pyBinder: Label-free Quantitation to Advance Affinity Selection-Mass Spectrometry

Joseph S. Brown[1†], Michael A. Lee[1†], Wayne Vuong[1], Andrei Loas[1], Bradley L. Pentelute[1-4*]

[1] Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States
[2] The Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, United States
[3] Center for Environmental Health Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States
[4] Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, United States
*Email: blp@mit.edu
†Equal contribution

1 Abstract

Affinity selection-mass spectrometry (AS-MS) is a ligand discovery platform that relies upon mass spectrometry to identify molecules bound to a biomolecular target. When utilized with large peptide libraries (10^8 members), AS-MS sample complexity can surpass the sequencing capacity of modern mass spectrometers, resulting in incomplete data, identification of few target-specific ligands, and/or incomplete sequencing. To address this challenge, we introduce pyBinder to apply label-free quantitation (LFQ) to AS-MS data to process primary MS^1 data and develop two scores to rank the peptides from the integration of their peak area: target selectivity and concentration-dependent enrichment. We benchmark pyBinder utilizing AS-MS data developed against a protein, anti-hemagglutinin antibody 12ca5, revealing that peptides that contain a motif known for target-specific high-affinity binding are well characterized by these two scores. AS-MS data from a second protein target, WD Repeat Domain 5 (WDR5), is analyzed to confirm the two pyBinder scores reliably capture the target-specific motif-containing peptides. From the results delivered by pyBinder, a list of target-selective ions is developed and fed back into subsequent MS experiments to facilitate expanded data generation and the targeted discovery of selective ligands. pyBinder analysis resulted in a four-fold increase in motif-containing sequence identification for WDR5 (from 3 ligands discovered to 14 discovered), showing the utility of the two scores. This work establishes an improved approach for AS-MS to enable discovery outcomes (i.e., more ligands identified), but also a way to compare AS-MS data across samples, protocols, and conditions broadly.

2 Introduction

Affinity selection-mass spectrometry (AS-MS) discovers high-affinity ligands to biomolecular targets using mass spectrometry for ligand identification.[1–3] The selection principles of AS-MS are highly similar to phage and mRNA display,[4–6] though AS-MS generally utilizes a single enrichment step without genetic amplification. AS-MS typically utilizes synthetic peptide libraries, providing unrestricted access to non-natural amino acids and a facile opportunity to tailor library design to the biomolecule target. Thus, one of the main applications of AS-MS is the selection of small combinatorial libraries (10^3-10^6 members) biased or 'focused' toward the target to gain structure-activity relationship (SAR) information.[7–10] These approaches can accelerate medicinal chemistry efforts by the rapid identification of 'hot-spot' residues critical for activity as well as the combinatorial sampling of the chemical space available to non-natural amino acids.[7,11,12] Beyond these focused efforts, recent advancements have demonstrated *de novo* ligand discovery with AS-MS from fully randomized peptide and peptidomimetic libraries up to 10^8 members against several targets.[13–16] Despite its overall use, AS-MS heavily

https://doi.org/10.26434/chemrxiv-2024-lfqjb **ORCID:** https://orcid.org/0000-0002-7242-801X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

depends on high-resolution mass spectrometry analysis and stands to benefit by leveraging methods from the field of MS-based proteomics.

Solutions developed to solve the data incompleteness from the field of proteomics could be used to improve AS-MS. MS-based proteomics has detailed the “missing value” problem, hallmarked by an incomplete set of peptides or proteins identified across samples or replicates.[17–20] A similar concept can be applied to AS-MS, where there is little-to-no overlap of the identified sequences of peptide ligands across replicates. This challenge is pronounced by the common practice in AS-MS to run the spectrometer in data dependent acquisition (DDA) mode. In DDA, individual precursor ions are selected from the primary mass spectrum (MS^1) for fragmentation by tandem mass spectrometry (MS^2). Rules for precursor selection are clearly programmed into the spectrometer method; however, the selection process is not robust. In other words, DDA is stochastic and can result in incomplete data collection. Due to the time required to perform DDA, the speed of the mass spectrometer can become insufficient for complex samples where peptides elute together quickly. High sample complexity combined with the stochasticity of precursor selection can lead to inconsistent peptide identifications across technical replicates. Together, the typical approach for mass spectrometry in AS-MS can lead to many peptides being missed, as detected later from retrospective analysis (see Figure S1).

Label-free quantitation (LFQ) is fully compatible with AS-MS and delivers a complete dataset without relying on peptide identification from tandem spectrometry. As a cornerstone in MS-based proteomics, LFQ utilizes MS^1 precursor ions to examine sample composition. This approach compares peptide ions without the need for sequence databases, stable isotope labeling, or chemical modifications seen in other methods[21–23] ensures AS-MS data can be directly analyzed. For unambiguous discernment of peptide ions, LFQ strongly and benefits from high-resolution instrumentation.[24,25] LFQ analysis of mass spectrometry data have become used as a part of open-source and commercial software and including MaxQuant,[22] Proteome Discoverer, and PEAKS Studio.[26–28] LFQ has been proven to increase data depth, sensitivity, and completeness with applications in biomarker discovery, disease profiling, elucidation of drug mechanisms, and single-cell proteomics. Thus, the versatility of LFQ is evident in both basic and applied research,[21,29] and might enhance the capabilities of AS-MS.

We demonstrate the integration of LFQ into AS-MS using Python, named ‘pyBinder,’ for the improved discovery of target-selective, high-affinity peptide ligands. The broader goal of AS-MS is to discover target-selective ligands, which can be understood from the MS^1 data. However, data processing methods in AS-MS have primarily focused on filtering peptide sequencing data derived from tandem MS^2 spectra.[30] To this end, pyBinder generates two scores for each peptide ion identified: (i) target selectivity, established by comparing target protein versus off-target samples; and, (ii) a concentration-dependent enrichment score (CDE), calculated by analyzing the correlation between peptide intensity and target concentration in the AS-MS experiments. We have validated our method by analyzing ligands targeting the anti-hemagglutinin antibody (12ca5) and WD repeat-containing protein 5 (WDR5), both known for their high-affinity binding motifs. The outcomes from pyBinder analysis indicate that peptides containing the 12ca5 and WDR5 motifs are highly-ranked based on target selectivity and CDE compared to other peptide features identified in the LFQ analysis. Lastly, re-measurement of the sample by targeting other highly ranked peptide ions increased the discovery of motif-containing peptide sequences. Thus, pyBinder appears might aid peptide therapeutic candidate discovery from AS-MS data.

https://doi.org/10.26434/chemrxiv-2024-lfqjb **ORCID:** https://orcid.org/0000-0002-7242-801X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

3 Results and Discussion

Improvements to mass spectrometry methods stand to improve AS-MS broadly. To understand AS-MS data, we define two terms: “sequencing coverage” and “sequencing fidelity.” First, sequencing coverage is defined as the percentage of all peptide ions isolated for MS^2 fragmentation. Sequencing coverage is calculated in post-processing analysis by comparing the number of MS^2 spectra gathered versus the total number of peptide ion features in the sample. A peptide ion feature is defined by its retention time, mass-to-charge ratio (m/z), and isotopic pattern. Low sequencing coverage means that the AS-MS sample was too complex. In this scenario, the number of peptides in the sample surpassed the spectrometer’s capability. In comparison, high sequencing coverage means the spectrometer gathered MS^2 spectra for almost all peptides in the sample. Our second term, sequencing fidelity, assesses the quality of the MS^2 spectra. It is defined as the percentage of all MS^2 spectra that result in high-quality assignment of a peptide sequence in post-processing analysis. In our work, *de novo* sequencing analysis was performed in PEAKS Studio where an Average Local Confidence (ALC) of ≥80 was considered for high-quality sequence assignment.[26] Low sequence fidelity can be due to several factors: low peptide abundance, co-isolation of multiple peptide precursors, poor fragmentation patterns or kinetics, and/or mistaken isolation of non-peptide molecules.[31–34] Thus, we can evaluate approaches to improve both the quantity and quality of AS-MS from the sequencing coverage and fidelity, respectively.

Retrospective analysis of a prior AS-MS discovery campaign estimated the sequence coverage and fidelity to be ~10-18% and ~1.7-8%, respectively (Figure 1), concretely describing the incompleteness of AS-MS data. We reanalyzed the raw data from our previously published ligand discovery campaign of a canonical 12-mer library against angiotensin-converting enzyme 2 (ACE2) with anti-hemagglutinin antibody 12ca5 used as a side-by-side off-target control.[14] The mass spectrometer gathered 3,468 (ACE2) and 5,895 (12ca5) MS^2 spectra out of 32,722 and 33,306 total identified peptide features respectively, meaning the sequence coverage was low at 10.6% for ACE2 and 17.7% for 12ca5. Most peptides (>80%) were not isolated for MS^2 fragmentation by the mass spectrometer. The low coverage is partly due to the use of both higher-energy collisional dissociation (HCD) and orthogonal electron-transfer dissociation (ETD), which have been previously seen to improve sequencing fidelity of *de novo* sequencing.[8,13,30] Using both fragmentation modes increases the cycle time to fragment each peptide precursor, with a rate of ~1.2 MS^2 spectra per second was observed here. Of the MS^2 spectra gathered, most were of poor quality with the sequence fidelity 1.7% for ACE2 and 8.0% for 12ca5. When sequencing coverage and fidelity are combined, only 0.18 – 1.4% of all peptides were successfully characterized by the mass spectrometer to yield a peptide sequence.

https://doi.org/10.26434/chemrxiv-2024-lfqjb **ORCID:** https://orcid.org/0000-0002-7242-801X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

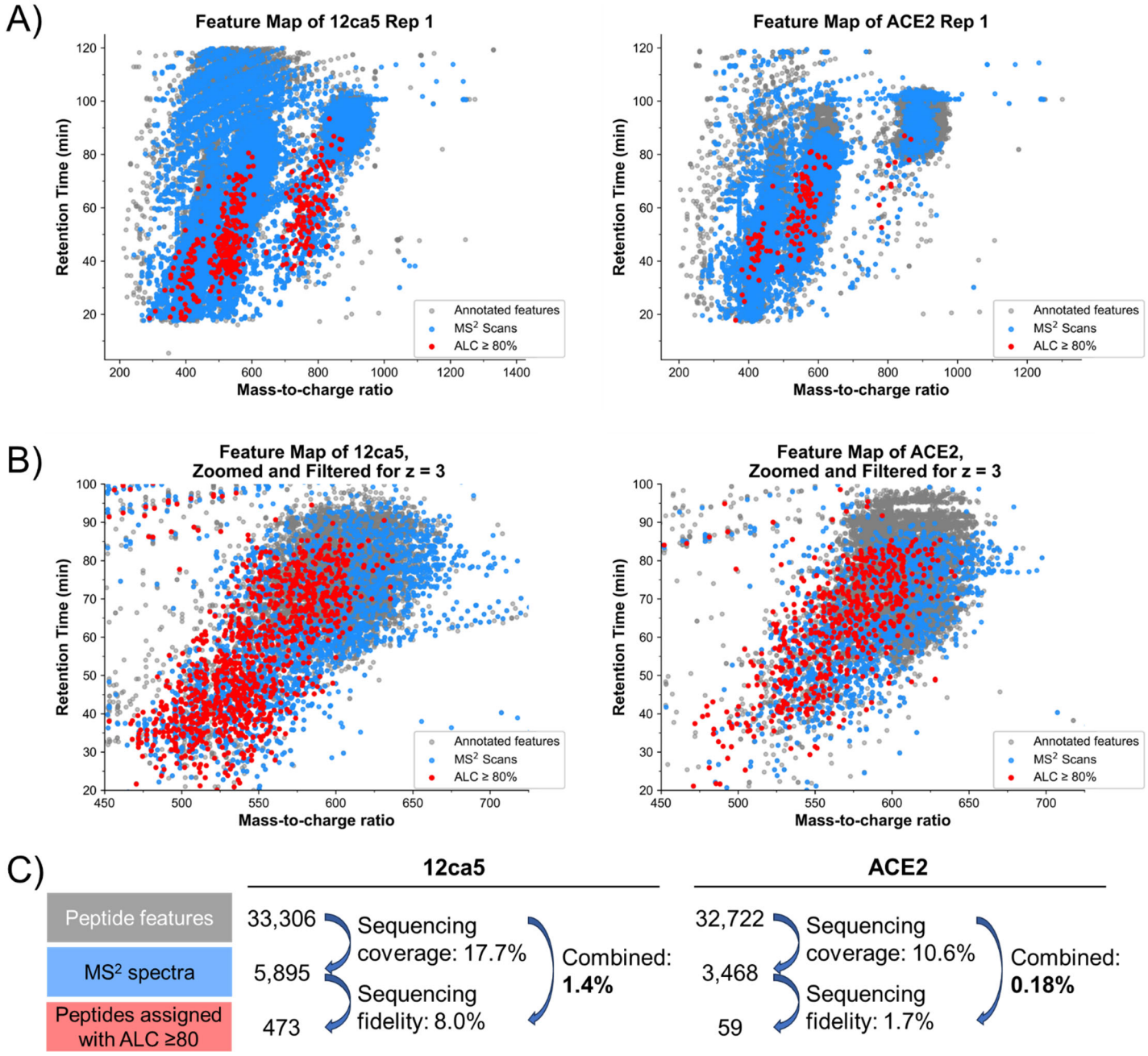

Figure 1. Retrospective analysis of previous AS-MS campaigns reveals the opportunity for deeper data analysis by LFQ. (A) The mass chromatogram of mass-to-charge ratio versus retention time with peptide features identified by PEAKS studio in black, all collected MS^2 scans in blue, and all MS^2 scans that resulted in a high-quality sequence assignment in red. High-quality sequence assignment was defined by having an ALC score calculated by PEAKS Studio ≥80 with a sequence that conforms to the synthetic library design. (B) A zoomed in portion of the mass-to-charge ratio versus retention time plot filtered to show only z states of 3 shows the low coverage of high confidence identifications during untargeted runs. (C) Statistics for each of the three groups, showing the percentages of the total number of features subjected to MS^2 and that resulted in high confidence sequences that conform to the synthetic library design.

https://doi.org/10.26434/chemrxiv-2024-lfqjb **ORCID:** https://orcid.org/0000-0002-7242-801X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

To improve the mass spectrometry performance in AS-MS, the common method of spectral database matching was considered first as it would improve sequencing fidelity. Database matching constrains the sequence assignment of the MS^2 spectra to a list of peptide sequences that may be in the sample. Thus, peptide sequences can be assigned from MS^2 spectra even if the spectra are of sub-optimal quality for de novo sequencing. However, spectral matching appears intractable for 10^8 combinatorial libraries because the corresponding database must be large. These types of libraries are designed to include numerous amino acids at multiple positions in an unbiased, randomized, distribution and thus samples from a much larger theoretical sequence space (e.g., a 10^8 library samples a 10^{15} sequence space).[13–16] Because the synthesis is unbiased, the database would need to include all sequences in the full theoretical sequence space. For a routinely used 10^8 library, this large database would result in a ~15000 TB FASTA file using a minimal UTF-8 encoding, unable to be handled by most MS analysis software and storage media. However, database matching may still be tractable for smaller, focused libraries, depending on their design. Because spectral matching is not possible for large libraries used in de novo discovery, other methods that rely on spectral matching appear intractable, including data-independent acquisition (DIA) spectral matching to improve the MS^2 deconvolution of co-isolated peptides.[35–37] Nevertheless, several strategies from MS-based proteomics appear compatible with AS-MS, including LFQ as previously mentioned.[21–23]

We introduce 'pyBinder' to combine LFQ with AS-MS to understand the quality and value of the ligands discovered for their target-selectivity (Figure 2). While standard software packages can accomplish LFQ analysis of MS data,[22,26–28] we sought to develop an open-source approach in Python and utilized pyOpenMS.[38] Starting with the ACE2 and 12ca5 dataset in Figure 1, pyOpenMS was used to identify peptide ion features by fitting the Averagine isotopic distribution[39] with z state filtering to compile a list of peptide features per AS-MS sample replicate. Optimization of the feature identification was performed by comparing the overlap in features identified between pyOpenMS and PEAKS Studio as a baseline, until both showed comparable feature detection capability. Details of the parameter optimization are given in Table S2. Because AS-MS experiments are completed in triplicate, the map of peptide features (retention time vs m/z) from each sample was aligned in retention time using the pose clustering algorithm as previously described.[40] The resulting aligned map generated a consensus list of all peptide ion features observed across all proteins and replicates. Comparison of each peptide feature can then discern the target selectivity of each.

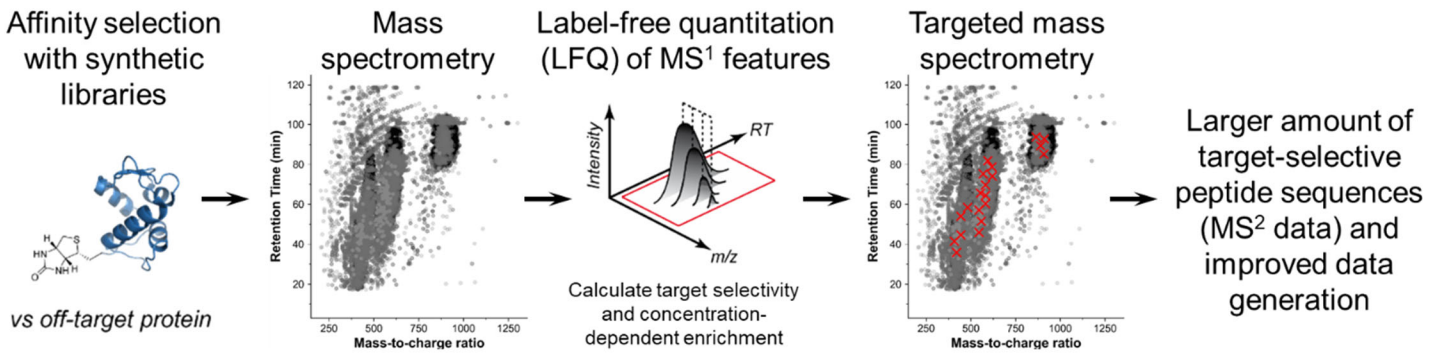

Figure 2. Label-free quantitation (LFQ) improves the affinity selection-mass spectrometry (AS-MS) discovery platform. LFQ performed by pyBinder enables the analysis of AS-MS data from the MS^1 peptide features without relying on tandem sequencing results (MS^2 data). Thus, the success of the affinity selection can be robustly judged by the enrichment level of peptides identified from MS^1 features. The MS^1 features can be evaluated for the target-selectivity as well as target concentration-dependent enrichment (CDE). With the target-selectivity and CDE scores, a list of promising peptide features can be generated by pyBinder and fed back into

https://doi.org/10.26434/chemrxiv-2024-lfqjb **ORCID:** https://orcid.org/0000-0002-7242-801X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

a subsequent targeted mass spectrometry experiment to reveal a larger amount of target-selective peptide ligands.

To discern target selectivity, pyBinder analyzes extracted ion chromatograms (EICs) for all peptide ion features discovered. EICs shows the signal intensity as a function of retention time, where the signal is the intensity from the spectrometer at the mass of the peptide ion ± 0.005 m/z (10 ppm error on 500 m/z). In the EICs, the peptide ion features are unique given the high precision of the Orbitrap spectrometer utilized when combined with a specified retention time window (10 minutes). From these EICs, all consensus features are quantitated by integration. Integrated peak areas were gathered after a Savitsky-Golay noise filter was applied. Detection of the peak was done independently by using the PeakUtils Python package within the EIC window to account for retention time drift across AS-MS replicates. The smoothed, identified peaks were then integrated numerically using cumulative trapezoids, as this method accounts for abnormal peak shape while also retaining a short computing time.

From the integrated peak areas, two scores were developed to rank and prioritize peptides for their value as ligands: target selectivity and concentration-dependent enrichment (CDE, Figure 3). Target selectivity is a critical property at play in all ligand discovery platforms. While experimental controls and protocols are optimized, the discovery of nonselective or non-specific ligands impedes discovery efforts.[41,42] By comparing the integrated peak areas from experimental replicates, the selectivity of each prospective ligand towards the target protein versus off-target proteins is immediately assessed. As illustrated in Figure 3A, the target selectivity score for a specific protein concentration is determined by the fraction of the total peak area contributed by that protein, assigning a selectivity score to each peptide feature for every protein, with all scores summing to one. A target selective ligand will appear only in the AS-MS samples that contain the target, whereas a nonselective ligand will have a target selectivity equal to the reciprocal of the total number of targets. Thus, selectivity scores differentiate between target-selective and nonselective ligands. With multiple AS-MS replicates, statistical significance of the target selectivity is assigned (SI Section 11.6).

The second score calculated in pyBinder is concentration-dependent enrichment (CDE). CDE was inspired by the connection between concentration dependence in binding interactions and selectivity.[43,44] In pyBinder, CDE measures the change in the integrated intensity of a peptide feature relative to the amount of target protein used in the affinity selection experiment (Figure 3B). To enable this analysis, affinity selections were completed using varying quantities of target-labelled magnetic beads, as well as a negative control with beads lacking the target protein. We calculated the integrated peak areas for each protein loading scenario and assigned a CDE score based on the formula depicted in Figure 3B. The sign and magnitude of the CDE score is reported to gauge the target selectivity of each peptide feature.

Beyond target selectivity, CDE scores can provide insight into the relative ligand binding affinity (apparent dissociation constant, K_D), with theoretical scenarios given assumed K_D values shown in Figure S2. High CDE scores indicate strong peptide enrichment from the affinity selection due to the target protein. Meanwhile, low CDE scores (e.g., near zero) indicate peptide enrichment regardless of target protein concentration, explained by nonspecific binding or poor affinity. Another potential case is a negative CDE score that could indicate that the target protein reduces peptide enrichment, possibly by reducing nonspecific binding.

By utilizing these two scores, peptides are prioritized based on their potential as target-selective ligands. If known, the peptide sequences can delineate structure-activity relationships with respect to the target protein. If unknown, the peptide ion features can be formulated into a targeted list to perform subsequent targeted mass spectrometry. A much larger amount of data could then be revealed, greatly improving the data generation capabilities of AS-MS as a ligand discovery platform.

https://doi.org/10.26434/chemrxiv-2024-lfqjb **ORCID:** https://orcid.org/0000-0002-7242-801X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

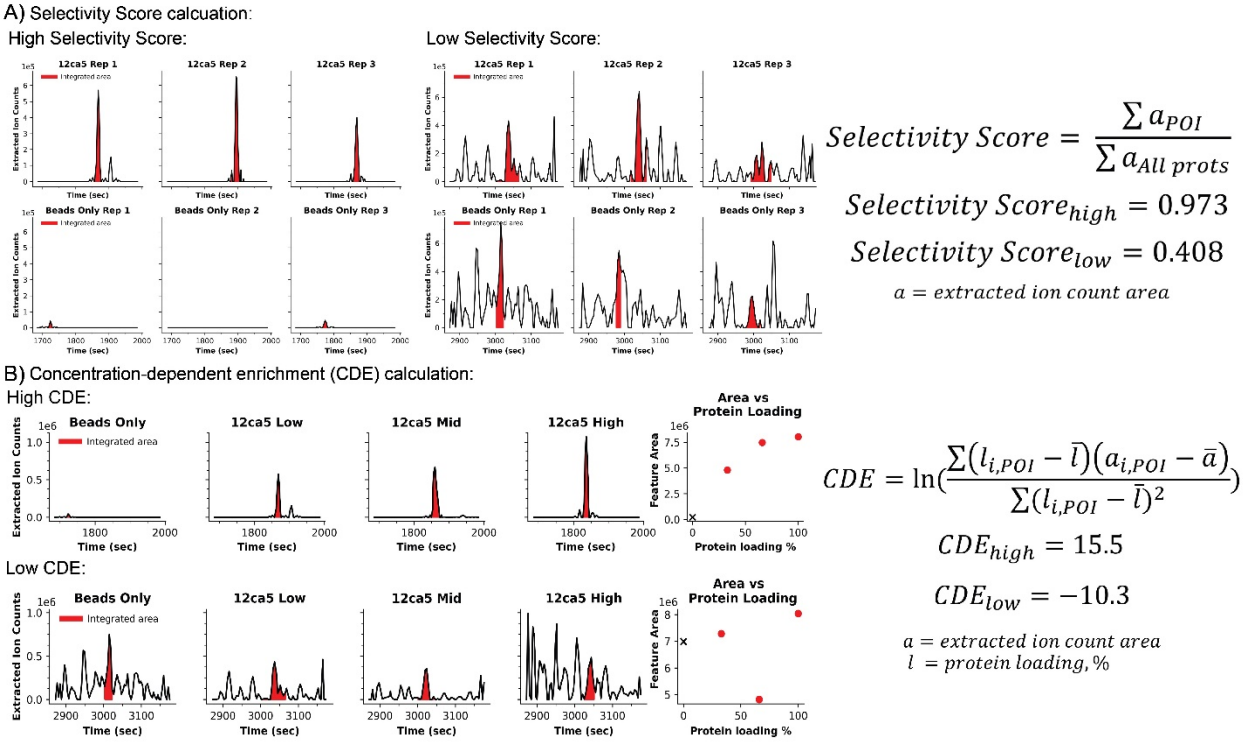

Figure 3. Target selectivity and concentration-dependent enrichment (CDE) scores are used to evaluate peptide features. (A) The selectivity score is calculated by comparing the area for a given feature with respect to a single protein and the total feature area measured across all proteins. A high selectivity score reflects a protein-specific feature, while a selectivity score near the reciprocal of the total number of proteins reflects a nonspecific binding feature. (B) The CDE score is calculated using the extracted feature area across several protein concentrations using the formula shown at the right. A high CDE score shows a strong selection pulldown of the peptide feature even at lower protein concentrations, while a low CDE score shows a lack of relationship between protein concentration and peptide pulldown.

To evaluate the performance of LFQ analysis of AS-MS data by pyBinder, an affinity selection was completed using 12ca5 compared to unlabeled magnetic beads. The anti-hemagglutinin antibody 12ca5 was chosen for its known binding motif, where peptides containing the sequence D**DY(A/S) often exhibit high affinity binding (e.g., K_D < 300 nM).[13,45] A 10^8-membered $X_{12}K$ library was used, where X denotes the set of 20 natural amino acids except cysteine (to exclude disulfide formation) and isoleucine (indistinguishable from leucine by MS). The selection was performed using three different amounts of 12ca5 loaded on the beads to enable CDE score calculations with either 0 (beads only), 55, 110, or 180 pmol of 12ca5 utilized. Selectivity scores were calculated using the beads only control as the off-target protein. After selection, peptide sequencing was performed with the standard intensity-ranked DDA approach, as in the 12ca5/ACE2 campaign. The list of sequenced peptides was filtered to match the library design and peptides containing the 12ca5 binding motif assigned with high confidence were compiled for analysis. This list of motif-containing peptides was then compared to the results from pyBinder for the high-priority peptide features.

Both the selectivity and CDE scores from pyBinder were high for 12ca5 motif-containing peptides, which are expected to have high-affinity, target-selective binding (Figure 4). Independently, the motif-containing peptides were color-coded and visualized for target selectivity and CDE scores (Figure 4A and 4B). While their statistical significance, denoted by -$\log_{10}$(P-value), was less discerning than the scores themselves, the target selectivity and CDE scores clearly indicate the high performance of the motif-containing peptides in the affinity

https://doi.org/10.26434/chemrxiv-2024-lfqjb **ORCID:** https://orcid.org/0000-0002-7242-801X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

selection experiment. Also, as expected, many peptide features were not sequenced (shown in gray) due to the low sequence coverage and low sequencing fidelity. Last, combining the two scores (Figure 4C) presented a high density of motif-containing peptides in the top right quadrant of the graph. Thus, this analysis in pyBinder, rooted in LFQ, demonstrated clear potential to efficiently analyze AS-MS data and distinguish ligands that are expected to be target-selective and high-affinity.

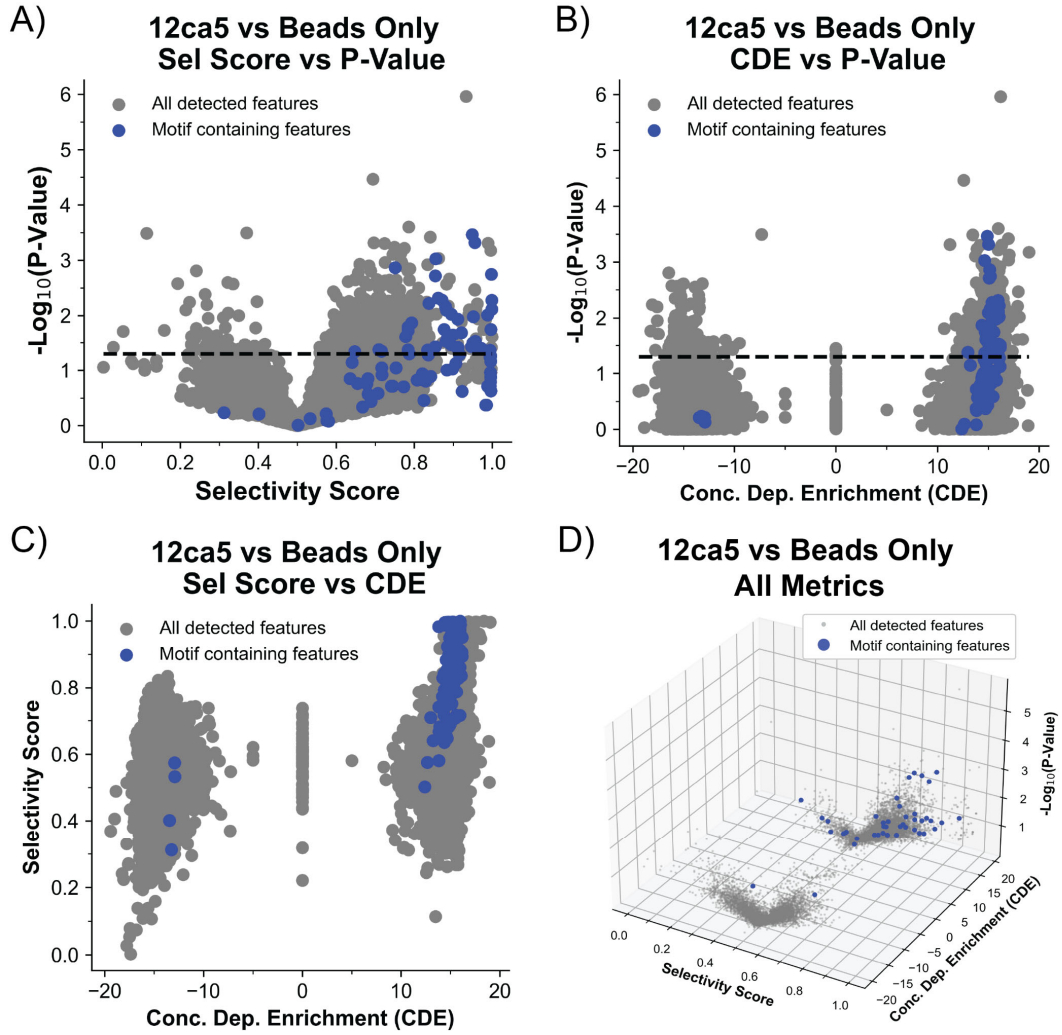

Figure 4. The target selectivity and CDE scores of 12ca5 motif-containing peptides demonstrate the ability of pyBinder to distinguish target-selective, high-affinity peptides. Motif-containing peptides are shown in blue in each graph, with all other detected features are shown in gray. (A) A comparison of the selectivity score with respect to 12ca5 and the statistical significance as shown by the p-value. (B) A comparison of the CDE score and the statistical significance as shown by the p-value. (C) A comparison of the selectivity score and the CDE score. (D) A comparison of selectivity score, CDE score, and p-value.

With 12ca5, we applied pyBinder to AS-MS data collected on a novel target protein, WDR5, using a similar motif-based analysis for validation when selected against the $X_{12}K$ peptide library. WDR5, like 12ca5, also has a known set of motifs associated with ligand binding at its 'WIN' site based on arginine-containing tripeptide sequences (e.g., ART and ARA) at the N-terminus of the peptide.[46,47] From the AS-MS data, target selectivity and CDE were calculated and sequence assignments were gathered from the standard tandem sequencing of the 12ca5 and WDR5 samples. Motif-containing peptide sequences for both 12ca5 and WDR5 assigned from the data (ALC ≥70) were matched back to their respective scores in pyBinder by mass and were plotted

https://doi.org/10.26434/chemrxiv-2024-lfqjb **ORCID:** https://orcid.org/0000-0002-7242-801X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

according to their selectivity scores, CDE scores, and p-values in Figure 5. For this case, the CDE score appeared to be a more effective filter than target selectivity. A range of target selectivity scores were observed across all the motif-containing peptides, suggesting a degree of nonspecific interactions with 12ca5 or possible sample carry-over in the mass spectrometer. Last, the low p-value cutoff ($p < 0.05$) appeared to hinder the prioritization of motif-containing peptides, consistent with the observations from the 12ca5 vs beads experiment in Figure 4A and B. For both cases, these results indicate that the peak detection and integration could potentially be improved to decrease the noise of the peak areas gathered.

Given its potential, target-selective peptide features from pyBinder were used in a second round of mass spectrometry to reveal a larger amount of peptide ligands compared to the standard approach for WDR5 (Figure 5). The output from pyBinder allows the quick prioritization of peptide features observed from the AS-MS experiment using the target selectivity and CDE scores to construct a list of features for tandem sequencing. With the same samples, additional mass spectrometry to the m/z and retention time of promising peptide features was completed. For WDR5, this approach increased the number of ligands discovered (ALC >80) from 2 to 7, or an increase from 3 to 14 ligands if using a slightly relaxed sequencing confidence (ALC >70). Full lists of the identified motif-containing sequences for the untargeted and targeted analysis methods for 12ca5 and WDR5 are given in Tables S3 to S6. This increase of WDR5 motif-containing sequences demonstrates the application of pyBinder to increase the target-selective ligand identification rate and quality of data generated from AS-MS.

https://doi.org/10.26434/chemrxiv-2024-lfqjb **ORCID:** https://orcid.org/0000-0002-7242-801X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

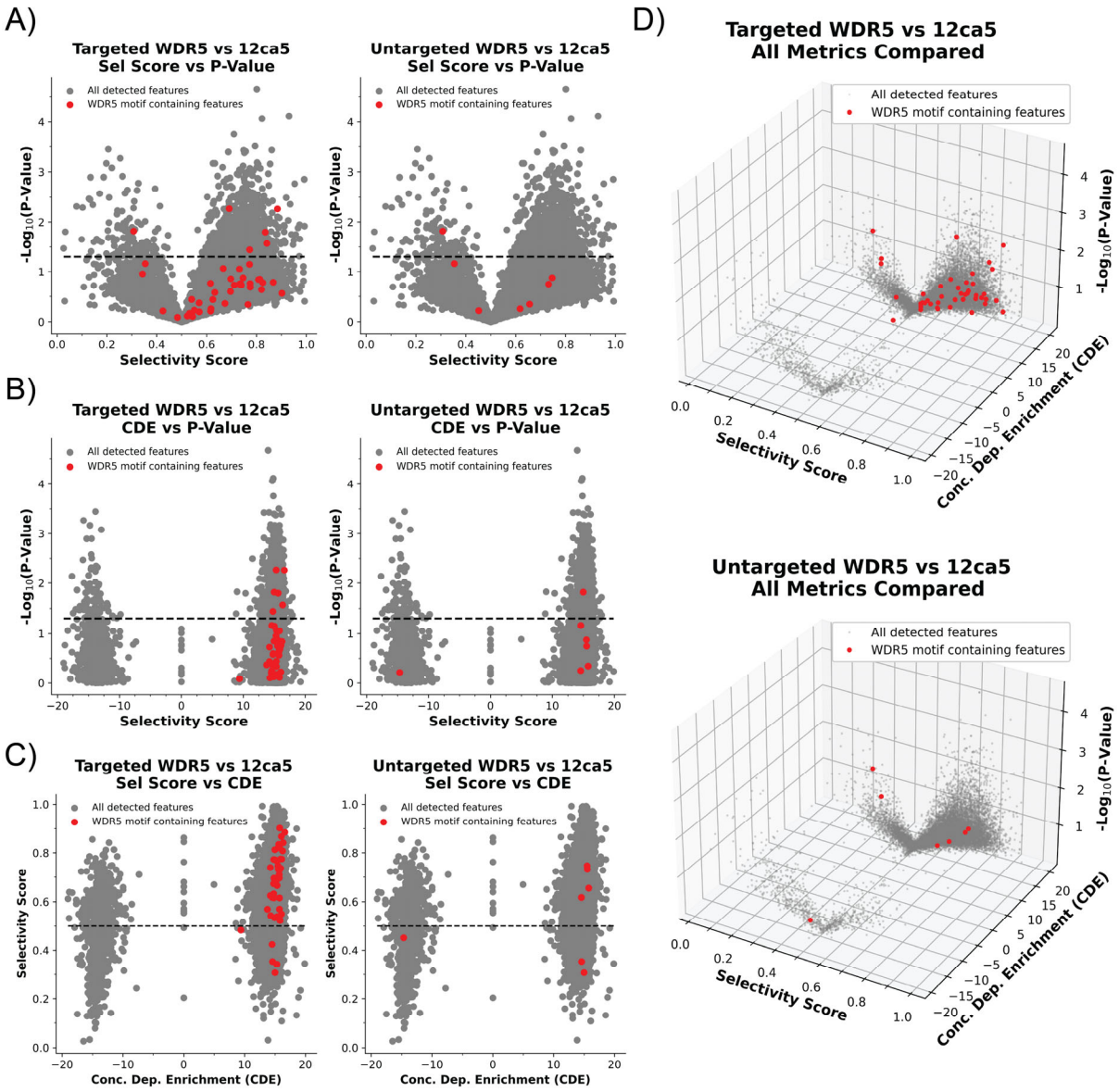

Figure 5. The application of pyBinder in targeted AS-MS increases the discovery rate of peptides containing the WDR5 binding motifs compared to untargeted methods. Plots shown highlight WDR5 motif-containing sequences that were successfully sequenced with high enough confidence, defined as an ALC score ≥70. Gray points reflect extracted features that either were not sequenced or had too low confidence in the sequence assignment. Motif-containing peptides trend towards having high selectivity scores and high CDE scores. Scatterplots comparing relationships between all the scores used are shown, where (A) shows selectivity score against statistical confidence, (B) shows CDE score against statistical confidence, (C) shows selectivity score against CDE score, and (D) shows all three values compared simultaneously. A tolerance value of 0.005 in mass-to-charge ratio was used to match sequence assignments back to features annotated by pyBinder, causing potential double assignments.

https://doi.org/10.26434/chemrxiv-2024-lfqjb **ORCID:** https://orcid.org/0000-0002-7242-801X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

4 Conclusion

We present pyBinder as a workflow to perform LFQ on AS-MS data that introduces two scores to characterize the value of identified peptide features: target selectivity and concentration-dependent enrichment (CDE). Starting from the results gathered from LFQ of AS-MS data, target-selective ligands can be identified without the need for isobaric labeling, stable-isotope labeling, or observation of MS^2-based mass tags. Trends in the two scores were shown to distinguish ligands that were target-selective for two target proteins, 12ca5 and WDR5. Because they are connected to the ligand affinity, CDE scores can be combined with peptide sequence information in machine learning models to discover ligands. However, we did observe that the statistical significance of the two scores was less discerning. Aside from improvements to the data quality, we expect this challenge to be remedied with improvements to the peak detection and integration methods; however, the current method provides sufficiently powerful characterization of the data.

From the two pyBinder scores, a list of prioritized peptide features could be enumerated for successful targeting in subsequent selection rounds to expand the data gathered from AS-MS. Lists of peptide features that exhibit high target selectivity and CDE can be fed back into targeted mass spectrometry methods by their mass-to-charge ratio and retention time extracted from MS^1 data. This approach of targeted mass spectrometry enabled by pyBinder remedies the challenge of high sample complexity and low sequencing coverage by focusing the MS sequencing capacity toward promising ligands. Carried further, the targeting enabled by pyBinder allows the deliberate use of increased mass spectrometer time per peptide to potentially increase sequencing fidelity. Thus, pyBinder appears able to overcome the two bottlenecks that limit AS-MS, sequence coverage and sequence fidelity, originally revealed in our retrospective analysis.

We expect this work to improve the robustness of AS-MS ranging from increasing the number of target-selective ligands discovered to evaluating affinity selection conditions and peptide libraries. We demonstrated the ability of pyBinder to increase the amount of data generated from AS-MS experiments for the purpose of target-selective ligand discovery. pyBinder removes the reliance on sequencing results, which can be poor due to multiple reasons, and instead reports the quality of the AS-MS data using LFQ of MS^1 information. Thus, pyBinder can analyze the general enrichment achieved by the affinity selection and be used to evaluate experimental designs and the suitability of peptide libraries to new targets. We expect pyBinder to improve AS-MS by minimizing the identification rate of nonspecific ligands, improving the ability to establish structure-activity relationships (SAR), and estimate of binding affinity (K_D) directly from ligand discovery experiments.

5 Software Availability

All code used in this work is available at https://github.com/malee97/pyBinder. A Jupyter notebook facilitating the usage of pyBinder is present in the repository and is the primary method of using pyBinder. Additional instructions and required modules are also included.

6 Acknowledgements

The authors thank Harrison Specht for advice and suggestions on the methods developed in this manuscript. Funding for this work was provided by Novo Nordisk A/S. We thank Dr. Thomas E. Nielsen and Dr. Uli Stilz for their helpful discussion in support of our work. Joseph S. Brown acknowledges support from Pharmaceutical Research and Manufacturers of America (PhRMA) Foundation through the Postdoctoral Fellowship in Drug Discovery. Michael A. Lee acknowledges support from the MIT Dean of Science Fellowship.

https://doi.org/10.26434/chemrxiv-2024-lfqjb **ORCID:** https://orcid.org/0000-0002-7242-801X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

7 Conflict of Interest Statement

The authors declare the following competing interests: Bradley L. Pentelute is a co-founder and/or member of the scientific advisory board of several companies focusing on the development of protein and peptide therapeutics.

8 References

1. Zuckermann, R. N., Kerr, J. M., Siani, M. A., Banville, S. C. & Santi, D. V. Identification of highest-affinity ligands by affinity selection from equimolar peptide mixtures generated by robotic synthesis. *Proc. Natl. Acad. Sci. U. S. A.* **89**, 4505–4509 (1992).
2. Kaur, S., Mcguire, L., Tang, D., Dollinger, G. & Huebner2', V. *Affinity Selection and Mass Spectrometry-Based Strategies to Identify Lead Compounds in Combinatorial Libraries*. *Journal of Protein Chemistry* vol. 16 (1997).
3. Prudent, R., Annis, D. A., Dandliker, P. J., Ortholand, J. Y. & Roche, D. Exploring new targets and chemical space with affinity selection-mass spectrometry. *Nat. Rev. Chem. 2020 51* **5**, 62–71 (2020).
4. Josephson, K., Ricardo, A. & Szostak, J. W. mRNA display: from basic principles to macrocycle drug discovery. *Drug Discov. Today* **19**, 388–399 (2014).
5. Wilson, D. S., Keefe, A. D. & Szostak, J. W. The use of mRNA display to select high-affinity protein-binding peptides. *Proc. Natl. Acad. Sci.* **98**, 3750–3755 (2001).
6. Smith, G. P. & Petrenko, V. A. *Phage Display*. https://pubs.acs.org/sharingguidelines (1997).
7. Touti, F., Gates, Z. P., Bandyopdhyay, A., Lautrette, G. & Pentelute, B. L. In-solution enrichment identifies peptide inhibitors of protein–protein interactions. *Nat. Chem. Biol.* **15**, 410–418 (2019).
8. Gates, Z. P. *et al.* Xenoprotein engineering via synthetic libraries. *Proc. Natl. Acad. Sci.* **115**, E5298–E5306 (2018).
9. Silvestri, A. P. *et al.* DNA-Encoded Macrocyclic Peptide Libraries Enable the Discovery of a Neutral MDM2–p53 Inhibitor. *ACS Med. Chem. Lett.* **14**, 820–826 (2023).
10. Garrigou, M. *et al.* Accelerated Identification of Cell Active KRAS Inhibitory Macrocyclic Peptides using Mixture Libraries and Automated Ligand Identification System (ALIS) Technology. *J. Med. Chem.* **65**, 8961–8974 (2022).
11. Weiss, G. A., Watanabe, C. K., Zhong, A., Goddard, A. & Sidhu, S. S. Rapid mapping of protein functional epitopes by combinatorial alanine scanning. *Proc. Natl. Acad. Sci.* **97**, 8950–8954 (2000).
12. Ye, X. *et al.* Binary combinatorial scanning reveals potent poly-alanine-substituted inhibitors of protein-protein interactions. *Commun. Chem.* **5**, 1–10 (2022).
13. Quartararo, A. J. *et al.* Ultra-large chemical libraries for the discovery of high-affinity peptide binders. *Nat. Commun. 2020 111* **11**, 1–11 (2020).
14. Zhang, G. *et al.* Rapid de novo discovery of peptidomimetic affinity reagents for human angiotensin converting enzyme 2. *Commun. Chem. 2022 51* **5**, 1–10 (2022).
15. Pomplun, S. *et al.* De Novo Discovery of High-Affinity Peptide Binders for the SARS-CoV-2 Spike Protein. *ACS Cent. Sci.* **7**, 156–163 (2021).
16. Pomplun, S., Gates, Z. P., Zhang, G., Quartararo, A. J. & Pentelute, B. L. Discovery of Nucleic Acid Binding Molecules from Combinatorial Biohybrid Nucleobase Peptide Libraries. *J. Am. Chem. Soc.* **142**, 19642–19651 (2020).
17. Jin, L. *et al.* A comparative study of evaluating missing value imputation methods in label-free proteomics. *Sci. Rep.* **11**, 1760 (2021).
18. Kong, W., Hui, H. W. H., Peng, H. & Goh, W. W. B. Dealing with missing values in proteomics data. *PROTEOMICS* **22**, 2200092 (2022).

https://doi.org/10.26434/chemrxiv-2024-lfqjb **ORCID:** https://orcid.org/0000-0002-7242-801X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

19. Lazar, C., Gatto, L., Ferro, M., Bruley, C. & Burger, T. Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. *J. Proteome Res.* **15**, 1116–1125 (2016).
20. Liu, M. & Dongre, A. Proper imputation of missing values in proteomics datasets for differential expression analysis. *Brief. Bioinform.* **22**, bbaa112 (2021).
21. Bantscheff, M., Schirle, M., Sweetman, G., Rick, J. & Kuster, B. Quantitative mass spectrometry in proteomics: a critical review. *Anal. Bioanal. Chem.* **389**, 1017–1031 (2007).
22. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
23. Weisser, H. *et al.* An Automated Pipeline for High-Throughput Label-Free Quantitative Proteomics. *J. Proteome Res.* **12**, 1628–1644 (2013).
24. Al Shweiki, M. R. *et al.* Assessment of Label-Free Quantification in Discovery Proteomics and Impact of Technological Factors and Natural Variability of Protein Abundance. *J. Proteome Res.* **16**, 1410–1424 (2017).
25. Wong, J. W. H., Sullivan, M. J. & Cagney, G. Computational methods for the comparative quantification of proteins in label-free LCn-MS experiments. *Brief. Bioinform.* **9**, 156–165 (2008).
26. Ma, B. *et al.* PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **17**, 2337–2342 (2003).
27. Mueller, L. N., Brusniak, M.-Y., Mani, D. R. & Aebersold, R. An Assessment of Software Solutions for the Analysis of Mass Spectrometry Based Quantitative Proteomics Data. *J. Proteome Res.* **7**, 51–61 (2008).
28. Röst, H. L. *et al.* OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat. Methods 2016 139* **13**, 741–748 (2016).
29. Ong, S.-E. & Mann, M. Mass spectrometry–based proteomics turns quantitative. *Nat. Chem. Biol.* **1**, 252–262 (2005).
30. Vinogradov, A. A. *et al.* Library Design-Facilitated High-Throughput Sequencing of Synthetic Peptide Libraries. *ACS Comb. Sci.* **19**, 694–701 (2017).
31. You, Z., Wen, Y., Jiang, K. & Pan, Y. Fragmentation mechanism of product ions from protonated proline-containing tripeptides in electrospray ionization mass spectrometry. *Chin. Sci. Bull.* **57**, 2051–2061 (2012).
32. Paizs, B. & Suhai, S. Fragmentation pathways of protonated peptides. *Mass Spectrom. Rev.* **24**, 508–548 (2005).
33. König, S., Marco, H. G. & Gäde, G. The proline effect and the tryptophan immonium ion assist in de novo sequencing of adipokinetic hormones. *Sci. Rep.* **13**, 10894 (2023).
34. Breci, L. A., Tabb, D. L., Yates, J. R. & Wysocki, V. H. Cleavage N-Terminal to Proline: Analysis of a Database of Peptide Tandem Mass Spectra. *Anal. Chem.* **75**, 1963–1971 (2003).
35. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S. & Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods* **17**, 41–44 (2020).
36. Sinitcyn, P. *et al.* MaxDIA enables library-based and library-free data-independent acquisition proteomics. *Nat. Biotechnol.* **39**, 1563–1573 (2021).
37. Fernández-Costa, C. *et al.* Impact of the Identification Strategy on the Reproducibility of the DDA and DIA Results. *J. Proteome Res.* **19**, 3153–3161 (2020).
38. Röst, H. L., Schmitt, U., Aebersold, R. & Malmström, L. pyOpenMS: A Python-based interface to the OpenMS mass-spectrometry algorithm library. *PROTEOMICS* **14**, 74–77 (2014).
39. Senko, M. W., Beu, S. C. & McLafferty, F. W. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J. Am. Soc. Mass Spectrom.* **6**, 229–233 (1995).
40. Lange, E. *et al.* A geometric approach for the alignment of liquid chromatography—mass spectrometry data. *Bioinformatics* **23**, i273–i281 (2007).
41. Baell, J. B. & Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **53**, 2719–2740 (2010).

https://doi.org/10.26434/chemrxiv-2024-lfqjb **ORCID:** https://orcid.org/0000-0002-7242-801X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

42. Payne, D. J., Gwynn, M. N., Holmes, D. J. & Pompliano, D. L. Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nat. Rev. Drug Discov.* **6**, 29–40 (2007).
43. Stock, L., Hosoume, J. & Treptow, W. Concentration-Dependent Binding of Small Ligands to Multiple Saturable Sites in Membrane Proteins. *Sci. Rep.* **7**, 5734 (2017).
44. Xu, M. *et al.* Concentration-Dependent Enrichment Identifies Primary Protein Targets of Multitarget Bioactive Molecules. *J. Proteome Res.* **22**, 802–811 (2023).
45. Houghten, R. A. *et al.* Generation and use of synthetic peptide combinatorial libraries for basic research and drug discovery. *Nature* **354**, 84–86 (1991).
46. Aho, E. R. *et al.* Displacement of WDR5 from Chromatin by a WIN Site Inhibitor with Picomolar Affinity. *Cell Rep.* **26**, 2916-2928.e13 (2019).
47. Ding, J. *et al.* Discovery of Potent Small-Molecule Inhibitors of WDR5-MYC Interaction. *ACS Chem. Biol.* **18**, 34–40 (2023).

https://doi.org/10.26434/chemrxiv-2024-lfqjb **ORCID:** https://orcid.org/0000-0002-7242-801X Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0