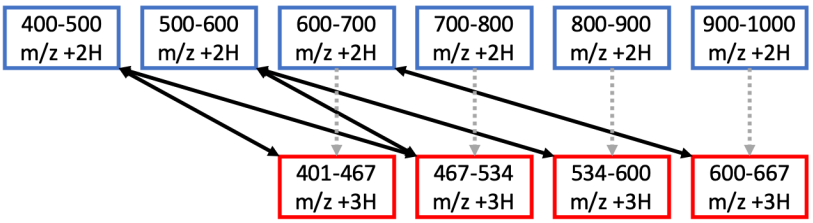
Supplementary Note 2: Frequently Asked Questions

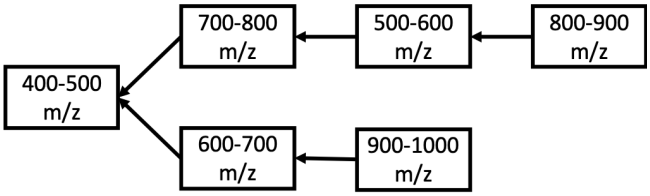
1) GAS-PHASE FRACTIONATION AND CHROMATOGRAM LIBRARIES

Question: Is it possible to align between GPF-DIA runs to generate a chromatogram library?

Retention time alignment between gas phase fractions is complicated by the fact that the fractions typically do not contain any of the same peptide detections. Aside from peptides that "sneak" into a different window due to splitting the isotopic envelope, the only exception is potentially different charge states of the same peptide (e.g. the 900-1000 m/z fraction will contain some +2H peptides with +3H precursors between 600-667 m/z):



where blue +2H peptides could be aligned to red +3H peptides through black connections. Theoretically it is possible to use these equivalences to construct a multiple alignment strategy:



However, this requires as much as 5x alignment inferences (from the 800-900 m/z fraction to the 900-1000 m/z fraction), which may severely compromise the quality of the retention time library and requires the detection of multiple charge states for a large number of peptides, particularly in the 700-800 m/z gas phase fraction.

Question: What happens to samples with severe retention time deviations, such as those acquired at the very beginning of a new experiment?

Retention time shifts happen whenever a new column is used, or when switching between sample types (e.g. acquiring *E. coli* lysate after human cell lysates). We find that with the same column, retention times may shift but retention time ordering between peptides typically does not. EncyclopeDIA was designed to take advantage of this phenomenon. EncyclopeDIA differs from most other DIA peptide-centric tools in that it always performs a first-pass peptide search without using any knowledge of retention time and peptide detections are never explicitly filtered based on retention time windows. After the first-pass search, all of the high confidence detections become the equivalent of typically hundreds of iRT anchors in simple mixtures or tens of thousands of anchors in complex samples. Because retention time ordering stays constant with the same column, it is very easy to precisely retention time align peptides between

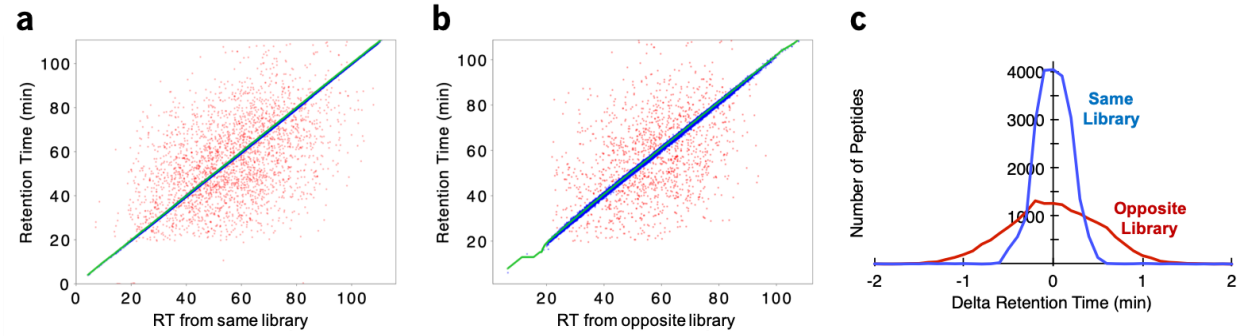
two runs on the same column with EncyclopeDIA, including samples that are acquired early in the column's lifetime. However, this process assumes that there are no retention time deviations between GPF-DIA library-building injections, which cannot easily be aligned together (see earlier). Consequently, we find it better to run the GPF-DIA injections in the middle of the acquisition queue after the column retention properties have stabilized.

Question: Why is it not recommended to "reuse" chromatogram libraries on other columns or instruments?

While it is possible to share chromatogram libraries, we typically do not recommend it because small variations between columns can change retention time characteristics. For example, we repeated an experiment using the same yeast sample on the same Thermo QE-HF instrument and Waters NanoAcquity HPLC, but with two different columns. The number of detections dropped by approximately 25% when single-injection DIA acquisitions were searched with libraries from the opposite experiment:

	Experiment 1	Experiment 2
Total Library Size	25861	29817
vs Same Library	21704	23787
vs Opposite Library	16439	17589

One reason for this drop is due to minor reordering in retention times between peptides on different columns. Below is a scatter plot of Experiment 2 peptides when searching (**a**) the same library, or (**b**) the opposite library from Experiment 1. The distribution of detected peptides (**c**) indicates that when searching the correct library 90% of peptides fit within +/-30 seconds, while when searching the incorrect library, 90% of peptides fit within +/-66 seconds:



Question: How does GPF-DIA compare to strong cation exchange or high-pH reverse-phase fractionated DDA?

Using data presented in another manuscript in press (Searle 2019 bioRxiv, ProteomeXchange PXD: PXD017705), this figure shows an Euler plot of the overlap of unique yeast peptide sequences detected between 6x GPF-DIA fractions searched with Pecan, and 8x injections of high-pH reverse phase fractions (6x fractions, wash, flow through) searched with Comet using the Trans Proteomic Pipeline:

Both datasets were filtered to a 1% peptide-level FDR. To avoid conflict with pre-published data, we have opted to cite the preprint of this manuscript directly, rather than present this data here, because it demonstrates a similar conclusion.

2) DIA METHODS DESIGN

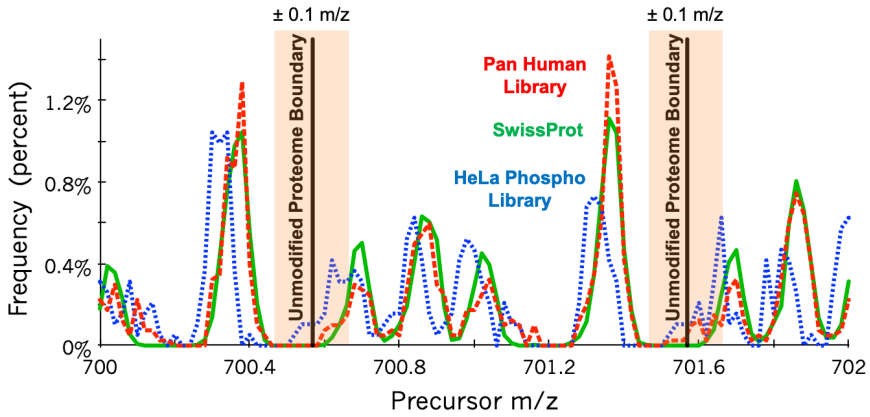
Question: How do you calculate "forbidden zones" for PTM enriched samples?
The equation for forbidden zones is based on Egertson et al. 2012:

$$ceil(\frac{nominal\ mass}{optimal\ m/z\ increment}) * optimal\ m/z\ increment + optimal\ m/z\ constant$$

By inverting the formula shown in the manuscript, we propose the following equation for estimating the optimal m/z constant for PTM enriched samples:

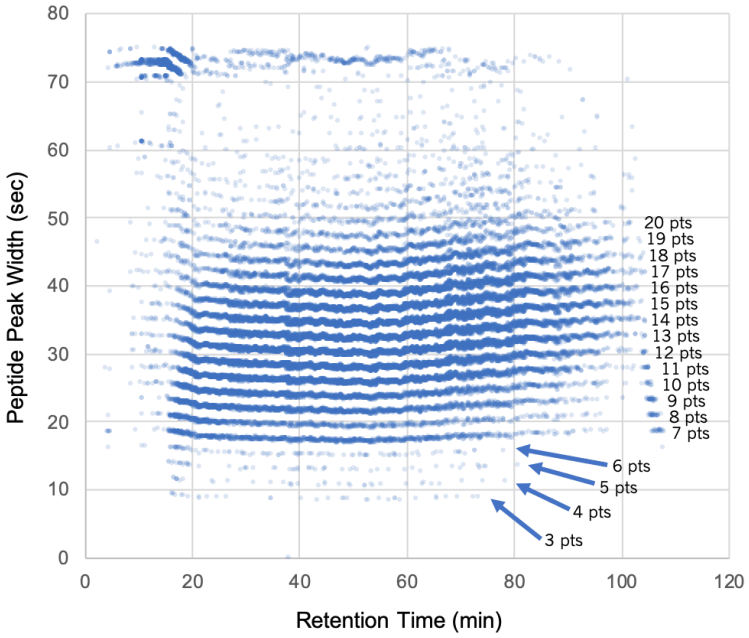
$$optimal\ m/z\ constant = 0.25 + \left(\frac{PTM\ mass}{optimal\ m/z\ increment} - ceil(\frac{PTM\ mass}{optimal\ m/z\ increment})\right)$$

Because the addition of phosphate does not consistently change peptide charge state, the adjustment for phospho-enriched samples can be calculated as 97.976896 / 1.00045475 - ceil(97.976896) = -0.067638741. Relative to the original 0.25 constant, 0.25 - 0.07 = 0.18. This negative shift agrees with precursor mass frequencies of the Pan-Human library and the HeLa phospho library from Figure 1 compared to the distribution of tryptic peptides predicted from SwissProt:



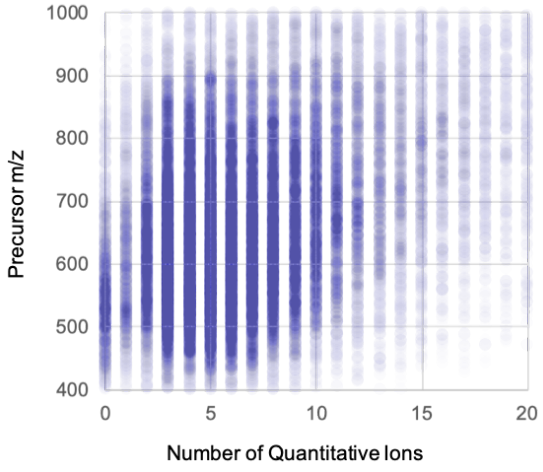
While this calculation appears to work for phosphopeptides, which exhibit an unusual mass defect, we strongly recommend using the precursor mass distribution from a large library to validate the shift.

Question: What percentage of peptides have more than 10 chromatographic points?
In the quantitative experiment from Figure 9 showing technical replicates, we find that 91% of peptides have 10 or more points across the peak and that 99.8% of peptides have 6 or more points:



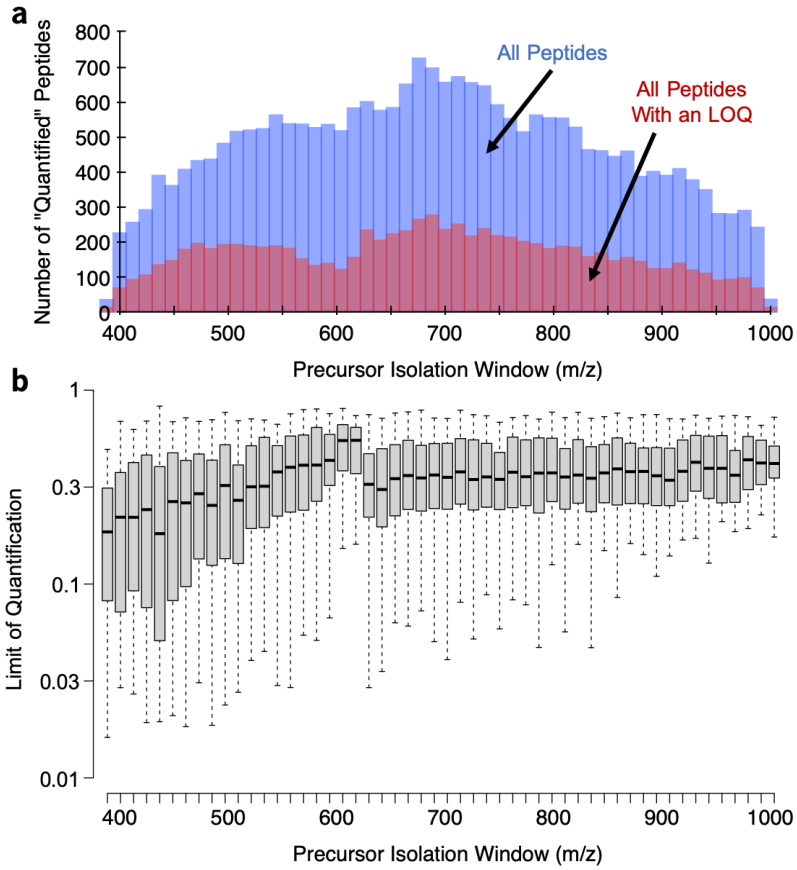
The distribution of these points show interesting behavior across retention time. Many very early eluting peptides that do not stick to the column tend to smear with wide peak widths, however most early eluting peaks have narrower peak widths than those later in the gradient. The exception are very late eluting peaks (>100 mins), which get progressively more narrow.

Question: How does the number of interference-free fragment ions change across m/z?
Peptides with low precursor m/zs tend to have a lower number of interference-free fragments, likely due to shorter peptides producing fewer fragment ions. However, the majority of peptides with fewer than 3 transition-free fragment ions occur in the congested 500-700 m/z range:



Question: How does quantitative accuracy change across m/z?

The true accuracy of quantification is hard to estimate; in the experiments presented in this manuscript we can only use variability metrics like CV as a proxy for accuracy. To answer this question we turn to our recent paper (Pino et al 2020), where we demonstrate a method for estimating the limit of quantification (LOQ) for every peptide in a proteome. Using the yeast dataset from this paper, we considered both (**a**) the number of peptides with an LOQ and (**b**) the actual LOQ values across precursor m/z windows:



We find that while the percentage of peptides with an LOQ is relatively constant across windows (median=36%, stdev=4%), peptides with low *m/z* precursor masses tend to have lower (more sensitive) LOQs. Because low m/z ions generally tend to have higher intensity, this matches the expectation that LOQ tracks with measured intensity.