

# FLiPPR: A Processor for Limited Proteolysis (LiP) Mass Spectrometry Data Sets Built on FragPipe

Edgar Manriquez-Sandoval, Joy Brewer, Gabriela Lule, Samanta Lopez, and Stephen D. Fried\*



Cite This: *J. Proteome Res.* 2024, 23, 2332–2342



Read Online

ACCESS |

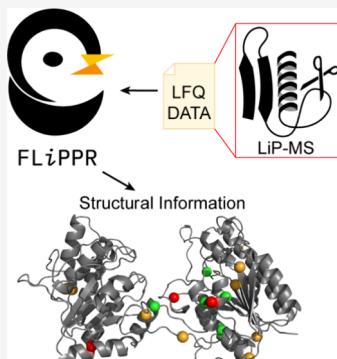
Metrics & More

Article Recommendations

Supporting Information

**ABSTRACT:** Here, we present FLiPPR, or FragPipe LiP (limited proteolysis) Processor, a tool that facilitates the analysis of data from limited proteolysis mass spectrometry (LiP-MS) experiments following primary search and quantification in FragPipe. LiP-MS has emerged as a method that can provide proteome-wide information on protein structure and has been applied to a range of biological and biophysical questions. Although LiP-MS can be carried out with standard laboratory reagents and mass spectrometers, analyzing the data can be slow and poses unique challenges compared to typical quantitative proteomics workflows. To address this, we leverage FragPipe and then process its output in FLiPPR. FLiPPR formalizes a specific data imputation heuristic that carefully uses missing data in LiP-MS experiments to report on the most significant structural changes. Moreover, FLiPPR introduces a data merging scheme and a protein-centric multiple hypothesis correction scheme, enabling processed LiP-MS data sets to be more robust and less redundant. These improvements strengthen statistical trends when previously published data are reanalyzed with the FragPipe/FLiPPR workflow. We hope that FLiPPR will lower the barrier for more users to adopt LiP-MS, standardize statistical procedures for LiP-MS data analysis, and systematize output to facilitate eventual larger-scale integration of LiP-MS data.

**KEYWORDS:** *limited proteolysis, structural proteomics, computational tools, data analysis workflow, bioinformatics*



## INTRODUCTION

Structural proteomics is an expanding subfield within the space of proteomics that aims to explore protein structure, dynamics, and stability in a global, unbiased manner. The field is defined by several emerging methods that convert structural information into mass and abundance, enabling mass spectrometry data to measure structural information of many distinct proteins in a complex mixture. Four leading structural proteomic methods include hydrogen–deuterium exchange (HDX),<sup>1–3</sup> methionine oxidation methods (e.g., SPROX),<sup>4,5</sup> fast photochemical oxidation of proteins (FPOP),<sup>6,7</sup> and limited proteolysis mass spectrometry (LiP-MS).<sup>8–10</sup> In these methods, regions within proteins that are solvent-accessible are labeled, respectively, by deuteration at backbone amides, by oxidation at methionine from H<sub>2</sub>O<sub>2</sub>, by oxidation at any amino acid from HO radicals, or by cleavage from a broad-spectrum protease.<sup>11</sup> Cross-linking mass spectrometry (XL-MS) is another method that can provide rich structural information in the form of residue–residue contacts (which can serve as distance restraints),<sup>12–15</sup> though it differs from the other methods in that it requires sequencing low-abundance cross-linked peptides, creating a unique set of technical challenges.<sup>16,17</sup>

Among the labeling-based methods, LiP-MS has emerged as a popular structural approach because it has a few key advantages (e.g., residue-level resolution, proteome-wide coverage) without some drawbacks that affect other methods

(e.g., need for specialized purpose-built instruments, need to identify rare low-abundance species). In its modern form (devised by Feng et al.<sup>9</sup>), the experimentalist subjects a complex mixture of proteins to a pulse of proteolysis with a broad-spectrum protease (typically Proteinase K, also referred to as PK) under native conditions, causing solvent-accessible or unstructured portions within proteins to get cleaved (Figure 1A).<sup>9</sup> Afterward, the sample is subjected to in-solution trypsin digest under denaturing conditions, which produces a mixture of tryptic and half-tryptic peptides that are sequenced by LC-MS/MS. The nontryptic cut-site of each half-tryptic peptide reveals a residue that was solvent-accessible in the parental protein. LiP-MS is a very accessible structural proteomic method because, in most samples, half-tryptic peptides are numerous, abundant, and do not require specialized search settings to identify. Data-dependent acquisition (DDA)<sup>8,18</sup> and data-independent acquisition (DIA)<sup>10,19</sup> workflows have both been implemented. So far, it has been applied to probe a range of biological and biophysical questions at the proteome scale, such as metabolic rewiring in response to nutrients,<sup>19</sup> aging in

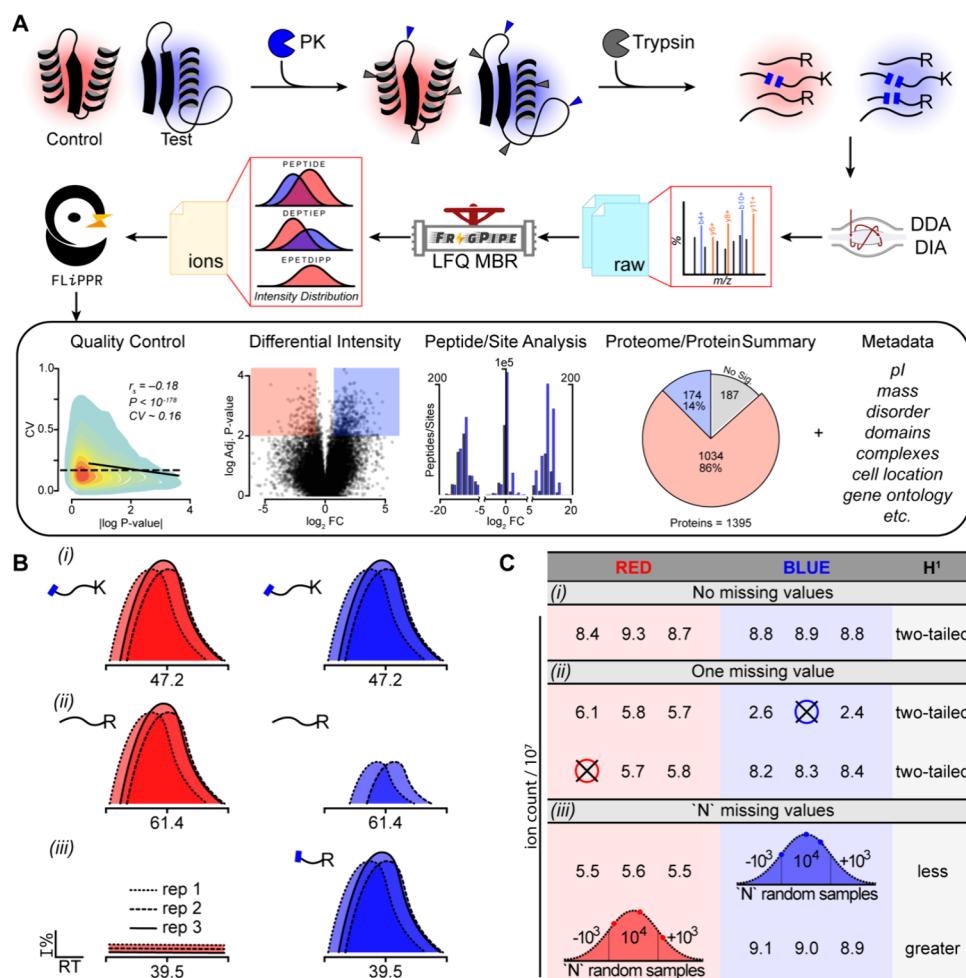
Received: December 14, 2023

Revised: February 23, 2024

Accepted: May 16, 2024

Published: May 24, 2024





**Figure 1.** Summary of LiP-MS workflow and processing in FLiPPR. (A) The top row is a simple schematic of LiP-MS sample preparation, featuring a control condition (red) and a test condition (blue) simultaneously subjected to limited proteolysis (LiP) by Proteinase K (PK) and then complete trypsinolysis. The second row represents data handling in the FLiPPR pipeline: raw mass spectra are searched and quantified in FragPipe, and the ions file is processed by FLiPPR, which produces a range of outputs featured in the third row. (B and C) How FLiPPR treats missing data at the ion level. Case *i*: if quantifications are available for an ion in all replicates of test and control, then averages are calculated, and the *P*-value is assessed with a two-tailed *t* test. Case *ii*: if one quantification is missing in either the test or the control, the missing value is dropped, and the *P*-value is assessed with a two-tailed *t* test comparing *n* to *n* – 1 values. Case *iii*: if all the values are missing in either the control or the test (in the example shown in B, it is the control) and all the values are present in the other condition, then the ion is considered “all-or-nothing” (AON). The *n* missing values are filled by Gaussian imputation, then averages are calculated as in case *i*; *P*-values are calculated with a one-tailed test. All other permutations of missing data result in an ion being disqualified.

yeast,<sup>20</sup> aging in rodents,<sup>21</sup> thermostability,<sup>22</sup> and protein folding,<sup>18,23</sup> among others.

A LiP-MS study is usually designed as a quantitative experiment in which two (or more) closely related samples are generated and subjected to the same workflow; half-tryptic peptides that are present with significantly different abundances between two sample types represent locations associated with a structural change (strictly speaking, a change in proteolytic susceptibility) between the respective conditions. Most studies to date have employed area-under-the-curve label-free quantification (LFQ) to assess these abundance differences, with a few studies applying SILAC quantification or isobaric mass tag methods instead.<sup>5,8,24,25</sup> Some key advantages of LFQ-based quantification are its high dynamic range and remarkable ability to independently assess missing features in distinct conditions (or replicates),<sup>26,27</sup> which has outsized importance in LiP-MS and can provide very insightful information if used judiciously.

The experimental details to prepare samples for LiP-MS have been carefully developed and advanced by Picotti and co-workers,<sup>10</sup> and we have used them with minor modifications in our adaptation of the method to study protein refolding (see the *Experimental Section*). On the data analysis side, however, we have found that several developments from our lab have been valuable. While the relevance of these improvements was first realized in the context of our ongoing studies applying LiP-MS to protein folding, we believe they would be useful for LiP-MS studies in general, and here we report a computational tool called FLiPPR (FragPipe LiP-MS Processor) to formalize our data analysis workflow and facilitate its adoption by the structural proteomics community.

First, FLiPPR seamlessly accepts output from FragPipe (Figure 1A), an open software suite developed by the Nesvizhskii lab that has fast, powerful, and state-of-the-art algorithms for spectral search (MSFragger)<sup>28,29</sup> and label-free quantification with FDR-controlled match between runs (IonQuant) to analyze both DDA- and DIA-based studies.<sup>28,30</sup>

Most DDA-based LiP-MS studies to date (including ours) have employed Proteome Discoverer<sup>31</sup> to perform search and LFQ (via the Minora feature detector node), though we have found that this workflow has been rate-limiting, particularly for studies that cross-compare an extensive set of conditions (e.g., LFQs with  $>9$  raw files) which run very slowly. Indeed, our initial motivation for developing FLiPPR was to facilitate a shift to FragPipe to analyze LiP-MS data.

Second, FLiPPR implements a unique treatment of missing data tailored to LiP-MS. Sometimes, when a protein undergoes a large structural transition, core regions that were completely inaccessible to Proteinase K can become proteolytically susceptible (as shown in the case of Figure 1A); this results in a situation in which a half-tryptic peptide will be detected only in samples containing the structurally altered conformation but will be absent in samples containing the native conformation, resulting in missing data. A process is necessary to distinguish this scenario (where the missing data are informative (e.g., Figure 1B, case *iii*)) from scenarios where missing data should instead prevent a feature from being included in an analysis.

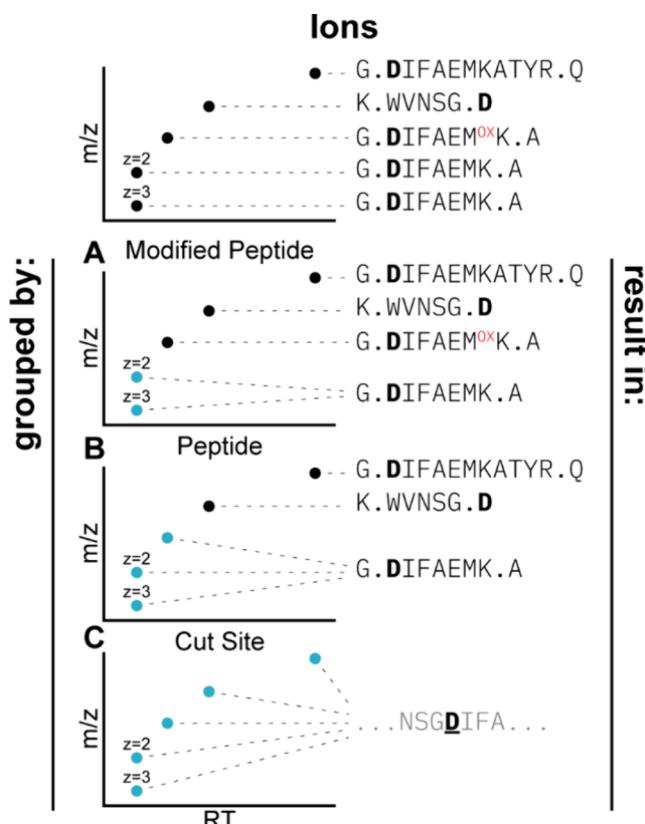
Third, LiP-MS is fundamentally a “peptide-centric” technique, and so data compression, propagation of error, and data disagreement must be considered from the level of ions to modified peptides to peptides, and as we argue here, ultimately to cut-sites (cf. Figure 2). Since most quantitative proteomics studies focus on the protein level and have an additional “protective” layer to buffer against noise by averaging across peptides, data processing for LiP-MS raises unique concerns not addressed by commercial software packages.

Finally, we have found that correlating the outcome of LiP-MS experiments with biophysical and structural features (such as percent disorder, isoelectric point, and domain structure) has helped illuminate key trends, and so far we have accumulated these metadata on a case-by-case basis. However, since many of these biophysical traits can now be calculated (or predicted) from sequence alone, valuable metadata can be generated in an automated manner, which we anticipate will be constructive for scaling up the interpretation of structural proteomics across species and clades. In summary, FLiPPR creates a pipeline that formalizes some of the subtle data analysis considerations we have devised in our experience of working with LiP-MS data. Hence, we expect FLiPPR will contribute to standardizing the analysis workflow for this emerging and exciting proteomic technique by building off a popular, free platform.

## COMPUTATIONAL SECTION

### Inputs

The inputs into FLiPPR are the output directories generated by LFQs in FragPipe using IonQuant (Figure 1A). Suggested modifications to FragPipe’s default parameters specific for LiP-type experiments are provided in Table S1 and are discussed further in the Supporting Information (cf. Figure S1). The LFQ experiment should minimally consist of six separate raw spectra files comprising three replicates of at least two conditions, a test and a control condition each subjected to the limited proteolysis sample preparation protocol (see the Experimental Section). Although FLiPPR is compatible with any number of replicates per condition ( $N \geq 1$ ), lower statistical significance is expected from analyses with fewer



**Figure 2.** Three merging schemes for ion data. The top row represents five distinct ions that can be grouped together in three ways. (A) Ions that differ in charge state for the same modified peptide can be merged. (B) Modified peptides that differ in a specific modification (such as methionine oxidation) but correspond to the same base peptide can be merged. (C) Peptides that correspond to the same Proteinase K cut-site can be merged. LiP-MS analyses typically merge ions to the peptide level; FLiPPR introduces further merging to the cut-site level.

than three replicates per condition. Experimental designs can sometimes involve a single test condition or multiple test conditions that are all compared to a common control condition. FLiPPR can process both situations; in either case, LFQs in FragPipe are calculated with all raw spectra files and submitted with the naming convention shown in Figure S2.

LiP-MS studies are frequently conducted as two parallel experiments: the “LiP experiment”, which consists of samples subjected to pulse-proteolysis with Proteinase K and then fully digested with trypsin, as well as a parallel “TrP experiment” in which the samples are treated identically except pulse-proteolysis is skipped and they are only digested with trypsin. The TrP experiment is analyzed at the protein level and addresses whether protein abundances have changed between the test and control conditions.<sup>9,10</sup> This TrP experiment is important because the abundance differences between the half-tryptic peptides in the LiP experiment convolve changes in proteolytic susceptibility with changes in overall protein abundance, an effect that must be controlled. If there is a significant change in protein abundance between a test and control condition, ion fold changes are normalized by the protein fold changes. Figure S2 illustrates how FragPipe output from the TrP normalization experiment is incorporated into the FLiPPR pipeline.

## Missing Data Imputation and Differential Intensity Analysis

FragPipe provides raw integrated ion counts for each confidently sequenced precursor ion for all spectra files, applying an FDR-based validation criteria to assign ion counts to precursor ions whose peaks were not identified in the parental spectra file but could be confidently “matched between runs” across other spectra files. At this level, FLiPPR applies a particular heuristic to managing missing data. The heuristic is applied to the set of  $2N$  ion counts (where  $N$  is the number of replicates per condition that were conducted) corresponding to each pairwise comparison between a test condition and the control (Figure 1C). If all  $2N$  ion counts are available, a simple ratio of averages is tabulated, and a  $P$ -value is calculated using a two-tailed  $t$  test with Welch’s correction for unequal population variances (Figure 1C, case *i*). If one ion count is missing, a ratio is still calculated, but rather than treating the missing value as a zero (as commonly employed), it is dropped, and the ratio and  $P$ -value are calculated using one fewer replicate for either the test or control condition (Figure 1C, case *ii*). This treats the single missing value as missing at random (MAR) and avoids artificially decreasing the calculated mean.<sup>32,33</sup> The maximum number of allowable missing values for an ion to be further considered can be modified by the user (default is one) and may be recommended in studies with greater than three replicates per condition.

A special case arises if *all* the ion counts are missing for all the replicates of one condition (either test or control; Figure 1C, case *iii*). This can be interpreted as arising from a scenario where a portion of the protein is entirely inaccessible to Proteinase K in the control condition and then becomes accessible in the test condition (or vice versa). In our experience with LiP-MS data, these situations carry among the richest information since they report on large structural changes.<sup>18,23</sup> However, because arguing for an effect based on missing data is fraught, a safeguard is employed requiring that the ion missing in *all* replicates of one condition must *also* be observed in *all* replicates of the other condition, for the former to be considered missing not at random (MNAR). We refer to these as “all-or-nothing” ions (Figure 1B, case *iii*). Any other combination of missing data (either because there are two missing ion counts or because one condition is missing all data, and the other condition is missing even one ion count) results in the ion being discarded and not used for quantification for this test vs control comparison. For all-or-nothing ions, a ratio is calculated as a ratio of averages after Gaussian imputation for the three missing values (selected from a normal distribution with a standard deviation of  $10^3$  and a mean of  $10^4$ ; the mean is an approximation for the limit of detection of a feature in a high-resolution MS1 scan on a Q Exactive instrument, and the standard deviation reflects a typical coefficient of variation for assigned features, cf. Figure S1D). The parameters of the Gaussian imputation can be changed by users based on their instrument’s sensitivity. The  $P$ -value is calculated by a one-tailed Welch’s  $t$  test. At the end of this step, a (ratio,  $P$ -value) pair is assigned to each “valid” ion for each pairwise test/control comparison. Ions with too many missing data are discarded.

## Merging Data from Ions to Cut-Sites

Most proteomics experiments seek to measure quantitative differences in protein abundance across conditions, but LiP-MS seeks to measure quantitative differences in proteolytic

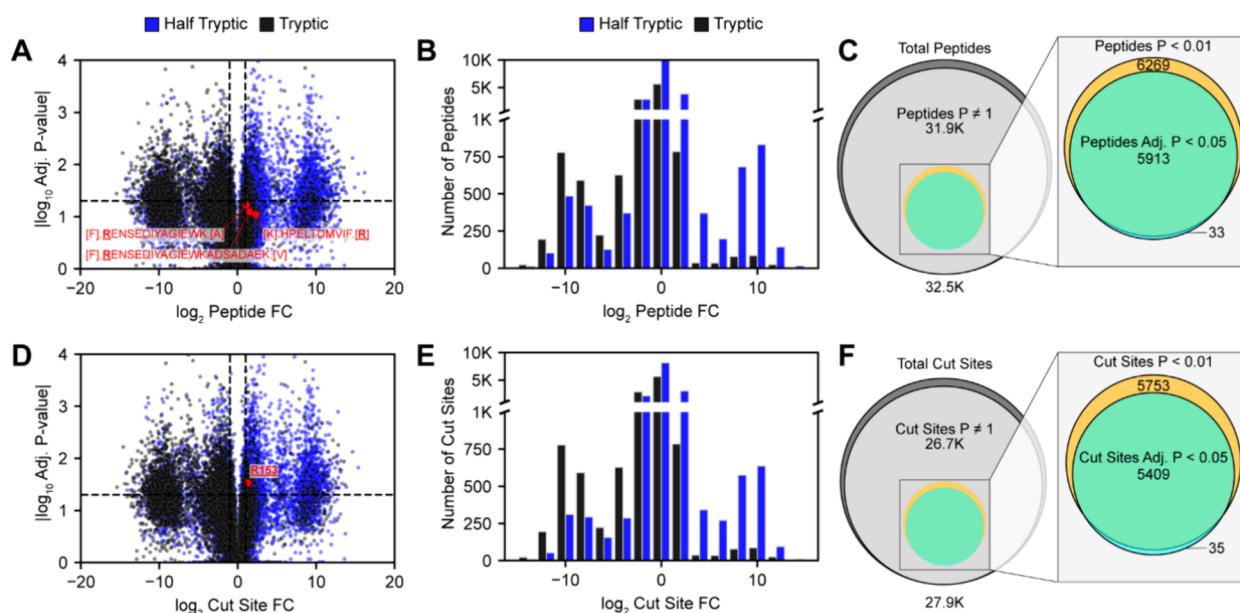
susceptibility at a particular cut-site. Consequently, there are several distinct ions that can be combined and averaged from the raw ion count level to the cut-site level (Figure 2). These correspond to all the ions that map to a given modified peptide (e.g., charge state 2+ or 3+), all the modified peptides that map to a given peptide (e.g., oxidized-methionine or not), and all the peptides that map to the same cut-site (e.g., the peptides [G].D<sup>104</sup>IFAEMKATYR<sup>114</sup>.[Q] and [G].D<sup>104</sup>IFAEMK<sup>110</sup>.[A] both report on the activity of Proteinase K between G103 and D104 because they differ in regard to whether the subsequent trypsin digest missed the cleavage at K110). An alternative scenario where multiple peptides can encode the same cut-site arises when both half-tryptic peptides created by Proteinase K are sequenced (e.g., [G].D<sup>104</sup>IFAEMK<sup>110</sup>.[A] and [K].W<sup>99</sup>VNSG<sup>103</sup>.[D]).

For all the ions that can inform on the susceptibility at a given cut-site, the (ratio,  $P$ -value) pairs for those ions are collectively considered. If they all agree in direction (i.e., the signs of the  $t$  test statistics are all the same), then the “overall” ratio for the cut-site is calculated by taking the median of the ratios of all ions that map to it, and the  $P$ -values are combined with Fisher’s method to provide an updated (ratio,  $P$ -value) for the cut-site. If there are two independent ions and they disagree (e.g., the ion is more abundant in the test condition in the 2+ charge state but more abundant in the control condition in the 3+ charge state), then a median is still calculated, but the  $P$ -value is set to 1, implying there is no confidence as to whether this cut-site was more susceptible in the test or control condition. These cut-sites are discounted from the tally of the total valid cut-sites. If there are three ions, then a “majority rules” heuristic is applied: the disagreeing ion is disregarded, and the (ratio,  $P$ -value)s are only combined for the agreeing ions. In practice, it is relatively rare for more than three ions to be mapped to the same cut-site, but where this occurs, if a majority (or all) of the ions agree in direction, they are combined. If there is a tie, the  $P$ -value is set to 1.

In experimental designs with multiple test conditions, this compression is carried out separately for each test-to-control comparison. As shown in Figure 2, FLiPPR implements merging at all three levels (e.g., merge all ions together that map to the same modified peptide, peptide, or cut-site). The lowest level of merging (to a modified peptide) could be useful for studies focusing on PTMs and their effect on protein structure. The middle level of merging (to a peptide) is the one that historically we have used,<sup>18,23</sup> as well as others.<sup>8–10,20–22</sup> The highest level of merging (to cut-site), to the best of our knowledge, is novel to this analysis workflow.

## Metadata Integration

When provided with the same fasta file used as the search database during the LFQ analysis, FLiPPR can supplement the protein summary data with an array of metadata using open-source packages available in Python. Mass, molecular weight, length, and pI are obtained using Biopython.<sup>34</sup> Protein disorder is predicted using Metapredict<sup>35</sup> with the default disorder score thresholds. Optionally, protein domain information can be supplemented from DomainMapper,<sup>36</sup> though this requires that a user provide a DomainMapper output. Moreover, when provided with DomainMapper outputs, FLiPPR can perform advanced structural analyses by mapping cut-sites to regions within domains or linkers.



**Figure 3.** Analysis of LiP-MS at the peptide or cut-site level. (A) Representative volcano plot in which each dot represents a peptide quantified in the experiment. Three specific peptides from *E. coli* isocitrate dehydrogenase are highlighted in red. Data are from the 1 min time point from refolding assays on *E. coli* extracts and reflect the abundance ratio of refolded/native. Blue corresponds to half-tryptic peptides, and black to tryptic peptides. (B) Histogram of peptide abundance ratios. Enrichment of half-tryptic all-or-nothings in the refolded form is frequently encountered after global refolding or other test conditions that result in perturbed structures. (C) Accounting of the peptides. Dark gray = all peptides with sufficient data to quantify; light gray = all “valid” peptides (e.g., discounting those with inconsistent ions); mustard = significant peptides based on effect-size ( $>2$ -fold) and  $P$ -value ( $<0.01$ , or  $<0.016$  if effect-size  $> 64$ ); teal = significant peptides based on effect-size ( $>2$ -fold) and BH-adjusted  $P$ -value ( $<0.05$ ). (D–F) Same as panels A–C, except at the cut-site level rather than the peptide level. The cut-site at R153 in isocitrate dehydrogenase that results from the merger of the three peptides shown in A is highlighted in red in panel D.

### Output and Interpretation

FLiPPR produces five key output files for each test vs control comparison: an ions file, a modified peptides file, a peptides file, a cut-sites file, and a protein summary file.

Merged abundance ratios for each modified peptide, peptide, and cut-site (as well as normalized abundance ratios) are provided along with merged  $P$ -values and FDR corrected  $P$ -values by the Benjamini–Hochberg (BH) procedure for multiple hypothesis testing.<sup>37</sup> BH correction is applied on a per-protein basis, meaning that the set of  $P$ -values for all quantified ion counts for a given protein are subjected to FDR correction, and this process is iterated separately (and independently) for each protein. The logic for applying multiple hypothesis correction in this way is the following: our null hypothesis is that a protein is *not* structurally perturbed by the treatment in the test condition relative to the control. Each quantified ion provides an opportunity to reject this hypothesis. However, the more ions that are quantified for a given protein, the higher the likelihood that one of them will show a significant effect due to chance. Hence, BH correction ensures that proteins are not artificially easier to call structurally altered simply by having higher coverage.

The protein summary file provides a high-level view of which proteins were structurally altered by the treatment and which ones were not. To do this, it counts the total number of cut-sites (or peptides) that were quantified and how many were significant according to a  $P$ -value and effect-size cutoff. We generally use 2-fold as an effect-size cutoff ( $|\log_2(\text{fold-change})| > 1$ ) and 0.01 as a  $P$ -value cutoff (or alternatively, 0.05 for BH-adjusted  $P$ -values). Historically, we granted large effect sizes ( $|\log_2(\text{fold-change})| > 6$ ) a relaxed  $P$ -value cutoff (of 0.016) and for consistency maintain this practice for the

non-BH-adjusted  $P$ -values, though this relaxation is not necessary for BH-adjusted  $P$ -values. The user can decide whether to make the assessment at the level of peptides or cut-sites and whether to use adjusted or normal  $P$ -values. Our current recommendation is to use adjusted  $P$ -values at the cut-site level. To “call” a protein structurally altered, we typically require that two or more cut-sites (or peptides) be significant.

### Installing FLiPPR

Readers interested in using FLiPPR should clone the repository from GitHub at <https://github.com/FriedLabJHU/FragPipe-Limited-Proteolysis-Processor>. FLiPPR is built in Python and all its dependencies can be installed through Python distributions and package managers.

## RESULTS

### Merging Peptides into Cut-Sites Improves Statistical Confidence

Figure 3A shows a representative peptide-level volcano plot from one of our LiP-MS experiments refolding the proteome of *E. coli*. Each point represents a confidently identified and quantified peptide. We typically find that half-tryptic peptides are more abundant in refolded samples compared to native (Figure 3B), a characteristic that is qualitatively consistent with the notion that misfolded proteins are less well-packed and are therefore more susceptible to PK than natively folded proteins. In general, we find this feature any time the test condition is one that “perturbs” proteins and have seen it in a number of ongoing studies focusing on other variables. Following the fate of the ca. 32,500 peptides quantified in this experiment (Figure 3C), we find that a relative minority (602) are discounted due to the ions that merge into it showing inconsistent signals.

FLiPPR keeps these “invalidated” peptides in the output for completeness but sets their *P*-value to 1. A minority of these peptides are deemed significant (mustard circle) by the cutoffs required to call it structurally altered, and these are further culled to a smaller set by FDR correction (teal circle).

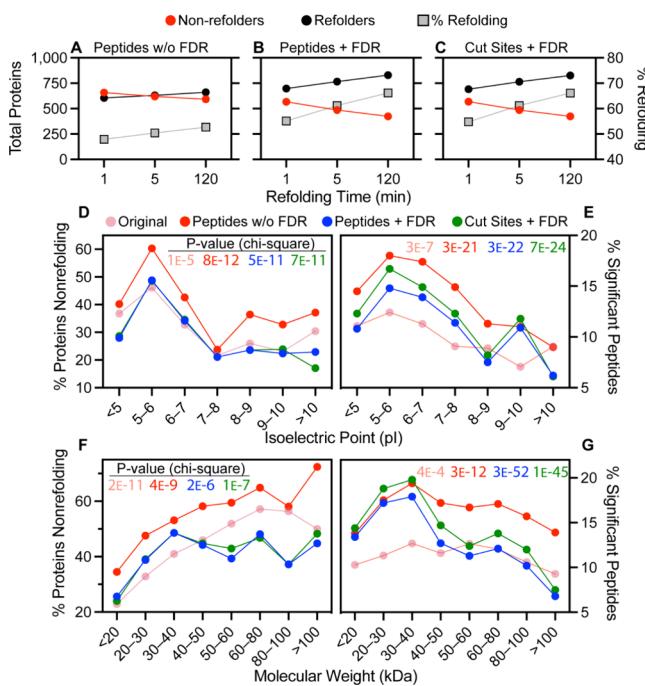
FLiPPR introduces the idea of merging together peptides that map to the same PK cut-site. Following this process, we redraw the volcano plot and relative abundance histogram (Figures 3D and 3E), where the points represent cut-sites instead of peptides. One example, highlighted in red, shows a scenario in which three separate peptides, none of which were statistically significant on their own, admitted a significant cut-site once the ions were merged by Fisher’s method. In this LiP-MS experiment, there are 27,900 cut-sites (instead of 32,500 peptides), but 5,753 cut-sites are assessed as significant (Figure 3F; instead of 6,269 peptides). This slight reduction arises primarily from “removing” duplicates, but it is noteworthy that only 8% of significant peptides are lost even though 14% of the peptides were merged. This difference occurs because some peptides cross the threshold to significance when their data are merged. Hence, we assess that using cut-sites instead of peptides results in a data set with fewer duplications and more “unique” significant sites detected.

### Protein-Level Trends in LiP-MS Data Sets Become Sharper and More Confident with FLiPPR Analysis

Our original work on refolding the *E. coli* proteome highlighted the prevalence of nonrefoldability among soluble proteins<sup>18</sup> and estimated that after 1 min of refolding, 56% of *E. coli* proteins could return to native-like structures, a figure that rises to 67% after providing 2 h to refold. To et al. also called attention to the fact that differences in protein coverage are a source of bias: proteins with more identified peptides are more likely to be labeled as “structurally altered” because there are more opportunities for a significant effect to be detected.

We sought to compare how these overall outcomes of the experiment are affected by some of the differences in the analysis implemented in FLiPPR (Figures 4A–4C). If we analyze the data in a manner analogously to the original work (Figure 4A), we find that refoldability levels go down (48% at 1 min and 53% at 2 h). This difference is in part due to coverage bias, arising because FragPipe’s search (with MSFragger) produces more identifications than ProteomeDiscoverer v2.3 (with Sequest) does: namely, 31,900 and 31,800 peptides at 1 min and 2 h, respectively, versus 28,700 and 28,200 in ProteomeDiscoverer. As mentioned, proteins with more identifications are “easier” to label as structurally altered, which can be corrected by adjusting for multiple hypothesis testing. Using adjusted *P*-values (Figure 4B), the refolding propensities become closer to originally reported values (55% at 1 min and 66% at 2 h), and these overall trends are not appreciably changed if cut-sites are used instead of peptides (Figure 4C).

Global protein refolding assays also demonstrated that refoldability possessed clear correlations with other biophysical and biochemical variables. For instance, from all the studies we have conducted to date, a general theme has emerged that proteins at the extrema of the pI range (the most acidic and the most basic) refold more often than those whose pI values are between 5–6 (mildly acidic proteins). Impressively, the pI trend become much sharper when the exact same data were analyzed in FLiPPR, taking advantage of the two key improvements we have implemented. Based on the 5 min



**Figure 4.** Reanalysis of refoldability data with FLiPPR sharpens trends. (A–C) Plots show the number of proteins (left *y*-axis) labeled refoldable (black) and nonrefoldable (red) and the fractional refoldability (right *y*-axis, gray boxes) as a function of refolding time, depending on whether calls are made based on peptides without Benjamini–Hochberg FDR correction (A), with FDR correction (B), or with FDR correction and after merging peptides to cut-sites (C). Data correspond to *E. coli* refolding assays from ref 18 (PRIDE: PXD025926). (D) Fraction of proteins nonrefolding after 5 min as a function of protein isoelectric point (pI), based on the original analysis (pink) or using FLiPPR using one of three schemes. Inset shows that the *P*-value from the chi-square test against the null hypothesis pI does not explain differences in refoldability. (E) Fraction of peptides (or cut-sites for green symbols) that are assessed as significantly different after 5 min of refolding, as a function of the pI of the protein from which they came, as based on the original analysis (pink) or using FLiPPR using one of three schemes. Inset shows that the *P*-value from the chi-square test against the null hypothesis pI does not explain differences in refoldability. (F) Same as D except proteins are divided by molecular weight. (G) Same as E except peptides are divided by the molecular weight of the protein from which they came.

time point, the original analysis showed that proteins pI values are between 5–6 did not refold 46% of the time, a fraction that decreases for the proteins in the <5 (37%, 0.80-fold) or >10 (30%, 0.65-fold) pI ranges. Using FLiPPR (Figure 4D, red trace), the nature of acidic and basic proteins to refold better is more striking. As before, the peak of nonrefoldability (60%) occurs at pI 5–6, but this drops off more dramatically for proteins with pI values <5 (40%, a 0.66-fold) or >10 (37%, a 0.62-fold). The slopes become even steeper after FDR correction has been incorporated (Figure 4D, blue trace). Now, pI < 5 (pI > 10) proteins have nonrefoldability rates that are 0.57-fold (0.47-fold) that of the maximum. If we accept as a ground-truth that acidic and basic proteins are more refoldable than those with pI 5–6 (as we have seen in all our studies to date),<sup>18,23</sup> then this implies that analyzing LiP-MS data with FLiPPR provides a view that coheres much more closely to physical reality.

Another metric that shows the FLiPPR-analyzed data set has more discriminating power is the chi-square test on the null hypothesis that pI does *not* explain differences in protein refoldability. In the original publication, the *P*-value was placed at  $1.3 \times 10^{-5}$ ; with FLiPPR it becomes closer to  $10^{-11}$  (depending on whether peptides or cut-sites are used). We emphasize here that this is purely from reanalyzing the exact same raw mass spectra.

One of the signs that the pI-refoldability trend is robust is that it is apparent at both the protein level (i.e., assessing proteins as refoldable or not after grouping peptides/cut-sites by protein) and at the individual peptide/cut-site level (i.e., not grouping these by protein and simply calculating the percent that are significant for each category). At the peptide/cut-site level, we find that FLiPPR produces much sharper trends than the original analysis (Figure 4E), and the trends become even more apparent as we introduce FDR correction and peptide to cut-site merging. The strengthening of the correlation is apparent in the chi-square tests as well.

The original work studying refoldability of *E. coli* proteins documented an apparent trend whereby proteins of greater molecular weight were more likely not to refold (Figure 4F, pink trace). However, it was noted that this trend could be the result of coverage bias. In general, the peptides assigned to larger proteins were not more likely to be significant (Figure 4G, pink trace), though larger proteins do generally get more peptides mapped to them, thereby making it more likely that one of them would be significant. Reexamining this trend is therefore a stringent test for whether our FDR correction can “catch” this problem. Indeed, using FLiPPR with FDR corrections (blue and green traces in Figure 4F), we find that there is some correlation between molecular weight and refoldability up to 40 kDa, but afterward, the trend is flat. This is very well recapitulated at the peptide/cut-site level (Figure 4G), in which peptides are indeed more likely to be significant in 30–40 kDa proteins than in proteins  $<20$  kDa, but then the trend reverts. This finding makes sense because 30–40 kDa is the size of the largest single domains and our previous findings have found that larger domains typically refold poorly; proteins larger than this typically contain multiple domains. Hence, we conclude that the more reliable and confident quantification offered by FLiPPR improves downstream analysis of LiP-MS data, and in the case of the present reanalysis of our original study, firms up one finding (pI) and offers a reinterpretation of another (molecular weight).

### FLiPPR Enables Rapid Analysis of Data from Non-Model Species, Facilitating Cross-Species Comparisons

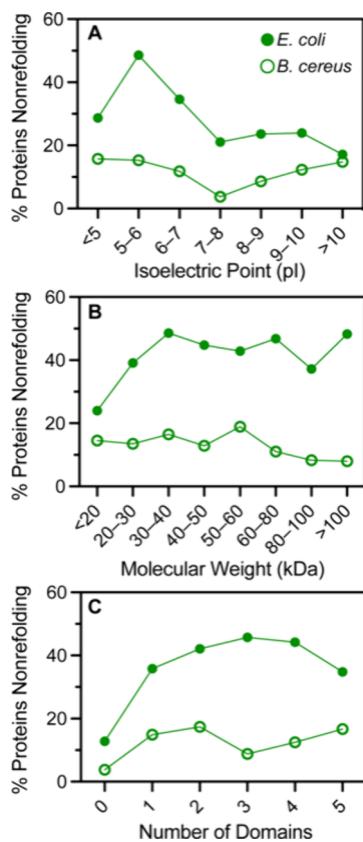
Work in our laboratory has focused on using LiP-MS as a tool to interrogate protein folding on the proteome scale,<sup>18,23</sup> with a focus on addressing how “refoldable” proteins are; namely, if they are globally unfolded in 6 M guanidinium chloride and compelled to refold by dilution, how many proteins succeed in this challenge and can assume a conformation identical to their native ones that were never unfolded? We have found this can be assessed in a high-throughput manner through LiP-MS, in which the proteolysis pattern of a native extract and one that is subjected to an unfolding/refolding cycle are compared. A current goal of ours is to perform these assays on many organisms to search for evolutionary trends in refoldability, which in turn prompted us to build computational workflows with higher speed and reliability. For this study, we have refolded the proteome of *Bacillus cereus* (ATCC 14579)<sup>38</sup>

following an approach similar to those previously described.<sup>18,23</sup> The methods are described in full in the **Experimental Section**. In brief, we grew triplicate cultures of *B. cereus* to a final  $OD_{600}$  of 1.0, lysed by cryogenic pulverization, subjected the lysates to global unfolding and refolding, and performed limited proteolysis on the native extract and the refolded ones. Refolding was monitored at three time points following dilution: 1, 5, and 30 min.

In previous studies, acquisition of biochemical and biophysical metadata for each protein provided an important set of metadata which enabled correlates with refoldability to be determined from LiP-MS experiments. However, we relied on databases (EcoCyc<sup>39</sup> for *E. coli* and SGD<sup>40</sup> for yeast), which are not available in general for nonmodel organisms. To make metadata acquisition more streamlined and species-agnostic, we have implemented in FLiPPR a basic infrastructure to automate the acquisition of metadata using BioPython prediction tools and UniProt. When we analyzed the refolding LiP-MS data from *B. cereus* in FLiPPR, it was clear that although its proteins refold on the whole more efficiently than *E. coli*’s, several trends are conserved between the two organisms while others are different. For instance, we find in *B. cereus* that mildly acidic proteins (pI 5–6) are still among the worst refolders (like *E. coli*), although one major point of difference is that proteins at the extremes of the pI range ( $<5$ ,  $>10$ ) are also relatively poor refolders, a behavior quite distinct from *E. coli* (Figure 5A). Like *E. coli*, we find a relatively flat dependence on molecular weight for proteins larger than 30 kDa (Figure 5B). Moreover, in both *E. coli* and *B. cereus*, we find that proteins with no globular domains refold the best, and there is a “shallow” additional challenge to refold as the number of domains increases beyond one (Figure 5C). Whether or not bacterial proteomes generally refold quite poorly (like *E. coli*) or quite efficiently (like *B. cereus*) is a question of current active research.

## ■ DISCUSSION

In recent years, FragPipe has become recognized as a leading proteomics analysis platform which combines a fast and sensitive search engine and several quantification algorithms (with support for LFQ, SILAC, and TMT). It is also free, easy to use, and supports both data-dependent acquisition (DDA) and data-independent acquisition (DIA) modalities. Structural proteomics, an emerging field within proteomics, currently suffers from having a panoply of software pipelines which hampers interoperability and standardization. For instance, there are at least 10 different packages for analyzing cross-linking mass spectrometry data.<sup>15</sup> LiP-MS studies have historically employed primarily ProteomeDiscoverer (for DDA) and Spectronaut (for DIA) with less standardization in how raw quantifications are processed. Likewise, FPOP data has historically been analyzed in ProteomeDiscoverer with user-defined workflows to handle the experiment’s specific needs. Recently, a standardized workflow for FPOP experiments in FragPipe was proposed<sup>42</sup> showing that this platform is well-suited to bring its numerous other strengths to bear in structural proteomics. Our original goal in developing FLiPPR was to create a pipeline that would facilitate using FragPipe to analyze LiP-MS data. Following this goal, we have aimed to build a tool that is easy to use, compatible with various experimental designs, provides a range of useful outputs (including those for quality control), and facilitates metadata integration.



**Figure 5.** Comparison of nonrefolding rates between two species. (A) Fraction of proteins nonrefolding after 5 min as a function of protein isoelectric point (pI), based on cut-sites with FDR correction for two organisms, *E. coli* (ref 18) and *B. cereus* (data reported in this study). pI is calculated in BioPython. (B) Fraction of proteins nonrefolding after 5 min as a function of molecular weight (MW), based on cut-sites with FDR correction for *E. coli* and *B. cereus*. MW is calculated in BioPython. (C) Fraction of proteins nonrefolding after 5 min as a function of number of domains, based on cut-sites with FDR correction for *E. coli* and *B. cereus*. Domain assignment is determined in DomainMapper.<sup>41</sup>

LiP-MS has been applied to answer a range of biological and biophysical questions. Our lab's focus on protein refolding, which can produce very divergent protein conformations, has made us particularly sensitive to the importance of missing data in LiP-MS and how much useful information can be gained from them if used carefully. FLiPPR formalize these heuristics into software that can be widely used for structural proteomics. Compared to our previous work, we have incorporated two further improvements: (1) a protein-centric FDR correction to control for coverage bias and (2) a hierarchy of data merging rules that enables quantification from the ion level to the cut-site level. These two additions create more robust and less redundant data sets. The result is greater accuracy, which can be attested by sharper trends and correlations upon reanalyzing earlier work on the *E. coli* proteome (cf. Figure 5). As we proceed with LiP-MS experiments on proteomes from more diverse organisms (such as *B. cereus*, discussed here), we expect that the speed and standardization offered by the FragPipe/FLiPPR pipeline will prove invaluable.

There are several improvements we foresee adding to future versions of FLiPPR. The Picotti lab introduced a mixed linear model to perform normalizations that considers data from LiP

and TrP experiments collectively, rather than basing decisions on whether to normalize by the TrP experiment unilaterally.<sup>10</sup> In practice, we have found that when LiP-MS is applied to biophysical questions (and the test and control conditions arise from the same biological source but differentiate from each other by treatments performed *in vitro*), normalization is less critical and in some cases weakens the data by propagating more error to the quantifications. On the other hand, we acknowledge that careful normalization is required for some biological studies and that this option should be added to increase the generality of FLiPPR to other LiP-MS applications.

A second outstanding question is how to apply LiP-MS to interrogate the effect of post-translational modifications (PTMs) on protein structure. PTMs can profoundly impact protein structure,<sup>43,44</sup> and by extension their limited proteolysis pattern, making LiP-MS a potential technique to map this effect. For now, FLiPPR provides an output that restricts merging to the modified peptide level, and users can interrogate these files to see (for instance) if a phosphorylation within a peptide results in a different outcome compared to the same unmodified peptide. In practice, allostery can induce structural changes far from the PTM site, which creates a "proteoform problem": a half-tryptic peptide associated with a PTM need not be close to the PTM itself, and so sequencing such a half-tryptic peptide would not provide enough information to trace it back to the specific (set of) PTM(s) that induced it. Due to this ambiguity, we recommend merging over peptide modifications and regard LiP-MS experiments as probing the "average" of all proteoforms for a given protein. While separating peptide modifications more explicitly into proteoform-specific categories is a functionality we plan to add, we emphasize that new LiP-MS experiments (potentially of a middle-down nature) will be needed to confidently assess the effect of PTMs on structure.

In summary, we anticipate growth in the number and variety of applications for LiP-MS and hope that FLiPPR will contribute by lowering the barrier to adopt this structural proteomic approach and standardizing statistical procedures for LiP-MS data analysis.

## EXPERIMENTAL SECTION

### Culture and Lysis of *Bacillus cereus* (ATCC 14579)

Saturated overnight cultures of *B. cereus* cells (ATCC 14579) were used to inoculate  $3 \times 100$  mL (biological triplicate) cultures in ATCC Medium 3 in 250 mL flasks at a starting  $OD_{600}$  of 0.05. Cells were incubated at 37 °C with agitation (220 rpm) to a final  $OD_{600}$  of 1.0, followed by centrifugation at 3000 g for 10 min at 4 °C. Supernatants were removed, and cell pellets were flash frozen with liquid nitrogen for 30 s and stored at -80 °C until further use.

Cell pellets were resuspended in 1 mL of native buffer (20 mM Tris pH 8.2, 150 mM KCl, 2 mM MgCl<sub>2</sub>) with the addition of 20 units of DNase I (NEB M0303S) and protease inhibitors (500 μM PMSF; Thermo Scientific 36978, 15 μM E-64; Millipore Sigma E3132, 50 μM Bestatin; Millipore Sigma B8385). Resuspended cells were flash frozen by slow drip over liquid nitrogen followed by cryogenic pulverization with a freezer mill (SPEX Sample Prep) over eight cycles consisting of 1 min of grinding, 9 Hz, and 1 min of cooling. Pulverized lysates were transferred to a 50 mL centrifuge tube and thawed at 4 °C. Lysates were then transferred into 1.5 mL

microcentrifuge tubes and clarified at 16,000  $g$  for 15 min at 4  $^{\circ}\text{C}$  to remove cellular membrane and debris. Clarified lysates were then transferred into 3 mL Beckman Coulter Konical tubes (Beckman Coulter C14307) in preparation for ultracentrifugation at 33,300 rpm at 4  $^{\circ}\text{C}$  for 90 min using a SW55-Ti rotor without a sucrose cushion to deplete ribosomes. These clarified supernatants were then transferred into new 1.5 mL microcentrifuge tubes and protein concentrations were determined using a bicinchoninic acid assay (Rapid Gold BCA Assay, Pierce) in a clear 96-well plate (Corning Falcon 353075) with a plate reader (Molecular Devices iD3). Using the results from the BCA assay, the clarified lysates were normalized to a protein concentration of 2.0 mg mL $^{-1}$  using the same native buffer (20 mM Tris pH 8.2, 150 mM KCl, 2 mM MgCl $_2$ ). The normalized lysate is used as the starting point for all the following workflows.

### Preparation of Native and Refolded Lysates

Native lysates were prepared by diluting 58  $\mu\text{L}$  of normalized lysate with native buffer supplemented with 5.7 mg of guanidinium chloride (GdmCl) and 15.4  $\mu\text{g}$  of dithiothreitol (DTT) such that the final composition of the native lysates was 0.116 mg mL $^{-1}$  protein, 20 mM Tris pH 8.2, 150 mM KCl, 2 mM MgCl $_2$ , 0.1 mM DTT, and 0.06 M GdmCl in a 1 mL volume. These samples were prepared in biological triplicate and allowed to incubate for at least 1 h at 25  $^{\circ}\text{C}$  prior to limited proteolysis.

Unfolded lysates were prepared by concentrating 290  $\mu\text{L}$  of normalized lysate supplemented with 28.7 mg of GdmHCl and 77.1  $\mu\text{g}$  of DTT (from a solution stock) in a Vacufuge Plus (Eppendorf) to a final volume of 50  $\mu\text{L}$ . The final composition of the unfolded lysates was 11.6 mg mL $^{-1}$  protein, 116 mM Tris pH 8.2, 870 mM KCl, 11.6 mM MgCl $_2$ , 10 mM DTT, and 6 M GdmCl. These samples were prepared in biological triplicate and allowed to incubate for at least 16 h at 25  $^{\circ}\text{C}$  prior to refolding and subsequent limited proteolysis.

Refolded lysates were prepared by diluting 5  $\mu\text{L}$  of unfolded lysate into 495  $\mu\text{L}$  of refolding buffer (19 mM Tris pH 8.2, 143 mM KCl, 1.9 mM MgCl $_2$ ). The final composition of the refolded lysates was 0.116 mg mL $^{-1}$  protein, 20 mM Tris pH 8.2, 150 mM KCl, 2 mM MgCl $_2$ , 0.1 mM DTT, and 0.06 M GdmCl in a 500  $\mu\text{L}$  volume. These samples were prepared in biological triplicate and allowed to incubate for 1, 5, and 30 min at 25  $^{\circ}\text{C}$  prior to limited proteolysis.

### Limited Proteolysis and Mass Spectrometry Sample Preparation

A stock of PK solution was freshly prepared at a concentration of 0.116 mg mL $^{-1}$  Proteinase K in native buffer with 10% glycerol. Triplicate limited proteolysis (LiP) samples of the native and refolded conditions, for all time points, were then generated by aliquoting 200  $\mu\text{L}$  of each native or refolded sample into new 1.5 mL microcentrifuge tubes containing 2  $\mu\text{L}$  of PK solution (1:100 enzyme:substrate mass ratio), quickly aspirated 10 times, and incubated for exactly 1 min at 25  $^{\circ}\text{C}$ . Samples were then transferred to a 110  $^{\circ}\text{C}$  mineral oil bath for 5 min to inactivate Proteinase K. Boiled samples were then transferred into new 1.5 mL microcentrifuge tubes containing 150 mg of urea to achieve a final volume of 312  $\mu\text{L}$  and 8 M urea concentration. Triplicate trypsin-only (TrP) controls of the native and refolded samples were generated in the same way without the addition of PK solution. This process generated the following 18 samples: 3  $\times$  Native\_TrP, 3  $\times$

Refolded\_TrP, 3  $\times$  Native\_LiP, 3  $\times$  Refolded\_1 min\_LiP, 3  $\times$  Refolded\_5 min\_LiP, and 3  $\times$  Refolded\_30 min\_LiP.

All samples then received 6.24  $\mu\text{L}$  of freshly prepared 500 mM DTT (final concentration 10 mM DTT) to reduce disulfide bonds and were incubated for 30 min at 37  $^{\circ}\text{C}$  with agitation at 700 rpm followed by 17  $\mu\text{L}$  of freshly prepared 750 mM IAA (final concentration 40 mM IAA) and incubated for 45 min in the dark at 25  $^{\circ}\text{C}$  to alkylate reduced cysteines. Next, 2  $\mu\text{L}$  of 0.116 mg mL $^{-1}$  of LysC (NEB P8109) were added to each sample and incubated for 2 h at 37  $^{\circ}\text{C}$ . Samples were then diluted with 1005  $\mu\text{L}$  of 100 mM ammonium bicarbonate followed by 4  $\mu\text{L}$  of 0.116 mg mL $^{-1}$  of trypsin-ultra (NEB P8101) and incubated at 25  $^{\circ}\text{C}$  for 16–24 h at 25  $^{\circ}\text{C}$ . Trypsin was then quenched with 1% v/v trifluoroacetic acid prior to desalting with Sep-Pak C18 1 cm $^3$  cartridges (Waters).

Cartridges were first conditioned (2  $\times$  1 mL 80% ACN, 0.5% TFA) and equilibrated (4  $\times$  1 mL 0.5% TFA) before samples were slowly loaded under a weak vacuum. The columns were then washed (4  $\times$  1 mL 0.5% TFA), and peptides were eluted by the addition of 1 mL of elution buffer (80% ACN, 0.5% TFA). During elution, vacuum cartridges were suspended above 15 mL conical tubes, placed in a swing-bucket rotor (Eppendorf 5910R), and spun for 5 min at 300 g. Eluted peptides were transferred from Falcon tubes back into new 1.5 mL microcentrifuge tubes and dried using a Vacufuge Plus (Eppendorf). Dried peptides were stored at –80  $^{\circ}\text{C}$  until analysis. For analysis, samples were vigorously resuspended in 0.1% FA in Optima water (ThermoFisher) to a final concentration of 1 mg mL $^{-1}$ .

### Mass Spectrometry Acquisition

Chromatographic separation of digests was carried out on a Thermo UltiMate3000 UHPLC system with an Acclaim Pepmap RSLC, C18, 75  $\mu\text{m} \times 25$  cm, 2  $\mu\text{m}$ , 100  $\text{\AA}$  column. Approximately 1  $\mu\text{g}$  of protein was injected onto the column. The column temperature was maintained at 40  $^{\circ}\text{C}$ , and the flow rate was set to 300 nL min $^{-1}$  for the duration of the run. Solvent A (0.1% FA) and solvent B (0.1% FA in ACN) were used as the chromatography solvents.

The samples were run through the UHPLC system as follows: peptides were allowed to accumulate onto the trap column (Acclaim PepMap 100, C18, 75  $\mu\text{m} \times 2$  cm, 3  $\mu\text{m}$ , 100  $\text{\AA}$  column) for 10 min (during which the column was held at 2% solvent B). The peptides were resolved by switching the trap column to be in-line with the separating column, quickly increasing the gradient to 5% B over 5 min and then applying a 95 min linear gradient from 5% B to 25% B. Subsequently, the gradient was increased from 25% B to 40% B over 25 min and then increased again from 40% B to 90% B over 5 min. The column was then cleaned with a sawtooth gradient to purge residual peptides between runs in a sequence.

A Thermo Q-Exactive HF-X Orbitrap mass spectrometer was used to analyze protein digests. A full MS scan in positive ion mode was followed by 20 data-dependent MS scans. The full MS scan was collected using a resolution of 120,000 (at  $m/z$  200), an AGC target of 3E6, a maximum injection time of 64 ms, and a scan range from 350 to 1500  $m/z$ . The data-dependent scans were collected with a resolution of 15,000 (at  $m/z$  200), an AGC target of 1E5, a minimum AGC target of 8E3, a maximum injection time of 55 ms, and an isolation window of 1.4  $m/z$  units. To dissociate precursors prior to their reanalysis by MS2, peptides were subjected to an HCD of 28% normalized collision energies. Fragments with charges of

1, 6, 7, or higher and unassigned were excluded from analysis, and a dynamic exclusion window of 30.0 s was used for the data-dependent scans.

### Analyzing MS Data in FragPipe

FragPipe v20.0, along with MSFragger v3.8, IonQuant v1.9.8, and Philosopher v5.0, were used to analyze raw mass spectra with label-free quantification with match between runs enabled. Default settings were used except those delineated in Table S1. Namely, the peptide digest pattern was set to semienzymatic, methionine oxidation and N-terminal acetylation were set as dynamic modification, and cysteine carbamidomethylation was set as a static modification. The workflows for *B. cereus* refolding and the reanalysis of PXD025926 were set up according to the conventions set forth in Figure S2B, wherein multiple LiP control–test pairs are normalized to a single TrP control–test pair. FragPipe outputs were then passed to FLiPPR v0.0.7 with the command shown in Figure S2B. Output files from FLiPPR were then further processed in Python to create summaries and graphical representations of both data sets.

## ■ ASSOCIATED CONTENT

### Data Availability Statement

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the data set identifier PXD047776. Quantification for all peptides and all cut sites, as well as Jupyter Notebooks for processing data to generate Figures 4 and 5, have been deposited on Zenodo (10.5281/zenodo.10372617).

### ■ Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jproteome.3c00887>.

Figure S1, the match-between-run (MBR) false discovery rate (FDR) controls sensitivity to all-or-nothing peptides; Figure S2, filename conventions and implementing normalization in FLiPPR; Table S1, modifications to default FragPipe settings for refolding experiments reported in this study; and Note S1, the match-between-run (MBR) false discovery rate (FDR) controls sensitivity to all-or-nothing peptides (PDF)

Summary data for all proteins assessed in the refolding experiment on *E. coli* (reanalyzed from PXD025926) (XLSX)

Summary data for all proteins assessed in the refolding experiment on *B. cereus* (PXD047776) (XLSX)

## ■ AUTHOR INFORMATION

### Corresponding Author

Stephen D. Fried — Department of Chemistry, Johns Hopkins University, Baltimore, Maryland 21218, United States; T. C. Jenkins Department of Biophysics, Johns Hopkins University, Baltimore, Maryland 21218, United States;  [orcid.org/0000-0003-2494-2193](https://orcid.org/0000-0003-2494-2193); Email: [sdfried@jhu.edu](mailto:sdfried@jhu.edu)

### Authors

Edgar Manriquez-Sandoval — Department of Chemistry, Johns Hopkins University, Baltimore, Maryland 21218, United States; T. C. Jenkins Department of Biophysics, Johns Hopkins University, Baltimore, Maryland 21218, United States

Joy Brewer — Department of Chemistry and Biochemistry, Old Dominion University, Norfolk, Virginia 23529, United States  
Gabriela Lule — Department of Chemistry, Johns Hopkins University, Baltimore, Maryland 21218, United States  
Samanta Lopez — Department of Chemistry, Johns Hopkins University, Baltimore, Maryland 21218, United States

Complete contact information is available at: <https://pubs.acs.org/10.1021/acs.jproteome.3c00887>

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

We would like to thank the NSF Division of Molecular and Cellular Biology for a CAREER grant (2045844) and NIH/NIGMS for a New Innovator Award (DP2-GM140926), both to S.D.F. E.M.S. thanks the Program in Molecular Biophysics training grant (NIH-T32GM135131). J.B. thanks the support from a training grant issued from NIH/NIGMS (NIH-1T34GM118259-01A1). Matching funds from Old Dominion University are also acknowledged. Funds from the NSF CAREER grant, a Camille Dreyfus Teacher-Scholar Award, and a Cottrell Scholars Award supported the undergraduates involved in the project (J.B., G.L., and S.L.). We thank Alexey Nesvizhskii for critical manuscript reading and helpful feedback.

## ■ REFERENCES

- (1) Hyung, S. J.; Ruotolo, B. T. Integrating mass spectrometry of intact protein complexes into structural proteomics. *Proteomics* **2012**, *12*, 1547–1564.
- (2) Engen, J. R.; Smith, D. L. Investigating the higher order structure of proteins. Hydrogen exchange, proteolytic fragmentation, and mass spectrometry. *Methods Mol. Biol.* **2000**, *146*, 95–112.
- (3) Costello, S. M.; Shoemaker, S. R.; Hobbs, H. T.; Nguyen, A. W.; Hsieh, C.; Maynard, J. A.; McLellan, J. S.; Pak, J. E.; Marqusee, S. The SARS-CoV-2 spike reversibly samples an open-trimer conformation exposing novel epitopes. *Nature Structural & Molecular Biology* **2022**, *29*, 229–238.
- (4) Walker, E. J.; Bettinger, J. Q.; Welle, K. A.; Hryhorenko, J. R.; Ghaemmaghami, S. Global analysis of methionine oxidation provides a census of folding stabilities for the human proteome. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 6081–6090.
- (5) Tran, D. T.; Adhikari, J.; Fitzgerald, M. C. Stable Isotope Labeling with Amino Acids in Cell Culture (SILAC)-based strategy for proteome-wide thermodynamic analysis of protein-ligand binding interactions. *Mol. Cell. Proteomics* **2014**, *13*, 1800–1813.
- (6) Jones, L. M.; Sperry, J. B.; Carroll, J. A.; Gross, M. L. Fast photochemical oxidation of proteins for epitope mapping. *Anal. Chem.* **2011**, *83*, 7657–7661.
- (7) Johnson, D. T.; Di Stefano, L. H.; Jones, L. M. Fast photochemical oxidation of proteins (FPOP): A powerful mass spectrometry-based structural proteomics tool. *J. Biol. Chem.* **2019**, *294*, 11969–11979.
- (8) Liu, F.; Fitzgerald, M. C. Large-Scale Analysis of Breast Cancer-Related Conformational Changes in Proteins Using Limited Proteolysis. *J. Proteome Res.* **2016**, *15*, 4666–4674.
- (9) Feng, Y.; De Franceschi, G.; Kahraman, A.; Soste, M.; Melnik, A.; Boersema, P. J.; de Laureto, P. P.; Nikolaev, Y.; Oliveira, A. P.; Picotti, P. Global analysis of protein structural changes in complex proteomes. *Nat. Biotechnol.* **2014**, *32*, 1036–1044.
- (10) Malinovska, L.; Cappelletti, V.; Kohler, D.; Piazza, I.; Tsai, T. H.; Pepejnjak, M.; Stalder, P.; Dörig, C.; Sesterhenn, F.; Elsässer, F.; Kralickova, L.; Beaton, N.; Reiter, L.; de Souza, N.; Vitek, O.; Picotti, P. Proteome-wide structural changes measured with limited

proteolysis-mass spectrometry: an advanced protocol for high-throughput applications. *Nature Protocols* 2022 18:3 2023, 18, 659–682.

(11) Kaur, U.; Meng, H.; Lui, F.; Ma, R.; Ogburn, R. N.; Johnson, J. H. R.; Fitzgerald, M. C.; Jones, L. M. Proteome-Wide Structural Biology: An Emerging Field for the Structural Analysis of Proteins on the Proteomic Scale. *J. Proteome Res.* 2018, 17, 3614–3627.

(12) O'Reilly, F. J.; Rappaport, J. Cross-linking mass spectrometry: methods and applications in structural, molecular and systems biology. *Nature Structural & Molecular Biology* 2018, 25, 1000–1008.

(13) Chavez, J. D.; Bruce, J. E. Chemical cross-linking with mass spectrometry: a tool for systems structural biology. *Curr. Opin. Chem. Biol.* 2019, 48, 8–18.

(14) Sinz, A. Cross-Linking/Mass Spectrometry for Studying Protein Structures and Protein-Protein Interactions: Where Are We Now and Where Should We Go from Here? *Angew. Chem., Int. Ed.* 2018, 57, 6390–6396.

(15) Iacobucci, C.; Götze, M.; Sinz, A. Cross-linking/mass spectrometry to get a closer view on protein interaction networks. *Curr. Opin. Biotechnol.* 2020, 63, 48–53.

(16) Faustino, A. M.; Sharma, P.; Manriquez-Sandoval, E.; Yadav, D.; Fried, S. D. Progress toward Proteome-Wide Photo-Cross-Linking to Enable Residue-Level Visualization of Protein Structures and Networks In Vivo. *Anal. Chem.* 2023, 95, 10670–10685.

(17) Lee, K.; O'Reilly, F. J. Cross-linking mass spectrometry for mapping protein complex topologies in situ. *Essays Biochem.* 2023, 67, 215–228.

(18) To, P.; Whitehead, B.; Tarbox, H. E.; Fried, S. D. Nonrefoldability is Pervasive Across the *E. coli* Proteome. *J. Am. Chem. Soc.* 2021, 143, 11435–11448.

(19) Piazza, I.; Kochanowski, K.; Cappelletti, V.; Fuhrer, T.; Noor, E.; Sauer, U.; Picotti, P. A Map of Protein-Metabolite Interactions Reveals Principles of Chemical Communication. *Cell* 2018, 172, 358–372.

(20) Paukštė, J.; López Cabezas, R. M.; Feng, Y.; Tong, K.; Schnyder, D.; Elomaa, E.; Gregorova, P.; Doudin, M.; Särkkä, M.; Sarameri, J.; Lippi, A.; Viihinen, H.; Juutila, J.; Nieminen, A.; Törönen, P.; Holm, L.; Jokitalo, E.; Krisko, A.; Huiskonen, J.; Sarin, L. P.; Hietakangas, V.; Picotti, P.; Barral, Y.; Saarikangas, J. Global analysis of aging-related protein structural changes uncovers enzyme-polymerization-based control of longevity. *Mol. Cell* 2023, 83, 3360–3376.

(21) Shuken, S. R.; Rutledge, J.; Iram, T.; Losada, P. M.; Wilson, E. N.; Andreasson, K. I.; Leib, R. D.; Wyss-Coray, T. Limited proteolysis-mass spectrometry reveals aging-associated changes in cerebrospinal fluid protein abundances and structures. *Nature Aging* 2022, 2, 379–388.

(22) Leuenberger, P.; Ganscha, S.; Kahraman, A.; Cappelletti, V.; Boersema, P. J.; von Mering, C.; Claassen, M.; Picotti, P. Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability. *Science* 2017, 355, No. eaai7825.

(23) To, P.; Xia, Y.; Lee, S. O.; Devlin, T.; Fleming, K. G.; Fried, S. D. A proteome-wide map of chaperone-assisted protein refolding in a cytosol-like milieu. *Proc. Natl. Acad. Sci. U. S. A.* 2022, 119, No. e2210536119.

(24) Ma, R.; Meng, H.; Wiebelhaus, N.; Fitzgerald, M. C. Chemo-Selection Strategy for Limited Proteolysis Experiments on the Proteomic Scale. *Anal. Chem.* 2018, 90, 14039–14047.

(25) Wiebelhaus, N.; Singh, N.; Zhang, P.; Craig, S. L.; Beratan, D. N.; Fitzgerald, M. C. Discovery of the Xenon-Protein Interactome Using Large-Scale Measurements of Protein Folding and Stability. *J. Am. Chem. Soc.* 2022, 144, 3925–3938.

(26) Taverna, D.; Gaspari, M. A critical comparison of three MS-based approaches for quantitative proteomics analysis. *J. Mass Spectrom.* 2021, 56, No. e4669.

(27) Hogrebe, A.; von Stechow, L.; Bekker-Jensen, D.; Weinert, B. T.; Kelstrup, C. D.; Olsen, J. V. Benchmarking common quantification strategies for large-scale phosphoproteomics. *Nat. Commun.* 2018, 9, 1045.

(28) Yu, F.; Haynes, S. E.; Teo, G. C.; Avtonomov, D. M.; Polasky, D. A.; Nesvizhskii, A. I. Fast Quantitative Analysis of timsTOF PASEF Data with MSFragger and IonQuant. *Molecular & Cellular Proteomics* 2020, 19, 1575–1585.

(29) Kong, A. T.; Leprevost, F. V.; Avtonomov, D. M.; Mellacheruvu, D.; Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods* 2017, 14, 513–520.

(30) Yu, F.; Haynes, S. E.; Nesvizhskii, A. I. IonQuant Enables Accurate and Sensitive Label-Free Quantification With FDR-Controlled Match-Between-Runs. *Molecular & Cellular Proteomics* 2021, 20, 100077.

(31) Palomba, A.; Abbondio, M.; Fiorito, G.; Uzzau, S.; Pagnozzi, D.; Tanca, A. Comparative Evaluation of MaxQuant and Proteome Discoverer MS1-Based Protein Quantification Tools. *J. Proteome Res.* 2021, 20, 3497–3507.

(32) Kong, W.; Hui, H. W. H.; Peng, H.; Goh, W. W. B. Dealing with missing values in proteomics data. *Proteomics* 2022, 22, 2200092.

(33) Harris, L.; Fondrie, W. E.; Oh, S.; Noble, W. S. Evaluating Proteomics Imputation Methods with Improved Criteria. *J. Proteome Res.* 2023, 22, 3427–3438.

(34) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J. L. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 2009, 25, 1422–1423.

(35) Emenecker, R. J.; Griffith, D.; Holehouse, A. S. Metapredict: a fast, accurate, and easy-to-use predictor of consensus disorder and structure. *Biophys. J.* 2021, 120, 4312–4319.

(36) Manriquez-Sandoval, E.; Fried, S. D. DomainMapper: Accurate domain structure annotation including those with non-contiguous topologies. *Protein Sci.* 2022, 31, No. e4465.

(37) Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 1995, 57, 289–300.

(38) Frankland, G. C.; Frankland, P. F. Studies on Some New Micro-Organisms Obtained from Air. *Philosophical Transactions of the Royal Society B* 1887, 178, 257–287.

(39) Keseler, I. M.; Mackie, A.; Santos-Zavaleta, A.; Billington, R.; Bonavides-Martinez, C.; Caspi, R.; Fulcher, C.; Gama-Castro, S.; Kothari, A.; Krummenacker, M.; Latendresse, M.; Muñiz-Rascado, L.; Ong, Q.; Paley, S.; Peralta-Gil, M.; Subhraveti, P.; Velázquez-Ramírez, D. A.; Weaver, D.; Collado-Vides, J.; Paulsen, I.; Karp, P. D. The EcoCyc database: reflecting new knowledge about *Escherichia coli* K-12. *Nucleic Acids Res.* 2017, 45, D543–D550.

(40) Wong, E. D.; Miyasato, S. R.; Aleksander, S.; Karra, K.; Nash, R. S.; Skrzypek, M. S.; Weng, S.; Engel, S. R.; Cherry, J. M. *Saccharomyces* genome database update: server architecture, pan-genome nomenclature, and external resources. *Genetics* 2023, 224, iyac191.

(41) Manriquez-Sandoval, E.; Fried, S. D. DomainMapper: Accurate domain structure annotation including those with non-contiguous topologies. *Protein Sci.* 2022, 31, No. e4465.

(42) Rojas Ramirez, C.; Espino, J. A.; Jones, L. M.; Polasky, D. A.; Nesvizhskii, A. I. Efficient Analysis of Proteome-wide FPOP Data by FragPipe. *bioRxiv* 2023, 95, 16131.

(43) Kumar, A.; Narayanan, V.; Sekhar, A. Characterizing Post-Translational Modifications and Their Effects on Protein Conformation Using NMR Spectroscopy. *Biochemistry (N. Y.)* 2020, 59, S7–S73.

(44) Huang, J.; Chu, X.; Luo, Y.; Wang, Y.; Zhang, Y.; Zhang, Y.; Li, H. Insights into Phosphorylation-Induced Protein Allostery and Conformational Dynamics of Glycogen Phosphorylase via Integrative Structural Mass Spectrometry and In Silico Modeling. *ACS Chem. Biol.* 2022, 17, 1951–1962.