SUPPORTING INFORMATION

Advancing top-down analysis of the human proteome using a benchtop quadrupole-Orbitrap mass spectrometer

Luca Fornelli, Kenneth R. Durbin, Ryan T. Fellers, Bryan P. Early, Joseph B. Greer, Richard D. LeDuc, Philip D. Compton, and Neil L. Kelleher*

Table of contents.

belonging to the charge state envelope of a certain protein should have a similar resolution factor (within experimental error). Panels A and B are based on 22824 proteoforms from the theoretical proteome of *Pseudomonas aeruginosa*. (A), a linear correlation is apparent between corresponding monoisotopic and average masses of theoretical proteoforms (ranging from ~4 to 570 kDa) (B), the difference between average and monoisotopic mass, expressed in ppm (by dividing the "delta mass" by the monoisotopic mass), remains approximately constant throughout the entire mass range, with the average value being 629.5 ppm. (C), experimental results based on short transient (8 ms) measurement of 5 proteins: ubiquitin (8.5 kDa), superoxide dismutase (16 kDa), myoglobin (17 kDa), carbonic anhydrase (29 kDa) and enolase (47 kDa). The y-axis shows the mass difference between monoisotopic (theoretical) and average mass (experimental, measured using the peak apex), expressed in ppm as in panel B; on the x-axis we report the 'resolution factor' for the peaks of several charge states of these five proteins, calculated as the spectral resolution expressed as full width at half maximum (FWHM) multiplied by the corresponding charge state. The plot shows that when the molecular weight of a protein approaches ~40 kDa, the peak apex corresponds closely to the position of the average mass, with very low dependence on the considered charge state. For enolase, the average difference between monoisotopic and average mass is 625.3 ppm (calculation based on 33 different charge states). Conversely, for smaller proteins the calculation of the average mass based on the peak apex results in the underestimation of the average mass itself, due to the more pronounced asymmetry of the underlying distribution of isotopomers. Note that a 'resolution factor' of 1 indicates that peaks are isotopically resolved (at 50% peak height), whereas larger values describe non-isotopically resolved species. These larger resolution factor values can be interpreted as the fold-increase in resolution needed to obtain isotopically resolved

charge states at FWHM. (D), zoomed-in view on the proteoform PFR20440 displayed in Figure 4. The left panel includes charge states 45-48+, while a detail of charge state 45+ is shown in the inset on the right.

Supplementary Figure S-3. An example of identifying a low-abundance proteoform via SIM marching. (A), the grey column indicates the applied isolation window (width = 3 *m/z* units). The precursor corresponding to the isolated, fragmented and identified proteoform is highlighted in red in this MS^1 spectrum obtained during an automated SIM march. Metrics for the protein identification and characterization of the proteoform, PFR13634, appear in the inset (upper right of the top panel). (B), graphical fragment map based on a HCD tandem mass spectrum.

Supplementary Figure S-4. Graphical output of STRING gene ontology analysis based on accession numbers of larger proteins identified by experiments run with "medium/high" data acquisition logic. The plot was obtained via the STRING graphic tool available on-line. The three main protein clusters, organized according to the KEGG pathway catalog, are indicated by red circles. The Glycolysis/Gluconeogenesis protein group (pathway ID: 00010) is represented here by 5 gene products; an additional four genes in this cluster are from the related Fructose and Mannose Metabolism pathway (ID: 00051). The gene products identified as members of the Hippo Signaling Pathway can be considered here as part of a larger group of DNA-bound proteins.

Supplementary Figure S-5. Distribution of all 1952 proteoforms identified at a 1% proteoform-level FDR. This Venn diagram includes also the 80 proteoforms that did not map to any of the 393 Accessions identified at 1% FDR cutoff at the protein level. These 80 proteoforms, corresponding to 4.1% of the total, were mostly identified

in data-dependent experiments and their average C-score was 52.6 (versus an average of 89.7 overall).

Supplementary Figure S-6. Correlation between *q*-values and C-scores for the proteoforms identified by AUTOPILOT high/high experiments at 1% proteoform-level FDR. (A), considering all 990 proteoforms identified in the 15 RAW files generated by AUTOPILOT, there is no clear linearity between the two scores. (B), limiting linear regression to the top 167 proteoforms with –LOG(*q*-value) >30, the coefficient of determination R^2 does not improve significantly relative to that determined for all proteoforms. Note that the analogous graphs for datasets obtain using the two other two data acquisition modes show very similar trends (data not shown).

Supplementary Table S-1. List of identifications including entries (with assigned UniProt accession numbers) and proteoforms (with related Proteoform Record numbers) for data-dependent high/high experiments. (XLSX)

Supplementary Table S-2. List of identifications including entries (with assigned UniProt accession numbers) and proteoforms (with related Proteoform Record numbers) for AUTOPILOT high/high experiments. (XLSX)

Supplementary Table S-3. List of identifications including entries (with assigned UniProt accession numbers) and proteoforms (with related Proteoform Record numbers) for data-dependent medium/high experiments. (XLSX)

Figure S-1

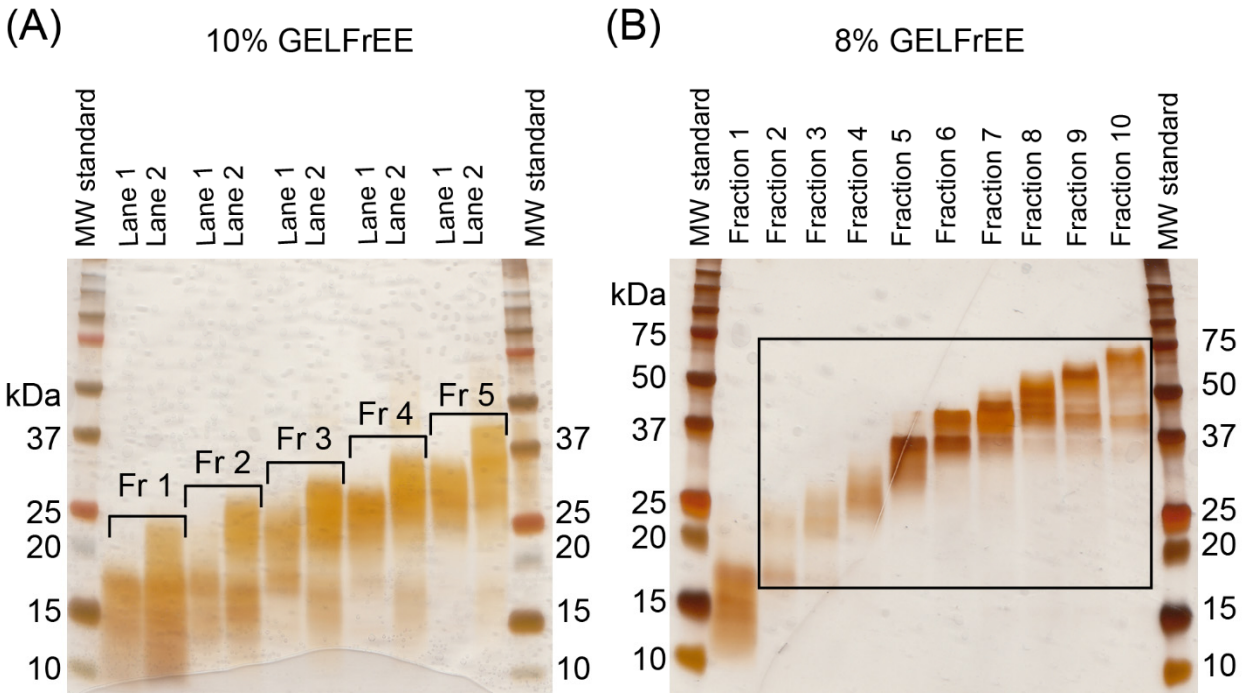
(A)
10% GELFrEE
MW standard
Lane 1
Lane 2
Lane 1
Lane 2
Lane 1
Lane 2
Lane 1
Lane 2
Lane 1
Lane 2
MW standard
kDa
37
25
20
15
10
Fr 1
Fr 2
Fr 3
Fr 4
Fr 5
37
25
20
15
10
(B)
8% GELFrEE
MW standard
Fraction 1
Fraction 2
Fraction 3
Fraction 4
Fraction 5
Fraction 6
Fraction 7
Fraction 8
Fraction 9
Fraction 10
MW standard
kDa
75
50
37
25
20
15
10
75
50
37
25
20
15
10

Figure S-2

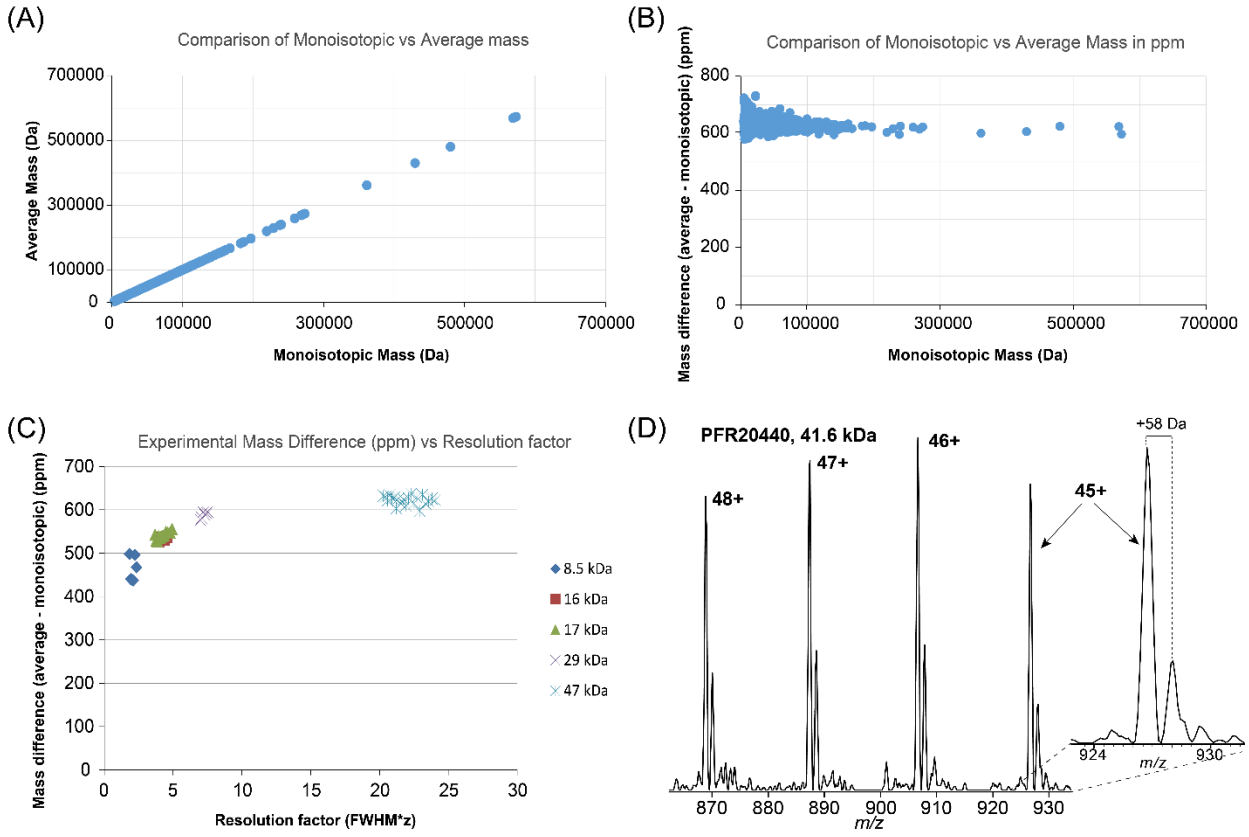
(A)
Comparison of Monoisotopic vs Average mass
Average Mass (Da)
Monoisotopic Mass (Da)
(B)
Comparison of Monoisotopic vs Average Mass in ppm
Mass difference (average - monoisotopic) (ppm)
Monoisotopic Mass (Da)
(C)
Experimental Mass Difference (ppm) vs Resolution factor
Mass difference (average - monoisotopic) (ppm)
Resolution factor (FWHM*z)
8.5 kDa
16 kDa
17 kDa
29 kDa
47 kDa
(D)
PFR20440, 41.6 kDa
48+
47+
46+
45+
+58 Da
m/z

Figure S-3

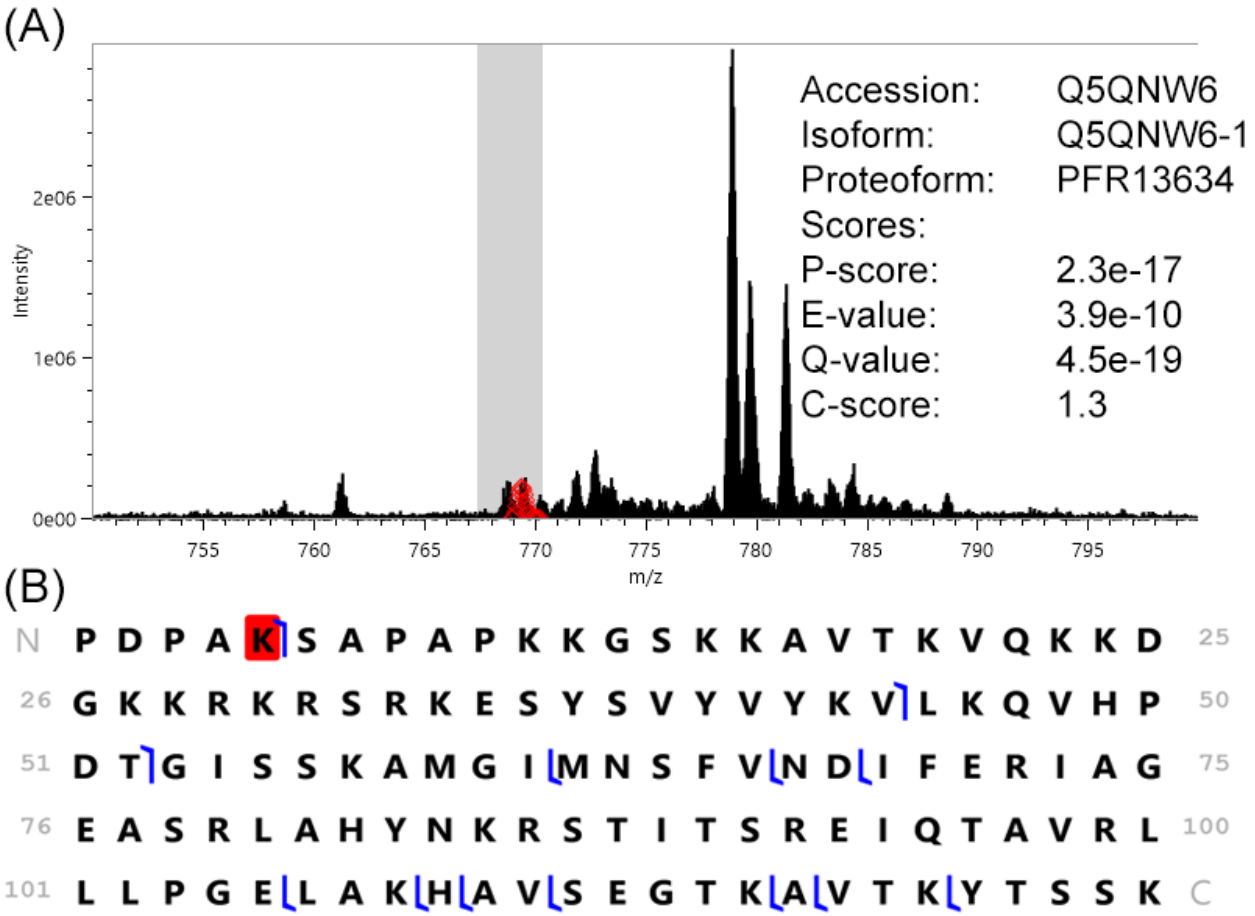
(A)
Accession: Q5QNW6
Isoform: Q5QNW6-1
Proteoform: PFR13634
Scores:
P-score: 2.3e-17
E-value: 3.9e-10
Q-value: 4.5e-19
C-score: 1.3
Intensity
2e06
1e06
0e00
755
760
765
770
775
780
785
790
795
m/z
(B)
N P D P A K S A P A P K K G S K K A V T K V Q K K D 25
26 G K K R K R S R K E S Y S V Y V Y K V L K Q V H P 50
51 D T G I S S K A M G I M N S F V N D I F E R I A G 75
76 E A S R L A H Y N K R S T I T S R E I Q T A V R L 100
101 L L P G E L A K H A V S E G T K A V T K Y T S S K C

Figure S-4

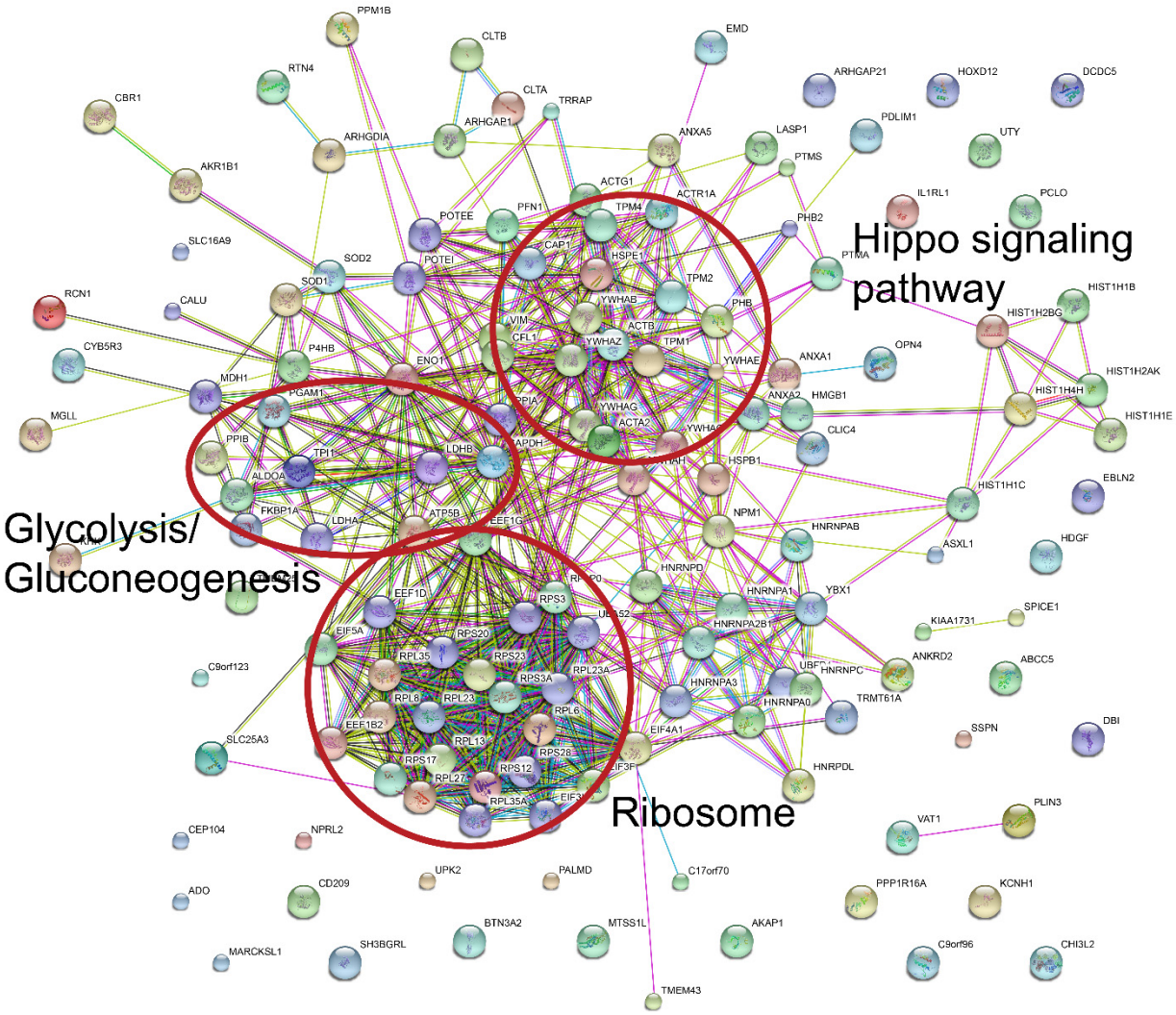
Hippo signaling pathway
Glycolysis/ Gluconeogenesis
Ribosome
PPM1B
CLTB
EMD
RTN4
CLTA
TRRAP
ARHGAP21
HOXD12
DCDC5
CBR1
ARHGAP1
ANXA5
LASP1
PDLIM1
UTY
ARHGDIA
PTMS
AKR1B1
ACTG1
ACTR1A
IL1RL1
PCLO
POTEE
PFN1
TPM4
PHB2
SLC16A9
CAP1
HSPE1
PTMA
POTEI
SOD2
TPM2
SOD1
YWHAB
HIST1H1B
RCN1
CALU
PHB
HIST1H2BG
VIM
ACTB
CFL1
YWHAZ
TPM1
CYB5R3
P4HB
YWHAE
ANXA1
OPN4
ENO1
HIST1H2AK
MDH1
PGAM1
ANXA2
HMGB1
HIST1H4H
PPIA
YWHAG
MGLL
ACTA2
YWHAQ
HIST1H1E
CLIC4
PPIB
TPI1
LDHB
APDH
HSPB1
ALDOA
HIST1H1C
EBLN2
FKBP1A
LDHA
ATP5B
EEF1G
NPM1
HNRNPAB
ASXL1
HDGF
HNRNPD
EEF1D
RPS3
HNRNPA1
YBX1
SPICE1
EIF5A
RPS20
HNRNPA2B1
KIAA1731
ANKRD2
ABCC5
C9orf123
RPL35
RPS23
RPL23A
UBF1
HNRNPC
RPS3A
HNRNPA3
TRMT61A
RPL8
RPL23
HNRNPA0
RPL6
SSPN
DBI
EEF1B2
EIF4A1
SLC25A3
RPL13
RPS28
RPS17
RPS12
EIF3F
HNRNPDL
RPL27
RPL35A
EIF3
PLIN3
VAT1
CEP104
NPRL2
UPK2
PALMD
C17orf70
ADO
CD209
PPP1R16A
KCNH1
BTN3A2
MTSS1L
AKAP1
MARCKSL1
SH3BGRL
C9orf96
CHI3L2
TMEM43

Figure S-5

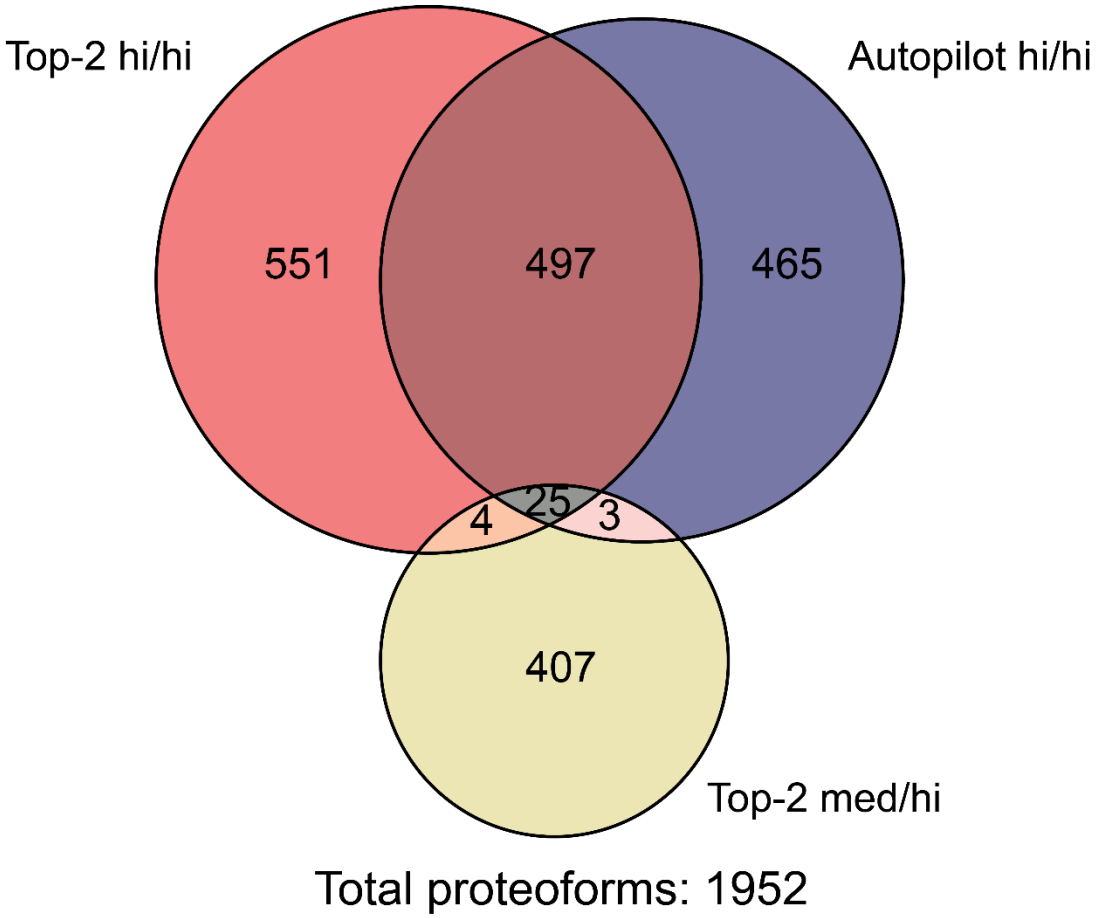
Top-2 hi/hi
Autopilot hi/hi
551
497
465
4
25
3
407
Top-2 med/hi
Total proteoforms: 1952

Figure S-6