prior to any sample preparation, the 3 highest abundant proteins were found to
261 produce similar ratios in the acetone-MCX experiment, but serotransferrin, which was spiked at 10
262 fmol per μ g of sample, produced a much higher ratio. The deviation of the ratios suggests a
263 difference in recovery rates as a result of inconsistencies during either the precipitation or digestion,
264 but the remaining proteins allowed for compensation of this. The suspension trap experiment
265 showed a higher degree of variance between the ratios of all reference proteins throughout the

266 entire range. It is possible that these variations are a result of the smaller precipitated amount, the
267 precipitation time or the temperature, compared to the acetone-MCX strategy. The overall ratio
268 produced from all reference proteins is quite consistent with the other experiments, however,
269 inconsistent recovery rates will complicate the quantification. The acetone-MCX and suspension trap
270 Top3 values were normalized against an internal reference protein and based on the correlation of
271 the results, the recovery rates of the majority of HCPs was found to be consistent. Figure 2 shows
272 the individual and collected ratios of the reference proteins for the different spiking strategies.

273 **Evaluation of calculated amounts by SIS-peptides**

274 Absolute protein quantification by spiking with heavy-isotope peptides is a very accurate strategy
275 and is often performed using MRM on a triple-quadrupole instrument²⁹. Recent studies show that
276 SWATH-acquisition is, to a high degree, comparable to MRM in obtaining accurate quantitative
277 results^{17,30,31}. Of the 6 target-HCPs, A2RHZ5, A2RKI9, A2RMQ7 and P0A3J1 were repeatedly
278 quantified with 3 or more peptides in the label-free experiment. The protein A2RL56 could not be
279 identified in the analyses. A2RMV9 was only identified with two peptides, and did therefore not
280 meet the requirements for quantification. A2RMV9 was however still included exclusively in the
281 comparative analysis presented in Table 1, for evaluation purposes (Instead of a Top3 signal, the
282 intensity of the two identified peptides were used to construct a “Top2 signal”). When calculating
283 the absolute amounts of the SIS peptides, all peptides belonging to the same protein were found to
284 be present in the same range, differing by a factor of less than 3.5. This variation may be due to
285 inconsistencies in the sample preparation protocol, incomplete digestion³² or loss of the SIS-peptide,
286 either by contamination or lack of solubility. All the calculated peptide amounts along with the
287 protein amount calculated by Top3 are listed in Table 1 together with the CV between the methods.
288 A high level of agreement between the Top3 strategy and most of the SIS-peptides was observed.
289 Only the protein A2RMQ7 produced results which were not in agreement with the Top3 strategy.
290 This protein was, however, only identified in a single replicate and only by 3 peptides. A higher CV
291 was generally observed for low-abundant proteins and proteins identified with a low number of
292 peptides, which is in accordance with previous findings³³. This suggests that increasing the number
293 of peptide candidates also increases the accuracy and reliability of the Top3 quantification strategy.
294 The deduced amount was found to be present in the same range as the SIS-peptides.

295

296 **Absolute quantification of HCPs**

297 In order to compensate for losses during sample preparation, the reference proteins were spiked
298 into the sample before purification. In this case all losses of the reference proteins would reflect
299 losses of HCPs. The quantitative SWATH data was combined with the previously obtained spectral

300 libraries. The general workflow for absolute quantification of the HCPs is shown in Figure 3. 100.8 μg
301 of sample was precipitated with acetone and subjected to MCX purification (see Materials and
302 Methods). This sample was compared to 50 μg of sample, which was purified by the suspension trap
303 strategy. Both samples were spiked with the intact protein ladder before any initial sample
304 preparation. 4 μg of each sample was analyzed (assuming no loss). Using the acetone-MCX strategy,
305 150 proteins could be quantified, while 115 proteins were quantified using the suspension trap
306 strategy. The determined amounts were found to correlate well between the two methods
307 ($r^2=0.936$). However, proteins were generally determined to be in higher abundance from the
308 acetone-MCX data, which may be a result of the precipitation as described above. Figure 4A shows
309 the correlation between the two sample preparation strategies. Table 2 shows the 20 most
310 abundant HCPs along with the calculated amount/ μg of starting material and the coefficient of
311 variation (CV) between the two analyses. An average CV of 5.7 % was found between replicates for
312 the acetone-MCX samples, and 5.9 % for the suspension trap experiments. When comparing the
313 quantities between the two methods, an average CV of 23.1 % was observed.

314 The quantified HCPs were plotted against their theoretical pI and MW, which was calculated using
315 the *Compute pI/MW tool*, available on the ExPASy server³⁴. Figure 5 shows a bubble plot of the
316 quantified HCPs. It is seen that intense HCPs are present at MWs and pIs both below and above that
317 of BP, which is in accordance with the TFF purification strategy, which removes components below a
318 certain size. This information may be used for evaluation of the purification strategy, if certain HCPs
319 can be identified as risk factors.

320 From the quantitative data, the internal dynamic ranges were investigated. It was found that the
321 majority of the overall HCP signal arose from less than 10 proteins. 80 % of the cumulative amount
322 was determined to consist of only 31 proteins for the acetone-MCX data and 28 proteins for the
323 suspension trap data. The remaining proteins may not contribute much to the overall HCP amount,
324 however, these may still induce unwanted immune reactions. It is therefore essential to know the
325 identity and amounts of these proteins. Figure 4B shows the cumulative percentages of the
326 quantified HCPs. The HCP measured amounts are presented in Supporting Table S-3. The presented
327 data suggest that the acetone-MCX strategy is most suitable for HCP-quantitation, as reference
328 proteins showed a better linearity and allowed for quantification of more HCPs than the suspension
329 trapping strategy. The higher yield of HCPs and better linearity of HCPs is believed to be a result of
330 the powerful precipitation conditions along with the high capacity of the MCX cartridges. However, it
331 is possible that suspension trapping may be equally efficient with minor method optimization. This

332 method may be of interest, as it offers protein purification and digestion with minimal costs in
333 chemicals and time.

334

335 **Unspecific cleavage products**

336 Even though trypsin is described as a highly specific protease, non-specific cleavage tends to occur,
337 leading to semi-tryptic and non-specific peptides, which further complicates the process of database
338 searching³⁵. These non-specific cleavages occur at a much lower rate, leading to much lower
339 intensities of these peptides, compared to their fully tryptic counterparts^{36,37}. If the HCPs are
340 subjected to non-specific cleavage it may not affect the results, but if the BP is subjected to non-
341 specific cleavage, these cleavage products may be in the intensity range of the HCPs and be
342 misinterpreted as tryptic HCP peptides. This can lead to severe overestimation of contaminants and
343 false identifications.

344 It was determined that 13 % of the identified HCP peptides were semi-tryptic and 1 % were non-
345 specific. For the BP, 73 % of the identified peptides by count were found to be semi-tryptic and 23 %
346 were found to be non-specific. The relatively low number of fully tryptic BP peptides is most likely a
347 result of the high amount of substrate for unspecific cleavages. The non-specific BP cleavage
348 products were found in such high intensity that they accounted for most of the previously
349 unexplained peaks of the base peak chromatogram. Supporting Figure S-2 shows the elution profiles
350 for tryptic and non-tryptic BP peptides along with the cleavage specificity for BP and the HCPs. No
351 significant chymotryptic activity was observed.

352 The semi-tryptic and non-specific BP peptides were cross-referenced with the fully tryptic HCP
353 peptides identified by MASCOT in order to identify BP-peptides misinterpreted as HCP-peptides, and
354 the peptides were evaluated manually using their elution times and fragment spectra. The non-
355 specific BP peptides *SLLDEGK* and *WDATATELNNALQ* were assigned as *SIEDLGK* (A2RMJ0) and
356 *WDGAENTQVEAVK* (A2RN08), respectively. The peak areas of these two peptides were among the
357 largest observed in the sample, even though both proteins were only identified by a single peptide,
358 which is quite uncommon as the number of identified peptides commonly reflects the amount of a
359 given protein³⁸. This illustrates the importance of identifying non-specific cleavage products,
360 especially when dealing with high-dynamic range samples such as biopharmaceutical proteins. For
361 future analyses, searches should be performed in multiple passes, first excluding all peptide-
362 spectrum matches deriving from BP peptides and then searching the unassigned spectra for HCP
363 information. Aside from reassigning BP-peptides, this will in general increase the number of
364 identifications³⁹.

365 **Assessing purity of the biopharmaceutical protein**

366 Proteins represented by less than 3 peptides were excluded, as these were deemed not to be
367 reliably quantifiable. It was found that proteins represented by either one or two peptides
368 comprised 8 % of the total HCP signal (measured against Top3 signals of HCPs represented with ≥ 3
369 peptides). It was assessed that these proteins did not contribute significantly to the overall purity
370 and could therefore be excluded from the quantitative analysis, but may still act as immunogenic
371 targets, and should therefore still be reported as identified.

372 The purity of the sample was calculated based on the sum of the quantified HCPs. The MW of the
373 HCPs was used to calculate the measure of *total HCP mass / total mass* (which was determined by
374 AAA).

375 Since the proteins may exist as fragments, this could lead to possible overestimation of the HCP
376 amounts. However, by requiring at least 3 peptides from each protein, proteins with broad
377 sequence coverages are selected. The quantitative data showed a total HCP amount of 113 ng
378 HCP/ μg of sample (11.3 %) calculated using the acetone-MCX strategy and 66 ng HCP/ μg of sample
379 (6.6 %) calculated using the suspension trap strategy. The HCP content has previously been
380 determined by ELISA assays by SSI to be between 7.3 and 11.4 %, which is in agreement with the
381 obtained results.

382 **Conclusions**

383 This work presents a reproducible strategy for LC-MS/MS-based absolute quantification of host cell
384 proteins associated with a biopharmaceutical protein. The entire workflow may be performed within
385 a few days. By combining samples from different steps in the purification process, extended spectral
386 libraries were created and used to guide the SWATH-based quantification. When combining several
387 various libraries, additional proteins were quantified in the low-abundant range in the final product.
388 Non-tryptic cleavage products were evaluated in order to achieve higher certainty in the identified
389 HCPs, which was determined to be easily influenced by minor side-cleavage products of the BP. This
390 issue, although rarely addressed, is applicable to all proteomics experiments with high dynamic-
391 range samples.

392 Label-free absolute quantification was performed by spiking with reference proteins and quantifying
393 with Top3. The quantification strategy was verified by SIS-peptides.

394 The results were used to study the identity and quantity of >100 HCPs and determine the internal
395 dynamic range. It was determined that proteins represented by less than 3 peptides did not add

396 significantly to the total intensity, and could therefore be excluded from the quantitative analyses.
397 The HCP content of the sample was thus assessed to be 6.6 to 11.3 % of the total protein amount.
398 The acetone-MCX strategy was found to yield more identifications than the suspension trapping
399 strategy, resulting in a better linearity of the reference proteins. In conclusion, the strategy may be
400 used to study the HCP content of new biopharmaceutical proteins in the pipeline as an orthogonal
401 method to ELISA assays. Due to the speed and depth of information, the present method is
402 particularly suitable for obtaining information of in-process samples and thus guide the purification
403 process.

404 **Associated content**

405 Supporting information available:

- 406 • Purification and quantification of crude stable isotope labeled standard peptides.
- 407 • List of identified HCPs.
- 408 • Calibration plots for the different spiking strategies.
- 409 • List of quantified HCPs.
- 410 • Description of cleavage specificity for the BP and associated HCPs.

411 **Acknowledgements**

412 This work was supported by Innovation Fund Denmark (grant number 34-214-1). The authors thank
413 the employees at Sciex, Warrington, UK for introduction to instrumentation, especially Milla
414 Neffling, Steve Taylor and Nick Morrice. The authors furthermore thank the Villum Center for
415 Bioanalytical Sciences for providing the instrumentation on which preliminary analyses were
416 performed and Ingrid Kromann from SSI for providing the material.

417

418 **Conflict of interests**

419 The authors declare no conflict of interest.

420 **List of abbreviations**

- 421 • BP: Biopharmaceutical protein
- 422 • DDA: Data dependent acquisition
- 423 • DIA: Data independent acquisition
- 424 • DTT: Dithiothreitol
- 425 • ELISA: Enzyme-linked immunosorbent assay
- 426 • HCP: Host cell protein

- 427 • IAA: Iodoacetamide
- 428 • LC-MS: Liquid chromatography-Mass spectrometry
- 429 • MW: Molecular weight
- 430 • P10: Nonpurified harvest media containing the biopharmaceutical protein
- 431 • P15: Biopharmaceutical protein after all purification steps.
- 432 • SIS: Stable isotope standard
- 433 • SWATH: Sequential Window Acquisition of all Theoretical Mass Spectra
- 434 • TEA: Triethylamine
- 435 • TFF: Tangential flow filtration

- 436

437 **References**

- 438 (1) Eaton, L. C. J. *Chromatogr. A* 1995, 705, 105–114.
- 439 (2) Gao, S. X.; Zhang, Y.; Stansberry-Perkins, K.; Buko, A.; Bai, S.; Nguyen, V.; Brader, M. L.
440 *Biotechnol. Bioeng.* 2011, 108, 977–982.
- 441 (3) Hogwood, C. E. M.; Bracewell, D. G.; Smales, C. M. *Curr. Opin. Biotechnol.* 2014, 30, 153–160.
- 442 (4) Wang, X.; Hunter, A. K.; Mozier, N. M. *Biotechnol. Bioeng.* 2009, 103, 446–458.
- 443 (5) Zhu-Shimoni, J.; Yu, C.; Nishihara, J.; Wong, R. M.; Gunawan, F.; Lin, M.; Krawitz, D.; Liu, P.;
444 Sandoval, W.; Vanderlaan, M. *Biotechnol. Bioeng.* 2014, 111, 2367–2379.
- 445 (6) Krawitz, D. C.; Forrest, W.; Moreno, G. T.; Kittleson, J.; Champion, K. M. *Proteomics* 2006, 6,
446 94–110.
- 447 (7) Xu, D.; Mane, S.; Sosic, Z. *Electrophoresis* 2015, 36, 363–370.
- 448 (8) Reisinger, V.; Toll, H.; Mayer, R. E.; Visser, J.; Wolschin, F. *Anal. Biochem.* 2014, 463, 1–6.
- 449 (9) Levy, N. E.; Valente, K. N.; Choe, L. H.; Lee, K. H.; Lenhoff, A. M. *Biotechnol. Bioeng.* 2014, 111,
450 904–912.
- 451 (10) Zougman, A.; Selby, P. J.; Banks, R. E. *Proteomics* 2014, 14, 1006–1000.
- 452 (11) Arike, L.; Valgepea, K.; Peil, L.; Nahku, R.; Adamberg, K.; Vilu, R. J. *Proteomics* 2012, 75, 5437–
453 5448.
- 454 (12) Schwanhäusser, B.; Busse, D.; Li, N.; Dittmar, G.; Schuchhardt, J.; Wolf, J.; Chen, W.; Selbach,
455 M. *Nature* 2011, 473, 337–342.
- 456 (13) Ahrné, E.; Molzahn, L.; Glatter, T.; Schmidt, A. *Proteomics* 2013, 13, 2567–2578.
- 457 (14) Silva, J. C.; Gorenstein, M. V.; Li, G.-Z.; Vissers, J. P. C.; Geromanos, S. J. *Mol. Cell. Proteomics*
458 2005, 5, 144–156.
- 459 (15) Doneanu, C. E.; Anderson, M.; Williams, B. J.; Lauber, M. A.; Chakraborty, A.; Chen, W. *Anal.*
460 *Chem.* 2015, 87, 10283–10291.
- 461 (16) Doneanu, C. E.; Xenopoulos, A.; Fadgen, K.; Murphy, J.; Skilton, S. J.; Prentice, H.; Stapels, M.;
462 Chen, W. *MAbs* 2012, 4, 24–44.
- 463 (17) Gillet, L. C.; Navarro, P.; Tate, S.; Rost, H.; Selevsek, N.; Reiter, L.; Bonner, R.; Aebersold, R.
464 *Mol. Cell. Proteomics* 2012, 11, O111.016717-O111.016717.
- 465 (18) Egertson, J. D.; MacLean, B.; Johnson, R.; Xuan, Y.; MacCoss, M. J. *Nat. Protoc.* 2015, 10, 887–
466 903.
- 467 (19) Gilar, M.; Olivova, P.; Daly, A. E.; Gebler, J. C. *Anal. Chem.* 2005, 77, 6426–6434.
- 468 (20) Batth, T. S.; Francavilla, C.; Olsen, J. V. J. *Proteome Res.* 2014, 13, 6176–6186.
- 469 (21) Højrup, P. *Methods in Molecular Biology*; Springer: New York, 2015
- 470 (22) Waters Corporation. [*CARE AND USE MANUAL*] *Hi 3 STANDARDS*; Milford, MA, USA, 2012.

- 471 (23) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* 1999, 20, 3551–
472 3567.
- 473 (24) Cox, J.; Mann, M. *Nat. Biotechnol.* 2008, 26, 1367–1372.
- 474 (25) MacLean, B.; Tomazela, D. M.; Shulman, N.; Chambers, M.; Finney, G. L.; Frewen, B.; Kern, R.;
475 Tabb, D. L.; Liebler, D. C.; MacCoss, M. J. *Bioinformatics* 2010, 26, 966–968.
- 476 (26) R Core Team. R Core Team (2016): R Foundation for Statistical Computing, Vienna, Austria
477 2016.
- 478 (27) Zi, J.; Zhang, S.; Zhou, R.; Zhou, B.; Xu, S.; Hou, G.; Tan, F.; Wen, B.; Wang, Q.; Lin, L.; Liu, S.
479 *Anal. Chem.* 2014, 86, 7242–7246.
- 480 (28) Schubert, O. T.; Gillet, L. C.; Collins, B. C.; Navarro, P.; Rosenberger, G.; Wolski, W. E.; Lam, H.;
481 Amodei, D.; Mallick, P.; MacLean, B.; Aebersold, R. *Nat. Protoc.* 2015, 10, 426–441.
- 482 (29) Elliott, M. H.; Smith, D. S.; Parker, C. E.; Borchers, C. J. *Mass Spectrom.* 2009, 44, 1637–1660.
- 483 (30) Liu, Y.; Hüttenhain, R.; Surinova, S.; Gillet, L. C. J.; Mouritsen, J.; Brunner, R.; Navarro, P.;
484 Aebersold, R. *Proteomics* 2013, 13, 1247–1256.
- 485 (31) Nakamura, K.; Hirayama-Kurogi, M.; Ito, S.; Kuno, T.; Yoneyama, T.; Obuchi, W.; Terasaki, T.;
486 Ohtsuki, S. *Proteomics* 2016, 16, 2106–2117.
- 487 (32) Percy, A. J.; Yang, J.; Chambers, A. G.; Simon, R.; Hardie, D. B.; Borchers, C. H. *J. Proteome*
488 *Res.* 2014, 13, 3733–3747.
- 489 (33) Anderle, M.; Roy, S.; Lin, H.; Becker, C.; Joho, K. *Bioinformatics* 2004, 20, 3575–3582.
- 490 (34) Wilkins, M. R.; Gasteiger, E.; Bairoch, A.; Sanchez, J.; Williams, K. L.; Appel, R. D.;
491 Hochstrasser, D. F. *Methods Mol. Biol.* 2005, 112, 531–552.
- 492 (35) Bunkenborg, J.; Espadas, G.; Molina, H. J. *Proteome Res.* 2013, 12, 3631–3641.
- 493 (36) Picotti, P.; Aebersold, R.; Domon, B. *Mol. Cell. Proteomics* 2007, 6, 1589–1598.
- 494 (37) Burkhart, J. M.; Schumbrutzki, C.; Wortelkamp, S.; Sickmann, A.; Zahedi, R. P. *J. Proteomics*
495 2012, 75, 1454–1462.
- 496 (38) Ishihama, Y.; Oda, Y.; Tabata, T.; Sato, T.; Nagasu, T.; Rappsilber, J.; Mann, M. *Mol. Cell.*
497 *Proteomics* 2005, 4, 1265–1272.
- 498 (39) Kertesz-Farkas, A.; Keich, U.; Noble, W. S. *J. Proteome Res.* 2015, 14, 3027–3038.

499

500 **Tables**

501

502 Table 1 - Amounts of selected HCP peptides determined by spiking with SIS-peptides. The absolute
503 amounts of the individual proteins as determined by Top3 quantification are also listed. A2RMV9

504 was only identified with 2 peptides in the label-free analyses, but was included in the table for
 505 evaluation purposes.

Protein Accession	Amount in fmol (Top3)	Amount in fmol calculated by SIS peptides [fmol] (CV in %)		
		Peptide 1	Peptide 2	Peptide 3
A2RHZ5	652.24	736.68 (6.08)	2215.51 (54.51)	643.83 (0.65)
A2RKI9	926.33	937.84 (0.62)	1920.28 (34.92)	
P0A3J1	125.92	96.04 (13.46)	41.71 (50.24)	145.92 (7.36)
A2RMQ7	22.62	132.51 (70.84)	304.80 (86.18)	93.87 (61.16)
A2RMV9	*26.21	67.08 (43.81)	43.19 (24.47)	

506

507 Table 2 – The 20 most abundant HCPs associated with the BP (Sorted according to abundance
 508 measured by the acetone-MCX strategy). The amounts were determined by Top3 quantification
 509 using a set of reference proteins spiked in before digestion for both the acetone-MCX (AM) and
 510 suspension trapping (ST) strategy.

Protein name	Accession	fmol / μ g - AM	fmol/ μ g - ST	CV [%]
Uncharacterized protein	A2RM18	492	367	15
HU-like DNA-binding protein	A2RIK5	323	181	28
Non-heme iron-binding ferritin	A2RNH7	258	158	24
Secreted 45 kDa protein	P22865	250	90	47
50S ribosomal protein L7/L12	A2RKI9	181	93	32
Probable N-acetylmuramidase	A2RHZ5	133	110	10
Oligopeptide-binding protein oppA	A2RJ53	128	88	18
Basic membrane protein A	A2RK47	112	37	50
Serine protease Do-like HtrA	A2RNT9	111	61	29
Manganese ABC transporter substrate binding protein	A2RKC2	79	14	70
Mid-cell-anchored protein Z	A2RJC3	78	48	24
Phosphonate ABC transporter, phosphonate-binding protein	A2RI25	78	67	7
Cell surface antigen I/II	A2RL18	67	35	31
Putative secreted protein	A2RI00	64	41	22
Putative secreted protein	A2RJA8	53	N/A	N/A
50S ribosomal protein L11	A2RNF4	43	30	18
50S ribosomal protein L18	A2RNN7	42	27	22
Uncharacterized protein	A2RN73	42	28	21

<i>Chaperone protein DnaK</i>	P0A3J1	39	41	3
<i>Foldase protein PrsA</i>	A2RME7	37	9	61

511

512

513 **Figure legends**

514 Figure 1 – (A) Venn diagram showing the number of identified host cell proteins. (B) Venn diagram
 515 showing quantifiable host cell proteins when using different DDA files as spectral libraries.
 516 Identifications were performed using MASCOT and quantifiable proteins/peptides were selected
 517 based on the criteria described in the text.

518 Figure 2 – Signal/fmol ratios of reference proteins spiked into the BP sample. The ratios are shown
 519 for myoglobin (blue), carbonic anhydrase (red), alcohol dehydrogenase (green) and serotransferrin
 520 (purple). The ratio used for quantification is based on all proteins and is shown in cyan.

521 Figure 3 – Flowchart describing the general workflow applied for absolute quantification of HCPs.

522 Figure 4 - A) The absolute HCP amounts quantified in the acetone-MCX strategy plotted against the
 523 absolute HCP amounts quantified in the suspension trap strategy. B) The cumulative molar amounts
 524 of the 100 most abundant HCPs plotted against the number of proteins. The quantities are based on
 525 the Top3 signal of quantified HCPs, normalized against the internal reference protein calibration
 526 curve. The figure shows amounts determined both for the acetone-MCX (black) and suspension
 527 trapping (white) strategy.

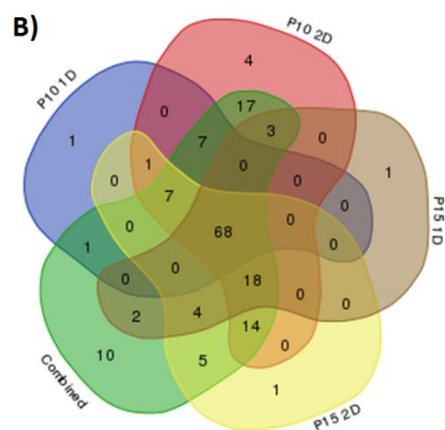
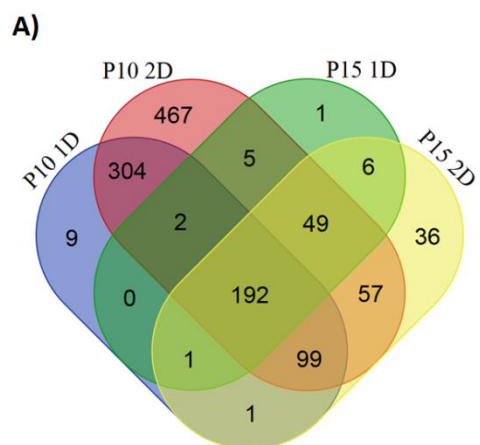
528 Figure 5 – Bubble plots showing the pI (x-axis), MW (y-axis) and amounts (bubble size) of the HCPs
 529 using the acetone-MCX strategy (A, blue) and the suspension trapping strategy (B, red). The BP is
 530 represented by the purple dot. The amount of BP does not correspond to the size of the bubble

531

532

533 **Figures**

534 Figure 1



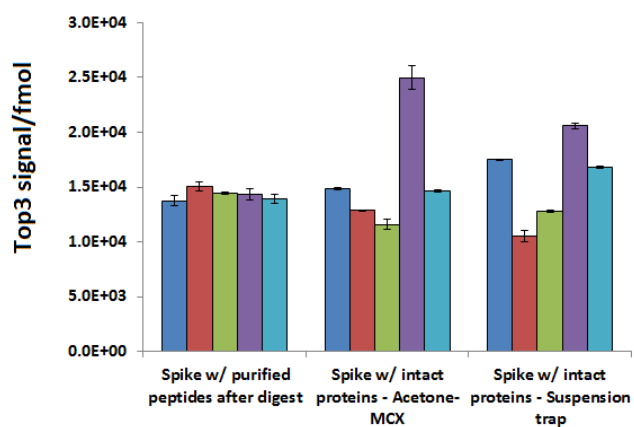
535

536

537 Figure 2

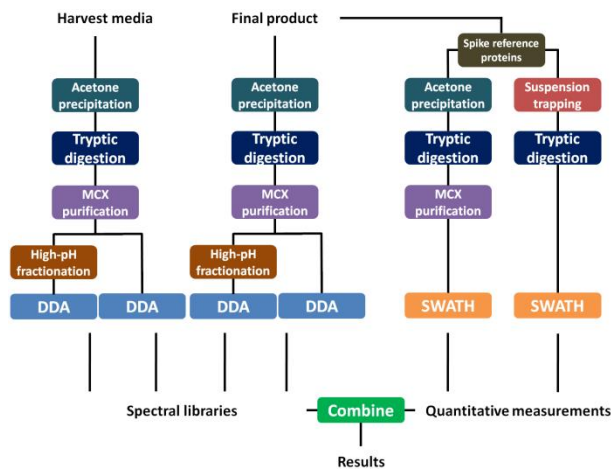
538

539



540

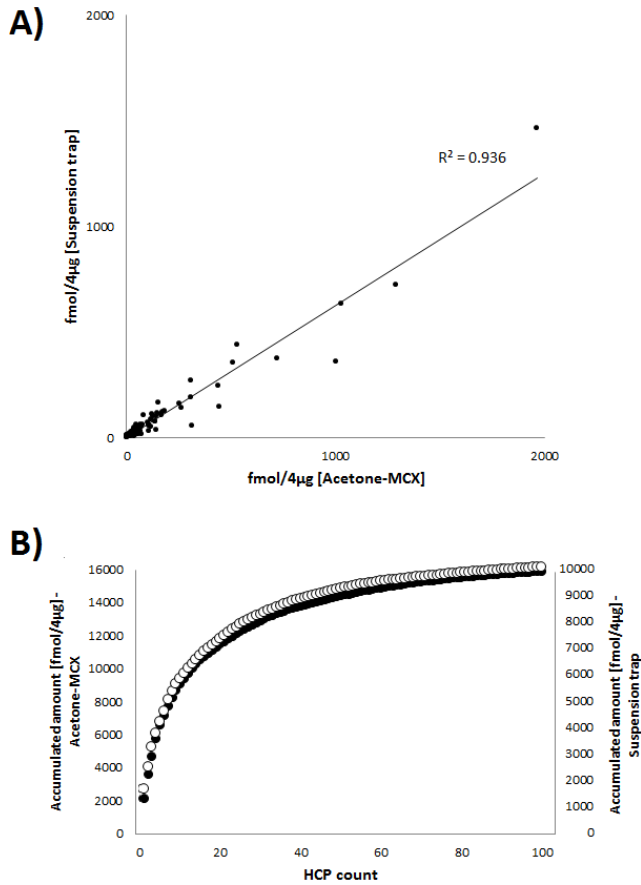
541 Figure 3



542

543

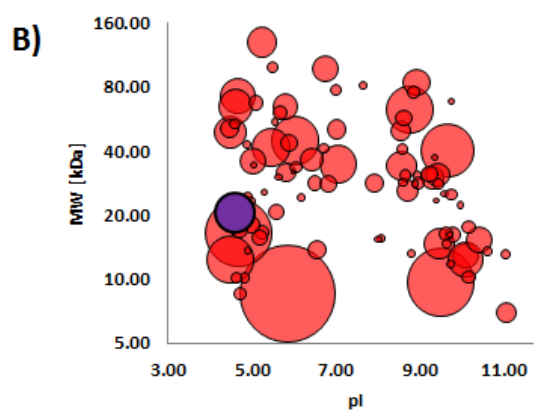
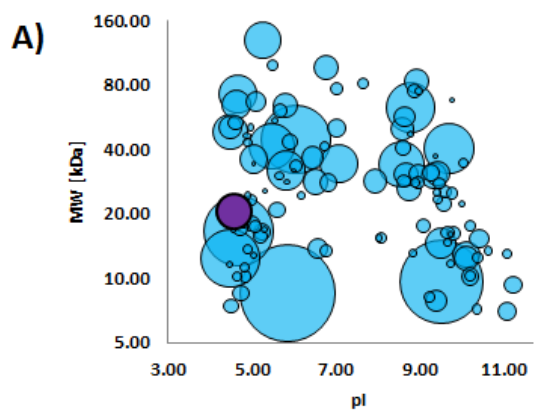
544 Figure 4



545

546

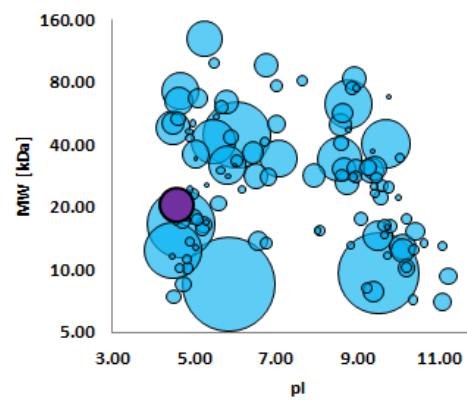
547 Figure 5



548

549

550 For TOC graphic only



551

552

553