Experimental and Computational Advancements in Peptidomimetic Ligand Discovery

by

Michael Alan Lee

B.S. Chemical Engineering, Brigham Young University

Submitted to the Department of Chemistry in partial fulfillment of the requirements of the degree of

Doctor of Philosophy in Chemistry

at the

Massachusetts Institute of Technology

May 2024

Authored by: ………………………………………………………………………………

Michael Alan Lee
Department of Chemistry
May 8th, 2024

Certified by: ………………………………………………………………………………

Bradley L. Pentelute
Professor of Chemistry
Thesis Supervisor

Accepted by: ………………………………………………………………………………

Adam P. Willard
Professor of Chemistry
Graduate Officer

This doctoral thesis has been examined by a committee of the Department of Chemistry as follows:

Matthew D. Shoulders..……………………………………………………………..
Thesis Committee Chair
Professor of Chemistry

Bradley L. Pentelute………………………………………………………………
Thesis Supervisor
Professor of Chemistry

Alex K. Shalek……………………………………………………………………..
Thesis Committee Member
J. W. Kieckhefer Professor of Chemistry

Experimental and Computational Advancements in Peptidomimetic Ligand Discovery

by

Michael Alan Lee

Submitted to the Department of Chemistry on May 8th, 2024 in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Chemistry

Abstract

The usage of peptides as therapeutics is a growing area of interest within the pharmaceutical industry for the facilitation of protein-protein interactions (PPIs). Peptides inhabit a unique therapeutic space because of their high levels of chemical customization balanced with their potential for high specificity due to a wide variety of potential structures. At the same time, discovery tools for finding peptides that modify PPIs have evolved, including advances in affinity selection techniques and combinatorial chemistry. Specifically, the usage of solid phase peptide synthesis for split-and-pool chemistry allows for rapid access to highly diverse (>10^8 total sequences) compound libraries for use in ligand discovery. A primary technique for *in vitro* ligand discovery is affinity selection-mass spectrometry (AS-MS), which utilizes tandem mass spectrometry to decode complex mixtures of peptide ligands pulled down from a peptide library through affinity selection. This approach provides unique advantages due to the high levels of chemical customization that can be performed on synthetic peptide libraries, including the incorporation of unnatural amino acids or the modification of library structure through macrocyclization.

This thesis will focus on the development of experimental and computational tools to analyze affinity selection datasets more efficiently and thoroughly. We demonstrate the synthesis of macrocyclic peptide libraries that increases the diversity of synthetic macrocyclic libraries while utilizing accessible, efficient chemistry for cyclization. These libraries are then used for the discovery of novel ligands to two proteins. Structure activity relationships are established for one of these ligands and its affinity is matured through the usage of focused libraries containing a variety of unnatural amino acids. Additionally, we investigate a variety of resins used for solid phase peptide synthesis, particularly in the synthesis of small domain proteins or difficult peptide sequences.

Because of the high amounts of peptides synthesized and pulled down by AS-MS experiments, efficient computational methods are crucial for effective ligand discovery efforts. Here, we discuss two methods of expanding data analysis, first by a sequence-independent enrichment quantification. AS-MS experiments operate using the decoded peptide sequence from tandem MS/MS data to nominate potential hit peptides, but that process depends on the efficient fragmentation of a target peptide and the quality of the subsequent MS^2 spectrum. We utilize techniques to identify putative hits through the comparison of peptide enrichment based only off mass-to-charge ratio without an assigned sequence, allowing for label free MS^1 quantification. The second method utilizes machine learning techniques to rationalize trends in successfully sequenced peptide sequences from AS-MS experiments with respect to target proteins. This approach allows for the creation of a ligand “sequence space”, which allows for the incorporation of unnatural amino acids in ligand discovery.

Overall, this thesis presents a variety of methods to enhance the scope of peptide-based drug discovery. We anticipate this work to accelerate the process of drug discovery through a diversification of peptide structure combined with more powerful computational analytics.

Thesis supervisor: Bradley L. Pentelute

Professor of Chemistry

Acknowledgements

I'd like to thank Professor Brad Pentelute for allowing me a very enriching and engaging experience in graduate school. I joined his lab having very little background in biology, but he still excitedly took me into his group and helped me learn the necessary skills and acquire the right mindset to have success in chemical biology. Brad's emphasis on learning allowed me to gain experience in many different aspects, ranging from recombinant protein expression to instrument maintenance to machine learning. I'd also like to thank Dr. Andrei Loas for making sure the group is running on the day-to-day and providing valuable feedback on manuscripts and presentations. Thank you to my committee members Professor Matt Shoulders and Professor Alex Shalek for challenging my scientific thinking and helping me become a more well-rounded scientist.

Thank you to all the Pentelute lab members for making the past four years a unique and enjoyable experience. Special thanks to the Novo team, Dr. Joe Brown, Dr. Sarah Antilla, Dr. Tom Wood, Dr. Genwei Zhang, Dr. Chengxi Li, and Roman Misteli. To Joe especially, a huge thank you for being an incredible mentor and providing constant guidance and assurance; he has had an enormous hand in refining my abilities in the lab. Another huge thank you to Dr. Charlie Farquhar for their incredible empathy and patience. Their support during my graduate school experience has been a cornerstone of stability during chaotic times. I would also like to thank the many friends I've made in the lab, including Dr. Wayne Vuong, Gha Young Lee, Dr. Jeff Wong, Dr. Ed Miller, Dr. Amanda Cowfer, Dr. Azin Saebi, Dr. Corey Johnson, Sydney Kelly, Dr. Nate Dow, Dr. Alex Callahan, Dr. Jacob Rodriguez, Dr. Dio Dieppa-Matos, Dr. Katsushi Kitahara, Dr. Hiro Takeuchi, Professor Kohei Sato, Yehlin Cho and everyone I've had the pleasure to work with. Also, shoutouts to 7/11 for providing a concerning amount of office snacks and being a great distraction for when I'm having a bad day.

I'd also like to thank Professor Steven Castle at BYU for being the start of my journey in chemistry. It was in his group that I discovered how much I enjoyed doing lab work and research, and it was my time in his group that ultimately convinced me to pursue chemistry. Thank you to Dr. Concordia Lo, Dr. Jatinder Singh, Stephen White, D.D.S, Dr. Daniel Joaquin, Dr. Diego Moyá, Dr. Alex Ramos, Dr. Ankur Jalan, Austin LeSuer, and Taylor Talentino for many fun days in the lab as well as important lessons in chemistry. Also, a special thanks to Jennah Mumford – working on p chem homework together not only helped grow my appreciation for chemistry but also gained me an invaluable friend.

I also thank the Longfellow Park crew for their support during my time in Boston. Knowing I had a friend from undergraduate in the legendary Darren Miller move to Boston at the same time I did was invaluable, especially since this was during the height of COVID lockdowns where I didn't have much opportunity for social interaction being locked in my sad, crooked box in Back Bay thousands of miles away from anyone else I knew. I'm grateful that Darren introduced me to

many of my friends at LP, even if he thought that I wouldn't get along with some of them. Thank you to Trish Franks, Amanda Strong, Keith Tambe, Dylan Ottney, Sarah Taggart, Emily Loveland, Rowan Cheney, Kristen Vellinga, Eric Moss, Liam Tan, Edz Cabral, Caleb Lindgren and everyone else from the ward that makes it such a welcoming and loving place to be. All the movie nights, theme parties, and game nights provided a well-needed respite from the doldrums and stress of graduate school. In addition to Darren, I'd also like to thank my friends from BYU, including Smoom, Jack Wright, Tommy Doxey, Jordan Johnson, Dave Christenson, THE Zach Gormley, Gibson Ainge, Nick Clarke, Trevor Harmon, Aaron Kim, Shean Gass, Connor Finuf, Davy Clarke, TJ Knotts, and Dan Brown for many fun game nights, especially during lockdown when that was about as much social interaction as you were legally allowed to get.

I want to thank Parin Rau and Chris Harrison for over a decade of my closest friendship. They always have a way to put a smile on my face, usually through something incredibly stupid. Thank you also to Trever Speck, Emily Hanley, Kyle Reynolds, and Morena Morales for always being great friends.

Finally, I'd like to thank my family. To my parents, thank you for your unconditional love and support throughout my entire life. They have always been an example of hard work, integrity, and charity ever since I was a baby. Thank you to my brothers for putting up with your baby brother and always being role models for me to look up to. Thank you to Nei Nei and Yeh Yeh for emphasizing the importance of education to me and supporting me greatly financially through this journey. Thank you to all my family for the warm memories and loving support.

Table of Contents

List of Figures

List of Tables

1. Background and Overview

1.1. Peptides as therapeutics

Peptides have emerged as a powerful force in facilitating protein-protein interactions (PPIs) within the pharmaceutical industry for their ability to balance many crucial properties for drug efficacy.[1,2] Peptides occupy a therapeutic space in between that of small molecules and large protein biologics, where peptides are highly amenable to precise and diverse chemical modifications but are able to utilize the specificity inherent to larger biological structures like in PPIs.[3] However, there have historically been drawbacks to peptide-based therapeutics, such as poor proteolytic stability, cell permeability, and oral bioavailability that have led to peptides trailing in the market share of the global pharmaceutical market behind biologics and small molecule drugs.[4] To address this, chemical modifications of potential peptide therapeutics have been essential in optimizing the pharmacological properties of a candidate, such as macrocyclization to increase cell permeability or incorporation of unnatural amino acids to enhance target specificity and resistance to proteolysis.[5–8] However, the breadth of potential structures and chemical moieties utilized in drug discovery campaigns still has potential for expansion.

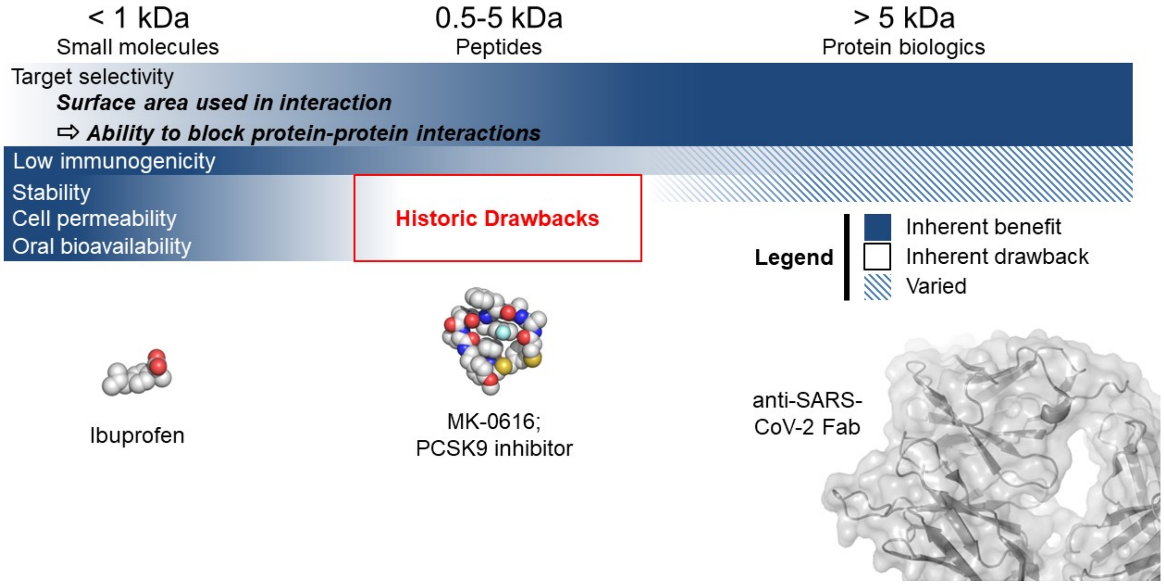


***Figure 1.1**: Peptides occupy a promising therapeutic "middle space". Small molecule drugs often show high cell permeability and oral bioavailability due to their size, while larger protein biologics can leverage superior target selectivity. Peptides hold the potential to bridge the gap between small molecules and biologics to leverage the strengths of both. Examples of each class of drug are shown.*

1.2. Synthesis of peptides

Peptide production is highly amenable to a variety of modifications through the usage of solid phase peptide synthesis (SPPS).[9] Peptide sequences up to 200 amino acids in length can be readily accessed through automated fast-flow synthesizers.[10–12] The peptide chain is anchored to a solid PEG-polystyrene resin, where all reactions are performed heterogeneously with the resin being washed thoroughly with solvent between steps.[9,13] The development of the 9-fluorenylmethyloxycarbonyl (Fmoc)/*tert*-butyl protecting group strategy allowed for efficient coupling of amino acids to the growing peptide chain which minimizing side reactivity with the amino acid side chains.[14–17] In this strategy, the side chains are masked using acid labile protecting groups, while the N-termini of the amino acid monomers are protected with the base labile Fmoc group, allowing for cyclic amino acid coupling and Fmoc deprotection to build the peptide chain.

SPPS methods also enable a variety of modifications that would be difficult for biologics produced using standard biological methods, such as the incorporation of unnatural amino acids[18,19], macrocyclization,[20,21] or chemical stapling.[22,23] A major example of the power of this chemical flexibility is the recent development of many Glucagon-Like Peptide-1 receptor agonists, such as dulaglutide (brand name Trulicity) and semaglutide (brand name Ozempic), which feature both the incorporation of unnatural amino acids and the addition of a fatty acid chain to a central lysine residue.[24,25] These techniques open a large potential for applications of peptides as therapeutics.

Additionally, the attachment of the peptide chain to a solid support during synthesis enables powerful combinatorial chemistry for the generation of large peptide libraries.[26–28] Resin can be split into individual aliquots, reacted with a variety of different reagents, then be pooled to rapidly synthesize peptide libraries (see Figure 1.2). This split-and-pool synthesis method is key for the *in vitro* generation of compound libraries and has been used to generate libraries containing >10^9 peptide sequences.[29] This method is also amenable to the incorporation of a variety of unnatural amino acids due to its chemical approach, whereas recombinantly produced biologics require complex cellular engineering.[18,30,31]

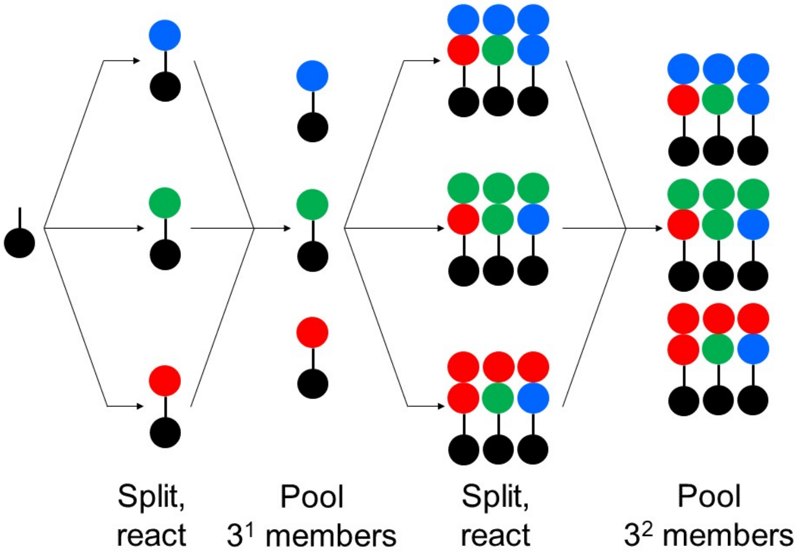


Figure 1.2: *Split-and-pool synthesis enables access to high diversity compound libraries. Peptide resin (represented in black) can be split into multiple aliquots, coupled with a new amino acid, and pooled together iteratively to exponentially increase the total number of peptide sequences.*

1.3. Affinity selection-mass spectrometry

There are a variety of techniques used for discovering peptidomimetic and small molecule ligands to desired protein targets, including affinity selection-mass spectrometry (AS-MS), genetically encoded methods like phage display,[32] mRNA display[33], and DNA encoded libraries,[34] as well as many others.[35–37] These techniques utilize large libraries of compounds (ranging from 10^4 to 10^{12} total variants[38,39] depending on platform) to efficiently survey potential hits. The total number of sequences present during an affinity selection is critical to hit identification, where a higher number of variants allows for a greater coverage of the sequence space.[29,40,41] A protein or protein complex of interest is assayed against this library of compounds, where the selection process will pull down ligands of interest from the compound library, referred to as putative hits. The identity of the hit, the sequence of the peptide, is then identified and validated through a biophysical assay or protein functional assay.[42–44] For AS-MS, the decoding process is done through tandem mass spectrometry, while genetically encoded methods utilize next-generation sequencing.[45]

Tandem mass spectrometry is a key method for hit identification in AS-MS. Utilizing a data dependent acquisition (DDA) protocol, the mass spectrometer will identify peptides based on their characteristic isotopic pattern and accumulate the highest intensity peptide ion for sequencing.[46–48] The sequencing process can be performed through a handful of methods, commonly through higher-energy collision dissociation (HCD) and electron transfer dissociation (ETD).[49–51] The HCD process accelerates peptide ions within the mass spectrometer using electric potential until the ion collides with a neutral molecule, often a noble gas, causing the kinetic energy to be rapidly converted to internal energy and fragmenting the peptide at the amide C-N bond.[52] ETD follows similar principles, but first transfers an electron to the ion of interest from a donor reagent to generate an unstable radical cation, which will then undergo fragmentation at the α-C-N bond.[53,54] These fragmentation processes break a peptide into small fragments of unique masses, allowing for the deconvolution of the sequence.

A major challenge in the drug discovery process is the effective processing of sequencing data. Samples generating from affinity selections are often highly complex mixtures of peptides, residual protein, and various small molecules like buffering agents and detergents. Additionally, DDA methods choose ion candidates for fragmentation in real-time, meaning that samples that are too complex will not fragment every peptide, leading to a data completeness problem. This is further compounded by the reliance on high fidelity sequencing for downstream data analysis, where peptides that are recalcitrant to fragmentation will give poor data regardless of abundance, resulting in missed putative hits. While AS-MS has a history of discovering high affinity ligands,[55–57] data analysis methods that work independently of sequencing (i.e. at the MS^1 level) are desired for a deeper analysis of affinity selection data. Comparing relative abundances of notable mass-to-charge ratios identified between target protein and off-target control samples can generate a list of focused ions for selective sequencing, optimizing spectrometer time while also giving longer observation times per ion.

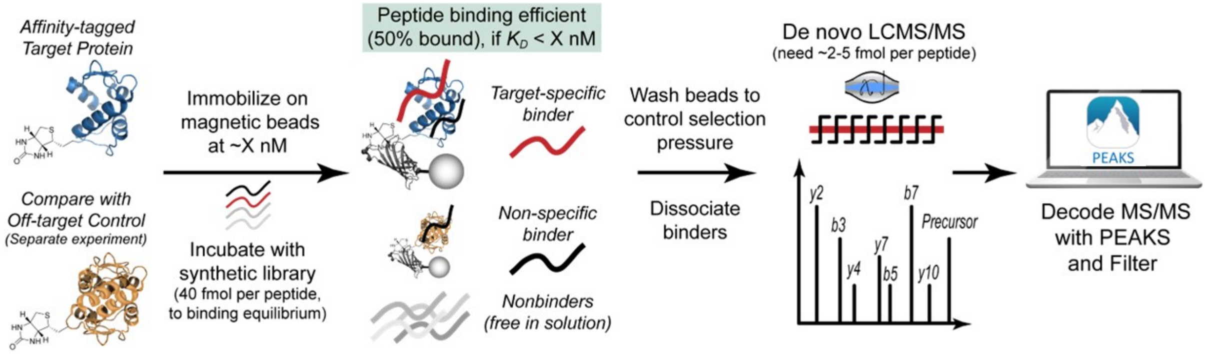


Figure 1.3: *AS-MS enables isolation of high affinity binders to a protein of interest. A target protein and off-target control are immobilized onto a magnetic bead and incubated with a synthetic peptide library. The beads are then washed, and the bound peptides eluted from the protein before sequencing using tandem mass spectrometry (MS/MS).*

One common method of combating the data completeness problem is through usage of data independent acquisition (DIA). While DDA isolates and analyzes individual peptide features, DIA instead isolates ranges of mass-to-charge ratio and fragments all isolated peptides.[58,59] This method increases the throughput of sequencing, but MS^2 spectra now show multiple peptides of instead of a single peptide sequence, making identification more difficult. The difficulty is remedied through usage of database matching; this method involves specifying a set of peptides that could be contained in the sample to which sequencing software simulates what each peptide’s MS^2 spectrum should look like, allowing for identification from the experimental spectra. This is in contrast to *de novo* sequencing, which relies solely on the information gained from the experimental MS^2 spectra.[60–62] While powerful, DIA and database matching approaches require prior knowledge of the sample, which limits its use in AS-MS. If a peptide library is synthesized combinatorially, this makes the total list of sequences exponentially large. For example, a library of X_8K (X = any canonical amino acid except Cys or Ile), then the total possible sequences is about 10^{10}, which is more than the age of the earth.[63] This size of database is not feasible for use, making *de novo* sequencing necessary for high diversity peptide libraries. Additionally, this means that not every sequence will be present from the theoretical set, meaning the entirety of the library would also need to be identified experimentally. These factors make database matching and DIA approaches difficult for high diversity synthetic peptide libraries.

1.4. Machine learning in drug discovery

Machine learning techniques are positioned to transform drug discovery.[64–68] There are two primary categories of machine learning: unsupervised and supervised methods. Unsupervised methods deduce underlying trends in unlabeled datasets and are often useful for either exploratory purposes or for simplification of the outputs from a more complex supervised model.[69–71] Clustering is commonly employed to group similar peptide sequences for the identification of important physicochemical properties or sequence motifs.[72] Supervised methods that take labeled training data (e.g. binding affinity, protein activity) and interpolate or extrapolate properties to allow for prediction.[73–75] Machine learning models have already been used for various steps of the drug development process, including hit discovery and activity prediction.[76–78]

Both machine learning methods require input data to be encoded into a mathematical format, which will significantly influence the capabilities of the model. The choice of encoding format can be based on a variety of properties, such as amino acid sequence[79] or chemical substructures.[80] The power of machine learning can be leveraged with the incorporation of unnatural amino acids from AS-MS datasets. Unnatural amino acids can utilize chemical moieties not found in the canonical twenty monomers, opening new chemical space for the design of peptide therapeutics. However, effective machine learning models require large amounts of data, especially the peptide sequence for the effective encoding of peptide properties. This presents an open challenge, where more powerful AS-MS data analysis methods will allow for the training of more powerful and accurate machine learning models, driving the field of drug discovery.

1.5. Thesis overview

This thesis presents work done to expand the scope of current drug discovery technology with respect to AS-MS and synthetic peptide libraries. **Chapter 1** reviews the emerging importance of peptides as therapeutics as well as methods for peptide-based drug discovery. **Chapter 2** outlines a method for the synthesis of high diversity macrocyclic peptide libraries and their use in novel ligand discovery to two proteins. **Chapter 3** surveys various resin options for SPPS in the synthesis of difficult peptide sequences. **Chapter 4** discusses an MS^1-based analysis of AS-MS data for a quantitative evaluation of putative hits. **Chapter 5** describes an unsupervised machine learning approach for the design of peptide ligands containing unnatural amino acids. Finally, the **Appendix** outlines a supervised machine learning approach for the design of peptide ligands.

1.6. References

1. Zompra, A. A.; Galanis, A. S.; Werbitzky, O.; Albericio, F. Manufacturing Peptides as Active Pharmaceutical Ingredients. *Future Med. Chem.* **2009**, *1* (2), 361–377. https://doi.org/10.4155/fmc.09.23.
2. Wieland, T.; Bodanszky, M. *The World of Peptides: A Brief History of Peptide Chemistry*; Springer: Berlin, Heidelberg, 1991. https://doi.org/10.1007/978-3-642-75850-8.
3. Wang, L.; Wang, N.; Zhang, W.; Cheng, X.; Yan, Z.; Shao, G.; Wang, X.; Wang, R.; Fu, C. Therapeutic Peptides: Current Applications and Future Directions. *Signal Transduct. Target. Ther.* **2022**, *7* (1), 1–27. https://doi.org/10.1038/s41392-022-00904-4.
4. Muttenthaler, M.; King, G. F.; Adams, D. J.; Alewood, P. F. Trends in Peptide Drug Discovery. *Nat. Rev. Drug Discov.* **2021**, *20* (4), 309–325. https://doi.org/10.1038/s41573-020-00135-8.
5. Goto, Y.; Suga, H. The RaPID Platform for the Discovery of Pseudo-Natural Macrocyclic Peptides. *Acc. Chem. Res.* **2021**, *54* (18), 3604–3617. https://doi.org/10.1021/acs.accounts.1c00391.
6. Vinogradov, A. A.; Yin, Y.; Suga, H. Macrocyclic Peptides as Drug Candidates: Recent Progress and Remaining Challenges. *J. Am. Chem. Soc.* **2019**, *141* (10), 4167–4181. https://doi.org/10.1021/JACS.8B13178/SUPPL_FILE/JA8B13178_SI_001.PDF
7. Ding, Y.; Ting, J. P.; Liu, J.; Al-Azzam, S.; Pandya, P.; Afshar, S. Impact of Non-Proteinogenic Amino Acids in the Discovery and Development of Peptide Therapeutics. *Amino Acids* **2020**, *52* (9), 1207–1226. https://doi.org/10.1007/s00726-020-02890-9.
8. Rossino, G.; Marchese, E.; Galli, G.; Verde, F.; Finizio, M.; Serra, M.; Linciano, P.; Collina, S. Peptides as Therapeutic Agents: Challenges and Opportunities in the Green Transition Era. *Molecules* **2023**, *28* (20), 7165. https://doi.org/10.3390/molecules28207165.
9. Merrifield, R. B. Solid Phase Peptide Synthesis. I. The Synthesis of a Tetrapeptide. *J. Am. Chem. Soc.* **1963**, *85* (14), 2149–2154. https://doi.org/10.1021/ja00897a025.
10. Hartrampf, N.; Saebi, A.; Poskus, M.; Gates, Z. P.; Callahan, A. J.; Cowfer, A. E.; Hanna, S.; Antilla, S.; Schissel, C. K.; Quartararo, A. J.; Ye, X.; Mijalis, A. J.; Simon, M. D.; Loas, A.; Liu, S.; Jessen, C.; Nielsen, T. E.; Pentelute, B. L. Synthesis of Proteins by Automated Flow Chemistry. *Science* **2020**, *368* (6494), 980–987. https://doi.org/10.1126/SCIENCE.ABB2491/SUPPL_FILE/ABB2491_MDAR_REPRODUCIBILITY_CHECKLIST.PDF.
11. Saebi, A.; Brown, J. S.; Marando, V. M.; Hartrampf, N.; Chumbler, N. M.; Hanna, S.; Poskus, M.; Loas, A.; Kiessling, L. L.; Hung, D. T.; Pentelute, B. L. Rapid Single-Shot Synthesis of the 214 Amino Acid-Long N-Terminal Domain of Pyocin S2. *ACS Chem. Biol.* **2023**, *18* (3), 518–527. https://doi.org/10.1021/acschembio.2c00862.

12. Callahan, A. J.; Gandhesiri, S.; Travaline, T. L.; Reja, R. M.; Lozano Salazar, L.; Hanna, S.; Lee, Y.-C.; Li, K.; Tokareva, O. S.; Swiecicki, J.-M.; Loas, A.; Verdine, G. L.; McGee, J. H.; Pentelute, B. L. Mirror-Image Ligand Discovery Enabled by Single-Shot Fast-Flow Synthesis of D-Proteins. *Nat. Commun.* **2024**, *15* (1), 1813. https://doi.org/10.1038/s41467-024-45634-z.
13. Moss, J. A. Guide for Resin and Linker Selection in Solid-Phase Peptide Synthesis. *Curr. Protoc. Protein Sci.* **2005**, *40* (1), 18.7.1-18.7.19. https://doi.org/10.1002/0471140864.ps1807s40.
14. Carpino, L. A.; Han, G. Y. 9-Fluorenylmethoxycarbonyl Amino-Protecting Group. *J. Org. Chem.* **1972**, *37* (22), 3404–3409. https://doi.org/10.1021/jo00795a005.
15. Bodanszky, M.; Deshmane, S. S.; Martinez, J. Side Reactions in Peptide Synthesis. 11. Possible Removal of the 9-Fluorenylmethyloxycarbonyl Group by the Amino Components during Coupling. *J. Org. Chem.* **1979**, *44* (10), 1622–1625. https://doi.org/10.1021/jo01324a008.
16. Behrendt, R.; White, P.; Offer, J. Advances in Fmoc Solid-phase Peptide Synthesis. *J. Pept. Sci.* **2016**, *22* (1), 4. https://doi.org/10.1002/psc.2836.
17. Musaimi, O. A.; Torre, B. G. de la; Albericio, F. Greening Fmoc/tBu Solid-Phase Peptide Synthesis. *Green Chem.* **2020**, *22* (4), 996–1018. https://doi.org/10.1039/C9GC03982A.
18. Ovaa, H.; wals, kim. Unnatural Amino Acid Incorporation in E. Coli: Current and Future Applications in the Design of Therapeutic Proteins. *Front. Chem.* **2014**, *2*.
19. Wang, X.; Yang, X.; Wang, Q.; Meng, D. Unnatural Amino Acids: Promising Implications for the Development of New Antimicrobial Peptides. *Crit. Rev. Microbiol.* **2023**, *49* (2), 231–255. https://doi.org/10.1080/1040841X.2022.2047008.
20. Aimetti, A. A.; Shoemaker, R. K.; Lin, C.-C.; Anseth, K. S. On-Resin Peptide Macrocyclization Using Thiol–Ene Click Chemistry. *Chem. Commun.* **2010**, *46* (23), 4061–4063. https://doi.org/10.1039/C001375G.
21. Fotsch, C.; Kumaravel, G.; Sharma, S. K.; Wu, A. D.; Gounarides, J. S.; Nirmala, N. R.; Petter, R. C. On-Resin Macrocyclization of Peptides via Intramolecular SnAr Reactions. *Bioorg. Med. Chem. Lett.* **1999**, *9* (15), 2125–2130. https://doi.org/10.1016/S0960-894X(99)00356-X.
22. Moiola, M.; Memeo, M. G.; Quadrelli, P. Stapled Peptides—A Useful Improvement for Peptide-Based Drugs. *Molecules* **2019**, *24* (20), 3654. https://doi.org/10.3390/molecules24203654.
23. Spokoyny, A. M.; Zou, Y.; Ling, J. J.; Yu, H.; Lin, Y.-S.; Pentelute, B. L. A Perfluoroaryl-Cysteine SNAr Chemistry Approach to Unprotected Peptide Stapling. *J. Am. Chem. Soc.* **2013**, *135* (16), 5946–5949. https://doi.org/10.1021/ja400119t.
24. Chavda, V. P.; Ajabiya, J.; Teli, D.; Bojarska, J.; Apostolopoulos, V. Tirzepatide, a New Era of Dual-Targeted Treatment for Diabetes and Obesity: A Mini-

Review. *Molecules* **2022**, *27* (13), 4315. https://doi.org/10.3390/molecules27134315.

25. Mahapatra, M. K.; Karuppasamy, M.; Sahoo, B. M. Semaglutide, a Glucagon like Peptide-1 Receptor Agonist with Cardiovascular Benefits for Management of Type 2 Diabetes. *Rev. Endocr. Metab. Disord.* **2022**, *23* (3), 521–539. https://doi.org/10.1007/s11154-021-09699-1.
26. FURKA, Á.; SEBESTYÉN, F.; ASGEDOM, M.; DIBÓ, G. General Method for Rapid Synthesis of Multicomponent Peptide Mixtures. *Int. J. Pept. Protein Res.* **1991**, *37* (6), 487–493. https://doi.org/10.1111/J.1399-3011.1991.TB00765.X.
27. Shin, D.-S.; Kim, D.-H.; Chung, W.-J.; Lee, Y.-S. Combinatorial Solid Phase Peptide Synthesis and Bioassays. *J. Biochem. Mol. Biol.* **2005**, *38* (5), 517–525. https://doi.org/10.5483/bmbrep.2005.38.5.517.
28. Gao, Y.; Kodadek, T. Split-and-Pool Synthesis and Characterization of Peptide Tertiary Amide Library. *J. Vis. Exp. JoVE* **2014**, No. 88, 51299. https://doi.org/10.3791/51299.
29. Quartararo, A. J.; Gates, Z. P.; Somsen, B. A.; Hartrampf, N.; Ye, X.; Shimada, A.; Kajihara, Y.; Ottmann, C.; Pentelute, B. L. Ultra-Large Chemical Libraries for the Discovery of High-Affinity Peptide Binders. *Nat. Commun. 2020 111* **2020**, *11* (1), 1–11. https://doi.org/10.1038/s41467-020-16920-3.
30. Wang, Q.; Wang, L. Genetic Incorporation of Unnatural Amino Acids into Proteins in Yeast. *Methods Mol. Biol. Clifton NJ* **2012**, *794*, 199–213. https://doi.org/10.1007/978-1-61779-331-8_12.
31. Adhikari, A.; Bhattarai, B. R.; Aryal, A.; Thapa, N.; KC, P.; Adhikari, A.; Maharjan, S.; Chanda, P. B.; Regmi, B. P.; Parajuli, N. Reprogramming Natural Proteins Using Unnatural Amino Acids. *RSC Adv. 11* (60), 38126–38145. https://doi.org/10.1039/d1ra07028b.
32. Smith, G. P.; Petrenko, V. A. *Phage Display*; 1997. https://pubs.acs.org/sharingguidelines.
33. Roberts, R. W.; Szostak, J. W. RNA-Peptide Fusions for the in Vitro Selection of Peptides and Proteins. *Proc. Natl. Acad. Sci. U. S. A.* **1997**, *94* (23), 12297–12302.
34. Brenner, S.; Lerner, R. A. Encoded Combinatorial Chemistry. *Proc. Natl. Acad. Sci.* **1992**, *89* (12), 5381–5383. https://doi.org/10.1073/pnas.89.12.5381.
35. McMahon, C.; Baier, A. S.; Pascolutti, R.; Wegrecki, M.; Zheng, S.; Ong, J. X.; Erlandson, S. C.; Hilger, D.; Rasmussen, S. G. F.; Ring, A. M.; Manglik, A.; Kruse, A. C. Yeast Surface Display Platform for Rapid Discovery of Conformationally Selective Nanobodies. *Nat. Struct. Mol. Biol.* **2018**, *25* (3), 289–296. https://doi.org/10.1038/s41594-018-0028-6.
36. Clark, L. A.; Boriack-Sjodin, P. A.; Eldredge, J.; Fitch, C.; Friedman, B.; Hanf, K. J. M.; Jarpe, M.; Liparoto, S. F.; Li, Y.; Lugovskoy, A.; Miller, S.; Rushe, M.; Sherman, W.; Simon, K.; Van Vlijmen, H. Affinity Enhancement of an in Vivo Matured Therapeutic Antibody Using Structure-Based Computational Design. *Protein Sci. Publ. Protein Soc.* **2006**, *15* (5), 949–960. https://doi.org/10.1110/ps.052030506.

37. Laustsen, A. H.; Greiff, V.; Karatt-Vellatt, A.; Muyldermans, S.; Jenkins, T. P. Animal Immunization, *in Vitro* Display Technologies, and Machine Learning for Antibody Discovery. *Trends Biotechnol.* **2021**, *39* (12), 1263–1273. https://doi.org/10.1016/j.tibtech.2021.03.003.
38. Gold, L. mRNA Display: Diversity Matters during in Vitro Selection. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98* (9), 4825–4826. https://doi.org/10.1073/pnas.091101698.
39. Ponsel, D.; Neugebauer, J.; Ladetzki-Baehs, K.; Tissot, K. High Affinity, Developability and Functional Size: The Holy Grail of Combinatorial Antibody Library Generation. *Molecules* **2011**, *16* (5), 3675–3700. https://doi.org/10.3390/molecules16053675.
40. Newton, M. S.; Cabezas, Y.; Seelig, B. Advantages of mRNA Display. *ACS Synth. Biol.* **2020**, *9* (2), 181–190. https://doi.org/10.1021/acssynbio.9b00419.
41. Derda, R.; Tang, S. K. Y.; Li, S. C.; Ng, S.; Matochko, W.; Jafari, M. R. Diversity of Phage-Displayed Libraries of Peptides during Panning and Amplification. *Molecules* **2011**, *16* (2), 1776–1803. https://doi.org/10.3390/molecules16021776.
42. Holdgate, G.; Embrey, K.; Milbradt, A.; Davies, G. Biophysical Methods in Early Drug Discovery. *ADMET DMPK* **2019**, *7* (4), 222–241. https://doi.org/10.5599/admet.733.
43. Cagiada, M.; Bottaro, S.; Lindemose, S.; Schenstrøm, S. M.; Stein, A.; Hartmann-Petersen, R.; Lindorff-Larsen, K. Discovering Functionally Important Sites in Proteins. *Nat. Commun.* **2023**, *14* (1), 4175. https://doi.org/10.1038/s41467-023-39909-0.
44. Alberts, B.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K.; Walter, P. Analyzing Protein Structure and Function. In *Molecular Biology of the Cell. 4th edition*; Garland Science, 2002.
45. Qin, D. Next-Generation Sequencing and Its Clinical Application. *Cancer Biol. Med.* **2019**, *16* (1), 4–10. https://doi.org/10.20892/j.issn.2095-3941.2018.0055.
46. Goldfarb, D.; Lafferty, M. J.; Herring, L. E.; Wang, W.; Major, M. B. Approximating Isotope Distributions of Biomolecule Fragments. *ACS Omega* **2018**, *3* (9), 11383–11391. https://doi.org/10.1021/acsomega.8b01649.
47. Okawa, S.; Fischer, B.; Krijgsveld, J. Properties of Isotope Patterns and Their Utility for Peptide Identification in Large-Scale Proteomic Experiments. *Rapid Commun. Mass Spectrom.* **2013**, *27* (9), 1067–1075. https://doi.org/10.1002/rcm.6551.
48. Slawski, M.; Hussong, R.; Tholey, A.; Jakoby, T.; Gregorius, B.; Hildebrandt, A.; Hein, M. Isotope Pattern Deconvolution for Peptide Mass Spectrometry by Non-Negative Least Squares/Least Absolute Deviation Template Matching. *BMC Bioinformatics* **2012**, *13* (1), 291. https://doi.org/10.1186/1471-2105-13-291.
49. Johnson, A. R.; Carlson, E. E. Collision-Induced Dissociation Mass Spectrometry: A Powerful Tool for Natural Product Structure Elucidation. *Anal.*

Chem. **2015**, *87* (21), 10668–10678. https://doi.org/10.1021/acs.analchem.5b01543.

50. Martin, D. B.; Eng, J. K.; Nesvizhskii, A. I.; Gemmill, A.; Aebersold, R. Investigation of Neutral Loss during Collision Induced Dissociation of Peptide Ions. *Anal. Chem.* **2005**, *77* (15), 4870–4882. https://doi.org/10.1021/ac050701k.
51. Jones, A. W.; Cooper, H. J. Dissociation Techniques in Mass Spectrometry-Based Proteomics. *Analyst* **2011**, *136* (17), 3419–3429. https://doi.org/10.1039/C0AN01011A.
52. Medzihradszky, K. F.; Burlingame, A. L. The Advantages and Versatility of a High-Energy Collision-Induced Dissociation-Based Strategy for the Sequence and Structural Determination of Proteins. *Methods* **1994**, *6* (3), 284–303. https://doi.org/10.1006/meth.1994.1030.
53. Good, D. M.; Wirtala, M.; McAlister, G. C.; Coon, J. J. Performance Characteristics of Electron Transfer Dissociation Mass Spectrometry*. *Mol. Cell. Proteomics* **2007**, *6* (11), 1942–1951. https://doi.org/10.1074/mcp.M700073-MCP200.
54. Kim, M.-S.; Pandey, A. Electron Transfer Dissociation Mass Spectrometry in Proteomics. *Proteomics* **2012**, *12* (0), 530–542. https://doi.org/10.1002/pmic.201100517.
55. Koh, L. Q.; Lim, Y. W.; Gates, Z. P. Affinity Selection from Synthetic Peptide Libraries Enabled by De Novo MS/MS Sequencing. *Int. J. Pept. Res. Ther.* **2022**, *28* (2), 1–14. https://doi.org/10.1007/S10989-022-10370-9/FIGURES/11.
56. Zhang, G.; Brown, J. S.; Quartararo, A. J.; Li, C.; Tan, X.; Hanna, S.; Antilla, S.; Cowfer, A. E.; Loas, A.; Pentelute, B. L. Rapid de Novo Discovery of Peptidomimetic Affinity Reagents for Human Angiotensin Converting Enzyme 2. *Commun. Chem. 2022 51* **2022**, *5* (1), 1–10. https://doi.org/10.1038/s42004-022-00625-3.
57. 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.* **2020**, *142* (46), 19642–19651. https://doi.org/10.1021/jacs.0c08964.
58. Doerr, A. DIA Mass Spectrometry. *Nat. Methods* **2015**, *12* (1), 35–35. https://doi.org/10.1038/nmeth.3234.
59. Li, J.; Smith, L.; Zhu, H.-J. Data-Independent Acquisition (DIA): An Emerging Proteomics Technology for Analysis of Drug-Metabolizing Enzymes and Transporters. *Drug Discov. Today Technol.* **2021**, *39*, 49–56. https://doi.org/10.1016/j.ddtec.2021.06.006.
60. Ng, C. C. A.; Zhou, Y.; Yao, Z.-P. Algorithms for *de-Novo* Sequencing of Peptides by Tandem Mass Spectrometry: A Review. *Anal. Chim. Acta* **2023**, *1268*, 341330. https://doi.org/10.1016/j.aca.2023.341330.
61. Liu, K.; Ye, Y.; Li, S.; Tang, H. Accurate de Novo Peptide Sequencing Using Fully Convolutional Neural Networks. *Nat. Commun.* **2023**, *14* (1), 7974. https://doi.org/10.1038/s41467-023-43010-x.

62. Mai, Z.-B.; Zhou, Z.-H.; He, Q.-Y.; Zhang, G. Highly Robust de Novo Full-Length Protein Sequencing. *Anal. Chem.* **2022**, *94* (8), 3467–3475. https://doi.org/10.1021/acs.analchem.1c03718.
63. *Age of Earth Collection*. https://education.nationalgeographic.org/resource/resource-library-age-earth (accessed 2024-03-16).
64. Dara, S.; Dhamercherla, S.; Jadav, S. S.; Babu, C. M.; Ahsan, M. J. Machine Learning in Drug Discovery: A Review. *Artif. Intell. Rev.* **2022**, *55* (3), 1947–1999. https://doi.org/10.1007/s10462-021-10058-4.
65. Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; Zhao, S. Applications of Machine Learning in Drug Discovery and Development. *Nat. Rev. Drug Discov.* **2019**, *18* (6), 463–477. https://doi.org/10.1038/s41573-019-0024-5.
66. Askr, H.; Elgeldawi, E.; Aboul Ella, H.; Elshaier, Y. A. M. M.; Gomaa, M. M.; Hassanien, A. E. Deep Learning in Drug Discovery: An Integrative Review and Future Challenges. *Artif. Intell. Rev.* **2023**, *56* (7), 5975–6037. https://doi.org/10.1007/s10462-022-10306-1.
67. Carracedo-Reboredo, P.; Liñares-Blanco, J.; Rodríguez-Fernández, N.; Cedrón, F.; Novoa, F. J.; Carballal, A.; Maojo, V.; Pazos, A.; Fernandez-Lozano, C. A Review on Machine Learning Approaches and Trends in Drug Discovery. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 4538–4558. https://doi.org/10.1016/j.csbj.2021.08.011.
68. Nag, S.; Baidya, A. T. K.; Mandal, A.; Mathew, A. T.; Das, B.; Devi, B.; Kumar, R. Deep Learning Tools for Advancing Drug Discovery and Development. *3 Biotech* **2022**, *12* (5), 110. https://doi.org/10.1007/s13205-022-03165-8.
69. Gentleman, R.; Carey, V. J. Unsupervised Machine Learning. In *Bioconductor Case Studies*; Hahne, F., Huber, W., Gentleman, R., Falcon, S., Eds.; Use R!; Springer: New York, NY, 2008; pp 137–157. https://doi.org/10.1007/978-0-387-77240-0_10.
70. *Unsupervised Machine Learning: Examples and Use Cases*. AltexSoft. https://www.altexsoft.com/blog/unsupervised-machine-learning/ (accessed 2024-03-02).
71. Alloghani, M.; Al-Jumeily, D.; Mustafina, J.; Hussain, A.; Aljaaf, A. J. A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. In *Supervised and Unsupervised Learning for Data Science*; Berry, M. W., Mohamed, A., Yap, B. W., Eds.; Unsupervised and Semi-Supervised Learning; Springer International Publishing: Cham, 2020; pp 3–21. https://doi.org/10.1007/978-3-030-22475-2_1.
72. Voicu, A.; Duteanu, N.; Voicu, M.; Vlad, D.; Dumitrascu, V. The Rcdk and Cluster R Packages Applied to Drug Candidate Selection. *J. Cheminformatics* **2020**, *12* (1), 3. https://doi.org/10.1186/s13321-019-0405-0.
73. Jiang, T.; Gradus, J. L.; Rosellini, A. J. Supervised Machine Learning: A Brief Primer. *Behav. Ther.* **2020**, *51* (5), 675–687. https://doi.org/10.1016/j.beth.2020.05.002.

74. Nasteski, V. An Overview of the Supervised Machine Learning Methods. *HORIZONS.B* **2017**, *4*, 51–62. https://doi.org/10.20544/HORIZONS.B.04.1.17.P05.
75. Burkart, N.; Huber, M. F. A Survey on the Explainability of Supervised Machine Learning. *J. Artif. Intell. Res.* **2021**, *70*, 245–317. https://doi.org/10.1613/jair.1.12228.
76. Jeon, J.; Nim, S.; Teyra, J.; Datti, A.; Wrana, J. L.; Sidhu, S. S.; Moffat, J.; Kim, P. M. A Systematic Approach to Identify Novel Cancer Drug Targets Using Machine Learning, Inhibitor Design and High-Throughput Screening. *Genome Med.* **2014**, *6* (7), 57. https://doi.org/10.1186/s13073-014-0057-7.
77. Ferrero, E.; Dunham, I.; Sanseau, P. In Silico Prediction of Novel Therapeutic Targets Using Gene-Disease Association Data. *J. Transl. Med.* **2017**, *15* (1), 182. https://doi.org/10.1186/s12967-017-1285-6.
78. Riniker, S.; Wang, Y.; Jenkins, J. L.; Landrum, G. A. Using Information from Historical High-Throughput Screens to Predict Active Compounds. *J. Chem. Inf. Model.* **2014**, *54* (7), 1880–1891. https://doi.org/10.1021/ci500190p.
79. ElAbd, H.; Bromberg, Y.; Hoarfrost, A.; Lenz, T.; Franke, A.; Wendorff, M. Amino Acid Encoding for Deep Learning Applications. *BMC Bioinformatics* **2020**, *21* (1), 235. https://doi.org/10.1186/s12859-020-03546-x.
80. Mohapatra, S.; An, J.; Gómez-Bombarelli, R. Chemistry-Informed Macromolecule Graph Representation for Similarity Computation, Unsupervised and Supervised Learning. *Mach. Learn. Sci. Technol.* **2022**, *3* (1), 015028. https://doi.org/10.1088/2632-2153/ac545e.

2. Affinity Selection-Mass Spectrometry with Linearizable Macrocyclic Peptide Libraries

The work presented in this chapter has been reproduced and adapted from the following publication:

Lee, M.A.*; Brown, J.S.*; Farquhar, C.E.; Loas, A.; Pentelute, B.L. Affinity Selection-Mass Spectrometry with Linearizable Macrocyclic Peptide Libraries*. ChemRxiv Preprint. Manuscript submission for peer review in progress.* **2024**

**Denotes equal contribution*

2.1. Introduction

Macrocyclic peptides show therapeutic promise with advantages over small molecules to disrupt protein-protein interactions and over proteins to cross biological membranes barriers.[1–4] Specifically, macrocyclization can impart several potential benefits to linear precursors, including increased proteolytic stability, cell permeability, binding affinity, and oral bioavailability.[5,6] Proteases often engage and degrade peptides in extended β-strand conformations.[7,8] Macrocyclization can offer proteolytic resistance by limiting conformational accessibility of the peptide backbone to the enzyme active site, and enable the use of specific engineerable scaffolds (e.g., stapled α-helices).[9–12] Cyclization is central to the currently applied design principles to achieve passive cell permeability, in addition to strategies that modulate molecular weight, polar surface area, hydrogen bond interactions, and shape.[13–16] Combining proteolytic stability and passive permeability can impart oral bioavailability for peptide-based drug candidates, which can further be improved by pharmaceutical formulation.[17,18] For these reasons, macrocyclic libraries are preferred for screening with peptide ligand discovery platforms. In addition, the direct identification of macrocyclic peptide binders from these selections streamlines subsequent development by alleviating the need to optimize suitable cyclization sites. Lastly, the conformational constraint imparted by macrocyclization may improve discovery rates of ligands from libraries against challenging targets.[19–21]

Genetically-encoded discovery platforms generally access macrocyclic peptide libraries while focusing on high diversity (>10^8 members),[22–28] while synthetic libraries can access the non-natural chemical space at lower diversity (<10^8 members).[29–33] While more stable macrocyclization linkages are preferred (e.g., thioether or alkyl chain),[24,34] the disulfide linkage is suitable at the ligand discovery stage, and does not require any chemical modification or treatment that could compromise genetic amplification in some platforms.[28,35] The disulfide linkage has been used to create macrocyclic libraries for over two decades in phage display discovery platforms,[36–38] and is commonly encountered in clinically-approved drugs.[5,39] Synthetic libraries generally leverage the broader use of non-natural or abiotic functionalities, which have frequently appeared critical to the success of clinical peptide drug candidates including inhibitors to the interleukin-23 receptor (IL-23R), mouse double minute 2 (MDM2), β-catenin, and proprotein convertase subtilisin/kexin type 9 (PCSK9).[17,29,40–42] Because they cannot be genetically-encoded or amplified, synthetic libraries are screened directly[30] as in affinity selection decoded by mass spectrometry (AS-MS).[43,44] With a key exception of DNA-encoded libraries,[45] state-of-the-art synthetic macrocyclic libraries generally number below tens of thousands of individual compounds.[30]

The complexity of decoding macrocyclic peptide sequences in mass spectrometry is a historic limitation for the use of synthetic macrocyclic libraries in affinity selection discovery platforms. Experimental approaches for decoding macrocyclic libraries include computational processing of mass spectra[46–49] and chemically-triggered linearization.[50–55] Computational approaches process primary, secondary, and various tertiary mass spectra of cyclic peptide fragments, and have exceled where database matching is possible.[46,49] For de novo sequencing, the complexity of enumerating virtual spectra dramatically increases as the number of monomers and library size increases, and has only been demonstrated up to ~1,000-membered libraries.[47,56] Chemically-triggered linearization adds a synthetic step that must be near-quantitative and high-yielding to enable bottom-up sequencing of non-cyclic peptides, which has been demonstrated at very high diversities.[57] However, most chemical linearization treatments are harsh and/or rely on the inclusion of non-standard chemical functional groups at fixed positions, limiting library composition.[50–55] Moreover, these approaches have yet to be demonstrated on high-diversity libraries (~10^8 members or more).

We demonstrate here ligand discovery from high-diversity libraries (10^8 members) utilizing AS-MS against anti-hemagglutinin antibody clone 12ca5 (hereinafter abbreviated as 12ca5) and mouse cadherin-2. Cyclization by disulfide bond formation is accomplished using aqueous iodine. We verify the integrity of the library utilizing Ellman's assay and size-exclusion chromatography (SEC) to confirm near-quantitative intramolecular macrocyclization of synthetically-prepared combinatorial libraries. Linearization is accomplished by mild reduction with heat and 1,3-dithiothreitol (DTT) and confirmed by Ellman's assay to enable standard tandem MS sequencing. We apply these new macrocyclic high-diversity libraries containing natural and non-natural (or noncanonical) amino acids in an MS-based affinity selection platform for de novo peptide ligand discovery.

We demonstrate successful discovery of nanomolar ligands against 12ca5 and the ectodomain of cadherin-2. The 12ca5 protein binds peptides containing the sequence D**DY(A/S).[58,59] While 12ca5 has been used to benchmark linear AS-MS libraries,[33,60] we utilize it here to benchmark and additionally validate the use of the new high-diversity macrocyclic libraries. Cadherin-2 was considered as a second target because of the potential impacts for chemical biology that an affinity reagent could provide, ranging from basic cell adhesion, to neural synapses formation,[61] to the construction of intercalated discs of mammalian heart,[62] as well as potential drug delivery due to its relative tissue selectivity in the brain and heart.[63] These critical roles in biology are generally facilitated by homodimerization in domain 1 and 2.[61,64,65] Thus, we sought to discover ligands that bind to cadherin-2 domains 4 and 5 as they may not interfere with caherin-2 function. Outside of domain 1 and 2, there are no ligands to cadherin-2 to our knowledge.[64,65] Lastly, we demonstrate the incorporation of non-natural amino acids for second-generation ligand discovery in libraries designed with input gained by structure-activity relationship (SAR) data gathered on the initially discovered cadherin-binding peptide (CBP). Taken together, the successful discovery of macrocyclic ligands to both targets from AS-MS demonstrates the potential deployment of ultra-large synthetically-prepared macrocyclic libraries for peptide ligand discovery and development.

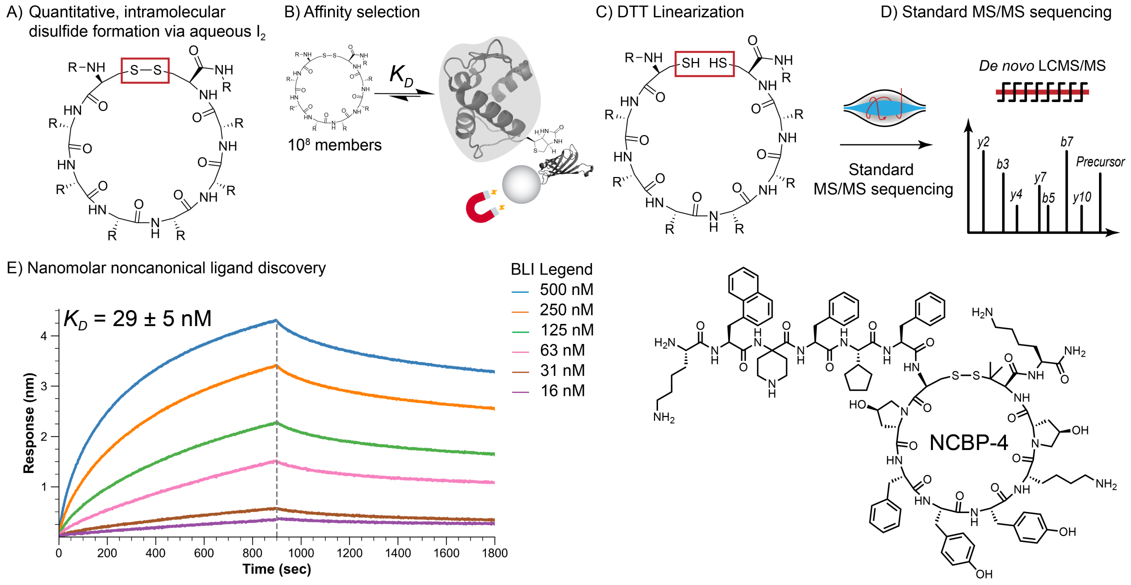


Figure 2.1: *Disulfide linkages allows for high-number diversity libraries compatible with decoding by tandem MS/MS. (A) Libraries of macrocyclic peptides are prepared for affinity selection by oxidation of cysteine analogs using aqueous iodine, providing a near-quantitative conversion to intramolecular macrocyclic peptides. (B) Affinity selection facilitates the isolation of high affinity ligands to a protein of interest. (C) After affinity selection, peptides can be quickly linearized using dithiothreitol (DTT). (D) Standard de novo LCMS/MS sequencing methods can be applied due to the linearization step. (E) Ligand affinity is confirmed by a biophysical assay (i.e. biolayer interferometry).*

2.2. Results and Discussion

Due to its demonstrated utility in genetically-encoded libraries, disulfide-induced macrocyclization of peptides has become a routine approach with a variety of existing methods to facilitate the oxidization step. Several different methods exist to form the disulfide linkage on single peptides, including oxidation using dimethyl sulfoxide,66 a gentle stream of air, or aqueous iodine with <5% methanol. Ideally, the macrocyclization step can be introduced during standard peptide library synthesis without incurring production delays or yield losses. The isolation of peptides in DMSO-containing solutions could be challenging as the solvent cannot be easily evaporated or lyophilized and solid-phase extraction could incur sample loss due to the DMSO content without further aqueous dilution.[67] In comparison, oxidation utilizing iodine presents itself as a rapid method compatible with mixtures of aqueous or organic solvents, and can even facilitate the formation of disulfide bonds on resin during solid-phase peptide synthesis.[43,68] However, longer reaction times on resin promote iodine-based side reactions, therefore rapid in-solution oxidation is preferred (< 15 min).[69]

Iodine facilitated formation of macrocyclic peptide libraries at 200-million membered scale. We synthesized macrocyclic libraries by split-and-pool solid-phase peptide synthesis using mono-sized 20 μm resin (8.33 g of resin, 2.00 mmol scale total), with each bead providing ~1 pmol of peptide. Two billion-membered libraries were prepared with the designs of CX12CK and X6CX6CK, where X = all canonical amino acids except Cys, to control disulfide formation, and Ile because it is isobaric in mass with Leu (18 amino acids) and C = cysteine, homocysteine, and penicillamine (Figure 2.3A). The libraries were split into five separate 200-million-membered aliquots and cleaved from the solid phase resin using a cleavage cocktail. After ether trituration and lyophilization, peptide libraries were cyclized in 5% acetonitrile in water (with 0.1% trifluoroacetic acid) at ~2 mg/mL (~1 mM) by dropwise addition of ~1 eq. iodine in methanol until a yellow-brown color persisted. After 5-10 minutes at room temperature in the dark, the reaction was quenched with aqueous ascorbic acid to provide a colorless solution again (3.5 eq.). These libraries were then characterized to verify the efficiency of the oxidation and linearization reactions as well as their structure (intramolecular vs intermolecular disulfide formation).

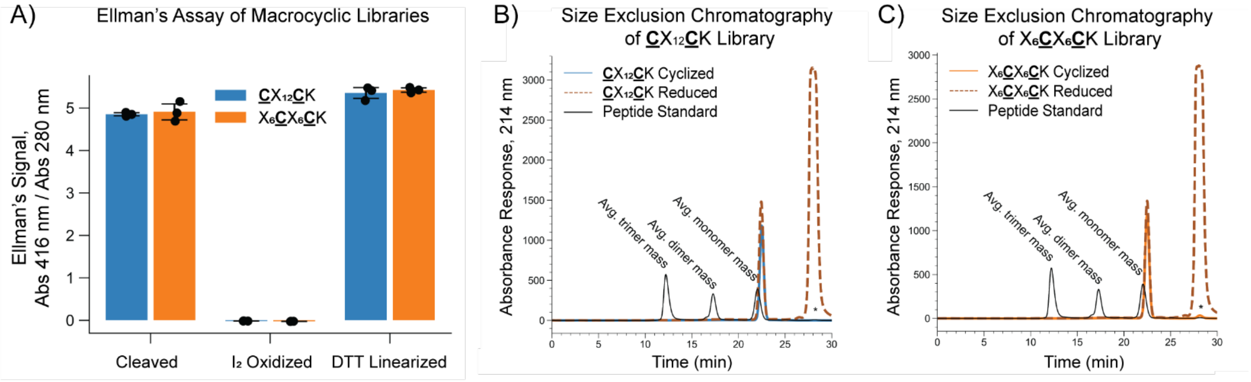


Figure 2.2: *Characterization of macrocyclic libraries based on size and thiol concentration showed near-quantitative formation of intramolecular disulfide bonds. (A) Ellman's assay showed expected changes in total thiol concentration of the peptides directly after cleavage from the solid phase resin, after oxidation by dropwise addition of ~1 eq. of 60 mM iodine in methanol to facilitate disulfide formation (room temperature, 5-10 minutes in the dark, subsequent quench with 3.5 eq. aqueous ascorbic acid), and after reduction using DTT (50 mg/mL, ~1000 eq at 60 ºC for 15 minutes). Free thiol was quantitatively consumed during the oxidation process and was restored after linearization to concentrations comparable to those determined directly after cleavage. (B,C) Size exclusion chromatograms of absorbance at 214 nm of two macrocyclic libraries compared to molecular weight standards corresponding to the average mass of monomeric, dimeric, and trimeric species. Library samples were ran using the cyclized form (later used in affinity selection experiments) and the DTT linearized form, demonstrating the formation of intramolecular disulfide bonds. Peaks marked with an asterisk (*) were residual elements from the sample buffer. C = cysteine, homocysteine, and penicillamine.*

Iodine-promoted cyclization was highly efficient and provided near-quantitative oxidization to disulfide by thiol quantification using Ellman's assay. We quantitated thiol oxidation by performing an Ellman's assay, normalized by the absorbance of the library at 280 nm (Figure 2.2A). The thiol content of the library was quantified by Ellman's reagent after cleavage, cold ether trituration, and solid-phase extraction (SPE), to remove any remaining reducing scavengers. Upon aqueous resuspension of the library, a strong thiol signal was observed. This signal was eliminated completely by the treatment of the library with iodine, ascorbic acid quench, SPE purification, and aqueous resuspension, consistent with the near-quantitative formation of disulfide bonds.

With macrocyclic libraries in hand, a MS-friendly protocol to reduce and linearize the peptides was devised to enable standard tandem sequencing approaches. Both DL-dithiothreitol (DTT) and tris(2-carboxyethyl)phosphine (TCEP) were considered for disulfide reduction. Compared to TCEP, DTT is smaller and more hydrophilic and less likely to be retained on reverse phase columns. Because of the concern for its retention on column, TCEP was utilized in a bead-immobilized form, whereas DTT was directly added to each sample just before mass spectrometry, reducing handling steps and potential sample loss. Due to its high solubility, DTT was utilized at 50 mg/mL, (~1000 eq) at 60 ºC for 15 minutes, whereas immobilized TCEP was utilized at 20 eq at room temperature for 25 minutes per manufacturer protocol.[70] For DTT-treated samples, an SPE purification was performed to remove excess DTT reagent, whereas samples treated with immobilized TCEP were isolated by centrifugation. While TCEP only provided incomplete reduction (40% for $X_6\underline{\mathbf{C}}X_6\underline{\mathbf{C}}K$ and 85% for $\underline{\mathbf{C}}X_{12}\underline{\mathbf{C}}K$), DTT provided near-quantitative disulfide reduction (~100% of original Ellman’s signal across all libraries, Figure 2.15). Additionally, the reduction efficiency by DTT was found to be similar at pH 3 and pH 8 (Figure 2.15). The direct reduction at pH 3 was performed to mimic a prepared sample in 0.1% formic acid in water, which could then directly be injected in the mass spectrometer for tandem sequencing. Overall, these data support the near-quantitative formation and reduction of disulfide bonds.

Size exclusion chromatography (SEC) confirmed disulfide bonds correspond to formation of intramolecular species, producing an almost-exclusively monomeric macrocyclic peptide library. Utilizing SEC to separate the library by its apparent molecular weight, we assessed if the iodine-facilitated disulfide bond formation was intramolecular or intermolecular. Library samples were injected on a SuperDex® 30 10/300 GL column, which can distinguish the molecular weight range of 100 to 7000 Da, to analyze the presence of monomeric, dimeric, and oligomeric peptides induced by iodine oxidation. Aliquots of 25 µg of library were injected after oxidation by iodine and purification by SPE, as well as in the reduced form after treatment with DTT. A custom low-molecular weight peptide standard was also prepared using peptides that correspond to the average molecular weights of monomeric, dimeric, and trimeric species along with the exclusion limit of the column; the components of the custom standard are given in Table 2.3. As shown in Figures 2.2B and 2.2C, only monomeric library species were observed when compared to the peptide standard, indicating that the disulfide bonds formed were intramolecular rather than intermolecular. This result, in tandem with the data from the Ellman's assay, asserted a nearly quantitative conversion of peptide thiols to intramolecular disulfide bonds and confirmed this technique provides a facile method for preparing high-diversity macrocyclic peptide libraries.

The high-diversity macrocyclic libraries verified to be compatible with standard tandem MS/MS sequencing protocols and DTT facilitated the expected recovery of high-confidence peptide sequencing. Small aliquots of library (~1000 beads equivalent to ~1000 sequences) were taken at various points in the affinity selection workflow from cyclization to linearization, including directly after cleavage from resin, after cyclization with iodine, and after linearization with DTT. About 8 μg of peptide library was purified via a C_{18} STAGE tip[71] from three each steps in the protocol: linear from cleavage, iodine-macrocyclized, and DTT-linearized. As expected, the library samples taken directly after cleavage and after linearization showed high sequencing confidence, described by *de novo* sequence IDs using PEAKS Studio 8.5 with an assigned local confidence (ALC) greater than 85% and an absolute mass error <5 ppm (see Figure 2.3B and Figure 2.17).[72] Conversely, the macrocyclized library sample showed poor sequencing confidence, consistent with unproductive fragmentation caused by the disulfide macrocycle. Lastly, the linearization by DTT in the mass spectrometry sample provided a significant recovery of the peptides discovered in standard tandem sequencing methods (Figure 2.3B).

Analysis of the MS/MS sequencing data demonstrated a balanced distribution of amino acid monomers throughout the library, as well as the incorporation of noncanonical cysteine analogs. Histograms showing the monomer distribution among high-confidence sequence assignments are given in Figure 2.3C. Although penicillamine and methionine are isobaric, the fixed positions of penicillamine allowed for effective filtering of sequences to prevent inaccurate assignments. Interestingly, the X_6**C**X_6**C**K library design sequenced with higher confidence overall compared to the **C**X_{12}**C**K library design, suggesting a benefit of the intermediate cysteine position during fragmentation events in sequence assignment. Overall, these results corroborate the successful split-and-pool synthesis, oxidation, reduction, and tandem sequencing decoding of the macrocyclic libraries.

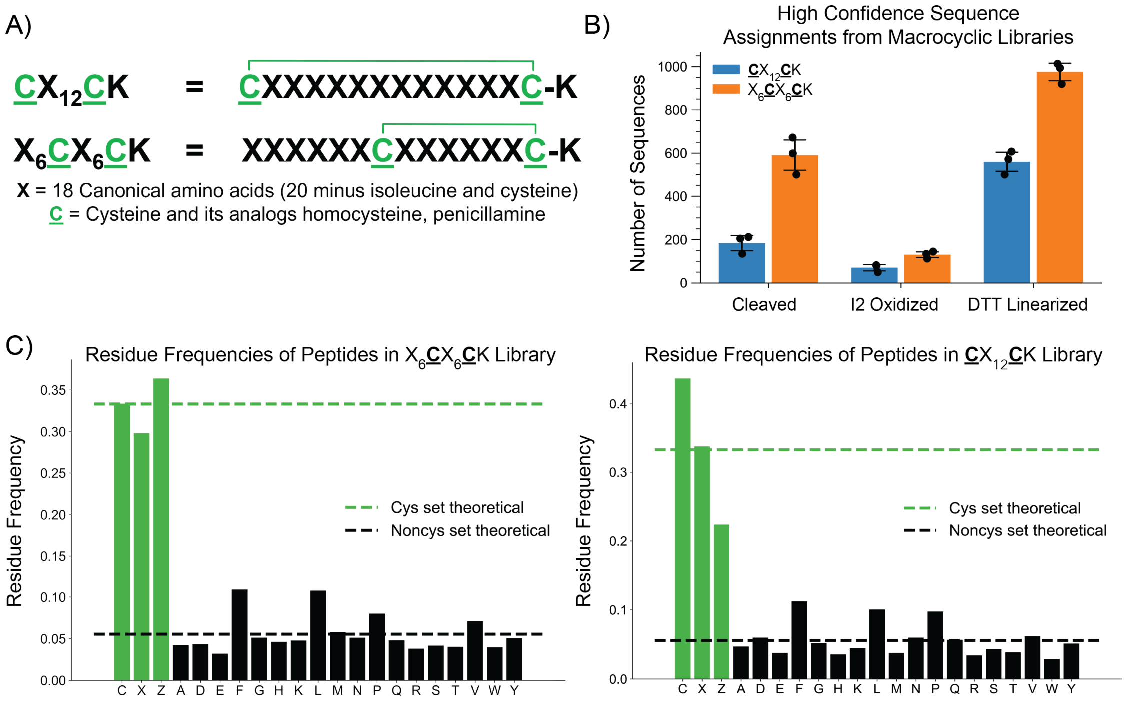


Figure 2.3: *Characterization of the macrocyclic peptide libraries by tandem MS/MS sequencing shows successful split-and-pool synthesis of 15-residue libraries. (A) The macrocyclic peptide libraries were synthesized according to two designs, with a large 12-member macrocycle or a smaller 6-member macrocycle. Additionally, cysteine analogs including homocysteine and penicillamine were used to increase the diversity around the resulting disulfide linkage. (B) High-confidence sequence assignments of ~1000-member library samples directly after cleavage, after oxidation by I_2, and after subsequence reduction using DTT show a loss and gain of sequencing capabilities with the macrocyclization process as expected (n = 3). High-confidence sequences were determined as having a calculated average local confidences (ALC) in PEAKS Studio 8.5 for the de novo sequence assignment by PEAKS Studio 8.5 greater than 85% and an*

absolute mass error <5 ppm. (C) Normalized residue frequencies assignments (as a fraction) show a balanced incorporation of amino acids in the variable positions (black) and cysteine analog positions (green, X is homocysteine, Z is penicillamine).

Macrocyclic low-nanomolar peptide ligands were discovered by affinity selection-mass spectrometry (AS-MS) performed against 12ca5 as a model protein target. The anti-hemagglutinin protein 12ca5 binds peptides containing the sequence D**DY(A/S) and has been used to benchmark AS-MS libraries.[33,58–60] Seven high affinity peptide ligands were pulled down from the X$_6$**C**X$_6$**C**K library design, while only one peptide was from the **C**X$_{12}$**C**K library (see Table 2.4 for all identified sequences). A select number of these sequences were synthesized and validated (see Figure 2.7) for their binding affinity using biolayer interferometry (BLI). All identified binders exhibited apparent dissociation constants (K_D) in the single-digit nanomolar to estimated high picomolar range, nearing the lower limit of detection for the instrument (see Figures 2.4 and 2.17). The binding motif was present in the same position or frameshift in all sequences found from the X$_6$**C**X$_6$**C**K library. Specifically, the cysteine analog in the middle of the library design was located inside of the 12ca5 motif at the third position (i.e., D*ΨDY(A/S), where Ψ was discovered to be cysteine, homocysteine, or penicillamine). Moreover, this trend also suggests the X$_6$**C**X$_6$**C**K library design is more amenable to enrich high affinity binders against 12ca5 relative to the **C**X$_{12}$**C**K library.

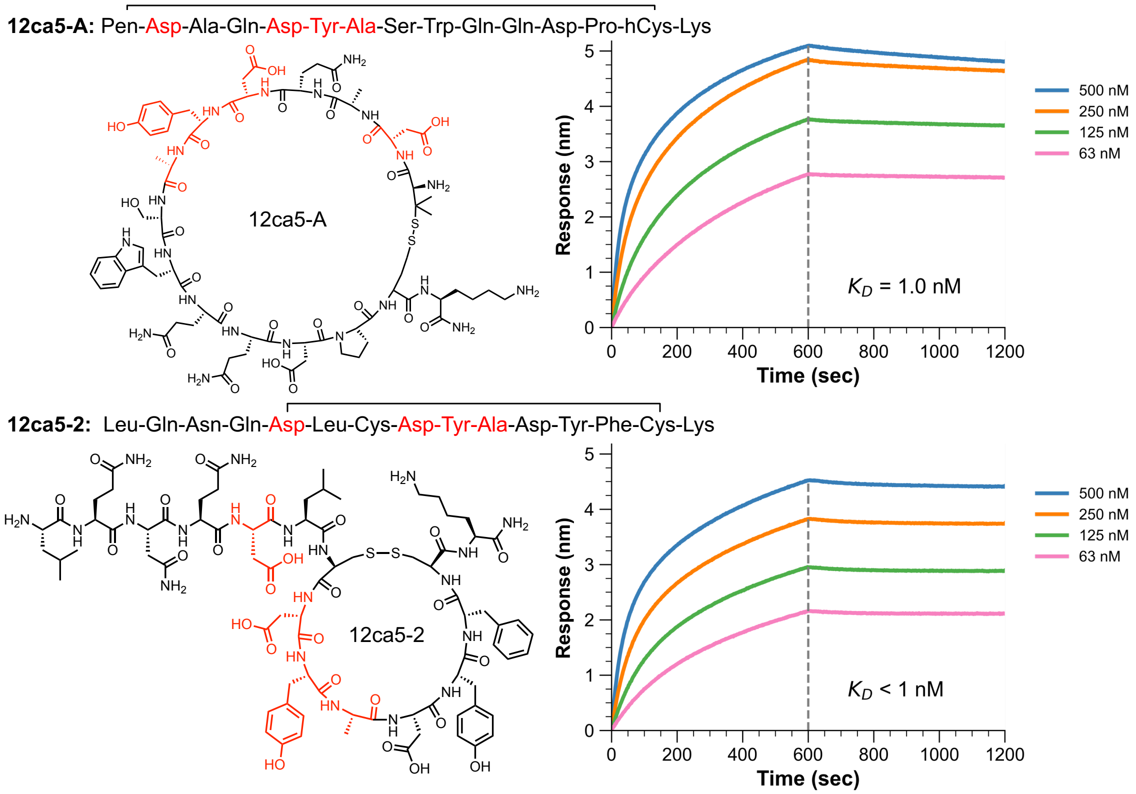


Figure 2.4: *High-affinity macrocyclic peptide ligands to 12ca5 were enriched and identified via affinity selection-mass spectrometry (AS-MS), and the binding affinity was confirmed and measured using biolayer interferometry (BLI). All peptides were prepared with Lys(Biotin)-aminohexanoic acid (Ahx) attached to the N-terminus of the shown sequence and immobilized onto the BLI tip. The BLI tip was then dipped into solutions containing varying concentrations of cadherin-2 to record the concentration-dependent association and dissociation events. The characteristic 12ca5-binding motif D**DY(A/S) is highlighted in red and appeared exclusively at a single position in the X$_6$CX$_6$CK library (7 discovered peptides). In comparison, only one motif-containing peptide was discovered from the CX$_{12}$CK library.*

Nanomolar affinity binders to cadherin-2 (CDH2) were discovered by AS-MS with the macrocyclic libraries. Due to the critical roles of CDH2 in adhesion in neural and cardiac junctions, the discovery of ligands outside of the protein homodimerization site could be of importance toward its study without affecting biological function. The homodimerization of cadherin-2 is largely driven by molecular interactions involving domains 1 and 2 of the ectodomain.[61,62,64,65] Based on these considerations, a fusion protein construct comprised of domains 4 and 5 of the CDH2 ectodomain (residues 498 to 724, Uniprot P15116) was used as the target for AS-MS selections. The fusion protein was cloned and expressed in mammalian cells, purified using SEC and Anti-Protein-C (clone HPC4) affinity tag purification, and verified by analytical SEC, reducing and non-reducing SDS-PAGE gels, as well as Western blotting (Figure 2.8). To our knowledge, no peptide or small molecular ligands have been reported to these CDH2 domains outside the homodimerization site.[64,65]

Nanomolar peptide ligands were discovered by AS-MS with both X_6**C**X_6**C**K and **C**X_{12}**C**K libraries against the CDH2[498-724] fusion protein. Only one peptide with high sequencing confidence was enriched against CDH2 (KMTFLFCNFTYKDZK, called cadherin-binding peptide, **CBP**, where Z is penicillamine disulfide bonded to C). Notably, previous ligand discovery efforts against CDH2 using linear X_{12}K libraries were unable to identify any ligands. **CBP** was synthesized and tested for its binding affinity to CDH2 by BLI, yielding a 53 nM K_D value (Figure 2.5). To verify sequence binding specificity, a scramble sequence that preserved the size of the macrocycle was synthesized and shown to have negligible binding response to CDH2 by BLI (Figures 2.9 and 2.19). The linearized form of **CBP** was also tested and demonstrated a greatly reduced binding response (0.2 nm versus 1.6 nm response for the macrocycle), with an affinity of K_D of 150 nM (Figures 2.20). The more rigid structure of the macrocyclic **CBP** thus appears to favor higher affinity.

Structure-activity relationships (SAR) were delineated to characterize **CBP** using single residue replacement studies (alanine and D-amino acid scans) and truncations. In all SAR studies the Cys7-Pen14 disulfide bond was maintained to provide a consistent macrocycle structure to optimize from. First, an alanine scan was performed by synthesis of 13 variants featuring individual alanine mutations, which were assayed by BLI against CDH2 (Table 2.5, Figure 2.10, Figure 2.21). This alanine scan revealed multiple residues to be important for binding (hot-spots) including Lys4, Phe6, Lys12, and Lys15, due to the complete ablation of binding to

CDH2 observed in BLI assays. Other residues including Met2, Thr3, Leu5, Thr10, Tyr11, and Asp13 had no effect on binding (cold-spots), suggesting they are not drivers of **CBP** binding to CDH2. Second, a truncation study focused at shortening **CBP** from the N-terminus, producing five additional peptides for BLI testing (Table 2.6, Figure 2.11 and Figure 2.22). BLI assays with these peptides confirmed the impact of the N-terminal residues on binding affinity, especially Phe6 as well as Leu5, while Met2 and Thr3 contributed minimally to CDH2 binding. Third, a D-amino acid scan of **CBP** was performed by iteratively replacing L-amino acids with D-amino acids to determine the impact of stereochemistry on the ligand interactions.[73,74] Notable hot-spots identified from the D-amino acid scan were Phe6 and Tyr11, further reinforcing the importance of the aromatic residues for binding (see Table 2.7, Figure 2.12 and Figure 2.23). In summary, these initial SAR studies outline the hot-spot residues that appear to drive the high affinity binding of CBP to

CDH2, including Lys4, Phe6, Lys12, and Lys15, while Met2, Thr3, Leu5, Thr10, and Asp13 are designated as cold-spots with a minimal effect on binding.

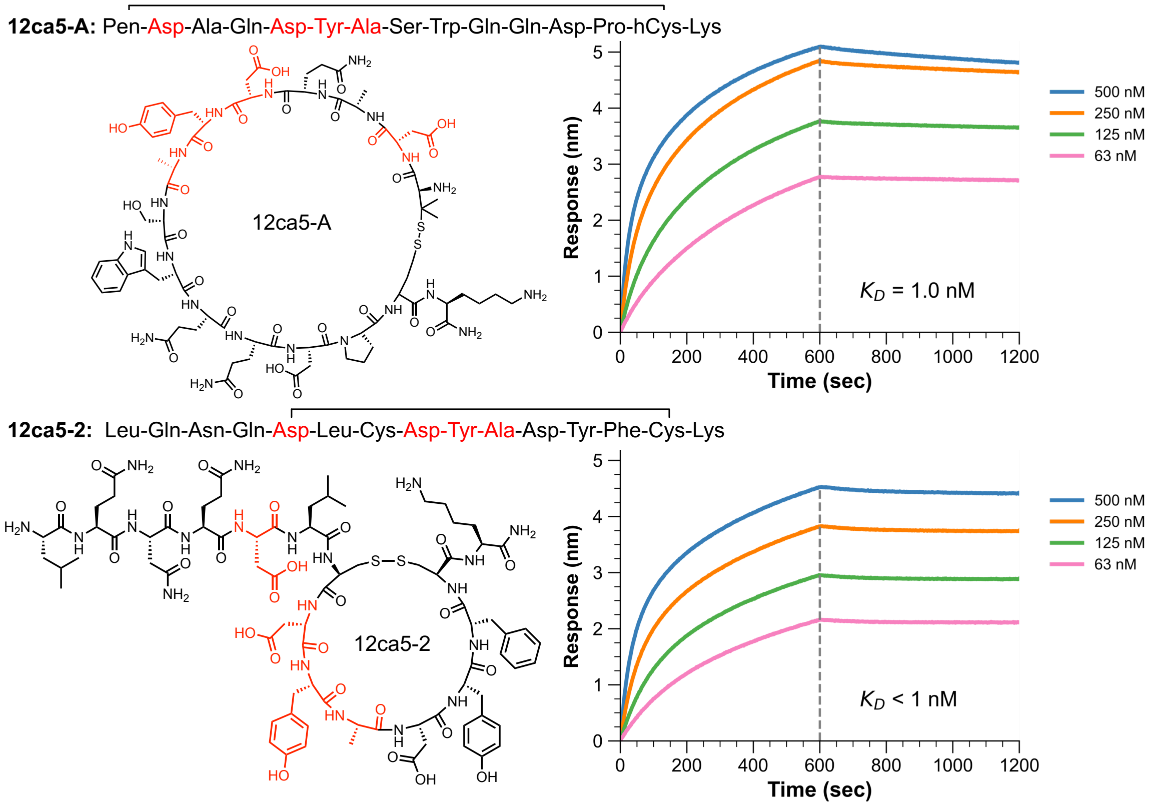


Figure 2.5: *Macrocyclic peptide libraries enabled discovery of a 53 nM peptide ligand to a portion of the ectodomain of cadherin-2. (A) Structure of CBP. (B) BLI experiment reports the affinity of CBP to CDH2 with K_D = 53 nM binding affinity. All peptides were prepared with Lys(Biotin)-Ahx attached to the N-terminus in addition to the sequence shown and immobilized onto the BLI tip. The BLI tip was then dipped into solutions containing varying concentrations of CDH2 to record the concentration-dependent association and dissociation events. (C) Summary of experimental (alanine scan, D-amino acid scan, and N-terminal truncation study) SAR data. "NB" denotes non-binding. This SAR information was used to inform the designation of CBP "hot-spots" and "cold-spots," which do or do not drive high affinity binding, respectively.*

SAR data informed the design of two focused libraries based on **CBP**: one to derivatize the high-affinity hot-spot residues, and the other to derivatize the cold-spot residues non-essential for binding. A previous approach to design noncanonical libraries is to diversify the hot-spots.[75,76] However, peptide development and optimization often also considers the cold-spot residues that do not drive high affinity binding to the target. These cold-spot residues can be non-intuitively critical to improving binding affinity, solubility, or proteolytic stability.[42,43] Thus, we chose to compare the results from both maturation strategies. The set of non-canonical amino acids for incorporation in both the hot- and cold-spot focused libraries were selected based on the consensus data provided by the docking, alanine scan, D-amino acid scan, and truncation studies (Figure 2.6A and 2.6B). These libraries were synthesized and subjected to validation by SEC as shown in Figure 2.16 to demonstrate the lack of apparent oligomerization after disulfide bond formation.

No high-affinity ligands were discovered from the hot-spot focused library by AS-MS, while the cold-spot library provided ten new high-affinity noncanonical macrocyclic cadherin-2 peptide binders (**NCBPs**). From the hot-spot library, only two candidates were identified with high sequencing confidence, which featured multiple mutations from the original **CBP** sequence and shared replacements including Thr10Msn, Tyr11Dph, and Lys15Arg (Table 2.8). However, the two hot-spot candidates (**NCBP-1** and **NCBP-2**) were synthesized and tested by BLI, revealing that they were non-binders to CDH2 under these conditions (Figure 2.13, 2.24). From the cold-spot library, ten noncanonical putative binders with high sequencing confidence were discovered and synthesized (Table 2.8, Figure 2.13). All discovered sequences from the cold-spot library (**NCBP-3 to NCBP-12**) were high-affinity binders to CDH2 in the BLI experiments, with determined K_D values between 20 and 50 nM (Figure 2.24). The significant improvement in binding affinities observed by BLI upon derivatization of the cold-spot residues support this strategy as an efficient avenue for further optimization.

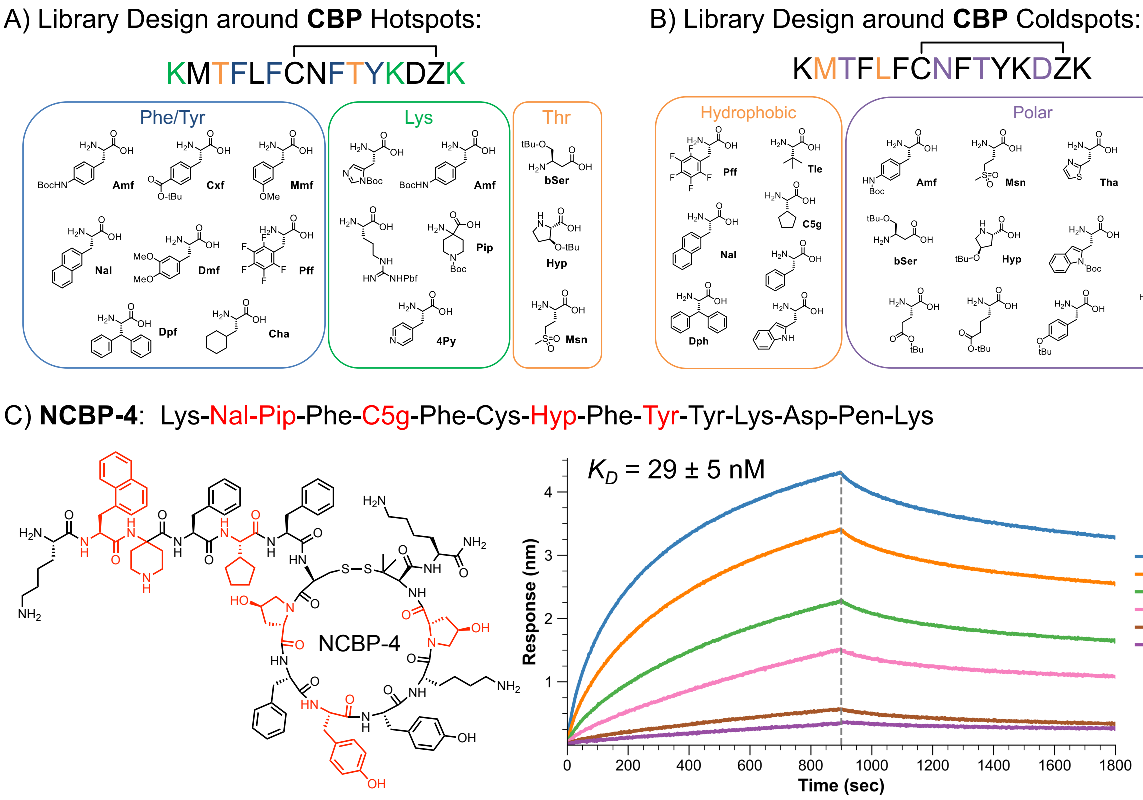


Figure 2.6: *Single-peptide SAR information informs combinatorial library design and affinity maturation with noncanonical amino acids. (A and B) From the single peptide SAR studies summarized in Figure 2.5C, two libraries were designed to perform affinity maturation. The first library focused on minimally derivatizing the hot-spots by matching the original natural amino acids properties (e.g., hydrophobicity or positive charge). The second library focused around diversifying the cold-spots to examine the possibility that the cold-spots could be mutated to improve the overall binding of the CBP peptide, either by pre-arranging the conformation of the peptide or facilitating new binding interactions with CDH2. Both libraries were prepared using split-pool synthesis, except the entirety of the theoretical sequence space was sampled by the library due to its smaller focused design. Specifically, the number of beads used in split-pool synthesis approximately matched the theoretical diversity: Hot-spot library total number of beads: 2.5 x 10^6 with theoretical sequence space diversity: 2.7 x 10^6 and cold-spot library total number of beads: 7.0 x 10^5 with theoretical sequence space diversity: 7.2 x 10^5. (C) Sequence and structure of NCBP-4 discovered from affinity selection and its BLI binding response, which exhibited high-affinity binding (K_D = 29 ± 5 nM). Nal = 3-(2-naphthyl)-L-alanine, Pip = 4-aminopiperidine-4-carboxylic acid, C5g = cyclopentylglycine, and Hyp = L-trans-4-hydroxyproline.*

The **NCBP-4** noncanonical binder exhibits nanomolar binding affinity to CDH2 (K_D = 29 ± 5 nM). The results from AS-MS of the cold-spot macrocyclic library show most cold-spot amino acids were replaced in the identified sequences (Table 2.8).

In all candidates, Met2 was replaced by a hydrophobic amino acid, with 3-(2-naphthyl)-L-alanine (Nal) appearing in six of the ten. Similarly, Leu5 was replaced by Phe, Nal, cyclopentylglycine (C5g), or pentafluoro-L-phenylalanine (PFf). For the polar subset of amino acids in the cold-spot library, the replacements made to **CBP** were more mixed. Asn8 and Thr10 were replaced with a diverse set of amino acids, possibly indicating their lack of contribution to the binding interaction. Thr3 was replaced by cationic 4-aminopiperidine-4-carboxylic acid (Pip) and polar L-β-Homoserine (bSer). And lastly, Asp13 demonstrated a preferred replacement to trans-4-hydroxyproline (Hyp), appearing in five of ten candidates. With these results in mind, **NCBP-4** was chosen for detailed investigation by BLI as it featured a consensus of amino acid replacements including Met2Nph, Thr3Pip, Leu5C5g, and Asp13Hyp. **NCBP-4** exhibited clear concentration-dependent binding to CDH2 and a resulting binding affinity of K_D = 29 ± 5 nM determined by BLI (Figure 2.6). Moreover, **NCBP-4** demonstrated a stronger response illustrated by a higher BLI signal (~4.3 nm), more than double the response seen with **CBP** (~1.6 nm, Figure 2.25).

A specific mutant of **NCBP-4** was constructed with serine mutated at the hot-spots to specifically examine if the cold-spot residues were effective in creating new binding interactions with CDH2 (sequence: Ser-Nal-Pip-Ser-C5g-Ser-Cys-Hyp-Ser-Tyr-Ser-Ser-Hyp-Pen-Ser-NH2). Interestingly, this mutant retained moderate binding to CDH2 with an apparent K_{D} of ~ 150 nM (Figure 2.25). This serine-substituted mutant was also tested against 12ca5 to assess non-specific binding. The mutant showed no affinity towards 12ca5 (Figure 2.25). This result suggests the cold spot residues of **CBP** could have been optimized further in **NCBP-4** in AS-MS, as this control ligand demonstrates some binding to CDH2 without nonspecific binding.

2.3. Conclusions

We established a protocol for the split-and-pool synthesis of high-diversity macrocyclic peptide libraries and demonstrated their use in the discovery of nanomolar ligands against two protein targets of different structure. Formation of the macrocycle is performed using a simple disulfide bond, which is rapidly installed in aqueous solution using iodine. Quantification of free thiol content via Ellman's assay in both the cyclized and linearized forms confirmed the complete conversion of thiols to disulfides in the library, while SEC revealed that the disulfide bonds were formed exclusively intramolecularly without oligomerization. While the disulfide bond is not the most robust linkage for cyclization, there are several approaches available to replace it when needed to improve stability toward therapeutic development.[11,79,80]

Affinity selection was able to identify protein-specific ligands from these synthetic macrocyclic libraries for both a model protein (12ca5) and a novel target, cadherin-2, which participates in basic biological adhesion of cells in neural and cardiometabolic function. Both the canonical **CBP** and non-canonical **NCBP-4** peptides demonstrated concentration-dependent binding to CDH2, binding specificity, and high affinity with K_D values of 50 and 29 nM, respectively. After examining the SAR of **CBP**, several hot-spot residues were revealed to be critical for binding to CDH2, featuring several hydrophobic and cationic residues.

From the SAR studies, two additional macrocyclic libraries containing a diverse set of noncanonical amino acids were synthesized focusing on the affinity-driving hot-spots and the non-essential cold-spots, respectively. The subsequent affinity selection experiments investigated the hypotheses of whether the hot-spots can be further refined or if the cold-spots can become meaningful contributors to the binding affinity upon maturation with focused library designs. Overall, AS-MS utilizing the hot-spot CDH2-focused library did not provide any binders or improvement to the original **CBP** peptide. However, AS-MS experiments utilizing the cold-spot library were able to provide several candidates individually validated to be high-affinity binders. Of these, **NCBP-4** was examined more closely for its high affinity (*K*$_D$ ~29 nM), specificity, and specific side-chain contributions demonstrated by the amino acids that replaced the original CBP cold-spot residues.

Utilizing noncanonical amino acids in combinatorically prepared macrocyclic libraries, we demonstrated the rapid affinity maturation of **CBP**. This process was successful through the replacement of cold-spot residues with noncanonical monomers. Overall, due to the improvements that macrocyclization often offer over linear peptide scaffolds, we expect this work to be fundamental to the impactful deployment of macrocyclic synthetic libraries for the advancement of peptide therapeutic discovery and development.

2.4. Materials

Canonical Fmoc-protected amino acids (FmocAla-OHxH2O, Fmoc-Arg(Pbf)-OH; Fmoc-Asn(Trt)-OH; Fmoc-Asp-(*O*-t-Bu)-OH; FmocCys(Trt)-OH; Fmoc-Gln(Trt)-OH; Fmoc-Glu(*O*-t-Bu)-OH; Fmoc-Gly-OH; Fmoc-His(Trt)- OH; Fmoc-Ile-OH; Fmoc-Leu-OH; Fmoc-Lys(Boc)-OH; Fmoc-Met-OH; Fmoc-Phe-OH; Fmoc- ProOH; Fmoc-Ser(But)-OH; Fmoc-Thr(t-Bu)-OH; Fmoc-Trp(Boc)-OH; Fmoc-Tyr(*O*-t-Bu)- OH; Fmoc-Val-OH) were purchased from Sigma Millipore (Novabiochem) and used as received. Fmoc-Lys(biotin)-OH was purchased from Sigma Millipore (Novabiochem) and used as received. Fmoc-L-His(Boc)-OH was purchased from Advanced ChemTech and used as received. O-(7-azabenzotriazol-1-yl)- *N,N,N',N'*-tetramethyluronium hexafluorophosphate (HATU, ≥97.0%) and (7-azabenzotriazol-1- yloxy)tripyrrolidinophospho-nium hexa-fluorophosphate (PyAOP, ≥97.0%) were purchased from P3 Biosystems. Fmoc-Rink amide linker (4-[(R,S)-(2,4-dimethoxyphenyl)(Fmoc-amino)methyl]phenoxyacetic acid) was purchased from Chem Impex Inc (Wood Dale, IL) and used as received. 1,4-dithio-DL-threitol (DTT, ≥99%) was purchased from Chem Impex, Inc. Iodine (crystalline, 99.5%) and L-(+)-ascorbic acid (99%) were purchased from Thermo Fisher Scientific.

Noncanonical amino acids used in this work with their associated protecting groups. All were purchased and used as received.

Table 2.1: *Noncanonical amino acids used for library synthesis.*

Noncanonical amino acid	Abbreviation	Source
Fmoc-3-(4'-pyridyl)-L-alanine	4Py	Chem Impex, Inc
Fmoc-6-aminohexanoic acid	Ahx	Chem Impex, Inc
Fmoc-4-(Boc-amino)-L-phenylalanine	Amf	Chem Impex, Inc
Fmoc-O-tert-butyl-L-β-homoserine	bSe	Chem Impex, Inc
Fmoc-L-cyclopentylglycine	C5g	Chem Impex, Inc
Fmoc-β-cyclobutyl-L-alanine	Cba	Chem Impex, Inc
Fmoc-(4-tert-butyloxycarbonyl)-L-phenylalanine	Cxf	Chem Impex, Inc
Fmoc-3,4-dimethoxy-L-phenylalanine	Dmf	Chem Impex, Inc
Fmoc-3,3-diphenyl-L-alanine	Dph	Chem Impex, Inc
Fmoc-O-tert-butyl-L-trans-4-hydroxyproline	Hyp	Chem Impex, Inc
Fmoc-3-methoxy-L-phenylalanine	Mmf	Chem Impex, Inc
Fmoc-L-methionine sulfone	Msn	Chem Impex, Inc
Fmoc-3-(1-naphthyl)-L-alanine	Nal	Chem Impex, Inc
Fmoc-pentafluoro-L-phenylalanine	Pff	Chem Impex, Inc
1-Boc-piperidine-4-Fmoc-amino-4-carboxylic acid	Pip	Chem Impex, Inc
Fmoc-3-(4-thiazolyl)-L-alanine	Tha	Chem Impex, Inc
Fmoc-L-α-tert-butylglycine	Tle	Chem Impex, Inc

Biosynthesis OmniSolv® grade N,N-dimethylformamide (DMF) was purchased from EMD Millipore (DX1732-1) and incubated with 1 pack of AldraAmine trapping agents (for 1000 – 4000 mL DMF, Sigma-Aldrich, catalog number Z511706) for 48 hours prior to use. Diisopropylethylamine (DIEA; 99.5%, biotech grade, catalog number 387649) and piperidine (ACS reagent, ≥99.0%) were purchased from Sigma-Aldrich. Formic acid (FA, 97%) was purchased from Beantown Chemical, Corp. Reaction vessels were purchased from Torviq equipped with a polypropylene frit. To each vessel was added a disc of Porex filter paper (0.025" thick, 7-12 micron) from Interstate Specialty Products. Trifluoroacetic acid (HPLC grade, ≥99.0%), Diethyl ether (anhydrous, ACS reagent, ≥99.0%), acetonitrile (HPLC grade, ≥99.9%), Omnisolv® acetonitrile (LC-MS grade, AX0156-1), Omnisolv® water (LC-MS grade, WX0001-1) and were purchased from Sigma-Aldrich. Methanol was purchased from Millipore Sigma. Formic acid Optima LC/MS (A117) was purchased from Fisher Chemical. Water was deionized using a Milli-Q Reference water purification system (Millipore). Nylon 0.22 µm syringe filters were TISCH brand SPEC17984.

H-Rink Amide-ChemMatrix® (0.49 mmol/g) resin was purchased from PCAS Biomatrix (St-Jean-sur-Richelieu, Quebec, Canada) and 20 µm TentaGel® M NH_2 Monosized Amino Microsphere resin was purchased from Rapp Polymere Inc. (Tübingen, Germany). HyClone™ Fetal Bovine Serum (SH30071.03HI, heat inactivated) was purchased from GE Healthcare Life Sciences (Logan, UT) Dynabeads MyOne Streptavidin T1 magnetic microparticles were purchased from Invitrogen (Carlsbad, CA). Phosphate buffered saline (10x, Molecular biology grade) was purchased from Corning. Sodium chloride (ACS grade) was purchased from Avantor. Guanidine hydrochloride (Cat BP178) and sodium phosphate monobasic monohydrate were purchased from Fisher Scientific.

Mouse anti-hemagglutinin antibody (clone 12ca5) was purchased from Columbia Biosciences Corporation (Cat: 00-1722, Frederick, Maryland) biotin-$(PEG)_4$-NHS ester and biotin-$(PEG)_4$-propionic acid were purchased from ChemPep Inc. (Wellington, FL). Biotinylation of 12ca5 was performed as previously described.[33]

Cadherin-2 plasmid DNA was supplied by Novo Nordisk A/S ($^{498\text{-}724}$CDH2-AviTag-HPC4). The Expi293 Expression System (A14635), Expi293 Expression Medium (A1435101), Opti-MEM™ I Reduced Serum Medium (31985070), ExpiFectamine™ 293 Transfection Kit (A14524), and Halt Protease Inhibitor Cocktail (100X, 78429) were purchased from Thermo Fisher Scientific. Sartoclear Dynamics® Lab V Clarification and Sterile Filtration Kits were purchased from Sartorius, Inc. HiTrap Q HP columns were purchased from Cytiva, Inc. Anti-Protein C Affinity Matrix (11815024001, HPC4, monoclonal Roche) was purchased from Millipore-Sigma. HPC4-Tag Antibody (68083) was purchased from Cell Signaling Technology. AviTag Biotinylation Kit (BirA500) was purchased from Avidity LLC. SuperDex 75 Increase 10/300 GL column (10 x 300 mm, 9 µm particle size, separation MW range 3000 and 70,000 Da) and Superdex 30 Increase 10/300 GL column (10 x 300 mm, 9 µm particle size, separation MW range 100 to 7000 Da) was purchased from Cytiva Life Sciences.

2.5. Methods

2.5.1. Split-and-pool synthesis of cyclized libraries

Synthesis of peptide libraries was performed using 20 µm Tentagel M NH_2 resin (0.31 mmol/g) for 10^8-member libraries. The resin was suspended in DMF and dividedly evenly between 18 syringes (all canonical amino acids except for cysteine and isoleucine) for the non-cysteine variable region, while the resin was split into 3 syringes for the variable cysteine analog positions (cysteine, homocysteine, penicillamine). Couplings were performed using the Fmoc-protected amino acid dissolved in DMF (10 eq, 0.40M) with PyAOP (0.9 eq relative to amino acid, 0.38M) activated with DIEA (1.1 eq relative to amino acid for histidine, 3 eq relative to amino acid for all others). Couplings were incubated for 1 hour. The resin was then recombined and washed with DMF three times. Fmoc deprotection was performed using 20% piperidine in DMF (1x flow wash, 2x 5 min batch treatments). The resin was washed again with DMF three times before being subjected to another split-couple-pool cycle until completion of all randomized positions.

2.5.2. Peptide Cleavage and global deprotection

Cleavage from solid phase and global deprotection was performed using a solution of 95% trifluoroacetic acid, 2.5% water, and 2.5% triisopropylsilane (~20 mL cleavage cocktail / g of resin). The solution was added until the resin was fully swelled and free flowing, then the resin was agitated on a nutating mixer for 3 hours. The peptides were triturated with 10:1 cold diethyl ether to cleavage solution. The precipitated solid was centrifuged into a pellet. The precipitate was washed with cold ethyl ether in the same manner an additional two times. The resulting solid pellet was dried gently using N2, suspended in 50% acetonitrile in water (0.1% trifluoroacetic acid), and lyophilized.

2.5.3. Solid-phase extraction

Peptides were adjusted to 5% acetonitrile in aqueous media (0.1% TFA) and purified using Supelclean™ LC-18 SPE Tube, bed wt. 1 g (Millipore Sigma Cat: 505471). The SPE tube was first conditioned with 3 CV of acetonitrile (0.1% TFA) and then equilibrated with 5 CV of 5% acetonitrile in water (0.1% TFA). Then, the suspended crude was loaded (Approximately 50 mg peptide loaded onto 1 g bed mass) and washed with 10 CV of 5% acetonitrile in water (0.1% TFA). Peptides were eluted with 70% acetonitrile (0.1% TFA, 1 CV) and lyophilized.

2.5.4. Stop and Go Extraction (STAGE) Tip preparation of library samples for nLC-MS/MS analysis

From the CDS Empore™ SDB-XC extraction disk, two cores of material were pressed using an 18 gauge blunt tip needle (each core binds 2-4 µg) and pressed into the tip of a 200 µL pipette tip. The STAGE tip was then fitted into a 1.5 mL microcentrifuge tube with a hole drilled in the center of the cap. The STAGE tip assembly was then wetted using 60 µL of 80% AcN in water (0.1% TFA) and centrifuged at 500g for 2 minutes. Then STAGE tip assembly was washed using 60 µL of 1.5% AcN in water (0.1% TFA) and centrifuged at 500g for 2 minutes. The sample was then loaded onto the STAGE tip and centrifuged at 500g in 3-minute intervals, checking the liquid level each time to ensure the tip does not run dry. The STAGE tip was then washed again using 60 µL of 1.5% AcN in water (0.1% TFA) and centrifuged at 500g for 2 minutes. The STAGE tip was then moved to a fresh microcentrifuge tube, and the peptides were eluted using 75 µL of 56% AcN in water (0.1% TFA) and centrifuged at 500g for 10 minutes. The eluted peptides were then dried using a vacuum centrifuge.

2.5.5. Rapid oxidation of peptide thiols for intramolecular cyclization using iodine

After cleavage and lyophilization, each peptide library (e.g., 10 mg, 4.2 µmol, average molecular weight ~ 2400 g/mol) was resuspended at 2 mg/mL in 5% AcN in Water (0.1% TFA) and treated with 10 µL portions of freshly prepared 60 mM I_2 in

MeOH until the solution remained yellow. For the example 10 mg scale, approximately 70 µL I_2 stock solution total was used, resulting in ~1 eq of I_2 with respect to the library. The iodine-treated library was incubated for 10 minutes in the dark at rt, upon which 15 uL of freshly prepared 1 M ascorbic acid was added (~3.6 eq of ascorbic acid with respect to the library example). This solution was immediately loaded onto a pre-equilibrated SPE column, SPE-purified to remove any remaining iodine and ascorbic acid, and lyophilized. The lyophilized powder was then resuspended at approximately 0.1 mM and its thiol concentration as quantified by Ellman's assay.

2.5.6. Reductive linearization of cyclized library

Peptide libraries were resuspended at 0.62 mg/mL to mimic the maximum concentration possibly isolated at the end of AS-MS, due to the maximum capacity of the STAGE tip (8 µg for a double plug, using 13 µL). As described in the Main Text, reduction after resuspension at pH 3 from 0.1% formic acid in ultrapure water was successful using 1,4-DL-dithiothreitol (DTT, Chem-Impex Cat: 00127). The reduction was also tested in 200 mM sodium phosphate, 5 mM EDTA, pH 8 with DTT, and immobilized Tris (2-carboxyethyl) phosphine (TCEP, Thermo Fisher Scientific, 77712). DTT was freshly prepared in a stock solution of 500 mg/mL and added to samples to provide 50 mg/mL final, 1000 eq and incubated at 60 ºC for 15 minutes. Samples were then SPE-purified, lyophilized, and Ellman's quantified upon resuspension. Immobilized TCEP beads (8 mM stock suspension) were washed three times before use with the assay buffer using centrifugation at 1000 rcf for 1 minute. Treatment of the library peptides with immobilized TCEP used 20 eq for 25 minutes at room temperature rocking on a nutating mixer. The supernatant was isolated from centrifugation, lyophilized, and Ellman's' quantified.

2.5.7. Sequencing validation of reduced and oxidized libraries

A small portion of library resin was measured and made into a 1 mg/mL stock solution in DMF. Several aliquots of ~1000 beads (for 20 µm resin, this will be about

~10 µL of 1 mg/mL stock) were taken, centrifuged, and aspirated of DMF. Each aliquot was cleaved from the solid phase support using 60 µL of 95% trifluoroacetic acid, 2.5% water and 2.5% triisopropylsilane at 60 °C for 15 minutes. Half of the liquid was then evaporated under a gentle stream of N_2, followed by dilution to a total volume of 240 µL using water (0.1% TFA). A third of the solution was aliquoted to represent the library before cyclization. The remaining solution was cyclized according to the Section 2.5.5. The cyclized peptide library and aliquoted linear library were both prepared for nLC/MS-MS analysis according to Section 2.5.11. The dried library samples were then reconstituted in water (0.1% TFA) at a concentration of 100 pg/µL/peptide (for example, prepare 8 µg of an aliquot of 1000 peptides in 80 µL). Half of the cyclized peptide sample was then linearized using DTT as described in Section 2.5.6. All three types of samples, peptide post-cleavage, cyclized, and cyclized then reduced, were subjected to nLC-MS/MS analysis as described in Section 2.5.11.

2.5.8. Ellman's thiol quantification assay

The thiol concentration of suspended peptides was completed using Ellman's reagent (Millipore-Sigma, 5,5′-Dithiobis(2-nitrobenzoic acid), D8130, ≥98%, BioReagent) using the following conditions. Ellman's stock solution was prepared at 10.0 mM and assay buffer was 1x PBS pH 8 1 mM EDTA. Nonsterile Greiner 96-well polystyrene plates (Millipore-Sigma, M2936) were used. Using the assay buffer to have a final 200 µL well volume, 3.6 µL of Ellman's stock solution was combined with the peptide solution to give a final peptide intended concentration of 0.1 mM, which was determined to be within the linear regime in which signal could be observed from a standard curve constructed using Cysteine (Millipore-Sigma, C7352, ≥98%, BioReagent). After combining, all materials were incubated for 7 minutes, and then read at 416 nm using a TECAN Spark Plate Reader. The concentration of the library was inferred by measuring its absorbance at 280 nm (NanoQuant Plate) and was used to normalize the Ellman's thiol concentration to account for slight variations in the intended resuspended concentrations.

Preparation of samples to measure thiol signal from cleavage and SPE of library (reduced library): Approximately 50 mg of peptide library (peptide + resin) was globally deprotected and cleaved from resin with 95% (v/v) TFA, 2.5% (v/v) water, and 2.5% (v/v) triisopropylsilane, for 15 minutes at 60 ºC (~ 20 mL / g of resin). Precipitated peptide was triturated (3 x 100 mL / g resin) with cold diethyl ether, resuspended in 5% acetonitrile in water (0.1% TFA) and solid-phase extracted. After lyophilization, this sample was resuspended in assay buffer at 0.1 mM and measured for its thiol concentration by Ellman's.

2.5.9. Size exclusion chromatography (SEC) of libraries

Size exclusion chromatography (SEC) was performed using an Agilent 1260 Infinity II LC System with a Superdex 30 Increase 10/300 GL column (10 x 300 mm, 9 µm particle size from Cytiva Life Sciences, separation MW range 100 to 7000 Da). 25 µg of library was injected in 200 µL of total solution. Cyclized peptide samples were aliquoted from the main stock prepared according to Sections 2.5.4 and 2.5.11. Linearized samples were prepared by adding 1000 eq from a 50 mg/mL DTT stock solution and heating the sample to 60 ^{o}C for 10 minutes before dilution to 200 µL using 1x PBS. Column conditions: isocratic 1x PBS for 1.5 column volumes at 0.8 mL/min. Buffer blanks were prepared for both cyclized and linearized samples and were subtracted from the library samples. A custom mass standard was prepared by adding 10 µg of a mixture of peptides corresponding to the following molecular weights: 1807, 3750, 5312, 8305 for average monomer mass, average dimer mass, average trimer mass, and the exclusion limit of the SEC column.

2.5.10. Affinity selection using cyclized libraries

Affinity selections were performed using a KingFisher^{TM} Duo Prime Purification System in 96 Deepwell Plates (Thermo Fisher Scientific, cat. #95040450) with the following setup:

***Table 2.2:** Plate Setup for KingFisher^{TM} Duo Prime*

Plate 1		
A	10 pM/member peptide library diluted into 1x PBS, 10% FBS	1 mL
B	Wash buffer (1x PBS, 10% FBS, 0.01% Tween20)	1 mL
C	Wash buffer (1x PBS, 10% FBS, 0.01% Tween20)	1 mL
D	Wash buffer (1x PBS, 10% FBS, 0.01% Tween20)	1 mL
E	Protein (1.5 eq) in Wash buffer (1x PBS, 10% FBS, 0.01% Tween20)	500 μL
F	Wash buffer (1x PBS, 10% FBS, 0.01% Tween20)	1 mL
G	Wash buffer (1x PBS, 10% FBS, 0.01% Tween20)	1 mL
H	1 mg of magnetic beads (100 uL) diluted in Wash buffer (1x PBS, 10% FBS, 0.01% Tween20)	1 mL

Plate 2		
A	1x PBS at 4 ^{o}C	1 mL
B	1x PBS at 4 ^{o}C	1 mL
C	1x PBS at 4 ^{o}C	1 mL
D	1x PBS at 4 ^{o}C	1 mL
E	1x PBS at 4 ^{o}C	1 mL
F	1x PBS at 4 ^{o}C	1 mL
G	Reserved for 12-tip Deepwell magnetic comb (Thermo Fisher, cat. #97003500)	

The program performed the following protocol:

1. Collect comb from Plate 2 Row G
2. Collect beads from Plate 1 Row H and wash for 30 sec at medium speed
3. Wash beads for 30 sec each at medium speed in Plate 1 Rows G and F
4. Incubate beads with biotinylated protein for 30 mins with slow mixing in Plate 1 Row E
5. Wash immobilized protein for 30 sec each at medium speed in Plate 1 Rows D, C, and B
6. Incubate immobilized protein for 1 hr at 10 ^{o}C with slow mixing in Plate 1 Row A
7. Wash immobilized protein for 2 mins each at medium speed in Plate 2 Rows A through E
8. Elute protein by mixing for 1 min at fast speed in Elution Strips 1 and 2

After affinity selection, samples were purified by STAGE Tip preparation and dried using a vacuum centrifuge. Dried samples were reconstituted into 10.8 µL of nLC-MS/MS mobile phase A and reduced using DTT as described in Section 2.5.6. 4 µL were injected per sample for nLC-MS/MS analysis as described in Section 2.5.11.

2.5.11. Nano-liquid chromatography-tandem mass spectrometry (nLC-MS/MS) analysis

Peptide sequencing was performed on an EASY-nLC 1200 (Thermo Fisher Scientific) nano-liquid chromatography system with an Orbitrap Fusion Lumos Tribrid Mass Spectrometer (Thermo Fisher Scientific). Samples were run on a PepMap RSLC C18 column (2 µm particle size, 15 cm x 50 µm ID; Thermo Fisher Scientific, cat. #ES801) with a nanoViper Trap Column (C18, 3 µm particle size, 100 A pore size, 20 mm x 75 µm ID; Thermo Fisher Scientific, cat. #164946) for desalting. Mobile phase A = water (0.1% FA) and mobile phase B = 80% AcN in water (0.1% FA). Method 1 was used for validation of library samples, and Method 2 was used for analysis of affinity selections.

The ion source voltage was set to 2200 volts in positive mode. Primary mass spectra were detected using the orbitrap at 120000 resolution with a scan range of

300-1400 (m/z), RF lens of 30%, a normalized AGC target of 250% with automatic injection time, and 1 microscan. Candidate ions were chosen for tandem mass spectrometry based on the following criteria: precursor mass range of 300-1200 (m/z), monoisotopic peak determination set to peptides, minimum intensity threshold of 4e4, charge states ranging from +2-+5, dynamic exclusion after 1 observation for 45 seconds with a ±10 ppm range. Fragmentation was done in the orbitrap using HCD followed by EThcD activation types with the following settings: 1.3 m/z isolation window, 30000 resolution, defined first mass of 120 m/z, 600% normalized AGC target with 100 ms maximum injection time, 2 microscans in centroid mode. HCD mode used 28% HCD collision energy and EThcD mode used 25% SA collision energy. Full cycle time for MS^1 and MS^2 scans was 3 seconds.

1. Gradient: linear gradient 1-40% B from 0-35 min; linear gradient 40-90% B from 35-38 min; isocratic 90% B from 38-45 min. Pre-column and analytical column were equilibrated before each run with 8 µL of mobile phase A before sample injection. Samples were loaded using 6 µL of mobile phase A. Mass data was recorded from 3-37 min.
2. Gradient: linear gradient 1-45% B from 0-120 min; linear gradient 45-90% B from 120-123 min; isocratic 90% B from 123-126 min; linear gradient 90-20% B from 126-129 min; isocratic 20% B from 129-132 min; linear gradient 20-90% B from 132-135 min; isocratic 90% B from 135-138 min; linear gradient 90-20% B from 138-141 min; isocratic 20% B from 141-144 min; linear gradient 20-90% B from 144-147 min; isocratic 90% B from 147-152 min. Pre-column and analytical column were equilibrated before each run with 8 µL of mobile phase A before sample injection. Samples were loaded using 12 µL of mobile phase A. Mass data was recorded from 3-120 min.

2.5.12. Automated fast-flow peptide synthesis

Before synthesis, all resins were allowed to swell in amine-free DMF for 15 minutes. Fmoc-Lys(Biotin) was manually coupled by preparing a solution in 1:1

DMF and NMP (5 eq, 0.20M) with PyAOP (0.9 eq relative to amino acid, 0.19M) activated with DIEA (3 eq relative to amino acid) and incubated for 3 hours. The resin was then washed with DMF 3 times before being moved to the automated synthesizer. Utilizing an automated synthesizer, amine-free DMF washed the resin before coupling, after coupling, and after deprotection (40 strokes, ~25 mL). Coupling was performed with HATU (single-coupling, 8 strokes, ~5 mL) except S and A with HATU (double-coupling, 21 strokes, ~10 mL) and C, H, N, Q, R, V, T with PyAOP (double-coupling, 21 strokes, ~10 mL). Deprotection was completed with 20% piperidine in amine-free DMF with 2% formic acid (13 pump strokes, ~5 mL). Amino acids were iteratively coupled and deprotected until the stepwise synthesis was complete. After automated synthesis, the resin was washed again with DMF (3 x 5 mL) and DCM (3 x 5 mL) then dried under reduced pressure. For a detailed explanation of the instrument setup and related chemistries, see Hartrampf et al.[81] or Mijalis et al.[82]

2.5.13. Purification of crude single peptides

Single peptides prepared by automated fast-flow peptide synthesis were cleaved from the resin according to Section 2.5.2. Lyophilized crude peptides were reconstituted in 10% AcN in water (0.1% TFA) at a concentration of 10 mg/mL. Peptides were purified using a Biotage Selekt purification system on a Biotage Sfär C18 Duo (12g, CV = 17 mL, 100 Å, 30 µm, cat. #FSUD-0401-0012) with mobile phase A = water (0.1% TFA) and mobile phase B = AcN (0.1% TFA). The flow rate was controlled according to the system pressure, with a maximum flowrate of 12 mL/min. The following gradient was used: isocratic 10% B for 2 CVs, linear gradient 10-50% B over 12 CVs, linear gradient 50-90% B over 1 CV, isocratic 90% B for 2.5 CVs, isocratic 10% B for 2.5 CVs. Fractions were collected based on absorbance at 214 nm, with a minimum absorbance threshold for collection at 25 mAU and fractionation set based on peak detection. Fractions were subjected to LC-MS analysis as described in Section 2.5.14 before combining the pure fractions and lyophilizing.

2.5.14. Liquid chromatography-mass spectrometry (LC-MS) analysis

LC-MS chromatograms and associated high resolution mass spectra were acquired using an Agilent 1290 Infinity HPLC coupled to an Agilent 6550 LC/Q-TOF mass spectrometer using a Phenomenex Jupiter C4 column (150 x 1.0 mm ID, 5 µm, 300Å silica) heated at 40 °C. Solvent compositions were 0.1% formic acid in water (mobile phase A) and 0.1% formic acid in acetonitrile (mobile phase B). Method 1 was used for fraction analysis after semi-preparative HPLC purification, and method 2 was used for characterization of pure material.

1. Column: Jupiter C4. Gradient: isocratic 1% B from 0-2 min; linear gradient 1-91% B from 2-8 min; isocratic 95% B from 8-10 min; post time 1% B for 1 min. Flow rate: 0.5 mL/min. MS data was collected from 2-8 min; MS was run in positive ionization mode, extended dynamic range (2 GHz), and standard mass range (m/z in the range of 300 to 3000 a.m.u.).
2. Column: Jupiter C4. Gradient: isocratic 1% B from 0-2 min; linear gradient 1-91% B from 2-47 min; isocratic 91% B from 47-49 min; post time 1% B for 2 min. Flow rate: 0.5 mL/min. MS data was collected from 2-47 min; MS was run in positive ionization mode, extended dynamic range (2 GHz), and standard mass range (m/z in the range of 300 to 3000 a.m.u.).

2.5.15. Analytical high-performance liquid chromatography (HPLC)

Analytical HPLC analysis was performed using an Agilent 1200 series system with UV detection at 214 nm on a Zorbax 300SB-C3 column (150 x 2.1 mm ID, 5 µm, 300Å silica) on an Agilent 1200 HPLC at room temperature. Solvent compositions were 0.1% trifluoroacetic acid in water (solvent A) and 0.08% trifluoroacetic acid in acetonitrile (solvent B). Gradient: linear gradient 5-65% B from 0-60 min; linear gradient 65-100% B from 60-61 min; isocratic 100% B from 61-66 min; linear gradient 100-5% B from 66-67 min; isocratic 5% B from 67-75 min. Flow rate: 0.400 mL/min.

2.5.16. Expression of $^{498\text{-}724}$CDH2

$^{498\text{-}724}$CDH2-Avi-HPC4 was expressed using the Expi293 Expression System (Thermo Fisher Scientific). Plasmid DNA was supplied from Novo Nordisk A/S. Cells were cultured in suspension at 37 °C, 8% CO_2 in 1L flasks agitated at 90 rpm. Expression was carried out according to manufacturer protocol. Protein was harvested on Day 6 post-transfection via centrifugation at 5000 rcf for 30 minutes at 4 °C. The supernatant was taken and adjusted to pH 8 using 2M Tris pH 9, then diluted by a factor of 2 with water. Halt Protease Inhibitor Cocktail was added to a final concentration of 1x and the resulting supernatant was filtered through a 0.22 µm filter (Sartoclear Dynamics® Lab V Clarification and Sterile Filtration Kits, Sartorius) and immediately subjected to purification.

2.5.17. Purification of $^{498\text{-}724}$CDH2

The supernatant from Section 2.5.16 was immediately loaded onto a HiTrap Q HP anion exchange chromatography column (5 mL) via a peristaltic pump at a rate of 2 mL/min. The supernatant was recycled through the column and allowed two full volume passes over the column. Protein was then eluted from the column using an ÄKTA Pure chromatography system with mobile phase A = 20 mM Tris, pH 9 and mobile phase B = 20 mM Tris, 1M NaCl, pH 9 using a linear gradient of 0-50% B over 25 CVs. Flowrate was controlled according to column backpressure with a maximum flowrate of 5 mL/min. The eluent was fractionated into 1.7 mL fractions that were then analyzed via western blot based on absorbance at 214 nm. Fractions containing $^{498\text{-}724}$CDH2 were concentrated using a 10K molecular weight cut-off centrifugal concentrator (apparent MW ~ 40000 Da).

The concentrated isolated protein was then subjected to anti-protein C affinity purification using anti-protein C affinity matrix from mouse IgG_1 κ (clone HPC4) according to the manufacturer protocol with column elution using EDTA. The protein was loaded and eluted for 5 cycles to maximize yield. Fractions were analyzed using western blot and concentrated using a 10K molecular weight cut-off centrifugal concentrator. The protein was split in half, with half being biotinylated and the other half proceeding to SEC purification. The concentration of protein was measure by absorbance at 280 nm (Ɛ = 25500 $M^{-1}cm^{-1}$).

A portion of protein was biotinylated using the BirA500 kit from Avidity LLC according to the manufacturer's protocol. The enzymatic reaction was incubated for a total of 24 hours. The remaining portion immediately subjected to SEC purification for use in ligand validation experiments.

The protein was then subjected to SEC purification using an ÄKTA Pure chromatography system with a SuperDex 75 Increase 10/300 GL column (10 x 300 mm, 9 µm particle size from Cytiva Life Sciences). Approximately 1 mg of protein in a volume of 300 µL of buffer was loaded onto the column with a mobile phase of 1x PBS. An isocratic gradient was run for 1.5 CVs at a flowrate of 0.8 mL/min and fractionated at 0.5 mL per fraction. Fractions were analyzed by western blot and concentrated using a 10K molecular weight cut-off centrifugal concentrator. Final protein concentration was measured using absorbance at 280 nm.

2.5.18. Anti-$^{498-724}$CDH2 Western Blot

Protein samples were subjected to SDS-PAGE by dilution into 4x Laemmli sample buffer followed by heating to 75 ^{o}C for 5 minutes. The resulting samples were allowed to cool before loading onto a Mini-PROTEAN TGX Stain-free gel using a Mini-PROTEAN Tetra Cell with a running buffer consisting of 25 mM Tris, 192 mM glycine, 0.1% SDS at pH 8.3. Gels were run at 105V for 75 minutes then washed with ddH_2O three times. The gel was then imaged using the Stain-Free setting of a ChemiDoc imaging system (Bio-Rad) to assess total protein content. Following imaging, the gel was then transferred to a 0.22 µm nitrocellulose membrane using a Trans-blot Turbo Transfer System (Bio-Rad). Following transfer, the blot was rinsed three times with ddH_2O and blocked with 5% nonfat dry milk in 1x PBS for 1 hour at room temperature. The liquid was then decanted and the blot was stained with 1:1000 HPC4 tag antibody (Cell Signaling Technology, cat.

#68083) for 1 hour at room temperature. The membrane was washed with 1x PBS with 0.1% Tween-20 5 times for 5 minutes each. The membrane was then stained with Goat anti-Rabbit IgG (H+L) secondary antibody, HRP (Thermo Fisher Scientific, cat. #32460) for 1 hour at room temperature. The membrane was washed with 1x PBS with 0.1% Tween-20 5 times for 5 minutes each then imaged using SuperSignal West Pico PLUS Chemiluminescent Substrate (Thermo Fisher Scientific, cat. #34580) on a ChemiDoc imager (Bio-Rad).

2.5.19. Biolayer interferometry (BLI)

Peptide binding validation was performed using a Gator Plus Next Generation Biolayer Interferometry instrument. All assays were run at 30 °C and agitated at 1000 rpm. Streptavidin-coated probes (Gator Bio cat. #160029) were dipped into 0.5 µg/mL solutions of biotinylated peptide solution in kinetic buffer (1x PBS with 0.02% BSA and 0.002% Tween-20) for immobilization for 5 minutes. All sequences have Lys(Biotin)-Ahx attached to the N-terminus in addition to the sequence shown. The probes were then moved to a dilution series of protein (500, 250, 125, and 62.5 nM) for 10 minutes to obtain the association curve. The tips were then moved into a new column of wells with kinetic buffer and incubated for 10 minutes to obtain the dissociation curve. Peptide-only and protein-only (concentration at 500 nM) were used for background subtraction. Apparent dissociation constants (KD) were calculated using the global Rmax unlinked algorithm with a 1:1 binding model as implemented in the Gator Bio data analysis software.

2.6. Appendix I: Synthesis and characterization data

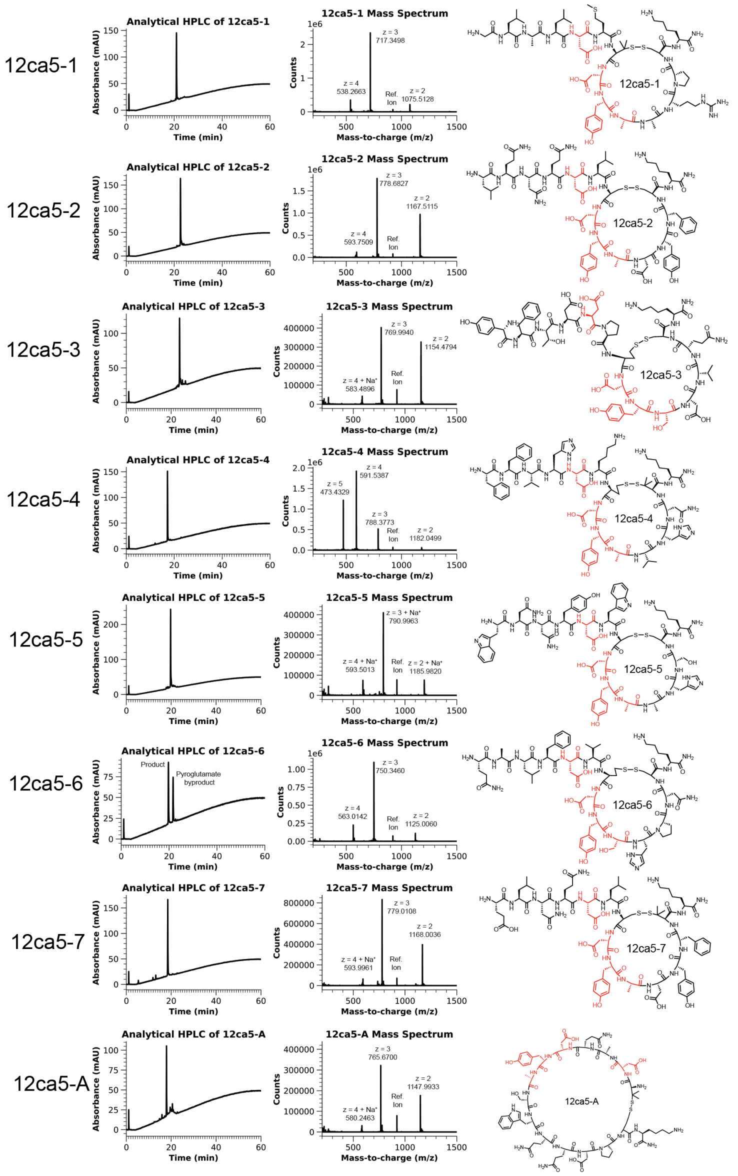


***Figure 2.7** LCMS and analytical HPLC data for the synthesis of identified 12ca5 binding peptides. All sequences have Gly-Ser-Lys(Biotin) attached to the C-terminus in addition to the structure shown. 12ca5-1: 95% pure, calc. mass 2148.9941, obs. mass 2149.02283 (+13.4 ppm); 12ca5-2: 85% pure, calc. mass 2332.9916, obs. mass 2333.01773 (+11.2 ppm); 12ca5-3: 75% pure, calc. mass 2306.9283, obs. mass 2306.9469 (+8.1 ppm); 12ca5-4: 91% pure, calc. mass 2362.0810, obs. mass 2362.0951 (+6.0 ppm); 12ca5-5: 84% pure, calc. mass 2346.9245, obs. mass 2346.9731 (+20.7 ppm); 12ca5-6: 62% pure; The minor peak is a result of pyroglutamate formation, a common side reaction for N-terminal glutamine residues, calc. mass 2247.9864, obs. mass 2248.0119 (+11.3 ppm); 12ca5-7: 88% pure, calc. mass 2333.9756, obs. mass 2333.9737 (-0.8 ppm); 12ca5-A: 70% pure, calc. mass 2293.9555, obs. mass 2293.9738 (+8.0 ppm).*

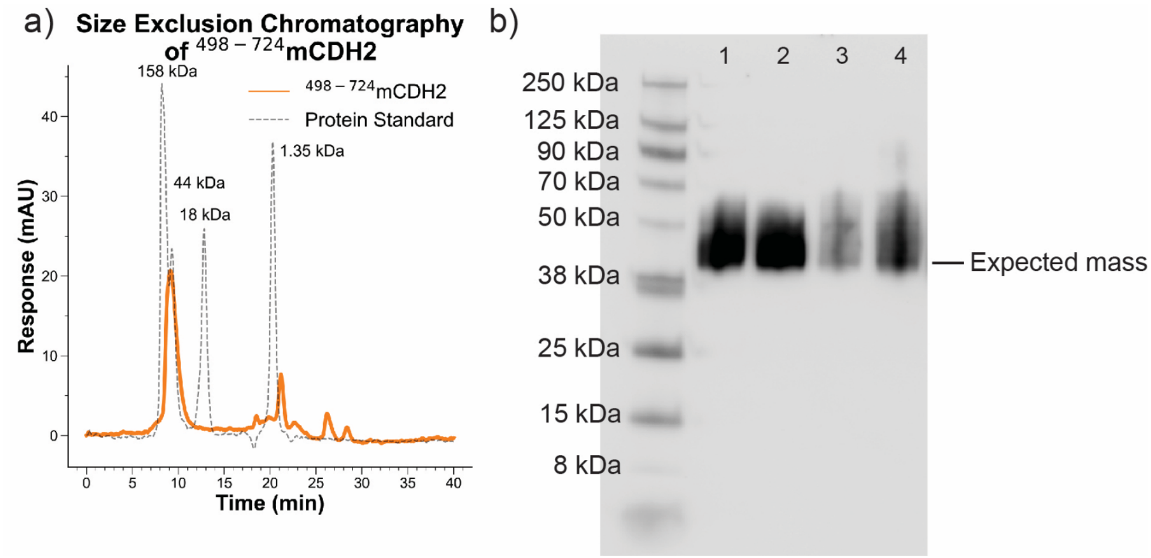


Figure 2.8: *Characterization of expressed $^{498\text{-}724}$CDH2. A) Size exclusion chromatogram of absorbance at 214 nm of $^{498\text{-}724}$CDH2 compared against BioRad Gel Filtration Standard (Cat. #1511901) shows the expected mass for the protein. Sample peaks eluting after 20 minutes were found to be buffer constituents. B) α-HPC4 tag Western blot against $^{498\text{-}724}$CDH2. Lane 1: biotinylated $^{498\text{-}724}$CDH2; lane 2: reduced biotinylated $^{498\text{-}724}$CDH2; lane 3: nonbiotinylated $^{498\text{-}724}$CDH2; lane 4: reduced nonbiotinylated $^{498\text{-}724}$CDH2.*

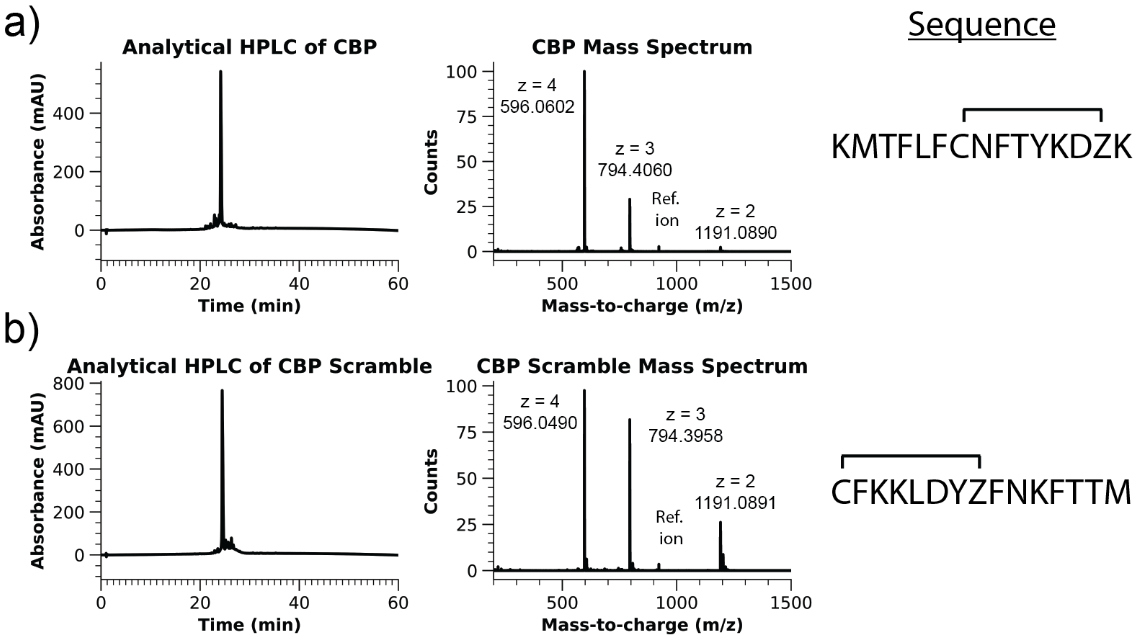


Figure 2.9: *Analytical HPLC and LCMS characterization of **CBP** and **CBP** sequence scramble. Calculated mass of both sequences: 2380.1605. **CBP**: 85% purity, observed mass 2380.1886 (+11.8 ppm); **CBP** Scramble: 84% purity, observed mass 2380.1635 (+1.3 ppm).*

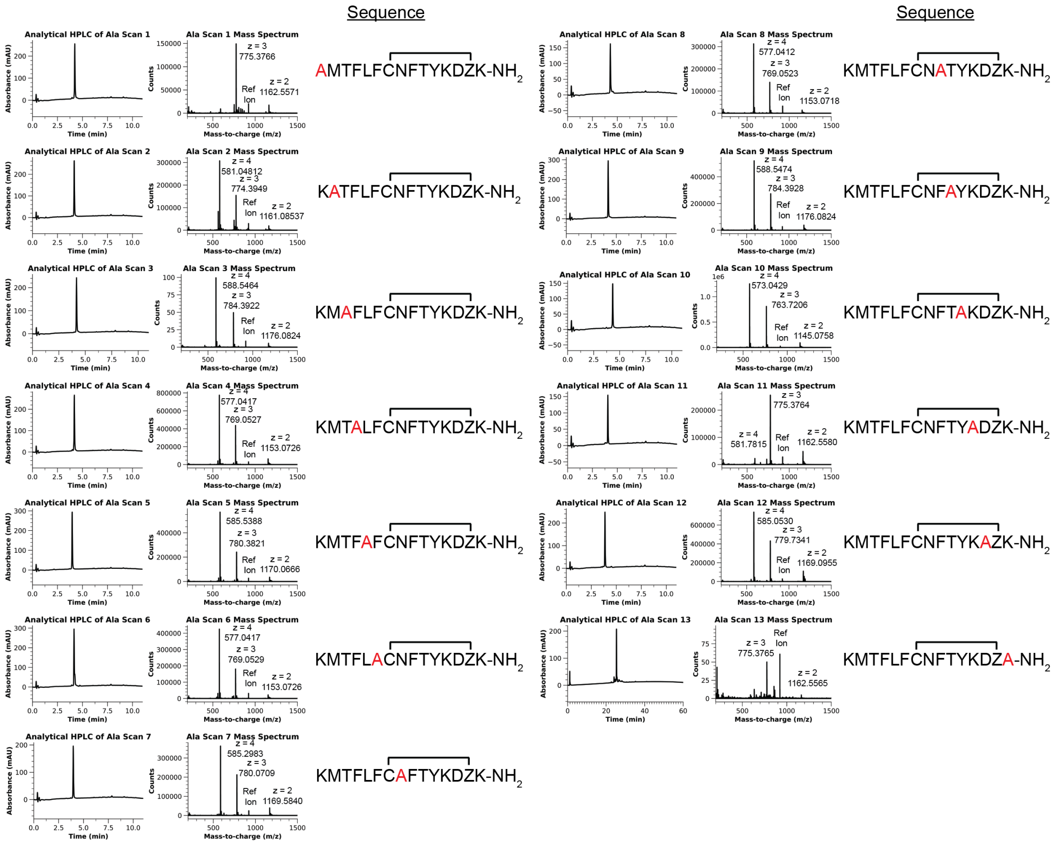


***Figure 2.10:** LCMS and analytical HPLC data for the synthesis of **CBP** alanine scan peptides. All sequences have Lys(Biotin)-Ahx attached to the N-terminus in addition to the sequence shown. Ala scan 1: 97% pure, calc. mass 2323.1027, obs. mass 2323.1023 (-0.2 ppm); ala scan 2: 97% pure, calc. mass 2320.1571, obs. mass 2320.1589 (+0.8 ppm); ala scan 3: 96% pure, calc. mass 2350.1499, obs mass 2350.1420 (+0.9 ppm); ala scan 4: 95% pure, calc. mass 2304.1292, obs. mass 2304.1330 (+1.7 ppm); ala scan 5: 99% pure, calc. mass 2338.1136, obs. mass 1338.1212 (+3.3 ppm); ala scan 6: 77% pure, calc. mass 2304.1292, obs. mass 2304.1332 (+1.7 ppm); ala scan 7: 99% pure, calc. mass 2337.1546, obs. mass 2337.1676 (+5.5 ppm); ala scan 8: 94% pure, calc. mass 2304.1292, obs. mass 2304.1314 (+1.0 ppm); ala scan 9: 96% pure, calc. mass 2350.1499, obs. mass 2350.1539 (+1.7 ppm); ala scan 10: 94% pure, calc. mass 2288.1343, obs. mass 2288.1380 (+1.6 ppm); ala scan 11: 94% pure, calc. mass 2323.1027, obs. mass 2323.1000 (-1.1 ppm); ala scan 12: 95% pure, calc. mass 2336.1707, obs. mass 2336.1781 (+3.2 ppm); ala scan 13: 75% pure, calc. mass 2323.1037, obs. mass 2323.1015 (-0.5 ppm).*

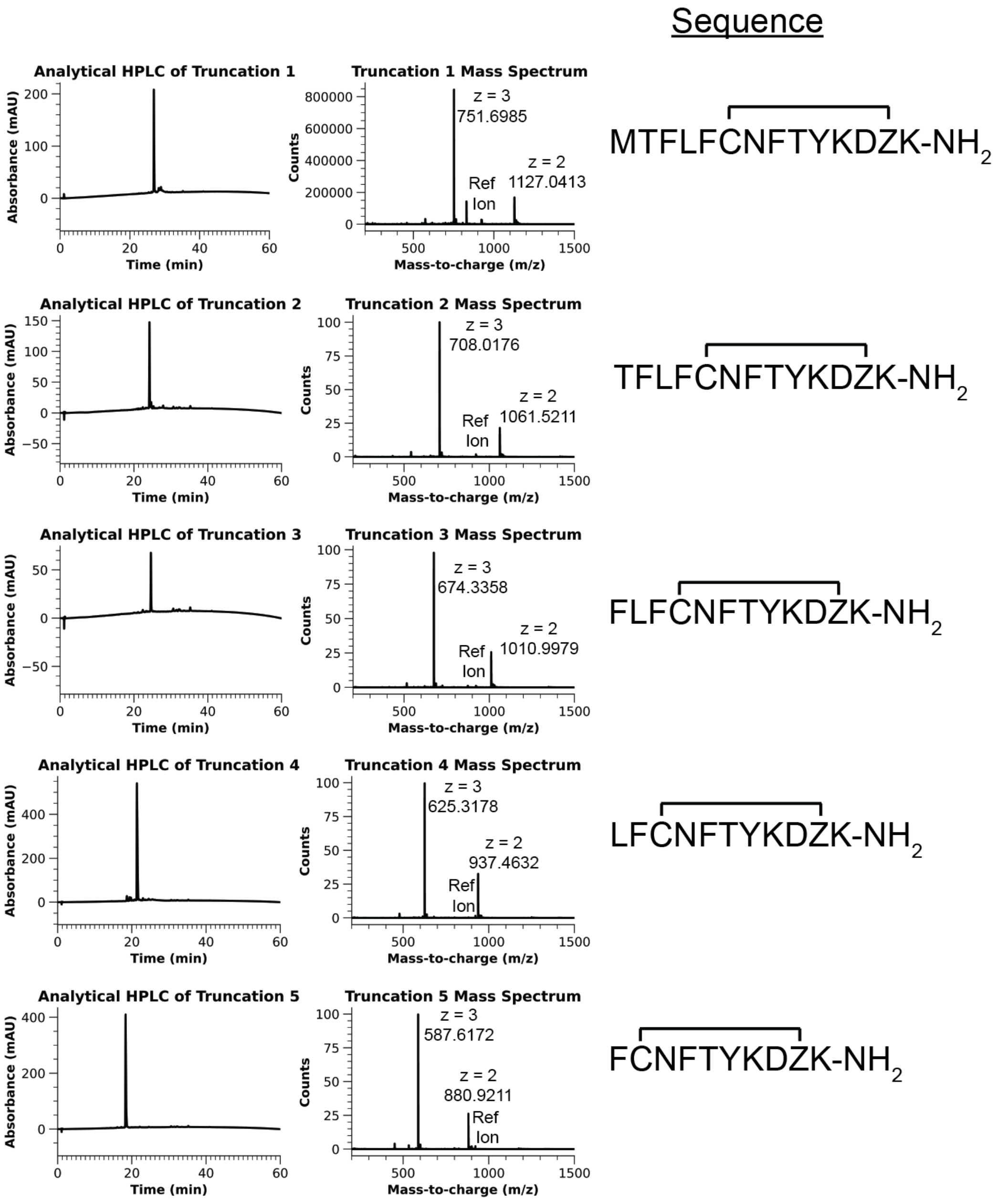


***Figure 2.11**. LCMS and analytical HPLC data for the synthesis of **CBP** truncation study peptides. All sequences have Lys(Biotin)-Ahx attached to the N-terminus in addition to the sequence shown. Truncation 1: 88% pure, calc. mass 2252.0655, obs. mass 2252.0693 (+1.7 ppm); truncation 2: 86% pure, calc. mass 2121.0251, obs. mass 2121.0278 (+1.3 ppm); truncation 3: 87% pure, calc. mass 2019.9774, obs. mass 2019.9819 (+2.2 ppm); truncation 4: 90% pure, calc. mass 1872.9090, obs. mass 1872.9202 (+6.0 ppm); truncation 5: 97% pure, calc. mass 1759.8249, obs. mass 1759.8272 (+1.3 ppm).*

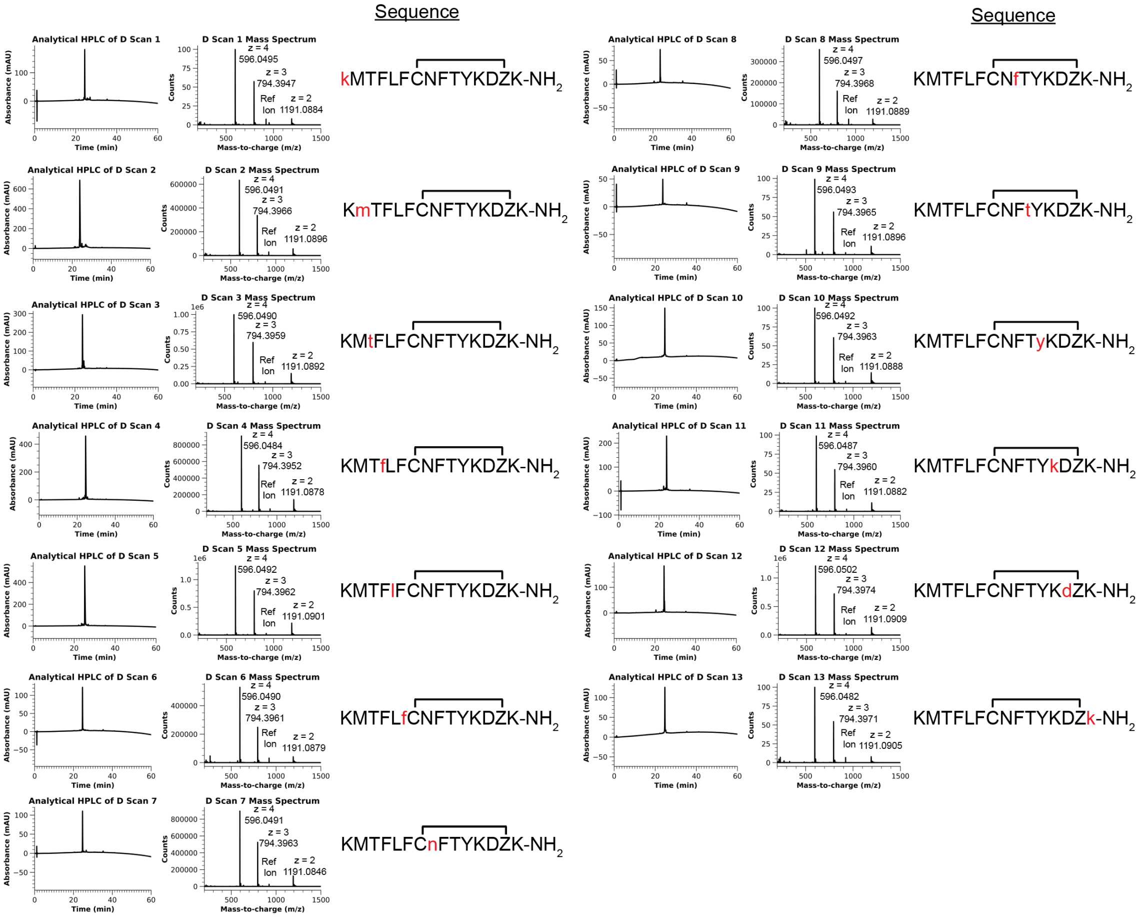


***Figure 2.12:** LCMS and analytical HPLC data for the synthesis of **CBP** d-amino acid scan peptides. All sequences have Lys(Biotin)-Ahx attached to the N-terminus in addition to the sequence shown. Calc. mass for all peptides: 2380.1605. D scan 1: 89% pure, obs. mass 2380.1626 (+0.9 ppm); D scan 2: 85% pure, obs. mass 2380.1648 (+1.8 ppm); D scan 3: 81% pure, obs. mass 2380.1637 (+1.3 ppm); D scan 4: 90% pure, obs. mass 2380.1612 (+0.3 ppm); D scan 5: 92% pure, obs. mass 2380.1782 (+7.4 ppm); D scan 6: 95% pure, obs. mass 2380.1630 (+1.1 ppm); D scan 7: 85% pure, obs. mass 2380.1611 (+0.3 ppm); D scan 8: 84% pure, obs. mass 2380.1653 (+2.0 ppm); D scan 9: 80% pure, obs. mass 2380.1649 (+1.9 ppm); D scan 10: 98% pure, obs. mass 2380.1641 (+1.5 ppm); D scan 11: 86% pure, obs. mass 2380.1627 (+0.9 ppm); D scan 12: 76% pure, obs. mass 2380.1679 (+3.1 ppm); D scan 13: 91% pure, obs. mass 2380.1647 (+1.8 ppm).*

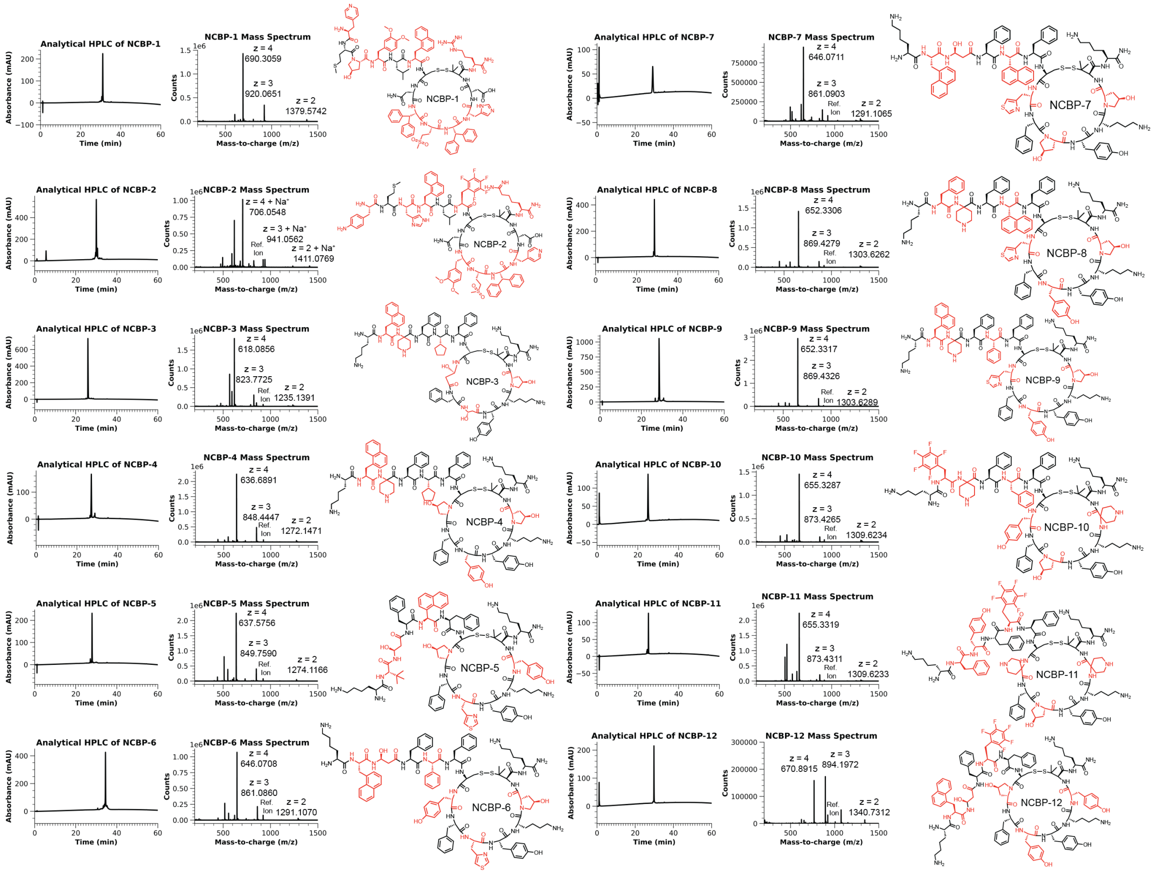


Figure 2.13: *LCMS and analytical HPLC data for the synthesis of* ***NCBP*** *peptides. All sequences have Lys(Biotin)-Ahx attached to the N-terminus. Residues mutated from the original* ***CBP*** *sequence are highlighted in red.* ***NCBP-1****: 86% pure, calc. mass 2757.1864, obs. mass 2757.1654 (-7.6 ppm);* ***NCBP-2****: 75% pure, calc. mass 2797.1670, obs. mass 2797.0775 (+32.0 ppm);* ***NCBP-3****: 95% pure, calc. mass 2468.2616, obs. mass 2468.2890 (+11.1 ppm);* ***NCBP-4****: 80% pure, calc. mass 2542.2773, obs. mass 2542.4379 (+63 ppm);* ***NCBP-5****: 88% pure, calc. mass 2546.2472 obs. mass 2546.2179 (+11 ppm);* ***NCBP-6****: 85% pure, calc. mass 2580.2025, obs. mass 2580.2280 (+10 ppm);* ***NCBP-7****: 98% pure, calc. mass 2580.2023, obs. mass 2580.2324 (+12 ppm);* ***NCBP-8****: 89% pure, calc. mass 2605.2340, obs. mass 2605.2625 (+11 ppm);* ***NCBP-9****: 88% pure, calc. mass 2605.2340, obs. mass 2605.2704 (+14 ppm);* ***NCBP-10****: 96% pure, calc. mass 2617.2309, obs. mass 2617.2567 (+10 ppm);* ***NCBP-11****: 79% pure, calc. mass 2617.2309, obs. mass 2617.2739 (+16 ppm);* ***NCBP-12****: 94% pure, calc. mass 2679.1986, obs. mass 2679.5163 (+120 ppm).*

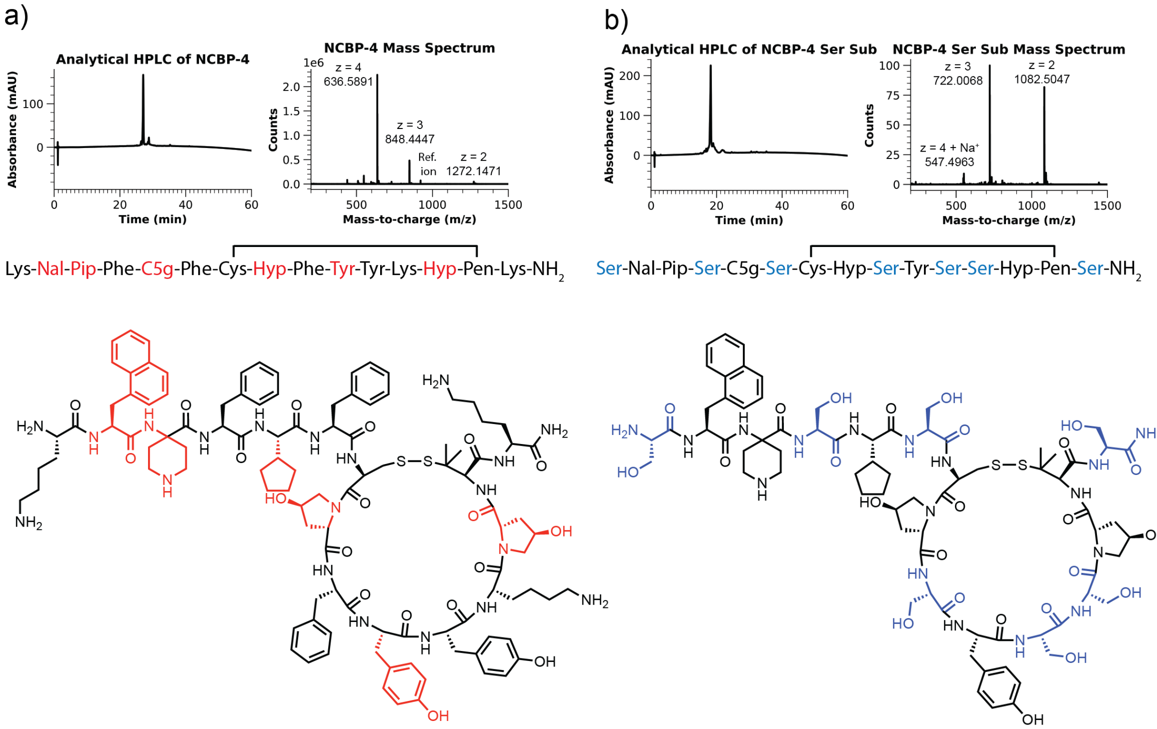


Figure 2.14: *Analytical HPLC and LCMS characterization of **NCBP-4** and **NCBP-4** serine substituted sequence. Red residues denote the mutated cold-spots from **CBP**, and blue residues denote the hot-spots from **CBP** mutated to serine. **NCBP-4**: 80% purity, calc. mass 2542.2769, obs. mass 2543.3046 (+10.9 ppm); NCBP-4 serine substituted sequence: 82% purity, calc. mass 2162.9475, obs. mass 2162.9847 (+17.2 ppm).*

2.7. Appendix II: Library validation data

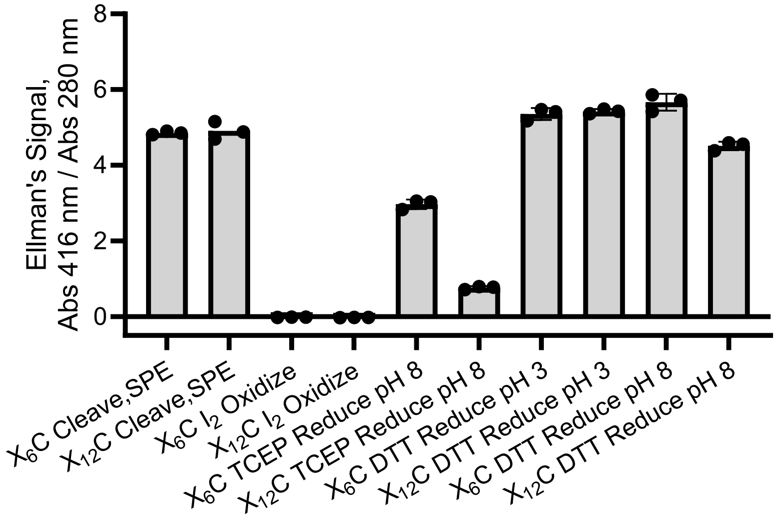


Figure 2.15: *Ellman's assay studies on reduction reaction conditions identify reduction using dithiothreitol (DTT) in acidic conditions as the most optimal for the reduction of disulfide bonds within the macrocyclic peptide libraries.*

Table 2.3: *Components of the custom peptide standard used for SEC validations. The average molecular weight of a member of either macrocyclic peptide library is 1802.*

Sequence	Approx. Molecular Weight
SQETFSDLWKLLPEN	1807
HGPATPRMAKFDQAAGDQYMAGMDKRKAGRAAGATL	3748
MNSTESIPLAQSTVAQSTVAGFTSELESTPVPSNETTCENWREIHHLVFHVA	5685

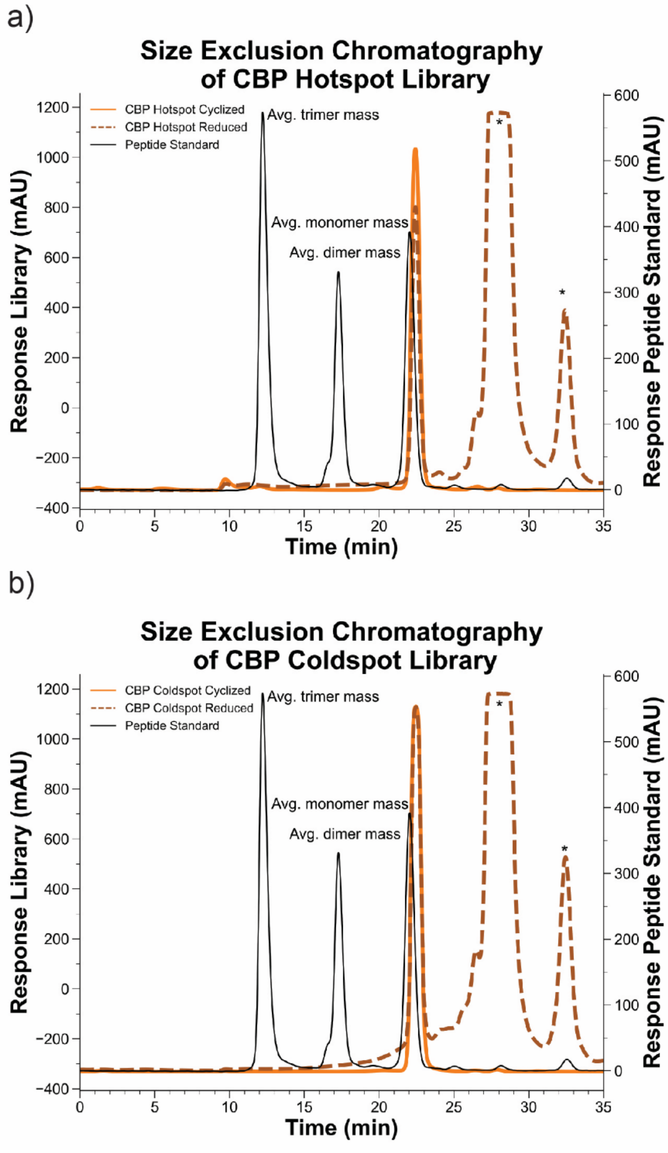


Figure 2.16: *Size exclusion chromatograms of absorbance at 214 nm of a)* ***CBP*** *hotspot and b)* ***CBP*** *coldspot macrocyclic libraries were compared to molecular weight standards corresponding to the average mass of monomeric, dimeric, and trimeric species. Library samples were ran using the cyclized form (later used in affinity selection experiments) and*

the DTT linearized form, demonstrating the formation of intramolecular disulfide bonds. Peaks marked with an asterisk (*) were residual elements from the sample buffer.

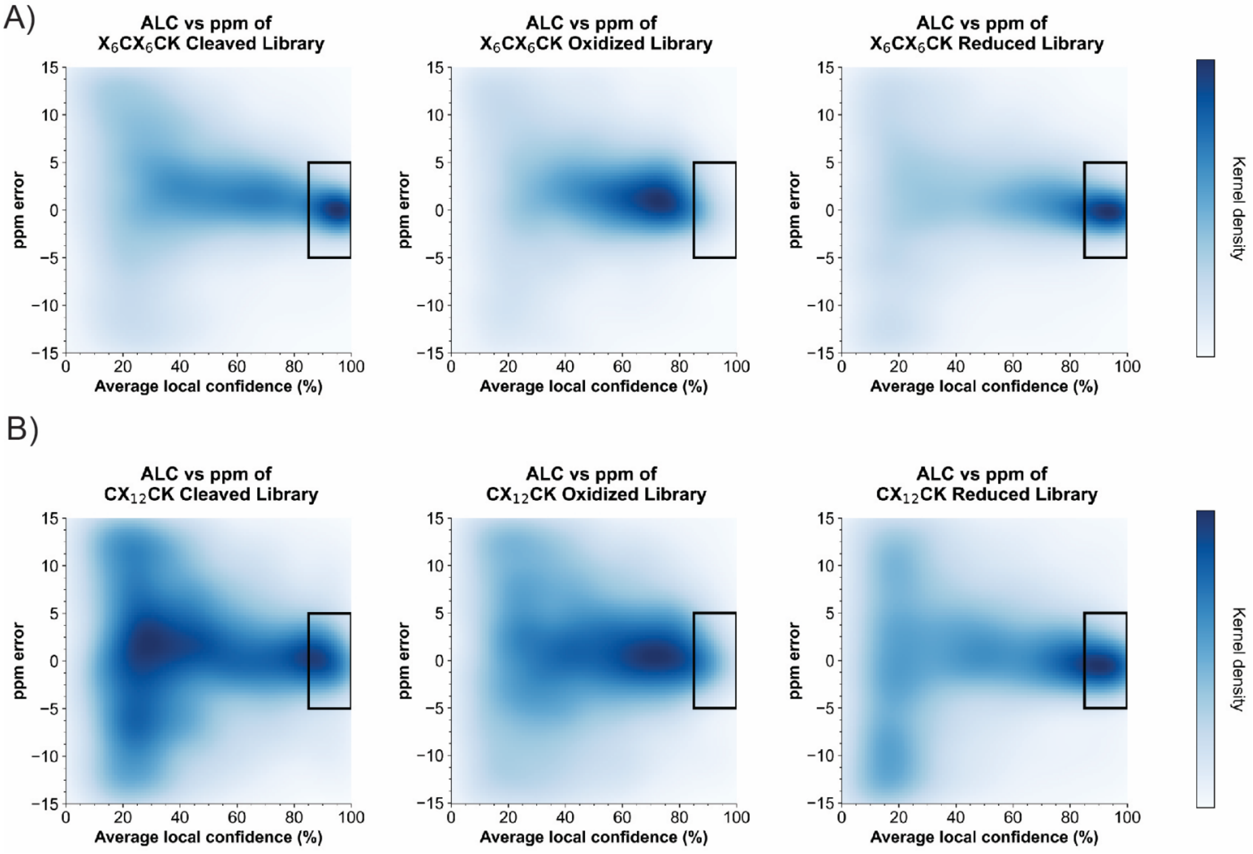


Figure 2.17: *Heatmaps comparing average local confidence (ALC %) of sequence assignment by PEAKS Studio 8.5 to instrument error (ppm) showed the loss and recovery of sequencing capabilities after oxidation and reduction respectively for the (a) X_6CX_6CK and (b) $CX_{12}CK$ macrocyclic peptide library designs. High density within the black box region were regarded as high fidelity de novo sequence assignments. A high density within this region is expected when the library is in a reduced form either directly post cleavage from the resin or after reduction using DTT. Conversely, a low density within the boxed region is expected when the library is in an oxidated state.*

2.8. Appendix III: Affinity selection and biolayer interferometry data

Table 2.4: *Sequences isolated by AS-MS that contain the characteristic binding motif to 12ca5, D**DY(A/S). Assigned local confidences (ALC) for each sequence assignment are given, as well as the ligand affinity to 12ca5 as measure by BLI. Retention time (RT, in minutes), mass-to-charge ratio (m/z), charge state (z), observed mass (Mass) and mass error in sequence assignment (ppm) are given.*

Peptide	Library	Sequence															ALC (%)	K_D, nM	RT	m/z	z	Mass	ppm
12ca5-1	X_6CX_6CK	Gly	Leu	Ala	Leu	Asp	Met	Pen	Asp	Tyr	Ala	Ala	Arg	Pro	Cys	Lys	99	4.5	33.28	557.2661	3	1668.7786	-1.2
12ca5-2	X_6CX_6CK	Leu	Gln	Asn	Gln	Asp	Leu	Cys	Asp	Tyr	Ala	Asp	Tyr	Phe	Cys	Lys	97	<1	59.07	919.3979	2	1836.781	0.2
12ca5-3	X_6CX_6CK	Tyr	Phe	Thr	Asp	Asp	Pro	hCys	Asp	Tyr	Ser	Asp	Val	Gln	Cys	Lys	99	1.5	43.89	906.3663	2	1810.7178	0.2
12ca5-4	X_6CX_6CK	Phe	Phe	Val	His	Asp	Lys	hCys	Asp	Tyr	Ala	Val	His	Gln	Pen	Lys	99	2.1	29.58	467.4745	4	1865.8706	-0.9
12ca5-5	X_6CX_6CK	Trp	Asn	Asn	Tyr	Asp	Trp	Cys	Asp	Tyr	Ala	Ala	His	Ser	Cys	Lys	91	3.1	35.18	625.583	3	1873.73	-1.5
12ca5-6	X_6CX_6CK	Gln	Ala	Leu	Phe	Asp	Val	hCys	Asp	Tyr	Ser	His	Pro	Asn	Cys	Lys	89	<1	41.67	584.9335	3	1751.7759	1.5
12ca5-7	X_6CX_6CK	Glu	Leu	Asn	Gln	Asp	Leu	Cys	Asp	Tyr	Ala	Asp	Tyr	Phe	Cys	Lys	98	<1	58.7	919.8975	2	1837.7651	8.4
12ca5-A	$CX_{12}CK$	Pen	Asp	Ala	Gln	Asp	Tyr	Ala	Ser	Trp	Gln	Gln	Asp	Pro	Pen	Lys	70	1.0	35.85	600.2608	3	1797.7451	8.6

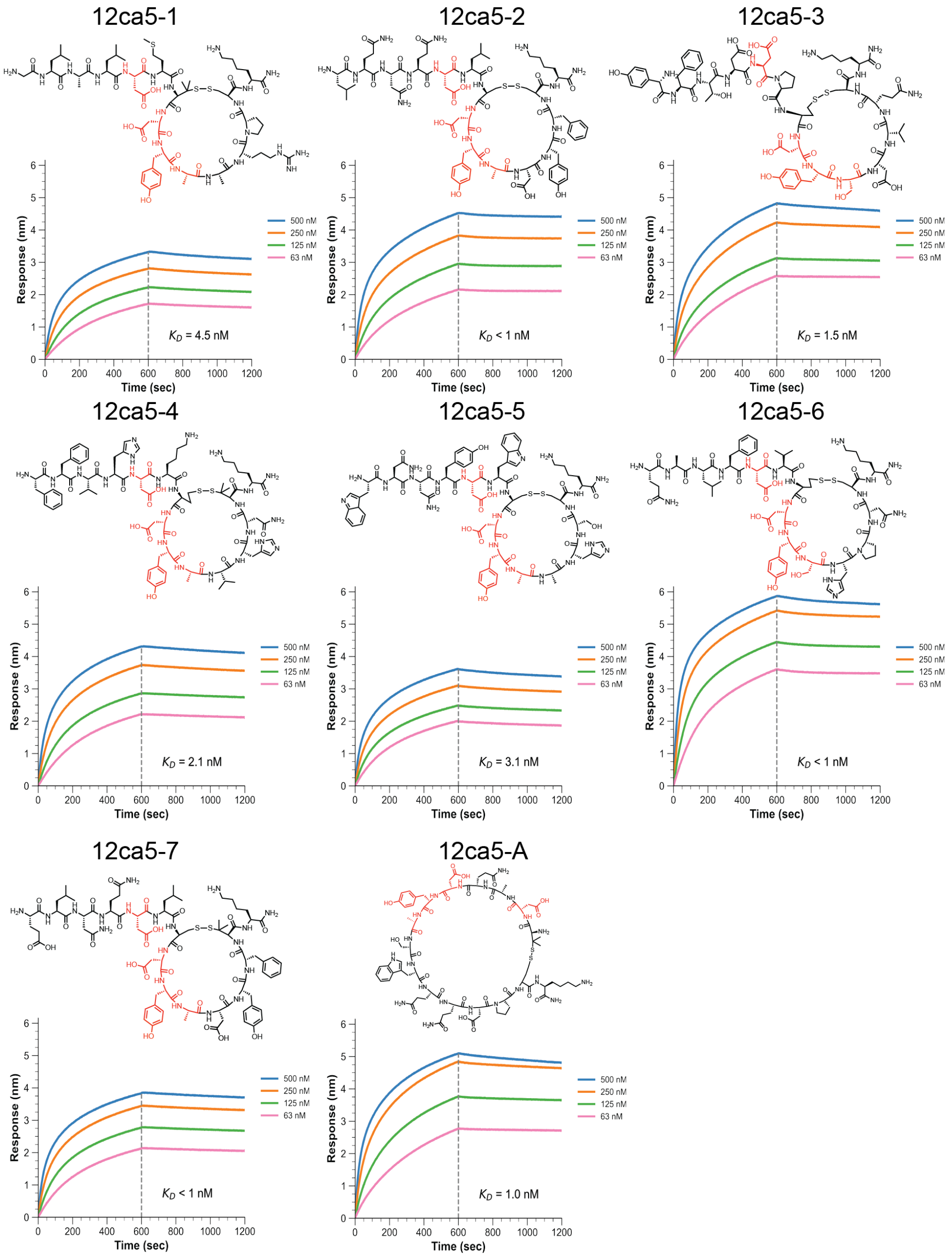


Figure 2.18: *Full BLI data for all identified sequences containing the 12ca5 binding motif, D**DY(A/S). The motif is highlighted in red within each structure. All sequences have Gly-Ser-Lys(Biotin) attached to the C-terminus in addition to the structure shown.*

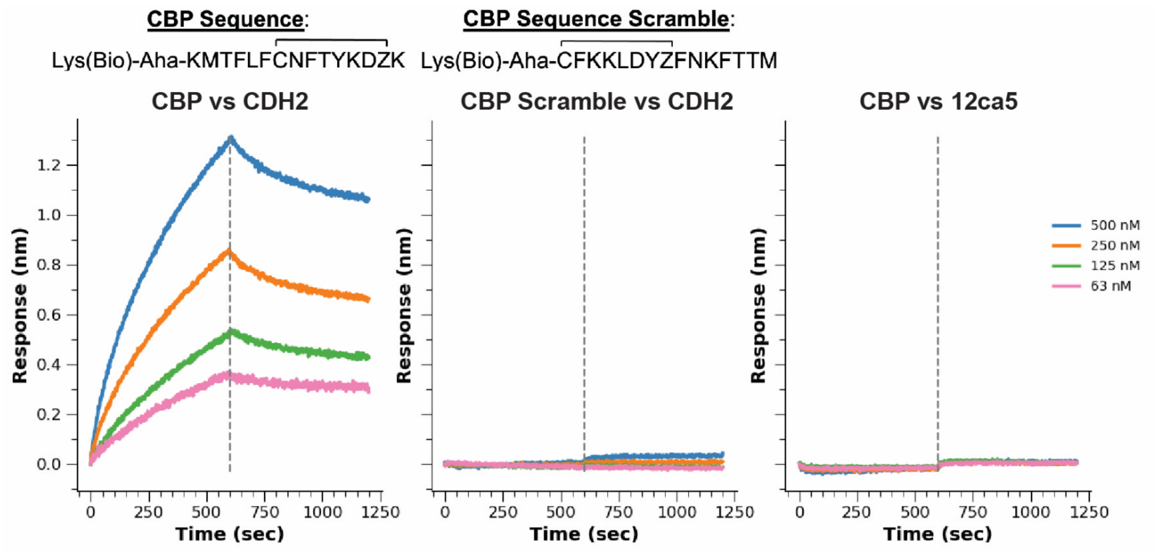


***Figure 2.19: CBP** sequence scramble and off-target controls show sequence and protein specificity of the ligand interaction towards CDH2 by BLI. BLI data for **CBP** tested against CDH2 is shown for reference.*

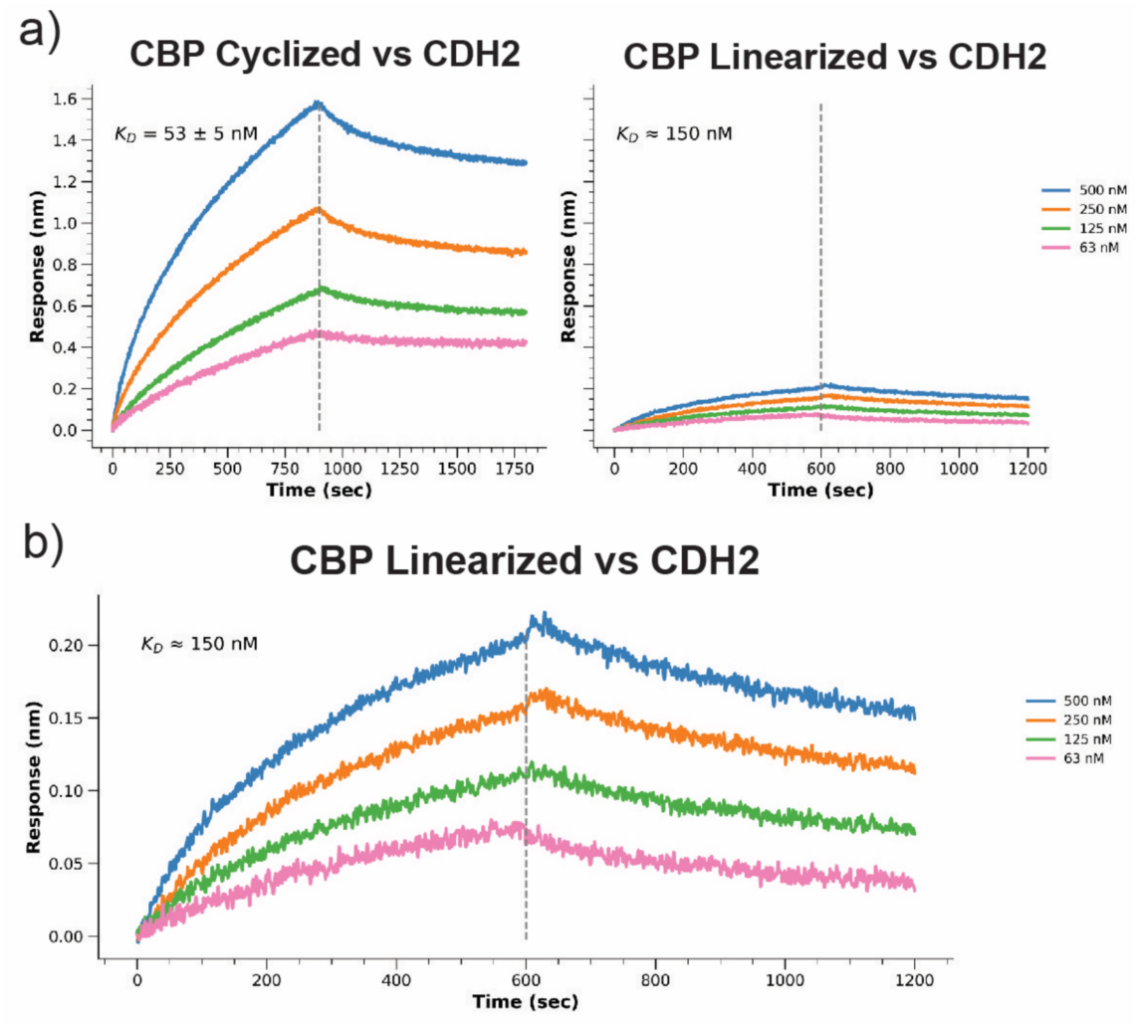


Figure 2.20: *The macrocyclic structure of CBP plays a role in the strength of the observed interactions, a) where a comparison of the cyclized CBP peptide and the DTT-linearized CBP sequence shows a large decrease in observed signal, as well as a drop in the observed dissociation constant. b) A magnified version of the BLI trace for the linearized CBP sequence is given.*

***Table 2.5:** Alanine scan of **CBP** shows critical residues for ligand interactions with $^{498\text{-}724}$CDH2. Each peptide is cyclized through disulfide bond formation of the side chains of cysteine (C) and penicillamine (Z).*

Modified Residue	Sequence	Approx. K_D (nM)
Original	KMTFLF$\overline{\text{CNFTYKDZ}}$K-$NH_2$	50
Lys1	**A**MTFLF$\overline{\text{CNFTYKDZ}}$K-$NH_2$	250
Met2	K**A**TFLF$\overline{\text{CNFTYKDZ}}$K-$NH_2$	70
Thr3	KM**A**FLF$\overline{\text{CNFTYKDZ}}$K-$NH_2$	60
Phe4	KMT**A**LF$\overline{\text{CNFTYKDZ}}$K-$NH_2$	No binding
Leu5	KMTF**A**F$\overline{\text{CNFTYKDZ}}$K-$NH_2$	80
Phe6	KMTFL**A**$\overline{\text{CNFTYKDZ}}$K-$NH_2$	No binding
Asn8	KMTFLF$\overline{\text{C}\textbf{A}\text{FTYKDZ}}$K-$NH_2$	100
Phe9	KMTFLF$\overline{\text{CN}\textbf{A}\text{TYKDZ}}$K-$NH_2$	120
Thr10	KMTFLF$\overline{\text{CNF}\textbf{A}\text{YKDZ}}$K-$NH_2$	50
Tyr11	KMTFLF$\overline{\text{CNFT}\textbf{A}\text{KDZ}}$K-$NH_2$	70
Lys12	KMTFLF$\overline{\text{CNFTY}\textbf{A}\text{DZ}}$K-$NH_2$	No binding
Asp13	KMTFLF$\overline{\text{CNFTYK}\textbf{A}\text{Z}}$K-$NH_2$	50
Lys15	KMTFLF$\overline{\text{CNFTYKDZ}}$**A**-$NH_2$	No binding

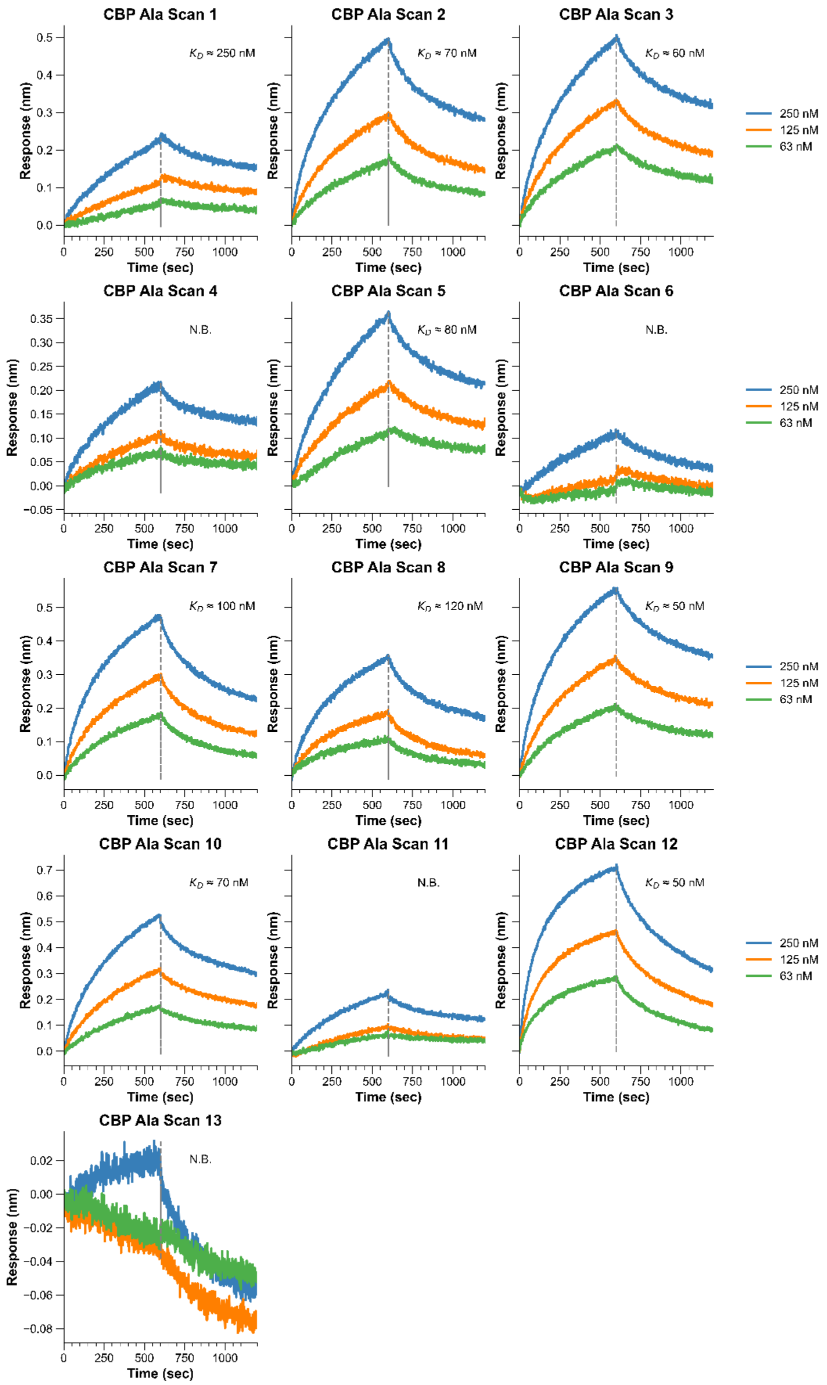


Figure 2.21: *BLI data of CBP alanine scan peptides against $^{498\text{-}724}$CDH2. Traces with minimal response were deemed non-binders.*

Table 2.6: *Truncation studies of **CBP** demonstrate the impact of the N-terminal residues in ligand interactions with $^{498\text{-}724}$CDH2. Each peptide is cyclized through disulfide bond formation of the side chains of cysteine (C) and penicillamine (Z).*

Modified Residue	Sequence	Approx. K_D (nM)
Original	KMTFLF$\overline{\text{CNFTYKDZ}}$K-NH$_2$	50
Lys1	MTFLF$\overline{\text{CNFTYKDZ}}$K-NH$_2$	60
Met2	TFLF$\overline{\text{CNFTYKDZ}}$K-NH$_2$	70
Thr3	FLF$\overline{\text{CNFTYKDZ}}$K-NH$_2$	100
Phe4	LF$\overline{\text{CNFTYKDZ}}$K-NH$_2$	190
Leu5	F$\overline{\text{CNFTYKDZ}}$K-NH$_2$	No binding

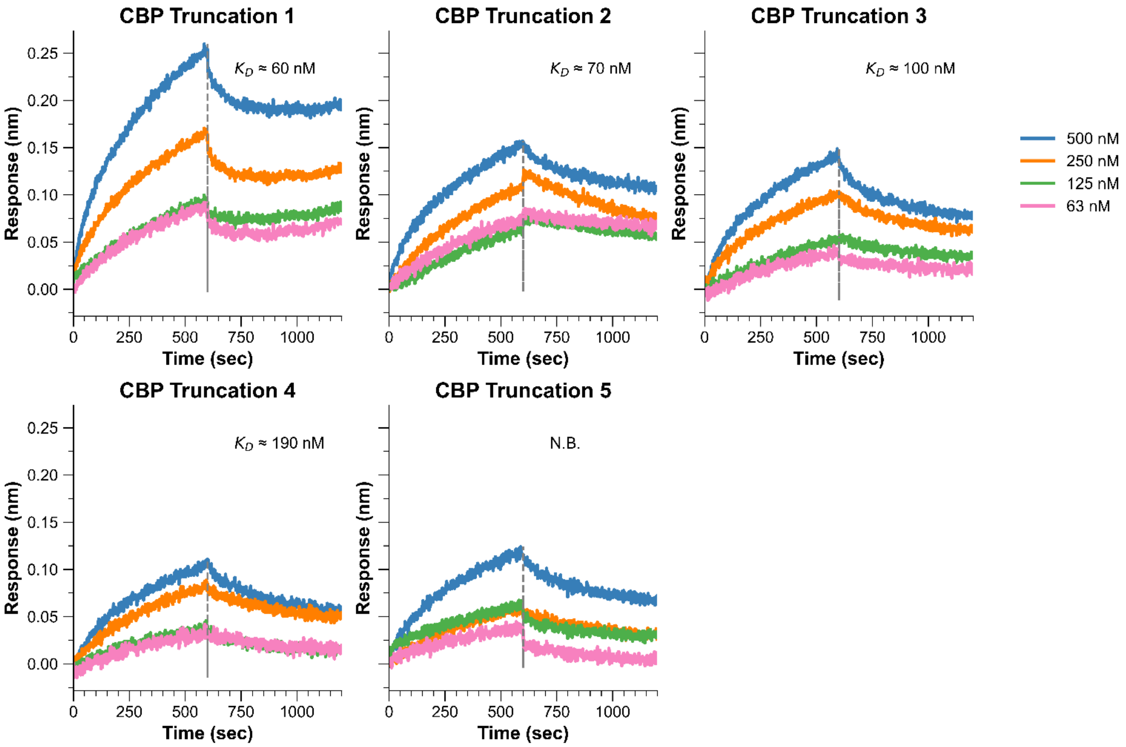


Figure 2.22: *BLI data of **CBP** truncated peptides against $^{498\text{-}724}$CDH2. Traces with minimal response were deemed non-binders.*

Table 2.7: *D-amino acid scan of* ***CBP*** *identifies critical stereocenters in ligand interactions with* $^{498\text{-}724}$*CDH2. Each peptide is cyclized through disulfide bond formation of the side chains of cysteine (C) and penicillamine (Z).*

Modified Residue	Sequence	Approx. K_D (nM)
Original	KMTFLF$\overline{\text{CNFTYKDZ}}$K-$NH_2$	50
Lys1	**k**MTFLF$\overline{\text{CNFTYKDZ}}$K-$NH_2$	50
Met2	K**m**TFLF$\overline{\text{CNFTYKDZ}}$K-$NH_2$	70
Thr3	KM**t**FLF$\overline{\text{CNFTYKDZ}}$K-$NH_2$	140
Phe4	KMT**f**LF$\overline{\text{CNFTYKDZ}}$K-$NH_2$	60
Leu5	KMTF**l**F$\overline{\text{CNFTYKDZ}}$K-$NH_2$	50
Phe6	KMTFL**f**$\overline{\text{CNFTYKDZ}}$K-$NH_2$	3000
Asn8	KMTFLF$\overline{\text{C}\textbf{n}\text{FTYKDZ}}$K-$NH_2$	60
Phe9	KMTFLF$\overline{\text{CN}\textbf{f}\text{TYKDZ}}$K-$NH_2$	150
Thr10	KMTFLF$\overline{\text{CNF}\textbf{t}\text{YKDZ}}$K-$NH_2$	60
Tyr11	KMTFLF$\overline{\text{CNFT}\textbf{y}\text{KDZ}}$K-$NH_2$	No binding
Lys12	KMTFLF$\overline{\text{CNFTY}\textbf{k}\text{DZ}}$K-$NH_2$	60
Asp13	KMTFLF$\overline{\text{CNFTYK}\textbf{d}\text{Z}}$K-$NH_2$	50
Lys15	KMTFLF$\overline{\text{CNFTYKDZ}}$**k**-$NH_2$	No binding

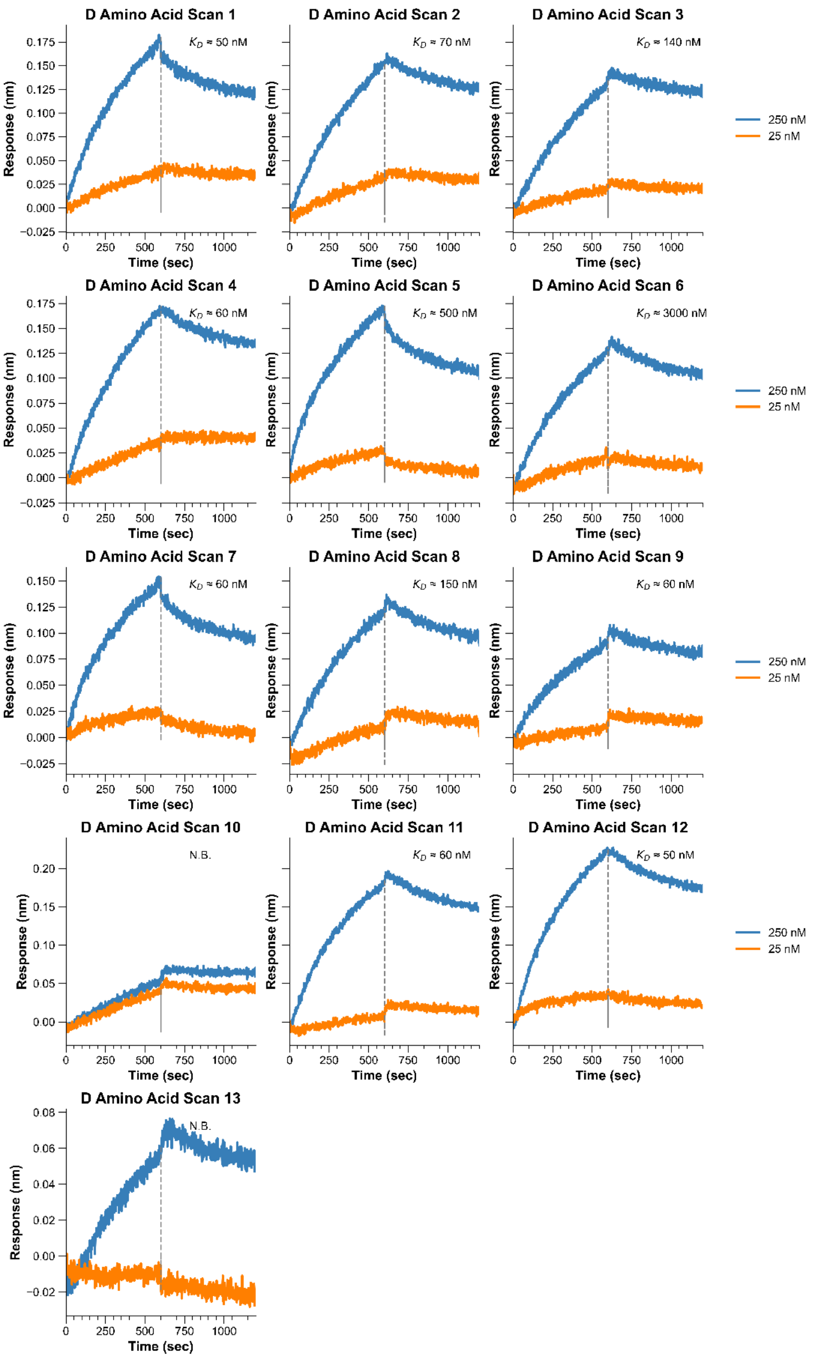


Figure 2.23: *BLI data of* ***CBP*** *d-amino acid scan peptides against* $^{498\text{-}724}$*CDH2. Traces with minimal response were deemed non-binders.*

Table 2.8: *List of top candidates identified from AS-MS experiments utilizing focused libraries against ligand $^{498\text{-}724}$CDH2. Assigned local confidences (ALC) are given for each sequence, as well as retention time (RT), mass-to-charge ratio observed (m/z), charge state (z), and observed parent mass (Mass) with mass error of the sequence assignment (ppm).*

Peptide	Library	Sequence															ALC (%)	RT	m/z	z	Mass	ppm
NCBP-1	Hotspot	4Py	Met	Hyp	Dmf	Leu	Nal	Cys	Asn	Dph	Msn	Dph	His	Asp	Pen	Arg	97	87.35	764.9832	3	2291.9443	-7.2
NCBP-2	Hotspot	Amf	Met	His	Nal	Leu	Pff	Cys	Asn	Dmf	Msn	Dph	4Py	Asp	Pen	Arg	89	80.69	778.965	3	2333.8499	9.9
NCBP-3	Coldspot	Lys	Nal	Pip	Phe	C5g	Phe	Cys	bSe	Phe	bSe	Tyr	Lys	Hyp	Pen	Lys	96	52.23	501.7623	4	2003.0195	0.3
NCBP-4	Coldspot	Lys	Nal	Pip	Phe	C5g	Phe	Cys	Hyp	Phe	Tyr	Tyr	Lys	Hyp	Pen	Lys	97	56.34	693.3494	3	2077.0352	-4.3
NCBP-5	Coldspot	Lys	Tle	bSe	Phe	Nal	Phe	Cys	Hyp	Phe	Tha	Tyr	Lys	Tyr	Pen	Lys	96	70.19	694.6673	3	2080.9758	2
NCBP-6	Coldspot	Lys	Nal	bSe	Phe	Phe	Phe	Cys	Tyr	Phe	Tha	Tyr	Lys	Hyp	Pen	Lys	92	73.5	705.9942	3	2114.9604	0.2
NCBP-7	Coldspot	Lys	Nal	bSe	Phe	Nal	Phe	Cys	Tha	Phe	Hyp	Tyr	Lys	Hyp	Pen	Lys	93	75.16	705.9943	3	2114.9602	0.4
NCBP-8	Coldspot	Lys	Phe	Pip	Phe	Nal	Phe	Cys	Tha	Phe	Tyr	Tyr	Lys	Hyp	Pen	Lys	83	62.65	536.0076	4	2139.9919	4.3
NCBP-9	Coldspot	Lys	Nal	Pip	Phe	Phe	Phe	Cys	Tha	Phe	Tyr	Tyr	Lys	Hyp	Pen	Lys	85	62.95	536.0094	4	2139.9919	7.7
NCBP-10	Coldspot	Lys	Pff	Pip	Phe	Phe	Phe	Cys	Tyr	Phe	Hyp	Tyr	Lys	Pip	Pen	Lys	85	50.47	539.0023	4	2151.9888	-4.1
NCBP-11	Coldspot	Lys	Phe	Tyr	Phe	Pff	Phe	Cys	Pip	Phe	Hyp	Tyr	Lys	Pip	Pen	Lys	87	48.75	539.0036	4	2151.9888	-1.6
NCBP-12	Coldspot	Lys	Nal	bSe	Phe	Pff	Phe	Cys	Hyp	Phe	Tyr	Tyr	Lys	Tyr	Pen	Lys	96	74.33	738.9948	3	2213.9565	2.7

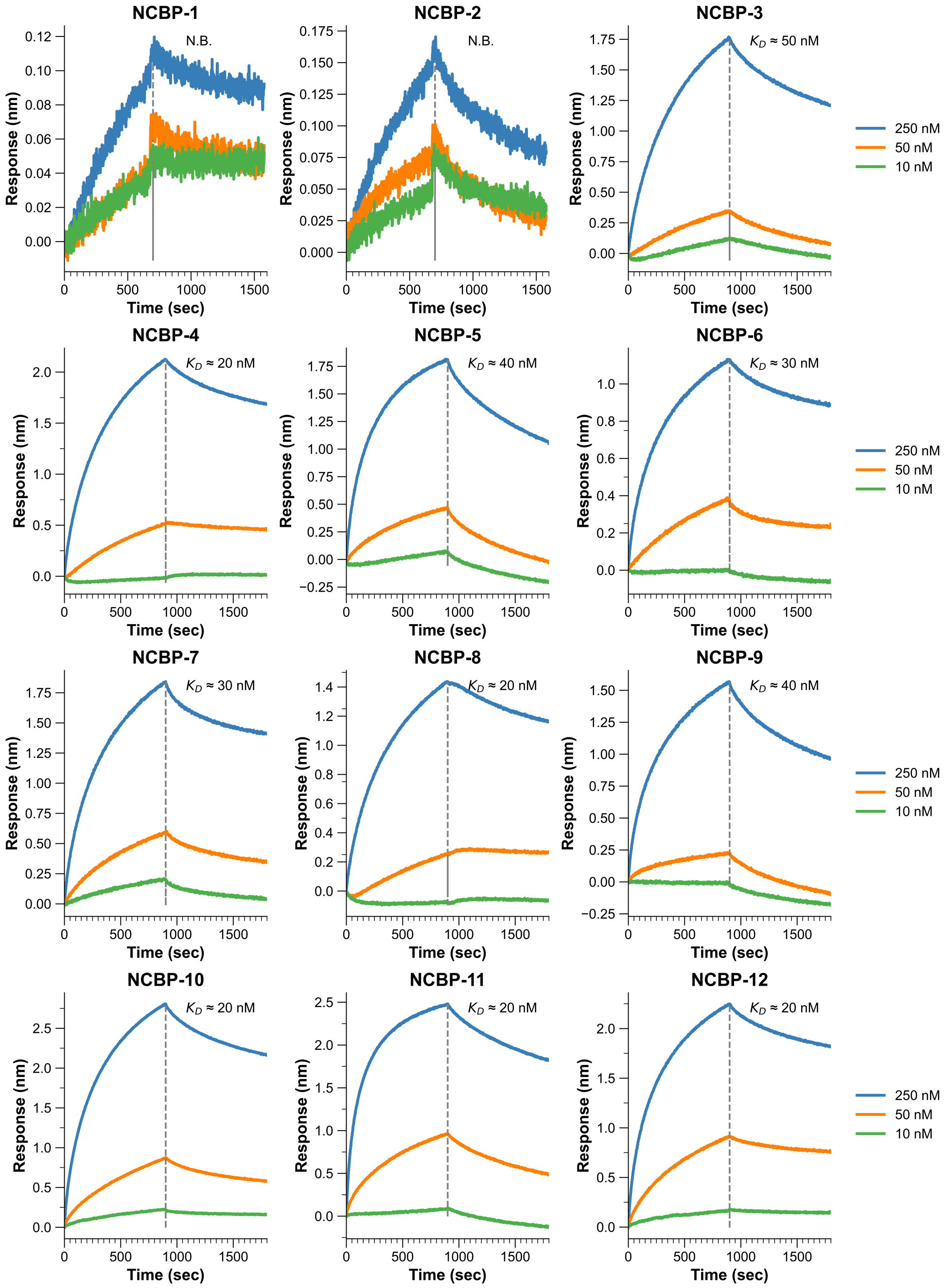


Figure 2.24: *BLI data of **NCBP** peptides against $^{498\text{-}724}$CDH2. Traces with minimal response were deemed non-binders.*

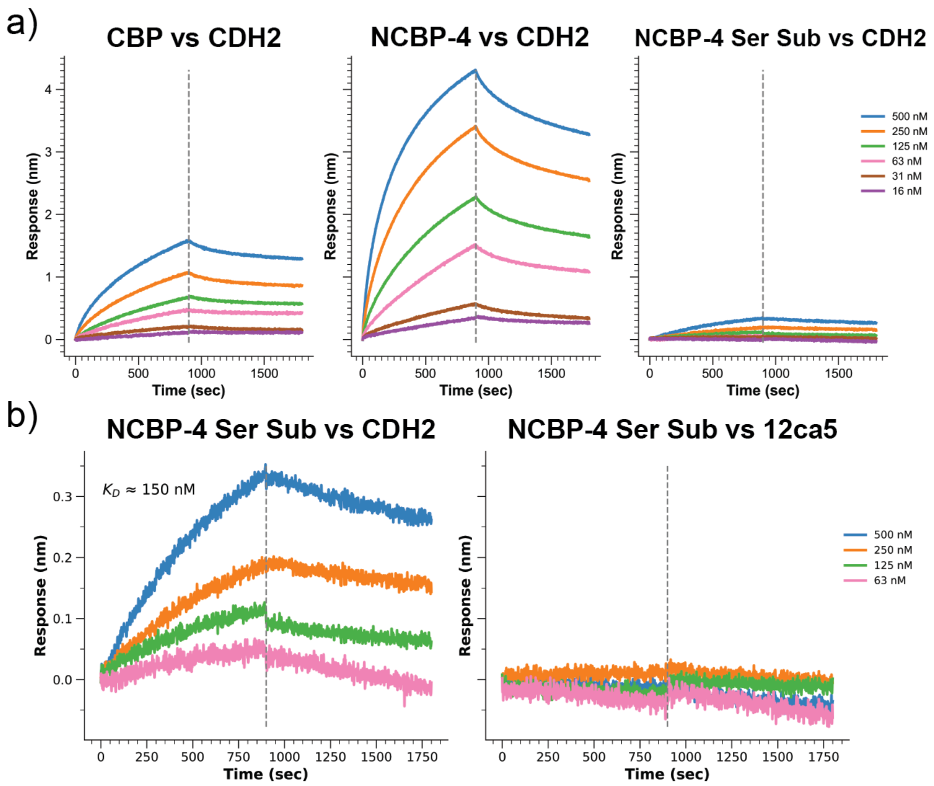


***Figure 2.25:** BLI data of the affinity-matured **NCBP-4** compared to the original **CBP** sequence, as well as the **NCBP-4** sequence with all identified hotspots mutated to serines (NCBP-4 Ser Sub). a) Substituting identified hotspots with serine in **CBP** still allows for modest affinity towards CDH2. b) Mutated residues from the NCBP-4 sequence do not contribute to nonspecific binding of off-target proteins, as demonstrated by BLI assays against 12ca5.*

2.9. Acknowledgements

The authors thank Anthony J. Quartararo, Arisa Shimada, Thomas M. Wood, Sarah A. Antilla, and Amanda E. Cowfer for advice and inspiration 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.

2.10. References

1. Cunningham, A. D., Qvit, N. & Mochly-Rosen, D. Peptides and peptidomimetics as regulators of protein–protein interactions. *Curr Opin Struct Biol* **44**, 59–66 (2017).
2. Wells, J. A. & McClendon, C. L. Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. *Nature* **450**, 1001–1009 (2007).
3. Henninot, A., Collins, J. C. & Nuss, J. M. The Current State of Peptide Drug Discovery: Back to the Future? *J Med Chem* **61**, 1382–1414 (2018).
4. Muttenthaler, M., King, G. F., Adams, D. J. & Alewood, P. F. Trends in peptide drug discovery. *Nat Rev Drug Discov* **20**, 309–325 (2021).
5. Góngora-Benítez, M., Tulla-Puche, J. & Albericio, F. Multifaceted roles of disulfide bonds. peptides as therapeutics. *Chem Rev* **114**, 901–926 (2014).
6. Vinogradov, A. A., Yin, Y. & Suga, H. Macrocyclic Peptides as Drug Candidates: Recent Progress and Remaining Challenges. *J Am Chem Soc* **141**, 4167–4181 (2019).
7. Tyndall, J. D. A., Nall, T. & Fairlie, D. P. Proteases universally recognize beta strands in their active sites. *Chem Rev* **105**, 973–999 (2005).
8. Fairlie, D. P. *et al.* Conformational selection of inhibitors and substrates by proteolytic enzymes: Implications for drug design and polypeptide processing. *J Med Chem* **43**, 1271–1281 (2000).
9. Wang, D. *et al.* Enhanced Metabolic Stability and Protein-Binding Properties of Artificial α Helices Derived from a Hydrogen-Bond Surrogate: Application to Bcl-xL. *Angewandte Chemie International Edition* **44**, 6525–6529 (2005).
10. Spokoyny, A. M. *et al.* A perfluoroaryl-cysteine SNAr chemistry approach to unprotected peptide stapling. *J Am Chem Soc* **135**, 5946–5949 (2013).
11. Tugyi, R., Mezö, G., Fellinger, E., Andreu, D. & Hudecz, F. The effect of cyclization on the enzymatic degradation of herpes simplex virus glycoprotein D derived epitope peptide. *Journal of Peptide Science* **11**, 642–649 (2005).
12. Walensky, L. D. *et al.* Activation of apoptosis in vivo by a hydrocarbon-stapled BH3 helix. *Science (1979)* **305**, 1466–1470 (2004).
13. Naylor, M. R., Bockus, A. T., Blanco, M. J. & Lokey, R. S. Cyclic peptide natural products chart the frontier of oral bioavailability in the pursuit of undruggable targets. *Curr Opin Chem Biol* **38**, 141–147 (2017).
14. Pye, C. R. *et al.* Nonclassical Size Dependence of Permeation Defines Bounds for Passive Adsorption of Large Drug Molecules. *J Med Chem* **60**, 1665–1672 (2017).
15. Bhardwaj, G. *et al.* Accurate de novo design of membrane-traversing macrocycles. *Cell* **185**, 3520-3532.e26 (2022).
16. Mizuno-Kaneko, M. *et al.* Molecular Design of Cyclic Peptides with Cell Membrane Permeability and Development of MDMX-p53 Inhibitor. *ACS Med Chem Lett* (2023) doi:10.1021/ACSMEDCHEMLETT.3C00102.
17. Iskandar, S. E. & Bowers, A. A. mRNA Display Reaches for the Clinic with New PCSK9 Inhibitor. *ACS Med Chem Lett* (2022) doi:10.1021/ACSMEDCHEMLETT.2C00319.

18. Bourne, G. *et al.* Oral peptide inhibitors of interleukin-23 receptor and their use to treat inflammatory bowel diseases. vol. 2 (2017).
19. Benhamou, R. I. *et al.* Macrocyclization of a Ligand Targeting a Toxic RNA Dramatically Improves Potency. *ChemBioChem* **21**, 3229–3233 (2020).
20. Hacker, D. E. *et al.* Direct, Competitive Comparison of Linear, Monocyclic, and Bicyclic Libraries Using mRNA Display. *ACS Comb Sci* **22**, 306–310 (2020).
21. Gao, Y. & Kodadek, T. Direct comparison of linear and macrocyclic compound libraries as a source of protein ligands. *ACS Comb Sci* **17**, 190–195 (2015).
22. Passioura, T., Katoh, T., Goto, Y. & Suga, H. Selection-Based Discovery of Druglike Macrocyclic Peptides. *https://doi.org/10.1146/annurev-biochem-060713-035456* **83**, 727–752 (2014).
23. Tavassoli, A. SICLOPPS cyclic peptide libraries in drug discovery. *Curr Opin Chem Biol* **38**, 30–35 (2017).
24. Goto, Y. & Suga, H. The RaPID Platform for the Discovery of Pseudo-Natural Macrocyclic Peptides. *Acc Chem Res* **54**, 3604–3617 (2021).
25. Deyle, K., Kong, X. D. & Heinis, C. Phage Selection of Cyclic Peptides for Application in Research and Drug Development. *Acc Chem Res* **50**, 1866–1874 (2017).
26. Bacon, K. *et al.* Isolation of Chemically Cyclized Peptide Binders Using Yeast Surface Display. *ACS Comb Sci* **22**, 519–532 (2020).
27. Wong, J. Y. K. *et al.* Genetically-encoded discovery of proteolytically stable bicyclic inhibitors for morphogen NODAL. *Chem Sci* **12**, 9694–9703 (2021).
28. Owens, A. E., Iannuzzelli, J. A., Gu, Y. & Fasan, R. MOrPH-PhD: An Integrated Phage Display Platform for the Discovery of Functional Genetically Encoded Peptide Macrocycles. *ACS Cent Sci* **6**, 368–381 (2020).
29. Biron, É., Vézina-Dawod, S. & Bédard, F. Synthetic Strategies for Macrocyclic Peptides. *Practical Medicinal Chemistry with Macrocycles* 205–241 (2017) doi:10.1002/9781119092599.CH9.
30. Habeshian, S. *et al.* Synthesis and direct assay of large macrocycle diversities by combinatorial late-stage modification at picomole scale. *Nature Communications 2022 13:1* **13**, 1–14 (2022).
31. 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).
32. Zhang, G. *et al.* Rapid de novo discovery of peptidomimetic affinity reagents for human angiotensin converting enzyme 2. *Commun Chem* **5**, 1–10 (2022).
33. Quartararo, A. J. *et al.* Ultra-large chemical libraries for the discovery of high-affinity peptide binders. *Nat Commun* **11**, 3183 (2020).
34. White, C. J. & Yudin, A. K. Contemporary strategies for peptide macrocyclization. *Nature Chemistry 2011 3:7* **3**, 509–524 (2011).
35. Ekanayake, A. I. *et al.* Genetically Encoded Fragment-Based Discovery from Phage-Displayed Macrocyclic Libraries with Genetically Encoded Unnatural Pharmacophores. *J Am Chem Soc* **143**, 5497–5507 (2021).
36. Wrighton, N. C. *et al.* Small Peptides as Potent Mimetics of the Protein Hormone Erythropoietin. *Science (1979)* **273**, 458–463 (1996).

37. Fairbrother, W. J. *et al.* Novel peptides selected to bind vascular endothelial growth factor target the receptor-binding site. *Biochemistry* **37**, 17754–17764 (1998).
38. DeLano, W. L., Ultsch, M. H., De Vos, A. M. & Wells, J. A. Convergent solutions to binding at a protein-protein interface. *Science (1979)* **287**, 1279–1283 (2000).
39. Zhang, H. & Chen, S. Cyclic peptide drugs approved in the last two decades (2001–2021). *RSC Chem Biol* **3**, 18–31 (2022).
40. Mortensen, K. T., Osberger, T. J., King, T. A., Sore, H. F. & Spring, D. R. Strategies for the Diversity-Oriented Synthesis of Macrocycles. *Chem Rev* **119**, 10288–10317 (2019).
41. Sayago, C. *et al.* Deciphering Binding Interactions of IL-23R with HDX-MS: Mapping Protein and Macrocyclic Dodecapeptide Ligands. *ACS Med Chem Lett* **9**, 912–916 (2018).
42. Guerlavais, V. *et al.* Discovery of Sulanemadlin (ALRN-6924), the First Cell-Permeating, Stabilized α-Helical Peptide in Clinical Development. *J Med Chem* **66**, 9401–9417 (2023).
43. 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).
44. Lim, S. *et al.* Discovery of cell active macrocyclic peptides with on-target inhibition of KRAS signaling. *Chem Sci* **12**, 15975–15987 (2021).
45. Fair, R. J., Walsh, R. T. & Hupp, C. D. The expanding reaction toolkit for DNA-encoded libraries. *Bioorg Med Chem Lett* **51**, 128339 (2021).
46. Behsaz, B. *et al.* De Novo Peptide Sequencing Reveals Many Cyclopeptides in the Human Gut and Other Environments. *Cell Syst* **10**, 99-108.e5 (2020).
47. Townsend, C. *et al.* CycLS: Accurate, whole-library sequencing of cyclic peptides using tandem mass spectrometry. *Bioorg Med Chem* **26**, 1232–1238 (2018).
48. Kavan, D., Kuzma, M., Lemr, K., Schug, K. A. & Havlicek, V. CYCLONE - A utility for de novo sequencing of microbial cyclic peptides. *J Am Soc Mass Spectrom* **24**, 1177–1184 (2013).
49. Mohimani, H. *et al.* Cycloquest: Identification of cyclopeptides via database search of their mass spectra against genome databases. *J Proteome Res* **10**, 4505–4512 (2011).
50. Lee, J. H., Meyer, A. M. & Lim, H. S. A simple strategy for the construction of combinatorial cyclic peptoid libraries. *Chemical Communications* **46**, 8615–8617 (2010).
51. Liang, X., Vézina-Dawod, S., Bédard, F., Porte, K. & Biron, E. One-Pot Photochemical Ring-Opening/Cleavage Approach for the Synthesis and Decoding of Cyclic Peptide Libraries. *Org Lett* **18**, 1174–1177 (2016).
52. Borges, A. *et al.* Facile de Novo Sequencing of Tetrazine-Cyclized Peptides through UV-Induced Ring-Opening and Cleavage from the Solid Phase. *ChemBioChem* **24**, e202200590 (2023).

53. Simpson, L. S. & Kodadek, T. A cleavable scaffold strategy for the synthesis of one-bead one-compound cyclic peptoid libraries that can be sequenced by tandem mass spectrometry. *Tetrahedron Lett* **53**, 2341–2344 (2012).
54. Elashal, H. E., Cohen, R. D., Elashal, H. E. & Raj, M. Oxazolidinone-Mediated Sequence Determination of One-Bead One-Compound Cyclic Peptide Libraries. *Org Lett* **20**, 2374–2377 (2018).
55. Menegatti, S. *et al.* Reversible cyclic peptide libraries for the discovery of affinity ligands. *Anal Chem* **85**, 9229–9237 (2013).
56. Novák, J., Lemr, K., Schug, K. A. & Havlíček, V. CycloBranch: De Novo Sequencing of Nonribosomal Peptides from Accurate Product Ion Mass Spectra. *J Am Soc Mass Spectrom* **26**, 1780–1786 (2015).
57. Tran, N. H. *et al.* Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. *Nature Methods 2018 16:1* **16**, 63–66 (2018).
58. Rini, J. M., Schulze-Gahmen, U. & Wilson, I. A. Structural evidence for induced fit as a mechanism for antibody-antigen recognition. *Science (1979)* **255**, 959–965 (1992).
59. Houghten, R. A. *et al.* Generation and use of synthetic peptide combinatorial libraries for basic research and drug discovery. *Nature 1991 354:6348* **354**, 84–86 (1991).
60. Brown, J. S. *et al.* Unsupervised machine learning leads to an abiotic picomolar peptide ligand. *ChemRxiv* (2023) doi:10.26434/CHEMRXIV-2023-TWS4N.
61. Takeichi, M. The cadherin superfamily in neuronal connections and interactions. *Nature Reviews Neuroscience 2006 8:1* **8**, 11–20 (2006).
62. Kostetskii, I. *et al.* Induced Deletion of the N-Cadherin Gene in the Heart Leads to Dissolution of the Intercalated Disc Structure. *Circ Res* **96**, 346–354 (2005).
63. Karlsson, M. *et al.* A single–cell type transcriptomics map of human tissues. *Sci Adv* **7**, (2021).
64. Williams, E. J., Williams, G., Gour, B., Blaschuk, O. & Doherty, P. INP, a novel N-cadherin antagonist targeted to the amino acids that flank the HAV motif. *Mol Cell Neurosci* **15**, 456–464 (2000).
65. Burden-Gulley, S. M. *et al.* Novel peptide mimetic small molecules of the HAV motif in N-cadherin inhibit N-cadherin-mediated neurite outgrowth and cell adhesion. *Peptides (N.Y.)* **30**, 2380–2387 (2009).
66. Tam, J. P., Wu, C. R., Liu, W. & Zhang, J. W. Disulfide Bond Formation in Peptides by Dimethyl Sulfoxide. Scope and Applications. *J Am Chem Soc* **113**, 6657–6662 (1991).
67. Hennion, M. C. Solid-phase extraction: method development, sorbents, and coupling with liquid chromatography. *J Chromatogr A* **856**, 3–54 (1999).
68. Yang, Y., Hansen, L., Fraczek, A., Badalassi, F. & Kjellström, J. DMF-Assisted Iodination Side Reaction during the Preparation of Disulfide Peptides, Its

Substrate/Solvent/pH Dependence, and Implications on Disulfide-Peptide Production. *Org Process Res Dev* **25**, 2090–2099 (2021).

69. Yang, Y., Hansen, L., Fraczek, A., Badalassi, F. & Kjellström, J. DMF-Assisted Iodination Side Reaction during the Preparation of Disulfide Peptides, Its Substrate/Solvent/pH Dependence, and Implications on Disulfide-Peptide Production. *Org Process Res Dev* **25**, 2090–2099 (2021).
70. Thermo Scientific (Pierce Biotechnology). *Instructions Immobilized TCEP Disulfide Reducing Gel (Cat 77712) Version 1325.3 MAN0011439*. (2023).
71. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nature Protocols 2007 2:8* **2**, 1896–1906 (2007).
72. Vinogradov, A. A. *et al.* Library Design-Facilitated High-Throughput Sequencing of Synthetic Peptide Libraries. *ACS Comb Sci* **19**, 694–701 (2017).
73. Simon, M. D. *et al.* D-Amino Acid Scan of Two Small Proteins. *J Am Chem Soc* **138**, 12099–12111 (2016).
74. Peeters, T. L. *et al.* d-Amino acid and alanine scans of the bioactive portion of porcine motilin. *Peptides (N.Y.)* **13**, 1103–1107 (1992).
75. Touti, F., Gates, Z. P., Bandyopdhyay, A. & Lautrette, G. In-solution enrichment identifies peptide inhibitors of protein – protein interactions protein-protein interactions. *Nat Chem Biol* 318–339 (2019) doi:10.1038/s41589-019-0245-2.
76. Ye, X. *et al.* Binary combinatorial scanning reveals potent poly-alanine-substituted inhibitors of protein-protein interactions. *Communications Chemistry 2022 5:1* **5**, 1–10 (2022).
77. Frutiger, A. *et al.* Nonspecific Binding - Fundamental Concepts and Consequences for Biosensing Applications. *Chem Rev* **121**, 8095–8160 (2021).
78. Ng, S. *et al.* De-risking Drug Discovery of Intracellular Targeting Peptides: Screening Strategies to Eliminate False-Positive Hits. *ACS Med Chem Lett* acsmedchemlett.0c00022 (2020) doi:10.1021/acsmedchemlett.0c00022.
79. Dias, R. L. A. *et al.* Protein ligand design: From phage display to synthetic protein epitope mimetics in human antibody Fc-binding peptidomimetics. *J Am Chem Soc* **128**, 2726–2732 (2006).
80. Kourra, C. M. B. K. & Cramer, N. Converting disulfide bridges in native peptides to stable methylene thioacetals. *Chem Sci* **7**, 7007–7012 (2016).
81. Hartrampf, N. *et al.* Synthesis of Proteins by Automated Flow Chemistry. *Science (1979)* **987**, 1–20 (2020).
82. Mijalis, A. J. *et al.* A fully automated flow-based approach for accelerated peptide synthesis. *Nat Chem Biol* **13**, 464–466 (2017).

3. Investigation of Commercially Available Resins for the Automated Flow Synthesis of Difficult or Long Peptide Sequences

The work presented in this chapter has been reproduced and adapted from the following publication:

Lee, M.A.; Brown, J.S.; Loas, A.; Pentelute, B.L. Investigation of Commercially Available Resins for the Automated Flow Synthesis of Difficult or Long Peptide Sequences*. Peptide Science.* **2024** e24344

3.1. Introduction

Solid phase peptide synthesis (SPPS) has been established as the mainstay method for the production of peptides and short proteins for a range of applications from basic research to the manufacture of clinically-approved peptide therapies.[1,2,3] This method utilizes a swellable crosslinked support functionalized with a chemical linker that allows for an iterative cycle of coupling amino acids and removal of protecting groups.[4] Since its introduction by Nobel Laureate Bruce Merrifield in 1963,[5] SPPS has seen rapid integration into standard research protocols and expansion of its capabilities to applications such as the synthesis of over 200 residue-long protein domains[6,7] or the split-and-pool synthesis of large combinatorial libraries.[8] A key feature for SPPS is the choice of resin solid support.[9]

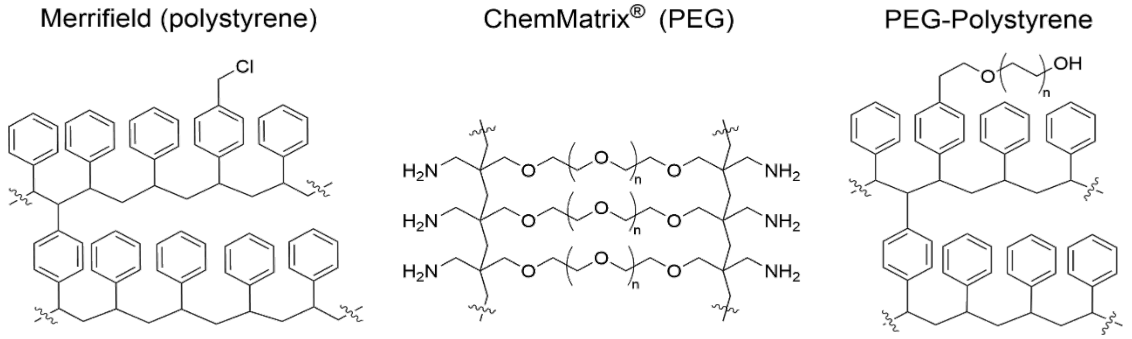


Figure 3.1: *The chemical structure of the resin solid support determines important physical properties for efficient peptide synthesis. Three common types of resin are the Merrifield (polystyrene) resins, ChemMatrix® PEG based resins, and PEG-PS resins.*

There are a variety of commercially available and in-house manufactured options for the solid support, ranging from the original polystyrene crosslinked (PS) resin to purely polyethyleneglycol- based (PEG) resins, with co-polymers of the two styles also being available. Each of these supports differ in the ability to swell and solvate the growing peptide chain.[10,11] Notably, there are several formulations comprised of crosslinked polystyrene supports with grafted polyethyleneglycol chains, termed PEG-PS, where the proportions of monomers and methods of grafting can generate a range of properties for SPPS.[12,13] These different backbone structures are shown in Figure 3.1.

The solid support and the elongating peptide chain are an important factor for SPPS.[14] The synthesis of long peptide sequences can be limited by steric hindrance of the peptide chain through the formation of secondary structures on resin or through aggregation of the peptide chain by unfavorable interactions with the polymer support, causing early terminations of growing peptide chains. These physical constraints impede reaction progress and can introduce additional mass transfer limitations into the coupling steps, reducing purity and yield.[1,11,15] Several approaches have improved SPPS in the attempts to overcome on-resin aggregation, ranging from chemical optimizations (i.e., backbone protection)[16,17] to physical parameters optimization (i.e., heat or microwave).[18,19] The chemical structure of the resin has also been investigated. For PS resins, aggregation has been hypothesized to be due to unwanted interactions with the polystyrene support from the side chains of the peptide, especially for longer peptide sequences as reported in literature.[14,20,21] PEG resins, on the other hand, can provide stabilizing polar interactions with the growing chain and can potentially stabilize its structure, allowing for the synthesis of longer sequences as reported in literature. Therefore, most advanced, commercially available resins incorporate a significant amount of PEG into the solid support.

Another critical property of a solid support for SPPS is the degree of swelling in the selected solvent, which reflects the rate of mass transport through the matrix.[22,23] The functionalized reaction sites of the resin are primarily spread throughout the bead, meaning that rapid transport of reagents into the matrix is critical to attain satisfactory yield and reaction rate.[24,25] Reagent transport rate has been historically challenging to engineer and has been primarily achieved through carefully tuning the cross-linking of the polymer support.[10] However, this mobility can also lead to unwanted side reactions and site-site interactions, where nearby peptide chains can interact with each other and potentially block the availability of the N terminus to chain extension through formation of secondary structure. A balance is necessary to optimize the swelling properties as seen by reaction rate studies performed on a range of cross-linking compositions by Rana et al.[11] These same principles apply to the resin loading, where higher loading will decrease the distance between peptide chains; therefore, resin loading is also critical to performance. Across a range of solvents, PEG resins provide excellent swelling properties and have been the solid support of choice for the synthesis of difficult, side-chain to backbone aggregation-prone sequences,[26] but they may not be suitable for all peptide syntheses. Additionally, PEG resins can be difficult to source commercially, especially in larger quantities.[27] PS resins show significantly lower amounts of swelling, which may introduce coupling limitations, such as a reduction in reaction rate.[26] Thus, hybrid PEG-PS resins appear as strong candidates for the general synthesis of difficult or long (>50 residues in length) peptide sequences.

Whereas many studies describe the optimization of resins for SPPS, there have been few side-by- side comparisons of the available resins that have been released to the market in recent years. In this work, we chose three commercially available PEG-PS resins recently introduced and compared their performance to previously published data using ChemMatrix® resin, a common choice of resin for the synthesis of complex peptides.[12] The three PEG-PS resins used for these studies are: OctaGel™ resin, Tentagel XV resin®, and ProTide® resin. These resins were chosen as representatives of the PEG-PS candidate design, their excellent swelling properties in common solvents including dichloromethane (DCM) and *N,N*-dimethylformamide (DMF), and for their reported efficiencies compared to ChemMatrix resins.[7,28] Each of these resins were used with an automated fast-flow peptide synthesizer (AFPS) for production of the following sequences: JR10, a 10-mer peptide known to aggregate during SPPS;[29,30] GLP-1, a 30-mer sequence previously characterized under automated flow conditions;[7] a fragment of the mouse double minute 2 (MDM2) N-terminal domain, an 84-mer sequence; and the tetranucleotide repeat domain of the E3 ubiquitin-protein ligase CHIP, a 133-mer sequence.[28]

3.2. Results and Discussion

The first peptide we synthesized was JR10, a ten amino acid peptide (WFTTLISTIM) that has been previously used to characterize synthesis conditions,[14,30,35] where the ProTide resin performed the best as evaluated by crude yield and purity. This sequence is known to aggregate during the coupling of Thr4, where early peptide chain terminations during synthesis were seen using an in-line UV detector. The observed absorbance at 310 nm quantitatively detects the fluorenylmethyloxycarbonyl (Fmoc) deprotection and its resulting byproduct dibenzofulvene.[14] Looking at the characteristics of the deprotection peaks, clear peak width broadening is observed during synthesis on all three candidate resins (see Figure 3.2).[15,36] The broadening of the deprotection peak width is indicative of side-chain to backbone aggregation, as the interactions of the peptide chain with other peptide chains or the solid support interferes with mass transport through the matrix.[30] This occurs as the peptide chain grows and interacts with the polymer support by impeding fluid flow through the resin, causing increased rates of axial mixing that will subsequently increase the residence time distribution as measured by the broadening of the UV absorbance peak of the Fmoc protecting group.[37] This could be approximated by increases of diameter around a reactive site as the peptide chains grow and aggregate, which would correlate to a decrease in the Péclet number via the Gunn correlation of axial and radial dispersion over fixed beds.[38,39] The increased mass of peptide chain on the surface of the bead interferes with flow of reagents across the resin bed. Raw data of the UV signal for amino acid coupling and Fmoc group deprotection is given in Figure 3.6. Decreased or hindered mass transport within the resin in turn leads to the reactive sites of the growing peptide chain to be less accessible to the reagents, increasing the potential for single residue deletions and/or early termination of peptide chain elongation as seen by the drop in peak area.

We cleaved JR10 from each of the resins and characterized the crude material by HPLC and LC-MS. For each sample, the full 10-mer sequence was obtained, but several truncated products were observed by HPLC and LC-MS and characterized in Figures 3.11-3.13. Octagel resin yielded a mixture of truncation products started after Thr4, with three major peptides corresponding to the W, WF, and WFT deletions identified through LC-MS. ProTide and Tentagel XV resins also yielded significant truncation products, corresponding to W and WF deletions at the N terminus. Based on this data, the purity of the JR10 peptide synthesized across the resins was comparable, with a slight advantage to the ProTide and Tentagel XV resins.

Crude yields across the resins show ProTide to be the best performing candidate for synthesis of JR10. Octagel, in addition to having the lowest purity of crude product, also gave the lowest yield off resin, with TentaGel XV being slightly lower than ProTide (see Table 3.1). This trend also correlates with the loading of each resin, where Octagel has the highest loading with 0.441 mmol/g, Tentagel XV the next highest with 0.27 mmol/g, and ProTide with the lowest at 0.20 mmol/g; all resins have bead sizes of 75-150 microns (200-100 mesh). Since side-chain to backbone aggregation is the key factor for the efficiency of JR10 synthesis, a lower loading on each bead of resin could allow for an increased spacing between peptide chains, potentially lowering the opportunity for aggregation.[40]

To investigate this hypothesis, we prepared two batches of Tentagel XV resin with 50% and 10% of the normal loading as well as three batches of Octagel resin at 63%, 31%, and 6% loading to roughly match the loading of the Tentagel XV resin. Reduction in resin loading was done by mixing acetic acid to cap the resin and Fmoc-methionine-OH (the first residue of JR10) in the respective molar amounts and coupling manually to the resin. After manual coupling, the resins were moved to the automated synthesizer to finish synthesis. HPLC analysis (see Figure 3.14) of the crude products from the Tentagel XV resin showed an improvement of crude purity, increasing to 61% from the original 32% from the full loading amount. Meanwhile, the Octagel resin was also able to achieve a similar purity of 60% for a similar loading amount to Tentagel XV (see Figure 3.15). Lower loading resins are often available commercially by request to the respective vendor, as is the case for all three resins.

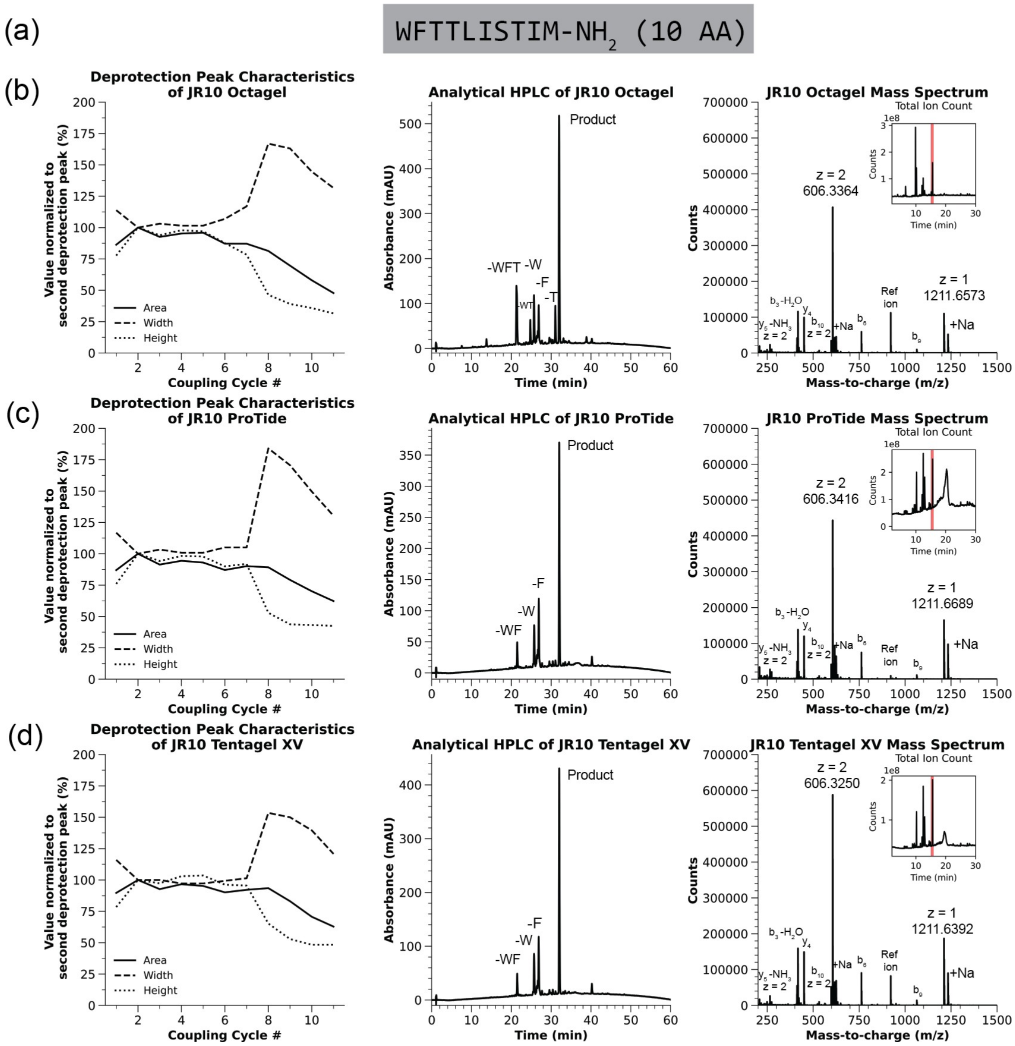


Figure 3.2: *Inline UV detection, analytical HPLC, and LC-MS of crude JR10 indicated no major differences between the candidate resins in automated flow synthesis. UV absorbance was gathered at 310 nm to quantity the Fmoc deprotection byproduct of dibenzofulvene in flow. Analytical HPLC and LC-MS spectra of the crude JR10 peptide (Calculated exact mass: 1210.6420) from each resin enabled performance evaluation. Known significant aggregation at Thr4 resulted in several truncation side products; asymmetries in side product peaks were found to be other side products and JR10 coeluting. (a) Sequence of JR10; (b) Data of JR10 synthesized using Octagel Resin. Resin loading: 0.441 mmol/g. Crude purity: 24%. Observed mass: 1210.654 (9.8 ppm error); (c) Data of JR10 synthesized using ProTide resin. Resin loading: 0.20 mmol/g. Crude purity: 32%. Observed mass: 1210.665 (19 ppm error); (d) Data of JR10 synthesized using Tentagel XV resin. Resin loading: 0.27 mmol/g. Crude purity: 32%. Observed mass: 1210.634 (-7.0 ppm error)*

The next sequence we synthesized was GLP-1, a 30-mer peptide hormone, where ProTide and Tentagel XV performed similarly well by crude purity while product from Octagel resin showed poor purity. This peptide was chosen as a routine length sequence characterized in a wealth of previous synthesis data.[7] Like JR10, GLP-1 also has a known aggregation point at Ala18.[14] This side-chain to backbone aggregation is clearly seen during synthesis using Octagel, where a sudden broadening of the deprotection peaks is observed at the coupling corresponding to Ala18 (see Figure 3.3). However, this side-chain to backbone aggregation is not as apparent for ProTide and Tentagel XV resins, where only a slight peak broadening is observed. Full UV data for the synthesis of GLP-1 is given in Figure 3.7 and characterization of side products is given in Figures 3.16-3.18.

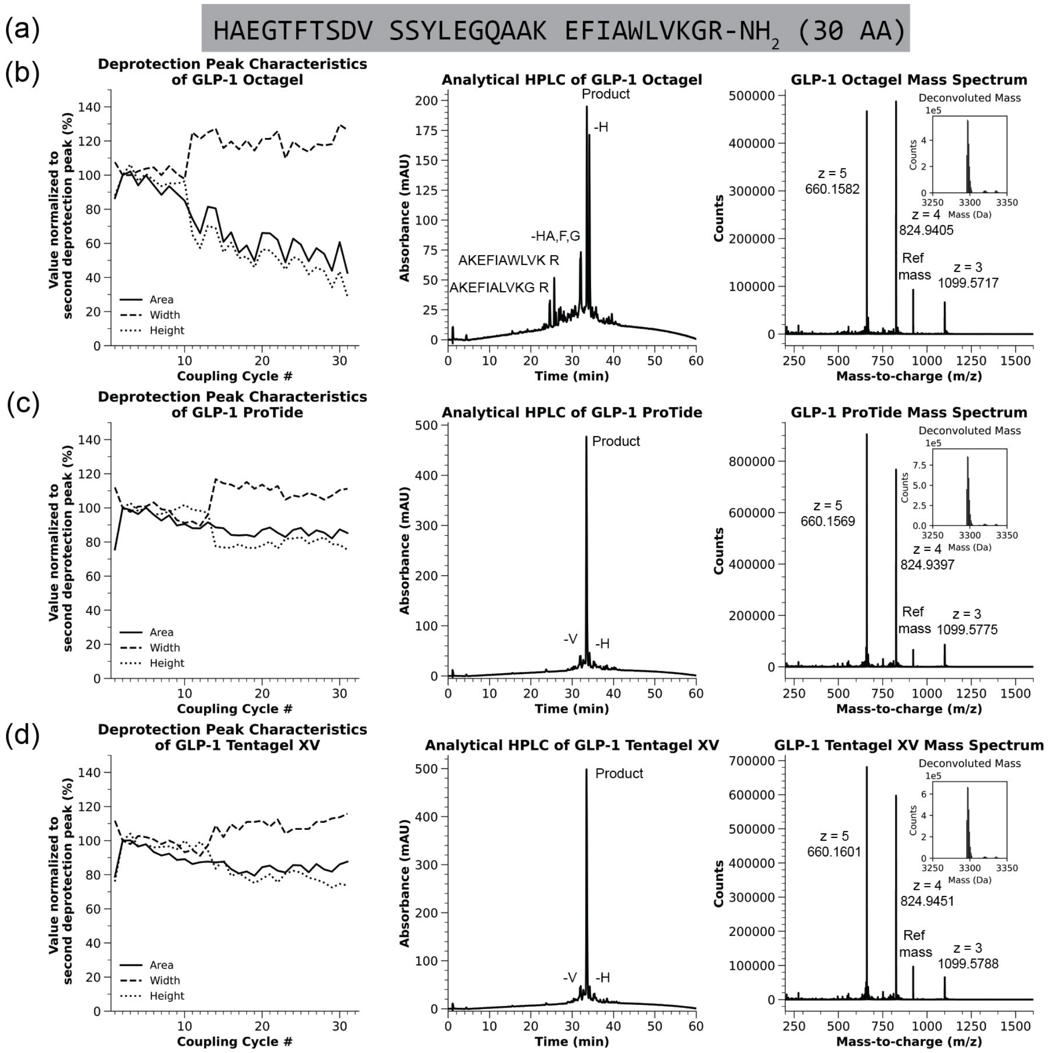


Figure 3.3: *Inline UV detection, analytical HPLC and LC-MS data for GLP-1 synthesis show excellent performance by ProTide and Tentagel XV resins, but suboptimal performance by Octagel resin. (a) Sequence of GLP-1; (b) Data of GLP-1 (Calculated deconvoluted mass: 3295.66) synthesized using Octagel resin. Resin loading: 0.441 mmol/g. Crude purity: 24%. Observed deconvoluted mass: 3295.72 (17 ppm error); (c) Data of GLP-1 synthesized using ProTide resin. Resin loading: 0.20 mmol/g. Crude purity: 54%. Observed deconvoluted mass: 3295.73 (20 ppm error); (d) Data of GLP-1 synthesized using Tentagel XV resin. Resin loading: 0.27 mmol/g. Crude purity: 49% Observed deconvoluted mass: 3295.74 (24 ppm error).*

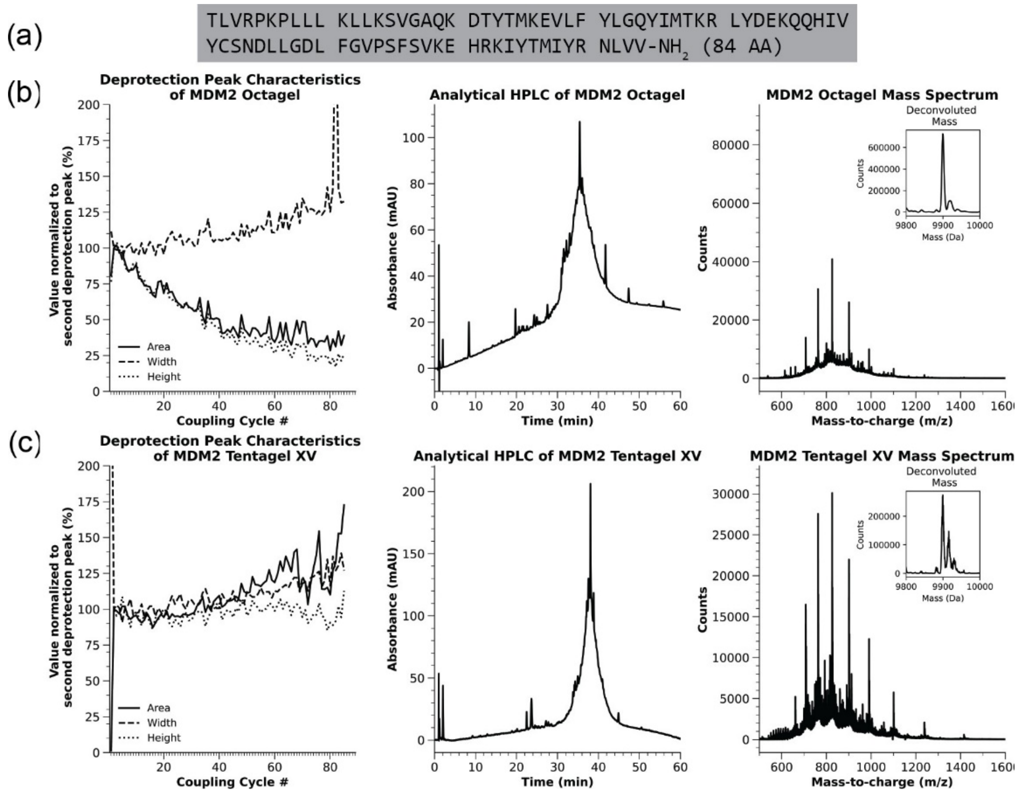
Cleavage of GLP-1 resulted in a clear difference between Octagel and the other two candidate resins. Crude HPLC of the ProTide and Tentagel XV products showed a pure product, but Octagel yielded a mixture of products. ProTide and Tentagel XV resins outperformed Octagel resin for producing higher purity crude polypeptide (54% and 49% versus 24% respectively). The GLP-1 crude isolated mass across all three resins were comparable (see Table 3.1). and Tentagel XV resins provided similarly pure material at comparable yields and purities, thus we observed either resin is suitable for the synthesis of intermediate length sequences (~30 residues).

Table 3.1: *Resin measurements and crude yields for the synthesis of JR10 and GLP-1 on Octagel, ProTide, and Tentagel XV resins. Theoretical mass is calculated based on the reported loading of the resin by the respective vendor.*

Peptide	Resin	Starting Resin (mg)	Crude Resin (mg)	Isolated Mass (mg)	Theoretical Mass (mg)	Cleaved Yield	Crude Purity
JR10	Octagel	43.6	49.4	8.7	23.3	37%	24%
JR10	ProTide	98.4	114.4	18.6	23.8	77%	32%
JR10	Tentagel XV	97.9	114.0	18.6	32.0	58%	32%
GLP-1	Octagel	49.0	92.2	37.8	71.3	43%	24%
GLP-1	ProTide	99.5	143.6	39.3	63.9	61%	54%
GLP-1	Tentagel XV	96.9	123.7	46.2	88.6	52%	49%

We designed experiments to test the ability of these resins to facilitate the chemical synthesis of single-domain proteins, illustrated by an 84 amino acid fragment of the N-terminal domain of MDM2 and the 134 amino acid tetranucleotide repeat domain of CHIP. During synthesis using the ProTide resin of CHIP, however, the instrument experienced an increase in back pressure over the course of synthesis, ultimately causing an early termination of the synthesis. This issue could have potentially been caused by an accumulation of fine particulates from the ProTide resin downstream of the reactor, causing a clog. Thus, the synthesis of MDM2 was not attempted with the ProTide resin out of caution for the condition of the instrument.

Synthesis of the 84 amino acid fragment of the N-terminal domain of MDM2 was evaluated. This protein fragment had been previously produced using the same AFPS technology, which allows for additional comparison of these three resins to the ChemMatrix resin.[7] A biotinylated glutamine residue was manually coupled to the C-terminus of each resin before automated flow synthesis. While ProTide resin was deemed not suitable for the AFPS system, Octagel and Tentagel XV resins were amenable to producing MDM2. Synthesis data of MDM2 using Octagel and Tentagel XV resin are given in Figure 3.4, with raw UV data given in Figure 3.8. Notably, the UV data of Tentagel XV resin shows a strong performance by the consistency of the deprotection peak shape throughout most of the synthesis. The deprotection peaks potentially increase in width near the end of the synthesis due to residual side chain deprotection of nucleophilic amino acids, such as histidine, leading to side chain elongation.[41] However, an initial deprotection peak was not observed, meaning that the initial manual coupling failed for the Tentagel XV resin. This outcome could be due to the lower degree of swelling in these resins compared to ChemMatrix resin, which was the basis for the deprotection protocol. For consistency, MDM2 without the biontinylated glutamine was synthesized on Octagel resin and is reported in Figure 3.4.



(a) TLVRPKPLLL KLLKSVGAQK DTYTMKEVLF YLGQYIMTKR LYDEKQQHIV YCSNDLLGDL FGVPSFSVKE HRKIYTMIYR NLVV-NH_2 (84 AA)

Figure 3.4: *Synthesis of MDM2 evaluated by the inline UV detection, analytical HPLC, and LC-MS data for Octagel and Tentagel XV resins show superior performance by Tentagel XV resin. Synthesis on ProTide resin was not completed as it caused significant backpressure to develop. (a) Sequence of MDM2; (b) Data of crude MDM2 synthesized on Octagel resin. Resin loading: 0.441 mmol/g. Purity post-HPLC purification: 80% Calculated deconvoluted mass: 9899.82; Observed deconvoluted mass: 19899.7 (-12 ppm error). Oxidized product also present. (c) Data of crude MDM2 synthesized on Tentagel XV resin. Resin loading: 0.27 mmol/g. Purity post-HPLC purification: 81%. Calculated observed mass: 9899.82; Observed deconvoluted mass: 9899.5 (-31 ppm error). Oxidized product also present.*

To further investigate this phenomenon, we compared the efficiency of this deprotection protocol using a short test peptide with a variety of initial monomers manually coupled to the resin. Tentagel XV resin was deprotected either using two five-minute reactions as was previously done for Chem- Matrix resin, or by washing with deprotection solution followed by three five-minute reactions. Analytical HPLC and LC-MS analysis following cleavage did not reveal any significant difference in the coupling for alanine, leucine, arginine, proline, or 2-aminoisobutyric acid (see Figure 3.19), suggesting the deletion was likely a C-terminal monomer-specific effect concerning the Tentagel XV resin, potentially due to the PEG based spacer present on the side chain interacting with the resin backbone's PEG network. Each of the resins come with their amine functional group Fmoc protected, and additional time may need to be taken at the beginning of synthesis to remove the Fmoc protecting group. This first Fmoc deprotection was not an issue under automated conditions due to the high temperature of the reactor (85-90 ºC), which expedited the deprotection. Octagel, while there were no issues with the initial deprotection at room temperature, demonstrated a poorer performance during synthesis. The deprotection peaks consistently broadened starting early in the synthesis, signaling continual peptide chain termination. As with JR10, this could potentially be ameliorated by reducing the total loading of the resin, as the loading of Octagel resin is significantly higher than what is generally used for single domain protein synthesis. In order to test this, we synthesized MDM2 again on Octagel resin with a reduced loading to 63% of the original value, which roughly matches the loading of Tentagel XV resin. This resulted in a more stable deprotection peak width (see Figure 3.20), a notable sign for increased synthesis quality.

Both samples of MDM2 were cleaved off the resin and subjected to preparative HPLC purification to compare yields of the pure polypeptide. Cleavage yields for Tentagel XV resin were high (see Table 3.2), outperforming both Octagel in these experiments and ChemMatrix resin reported in previous work.[28] Purification yields were also high, although formation of oxidized product was observed over the characterization process. While Tentagel XV resin performed better, Octagel resin still provided a pure product in acceptable and additionally succeeded in the initial manual coupling of biontinylated glutamine. However, Tentagel XV resin appeared to be the optimal choice for this synthesis, provided extra care can be taken for the initially coupled amino acids at room temperature, especially during initial deprotection and subsequent washing steps.

(a)
ZSPSAQELKE QGNRLFVGRK YPEAAACYGR AITRNPLVAV YYTNRALCYL
KMQQHEQALA DCRRALELDG QSVKAHFFLG QCQLEMESYD EAIANLQRAY
SLAKEQRLNF GDDIPSALRI AKKKRWNSIE ERR-NH$_2$ (133 AA, Z = Biotin-PEG$_4$)

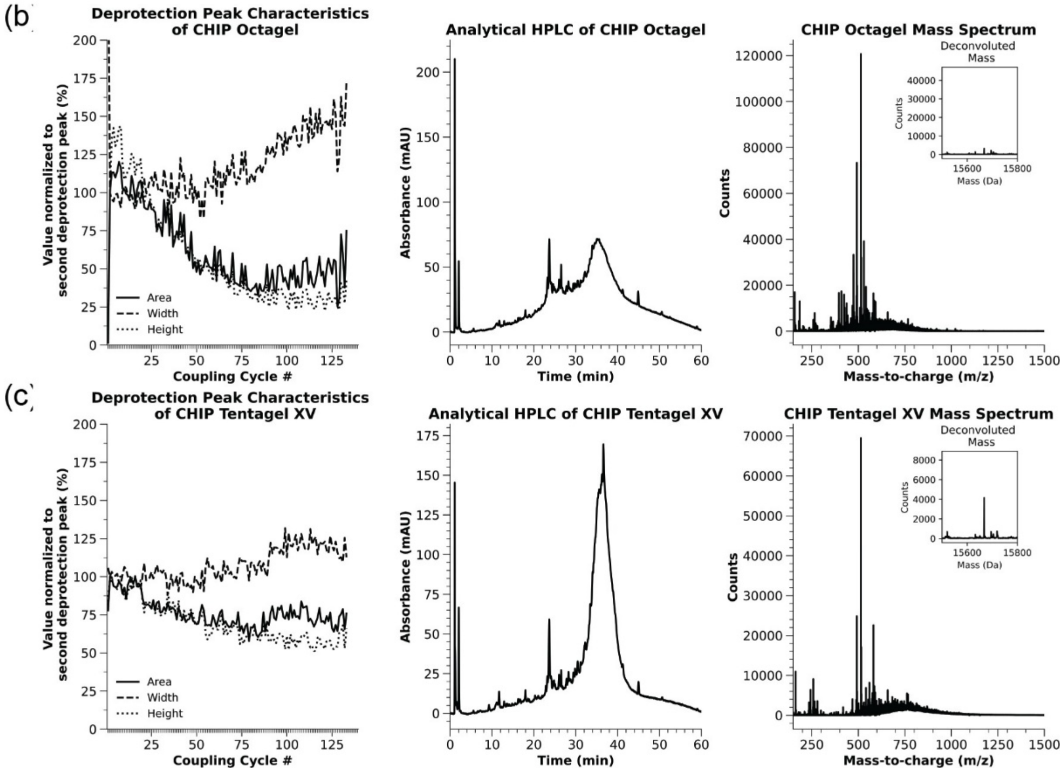


Figure 3.5: *Synthesis of CHIP evaluated by the inline UV detection, analytical HPLC, and LC-MS data for Octagel and Tentagel XV resin show improved performance by Tentagel XV resin. Synthesis on ProTide resin resulted in instrument failure and raw data is shown in Figure 3.10. (a) Sequence of CHIP; (b) Data of crude CHIP (Calculated deconvoluted mass: 15671.9) synthesized on Octagel resin. Resin loading: 0.441 mmol/g. Purity post-HPLC purification: No pure fractions found. Observed deconvoluted mass: Not found. (c) Data of crude CHIP synthesized on Tentagel XV resin. Resin loading: 0.27 mmol/g. Purity post-HPLC purification: 71%. Observed deconvoluted mass: 15672.0 (-3.8 ppm error).*

The longest sequence used to compare the three resins was the 134 amino acid tetranucleotide repeat domain of CHIP. This domain was chosen as a representative long sequence, which also has synthetic data recently made available using ChemMatrix resin.[28] Figure 3.5 shows the synthesis of CHIP using Octagel and Tentagel XV resins. Raw UV data for the synthesis of CHIP is shown in

Figure 3.9. Details on the reactor failure when using ProTide resin in Figure 3.10. The UV data for the remaining two resins recapitulate the observations for MDM2, where Octagel resin saw a steady deprotection peak broadening coupled with a decrease in total peak area. Tentagel XV saw a modest amount of deprotection peak broadening, ending the synthesis with deprotection peaks with a reduced area of about 25% and decreased height by 25%. After synthesis, the proteins were biotinylated using Biotin-PEG$_4$ propionic acid to test modifications of the N-terminus on each resin.

Cleavage of the protein domain from both resins showed Tentagel XV to be the preferred choice for the synthesis of long sequences. Tentagel XV resin resulted in a higher crude yield after cleavage by a margin similar to yields for MDM2 (65% versus 37% respectively, see Table 3.2). Crude LC-MS analysis also showed no significant amount of product for Octagel resin, while the full-length biotinylated sequence was observed in the crude product of Tentagel XV. This result supports the inline UV absorbance measurements of the Fmoc deprotection, which suggested a lower synthesis quality with the Octagel resin relative to the Tentagel XV resin. To confirm these results, both proteins were purified by HPLC, and fractions were analyzed by LC-MS. No fractions were found to contain the mass corresponding to the full-length CHIP domain after Octagel synthesis. However, Tentagel XV resin afforded reasonable isolated purification yield (8%), which was moderately higher than ChemMatrix resin for the synthesis and purification of CHIP (2%).[28]

Table 3.2: *Resin measurements, crude yields and purification yields for the synthesis of MDM2 and CHIP on Octagel, ProTide and Tentagel XV resins. Theoretical mass is calculated based on the reported loading of the resin by the respective vendor. Extrapolated yield refers to the total yield if all crude material was purified with yields similar to what was recovered from the isolated pure mass.*

Protein	Resin	Starting Resin (mg)	Crude Resin (mg)	Isolated Mass (mg)	Theor. Mass (mg)	Cleaved Yield	Purified Mass (mg)	Isolated Pure Mass (mg)	Purified Yield	Extrap. Yield (mg)	Overall Yield
MDM2	Octagel	49.9	200.5	71.3	230.9	31%	35.8	1.1	3%	2.2	1%
MDM2	ProTide	-[a]	-	-	-	-	-	-	-	-	-
MDM2	Tentagel XV	96.9	352.2	178.4	273.4	65%[b]	41.1	8.2	20%	35.6	13%
CHIP	Octagel	54.0	265.2	89.5	366.1	24%	38.0	0.0	0%	0.0	0%
CHIP	ProTide	105.7	0[c]	-	-	-	-	-	-	-	-
CHIP	Tentagel XV	99.0	553.0	192.4	418.6	46%	46.8	8.5	18%	34.9	8%

[a] Not synthesized due to observed increase in back pressure during CHIP synthesis

[b] Synthesis ended in reactor failure due to an increase of back pressure by leakage of fine particles into UV module

3.3. Conclusion

Here we report the performance of three different commercially available PEG-grafted polystyrene resins used for the synthesis of four peptide sequences of varying length and synthetic difficulty using an automated flow synthesizer. The synthesis quality of each synthesis was evaluated by inline UV absorbance to quantify Fmoc deprotection, yields after cleavage from the solid phase, and crude purities. While these studies were performed on a fast-flow automated peptide synthesizer, the same chemical principles should apply to batch synthesizers or manual synthesis assuming similar reaction conditions are used across each resin type. ProTide resin showed excellent performance for short to intermediate length peptides JR10 (10 residues) and GLP-1 (30 residues), but due to technical limitations was unable to be evaluated for synthesis of small proteins MDM2 and CHIP. The technical issues could be resolved through use of finer filter paper or a smaller pore size reactor frit, but it would require additional instrument optimization to ensure comparable fluid delivery and safe operating pressures. Octagel resin underperformed across all four sequences, showing significant side-chain to backbone aggregation during synthesis of JR10 and GLP-1, decreased yields for

MDM2, and an unsuccessful synthesis of CHIP. However, reduction of the loading amount significantly improved the performance of Octagel resin, making it another viable possibility. Tentagel XV resin showed acceptable to superior performance across all sequences and appears to be comparable to reported syntheses on ChemMatrix resin. Summaries of the yields for JR10 and GLP-1 are given in Table 3.1, while yields for MDM2 and CHIP are given in Table 3.2. Representative syntheses were shown here, but synthesis results are highly reproducible as supported by our previous work.7,28,33 Notably, the high reproducibility of automated flow syntheses enabled collection of high quality UV-Vis data for training a deep learning model that can predict synthesis outcomes.14 While not an exhaustive study for the available solid supports for peptide synthesis, this work seeks to provide a survey of current commercial PEG-PS options, recommends products tailored for specific sequence length ranges, and offers a template for evaluation of additional candidates.

3.4. Materials

Fmoc-Rink Amide OctaGel resin (0.441 mmol/g) was purchased from Aapptec, Tentagel XV RAM resin (0.27 mmol/g) was purchased from Rapp Polymere, and ProTide Rink amide resin (0.20 mmol/g) was purchased from CEM Corporation. Reaction vessels were purchased from Torviq equipped with a polypropylene frit. To each vessel was added a disc of Porex filter paper (0.025” thick, 7-12 micron) from Interstate Specialty Products. *N,N*-Dimethylformamide (DMF, biosynthesis grade) was purchased from Millipore Sigma (product DX1732-1). *N,N*-Diisopropylethylamine (DIEA; ReagentPlus ≥99%), piperidine (ACS reagent, ≥99.0%), trifluoroacetic acid (HPLC grade, ≥99.0%), triisopropylsilane (≥98.0%), acetonitrile (AcN, HPLC grade), formic acid (FA, ≥95.0%), 1,2- ethanedithiol (EDT, GC grade, ≥98.0%), and AldraAmine trapping agents (for 1000 - 4000 mL DMF, catalog number Z511706) were purchased from Sigma-Aldrich. Fmoc-protected amino acids (FmocAla-OHxH2O, Fmoc-Arg(Pbf)-OH; Fmoc-Asn(Trt)-OH; Fmoc-Asp-(O-t-Bu)-OH; FmocCys(Trt)-OH; Fmoc-Gln(Trt)-OH; Fmoc-Glu(O-t-Bu)-OH;

Fmoc-Gly-OH; Fmoc-His(Trt)- OH; Fmoc-Ile-OH; Fmoc-Leu-OH; Fmoc-Lys(Boc)-OH; Fmoc-Met-OH; Fmoc-Phe-OH; Fmoc- ProOH; Fmoc-Ser(But)-OH; Fmoc-Thr(t-Bu)-OH; Fmoc-Trp(Boc)-OH; Fmoc-Tyr(O-t-Bu)- OH; Fmoc-Val-OH); Fmoc-Glu(biotinyl-PEG)-OH (product 8.52102, CAS Num. 817169-73-6) were purchased from the Novabiochem product line of Millipore Sigma; Fmoc-His(Boc)-OH and Biotin-PEG$_4$-propionic acid were purchased from ChemPep, Inc. O-(7-azabenzotriazol-1-yl)-*N,N,N',N'*-tetramethyluronium hexafluorophosphate (HATU, ≥97.0%) and (7-azabenzotriazol-1-yloxy)tripyrrolidinophosphonium hexafluorophosphate (PyAOP, ≥97.0%) were purchased from P3 Biosystems. Glacial acetic acid (ACS grade) was purchased from VWR Chemicals. Water was deionized using a Milli-Q water purification system (Millipore). Nylon 0.22 µm syringe filters were TISCH brand SPEC17984.

3.5. Methods

3.5.1. Manual amino acid coupling

Sequences were synthesized using the following amounts of resin weighed into 5 mL Torviq syringes: 50 mg of Octagel resin, 100 mg of Tentagel XV resin, and 100 mg of ProTide resin. Before synthesis, all resins were allowed to swell in amine-free DMF for 15 minutes. For the manual coupling of biotinylated amino acids, resins were deprotected using 20% (v/v) piperidine in DMF (2 x 5 mL with 5 min incubation each time) and washed with DMF (3 x 5 mL) before addition of biotinylated acid (5 equivalents) dissolved in a solution of PyAOP (0.38M, 4.5 equivalents) and activation with DIEA (15 equivalents). Coupling solutions were stirred periodically and incubated for 2 hours. Resins were then washed with DMF (3 x 5 mL) and DCM (3 x 5mL), dried under reduced pressure and stored for later synthesis.

3.5.2. Automated fast-flow peptide synthesis

Before synthesis, all resins were allowed to swell in amine-free DMF for 15 minutes. Utilizing an automated synthesizer, amine-free DMF washed the resin before coupling, after coupling, and after deprotection (40 strokes, ~25 mL).

Coupling was performed with HATU (single-coupling, 8 strokes, ~5 mL) except S and A with HATU (double-coupling, 21 strokes, ~10 mL) and C, H, N, Q, R, V, T with PyAOP (double-coupling, 21 strokes, ~10 mL). Deprotection was completed with 20% piperidine in amine-free DMF with 2% formic acid (13 pump strokes, ~5 mL). Amino acids were iteratively coupled and deprotected until the stepwise synthesis was complete. After automated synthesis, the resin was washed again with DMF (3 x 5 mL) and DCM (3 x 5 mL) then dried under reduced pressure. For a detailed explanation of the instrument setup and related chemistries, see Hartrampf et al.,[7] Simon et al.,[31,32] Mijalis et al.,[33] and Mong et al.[34] Sequences synthesized were as follows:

1. JR10: WFTTLISTIM-NH_2

2. GLP-1: HAEGTFTSDV SSYLEGQAAK EFIAWLVKGR-NH_2

3. MDM2: TLVRPKPLLL KLLKSVGAQK DTYTMKEVLF YLGQYIMTKR LYDEKQQHIV YCSNDLLGDL FGVPSFSVKE HRKIYTMIYR NLVV-NH_2

4. CHIP: (Biotin-PEG_4)-SPSAQELKEQ GNRLFVGRKY PEAAACYGRA ITRNPLVAVY YT- NRALCYLK MQQHEQALAD CRRALELDGQ SVKAHFFLGQ CQLEMESYDE AIAN- LQRAYS LAKEQRLNFG DDIPSALRIA KKKRWNSIEE RR-NH_2

3.5.3. Integration of synthesizer UV signal

UV absorbance at 310 nm was continuously monitored over the course of synthesis using an Agilent G1315D 1260 variable length diode array detector. Deprotection peaks were identified based on syncing timepoints to pump steps and baseline corrected using the PeakUtils package (Version 1.3.3) in Python (Version 3.9.6). Integrated areas, heights, and widths of the deprotection peaks were normalized to the second deprotection peak. Integration was done using a cumulative sum of rectangles based on the time step of a single pump stroke (~1.3 sec).

3.5.4. Cleavage of peptides from the solid-phase support

Before cleavage, all dried resins were weighed to assess crude on-resin yield. Cleavage from solid phase and global deprotection was performed using a solution of 94% trifluoroacetic acid, 2.5% water, 2.5% ethanedithiol and 1% triisopropylsilane for JR10 and GLP-1 resins, and Reagent K (82.5% trifluoroacetic acid, 5% water, 5% phenol, 5% thioanisole, 2.5% ethanedithiol) was used for MDM2 and CHIP. Each resin was suspended in 2 mLs of their respective cleavage cocktail, with JR10 and GLP-1 resins left to incubate for 2 hours at room temperature and MDM2 and CHIP resins left to incubate for 3 hours at room temperature. The peptides were triturated with cold diethyl ether (3 x 15 mL for JR10 and GLP-1, 1 x 45 mL and 2 x 25 mL for MDM2 and CHIP), dried gently using N2, suspended in 50% acetonitrile in water (0.1% trifluoroacetic acid), and lyophilized. Lyophilized powders were weighed to give crude yields post-resin cleavage.

3.5.5. Liquid chromatography-mass spectrometry (LC-MS)

LC-MS chromatograms and associated high resolution mass spectra were acquired using an Agi- lent 1290 Infinity HPLC coupled to an Agilent 6550 Q-TOF iFunnel mass spectrometer using a Phenomenex Jupiter C4 column (150 x 1.0 mm ID, 5 µm, 300Å silica) heated at 40 °C or a Zorbax 300SB-C3 column (150 x 2.1 mm ID, 5 µm, 300Å silica) at 40 °C. Solvent compositions were 0.1% formic acid in water (solvent A) and 0.1% formic acid in acetonitrile (solvent B). Method 1 was used for characterization of crude material, and methods 2 and 3 were used for fraction analysis after semi-preparative HPLC purification.

1. Column: Zorbax 300SB-C3. Gradient: linear gradient 5-65% B from 0-30 min; isocratic 91% B from 30-32 min; post time 5% B for 3 min. Flow rate: 0.4 mL/min. MS data was collected from 1-30 min; MS was run in positive ionization mode, extended dynamic range (2 GHz), and standard mass range (m/z in the range of 300 to 3000 a.m.u.).

2. Column: Jupiter C4. Gradient: isocratic 1% B from 0-2 min; linear gradient 1-91% B from 2-8 min; isocratic 95% B from 8-10 min; post time 1% B for 1 min. Flow rate: 0.3 mL/min. MS data was collected from 2-8 min; MS was run in positive ionization mode, extended dynamic range (2 GHz), and standard mass range (m/z in the range of 300 to 3000 a.m.u.).

3. Column: Jupiter C4. Gradient: isocratic 1% B from 0-3 min; linear gradient 1-91% B from 3-15 min; isocratic 95% B from 15-18 min; post time 1% B for 2 min. Flow rate: 0.3 mL/min. MS data was collected from 3-15 min; MS was run in positive ionization mode, extended dynamic range (2 GHz), and standard mass range (m/z in the range of 300 to 3000 a.m.u.).

3.5.6. Analytical ultra high-performance liquid chromatography (UHPLC)

Analytical HPLC analysis was performed using an Agilent 1200 series system with UV detection at 214 nm on a Zorbax 300SB-C3 column (150 x 2.1 mm ID, 5 µm, 300Å silica) on an Agilent 1200 HPLC at room temperature. Solvent compositions were 0.1% trifluoroacetic acid in water (solvent A) and 0.08% trifluoroacetic acid in acetonitrile (solvent B). Gradient: linear gradient 5-65% B from 0-60 min; linear gradient 65-100% B from 60-61 min; isocratic 100% B from 61-66 min; linear gradient 100-5% B from 66-67 min; isocratic 5% B from 67-75 min. Flow rate: 0.400 mL/min. A solvent-only blank injection was subtracted from each run before determining purity through manual integration of all signals from 0 to 61 min.

3.5.7. Semi-preparative high-performance liquid chromatography of MDM2 and CHIP

Lyophilized crude sample of protein was weighed in batches of ~35 mg, dissolved in 10 mL of 6 M guanidinium chloride, 0.1 M dithiothreitol, in 50 mM sodium phosphate pH 7.5, vortexed briefly, 0.2 µm filtered, and subjected to RP-HPLC purification using an Agilent Zorbax 300SB-C18 PrepHT (9.4 × 250 mm, 5 µm) column with an Agilent C3 Zorbax SB 300 guard column heated at 60 °C at 4.0 mL/min with the gradients listed in the subsequent section. Purification was performed on an Agilent mass-directed purification system (1260 Infinity LC and 6130 Single Quad MS) with a Timberline Instrument TL105 HPLC column heater. Fractions showing high purity charge state series were combined, lyophilized, and analyzed via LC-MS and analytical HPLC.

1. MDM2: Isocratic 5% B from 0-2 min; linear gradient 5-30% B from 2-32 min; linear gradient 30-50% B from 32-132 min; linear gradient 50-65% B from 132-133 min; linear gradient 65-80% B from 133-136 min; post time 5% B for 10 min.

2. CHIP: Isocratic 5% B from 0-2 min; linear gradient 5-35% B from 2-32 min; linear gradient 35-55% B from 32-132 min; linear gradient 55-75% B from 132-133 min; linear gradient 75-90% B from 133-136 min; post time 5% B for 10 min.

3.6. Appendix I: UV signals from fast-flow synthesizer

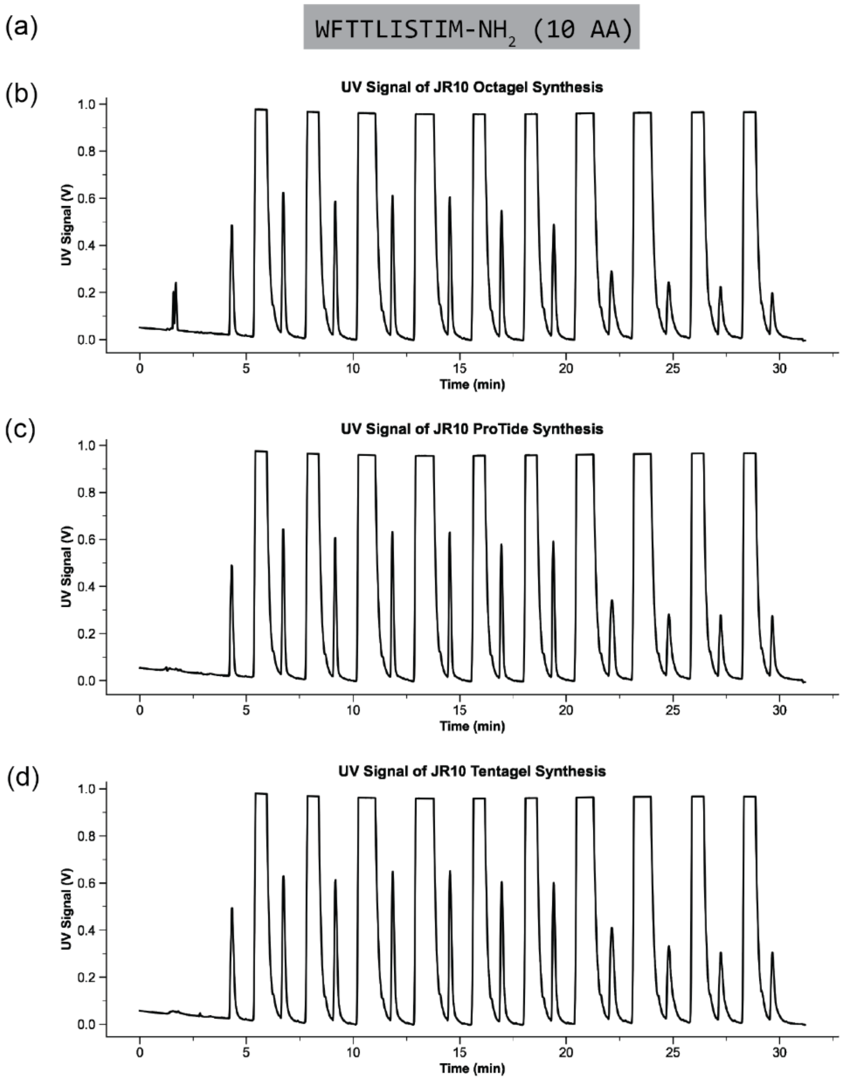


Figure 3.6: *Sequence of JR10 (a) and raw UV absorbance data for the synthesis of JR10 on (b) Octagel resin, (c) ProTide resin, and (d) Tentagel XV RAM resin. Each synthesis begins with washing the resin, followed by deprotection of the Fmoc protecting group as seen by the lower intensity peaks. The subsequent amino acid is then coupled, as shown by the high intensity saturated peaks.*

(a) HAEGTFTSDV SSYLEGQAAK EFIAWLVKGR-NH_2 (30 AA)

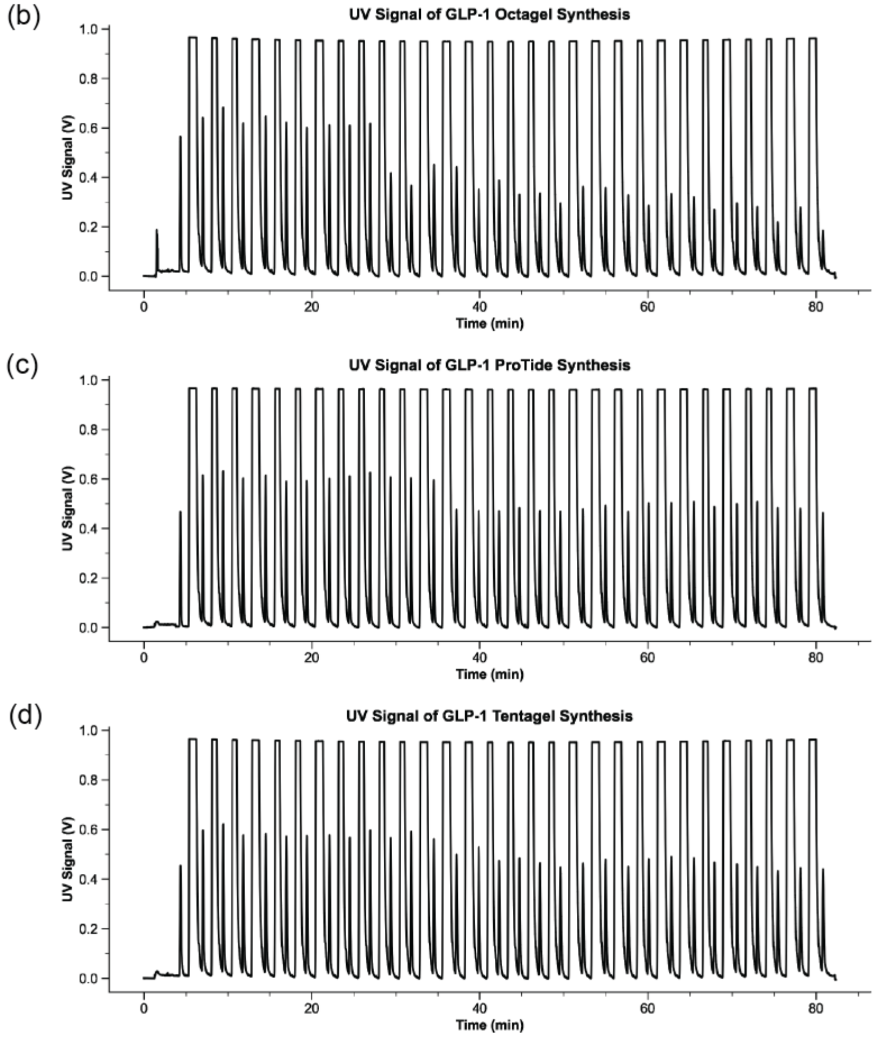
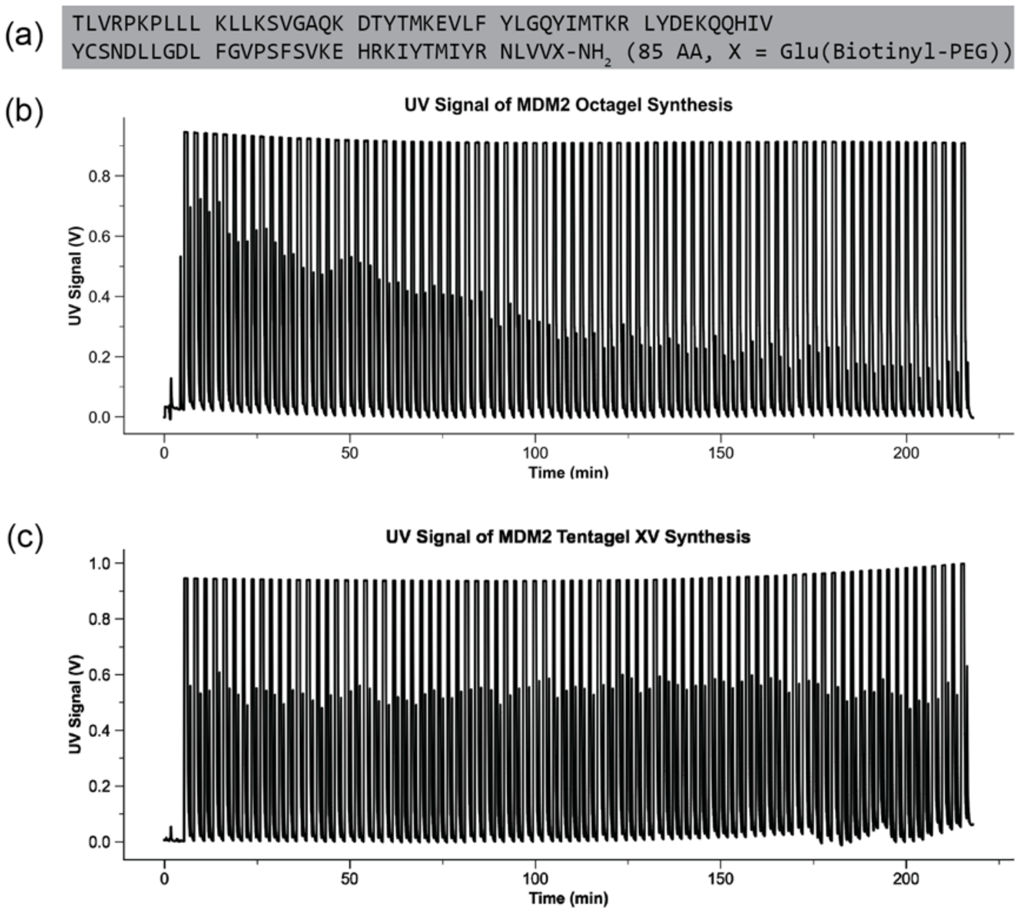


Figure 3.7: *Sequence of GLP-1 (a) and raw UV absorbance data for the synthesis of GLP-1 on (b) Octagel resin, (c) ProTide resin, and (d) Tentagel XV RAM resin. Each synthesis begins with washing the resin, followed by deprotection of the Fmoc protecting group as seen by the lower intensity peaks. The subsequent amino acid is then coupled, as shown by the high intensity saturated peaks.*



(a)

TLVRPKPLLL KLLKSVGAQK DTYTMKEVLF YLGQYIMTKR LYDEKQQHIV
YCSNDLLGDL FGVPSFSVKE HRKIYTMIYR NLVVX-NH_2 (85 AA, X = Glu(Biotinyl-PEG))

Figure 3.8: *Sequence of MDM2 (a) and raw UV absorbance data for the synthesis of MDM2 on (b) Octagel resin and (c) Tentagel XV RAM resin. Each synthesis begins with washing the resin, followed by deprotection of the Fmoc protecting group as seen by the lower intensity peaks. The subsequent amino acid is then coupled, as shown by the high intensity saturated peaks.*

(a)

```
ZSPSAQELKE QGNRLFVGRK YPEAAACYGR AITRNPLVAV YYTNRALCYL
KMQQHEQALA DCRRALELDG QSVKAHFFLG QCQLEMESYD EAIANLQRAY
SLAKEQRLNF GDDIPSALRI AKKKRWNSIE ERR-NH2 (133 AA, Z = Biotin-Peg4)
```

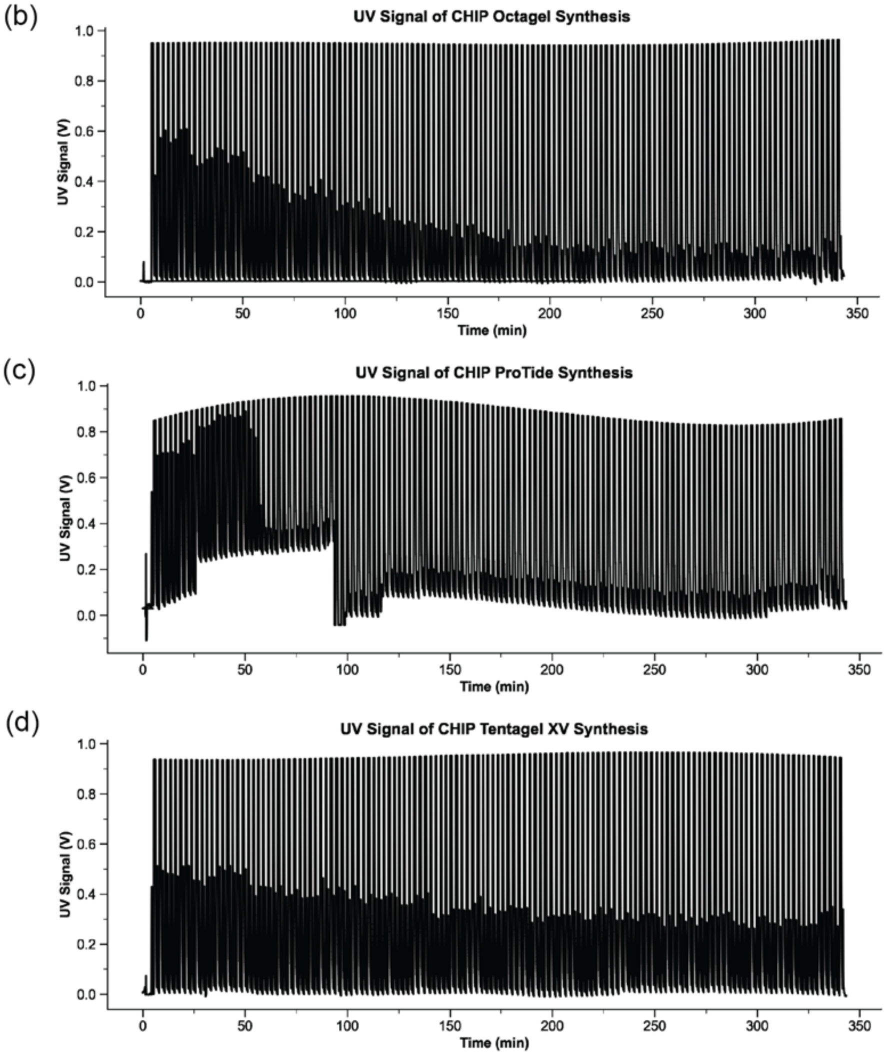


Figure 3.9: *Sequence of CHIP (a) and raw UV absorbance data for the synthesis of CHIP on (b) Octagel resin, (c) ProTide resin, and (d) Tentagel XV RAM resin. Each synthesis begins with washing the resin, followed by deprotection of the Fmoc protecting group as seen by the lower intensity peaks. The subsequent amino acid is then coupled, as shown by the high intensity saturated peaks.*

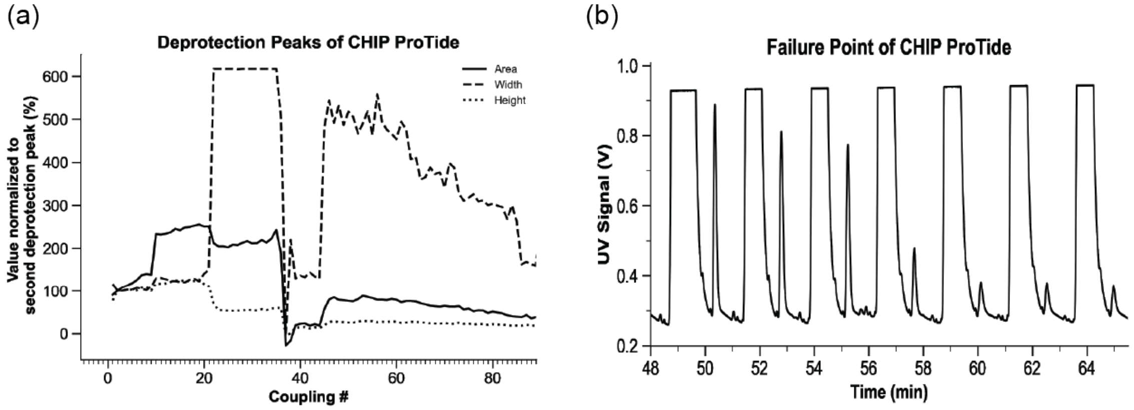


Figure 3.10: *Synthesis data of CHIP using ProTide resin during reactor failure. (a) Deprotection peak characterization during synthesis shows aberrant behavior early in the sequence with a point of failure less than 40 coupling cycles into the stepwise synthesis process. (b) Close up view of UV absorbance at time of reactor failure, resulting in an almost complete disappearance of deprotection peaks.*

3.7. Appendix II: LC-MS and UHPLC characterization data

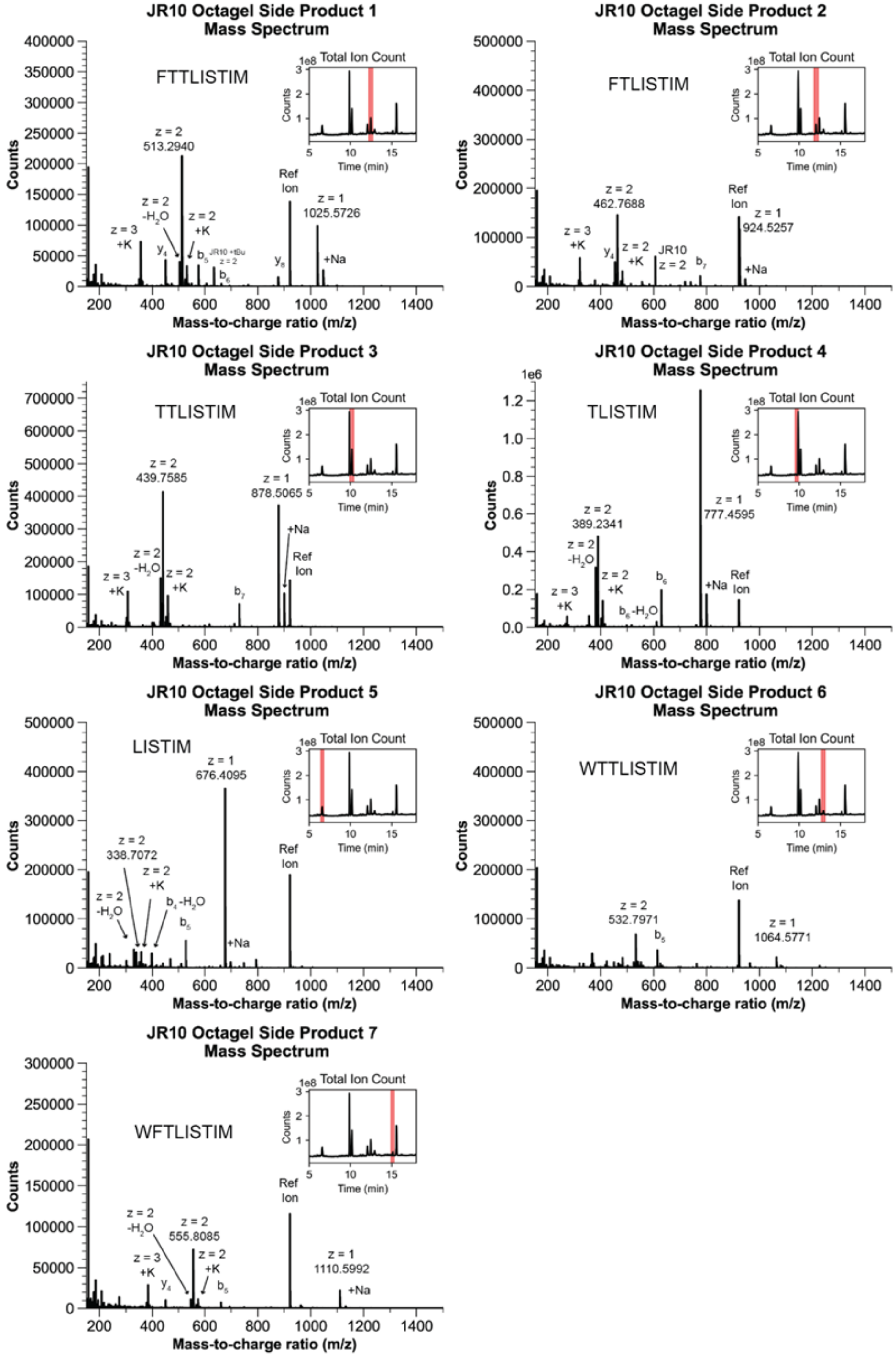


Figure 3.11: *LC-MS analysis of truncation products from the synthesis of JR10 on Octagel resin. The panels depict deconvoluted mass spectra of the bands highlighted in red from the corresponding total ion count chromatogram (TICC) shown in insets.*

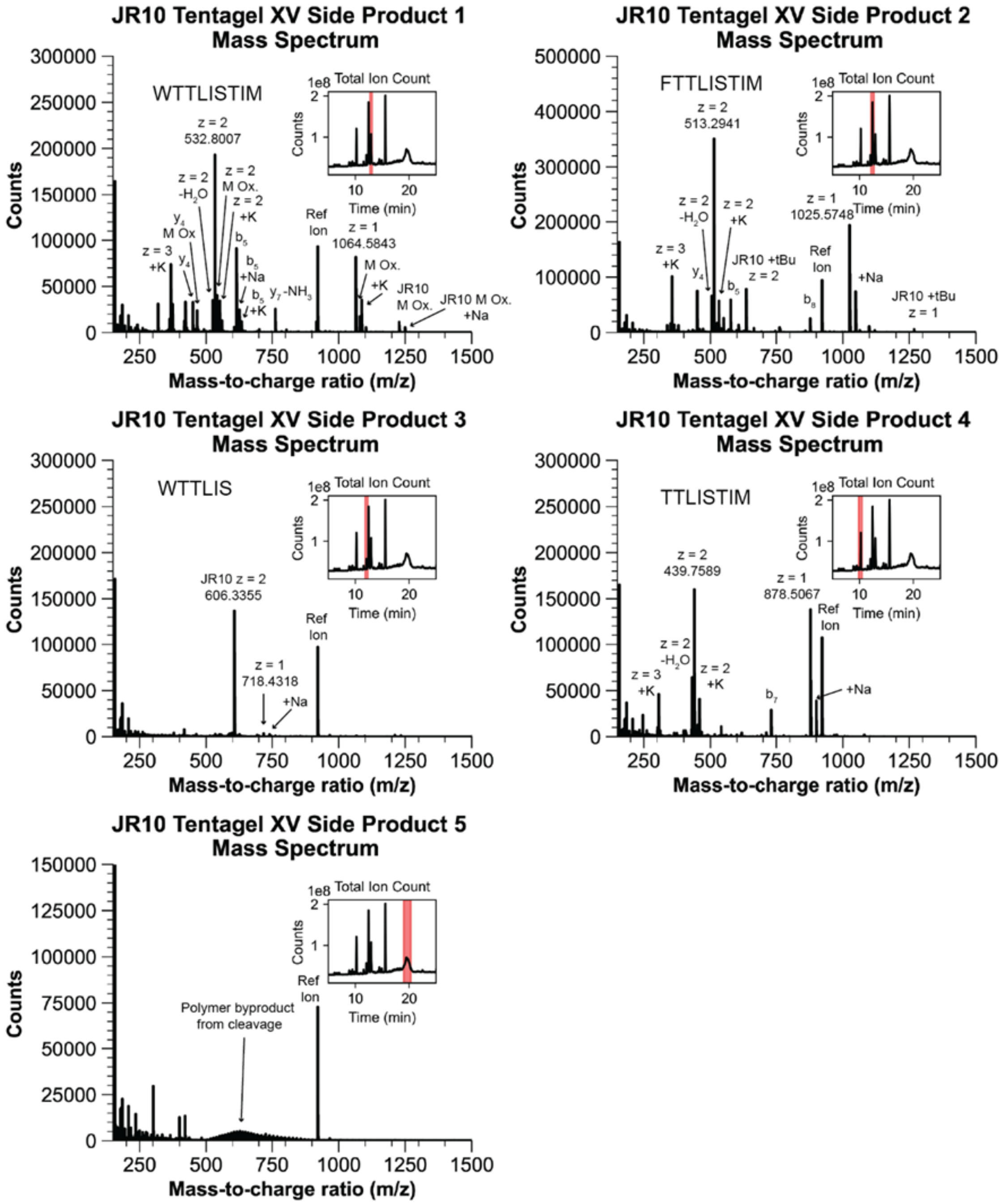


Figure 3.12: *LC-MS analysis of truncation products from the synthesis of JR10 on Tentagel XV RAM resin. The panels depict deconvoluted mass spectra of the bands highlighted in red from the corresponding total ion count chromatogram (TICC) shown in insets.*

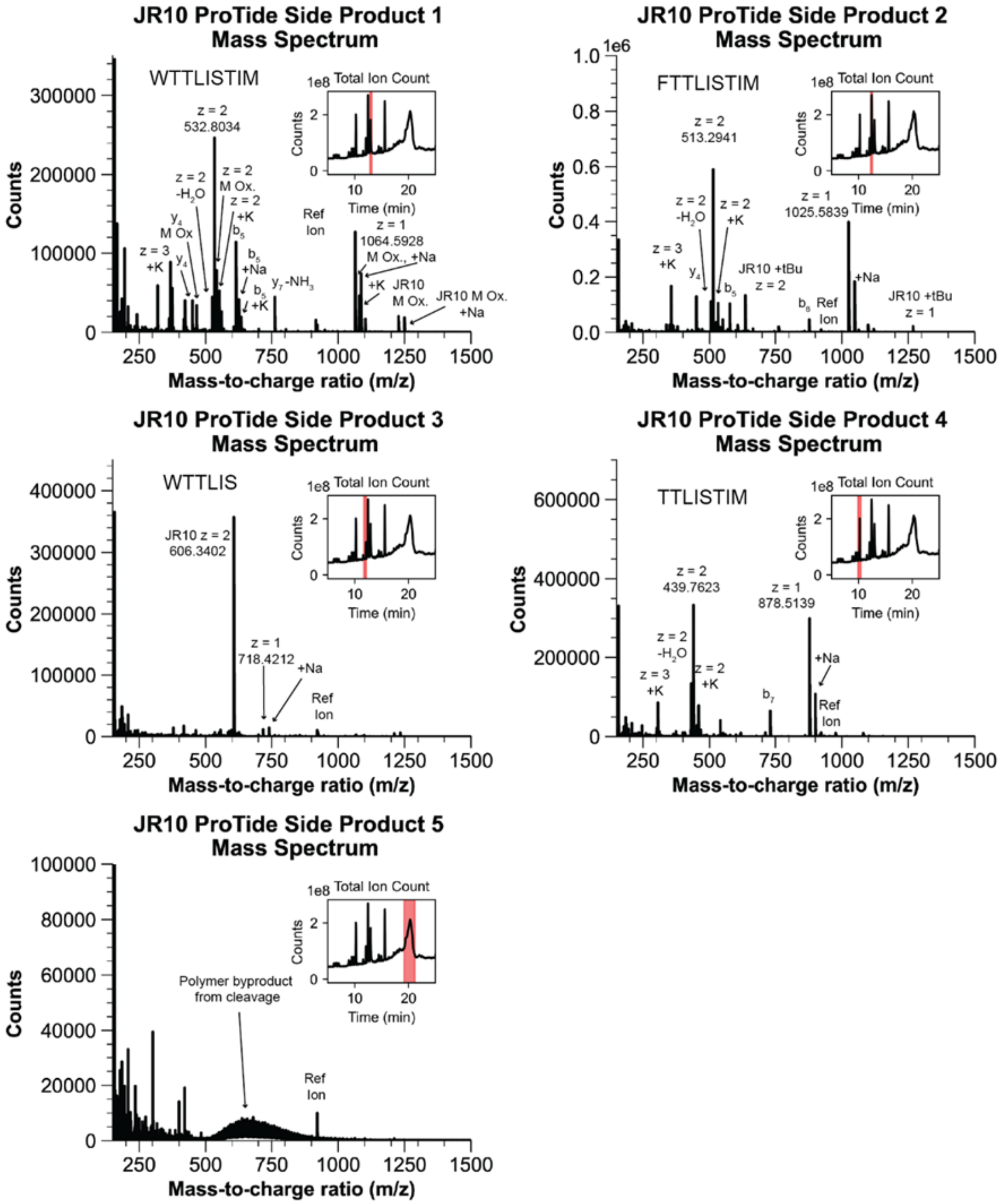


Figure 3.13: *LC-MS analysis of truncation products from the synthesis of JR10 on Tentagel XV RAM resin. The panels depict deconvoluted mass spectra of the bands highlighted in red from the corresponding total ion count chromatogram (TICC) shown in insets.*

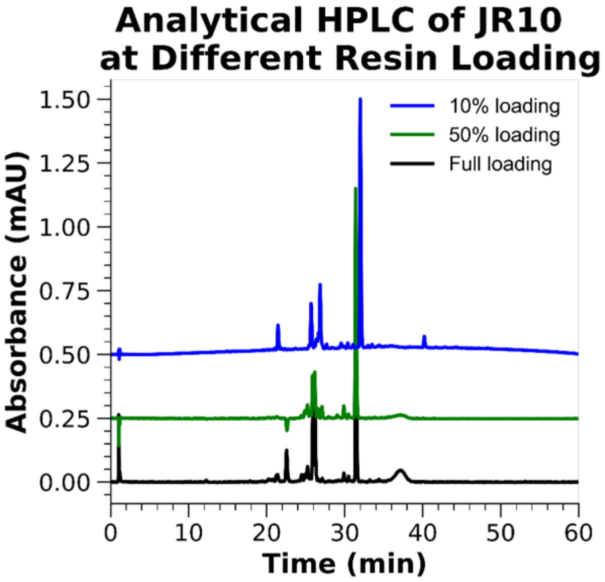


Figure 3.14: *Analytical HPLC analysis of JR10 syntheses done with Tentagel XV RAM resin at full loading, 50% total loading and 10% total loading after capping of the resin with the respective amounts of acetic acid. Crude purities for the 50% and 10% loading resins were 61% and 46%, respectively (compared to 32% for the full loading), showing that lowering the loading of the resin can help prevent aggregation.*

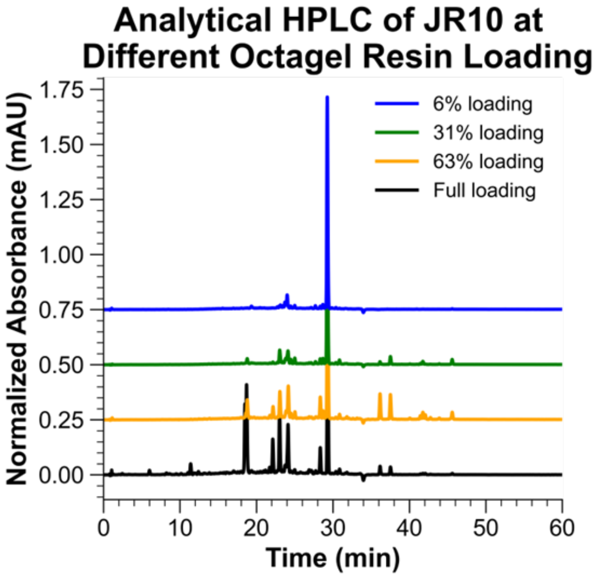


Figure 3.15: *Analytical HPLC analysis of JR10 syntheses done Octagel resin at full loading, 63% total loading, 31% total loading and 6% total loading after capping of the resin with the respective amounts of acetic acid. Crude purities for the 63%, 31% and 6% loading resins were 38%, 60% and 80%, respectively (compared to 24% for the full loading), showing that lowering the loading of the resin can help prevent aggregation.*

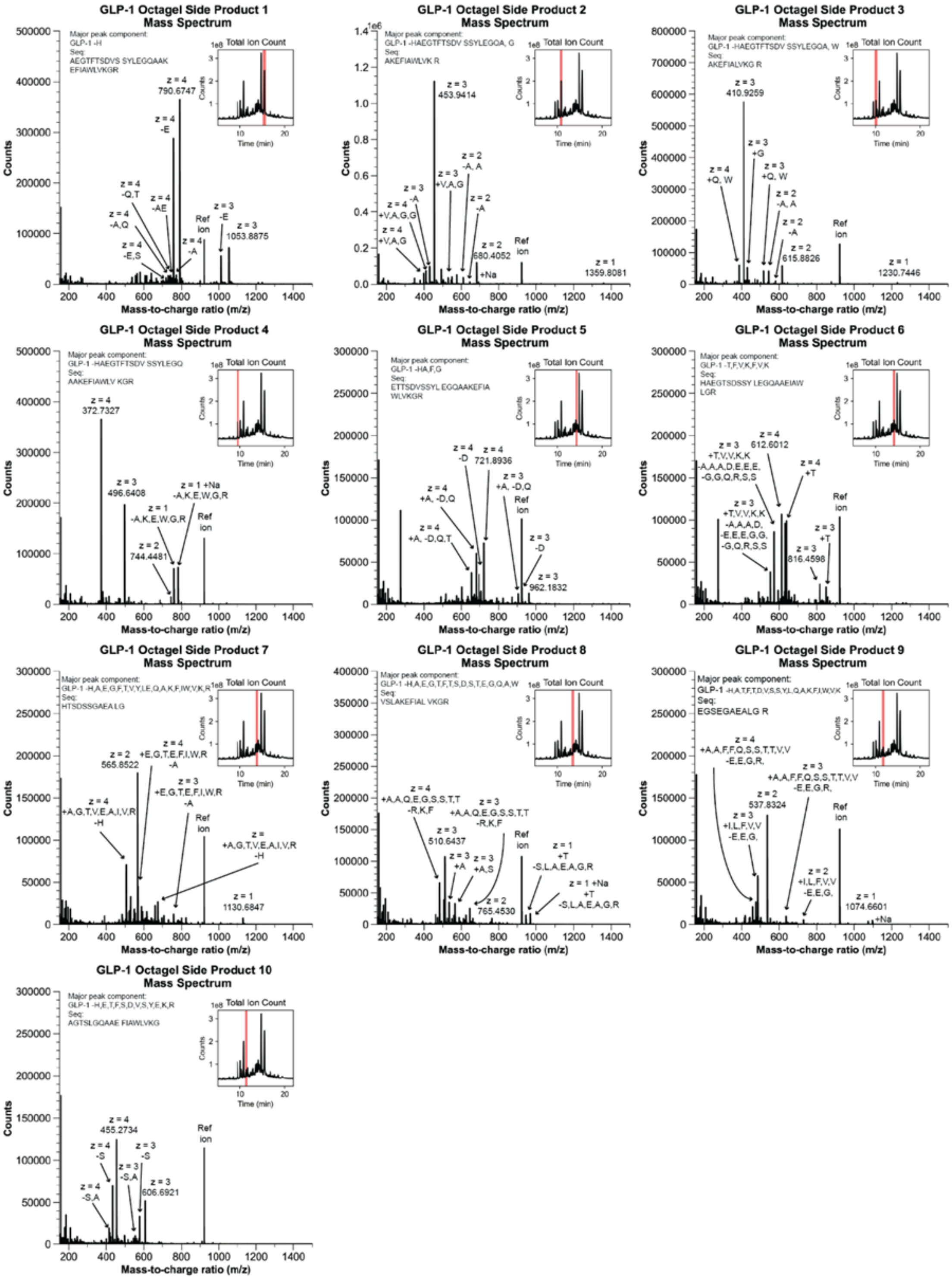


Figure 3.16: *LC-MS analysis of truncation products from the synthesis of GLP-1 on Octagel resin. The panels depict deconvoluted mass spectra of the bands highlighted in red from the corresponding total ion count chromatogram (TICC) shown in insets.*

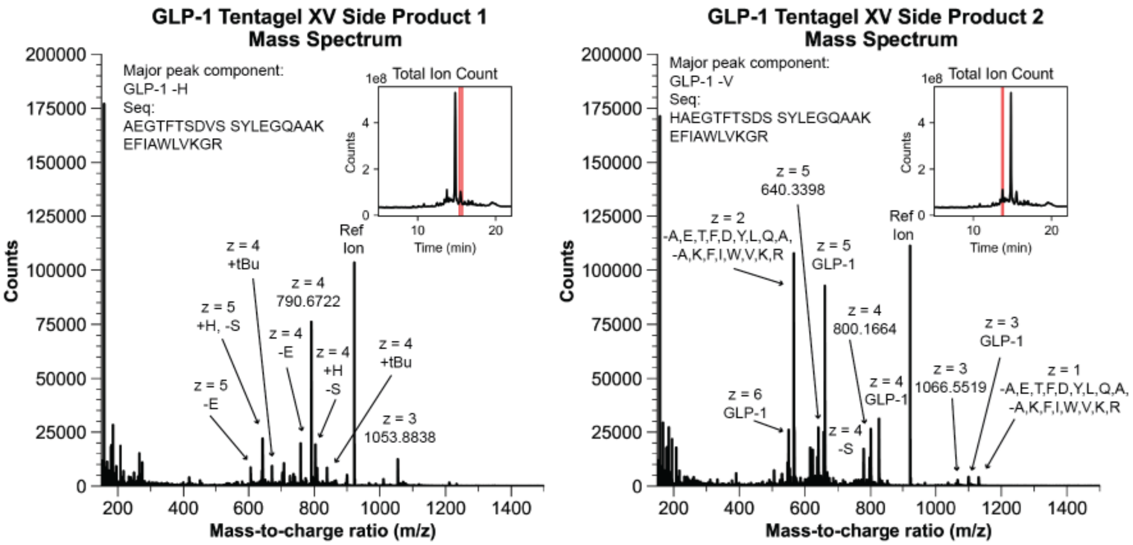


Figure 3.17: *LC-MS analysis of truncation products from the synthesis of GLP-1 on Tentagel XV RAM resin. The panels depict deconvoluted mass spectra of the bands highlighted in red from the corresponding total ion count chromatogram (TICC) shown in insets.*

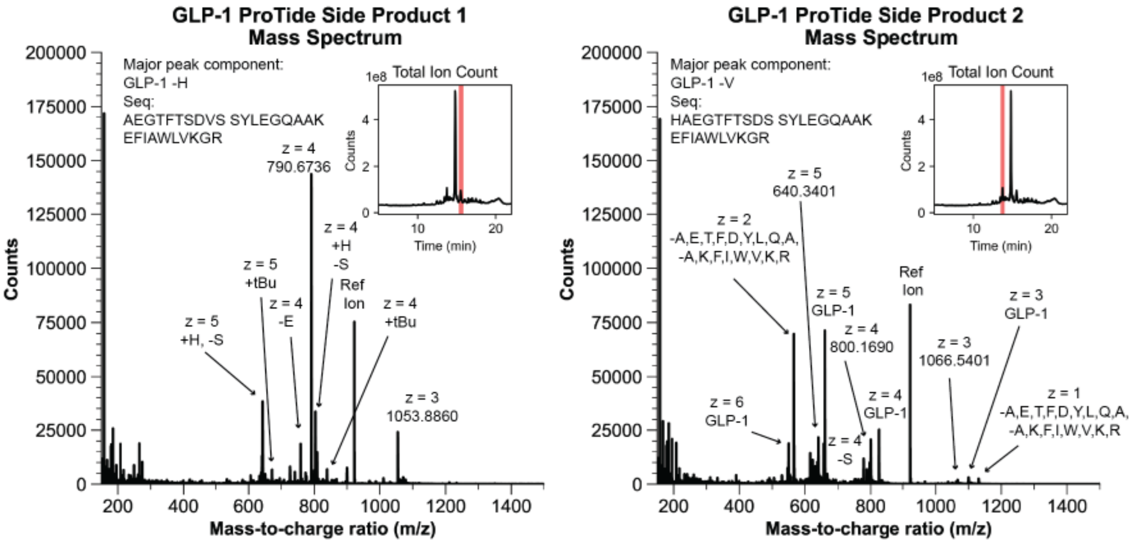


Figure 3.18: *LC-MS analysis of truncation products from the synthesis of GLP-1 on ProTide resin. The panels depict deconvoluted mass spectra of the bands highlighted in red from the corresponding total ion count chromatogram (TICC) shown in insets.*

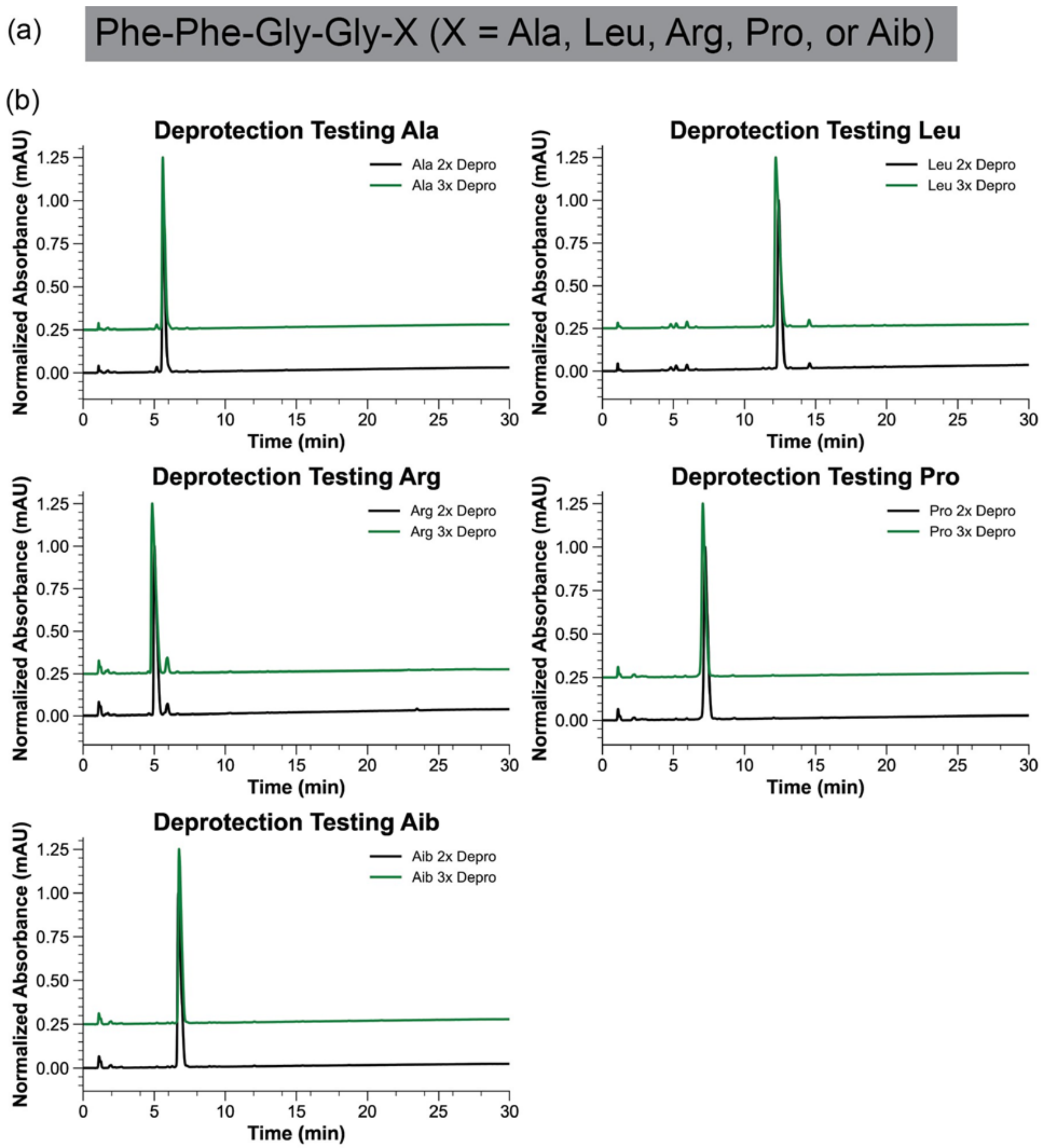


Figure 3.19: *Analytical HPLC analysis of syntheses to test for incomplete manual couplings dependent upon the success of the initial Fmoc deprotection. (a) Test sequence used for each synthesis, with a variable residue at the C terminus to test individual monomer effects. (b) Analytical HPLC data for five different C terminal monomers, showing no significant differences between two five-minute deprotection reactions compared to three five-minute deprotection reactions.*

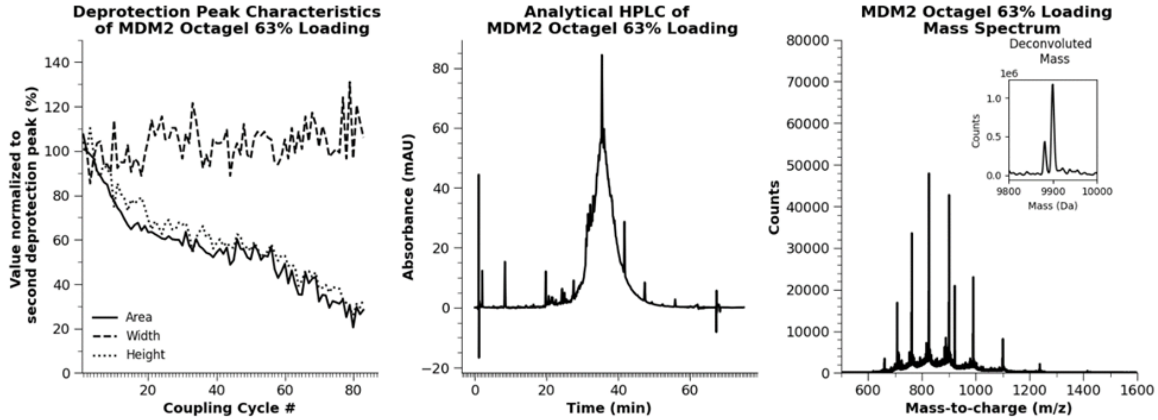


Figure 3.20: *UV absorbance, crude analytical HPLC, and crude LC-MS data for the synthesis of MDM2 on Octagel resin that has been reduced to 63% of its original loading, roughly matching the loading of Tentagel XV resin. Calculated deconvoluted mass: 9899.82; Observed deconvoluted mass: 9899.7 (-9.0 ppm error).*

3.8. Acknowledgements

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. We thank Charlie Farquhar and Maysa Ilamanova for their assistance with manuscript preparation.

3.9. References

1. Coin, I.; Beyermann, M.; Bienert, M. *Nat Protoc* **2007**, 2, 3247–3256.
4. Wang, L.; Wang, N.; Zhang, W.; Cheng, X.; Yan, Z.; Shao, G.; Wang, X.; Wang, R.; Fu, C. *Sig Transduct Target Ther* **2022**, 7, 1–27.
5. Zompra, A. A.; Galanis, A. S.; Werbitzky, O.; Albericio, F. *Future Medicinal Chemistry* **2009**, 1, 361–377.
6. Shelton, P. T.; Jensen, K. J. *In Peptide Synthesis and Applications*, Jensen, K. J., Tofteng Shelton, P., Pedersen, S. L., Eds.; Methods in Molecular Biology; Humana Press: Totowa, NJ, 2013, pp 23–41.
7. Merrifield, R. B. *J. Am. Chem. Soc.* **1963**, 85, 2149–2154.
8. Saebi, A.; Brown, J. S.; Marando, V. M.; Hartrampf, N.; Chumbler, N. M.; Hanna, S.; Poskus, M.; Loas, A.; Kiessling, L. L.; Hung, D. T.; Pentelute, B. L. *ACS Chem. Biol*. **2023**, 18, 518–527.
9. Hartrampf, N.; Saebi, A.; Poskus, M.; Gates, Z. P.; Callahan, A. J.; Cowfer, A. E.; Hanna, S.; Antilla, S.; Schissel, C. K.; Quartararo, A. J.; Ye, X.; Mijalis, A. J.; Simon, M. D.; Loas, A.; Liu, S.; Jessen, C.; Nielsen, T. E.; Pentelute, B. L. *Science* **2020**, 368, 980–987.
10. Furka, Á.; Sebestyén, F.; Asgedom, M.; Dibó, G. *International Journal of Peptide and Protein Research* **1991**, 37, 487–493.
11. Moss, J. A. *Current Protocols in Protein Science* **2005**, 40, 18.7.1–18.7.19.
12. Groth, T.; Grøtli, M.; Meldal, M. *J. Comb. Chem.* **2001**, 3, 461–468.
13. Rana, S.; White, P.; Bradley, M. *J. Comb. Chem*. **2001**, 3, 9–15.
14. De la Torre, B. G.; Jakab, A.; Andreu, D. Int J Pept Res Ther 2007, 13, 265–270.
15. Kates, S. A.; McGuinness, B. F.; Blackburn, C.; Griffin, G. W.; Solé, N. A.; Barany, G.; Albericio, F. *Peptide Science* **1998**, 47, 365–380.
16. Mohapatra, S.; Hartrampf, N.; Poskus, M.; Loas, A.; Gómez-Bombarelli, R.; Pentelute, B. L. *ACS Cent. Sci*. **2020**, 6, 2277–2286.
17. Atherton, E.; Dryland, A.; Shephard, R.; Wade, J. In Peptides: *Structure and Function*; Proceedings of the Eigth American Peptide Symposium 8; Pierce Chemical Company: 1983, pp 45–54.
18. Cardona, V.; Eberle, I.; Barthélémy, S.; Beythien, J.; Doerner, B.; Schneeberger, P.; Keyte, J.; White, P. D. *Int J Pept Res Ther* **2008**, 14, 285–292.
19. Haack, T.; Mutter, M. *Tetrahedron Letters* **1992**, 33, 1589–1592.
20. Bacsa, B.; Horváti, K.; Bõsze, S.; Andreae, F.; Kappe, C. O. *J. Org. Chem.* **2008**, 73, 7532–7542.
21. Bicciato, S.; Bagno, A.; Dettin, M.; Di Bello, C. *IFAC Proceedings Volumes* **1995**, 28, 12–16.
22. Bürgisser, H.; Williams, E. T.; Lescure, R.; Premanand, A.; Jeandin, A.; Hartrampf, N. A versatile “Synthesis Tag” (SynTag) for the chemical synthesis of aggregating peptides and proteins, *ChemRxiv Preprint*. DOI: 10.26434/chemrxiv-2023-7mz2c-v2, 2023.

23. Paradís-Bas, M.; Tulla-Puche, J.; Albericio, F. *Chem. Soc. Rev*. **2016**, 45, 631–654.
24. Fields, G. B., *Solid-Phase Peptide Synthesis,* 2nd; Methods in Enzymology, Vol. 289; Academic Press: 1997.
25. Miklos, B., *Principles of Peptide Synthesis*, 2nd; Springer Science & Business Media: 1993.
26. Amblard, M.; Fehrentz, J.-A.; Martinez, J.; Subra, G. *Mol Biotechnol* **2006**, 33, 239–254.
27. Meldal, M., *Solid-Phase Peptide Synthesis*; Methods in Enzymology, Vol. 289; Academic Press: 1997, pp 83–104.
28. García-Martín, F.; Quintanar-Audelo, M.; García-Ramos, Y.; Cruz, L. J.; Gravel, C.; Furic, R.; Côté, S.; Tulla-Puche, J.; Albericio, F. *J Comb Chem* **2006**, 8, 213–220.
29. ChemMatrix® Discontinued https://www.iris-biotech.de/en/blog/chemmatrixr-discontinued/ (accessed 10/13/2023).
30. Callahan, A. J.; Gandhesiri, S.; Travaline, T. L.; Salazar, L. L.; Hanna, S.; Lee, Y.-C.; Li, K.; Tokareva, O. S.; Swiecicki, J.-M.; Loas, A.; Verdine, G. L.; McGee, J. H.; Pentelute, B. L. Single-Shot Flow Synthesis of D-Proteins for Mirror-Image Phage Display, *ChemRxiv Preprint*. DOI: 10.26434/chemrxiv-2023-x86xp, 2023.
31. Carpino, L. A.; Krause, E.; Sferdean, C. D.; Schümann, M.; Fabian, H.; Bienert, M.; Beyermann, M. *Tetrahedron Letters* **2004**, 45, 7519–7523.
32. Sletten, E. T.; Nuño, M.; Guthrie, D.; Seeberger, P. H. *Chem. Commun*. **2019**, 55, 14598–14601.
33. Simon, M. D.; Mijalis, A. J.; Totaro, K. A.; Dunkelmann, D.; Vinogradov, A. A.; Zhang, C.; Maki, Y.; Wolfe, J. M.; Wilson, J.; Loas, A.; Pentelute, B. L. *In Total Chemical Synthesis of Proteins*; John Wiley & Sons, Ltd: 2021; Chapter 2, pp 17–57.
34. Simon, M. D.; Heider, P. L.; Adamo, A.; Vinogradov, A. A.; Mong, S. K.; Li, X.; Berger, T.; Policarpo, R. L.; Zhang, C.; Zou, Y.; Liao, X.; Spokoyny, A. M.; Jensen, K. F.; Pentelute, B. L. *ChemBioChem* **2014**, 15, 713–720.
35. Mijalis, A. J.; Thomas, D. A.; Simon, M. D.; Adamo, A.; Beaumont, R.; Jensen, K. F.; Pentelute, B. L. *Nature Chemical Biology* **2017**, 13, 464–466.
36. Mong, S. K.; Vinogradov, A. A.; Simon, M. D.; Pentelute, B. L. *ChemBioChem* **2014**, 15, 721–733.
37. Collins, J. M.; Porter, K. A.; Singh, S. K.; Vanier, G. S. *Org. Lett*. **2014**, 16, 940–943.
38. Gates, Z. P.; Hartrampf, N. *Peptide Science* **2020**, 112, e24198.
39. Bianchi, P.; Williams, J. D.; Kappe, C. O. *J Flow Chem* **2020**, 10, 475–490.
40. Catchpole, O. J.; Bernig, R.; King, M. B. *Ind. Eng. Chem. Res*. **1996**, 35, 824–828.
41. Gunn, D. J. *Chem. Eng. Sci*. **1987**, 42, 363–373.
42. Mueller, L. K.; Baumruck, A. C.; Zhdanova, H.; Tietze, A. A. *Frontiers in Bioengineering and Biotechnology* **2020**, 8, 1–16.

43. Pessi, A.; Mancini, V.; Filtri, P.; Chiappinelli, L. *Int J Pept Protein Res* **1992**, 39, 58–62.

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

The work presented in this chapter has been reproduced and adapted from the following publication:

Lee, M.A.*; Brown, J.S.*; Loas, A.; Pentelute, B.L. pyBinder: Label-free Quantitation to Advance Affinity Selection-Mass Spectrometry. *Manuscript in preparation*. **2024**

**Denotes equal contribution*

4.1. Introduction

Affinity selection-mass spectrometry (AS-MS) discovers high-affinity ligands to biomolecular targets using mass spectrometry for ligand identification.[1–3] The affinity selection of AS-MS is highly similar to phage and mRNA display,[4–6] though AS-MS generally utilizes a single enrichment step without genetic amplification. AS-MS utilizes synthetic libraries, providing unfettered access to non-natural amino acids and a facile design opportunity to tailor libraries toward the target. Thus, one of the primary uses of AS-MS is the selection of small combinatorial libraries (10^3-10^6) 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 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 prominence, AS-MS heavily depends on mass spectrometry analysis and stands to benefit by leveraging methods from the field of MS-based proteomics.

Solutions developed to combat data incompleteness in the field of proteomics could be highly valuable to improve AS-MS. MS-based proteomics has long detailed the "missing value" problem, hallmarked by an incomplete series of peptides or proteins expected across samples or replicates.[17–20] This challenge is pronounced in approaches that use data dependent acquisition (DDA), where precursors are selected from the mass spectrum (MS^1) for tandem MS^2 fragmentation. Precursor ions are often selected in order of their signal intensity in DDA, biasing the discovery of highly ionizing species. While the rules for precursor selection for MS^2 are clearly outlined, the precursor selection process is not perfectly reproducible and is instead stochastic. This stochasticity can hinder further data analysis, ranging from the identified peptides across technical replicates as well as statistical analysis for sample comparison.

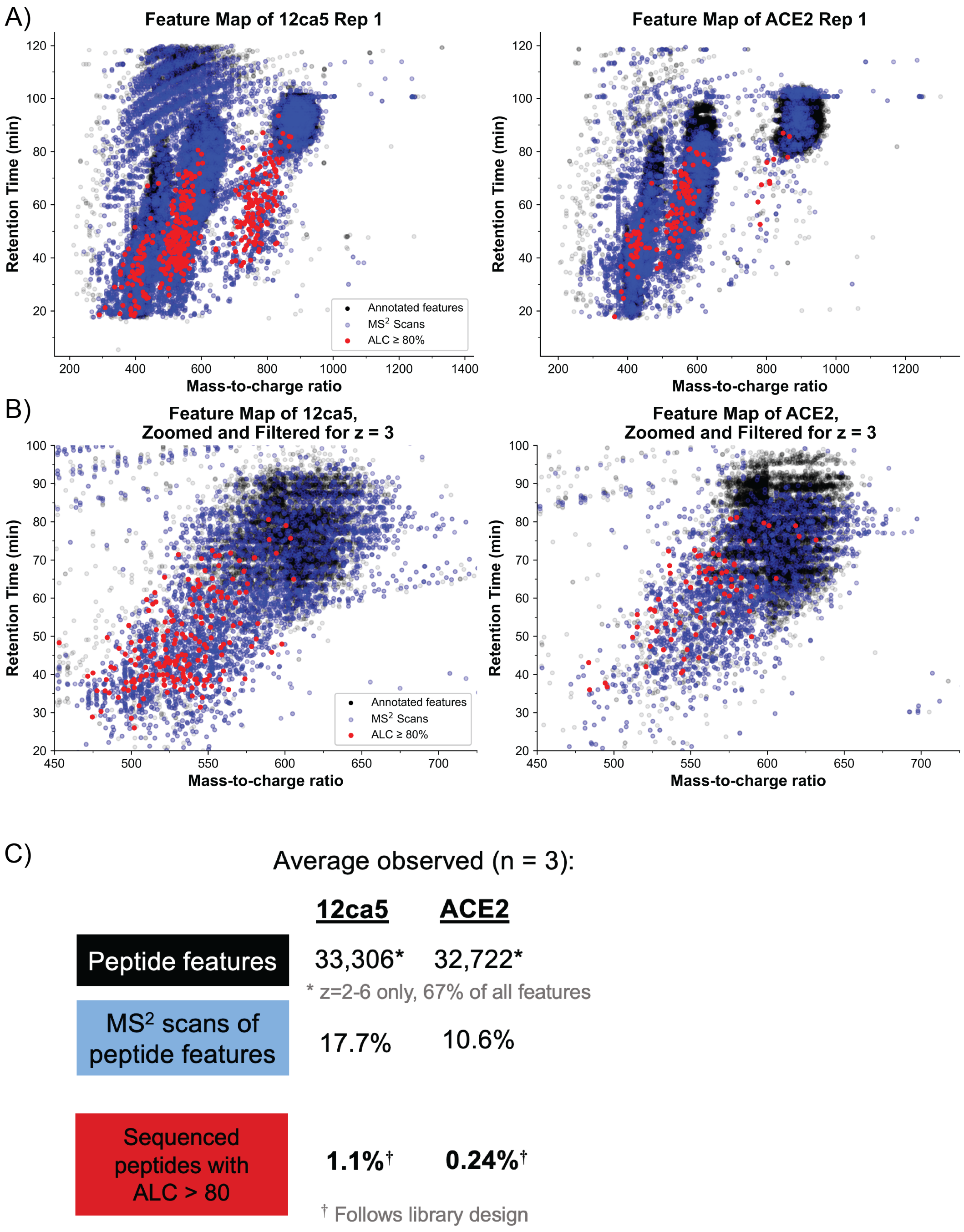
Label-free quantitation (LFQ) is directly compatible with AS-MS and provides a solution to incomplete data without relying on peptide sequence identification from MS^2 fragmentation. LFQ has long been a foundational method for the analysis of proteomic mixtures through the examination of peptide precursor ions. This approach makes LFQ highly compatible with AS-MS because it allows for the comparison of peptide ions without relying on sequence databases, stable isotope labeling, or chemical labeling.[21–23] The MS^1 spectra LFQ uses provide a larger dynamic range of ion detection, as opposed to quantification in tandem MS^2 spectra (e.g., tandem mass tags).[24] However, the quantitation capability of LFQ is strongly reliant on mass spectrometry resolution, with precise, high-resolution instruments demonstrating improved discernment between peptide features.[25,26] LFQ is also highly susceptible to variation in experimental conditions. Advances in computational analysis of mass spectrometry have become largely ubiquitous for LFQ, commonly seen in commercial and open-source software including MaxQuant,[22] Proteome Discoverer, and PEAKS Studio.[27–29] LFQ has thus been shown to increase data depth, sensitivity, and data completeness with applications in biomarker discovery, disease profiling, elucidation of drug mechanisms, and single-cell proteomics, underscoring its versatility and value in both basic and applied research.[21,30] Thus, LFQ is well-positioned to enhance the capabilities of AS-MS.

In this work, we demonstrate the integration of LFQ into AS-MS for the improved discovery of target-selective, high-affinity peptide ligands, named pyBinder. Data processing methods in AS-MS have primarily focused on filtering peptide sequencing data derived from MS^2 spectra, strongly increasing the dependence on mass spectrometry performance.[31] However, we seek to understand the target-selectivity of the ligand discovered, best done by comparing the MS^1 mass spectrometry data. Specifically, we use the result from LFQ of AS-MS samples to create two scores to understand the value of the peptides ligands discovered: i) target selectivity by comparing experimental samples (target versus off-target) and ii) concentration-dependent enrichment score (CDE), understood by comparing multiple samples from affinity selections completed at different target concentrations. We examine the ligands discovered from AS-MS against anti-hemagglutinin antibody (12ca5) and WD repeat-containing protein 5 (WDR5), which both have known high-affinity binding motifs that can validate the analysis method used. Overall, the outcome of pyBinder analysis demonstrates that 12ca5 and WDR5 motif-containing peptides are highly ranked by target selectivity and CDE versus other peptide features identified in the LFQ analysis. This result also enables targeted measurement of desired ions that show target selectivity and CDE. Thus, from discovery data, pyBinder appears poised to provide a variety of benefits for peptide drug discovery from AS-MS data ranging from minimizing the discovery of nonspecific ligands, structure activity relationships (SAR), to the estimation of binding affinity (K_D) direct from ligand discovery experiments.

4.2. Results and Discussion

Using peptide libraries, AS-MS performs an affinity selection against biomolecular targets and relies upon mass spectrometry to reveal the target-enriched peptide sequences. Thus, improvements to mass spectrometry protocols stand to improve AS-MS broadly. To understand AS-MS data, we considered what we term “sequencing coverage” and “sequencing fidelity.” Sequencing coverage is defined as the percentage of peptide precursor ions isolated for MS^2 fragmentation. Low sequencing coverage would indicate that the mass spectrometer was “overwhelmed” with peptides above its capabilities and/or the peptides were in low abundance necessitating long accumulation times. In comparison, high sequencing coverage would indicate that the spectrometer generated MS^2 spectra to most all peptides. Sequencing fidelity is defined as the percentage of MS^2 spectra that produce high-quality sequence assignment in its analysis. In our work, *de novo* sequencing analysis was performed in PEAKS Studio where an Average Local Confidence (ALC) of ≥ 80 was considered to be sufficient for high-quality sequence assignment.[27] Low sequence fidelity is generally due to poor or incomplete fragmentation patterns (i.e., b- and y-ions) due to low peptide abundance, co-isolation of multiple peptide precursors, poor fragmentation kinetics of the particular sequence (e.g. presence of C-terminal proline hindering nonspecific fragmentation[32–35]) and/or errant isolation of non-peptide library species by the mass spectrometer. Thus, by investigating the sequence coverage and sequence fidelity of our mass spectrometer, we can improve the data generated by AS-MS both in quantity and quality.

We performed retrospective analysis of a prior AS-MS discovery campaign to estimate sequence coverage and fidelity to be 10-18% and 0.2-1%, respectively (Figure 4.1), indicating a data incompleteness challenge in AS-MS currently. We reanalyzed the raw data from our previously published ligand discovery campaign of a natural 12-mer library against angiotensin-converting enzyme 2 (ACE2) with anti-hemagglutinin antibody 12ca5 used as a side-by-side off-target control.[14] Analysis of the raw data in PEAKS Studio enumerated the peptide features in the mass chromatogram (retention time versus mass-to-charge ratio, m/z), MS^2 spectra, and ALC of the sequence assignment per peptide feature. With over 30,000 peptide features, 3,468 (ACE2) and 5,895 (12ca5) MS^2 spectra were gathered, meaning the sequence coverage was low at 10.6% and 17.7% for ACE2 and 12ca5. Thus, a maximum rate of ~1.2 MS^2 spectra per second was observed. While modern mass spectrometers like the Orbitrap Fusion Lumos used here can perform faster, both higher-energy collisional dissociation (HCD) and electron-transfer dissociation (ETD) fragmentation methods were used and has been previously seen to improve the fidelity of *de novo* sequencing due to their orthogonality.[8,13,31] This sequencing coverage indicates that most peptides (> 80% of the ~33,000 peptides) were not isolated for MS^2 fragmentation by the mass spectrometer despite the use of a long 120-minute gradient. In addition, sequence fidelity was low at 0.24% and 1.1% for ACE2 and 12ca5, respectively, meaning that most all (> 95%) MS^2 mass spectra gathered did not produce a high-quality sequence assignment to the library used in the affinity selection experiment. This analysis clearly indicates that AS-MS samples are highly complex, and the mass spectrometer appeared “overwhelmed” with the number of peptides eluting given its throughput.



Average observed (n = 3):

	12ca5	**ACE2**
Peptide features	33,306*	32,722*
MS^2 scans of peptide features	17.7%	10.6%
Sequenced peptides with ALC > 80	**1.1%**†	**0.24%**†

* z=2-6 only, 67% of all features

† Follows library design

Figure 4.1: *Retrospective analysis of previous AS-MS campaigns reveals the opportunity for deeper data analysis by LFQ. (A) Total map 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 confidence sequence in red. High confidence sequences were defined by having an ALC score calculated by PEAKS Studio greater than 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 dearth 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.

To improve the mass spectrometry in AS-MS, MS-based proteomic methods were considered, first with the use of spectral database matching. The practice of spectral database matching is commonplace and would boost sequencing fidelity if applicable. This method matches MS^2 spectra with a database of peptide sequences that may be present in the sample, meaning peptide sequences can be confidently assigned to an incomplete MS^2 spectra. However, the use of spectral database matching appeared intractable for AS-MS using large (10^8) libraries for de novo discovery.[13] AS-MS libraries are prepared by split-and-pool synthesis to sample a vast theoretical sequence space.[36] For example, one common AS-MS library design uses an $X_{12}K$ design, where X is the 20 natural amino acids except cysteine (to exclude disulfide formation) and isoleucine (indistinguishable from leucine). For this $X_{12}K$ library, 10^8 beads are used in synthesis, resulting in a 10^8 peptide library; however, the theoretical sequence space is 10^{15} in total size.[13–16] Because the sampling of the 10^8 peptides from the 10^{15} is unbiased by design, database matching analysis of an $X_{12}K$ library would need to consider spectral matching against the full 10^{15} theoretical sequence space. This large number of sequences would result in a 15 PB/15000 TB FASTA file using a minimal UTF-8 encoding of each sequence and ignoring any additional sequence information commonly used in a FASTA file format, which is unable to be handled by most MS analysis software. Thus, database matching appeared intractable except for use with smaller, more-focused libraries (e.g., a 10^8 peptide library database using the same encoding method would be on the scale of 1-2 GB). For similar reasons due to the scale of the theoretical sequence space, DDA-based MS methods appear necessary, as DIA methods often rely upon spectral matching to improve the MS^2 deconvolution of co-isolated peptides.[37–39] Nevertheless, several strategies from MS-based proteomics appear compatible with AS-MS including (LFQ) as aforementioned were explored further.[21–23]

In an analysis method we call **pyBinder**, we combine LFQ with AS-MS data to understand the quality and value of the ligands discovered for their target-selectivity (Figure 4.2). While several standard proteomic software can accomplish LFQ analysis of mass spectrometry data,[22,27–29] we sought to develop an open-source approach in Python. Thus, the Python-based interface of OpenMS[29] (pyOpenMS[40]) was chosen to perform LFQ, with the ACE2/12ca5 AS-MS dataset for initial development. pyOpenMS was used to identify peptide features and prepare data for LFQ. Peptides were identified according to fitting to the Averagine isotopic distribution 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, until both showed comparable feature detection capability. Details of the parameter optimization are given in Table 4.2. 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 described in Lange et al[41]. The resulting aligned map was used to generate a consensus list of features across all proteins and replicates.

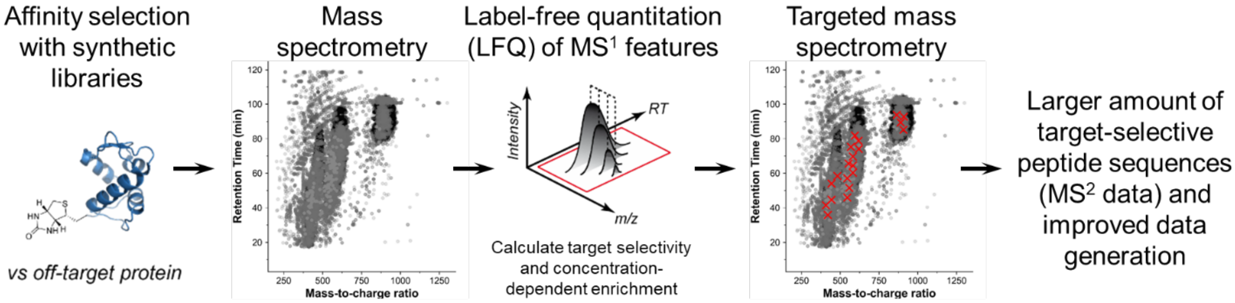

Affinity selection with synthetic libraries
vs off-target protein
Mass spectrometry
Retention Time (min)
120
100
80
60
40
20
250
500
750
1000
1250
Mass-to-charge ratio
Label-free quantitation (LFQ) of MS¹ features
Intensity
RT
m/z
Calculate target selectivity and concentration-dependent enrichment
Targeted mass spectrometry
Retention Time (min)
Mass-to-charge ratio
Larger amount of target-selective peptide sequences (MS² data) and improved data generation

Figure 4.2: *The combination of label-free quantitation (LFQ) and affinity selection-mass spectrometry (AS-MS) stands to provide an improved 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 a subsequent targeted mass spectrometry run to potentially reveal a larger amount of target-selective peptide ligands.*

To discern target-selectivity, pyBinder processes all peptide ion features discovered from AS-MS using extracted ion chromatograms (EICs), where an EIC shows all signal of a defined range of mass-to-charge ratio. In the EICs, the peptide ion features are highly unique given the high precision of the Orbitrap spectrometer utilized when combined with a specified retention time window. 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 remaining fast to compute.

From the integrated peaks areas, two scores were developed to rank and prioritize peptides for their value as ligands: target-selectivity and concentration-dependent enrichment (CDE, Figure 4.3). Target-selectivity of a ligand is a critical and elusive property at play in all ligand discovery platforms. While experimental controls and protocols are optimized, the discovery of nonselective or non-specific ligands plagues discovery efforts.[42,43] 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 4.3A, 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.

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 and specificity.[44,45] 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 4.3B). To enable this, 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 4.3B. 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 potential insight into ligand binding affinity (K_D), with theoretical scenarios given assumed K_D values shown in Figure 4.7. High CDE scores indicate strong peptide enrichment from the affinity selection due to the target protein. Meanwhile, low CDE scores (e.g., near zero) indicates 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 high-affinity, target-selective ligands. If known, the peptide sequences can provide insight into 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. Readdressing the ACE2/12ca5 AS-MS campaign, pyBinder revealed many peptide features that were target-specific, but not isolated for MS^2 sequencing, consistent with the low sequence coverage. Full results from pyBinder for both ACE2 and 12ca5 are given in Figures 4.8 and 4.9. The low overall sequence coverage of AS-MS samples left many potential ligands undiscovered, with >500 target-specific ligands to ACE2 and 12ca5 not isolated for MS^2 sequencing. These peptide features could be formulated into a targeted list and provide a strategy to overcome the stochastic nature of DDA-based tandem mass spectrometry of these complex AS-MS samples with improved sensitivity (Figure 4.2). A much larger amount of data could then be revealed, greatly improving the data generation capabilities of AS-MS as a ligand discovery platform.

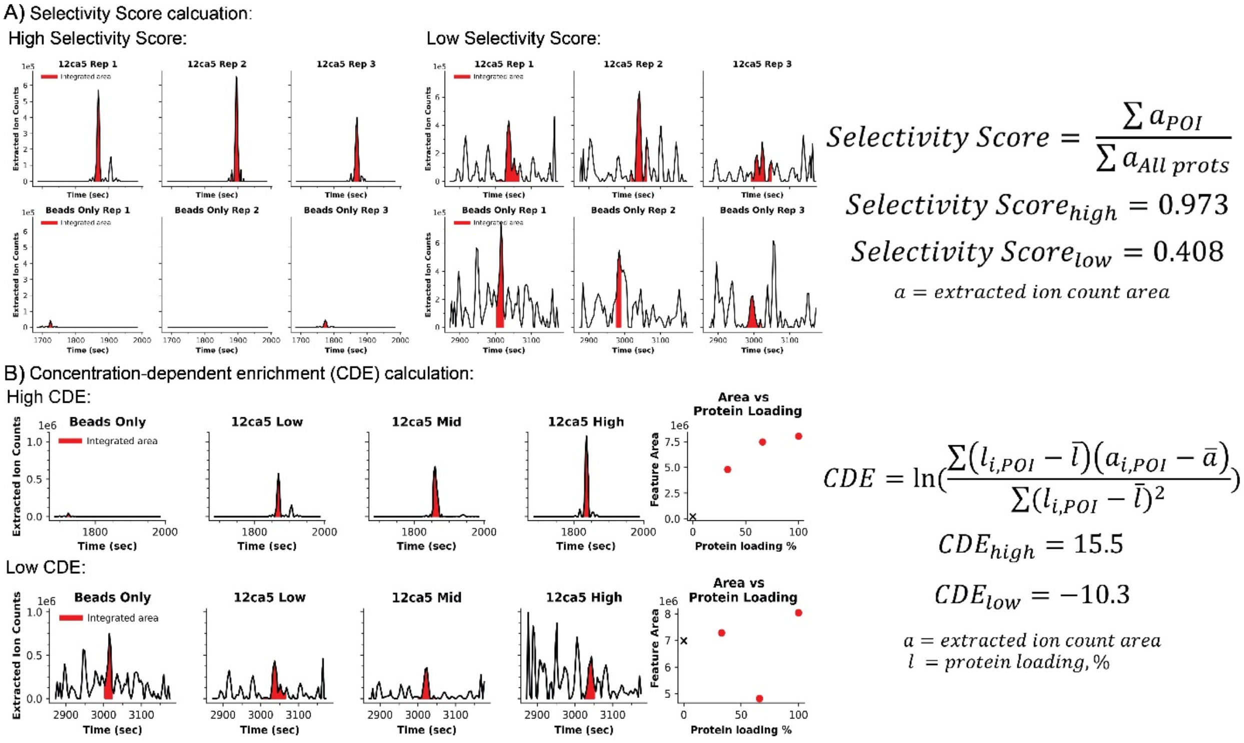


Figure 4.3: *Target selectivity and concentration-dependent enrichment (CDE) scores are used for the evaluation of the value of peptide features. (A) The selectivity score is calculated by comparing the area for a given feature with respect to a single protein and the total feature area measured across 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 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 by pyBinder of AS-MS data, an affinity selection was completed using 12ca5 compared against 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., $K_D < 200$ nM).[13,46] 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 4.4). Independently, the motif-containing peptides were color-coded and visualized for the target-selectivity and CDE (Figure 4.4A and 4.4B). While their statistical significance, denoted by -$\log_{10}$(P-value), was less discerning than the scores themselves, the target selectivity and CDE scores were clear to indicate the high performance of the motif-containing peptides in the affinity selection experiment. Also, as expected, many peptide features were not sequenced (shown in gray) due to the low sequence coverage or low sequencing fidelity. Last, combining the two scores (Figure 4.4C) 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 deeply analyze AS-MS data and distinguish ligands that are expected to be target-selective and high-affinity.

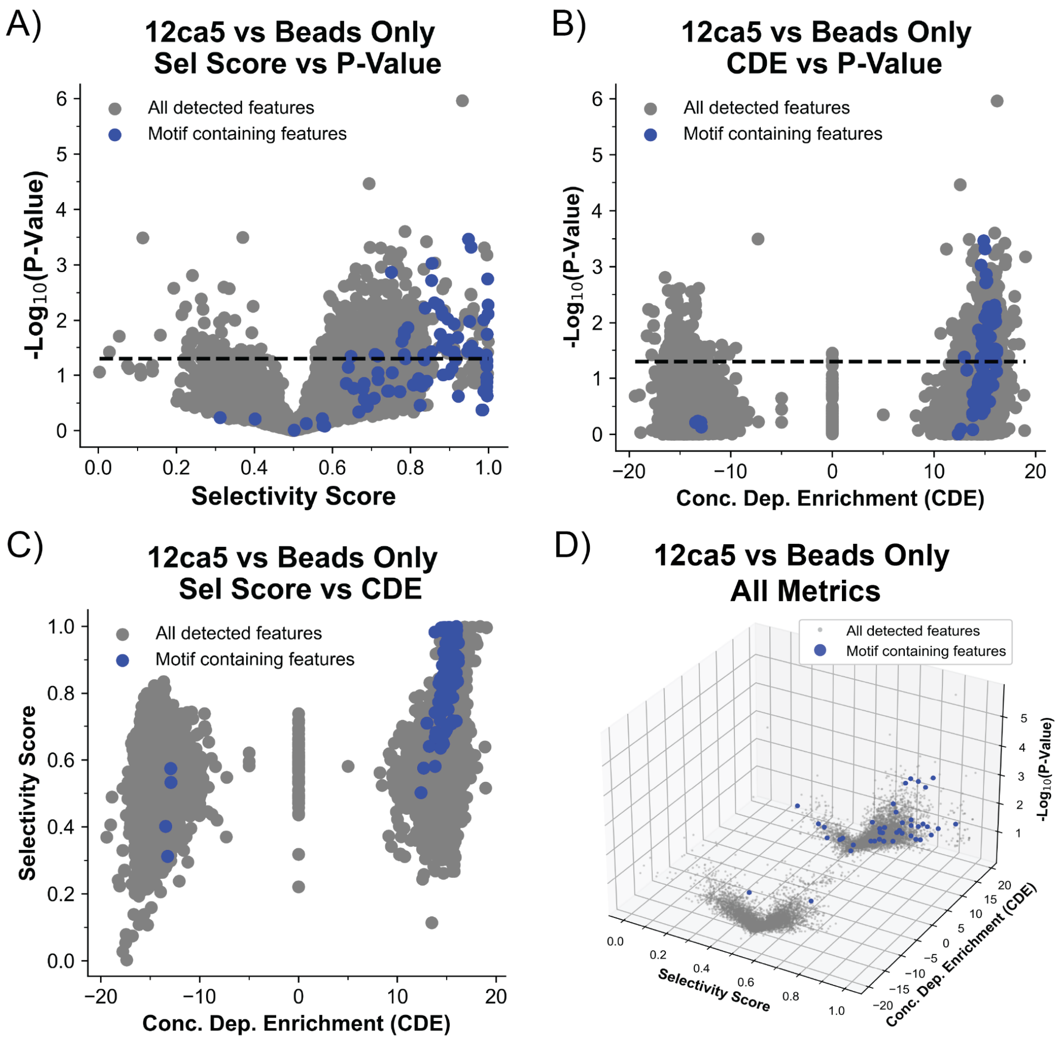


Figure 4.4: *The target selectivity and CDE scores of 12ca5 motif containing peptides demonstrate the ability of pyBinder to distinguish target-selective, high-affinity peptides due to the presence of their known motif. Motif-containing peptides are shown in blue in each graph, while all other detected features are shown in grey. (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 the proof-of-concept established with 12ca5, pyBinder demonstrated the ability to evaluate an AS-MS experiment for two target proteins, 12ca5 and WDR5, using a similar motif-based analysis for validation. WDR5, like 12ca5, also has a known set of sequence motifs that are common to several ligands and inhibitors to the WIN binding site based on arginine-containing tripeptide sequences (e.g., ART and ARA) at the N-terminus of the peptide.[47,48] 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 are plotted according to their selectivity scores, CDE scores, and p-values in Figure 4.5. For this case, the CDE score appeared 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 cutoffs ($p < 0.05$) appeared to hinder the prioritization of motif-containing peptides, consistent with the observations from the 12ca5 vs beads experiment in Figure 4.4A 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. Full pyBinder output is given in Figures 4.10 to 4.16.

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 4.5). 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 > 70%) from 3 to 14, demonstrating the application to of pyBinder to increase the data generated from AS-MS.

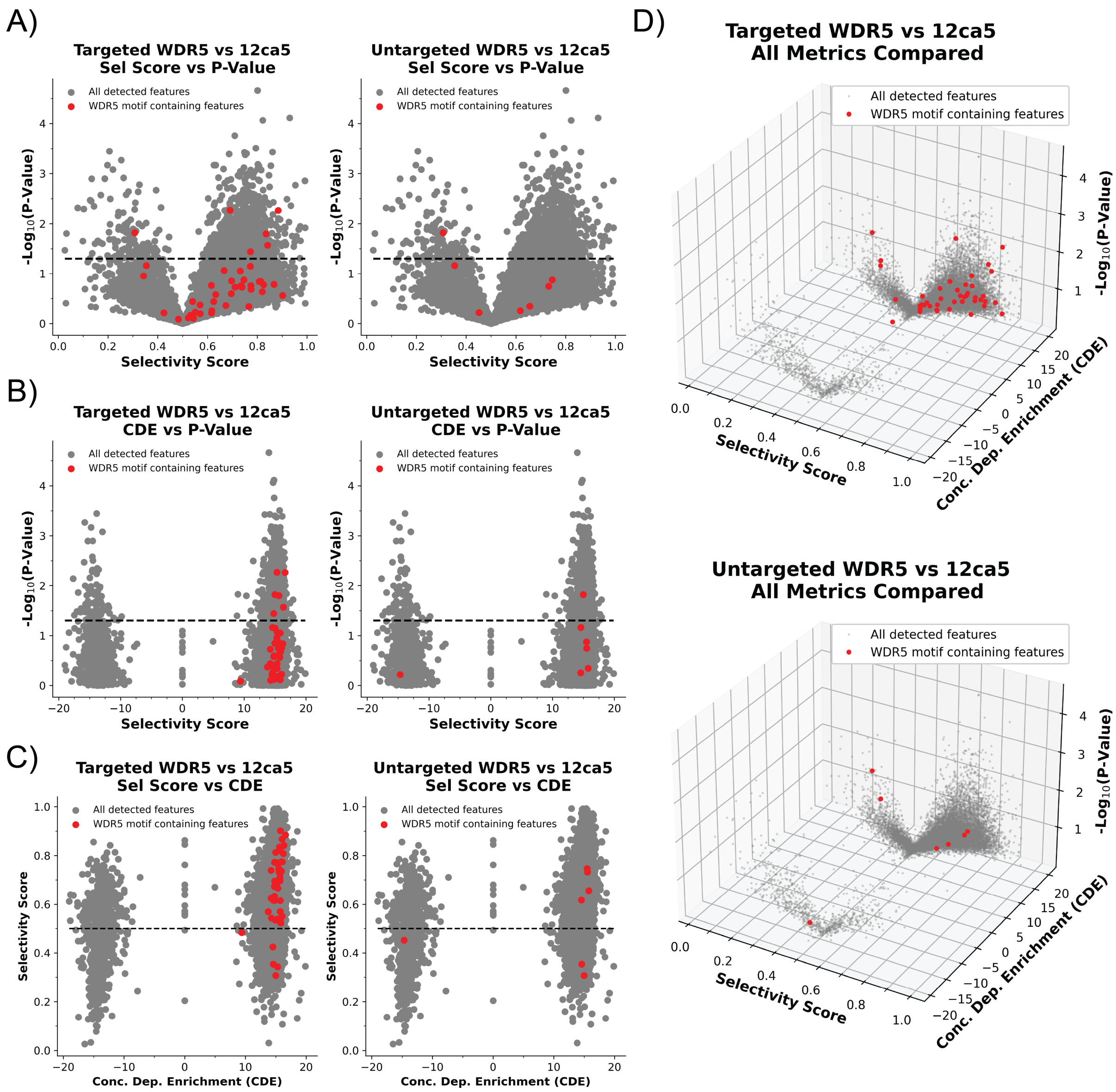


Figure 4.5: *The application of pyBinder in a second round of targeted mass spectrometry increases the discovery rate of peptides containing the WDR5 binding motifs compared to untargeted methods. Plots shown highlight WDR5 motif containing sequences that were successfully sequenced with high enough confidence, defined as an ALC score greater than or equal to 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.

4.3. Conclusion

We presented a workflow to perform LFQ on AS-MS data called pyBinder through the implementation of two scores of 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 high-affinity, target-selective ligands for two target proteins, 12ca5 and WDR5. Because they are connected to the ligand affinity, CDE scores are expected to be able to be combined with peptide sequence information in machine learning models discover and develop 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 could potentially 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 mass spectrometry 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 can enable the deliberate use of increased amounts of 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.

Overall, we expect this work to improve the robustness of AS-MS ranging from increasing the number of target-selective ligands discovered to the evaluation of affinity selection conditions and peptide libraries. We demonstrated in the ability of pyBinder to increase the amount of data generated from AS-MS experiments for the purpose of target-selective ligand discovery (Figure 4.5). pyBinder removes the reliance on sequencing results, which can be poor due to multiple reasons, and instead reports 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 significantly improve AS-MS for its ability to perform de novo ligand discovery and establishing structure activity relationships.

4.4. Materials

Canonical Fmoc-protected amino acids (FmocAla-OHxH2O, Fmoc-Arg(Pbf)-OH; Fmoc-Asn(Trt)-OH; Fmoc-Asp-(*O*-t-Bu)-OH; FmocCys(Trt)-OH; Fmoc-Gln(Trt)-OH; Fmoc-Glu(*O*-t-Bu)-OH; Fmoc-Gly-OH; Fmoc-His(Trt)- OH; Fmoc-Ile-OH; Fmoc-Leu-OH; Fmoc-Lys(Boc)-OH; Fmoc-Met-OH; Fmoc-Phe-OH; Fmoc- ProOH; Fmoc-Ser(But)-OH; Fmoc-Thr(t-Bu)-OH; Fmoc-Trp(Boc)-OH; Fmoc-Tyr(*O*-t-Bu)-OH; Fmoc-Val-OH) were purchased from Sigma Millipore (Novabiochem) and used as received. Fmoc-Lys(biotin)-OH was purchased from Sigma Millipore (Novabiochem) and used as received. Fmoc-L-His(Boc)-OH was purchased from Advanced ChemTech and used as received. *O*-(7-azabenzotriazol-1-yl)-*N,N,N',N'*-tetramethyluronium hexafluoro-phosphate (HATU, ≥97.0%) and (7-azabenzotriazol-1-yloxy)tripyrrolidinophosphonium hexafluorophosphate (PyAOP, ≥97.0%) were purchased from P3 Biosystems. Fmoc-Rink amide linker (4-[(R,S)-(2,4-dimethoxyphenyl)(Fmoc-amino)methyl]phenoxyacetic acid) was purchased from Chem Impex Inc (Wood Dale, IL) and used as received.

Biosynthesis OmniSolv® grade N,N-dimethylformamide (DMF) was purchased from EMD Millipore (DX1732-1) and incubated with 1 pack of AldraAmine trapping agents (for 1000 – 4000 mL DMF, Sigma-Aldrich, catalog number Z511706) for 48 hours prior to use. was purchased from Sigma-Aldrich. Diisopropylethylamine (DIEA; 99.5%, biotech grade, catalog number 387649) and piperidine (ACS reagent, ≥99.0%) were purchased from Sigma-Aldrich. Formic acid (FA, 97%) was purchased from Beantown Chemical, Corp. Reaction vessels were purchased from Torviq equipped with a polypropylene frit. To each vessel was added a disc of Porex filter paper (0.025" thick, 7-12 micron) from Interstate Specialty Products. Dichloromethane (DCM; ≥99.8%, HPLC grade, contains amylene as stabilizer, catalog number 34856), trifluoroacetic acid (HPLC grade, ≥99.0%), triisopropylsilane (98%, catalog number 233781), diethyl ether (anhydrous, ACS reagent, ≥99.0%), acetonitrile (HPLC grade, ≥99.9%), Omnisolv® acetonitrile (LC-MS grade, AX0156-1), and Omnisolv® water (LC-MS grade, WX0001-1) were purchased from Sigma-Aldrich. Methanol was purchased from Millipore Sigma. Formic acid Optima LC/MS (A117) was purchased from Fisher Chemical. Water was deionized using a Milli-Q Reference water purification system (Millipore). Nylon 0.22 μm syringe filters were TISCH brand SPEC17984.

20 μm TentaGel® M NH_2 Monosized Amino Microsphere resin was purchased from Rapp Polymere Inc. (Tübingen, Germany). Nestle Carnation instant nonfat dry milk (Code 12428935) was purchased from Nestle Professional (Solon, OH). Dynabeads MyOne Streptavidin T1 magnetic microparticles were purchased from Invitrogen (Carlsbad, CA). Phosphate buffered saline (10x, Molecular biology grade) was purchased from Corning. Sodium chloride (ACS grade) was purchased from Avantor. Guanidine hydrochloride (Cat BP178) and sodium phosphate monobasic monohydrate were purchased from Fisher Scientific.

Mouse anti-hemagglutinin antibody (clone 12ca5) was purchased from Columbia Biosciences Corporation (Cat: 00-1722, Frederick, Maryland) biotin-$(PEG)_4$-NHS ester and biotin-$(PEG)_4$-propionic acid were purchased from ChemPep

Inc. (Wellington, FL). Biotinylation of 12ca5 was performed as previously described.[13] WD repeat-containing protein 5 WDR5 was supplied by Civetta Therapeutics (Cambridge, MA).

4.5. Methods

4.5.1. Split-and-pool synthesis of linear $X_{12}K$ peptide library

Synthesis of peptide libraries was performed using 20 µm Tentagel M NH_2 resin (0.31 mmol/g) for a total of 2.4×10^9 sequences split into aliquots of 2×10^8 sequences. The resin was suspended in DMF and dividedly evenly between 18 syringes (all canonical amino acids except for cysteine and isoleucine) for variable regions. Couplings were performed using the Fmoc-protected amino acid dissolved in DMF (10 eq, 0.40M) with PyAOP (0.9 eq relative to amino acid, 0.38M) activated with DIEA (1.1 eq relative to amino acid for histidine, 3 eq relative to amino acid for all others). Couplings were incubated for 1 hour. The resin was then recombined and washed with DMF, DCM, and DMF. Fmoc deprotection was performed using 20% piperidine in DMF (1x flow wash, 3 x 5 min batch treatments). The resin was washed again with DMF, DCM, then DMF before being subjected to another split-couple-pool cycle until completion of all randomized positions.

4.5.2. Peptide cleavage and global deprotection

Cleavage from solid phase and global deprotection was performed using a solution of 95% trifluoroacetic acid, 2.5% water, and 2.5% triisopropylsilane (~20 mL cleavage cocktail / g of resin). The solution was added until the resin was fully swelled and free flowing, then the resin was agitated on a nutating mixer for 3 hours. The peptides were triturated with 10:1 cold diethyl ether to cleavage solution. The precipitated solid was centrifuged into a pellet. The precipitate was washed with cold ethyl ether in the same manner an additional two times. The resulting solid pellet was dried gently using N_2, suspended in 50% acetonitrile in water (0.1% trifluoroacetic acid), and lyophilized.

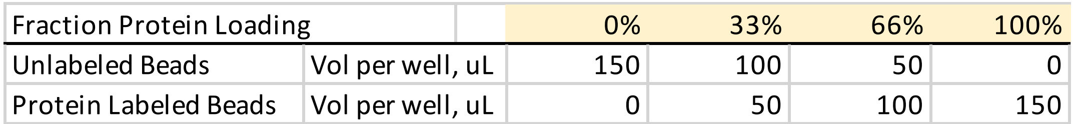
4.5.3. Solid-phase extraction of peptide library

Peptides were adjusted to 5% acetonitrile in aqueous media (0.1% TFA) and purified using Supelclean™ LC-18 SPE Tube, bed wt. 1 g (Millipore Sigma Cat: 505471). The SPE tube was first conditioned with 3 CV of acetonitrile (0.1% TFA) and then equilibrated with 5 CV of 5% acetonitrile in water (0.1% TFA). Then, the suspended crude was loaded (Approximately 50 mg peptide loaded onto 1 g bed mass) and washed with 10 CV of 5% acetonitrile in water (0.1% TFA). Peptides were eluted with 70% acetonitrile (0.1% TFA, 1 CV) and lyophilized.

4.5.4. Affinity selection

Dynabeads MyOne Streptavidin T1 magnetic microparticles (3 mg, 300 µL per replicate) were aliquoted and washed three times with Wash Buffer composed of 1x PBS, 2% nonfat dry milk (NFDM) and 0.01% Tween20. 100 µL per protein replicate of washed beads were aliquoted and incubated with biotinylated protein (1.2 eq) for 1 hour at 4 °C with agitation. At the same time, the peptide library dissolved in 1x PBS was combined with prewashed beads (150 µL per replicate) and supplemented with 10% NFDM in 1x PBS to a final concentration of 2% NFDM. The library mixture was then incubated for 1 hour at 4 °C with agitation. The beads were removed from the library mixture via magnetic rack and aliquoted to a 96 Deepwell plate as shown below. Protein labelled beads were washed three times with Wash Buffer and aliquoted into a 96 Deepwell plate as shown below in fractions based on the desired protein loading per replicate. For lower protein concentrations, additional unlabeled, prewashed beads were added to keep a constant total amount of beads used per sample. An example setup for variable protein concentration is shown below in Table 4.1.

Table 4.1: Example setup for an affinity selection utilizing variable protein concentrations using 150 uL of beads total per replicate



Fraction Protein Loading		0%	33%	66%	100%
Unlabeled Beads	Vol per well, uL	150	100	50	0
Protein Labeled Beads	Vol per well, uL	0	50	100	150

Affinity selections were then performed using a KingFisher^{TM} Duo Prime Purification System in 96 Deepwell Plates (Thermo Fisher Scientific, cat. #95040450) with the following setup:

Plate 1		
A	10 pM/member peptide library diluted into 1x PBS, 2% milk	0.5 mL
B	1.5 mg of magnetic beads (150 µL) diluted in Wash buffer (1x PBS, 2% NFDM, 0.01% Tween20)	1 mL
C	Wash buffer (1x PBS, 2% NFDM, 0.01% Tween20)	1 mL
D	Reserved for 12-tip Deepwell magnetic comb (Thermo Fisher, cat. #97003500)	1 mL

Plate 2		
A	1x PBS at 4 ^{o}C	1 mL
B	1x PBS at 4 ^{o}C	1 mL
C	1x PBS at 4 ^{o}C	1 mL
D	1x PBS at 4 ^{o}C	1 mL

The program performed the following protocol:

1. Collect comb from Plate 1 Row D
2. Collect beads from Plate 1 Row B and wash for 30 sec at low
3. Wash beads for 30 sec in Plate 1 Row C
4. Incubate immobilized protein for 1 h at 10 °C with slow mixing in Plate 1 Row A
5. Wash immobilized protein for 2 mins each at low speed in Plate 2 Rows A through D

6. Elute protein by mixing for 1 min at fast speed in Elution Strips 1 and 2 containing 100 µL of 6M guanidinium chloride in 50 mM sodium phosphate at pH 7.4.

After affinity selection, samples were purified by STAGE Tip preparation, split 40:60 for initial analysis and targeted analysis separately, and dried using a vacuum centrifuge. Dried samples for the initial scouting run were reconstituted into 16 µL of nLC-MS/MS mobile phase A and 1.778 µL of Pierce Retention Time Calibration Mixture (Thermo Fisher, catalog number 88321). Samples were centrifuged at 21.3k rcf for 10 minutes at 4 ºC. 4.5 µL were injected per sample for nLC-MS/MS analysis. Dried samples for targeted analysis were reconstituted into 24 µL of nLC-MS/MS mobile phase A and centrifuged at 21.3k rcf for 10 minutes at 4 ºC. 4.5 µL were injected per sample for nLC-MS/MS analysis.

4.5.5. Nano-liquid chromatography-tandem mass spectrometry (nLC-MS/MS) analysis

Peptide sequencing was performed on an EASY-nLC 1200 (Thermo Fisher Scientific) nano-liquid chromatography system with an Orbitrap Fusion Lumos Tribrid Mass Spectrometer (Thermo Fisher Scientific). Samples were run on a PepMap RSLC C18 column (2 µm particle size, 25 cm x 75 µm ID; Thermo Fisher Scientific, cat. #ES902) with a nanoViper Trap Column (C18, 3 µm particle size, 100 A pore size, 20 mm x 75 µm ID; Thermo Fisher Scientific, cat. #164946) for desalting. Mobile phase A = water (0.1% FA) and mobile phase B = 80% AcN in water (0.1% FA).

The ion source voltage was set to 2200 volts in positive mode. Primary mass spectra were detected using the orbitrap at 120000 resolution with a scan range of 300-1400 (m/z), RF lens of 30%, a normalized AGC target of 250% with automatic injection time, and 1 microscan. Candidate ions were chosen for tandem mass spectrometry based on the following criteria: precursor mass range of 300-1200 (m/z), monoisotopic peak determination set to peptides, minimum intensity threshold of 4e4, charge states ranging from +2-+6, dynamic exclusion after 1 observation for 30 seconds with a ±10 ppm range. Fragmentation was done in the orbitrap using HCD followed by EThcD activation types with the following settings: 1.3 m/z isolation window, 0.3 m/z offset, 30000 resolution, defined first mass of 120 m/z, 300% normalized AGC target with 100 ms maximum injection time, 2 microscans in centroid mode. HCD mode used 25% HCD collision energy and EThcD used 50% SA collision energy for $z = 2$ ions, 40% SA collision energy for $z = 3$ ions, and 35% SA collision energy for $z = 4$ to $z = 6$ ions. For targeted runs, a list of m/z values were supplied for each protein with start times 20 minutes before the reported retention time and stop times 20 minutes after the reported retention time with a tolerance of ±0.02 m/z. Full cycle time for MS^1 and MS^2 scans was 3 seconds.

The following gradient was used: linear gradient 1-45% B from 0-120 min; linear gradient 45-90% B from 120-123 min; isocratic 90% B from 123-126 min; linear gradient 90-20% B from 126-129 min; isocratic 20% B from 129-132 min; linear gradient 20-90% B from 132-135 min; isocratic 90% B from 135-138 min; linear gradient 90-20% B from 138-141 min; isocratic 20% B from 141-144 min; linear gradient 20-90% B from 144-147 min; isocratic 90% B from 147-152 min.

Pre-column and analytical column were equilibrated before each run with 8 µL and 12 µL of mobile phase A respectively xbefore sample injection. Samples were loaded using 12 µL of mobile phase A. Mass data was recorded from 3-120 min.

4.5.6. Analysis of AS-MS data using pyBinder

RAW files of the initial runs were converted to mzML file format using MSConvert from the ProteoWizard toolkit. Only MS^1 spectra were included in the conversion. A full guide for the inputs is given with the pyBinder source code. Briefly, the names of the proteins and concentrations used are input, as well as user-determined limits for selectivity scoring and statistical confidence (default $\alpha = 0.05$) along with parameters for peak detection. All inputs are checked for validity and output directories are created in the user-specified locations. mzML files are then opened using OpenMS (version 3.0.0) and centroided using the PeakPickerHiRes class with the default parameters. Plots comparing the profile and centroided versions are generated for inspection.

Following data centroiding, feature maps are generated for each mzML file using the FeatureFinder class with the following parameters: isotopic_pattern:mz_tolerance = 0.01, isotopic_pattern:charge_low = 2, isotopic_pattern:charge_high = 5, feature:max_rt_span = 3, mass_trace:min_spectra = 9, feature:rt_shape = asymmetric, seed:min_score = 0.5, feature:min_score = 0.5, mass_trace:max_missing = 4. All other parameters were used at their default value. All feature maps were exported to featureXML file format.

The feature maps were then aligned in retention time with the feature map with the greatest number of features as a reference. The MapAlignmentAlgorithmPoseClustering class was used with the following parameters: superimposer:mz_pair_max_distance = 0.5, pairfinder:distance_RT:max_difference = 300, superimposer:max_shift = 2000. All other parameters were used at their default value. Original retention times for each feature map were stored separately. Plots showing the retention times before and after alignment were generated. The aligned feature maps were then grouped into a consensus feature map using the FeatureGroupingAlgorithmQT class with the following parameters: distance_MZ:max_difference = 0.01, distance_RT:max_difference = 150. All other parameters were used at their default value. Retention time and mass-to-charge ratios were then extracted from the consensus map for use in further analysis, and the consensus map was also exported to a consensusXML file.

A mass filter is applied to the list of all features, defined as having a minimum mass of a sequence comprised of only glycine and a maximum mass of a sequence comprised of only tryptophan. Next, MS^1 data is read for all files and extracted ion counts (EICs) are taken for each peptide feature in each file, defined with mass-to-charge tolerance of 0.01 and a retention time window as specified by the user (default 2 minutes). The EIC signal is smoothed using a Savitsky-Golay noise filter from the SciPy module (version 1.11.1) with a window length of 19 and a polynomial order of 9. Peaks are then detected using the PeakUtils module (version 1.3.4) with thresholds defined by the maximum observed signal per EIC. The largest peak for each EIC is stored and integrated using the cumulative_trapezoid method from the SciPy module, and if a peak is not detected, a placeholder value near the limit of detection for the instrument is used.

Selectivity scores and feature p-values were calculated using the areas of the highest concentration of protein, where selectivity scores are calculated as shown in Figure 4.3. Welch's ANOVA test is used to determine statistical significance for more than two proteins, and a homoscedasticity test is performed to check the variances of each group. If the variances are equal, Tukey's test is then used to calculate p-values across pairs of proteins; if the variances are not equal, a Games-Howell post hoc test is used. If the p-value falls below the specified threshold, the protein areas that are statistically significant are labelled with the relevant protein.

Concentration-dependent enrichment scoring is performed using the different concentration levels of protein used, as well as the control run using unlabeled beads. The percentage of protein loading on the bead and corresponding areas are used to calculate the CDE score as shown in Figure 4.3. The results from both scores are then filtered as desired and exported into an inclusion list that can be exported directly into the Thermo Xcalibur Method Editor (Version 4.2.47) or into an Excel spreadsheet that displays the EICs and CDE score calculations for a specified number of top candidates.

4.6. Code 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.

4.7. Appendix: Full pyBinder Output

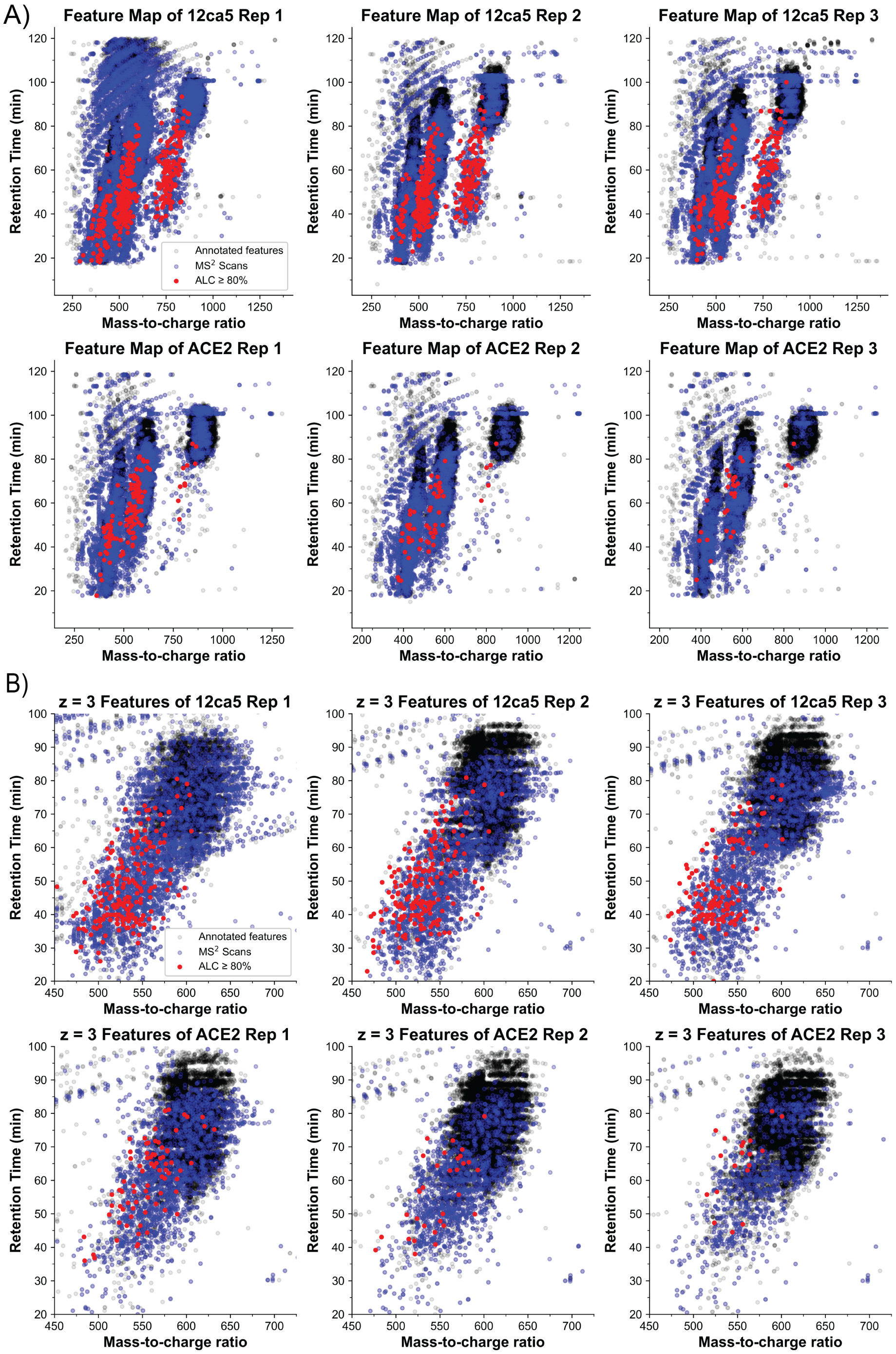
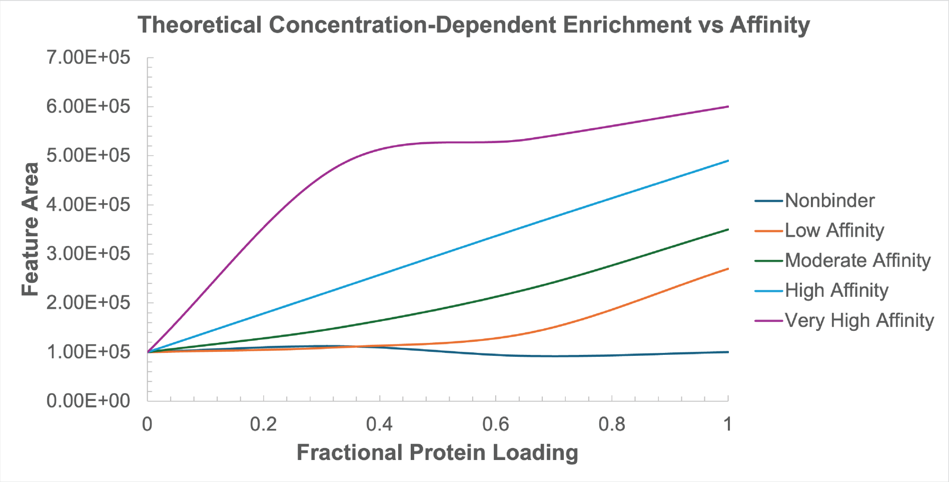


Figure 4.6: *Retrospective analysis of previously published AS-MS experiments reveals extent of "missing values" problem. Although high affinity ligands were discovered from these experiments, there is still a large volume of peptide features that are not analyzed during an untargeted run, leaving many potential ligands unobserved.*

Table 4.2: *Parameter optimization for the feature finding process in OpenMS. First round results were judged based on the percentage of features in common with the output from PEAKS 8.5 and balanced around the total number of annotated features, while second round results were judged based on the percentage of previously identified ligands to 12ca5 and ACE2 present in the OpenMS output.*

Feature finder parameter testing:																						From 12ca5	From ACE2	
Entry	Intensity: bins	mass_trace: mz_tolerance	mass_trace:min_spectra	mass_trace: max_missing	mass_trace: slope_bound	isotopic_pattern:charge_low	isotopic_pattern:charge_high	isotopic_pattern:mz_tolerance	isotopic_pattern: intensity_percentage	isotopic_pattern: intensity_precentage_optional	isotopic_pattern: optional_fit_improvement	isotopic_pattern: mass_window_width	seed: min_score	fit: max_iterations	feature: min_score	feature: min_isotope_fit	feature: min_trace_score	feature: min_rt_span	feature: max_rt_span	feature: rt_shape	feature: max_intersection	% ID from set	% ID from set	Total Features
First Round																								
1 (default)	10	0.02	5	1	0.5	2	5	0.04	10	0.1	2	25	0.8	500	0.5	0.5	0.5	0.333	2.5	asymmetric	0.7	97.45	100	351135
	10	0.02	5	1	0.5	2	5	0.04	0.1	0.1	2	25	0.8	500	0.5	0.5	0.5	0.333	2.5	asymmetric	0.7	0	0	0
	10	0.02	5	1	0.5	2	5	0.04	1	0.1	2	25	0.8	500	0.5	0.5	0.5	0.333	2.5	asymmetric	0.7	80.04	97.96	211364
	10	0.02	5	1	0.5	2	5	0.04	5	0.1	2	25	0.8	500	0.5	0.5	0.5	0.333	2.5	asymmetric	0.7	92.99	97.96	283895
	10	0.02	5	1	0.5	2	5	0.04	10	0	2	25	0.8	500	0.5	0.5	0.5	0.333	2.5	asymmetric	0.7	97.45	100	350761
	10	0.02	5	1	0.5	2	5	0.04	10	0.5	2	25	0.8	500	0.5	0.5	0.5	0.333	2.5	asymmetric	0.7	97.45	100	350388
	10	0.02	5	1	0.5	2	5	0.04	10	1	2	25	0.8	500	0.5	0.5	0.5	0.333	2.5	asymmetric	0.7	97.45	100	349979
	10	0.02	5	1	0.5	2	5	0.04	10	5	2	25	0.8	500	0.5	0.5	0.5	0.333	2.5	asymmetric	0.7	97.45	100	343139
	10	0.02	5	1	0.5	2	5	0.04	10	10	2	25	0.8	500	0.5	0.5	0.5	0.333	2.5	asymmetric	0.7	97.66	100	345999
	10	0.02	5	1	0.5	2	5	0.04	10	50	2	25	0.8	500	0.5	0.5	0.5	0.333	2.5	asymmetric	0.7	0	0	0
	10	0.02	5	1	0.5	2	5	0.04	10	0.1	2	2	0.8	500	0.5	0.5	0.5	0.333	2.5	asymmetric	0.7	97.45	100	350843
	10	0.02	5	1	0.5	2	5	0.04	10	0.1	2	5	0.8	500	0.5	0.5	0.5	0.333	2.5	asymmetric	0.7	97.45	100	350744
	10	0.02	5	1	0.5	2	5	0.04	10	0.1	2	10	0.8	500	0.5	0.5	0.5	0.333	2.5	asymmetric	0.7	97.45	100	351135
	5	0.02	5	1	0.5	2	5	0.04	10	0.1	2	25	0.8	500	0.5	0.5	0.5	0.333	2.5	asymmetric	0.7	97.24	100	349784
	20	0.02	5	1	0.5	2	5	0.04	10	0.1	2	25	0.8	500	0.5	0.5	0.5	0.333	2.5	asymmetric	0.7	97.45	100	351953
	10	0.02	5	2	0.5	2	5	0.04	10	0.1	2	25	0.8	500	0.5	0.5	0.5	0.333	2.5	asymmetric	0.7	97.66	100	354623
	10	0.02	5	1	0.5	2	5	0.04	10	0.1	5	25	0.8	500	0.5	0.5	0.5	0.333	2.5	asymmetric	0.7	97.45	100	349165
	10	0.02	5	1	0.5	2	5	0.04	10	0.1	10	25	0.8	500	0.5	0.5	0.5	0.333	2.5	asymmetric	0.7	97.24	100	346309
	10	0.02	5	1	1	2	5	0.04	10	0.1	2	25	0.8	500	0.5	0.5	0.5	0.333	2.5	asymmetric	0.7	96.82	97.96	145613
	10	0.02	5	1	2	2	5	0.04	10	0.1	2	25	0.8	500	0.5	0.5	0.5	0.333	2.5	asymmetric	0.7	97.45	95.92	79207
	10	0.02	5	1	5	2	5	0.04	10	0.1	2	25	0.8	500	0.5	0.5	0.5	0.333	2.5	asymmetric	0.7	96.82	97.96	65395
	10	0.02	5	1	10	2	5	0.04	10	0.1	2	25	0.8	500	0.5	0.5	0.5	0.333	2.5	asymmetric	0.7	96.6	97.96	64481
	10	0.02	5	1	1000	2	5	0.04	10	0.1	2	25	0.8	500	0.5	0.5	0.5	0.333	2.5	asymmetric	0.7	96.6	97.96	63997
	10	0.02	5	1	0.5	2	5	0.04	0.1	0.1	2	25	0.5	500	0.5	0.5	0.5	0.333	2.5	asymmetric	0.7	0	0	0
	10	0.02	5	1	0.5	2	5	0.04	15	0.1	2	25	0.5	500	0.5	0.5	0.5	0.333	2.5	asymmetric	0.7	98.3	100	373446
	10	0.02	5	1	0.5	2	5	0.04	20	0.1	2	25	0.5	500	0.5	0.5	0.5	0.333	2.5	asymmetric	0.7	98.09	100	390483
	10	0.02	5	1	0.5	2	5	0.04	25	0.1	2	25	0.5	500	0.5	0.5	0.5	0.333	2.5	asymmetric	0.7	98.09	100	419866
	10	0.02	5	1	0.5	2	5	0.04	10	20	2	25	0.5	500	0.5	0.5	0.5	0.333	2.5	asymmetric	0.7	99.15	100	446136
	10	0.02	5	1	0.5	2	5	0.04	10	30	2	25	0.5	500	0.5	0.5	0.5	0.333	2.5	asymmetric	0.7	98.51	100	375865
	10	0.02	5	1	0.5	2	5	0.04	10	40	2	25	0.5	500	0.5	0.5	0.5	0.333	2.5	asymmetric	0.7	0	0	0
	10	0.02	5	1	0.5	2	5	0.04	10	50	2	25	0.5	500	0.5	0.5	0.5	0.333	2.5	asymmetric	0.7	0	0	0
	10	0.02	5	1	0.5	2	5	0.04	10	0.1	1	25	0.5	500	0.5	0.5	0.5	0.333	2.5	asymmetric	0.7	97.45	100	352011
	10	0.02	5	1	0.5	2	5	0.04	10	0.1	2	25	0.5	500	0.5	0.3	0.5	0.333	2.5	asymmetric	0.7	97.45	100	367276
	10	0.02	5	1	0.5	2	5	0.04	10	0.1	2	25	0.5	500	0.5	0.4	0.5	0.333	2.5	asymmetric	0.7	97.45	100	361250
	10	0.02	5	1	0.5	2	5	0.04	10	0.1	2	25	0.5	500	0.5	0.6	0.5	0.333	2.5	asymmetric	0.7	97.88	100	334588
	10	0.02	5	1	0.5	2	5	0.04	10	0.1	2	25	0.5	500	0.5	0.7	0.5	0.333	2.5	asymmetric	0.7	97.45	100	351135
	10	0.02	5	3	0.5	2	5	0.04	10	0.1	2	25	0.5	500	0.5	0.5	0.5	0.333	2.5	asymmetric	0.7	97.88	100	356447
	10	0.02	5	4	0.5	2	5	0.04	10	0.1	2	25	0.5	500	0.5	0.5	0.5	0.333	2.5	asymmetric	0.7	97.88	100	357762
	10	0.02	5		0.25	2	5	0.04	10	0.1	2	25	0.5	500	0.5		0.5	0.333	2.5	asymmetric	0.7	98.51	100	621603
DONE ON DIFFERENT SET - THIS ONE COMPARED TO PEAKS FEATURES, NOT LIST OF D**DYAS AND ABPS																								
Second Round - matched to PEAKS Output																								
	10	0.004	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	0.35	25.32	22.7	5695
	10	0.004	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	0.7	25.47	22.73	5815
	10	0.004	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	0.5	25.39	22.74	5773
	10	0.004	10	2	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	0.35	30.36	28.12	7047
	10	0.004	10	3	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	0.35	35.51	33.36	8523
	10	0.004	10	4	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	0.35	41.2	39.81	9776
	10	0.004	10	5	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	0.35	46.23	45.37	10827
	10	0.004	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.5	0.5	0.333	2.5	symmetric	0.35	24.93	22.29	5373
	10	0.004	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.6	0.5	0.333	2.5	symmetric	0.35	25.31	22.8	5459
	10	0.004	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.7	0.5	0.333	2.5	symmetric	0.35	28.55	22.93	5583
	10	0.004	5	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	0.35	80.56	82.1	50537
	10	0.004	6	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	0.35	65.64	66.9	21188
	10	0.004	7	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	0.35	65.64	66.9	21188
	10	0.004	8	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	0.35	42.23	41.03	10425
	10	0.004	9	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	0.35	42.23	41.03	10425
	10	0.004	10	1	0.1	2	5	0.005	10	0.1	2	25	0.5	500	0.7	0.8	0.5	0.333	2.5	symmetric	0.35	44.83	44	9108
	10	0.004	10	1	0.1	2	5	0.005	10	0.1	2	25	0.6	500	0.7	0.8	0.5	0.333	2.5	symmetric	0.35	43.67	42.7	8779
	10	0.004	10	1	0.1	2	5	0.005	10	0.1	2	25	0.7	500	0.7	0.8	0.5	0.333	2.5	symmetric	0.35	38.34	37.26	7486
	10	0.004	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.5	0.8	0.5	0.333	2.5	symmetric	0.35	39.23	38.02	7522
	10	0.004	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.6	0.8	0.5	0.333	2.5	symmetric	0.35	35.58	34.04	7166
	10	0.004	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	asymmetric	0.35	27.06	24.89	5992
	5	0.004	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric		24.87	22.17	5366
	15	0.004	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric		25.05	22.31	5415
	20	0.004	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric		25.04	22.28	5415
	10	0.004	10	1	0.1	2	5	0.005	0.1	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric		5.02	5.16	896
	10	0.004	10	1	0.1	2	5	0.005	1	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric		9.57	8.75	1993
	10	0.004	10	1	0.1	2	5	0.005	5	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric		17.92	15.62	3944
	10	0.004	10	1	0.1	2	5	0.005	15	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric		25.23	22.51	5702
	10	0.004	10	1	0.1	2	5	0.005	20	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric		26.28	23.35	5985
	10	0.004	10	1	0.1	2	5	0.005	25	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric		26.49	23.65	6082
	10	0.004	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric		25.01	22.27	5405
	10	0.004	10	1	0.25	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric		27.97	24.96	6340
	10	0.004	10	1	1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric		19.84	16.34	4986
	10	0.004	10	1	2	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric		19.35	15.84	4914
	10	0.004	10	1	5	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric		19.35	15.84	4903
	10	0.004	10	1	10	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric		19.35	15.84	4902
	10	0.004	10	1	0.1	2	5	0.005	10	0	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	0.35	25.22	22.46	5388
	10	0.004	10	1	0.1	2	5	0.005	10	0.5	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	1.35	24.84	22.11	5406
	10	0.004	10	1	0.1	2	5	0.005	10	1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	2.35	24.89	22.13	5406
	10	0.004	10	1	0.1	2	5	0.005	10	5	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	3.35	25.06	22.51	5460
	10	0.004	10	1	0.1	2	5	0.005	10	10	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	4.35	24.8	21.92	5438
	10	0.004	10	1	0.1	2	5	0.005	10	20	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	5.35	23.54	19.89	6186
	10	0.004	10	1	0.1	2	5	0.005	10	50	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	6.35	0	0	0
	10	0.004	10	1	0.1	2	5	0.005	10	0.1	1	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	7.35	24.95	22.15	5421
	10	0.004	10	1	0.1	2	5	0.005	10	0.1	5	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	8.35	24.42	21.59	5355
	10	0.004	10	1	0.1	2	5	0.005	10	0.1	10	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	9.35	24.09	21.33	5316
	10	0.004	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	400	0.7	0.8	0.5	0.333	2.5	symmetric	10.35	25.01	22.27	5403
	10	0.004	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	600	0.7	0.8	0.5	0.333	2.5	symmetric	11.35	25.01	25.77	5400
	10	0.004	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	800	0.7	0.8	0.5	0.333	2.5	symmetric	12.35	25.01	22.27	5406
	10	0.004	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	1000	0.7	0.8	0.5	0.333	2.5	symmetric	13.35	25.03	22.34	5433
	10	0.004	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	1.5	symmetric	14.35	13.96	11.39	2624
	10	0.004	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2	symmetric	15.35	21.19	18.44	4206
	10	0.004	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	3	symmetric	16.35	25.93	23.17	6111
	10	0.004	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	5	symmetric	17.35	25.93	23.17	6112
	10	0.004	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.2	2.5	symmetric	18.35	24.97	22.24	5404
	10	0.004	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.5	2.5	symmetric	19.35	20.85	18.13	4126
	10	0.004	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.3	0.333	2.5	symmetric	20.35	25.18	22.56	5943
	10	0.004	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.4	0.333	2.5	symmetric	21.35	25.49	22.81	5693
	10	0.004	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.6	0.333	2.5	symmetric	22.35	23.66	20.73	4834
	10	0.004	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.7	0.333	2.5	symmetric	23.35	17.76	14.89	3570
	10	0.004	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.8	0.333	2.5	symmetric	24.35	8.03	5.23	1623
	10	0.004	10	1	0.1	2	5	0.001	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	0.35	11.69	8.36	4205
	10	0.004	10	1	0.1	2	5	0.002	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	0.35	15.9	12.4	5249
	10	0.004	10	1	0.1	2	5	0.003	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	0.35	18.29	14.69	5809
	10	0.004	10	1	0.1	2	5	0.004	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	0.35	21.36	17.85	6109
	10	0.004	10	1	0.1	2	5	0.006	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	0.35	26.38	23.25	6716
	10	0.004	10	1	0.1	2	5	0.007	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	0.35	30.44	27.66	7131
	10	0.004	10	1	0.1	2	5	0.008	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	0.35	34.26	32.11	7568
	10	0.004	10	1	0.1	2	5	0.009	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	0.35	38.15	36.23	8082
	10	0.004	10	1	0.1	2	5	0.01	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	0.35	41.73	40.44	8555
	10	0.001	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	0.35	9.92	8.59	2365
	10	0.002	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	0.35	19.34	16.17	5127
	10	0.003	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	0.35	22.08	18.61	5915
	10	0.005	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	0.35	24.24	20.54	6643
	10	0.006	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	0.35	25.85	22.39	6608
	10	0.007	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	0.35	26.95	23.78	6707
	10	0.008	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	0.35	28.22	25.19	6940
	10	0.009	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	0.35	28.68	25.55	7229
	10	0.01	10	1	0.1	2	5	0.005	10	0.1	2	25	0.8	500	0.7	0.8	0.5	0.333	2.5	symmetric	0.35	29.4	26.41	7340
Final Result:	10	0.004	9	4	0.1	2	5	0.01	10	0.1	2	25	0.5	500	0.5	0.8	0.5	0.333	3	asymmetric	0.35	93.38	95.25	130993



Simulated CDE Scores:
Nonbinder = -8.7
Low affinity = 12.0
Moderate affinity = 12.4
High affinity = 12.9
Very high affinity = 13.1

Figure 4.7: *Simulated values of concentration-dependent enrichment and their hypothesized affinities. A stronger affinity is expected to pull down a greater fraction of a ligand at lower protein concentrations, and that property is reflected in the CDE score.*

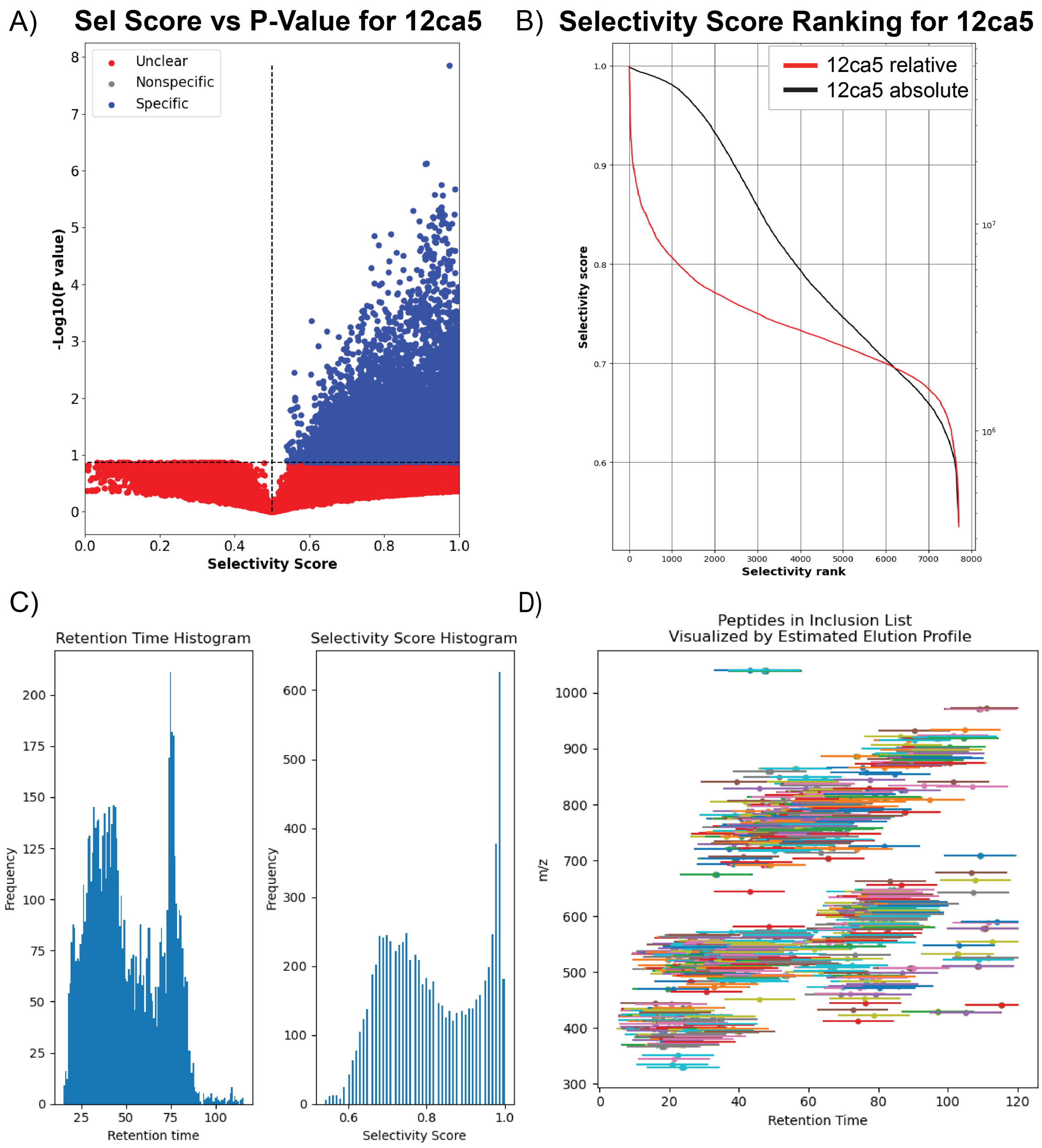


Figure 4.8: *Complete results for 12ca5 of the retrospective analysis of AS-MS performed against 12ca5 and ACE2 show extent of data incompleteness. Sequence coverage of the untargeted analysis was in the range of 10 – 18% of the total identified features, where sequence fidelity was as low as 0.24 – 1.1%. pyBinder analysis identified about 7700 target specific features for 12ca5, compared to the 373 peptides with an ALC ≥ 80 identified in the untargeted run. Because this analysis was done retrospectively, the concentration-dependent enrichment scores were unable to be calculated since the selection was performed with only one concentration of protein.*

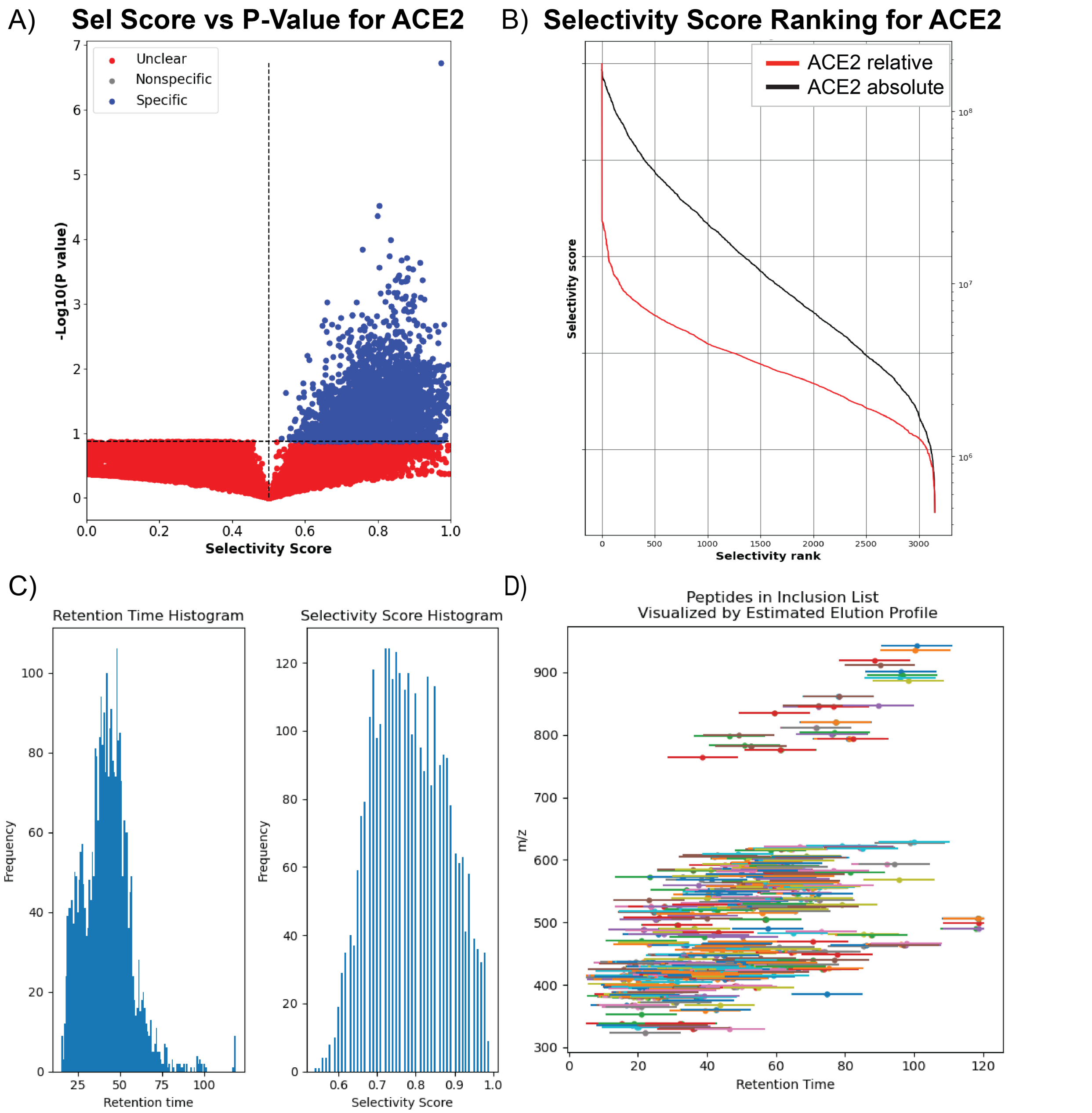


Figure 4.9: *Complete results for ACE2 of the retrospective analysis of AS-MS performed against 12ca5 and ACE2 show extent of data incompleteness. Sequence coverage of the untargeted analysis was in the range of 10 – 18% of the total identified features, where sequence fidelity was as low as 0.24 – 1.1%. pyBinder analysis identified about 3100 target specific features for ACE2, compared to the 80 peptides with an ALC ≥ 80 identified in the untargeted run. Because this analysis was done retrospectively, the concentration-dependent enrichment scores were unable to be calculated since the selection was performed with only one concentration of protein.*

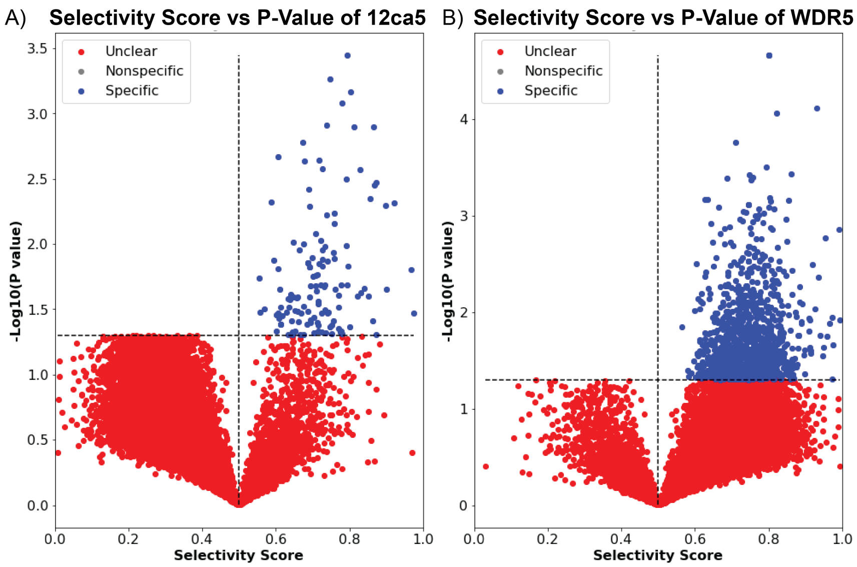


Figure 4.10: *Plots comparing the calculated selectivity score versus the p-value for features identified from selection against 12ca5 and WDR5. Data points in blue showed are identified as statistically significant (α = 0.05) when compared to the extracted areas in the opposing protein. Data points in red have a p-value above the threshold.*

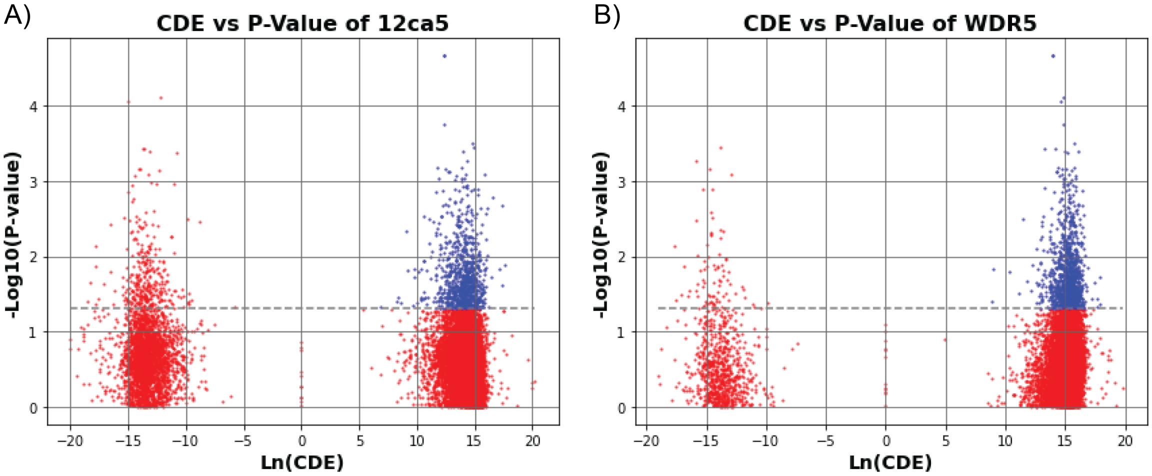


Figure 4.11: *Plots comparing the calculated CDE score versus the p-value for features identified from selection against 12ca5 and WDR5. Data points in blue showed are identified as statistically significant (α = 0.05) when compared to the extracted areas in the opposing protein. Data points in red have a p-value above the threshold.*

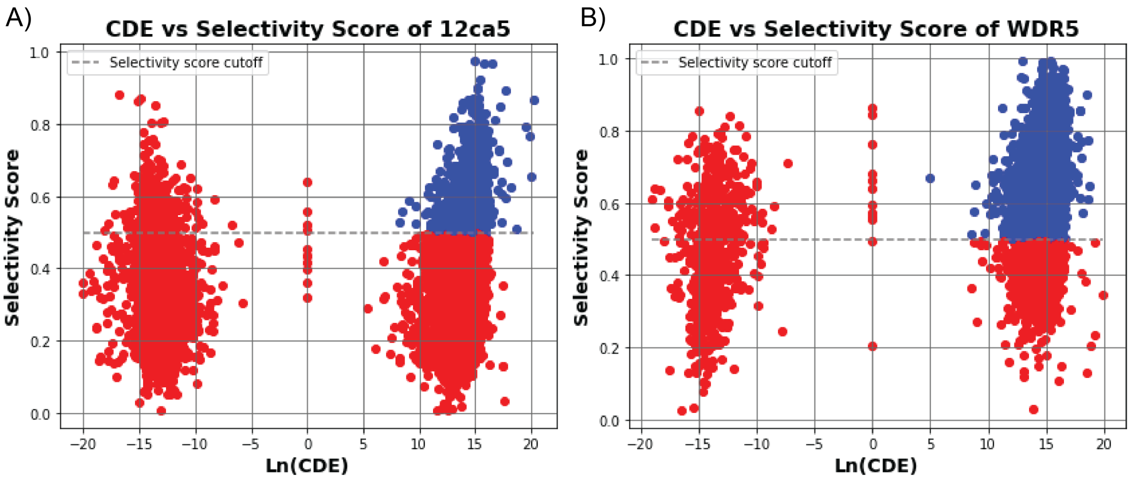


Figure 4.12: *Plots comparing the calculated selectivity score versus the selectivity score for features identified from selection against 12ca5 and WDR5. Data points in blue showed have a selectivity score greater than 0.5, which for an experiment comparing two proteins signifies a degree of selectivity for that protein.*

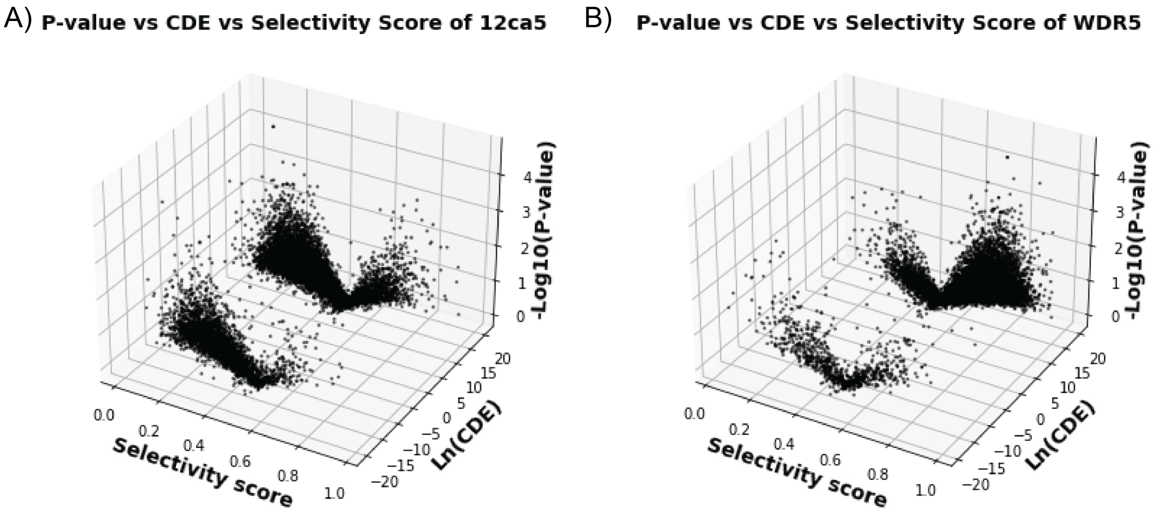


Figure 4.13: *Plot comparing the selectivity scores, CDE scores, and p-values calculated by pyBinder for all detected features.*

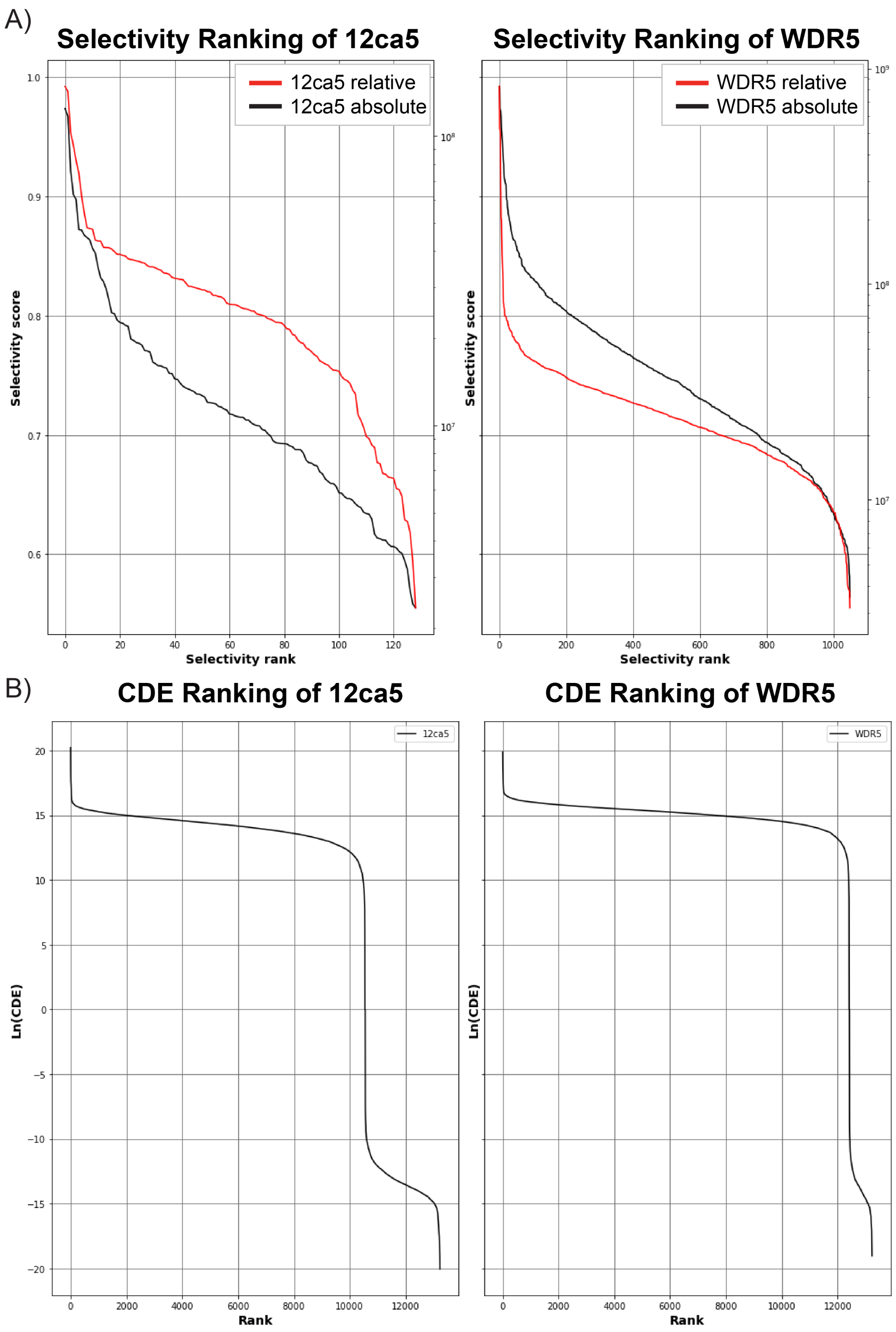


Figure 4.14: *Rankings of features based on their (A) selectivity scores and (B) CDE scores with respect to 12ca5 or WDR5. Relative values for selectivity score are calculated based on the fraction of total area observed as defined in Figure 4.3, while absolute values for selectivity score are based on the total area observed for a single protein and are not normalized relative to total area across all proteins.*

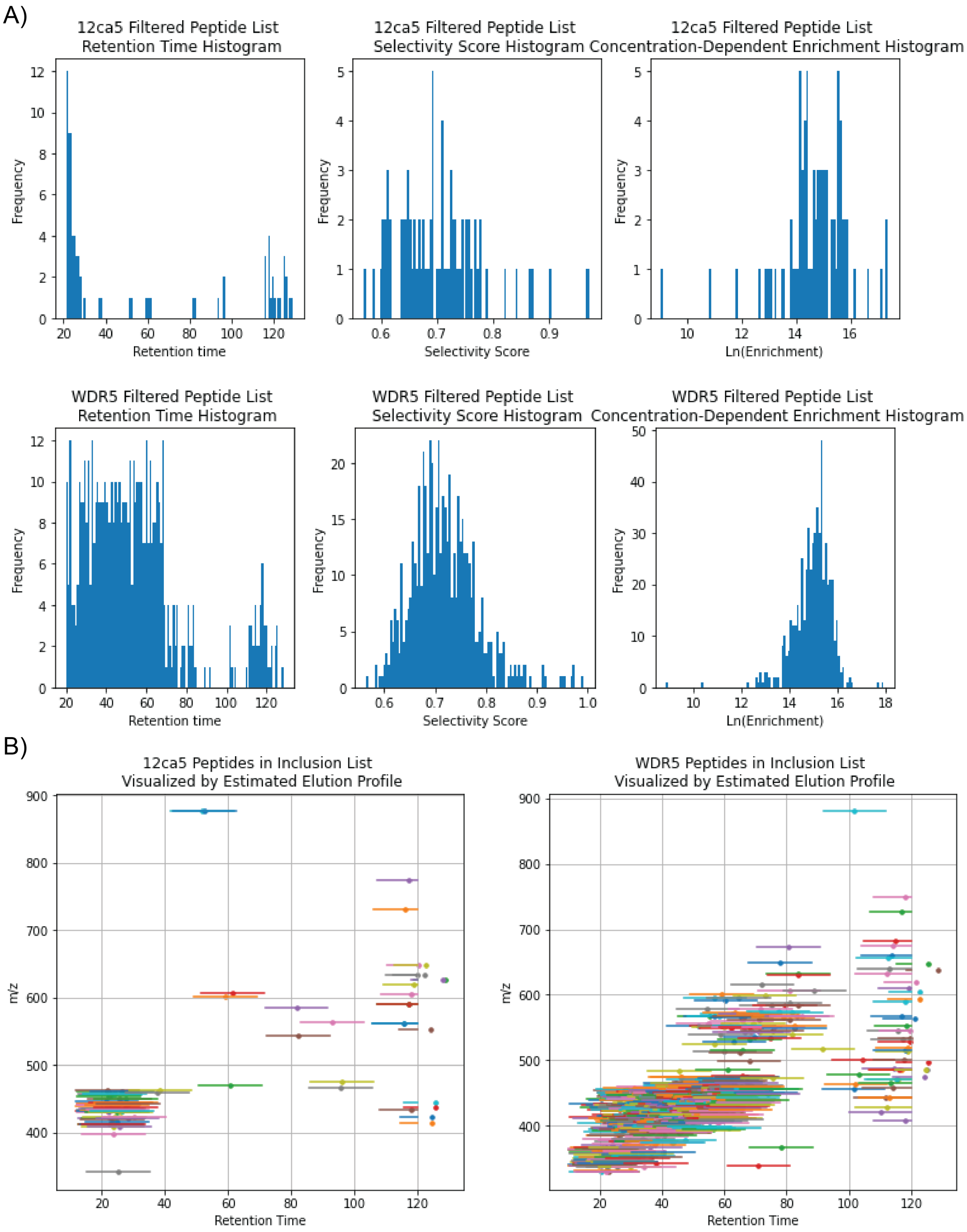


Figure 4.15: *Visualization of the retention times and scoring for the inclusion list of peptide features for each protein. (A) Histograms show the distribution of retention times, selectivity scores, and concentration-dependent enrichment scores for 12ca5 and WDR5. (B) Estimated feature maps for the inclusion lists for 12ca5 and WDR5, with the retention time windows used in the targeted sequencing runs shown for each feature.*

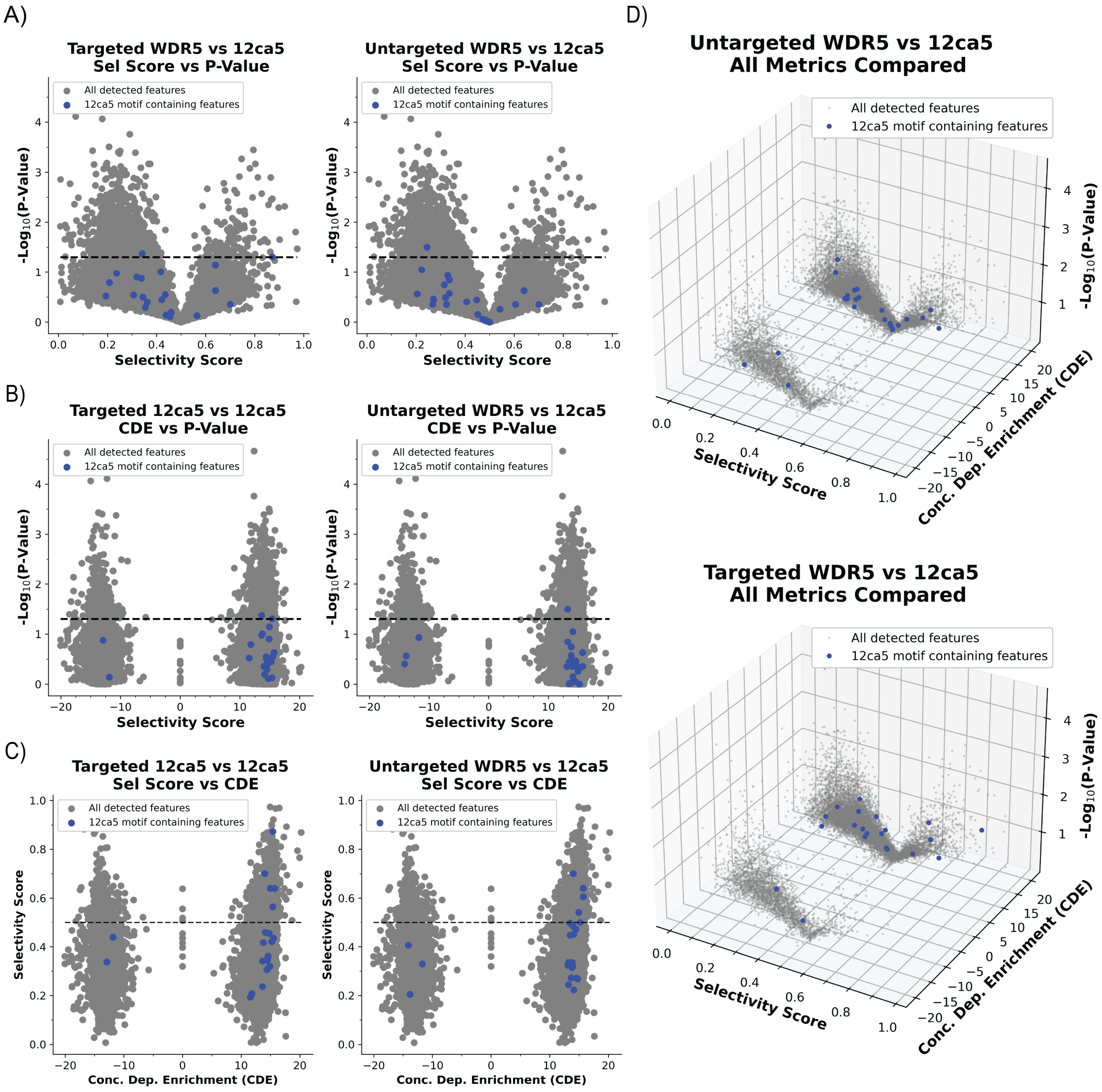


Figure 4.16: *Analysis of the scoring for peptides containing the characteristic 12ca5 binding motif D**DY(A/S). Fewer overall sequences were observed compared to WDR5, likely due to the increased stringency of binding introduced by utilizing a four-residue motif rather than three-residue motifs seen for WDR5. Overall, the trend of using the sign of the CDE score as a filter still applies, but the weakness of the p-value is also shown.*

4.8. 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).
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. Li, J. *et al.* TMTpro reagents: a set of isobaric labeling mass tags enables simultaneous proteome-wide measurements across 16 samples. *Nat. Methods* **17**, 399–404 (2020).
25. 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).
26. 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).
27. Ma, B. *et al.* PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **17**, 2337–2342 (2003).
28. 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).
29. 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).
30. Ong, S.-E. & Mann, M. Mass spectrometry–based proteomics turns quantitative. *Nat. Chem. Biol.* **1**, 252–262 (2005).
31. Vinogradov, A. A. *et al.* Library Design-Facilitated High-Throughput Sequencing of Synthetic Peptide Libraries. *ACS Comb. Sci.* **19**, 694–701 (2017).

32. 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).
33. Paizs, B. & Suhai, S. Fragmentation pathways of protonated peptides. *Mass Spectrom. Rev.* **24**, 508–548 (2005).
34. 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).
35. 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).
36. Lam, K. S., Lebl, M. & Krchňák, V. The “One-Bead-One-Compound” Combinatorial Library Method. *Chem. Rev.* **97**, 411–448 (1997).
37. 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).
38. Sinitcyn, P. *et al.* MaxDIA enables library-based and library-free data-independent acquisition proteomics. *Nat. Biotechnol.* **39**, 1563–1573 (2021).
39. 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).
40. 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).
41. Lange, E. *et al.* A geometric approach for the alignment of liquid chromatography—mass spectrometry data. *Bioinformatics* **23**, i273–i281 (2007).
42. 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).
43. 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).
44. Stock, L., Hosoume, J. & Treptow, W. Concentration-Dependent Binding of Small Ligands to Multiple Saturable Sites in Membrane Proteins. *Sci. Rep.* **7**, 5734 (2017).
45. Xu, M. *et al.* Concentration-Dependent Enrichment Identifies Primary Protein Targets of Multitarget Bioactive Molecules. *J. Proteome Res.* **22**, 802–811 (2023).
46. Houghten, R. A. *et al.* Generation and use of synthetic peptide combinatorial libraries for basic research and drug discovery. *Nature* **354**, 84–86 (1991).
47. 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).

48. Ding, J. *et al.* Discovery of Potent Small-Molecule Inhibitors of WDR5-MYC Interaction. *ACS Chem. Biol.* **18**, 34–40 (2023).

5. Unsupervised Machine Learning Rationalizes Abiotic Picomolar Ligand Discovery

The work presented in this chapter has been reproduced and adapted from the following publication:

Brown, J.S.; **Lee, M.A.**; Mohapatra, S.; Misteli, R.; Tseo, Y.; Grob. N. M.; Quartararo, A. J.; Loas, A.; Gomez-Bombarelli, R.; Pentelute, B.L. Unsupervised machine learning rationalizes abiotic picomolar ligand discovery. *ChemRxiv Preprint*. *Manuscript submission for peer review in progress.* **2024**

5.1. Introduction

In the chemical space between small molecules and biologics, peptidomimetic therapeutics have become more prevalent in regulatory approvals, with 80 drugs now in the market.[1,2] With higher clinical trial success rates than small molecules,[3] peptides and peptidomimetics can be inexpensive relative to other biologics,[1] while offering sufficiently large surface area for high-affinity interactions with shallow or difficult binding interfaces.[4–6] For initial *de novo* discovery of peptide ligands, several affinity selection techniques isolate high-affinity ligands to biomolecular targets including phage display,[7,8] mRNA display,[9] and recently affinity selection-mass spectrometry (AS-MS).[10–12]

However, the development of preclinical peptidomimetic candidates can be time-consuming due to the lack of guiding chemical design rules. Peptide ligands nominated from affinity selection or virtual screening are experimentally validated to confirm their affinity or activity. Then, peptides undergo optimization for lead generation through multiple cycles of synthesis, biophysical, and biological evaluations.[13–15] The goal of this optimization process is to increase activity and proteolytic stability, limit toxicity, and improve pharmacokinetics. Often, diverse abiotic or "noncanonical" amino acids and macrocyclization are utilized toward this goal.[1,2,9,16] This process can be especially challenging if the hit is of poor initial quality or affinity,[15] in a local activity minimum, or if it exhibits nonintuitive "activity cliffs."[17] At the end of this process, critical features or residues are understood to drive high-affinity binding or function, called a "motif,"[1–3,18–20] However, limited chemical design rules guide the exchange of canonical for noncanonical amino acids in this process, with peptidomimetics known for breaking rules that guide small molecule development.[21] As such, the most reliable approach is comprehensive sampling of the unexplored, noncanonical chemical space through time-consuming cycles of synthesis and experimental testing of individual compounds.[18,19] Thus, this process would greatly benefit from intuitive tools to translate functional knowledge across the canonical and noncanonical chemical spaces and predict peptidomimetic functionality before synthesis.[22]

Machine learning (ML) is poised to facilitate a paradigm shift in drug discovery and development.[22–24] Generally, unsupervised methods can identify patterns within unlabeled datasets, hence its common application in omics data analysis.[25–27] Additionally, unsupervised learning can evaluate embeddings from supervised models, as dimensionality reduction is typically performed with no weighting.[28–31] Supervised learning can interpolate and expand from labeled training data, where labels describe function or activity.[23] Where there are sufficiently large datasets of labeled peptide data, supervised learning has been applied toward the discovery of antimicrobial, cell-penetrant, or immunogenic peptides.[32–34] Though significant advancements have been made,[35] publicly-available peptide ligand data do not appear large or diverse enough to provide general ligand prediction, with little-to-no data available for peptidomimetics that include noncanonical amino acids. Thus, state-of-the-art peptide ligand discovery programs currently deploy a mixture of computational and ML analysis with expertly gathered experimental datasets. Recent examples largely include genetically-encoded discovery methods, where round-to-round enrichment serves as the label in supervised learning analysis.[36–38]

In contrast to genetically-encoded platforms, AS-MS has unparalleled use of abiotic chemical libraries, meaning its datasets could enable the development of an ML approach to broadly connect chemical space. AS-MS has historically been a screening tool;[39] but has been advanced with target-focused libraries,[16,40] and onward to *de novo* discovery with peptidomimetics against multiple biomolecular targets with large libraries (>10^8 members).[10–12] AS-MS experiments are rapid, utilizing only a single round of enrichment to identify high-affinity binders. However, this practical advantage limits the potential application of several commonly utilized supervised ML models. Direct modeling of binding affinity is not possible since AS-MS does not provide binding affinity or enrichment information about each reported compound, prohibiting regression. A binary classification model could be considered (i.e., binder vs nonbinder), but it would also lack any resolution of the continuum of binding across the range of reported compounds ($K_D \lesssim 300$ nM).[11] The utility of a classification model could be further compromised by target-dependent effects, which can strongly affect the resulting quality, quantity, and extent of nonspecific binder recovery within the data. Nevertheless, unsupervised learning techniques including dimensionality reduction remain well-suited to identify concealed patterns or clusters from affinity selection datasets.[31] In contrast to binary classification, dimensionality reduction may offer a more accessible, reliable, and unbiased representation of the data from ligand discovery experiments to aid expert interpretation. Toward this goal, two challenges remain: First, because of the noncanonical chemical diversity available to AS-MS, an optimal encoding representation of peptidomimetics should be investigated.[41] Second, the choice of dimensionality reduction method and optimization remains open and is usually referred to as an art more than a science.[25]

Herein, we demonstrate the utility of unsupervised learning to generate two-dimensional "maps" of the chemical space from peptide ligand discovery datasets. These maps visualize the chemical space of peptides and noncanonical peptidomimetics isolated from AS-MS protocols. For this study, we utilized anti-hemagglutinin antibody (12ca5) as the protein target. We surveyed five diverse representation methods ranging in complexity from low-dimensional one-hot encoding and physicochemical encoding, to high-dimensional protein language pretrained representations from the Evolutionary Scale Model-2 (ESM-2),[42] extended connectivity Fingerprints (ECFP_6), and N-grams encoding.[30,32,43] For dimensionality reduction, we primarily compared linear and nonlinear decomposition by principal component analysis (PCA) and uniform manifold approximation (UMAP), respectively. Clusters within the constructed maps enabled highly sensitive motif discovery by the isolation of the consensus and centroid sequence of the cluster. Lastly, we defined boundaries that separate regions of high-affinity peptides from the remaining chemical space, represented by nonspecific peptides and peptides sampled from the original library. While seen first in the canonical space, these boundaries are shown to be consistent in the broader peptidomimetic space, as supported by the experimental testing of all peptidomimetics discovered. AS-MS demonstrated its rapid ability to sample the noncanonical sequence space, with the discovery of mixed canonical-noncanonical peptidomimetic demonstrating a K_D of 210 pM. Thus, we expect these sequence space maps to inform the derivatization and generation of functional high affinity peptidomimetics.

5.2. Results and Discussion

5.2.1. Diverse representations and dimensionality reduction methods create chemical space maps of peptides discovered by AS-MS.

AS-MS experiments using anti-hemagglutinin antibody (12ca5) provided a ligand dataset for unsupervised learning analysis. Twelve libraries of $X_{12}K$ design each containing 200 million peptides (2.4 billion total) were synthesized, validated, and used in AS-MS, where X is any proteinogenic “canonical” amino acid, except cysteine and isoleucine. As an affinity selection, AS-MS only reports peptide ligands with sufficient binding affinity (4,104 peptides after filtering across all libraries), whereas nonbinding peptides are washed away and unidentified (Figure 5.1A). With less than 350 peptides identified per experiment, nearly all of the 200 million library peptides used do not bind. Thus, we sequenced a small subsample (e.g., ~500 peptides) of each $X_{12}K$ library to generate a dataset of nonbinders (5,047 peptide after filtering). The subsample of the libraries does not contain any motifs or pattern, as observed in Figure 5.11, and appears to be randomly dispersed over the $X_{12}K$ sequence space. Thus, by comparing target-enriched AS-MS ligands versus the subsampled library, we expected that overrepresentation of peptides with shared motifs within the AS-MS ligand could be due to target-based affinity enrichment.

Five different representations were used: one-hot, physicochemical property, latent embeddings from the evolutionarily-learned language model ESM-2, Fingerprint, and N-grams based encoding. All encoding methods strive to maintain human interpretability, while capturing sufficient ‘machine-readable’ detail.[23,24,44] First, one-hot and physicochemical encoding were used as common encoding methods (Table 5.5, see *Section 5.5.9*). Third, we utilized latent embeddings of the entire peptide from the protein language model ESM-2.[26,42] Evolutionarily-pretrained models including ESM-2 infer properties from the primary sequence, and could provide additional homological information outside of other encodings.[26,42] Fourth, extended connectivity Fingerprints (ECFP_6) from RDKit represented each amino acid as a vector, where each index indicates the presence (1) or absence (0) of a specific molecular substructure (see Figure 5.6). Similarities between amino acids including noncanonicals is encoded through shared substructures.[32,45] Fifth, N-grams encoding represented the entire peptide by its ungapped motifs, irrespective of position. The possible n-mer motifs were pre-calculated from the dataset to maintain computational practicality (Figure 5.8). Overall, these representations cover diverse aspects peptides across a range of dimensionalities (vector length, Figure 5.1B).

For dimensionality reduction, three methods were deployed based on the diversity of their theoretical underpinning including linear, non-linear reduction, and similarity mapping. Principal component analysis (PCA)[46] was used as the linear dimensionality reduction method. PCA is highly interpretable and deterministic, because components are built from the global variance of the data. Uniform manifold approximation (UMAP)[31] provided non-linear reduction and is user-friendly, requiring little hyperparameter optimization. The primary UMAP hyperparameter of n_neighbors was optimal at 5-10% of the dataset size to balance the representation (Figure 5.9). In Figure 5.1B, the AS-MS ligand data was combined with nonbinders from sampling the library without re-learning. The similarity mapping method, multidimensional scaling, showed poor ability to form any clusters (Figure 5.10) and the nonbinder library peptides could not be added without re-learning.

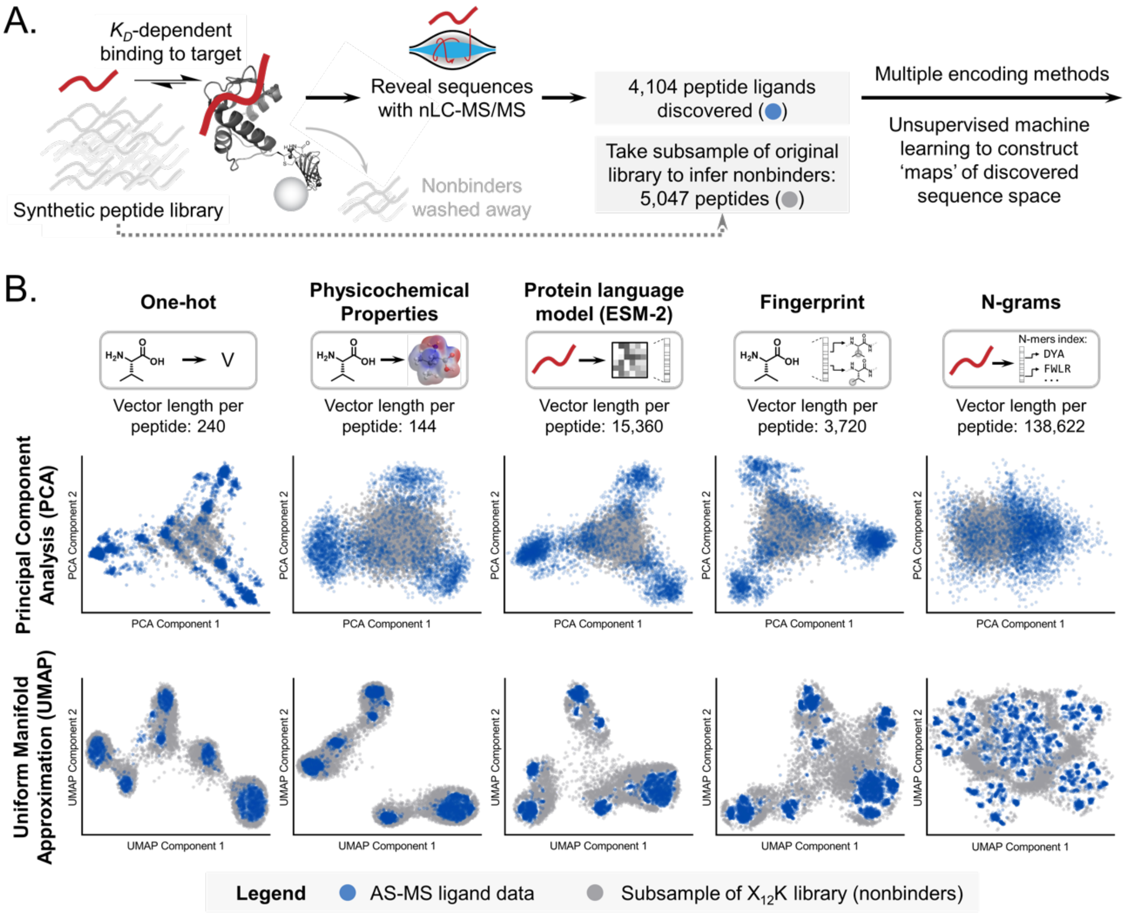


Figure 5.1: *Affinity selection-mass spectrometry (AS-MS) discovers peptides that appeared in separate regions from nonbinders in sequence space maps constructed by unsupervised dimensionality reduction across diverse encoding methods. (A) From AS-MS with 12ca5, 4,104 peptide ligands were identified after filtering and 5,047 peptides were identified as presumed nonbinders by directly sampling the original peptide library. (B) AS-MS peptides were encoded using one-hot, physicochemical property, latent space embeddings from the protein transformer language model (ESM-2), Fingerprint, and N-grams encoding methods, each describing different aspects and dimensional (vector) lengths. Dimensionality reduction of encoded peptides using PCA and UMAP constructed two-dimensional "maps" of the sequence space. Each peptide has a corresponding embedding (coordinate point) on the two-dimensional map, and several points densely grouped together form a cluster.*

5.2.2. Novel sequences are found at the edge of PCA maps and at the center of UMAP clusters.

The encoding and dimensionality reduction method strongly affect the resulting map (Figure 1B). Across all representations, PCA reduction placed the nonbinding library peptides at the center of the map, while most AS-MS ligands were pushed toward the edge. Only one-hot encoding provided distinct clusters, while physicochemical, ESM-2, and Fingerprint encoding provided 3 diffuse clusters at the map edge. In contrast to PCA, UMAP showed tight clusters with all encodings, while nonbinding library peptides filled the cluster interspace. All maps showed approximately 5-6 macro-clusters with varying resolution of smaller clusters. N-grams encoding showed a multitude of clusters grouped loosely into macro-clusters, far beyond all other methods. Color-coding 12ca5-specific labels further supported the localization of novel sequences at the PCA map edge and the center of UMAP clusters, separate from nonspecific or nonbinding library peptides (see Figure 5.12). Peptides containing the known high-affinity motif D**DY(A/S) were labeled as 12ca5-specific binders and were located mostly at the PCA map edge and in the center of UMAP clusters.[12,47,48] AS-MS peptides that did not contain the common motif in any form were labeled as nonspecific (Table 5.6, see *Section 5.7.2*) and were found near the nonbinding library peptides at the center of PCA maps and in the UMAP cluster interspace.

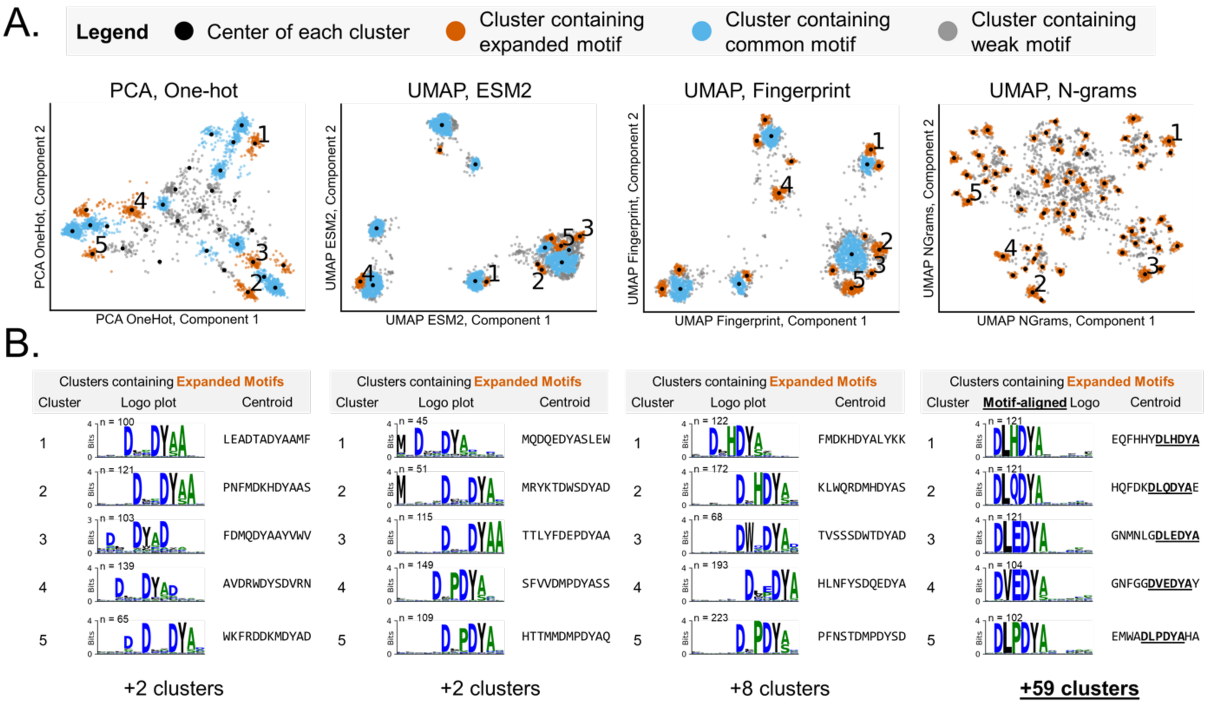


Figure 5.2: *In-depth motif discovery was made possible by high-dimensional peptide encoding techniques (ESM-2, Fingerprint, and N-grams) with UMAP, expanding upon the commonly known motif of D**DY(A/S). (A) All sequence space maps were analyzed to detect the sequences within dense groups of points (clusters), wherein each point represents a peptide. For each cluster, the geometric average of all peptide coordinates is marked with a black marker dot, defining the cluster center. Analysis of the sequences within each cluster allowed the assignment that the cluster contains a consensus sequence or motif that expands (orange) upon the common motif (blue), or is weak (gray). Maps constructed using UMAP and high-dimensional peptide encoders (ESM-2, Fingerprint, and N-grams encoding) provided most of the clusters with expanded motif information (see Figure 5.13 for all). (B) Corresponding logo plots, consensus sequences, and a centroid sequence for each method, corresponding to the numbered clusters in A. The centroid sequence of each cluster was reported as the peptide closest to the geometric center (black dot), with the option to report more sequences interspersed within the cluster available. Because N-grams encoding is irrespective of frameshift, sequences within each cluster were aligned by ClustalW2*[49] *to the second position in order to simply report the logo plot displaying the consensus sequence (motif) within the cluster. Logo plots were constructed using Logomaker.*[50] *Centroid sequences from N-grams can show the exact frameshift location of the motif (e.g., UMAP N-grams Cluster 1 motif is *DLHDYA*, which starts at frameshift 7). For brevity, the information from only five clusters are shown (see Section 5.7.4 for all).*

Each cluster was assigned a label based on its consensus sequence and logo plot to either expand upon (shown in orange) the common motif (D**DY(A/S), shown in blue) or contain weak motifs shown in gray (Figure 5.2). Clusters were algorithmically detected based on cluster density (Table 5.7).[51] Density-based spatial clustering of applications with noise (DBSCAN)[51] was found especially useful as clusters initialized at the dense cluster center, the known location of novel sequences. Most maps did not contain clusters with expanded motifs beyond the common motif (Figure 5.13). However, four maps showed clusters that contained expanded motifs and revealing more information depth from this discovery dataset (Figure 5.2). With PCA, one-hot encoding revealed motifs containing an additional alanine and additional aspartic acids (Table 5.8). However, of the maps that contained cluster with expanded motifs, the PCA one-hot map also showed the largest number of weak motifs.

5.2.3. High-dimensional descriptors provided most of the expanded motifs discovered in UMAP-constructed maps.

UMAP-constructed maps of the peptides encoded by high-dimensional descriptors (ESM-2, Fingerprint, and N-grams) provided additional clustering resolution to reveal many expanded motifs. ESM-2 encoding showed some similar motif results as the one-hot encoded PCA map, with additional weighting for N-terminal methionine. The one-hot, physicochemical, and Fingerprint encoded maps exhibited six ‘macro-clusters’ arising from the frameshifts of the common motif in a 12-mer variable region. However, the Fingerprint map provided more cluster resolution and expanded motifs. Moreover, of all the UMAP-constructed plots, the Fingerprint encoded map balanced cluster resolution, resulting chemical motif information depth, while maintaining a globally connected chemical space, with nonbinding library peptides contiguously filling the cluster interspace (Figure 5.1B).

By far, N-grams encoding with UMAP provided the highest cluster resolution and motif detection sensitivity, providing 64 clusters containing expanded motifs and more information depth than leading techniques. This high resolution is likely because the N-grams encoding only depends on the presence or absence of a n-mer or motif, irrespective of its frameshift. Thus, all frameshifts of a motif are encoded the same, combining to increase the sensitivity of detection. For motif detection comparison, the AS-MS ligand data was input to the MEME Suite using XSTREME,[52] one of the leading motif detection and discovery platforms.[53] The XSTREME analysis found only the common D**DY(A/S) motif within the dataset (see *Section 5.7.6*), except the expanded motif of *D**DYAD*. With our highly sensitive clustering approach, all motifs can be ranked and aid in building or contextualizing structure activity relationships from discovery datasets. For this target (12ca5), the preferred motif appeared to be *DΦΠDYA*, with Φ = L, V, and M, and Π = amino acids including Q, E, H, or P, consistent with literature.[12,47,48]

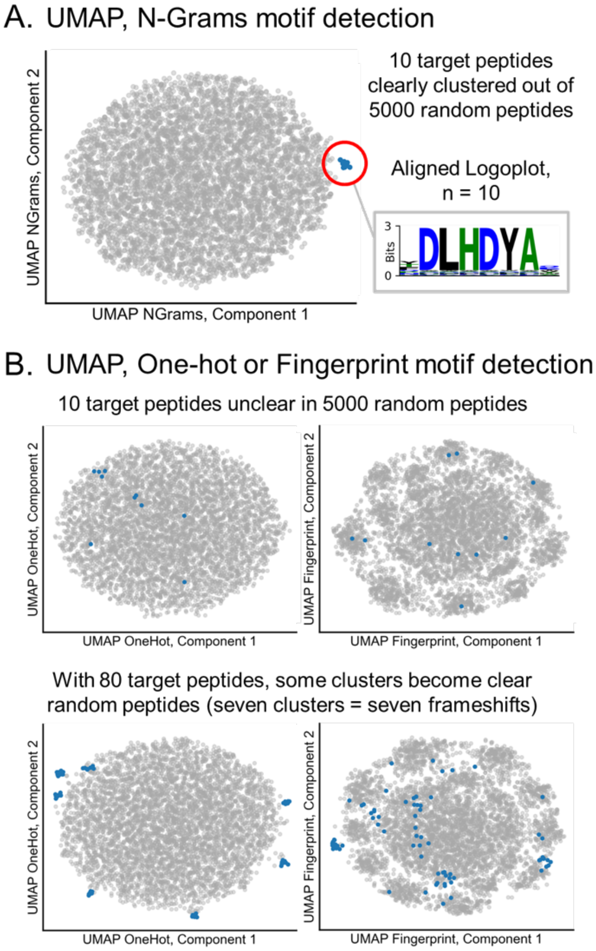


Figure 5.3: *N-grams encoding with UMAP provided highly sensitive clustering and identification of ≤10 unaligned target peptides, which contained a 6-mer target motif at random frameshifts in a large dataset of nonbinding library peptides. A. Target peptides containing a *DLHDYA* motif at random frameshifts from the AS-MS data were combined*

with library peptides for 5,000 total to test the sensitivity of successful clustering, detection, and identification. N-grams encoding with UMAP provided the lowest threshold of ≤10 peptides required for the formation of a tight, distinct cluster, primarily because N-grams encoding is agnostic of frameshift by design. The sensitivity of detection was compared to the MEME Suite using XSTREME,[52,53] which required 20 peptides before the motif was detected. B. One-hot and Fingerprint encoding were unable to produce distinct clusters with 10 target peptides. At 80 target peptides, one-hot and Fingerprint encoded UMAP maps exhibited 7 clusters each for the frameshifts of the target peptide motif (Figure 5.24). AS-MS peptides are shown in blue with library peptides in gray.

Toward further proving its motif detection sensitivity, N-grams encoding clustered as few as 10 unaligned motif-containing peptides in a large dataset, whereas the leading technique XSTREME required 20 sequences (see *Section 5.7.5 and 5.7.6.2*). The utility of this mapping approach hinges upon the sensitive clustering of novel sequences together in low-data regimes. Thus, we evaluated dimensionality reduction-encoding pairs to cluster and identify a small number of similar peptides containing an unaligned motif (*DLHDYA*) in a large dataset of library peptides (5,000 total). XSTREME analysis uses deterministic optimization based on the expectation maximization to perform motif detection.[53] Our clustering method could be categorized to perform enumeration (i.e., enumerate all n-mers), which is inherently more sensitive. Enumeration is computationally avoided; however, our approach is feasible because it precomputes n-mers from the input peptides, which are short relative to genes. For the same data, one-hot and Fingerprint encoding required ≥ 80 unaligned peptides (Figure 5.3B), but only required ≤ 10 peptides if the motif was placed at the same frameshift (Figure 5.25), meaning the clustering limit is primarily limited by UMAP.

Related to sensitivity, the augmentation of a small AS-MS ligand dataset with library peptides unexpectedly clarified the appearance and density of clusters. Maps prepared from a small number (~100) of peptides appeared dispersed, even for highly similar peptides. However, in these low data regimes, augmenting the dataset with nonbinding library peptides for co- learning improved cluster clarity of similar peptides (Figure 5.26). At a minimum, an augmented dataset should improve the contextualization of the original dataset but could also outline nearby chemical space to power Bayesian exploration.

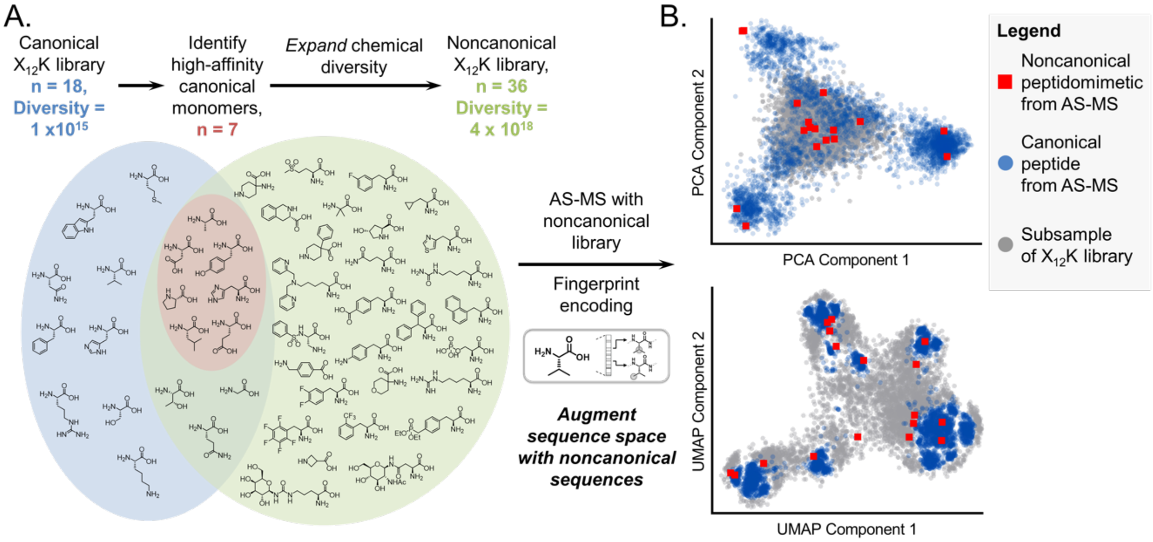

A.
Canonical $X_{12}K$ library
n = 18,
Diversity =
1 x10^{15}
Identify high-affinity canonical monomers,
n = 7
Expand chemical diversity
Noncanonical $X_{12}K$ library,
n = 36
Diversity =
4 x 10^{18}
AS-MS with noncanonical library
Fingerprint encoding
Augment sequence space with noncanonical sequences
B.
PCA Component 2
PCA Component 1
Legend
🟥 Noncanonical peptidomimetic from AS-MS
● Canonical peptide from AS-MS
● Subsample of $X_{12}K$ library
UMAP Component 2
UMAP Component 1

Figure 5.4: *The translation of canonically-understood binding function into the noncanonical chemical space was tested by the addition of peptidomimetics discovered from a highly noncanonical library. A. From unsupervised ML analysis of the canonical data, amino acids that drive high-affinity function over-represented in discovered motifs were identified and included in a high-diversity noncanonical library for peptidomimetic discovery and derivatization. B. Fingerprint encoding robustly encoded AS-MS peptidomimetics. Fingerprint encoding encodes using molecular functional groups regardless of their canonical or noncanonical identity, ensuring semantic consistency of the map. With robust encoding, the boundaries and regions of canonically-understood high-affinity function can be hypothesized to be maintained and translated across the canonical and noncanonical chemical space.*

5.2.4. Discovered peptidomimetics were proposed to test the sequence map fidelity across the canonical and noncanonical sequence space.

The ultimate utility of these sequence maps would be to maintain consistency of function across the broader chemical space. Thus, peptidomimetics were discovered from AS-MS using highly noncanonical libraries were augmented into the dataset (Figure 5.4). A common approach to design noncanonical libraries is to diversify the high-affinity motifs.[16] However, maintaining some randomization could discover other high-affinity motifs and/or improve other non-motif residues. Seven canonical monomers were chosen to be included in the noncanonical library because of their likely role in driving high affinity binding to maintain a likelihood of ligand discovery, while increasing the chemical diversity. Our noncanonical library utilized 36 monomers of the same $X_{12}K$ design, sampling a theoretical sequence space 4,000 times larger than the original library (see Table 5.1), including two synthetically prepared monomers: bis-pyridyl lysine and a galactosyl-citrulline for their structural diversity (see *Section 5.8.1*). From AS-MS experiment, seventeen peptidomimetics were identified from the noncanonical library (Table 5.18). This fewer number of identified compounds were expected because AS-MS still utilized 200 million compounds, meaning the sampling rate of the larger chemical space was lower relative to the canonical library.

AS-MS discovered noncanonical peptidomimetics were robustly augmented into the sequence maps using Fingerprint encoding. A semantically consistent map almost certainly requires similarities between all monomers to be encoded, eliminating one-hot and N-grams encoding for this this task. Encoding based on any proteome-based model (e.g., ESM-2) is unavailable as it lacks noncanonical training. Physicochemical encoding could be made possible by calculations or measurements but was not explored. In contrast, Fingerprint encoding was well-suited as the molecular similarities of all amino acids are readily apparent and captured at the same fidelity. Thus, Fingerprint encoded peptidomimetics were added to develop a co-learned sequence space.

Because the peptidomimetics appeared throughout the sequence space maps, we hypothesized that peptidomimetics localized near motif-containing canonical peptides would also be high-affinity binders. Half of the peptidomimetics were in or near high-affinity canonical clusters, with the other half located at clusters' edges or in the cluster interspace. However, it is unclear if the binding function of these peptidomimetics will be consistently connected to their location on the map, as is it for high-affinity canonical peptides. For example, noncanonical derivatization of a high-affinity canonical peptide would change its location on the map, even if the common motif were maintained.

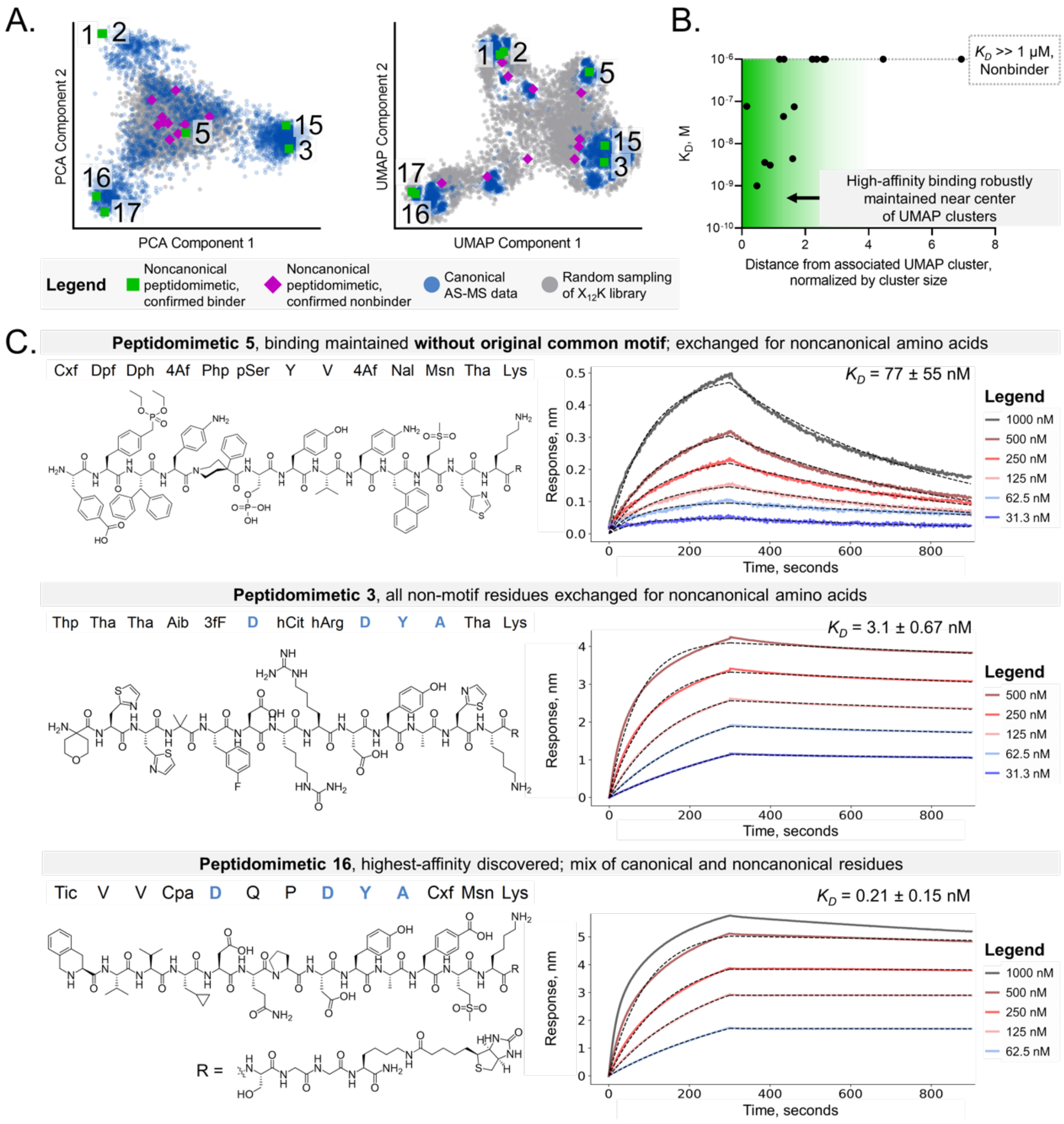


Figure 5.5: *Experimental binding validation of AS-MS peptidomimetics reveals a picomolar binder and reinforced the hypothesized regions of high-affinity binders, separated from nonbinders in the combined noncanonical and canonical sequence space. A. Labels of experimentally confirmed binder or nonbinder from biolayer interferometry (BLI) were overlayed onto the Fingerprint-encoded, UMAP and PCA maps. High affinity peptidomimetics were located at the PCA map edge, except for Peptidomimetic 5. However, with no exception, high affinity peptidomimetics were located in or closely near UMAP clusters, indicating the robust consistency of high affinity binding function between the canonical and noncanonical chemical space. B. A functional boundary can be visualized by plotting the distance of each peptidomimetic from its associated UMAP cluster versus its experimentally measured binding affinity (Dissociation constant, K_D). The distance from each cluster was normalized by the size of the characteristic cluster radius, which was determined by minimizing the summed error between all cluster points to the circle radius.*

C. Of the high-affinity peptidomimetics discovered, Peptidomimetics 3, 5, and 16 are highlighted for the effectiveness of AS-MS to rapidly sample the noncanonical sequence space while maintaining high-affinity binding function. Concentration-dependent binding observed on BLI sensorgrams of immobilized biotinylated peptidomimetics and unlabeled 12ca5 in solution were fit using a 1:1 binding model shown as a black dashed line on top of the data (see Section 5.8.3). Peptidomimetic 16 combines noncanonical and canonical amino acids for the highest affinity observed. The R group corresponds to a SGGK(Biotin) linker utilized in BLI immobilization (see Section 5.5.15).

5.2.5. Boundaries of high-affinity function were robustly consistent with UMAP across the chemical space, highlighted best by Peptidomimetic 5.

All seventeen of the peptidomimetics were tested for their experimental binding using biolayer interferometry (BLI, *Section 5.8.2, 5.8.3, and 5.5.15*). In general, high-affinity peptidomimetics were found in regions of high-affinity motif-containing canonical peptides. Of the high-affinity binders observed, three peptidomimetics (Peptidomimetics 3, 5, and 16) demonstrated the effectiveness of AS-MS to rapidly sample the noncanonical space while maintaining high-affinity binding function (Figure 5.5C). Peptidomimetic 3 exhibited high-affinity binding and was completely noncanonical except for its common motif. Peptidomimetic 16 displayed the highest binding affinity of K_D 210 ± 150 pM and was comprised of the common motif and noncanonical amino acids.

The discovery of Peptidomimetic 5 (K_D = 77 nM), which does not contain the common canonical motif, is significant for demonstrating of the utility of both AS-MS and the UMAP-constructed maps. Since the binding interaction mode is through anionic residues, the phosphoserine and 4-carboxy phenylalanine likely serve the role of aspartic acid in the common motif. Peptidomimetic 5 is evidence that AS-MS can rapidly sample the noncanonical space to discover a completely different abiotic binding motif. With UMAP, all peptidomimetics were localized in or close to canonical binding clusters with no exception. However, In the PCA map, Peptidomimetic 5 localized in the center of the map near nonspecific and nonbinding library peptides. Thus, the UMAP-constructed map robustly aggregated and clustered functional high-affinity binding compounds across both the canonical and noncanonical chemical space.

Across the chemical space, a functional ‘boundary’ between high-affinity binders and nonbinders can be seen when comparing the binding activity of the peptidomimetics to their normalized proximity to UMAP clusters (Figure 5.5B). While the boundary between high-affinity binders and nonbinders is not sharp, this result suggests that the binding activity of any peptide derivatization with noncanonicals can be predicted with a high degree of confidence. Thus, a functional design space can be defined in the combined canonical and noncanonical chemical space.

5.3. Conclusion

We applied unsupervised machine learning to peptide and peptidomimetic ligand discovery data for the visualization, clustering and in-depth extraction of motifs, and construction of functional boundaries between high-affinity binders and nonbinders. From comparison with nonbinding library and nonspecific peptides, novel sequences are found at the PCA map edge and the center of UMAP clusters (Figure 5.1), and further supported by 12ca5-specific labels (see Figure 5.12). While this analysis works well with a large dataset, small discovery datasets can be augmented with library peptides to contextualize discovered peptides, and potentially facilitate cluster formation (Figure 5.26). With UMAP, encodings that produced high-dimensionality descriptors resulted in sequence maps with increased cluster resolution, with frameshift-irrespective encoding by N-grams showing the highest sensitivity for motif discovery. From clusters, the consensus and centroid sequences identifies motifs and peptide binder families (Figure 5.2). This process can readily nominate representative peptides across the diversity of the dataset to “down-sample” and prioritize peptides for binding validation experiments and avoid nonspecific peptides. Thus, we expect this approach could readily be applied to accelerate any ligand discovery platform.

The experimental binding validation of AS-MS discovered peptidomimetics supported the ability to define functionally consistent chemical space, across canonical peptides and noncanonical peptidomimetics. AS-MS rapidly sampled the noncanonical space, exchanging and derivatizing canonical residues. We experimentally validated 7 peptidomimetics, with Peptidomimetic 5 discovered without the common canonical motif and with a Peptidomimetic 16 exhibiting K_D = 210 pM through natural and non-natural amino acids. With only 17 peptidomimetics and 4,014 peptides, unsupervised learning appeared unaffected in its ability to enable the prediction of peptidomimetic binding function and define a functional embedding space, despite the class imbalance.

Our results may imply that significant derivatization from the originally discovered hits decreases the likelihood of maintaining binding. However, significant derivatization, including a full exchange of the common canonical motif, was still rationalized from our analysis (e.g., Peptidomimetic 5). Overall, we expect this analysis to have a range of applications including the definition of functional chemical design spaces, prediction of peptidomimetic functionality before synthesis, and the ML-guided generation or “hallucination” of functional peptidomimetics to accelerate the discovery and development of therapeutics.

5.4. Materials

Table 5.1: *List of abbreviations used.*

Abbreviation	Full name
AGC	Automatic gain control
AggCl	Agglomerative clustering
ALC	Average local confidence
AS-MS	Affinity selection-mass spectrometry
BLI	Biolayer interferometry
Boc	tert-Butyloxycarbonyl
BSA	Bovine serum albumin
CID	Collision induced dissociation
CV	Column volume
Da	Dalton mass unit
DBSCAN	Density-Based Spatial Clustering of Applications with Noise

DCM	Dichloromethane
DIPEA or DIEA	*N,N*-diisopropylethylamine
DMF	*N,N*-dimethylformamide
ECFP_6	Extended connectivity Fingerprint
ESI	Electrospray ionization
ESM-2	Evolutionary scale model-2
EThcD	Electron-transfer dissociation with higher-energy collision
FBS	Fetal bovine serum
Fmoc	9-fluorenylmethyloxycarbonyl
HATU	1-[Bis(dimethylamino) methyl-ene]- 1H-1,2,3-triazolo[4,5-b]-pyridinium 3- oxide hexafluoro-phosphate
HCD	Higher-energy CID
HPLC	high pressure or high performance liquid chromatography
K Buffer	Kinetics buffer
LCMS	Liquid chromatography-mass spectrometry
MDS	Multidimensional scaling
MeCN	Acetonitrile
MEME	Multiple Em for Motif Elicitation
MeOH	Methanol
NHS	*N*-Hydroxysuccinimide
nLC	Nano liquid chromatography
PBS	Phosphate buffer saline
PCA	Principal component analysis
PEG	Polyethylene glycol
PTM	Post-translational modification
SA	Streptavidin
SAR	Structure activity relationship
STREME	Sensitive, Thorough, Rapid, Enriched Motif Elicitation
TFA	Trifluoroacetic acid
Trt	Trityl
UMAP	Uniform manifold approximation
XSTREME	Extreme Sensitive, Thorough, Rapid, Enriched Motif Elicitation

Canonical Fmoc-protected amino acids Fmoc-L-Ala-OH, Fmoc-L-Arg(Pbf)-OH; Fmoc-L-Asn(Trt)-OH; Fmoc-L-Gln(Trt)-OH; Fmoc-L-Leu-OH; Fmoc-L-Lys(Boc)-OH; Fmoc-L-Pro-OH; Fmoc-L-Ser(t-Bu)-OH; Fmoc-L-Tyr(t-Bu)-OH, Fmoc-L-Asp-(Ot-Bu)-OH; Fmoc-L-Glu(Ot-Bu)-OH; Fmoc-Gly-OH; Fmoc-L-Phe-OH; Fmoc-L-Thr(t-Bu)-OH; and Fmoc-L-Val-OH were purchased from Sigma Millipore (Novabiochem) and used as received. Fmoc-L-His(Boc)-OH was purchased from Advanced ChemTech and used as received. Fmoc-Rink amide linker (4-[(R,S)-(2,4-

dimethoxyphenyl)(Fmoc-amino)methyl]phenoxyacetic acid) was purchased from Chem Impex Inc (Wood Dale, IL) and used as received.

Table 5.2: *Noncanonical amino acids used in this work with their associated protecting groups. Unless specified as synthetically produced, all were purchased and used as received.*

Noncanonical amino acid	Abbreviation	1-Letter Abbreviation	Source
Fmoc-L-Phe(2-trifluoromethyl)-OH	2F3F	v	Chem Impex, Inc
Fmoc-3-fluoro-L-phenylalanine	3fF	m	Chem Impex, Inc
Fmoc-4-(Boc-amino)-L-phenylalanine	4AF	k	Chem Impex, Inc
Fmoc-Asn(GlcNAc(Ac)$_3$-β-D)-OH	Agn	X	Millipore Sigma
Fmoc-α-aminoisobutyric acid	Aib	b	Chem Impex, Inc
Fmoc-(4-aminomethyl) benzoic acid	Amb	h	Chem Impex, Inc
Fmoc-azetidine-3-carboxylic acid	Aza	a	Chem Impex, Inc
Fmoc-β-cyclopropyl-L-alanine	Cpa	d	Chem Impex, Inc
Fmoc-(4-tert-butyloxycarbonyl)-L-phenylalanine	Cxf	t	Chem Impex, Inc
Fmoc-3,4-difluoro-L-phenylalanine	DfF	r	Chem Impex, Inc
Fmoc-4-diethylphosphomethyl-L-phenylalanine	Dpf	z	Chem Impex, Inc
Fmoc-3,3-diphenyl-L-alanine	DPh	w	Chem Impex, Inc
Fmoc-L-HomoArg(Pbf)-OH	hArg	o	Chem Impex, Inc
Fmoc-L-homocitrulline	hCit	p	Chem Impex, Inc
Fmoc-O-tert-butyl-L-trans-4-hydroxyproline	Hyp	e	Chem Impex, Inc
Fmoc-L-methionine sulfone	Msn	l	Chem Impex, Inc
Fmoc-3-(1-naphthyl)-L-alanine	Nal	u	Chem Impex, Inc
Fmoc-pentafluoro-L-phenylalanine	PfF	y	Chem Impex, Inc
Fmoc-4-phenylpiperidine-4-carboxylic acid	Php	s	Chem Impex, Inc
1-Boc-piperidine-4-Fmoc-amino-4-carboxylic acid	Pip	f	Chem Impex, Inc
Fmoc-(S)3-amino-2-(phenylsulfonylamino)propionic acid	Psa	x	Chem Impex, Inc
Fmoc-O-benzylphospho-L-serine	pSer	n	Chem Impex, Inc
Fmoc-3-(4-thiazolyl)-L-alanine	Tha	i	Chem Impex, Inc
Fmoc-4-amino-tetrahydropyran-4-carboxylic acid	Thp	g	Chem Impex, Inc
Fmoc-(3S-)-1,2,3,4-tetrahydroisoquinoline-3-carboxylic acid	Tic	j	Chem Impex, Inc
Fmoc-Bispyridinolysine-OH	Bpl	B	Synthesized, see Noncanonical Monomer Synthesis
Fmoc-D-Galactosyl-L-citrulline	Git	Z	Synthesized, see Noncanonical Monomer Synthesis

For the synthesis of noncanonical monomers (Bpl and Git, see *Noncanonical Monomer Synthesis*), Fmoc-Lys-OH was purchased from Ambeed Inc. Sodium triacetoxyborohydride, 2-pyridinecarboxaldehyde, 1,2-dichloroethane, methanol, (D)-(+)-galactose, acetic anhydride and pyridine were purchased from MilliporeSigma. Fmoc-Cit-OH was purchased from Chem-Impex International Inc (Wood Dale, IL). Coupling agent O-(7-azabenzotriazol-1-yl)-*N,N,N’,N’*-

tetramethyluronium hexafluorophosphate (HATU, ≥97.0%) was purchased from P3 Biosystems (Lyndon, Kentucky).

Biosynthesis OmniSolv® grade *N,N*-dimethylformamide (DMF) was purchased from EMD Millipore (DX1732-1) and incubated with 1 pack of AldraAmine trapping agents (for 1000 – 4000 mL DMF, Sigma-Aldrich, catalog number Z511706) for 48 hours prior to use. Diisopropylethylamine (DIEA; 99.5%, biotech grade, catalog number 387649) and piperidine (ACS reagent, ≥99.0%) were purchased from Sigma-Aldrich. Formic acid (FA, 97%) was purchased from Beantown Chemical, Corp. Trifluoroacetic acid (HPLC grade, ≥99.0%), Diethyl ether (anhydrous, ACS reagent, ≥99.0%), acetonitrile (HPLC grade, ≥99.9%), Omnisolv® acetonitrile (LC-MS grade, AX0156-1), Omnisolv® water (LC-MS grade, WX0001-1) and were purchased from Sigma-Aldrich. Formic acid Optima LC/MS (A117) was purchased from Fisher Chemical. Water was deionized using a Milli-Q Reference water purification system (Millipore). Nylon 0.22 µm syringe filters were TISCH brand SPEC17984.

H-Rink Amide-ChemMatrix® (0.49 mmol/g) resin was purchased from PCAS Biomatrix (St-Jean-sur-Richelieu, Quebec, Canada) and 20 µm TentaGel® M NH_2 Monosized Amino Microsphere resin was purchased from Rapp Polymere Inc. (Tübingen, Germany). HyClone™ Fetal Bovine Serum (SH30071.03HI, heat inactivated) was purchased from GE Healthcare Life Sciences (Logan, UT) Dynabeads MyOne Streptavidin T1 magnetic microparticles were purchased from Invitrogen (Carlsbad, CA). Phosphate buffered saline (10x, Molecular biology grade) was purchased from Corning. Sodium chloride (ACS grade) was purchased from Avantor. Guanidine hydrochloride (Cat BP178) and sodium phosphate monobasic monohydrate were purchased from Fisher Scientific.

Mouse anti-hemagglutinin antibody (clone 12ca5) was purchased from Columbia Biosciences Corporation (Cat: 00-1722, Frederick, Maryland) biotin-$(PEG)_4$-NHS ester and biotin-$(PEG)_4$-propionic acid were purchased from ChemPep

Inc. (Wellington, FL). Biotinylation of 12ca5 was performed as previously described.[12]

5.5. Methods

5.5.1. Canonical peptide library synthesis

A total of three libraries were prepared, each portioned into 5 aliquots each (15 aliquots total), with 12 sampled in affinity selection-mass spectrometry experiments. The procedure below describes the synthesis of a single library.

Total number of beads:	1 x 10^9
Size:	20 micron Tentagel M NH2 (Cat: M30202)
Library design:	$X_{12}K\text{-}NH_2$
Variable Positions:	12
# of monomers:	18 (Canonical 20 minus Ile,Cys)

Ala, Asp, Glu, Phe, Gly, His, Lys, Leu, Met, Asn, Pro, Gln, Arg, Ser, Thr, Val, Trp, Tyr

Theoretical diversity:	1.16 x 10^{15}
Redundancy:	4.32 x 10^{-7}

Note: Redundancy is Total number of beads in each library / Theoretical diversity or 1.16 x 10^{15} / 1 x 10^8 and speaks to the sampling rate of the theoretical sequence space available

5.5.2. Noncanonical peptidomimetic library synthesis

A single library was prepared, each portioned into 5 aliquots (5 aliquots total), with 3 sampled in affinity selection-mass spectrometry experiments.

Total number of beads:	1 x 10^9
Size:	20 micron Tentagel M NH2 (Cat: M30202)

Library design: $X_{12}K-NH_2$

Variable Positions: 12

# of monomers: 36

Noncanonical monomers: Aze, Aib, Bpl, Cpa, Hyp, Pip, Thp, Amb, Tha, Tic, 4AF, Msn, 3fF, pSer, hArg, hCit, DfF, Php, CxF, NaI, 2F3F, DPh, Psa, Agn, PfF, Dpf, Git

Also, the following canonicals were included: Ala, Asp, Gly, His, Pro, Gln, Thr, Val, and Tyr, as well as Lys that was only included at the C-terminus position.

Theoretical diversity: 4.74×10^{18}

Redundancy: 1.06×10^{-10}

Note: For both canonical and noncanonical library synthesis, these libraries are highly ‘nonredundant,’ meaning the theoretical sequence is under-sampled. The successful discovery of high-affinity peptide binders is dependent on the presence of the minimal required motif / sequence required for binding. Low-complexity binding motifs defined by 3-5 amino acids are readily discovered because they are statistically common even within a highly nonredundant library. Since the library is highly nonredundant, sequence isomers can be confidently identified and removed (see *Curation of AS-MS Data*) as they are highly unlikely to exist.

5.5.3. Solid-phase peptide library synthesis by split-pool synthesis

4.2 g of 20 μm TentaGel M NH2 resin (0.26 mmol/g, 1.1 mmol, 1.0×10^9 beads) was swollen in and washed with DMF (3x) within a 250mL peptide synthesis vessel (medium frit, 10-15 μm pore size, ChemGlass CG-1866-05). Fmoc-Rink amide linker (2.9 g, 5.4 mmol, 5 eq) was dissolved in HATU solution (0.38 M in DMF, 12.9 mL, 4.5 mmol), activated with DIEA (2.7 mL, 16 mmol) immediately prior to coupling, and added to resin bed. Coupling was performed for 30 min and then washed with DMF (2 x 100 mL). Fmoc removal was completed with 20% piperidine in DMF (1 x 50 mL flow wash followed by 2 x 50 mL, 5 min batch treatments). Resin was then washed with DMF (3 x 150 mL). This process of coupling and Fmoc deprotection was repeated with the Fmoc-Lys(Boc)-OH (2.54 g, 5.4 mmol, 5 eq).

The resin was then **split** for the coupling of randomized (“X”) positions with the library amino acids. The resin was suspended in DMF (50 mL) and carefully divided evenly among HSW Norm-Ject syringes (Torviq) mounted on Restek Resprep SPE vacuum manifolds equipped (Cat 26077) with valves for coupling of each amino acid monomer in the library (i.e., for canonical synthesis: 18 syringes; for noncanonical synthesis 36 syringes).

With the Resprep valves closed, Fmoc-protected amino acids (0.6 mmol, 10 eq relative to resin) in HATU solution (0.38 M in DMF, 1.4 mL, 0.54 mmol, 0.9 eq relative to amino acid) were activated with DIEA (1.2 mmol, 2 eq relative to amino acid) and each added to their respective split resin (theory: ~260 mg resin, 60 μmol). Couplings proceed for one hour minimum. For Fmoc-Bpl-OH, 5.0 equiv. of DIEA relative to amino acid was used. For precious amino acids, lower equivalents were used: Fmoc-Blp-OH (6.6 equiv.), Fmoc-Git(OAc)$_4$-OH (4.7 equiv.), Fmoc-Dpf-OH (3.8 equiv.) and Fmoc-Agn(OAc)$_3$-OH (2.3 equiv.) with extended coupling times up to three hours. After coupling was completed, the Resprep valves were opened to remove the excess coupling solution from the resin.

All resin was then **pooled** into the 250 mL peptide synthesis vessel and the syringes were washed (3 x 5 mL) to recombine all resin. Additional wash (2 x 100 mL) and Fmoc deprotection (1 x 50 mL flow wash followed by 2 x 50 mL, 5 min batch treatments) with 20% piperidine in DMF. Resin was washed with DMF (3 x 100 mL) and was then ready again for the next split cycle. The cycle was iterated 12 times total to accomplish the $X_{12}K\text{-}NH_2$ design.

With the final N-terminal Fmoc group was removed, the resin was washed with DMF (150 mL), then suspended in DMF (~ 50 mL) and divided evenly among 5 aliquots in 20 mL syringes (2 x 10^8 peptides per aliquot). Then each were washed with DCM (3x) and dried under reduced pressure overnight. Resin was taken to perform experiment to validate the quality of the library, see *Library Validation Analysis*.

5.5.4. Cleavage from resin and stock solution preparation

Deacetylation of peracetylated noncanonical side-chains (Agn, Git) was carried out by treatment of resin with a solution of 5% anhydrous hydrazine in DMF for 16 h at ambient temperature. After deacetylation, the resin was washed with DMF (3x), DCM (3x), DMF (3x), MeOH (3x) and DCM (3x) and dried under reduced pressure.

Canonical libraries were globally deprotected and cleaved from resin with 94% (v/v) TFA, 2.5% (v/v) ethanedithiol, 2.5% (v/v) water, and 1.0% (v/v) triisopropylsilane, for 3 h at ambient temperature (~2 mL/mg of resin). Noncanonical libraries were globally deprotected and cleaved from resin with 85% (v/v) TFA, 5% (v/v) water, 5% (v/v) phenol and 5% (v/v) thioanisole for 2 h at ambient temperature (TIPS was found to reduce the GlcNAc of the Agn side chain).

The crude peptides were triturated with cold diethyl ether. Precipitated peptide was triturated (3x) with cold diethyl ether, dissolved in 50% acetonitrile in water (0.1% TFA), passed through a 0.2 µm nylon syringe filter, and lyophilized.

Crude lyophilized powders were resuspended in 5% acetonitrile in water (0.1% TFA) purified using Supelco Discovery® DSC-18 SPE Tubes (Millipore Sigma Cat: 52607-U). The SPE tube was first conditioned with 3 CV of acetonitrile (0.1% TFA) and then equilibrated with 5 CV of 5% acetonitrile in water (0.1% TFA). Then, the suspended crude was loaded (Maximum 150 mg crude peptide loaded onto 2 g bed mass) and washed with 10-12 CV of 5% acetonitrile in water (0.1% TFA). Peptides were eluted with 70% acetonitrile (0.1% TFA) and lyophilized.

Lyophilized, SPE-purified powders of libraries were each dissolved first in DMF and then diluted with 1x PBS to a final library concentration of 8 mM (~40 pM/member), and a final DMF concentration of 5% (v/v). Stock solutions were aliquoted out into low-bind tubes and stored at -80 °C. Aliquots were thawed on ice prior to use.

5.5.5. Library validation analysis

Canonical libraries were validated as previously described.[1] For the noncanonical library, 20 mg of resin was weighed out in a microcentrifuge tube and agitated for 16 h in 5% anhydrous hydrazine in DMF (100 mg/mL). The resin was then transferred to a 3 mL fritted Torviq syringe and washed with DMF (3x), DCM (3x), DMF (3x), MeOH (3x) and DCM (3x). The resin was suspended in DCM and transferred to a 15 mL conical tube and the solvent was evaporated under a stream of nitrogen.

For both the canonical and noncanonical libraries, 1.5 mg of dried resin was weighed out and suspended in DMF (5 mg/mL). From this stock suspension, 1.5 µL (estimated 877 beads) were transferred to a microcentrifuge tube, suspended in 200 µL cleavage solution. Canonical libraries were treated with 94% (v/v) TFA, 2.5% (v/v) ethanedithiol, 2.5% (v/v) water, and 1.0% (v/v) triisopropylsilane and heated to 60 ºC for 10 minutes. Noncanonical libraries were treated with 85% (v/v) TFA, 5% (v/v) water, 5% (v/v) phenol and 5% (v/v) thioanisole) and left at room temperature for 2 hours. The TFA was then evaporated under a stream of nitrogen and the remaining waxy oil was dissolved in 200 µL of 5% acetonitrile in water (0.1% TFA) and sonicated / vortex vigorously. The suspension was centrifuged at 21,300 rcf at room temperature. The supernatant was added onto a conditioned C18 STAGE tip (CDS Empore™ SDB-XC, Fisher Scientific Cat: 13-110-020) and purified according to the protocol of Rappsilber et al.[57] The eluting solvent was evaporated by vacuum centrifugation and the peptides were resuspended in 29 uL of 0.1% formic acid in water to enable the injection of 100 pg/peptide with 1 µL. The solution was centrifuged at 21,300 rcf at 4°C for 10 min and the supernatant was transferred to a MS vial for Orbitrap analysis. Upon analysis of the canonical and noncanonical libraries, the canonical library demonstrated near even monomer incorporation as previously reported.[1] However, within the noncanonical library, higher monomer variation was observed, with Bpl (Fmoc-Bispyridinolysine-OH) and PfF (Fmoc-pentafluoro-L-phenylalanine) showing poor incorporation at all positions. FfF (Fmoc-pentafluoro-L-phenylalanine) has previously been successfully incorporated into other noncanonical libraries. Additionally, the hydrazinolysis of for deacetylation of the glycan-mimetic functional groups (Agn, Git) was suspected to affect the slightly lower incorporation of Psa. Despite these shortcomings in the noncanonical library, it was used in AS-MS experiments as follows.

5.5.6. Affinity selection-mass spectrometry (AS-MS) experiments

Affinity selection-mass spectrometry (AS-MS) was performed manually as previously described with modifications[12] or with a KingFisher^{TM} Duo Prime (Thermo Fisher Scientific).

For manual AS-MS, 100 µL of magnetic beads (1 mg; 0.13 nmol IgG binding capacity, MyOne Streptavidin T1 Dynabeads, Thermo Fisher Scientific Cat: 65602) were transferred to 1.7 mL plastic centrifuge tubes and washed 3 times with blocking buffer (10% fetal bovine serum (FBS) in 1x PBS pH 7.4 and 0.01% Tween20, 0.2 µm filtered) using a magnetic separation rack (NEB Cat: S1506S). Then, 1.2 to 2 eq of biotinylated anti-hemagglutinin antibody (clone 12ca5, Columbia Biosciences Cat: 00-1722) was incubated with the magnetic beads at approximately 0.5 µM. The resulting suspensions were incubated on a nutating mixer for 30 min at 4 ºC and then washed 3 times with blocking buffer.

Next, the affinity selection samples were prepared. The peptide library was depleted of 'bead binders.' In a new tube, the following were combined for a 1mL sample and scaled if needed for multiple replicates using the library: 100 uL of neat FBS, 550 uL of 1x PBS, 250 uL of library stock solution to provide 10 fmol/peptide, and 50 uL of pre-washed magnetic beads. This sample was incubated for 1 hour at 4 ºC. Then, this sample was then centrifuged at 21,300 rcf and the supernatant aliquoted to a new tube to provide the library depleted of peptides that bind to the magnetic beads with high affinity. Then, 1 mg (100 uL volume in blocking buffer) of the washed magnetic beads with 12ca5 immobilized was mixed with the pre-depleted library solution to provide a solution concentration of 100-130 nM of 12ca5 final. These affinity selection samples were then incubated at 4 ºC for 1 hour on a nutating mixer. Then, the samples were washed 3-6 times with cold 1x PBS pH 7.4 using a magnetic rack (~10 minutes contact time with buffer). The isolated beads were eluted using 2 x 100 uL of 6 M guanidine, 50 mM sodium phosphate pH 7.

For automated selections, a KingFisher^{TM} Duo Prime was utilized with two (2) x 96 Deepwell Plates (Thermo Fisher, #95040450) in the following format, marked by rows. Three replicates were run by using three columns per library aliquot for 12 separate $X_{12}K$ libraries. The isolated peptides bound to the beads were eluted using 2 x 100 uL of 6 M guanidine, 50 mM sodium phosphate pH 7 in elution strips.

Table 5.3: *Plate layout for AS-MS using a KingFisher^{TM} Duo Prime system*

	Plate 1		Plate 2	
Row	Description	Vol, mL	Description	Vol, mL
A	Selection samples, see text	1	1x PBS, cold	1
B	Blocking buffer	1	1x PBS, cold	1
C	Blocking buffer	1	1x PBS, cold	1
D	Blocking buffer	1	1x PBS, cold	1
E	Biotinylated 12ca5	0.5	1x PBS, cold	1
F	Blocking buffer	1	1x PBS, cold	1
G	Blocking buffer	1	Comb for Kingfisher magnet	
H	Blocking buffer + beads	1		

	Elution strip 1		Elution strip 2	
Row	Description	Vol, mL	Elution strip 2	Vol, mL
N/A	6 M guanidine, 50 mM sodium phosphate, pH 7	0.1	6 M guanidine, 50 mM sodium phosphate, pH 7	0.1

For the “Selection samples” (Plate 1 Row A), the sample was prepared similarly to the manual selection. First, the peptide library was depleted of ‘bead binders.’ In a new tube, the following were combined for a each sample and scaled if needed for multiple columns / replicates: 100 uL of neat FBS, 550 uL of 1x PBS, 250 uL of library stock solution to provide 10 fmol/peptide, and 50 uL of pre-washed magnetic beads. This sample was incubated for 1 hour at 4 ºC. Then, this sample was then centrifuged at 21,300 rcf and the supernatant aliquoted to the 96 Deepwell plate to provide the library depleted of peptides that bind to the magnetic beads with high affinity.

For "Blocking buffer + beads" (Plate 1 Row H), 100 µL of magnetic beads were added to 900 uL of blocking buffer (10% fetal bovine serum (FBS) in 1x PBS pH 7.4 and 0.01% Tween20, 0.2 µm filtered).

For "Biotinylated 12ca5" (Plate 1 Row E), 500 uL of blocking buffer was added with the amount needed to provide 1.2-2 eq of 12ca5 from its stock solution (typically 10.4 uL of 12ca5 stock solution at 25 µM for 2 eq).

The following steps were programmed for affinity selection:

1. Collect comb from Plate 2, Row G
2. Wash beads by release beads (30 s, medium) in Plate 1, Row H, collect beads (3 x 1 second)
3. Wash beads as in Step 2 in Plate 1, Row G
4. Wash beads as in Step 2 in Plate 1, Row F
5. Release beads (20 s, medium) into Plate 1 Row E (30 minutes, mix slowly)
6. Wash beads as in Step 2 in Plate 1, Row D
7. Wash beads as in Step 2 in Plate 1, Row C
8. Wash beads as in Step 2 in Plate 1, Row B
9. Release beads into Plate 1, Row A, (1 hour, mix slowly)
10. Add plate 2, containing cold 1x PBS to the Kingfisher instrument
11. Collect beads from Plate 1, Row A (5 x 1 second)
12. Wash beads as in Step 2 in Plate 2, Row A
13. Wash beads as in Step 2 in Plate 2, Row B
14. Wash beads as in Step 2 in Plate 2, Row C
15. Wash beads as in Step 2 in Plate 2, Row D
16. Wash beads as in Step 2 in Plate 2, Row E
17. Wash beads as in Step 2 in Plate 2, Row F
18. Release beads into elution strip 1, 1 minute mix fast, collect beads (5 x 1 s)
19. Release beads into elution strip 2, 1 minute mix fast, collect beads (5 x 1 s)
20. Release beads and comb into Plate 2 Row G to end the program

Eluted peptide samples were then prepared for Orbitrap analysis by C18 STAGE tip (CDS Empore™ SDB-XC, Fisher Scientific Cat: 13-110-020) and purified according to the protocol of Rappsilber et al.[57] The eluting solvent was evaporated by vacuum centrifugation and the peptides were resuspended in 12-13 uL of 0.1% formic acid in water. The solution was centrifuged at 21,300 rcf at 4°C for 10 min and the supernatant was transferred (leave behind 1.5 uL) to a MS vial for Orbitrap analysis. Usually, 4-5 uL were injected onto the Orbitrap Fusion Lumos whereas 2-3 uL were injected onto the Orbitrap Eclipse.

5.5.7. Nanoscale liquid chromatography-tandem mass spectrometry (nLC-MS/MS)

Nanoscale liquid chromatography tandem mass spectrometry (nLC-MS/MS) was performed using an EASY-nLC 1200 (Thermo Fisher Scientific) nano-liquid chromatography handling system connected to an Orbitrap Fusion Lumos or an Orbitrap Eclipse Tribrid Mass Spectrometer (Thermo Fisher Scientific). Solvent A is water (0.1% formic acid) and solvent B is 80% acetonitrile in water (0.1% formic acid). Precolumn and analytical column equilibration with 8 µL of solvent A was performed at maximum of 1 µL/min or 600 bar. Samples were injected and loaded onto a nanoViper Trap Column (C18, 3 µm particle size, 100 A pore size, 20 mm x 75 µm ID; Thermo Fisher Scientific, Cat: 164946) for desalting with 12 µL of solvent A (maximum of 1 µL/min or 600 bar). The autosampler wash was 100 uL of solvent A. After trapping, samples were injected onto a PepMap RSLC C18 column (2 µm particle size, 15 cm x 50 µm ID; Thermo Fisher Scientific, Cat: ES901). The standard nano-LC method was run at 40 °C and a flow rate of 300 nL/min with the following gradient, expressed in % solvent B in solvent A: 1% to 41% over 120 minutes (*AS-MS Experiments*) or 90 minutes (*Library Validation Analysis* or other simple mixtures), move to 90% in 3 minutes, hold for 7 minutes, and then perform 2 “seesaw” washes (each comprising of moving to 20% over 3 minutes, holding at 20% for 3 minutes, moving to 90% for 3 minutes, and holding at 90% for 3 minutes).

Mass spectrometry acquisition was performed using an Orbitrap Fusion Lumos or an Orbitrap Eclipse Tribrid Mass Spectrometer (Thermo Fisher Scientific) with positive mode, where the ion source settings was set by the tune parameters (Spray voltage usually ~ 2200 V with no Arb gas). The method to perform data-dependent acquisition has been iteratively optimized.

The standard AS-MS MS analysis method analyzes from 3-120 minutes, with an expected LC peak width of 20 seconds, default charge state of 3, and no internal mass calibration. Primary spectra acquisition in positive mode was observed by the

Orbitrap with resolution = 120,000, using quadrupole isolation, 200-1400 m/z, RF Lens 30%, 250% AGC Target (auto injection time, usually < 10 ms), and 1 microscan. Secondary MS was performed with the following filters: Precursor selection range: 300-1200 m/z, MIPS: Peptide, Intensity threshold: 4e4, Charge state: 2-5 excluding undetermined charge states, Dynamic exclusion: exclude after 1 time for 30 seconds (10 ppm tolerance), Targeted mass exclusion of all peptides in the Pierce™ Peptide Retention Time Calibration Mixture (z = 2 and 3, Thermo Fisher Scientific, Cat: 88321). HCD and EThcD were completed. HCD used quadrupole isolation (1.3 m/z, no offset) at a fixed 28% collision energy and was observed on the Orbitrap with resolution = 30,000, Scan Range Mode: Define First Mass: 120 m/z, 600% AGC Target, maximum injection time 100 ms, and 2 microscans. EThcD used a charge filter of z ≥ 3, quadrupole isolation (1.3 m/z, no offset), using calibrated charge-dependent ETD activation, and supplemental HCD activation a fixed 25% collision energy and was observed on the Orbitrap with resolution = 30,000, Scan Range Mode: Define First Mass: 120 m/z, 600% AGC Target, maximum injection time 100 ms, and 2 microscans.

5.5.8. Curation of AS-MS data

De novo analysis of sequencing data was performed as described previously for canonical libraries using PEAKS Studio 8.5 (Bioinformatics Solutions, Inc, ON, Canada).[12] Mass precursor correction was used. Auto *de novo* sequencing was performed using a 15 ppm precursor mass error and 0.02 Da fragment mass error. For canonical libraries, the following PTM modifications were used: fixed C-terminal amidation (-.98 Da) on lysine, and variable oxidation on methionine (+15.99 Da). For noncanonical libraries, the PTMs used are shown in Table 5.4. 20 candidate sequences were obtained for each preprocessed scan. Post-*de novo* data analysis was performed as previously described[58] to convert the PTMs to 1-letter encoding also in Table 5.4.

Table 5.4: *Post-translational modification (PTM) utilized in PEAKS de novo sequencing analysis of noncanonical library. Where a single amino acid is modified (e.g., F modified to be F(+17.99) to represent 3fF), a fixed PTM is used. When the same amino acid can be modified to represent multiple noncanonical amino acids (e.g., alanine), a variable PTM was used.*

Monomer	**PTM**	**1-letter code**
Aze	A(+12.00)	a
Aib	A(+14.02)	b
Cpa	A(+40.03)	d
Hyp	A(+42.01)	e
Pip	A(+55.04)	f
Thp	A(+56.03)	g
Amb	A(+62.02)	h
Tha	A(+82.98)	i
Tic	A(+88.03)	j
4AF	A(+91.04)	k
Msn	M(+31.99)	l
3fF	F(+17.99)	m
pSer	S(+79.97)	n
hArg	R(+14.02)	o
hCit	N(+57.06)	p
hCit	A(+100.06)	c
DfF	C(+80.04)	r
Php	E(+58.06)	s
CxF	A(+120.02)	t
Nal	L(+84.00)	u
2F3F	A(+144.02)	v
DPh	W(+37.02)	w
Psa	A(+155.00)	x
PfF	A(+165.98)	y
Dpf	A(+226.08)	z
Bpl	A(+239.14)	B
Agn	A(+246.09)	X
Git	A(+248.10)	Z

After concatenating all data from *de novo* sequencing, the data was rigorously cleaned to remove poorly sequenced peptides and sequence isomers from the data, beyond what has previously been published.[58]

First, simple filters on the average local confidence of sequencing (ALC) and calculated ppm error of sequencing from PEAKS Studio 8.5 were applied: ALC > 85 (canonical) or > 80 (noncanonical) and absolute ppm error < 10 ppm were retained. Also, all duplicate peptides were removed.

Second, all sequences were compared pairwise and marked for removal if they had the same precursor mass within 0.01 Da or had specific differences in precursor mass corresponding to 1) incorrect monoisotopic precursor selection (absolute delta of 1, 2, or 3 Da), oxidation (absolute delta of 16, 32), or sodium adduct (absolute delta of 22). Additionally, the peptides must have some amount of sequence similarity (empirically seen to work well on trial datasets with a similarity of 0.69 by difflib.SequenceMatcher in Python). Retention time differences were not considered in case the data was acquired using different gradients. The highest ALC peptide was retained, with the lowest ppm sequencing error as tie-breaker.

Third, all remaining sequences were compared pairwise and marked for removal based only on a very high degree of sequence similarity. Again using difflib.SequenceMatcher in Python, a peptide similarity of > 0.92 was only seen for sequence isomers with either a single amino acid replacement or a dipeptide swap with the X12K type of peptides. While rigorous and potentially overly conservative, this step often removes < 5% of the remaining data after the second step is completed.

With the canonical library, 4104 peptides were uniquely identified from AS-MS with high sequencing fidelity for unsupervised learning analysis.

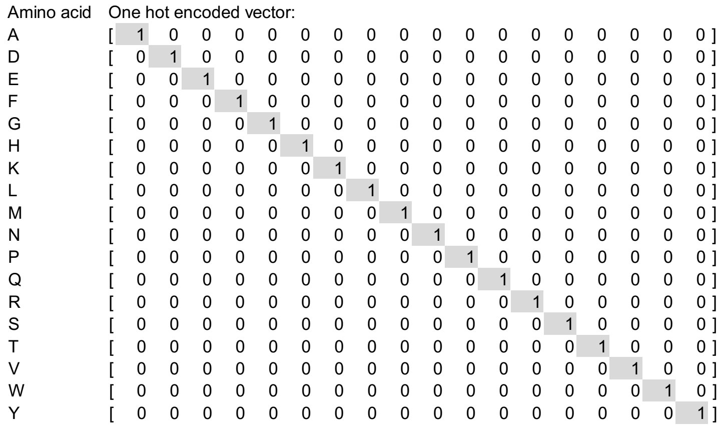
With the noncanonical library, 17 peptides were uniquely identified from AS-MS with high sequencing fidelity for unsupervised learning analysis.

5.5.9. Encoding of peptides for unsupervised analysis

5.5.9.1. One-hot encoding

Each amino acid was represented by the vectors seen below. A peptide was represented by concatenating these vectors together. Thus, each peptide was represented by a **vector 12 * 20 = 240 in length vector descriptor for each peptide**.

Table 5.5. *One-hot encoding vectors for canonical amino acids*



Amino acid	One hot encoded vector:																			
A	[	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	]
D	[	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	]
E	[	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	]
F	[	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	]
G	[	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	]
H	[	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	]
K	[	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	]
L	[	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	]
M	[	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	]
N	[	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	]
P	[	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	]
Q	[	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	]
R	[	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	]
S	[	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	]
T	[	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	]
V	[	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	]
W	[	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	]
Y	[	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	]

5.5.9.2. Physicochemical encoding

Each amino acid was represented by 12 physicochemical properties as reported from literature.[54]

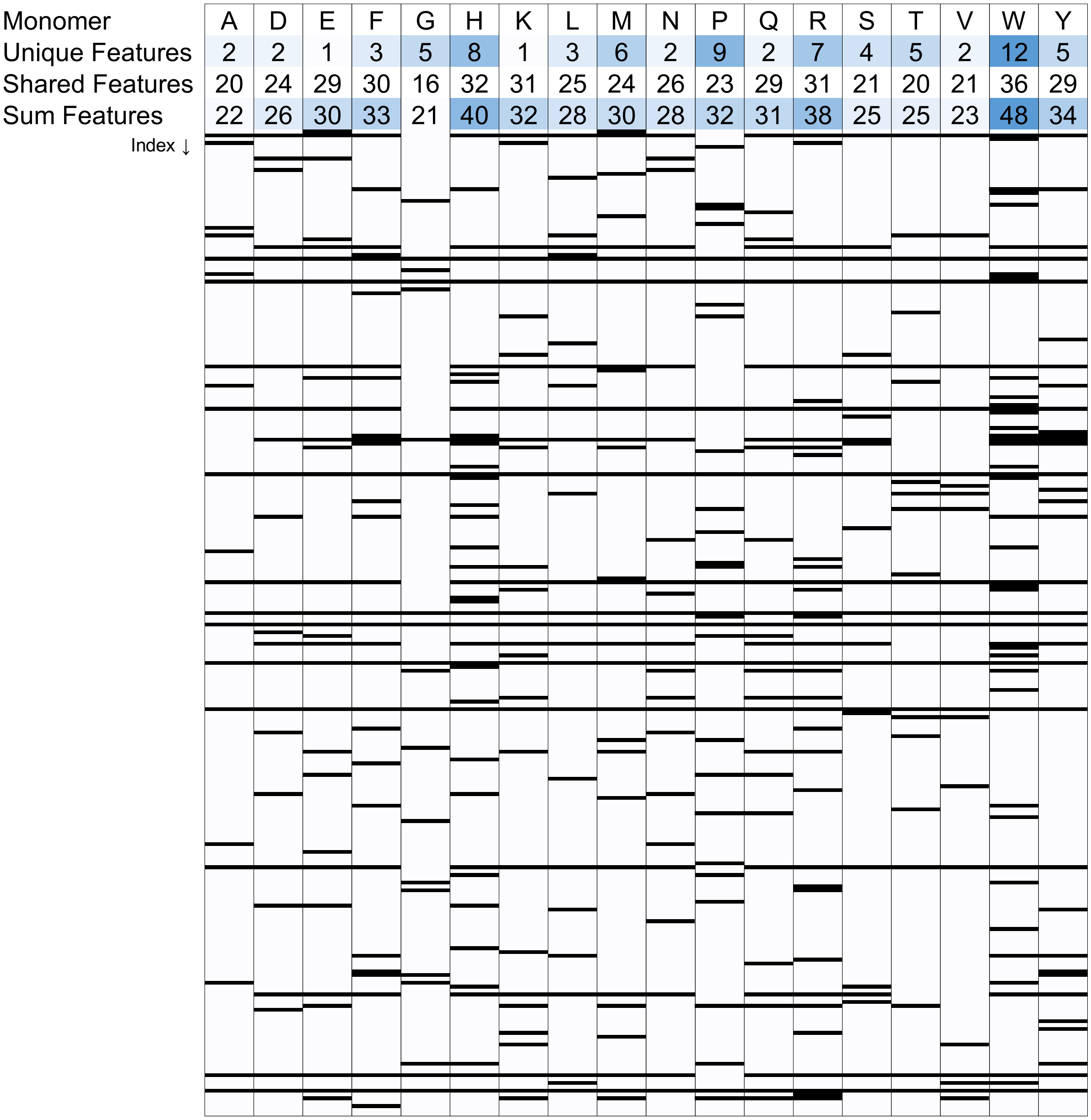
The reported properties were standardized before use. These properties included H11 and H12: hydrophobicity; H2: hydrophilicity; NCI: net charge index of side chains; P11 and P12: polarity; P2: polarizability; SASA: solvent-accessible surface area; V: volume of side chains; F: flexibility; A1: accessibility; E: exposed; T: turns; A2: antigenic. Hydrophobicity (H11 and H12) and polarity (P11 and P12) were calculated using two methods. The peptide was represented by concatenating the vectors of each amino acid together (**12 residues * 12 properties = 144 length vector descriptor for each peptide**)

5.5.9.3. ESM-2 encoding

ESM-2 is a protein language model that can be used for multiple applications where properties, structure, and function are derived from the input sequence, where the model was trained on the proteome (UniRef 50). Encoding was completed by extracting the amino acid embeddings of the peptides from 33rd layer of the pretrained “esm2_t33_650M_UR50D” model. From this layer, each embedding per amino acid is size 1280, and a peptide is represented by concatenating this output residue by residue, resulting in a **12 residues * 1280 sized embedding = 15,360 length vector descriptor for each peptide**. While this can seem large, N-grams encoding was also on this order of magnitude.

5.5.9.4. Fingerprint encoding

Extended connectivity Fingerprint encoding was used with bit-vectors of 256 length and radius = 3. Canonical and noncanonical amino acids were drawn in ChemDraw 21.0.0 with N-acetylation and N-methyl carboxamidation to replicate the featured of the amino acid integrated within a peptide. Histidine was drawn in its most common τ-tautomer form. Amino acids were exported as SMILES and canonicalized (standardized) in using molvs (standardize_smiles). The Fingerprint was the isolated using Chem.GetMorganFingerprintAsBitVect and Chem.MolFromSmiles. With an n-bit vector of 256, each peptide was represented as **12 residues * 256 bit-vector length = 3,072 length vector descriptor for each peptide**



Monomer	A	D	E	F	G	H	K	L	M	N	P	Q	R	S	T	V	W	Y
Unique Features	2	2	1	3	5	8	1	3	6	2	9	2	7	4	5	2	12	5
Shared Features	20	24	29	30	16	32	31	25	24	26	23	29	31	21	20	21	36	29
Sum Features	22	26	30	33	21	40	32	28	30	28	32	31	38	25	25	23	48	34

Figure 5.6: *The Fingerprint encoding illustrates the similarities and number of unique features in canonical amino acids. Specifically, one can see the similarity in specific substructure features between amino acids, as well as the number of unique features.*

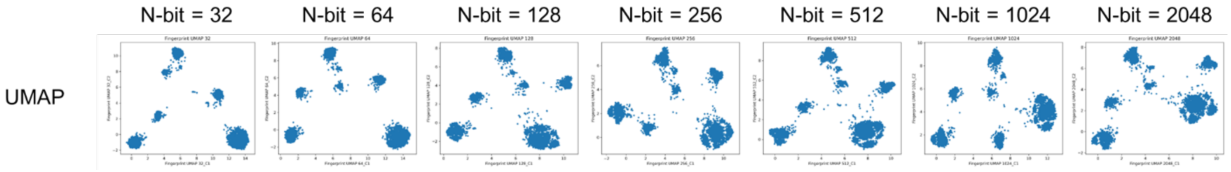


Figure 5.7: *The Fingerprint radius of 3 is generally set for extended connectivity Fingerprint encoding for ECFP_6. However, the bit-vector length can and was varied to see if it*

affected the data ranging from 2^5 (32 bit vector length) - 2^{11} (2048 bit vector length). Low bit-vector length minimized the appearance of distinct clusters in some analyses of the AS-MS data (e.g., UMAP Fingerprint shown above). The bit-vector length of 256 length was seen to provide more distinct clusters within some of the sequence maps, and above this value, no additional resolution was seen.

5.5.9.5. N-grams encoding

N-grams encoding was completed by pre-calculating the observed n-mers in the dataset up to a maximum n-mer length of the full peptide length (12 residues), as described below. As pre-calculated (Figure 5.7), the entire **peptide was represented at once as a 138,622 length vector**, where each index of the vector describes an n-mer motif that is either present (1) or absent (0) in the peptide.

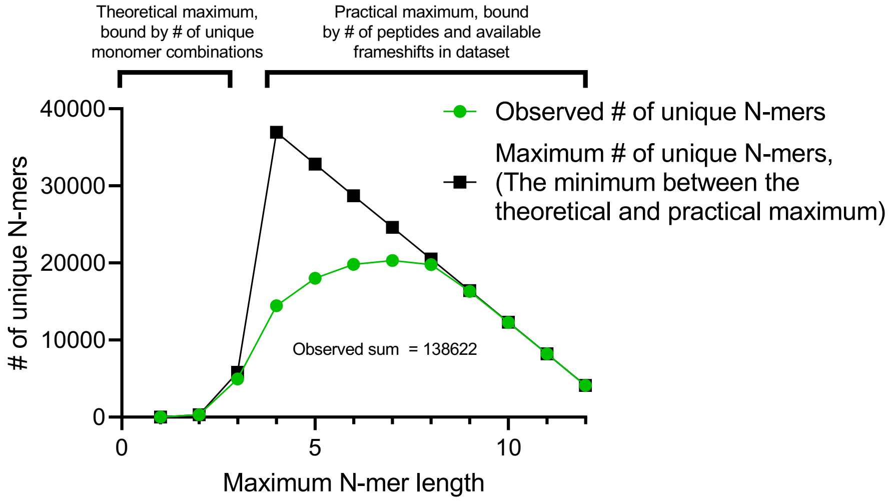


Figure 5.8: *The number of unique N-grams for encoding versus the maximum N-gram length used. N-Grams encoding proceeds first by predetermining all n-mers (sometimes called k-mers) within the dataset. The theoretical number of n-mers is bound by the number of unique combinations of monomers and the maximum N-gram length (i.e., [# of monomers]Maximum N-gram length), which up to a 12-mer length peptide would be 1015 n-mers. However, since the n-mer space is pre-calculated from the dataset, significantly fewer are actually observed than theoretically possible even with the maximum N-gram length set to the length of the peptides in the library. The practical maximum is the observed n-mers, bound by (# of peptides) x [1 + (Full Peptide Length – Maximum N-gram length)]. The true maximum is the minimum of the theoretical and practical maximum shown in the figure above in green.*

5.5.10. UMAP dimensionality reduction hyperparameter optimization

UMAP is a user-friendly, non-linear dimensionality reduction technique that requires minimal optimization to use. However, UMAP embedding results are generally stochastic. Thus the random seed state was always fixed. Some variation in the embeddings was noticed due to the UMAP version, which was 0.5.3 for this work. Lastly, UMAP embeddings are affected by the order of the data within the datafile used (see UMAP shuffle samples leads to quit different result · Issue #268 · lmcinnes/umap) likely because data seen first is weighted more in the initialization of the manifold. Thus, the sequences from AS-MS were randomly shuffled, and then used throughout this work. Additionally, we have observed that exact embedding results can vary from computer to computer but should remain generally similar.

The two main hyperparameters are n_neighbors and min_dist, and the distance metric setting.

First, n_neighbors balances the importance of the local vs global structure within the data. Low n_neighbors values (~1% of the dataset size) will provide results that focus on local structures, while large values seek to emphasize the global structures, losing fine local detail. This is observed by producing the UMAP embeddings versus n_neighbors (Figure 5.9).

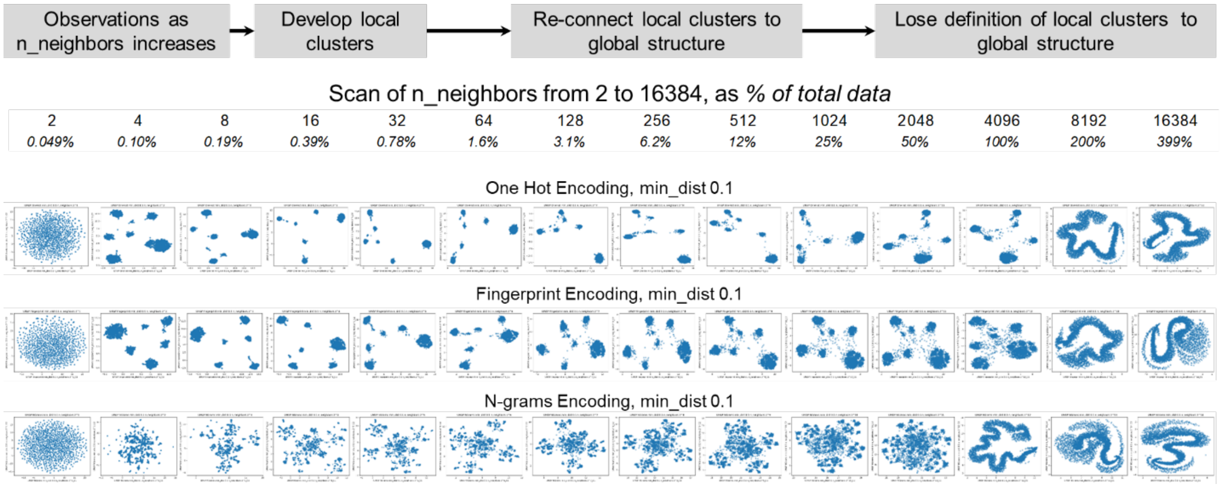


Figure 5.9: *Scan of n_neighbors with UMAP using one-hot, Fingerprint, and N-grams encoding. Local clusters are rapidly and initially developed. As n_neighbors increases, local clusters are reconnected to the global structure of the data at an optimal n_neighbors. As n_neighbors grows to a significant percentage of the dataset (> 50%), the clusters begin to be obscured in the global structure unifying the peptides. Stable embeddings results were seen at n_neighbors throughout the dataset from (1.5 - 25%), so 6.2% (n_neighbors = 256) was taken as an optimal value.*

Second, min_dist sets the minimum distance between points, meaning that tight local clusters are forced to be spread apart. The default of 0.1 was used for all analysis except for one-hot encoding, which showed exceptionally tight clusters, and so it was set to 0.4.

The distance metric was appropriately set based on the encoding type:[55] binary encoding method (one-hot, Fingerprint, and N-grams) used the Tanimoto distance metric, while continuous descriptors (evolutionarily-learned and physicochemical encoding) used the Euclidean distance metric.

5.5.11. Multidimensional scaling (MDS) dimensionality reduction

Multidimensional scaling (MDS)[56] was used as the similarity mapping method. However, it is currently unable to incorporate additional results without re-learning. Thus, the dataset of randomly sampled peptides could not be added as it would cause MDS to learn over random sequence space combined with the AS-MS discovered space. Specifically, MDS does not have a .transform function in the current version used (scikit-learn, version 1.0.2), see https://github.com/scikit-learn/scikit-learn/issues/2887, and https://github.com/scikit-learn/scikit-learn/issues/15808 .

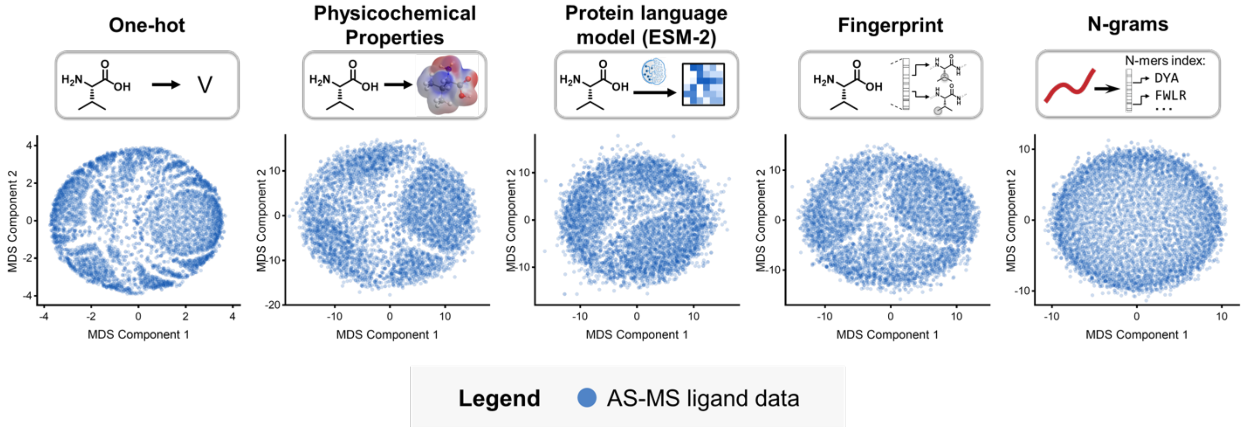


Figure 5.10: *MDS dimensionality reduction versus encoding method of the AS-MS data.*

5.5.12. Individual peptide synthesis and cleavage from resin

Peptides and peptidomimetic α-carboxamides were manually synthesized in batch using 100 mg of H-Rink Amide ChemMatrix resin (0.49 mmol/g). Resin was swollen in amine-free DMF for a minimum of 10 minutes in HSW Norm-Ject syringe (Torviq) syringes mounted on a Restek Resprep SPE vacuum manifolds equipped (Cat 26077) with valves. For each coupling cycle, Fmoc-protected amino acids (5 eq, 0.245 mmol) were dissolved at 0.4 M in 0.38 M HATU (4.75 eq relative to resin, 0.95 eq relative to Fmoc-protected amino acid) in amine free DMF and sonicated or vortexed as needed. Diisopropylethyl amine (DIEA; 10 eq, 0.49 mmol, 85.4 μL) was added and the solution, hand mixed to form the active ester, and confirmed to return being visually transparent as a clear light yellow solution. Using the Restek manifold, the excess DMF was drained from the DMF-swelled resin. Then the solution containing the activated Fmoc-amino acid ester was added to the resin and incubated at room temperature for 45 minutes. After which, the resin was drained and washed 3 x with amine free DMF. Fmoc deprotection was completed using 20% piperidine in DMF (2 x 5 minutes), and then washed 3 x with amine free DMF. Then the next amino acid coupling cycle could proceed. After synthesis was complete, resins were washed 5 x with amine free DMF, 3 x DCM, vacuum was pulled on the dry resin to remove the DCM (5 minutes), and then the resin was dried under vacuum before cleavage.

Cleavage was performed in HSW Norm-Ject syringe (Torviq) syringes by using the syringe plunger to pull the cleavage solution onto the resin with a blunt tip needle and then capping the syringe. Global side chain deprotection and cleavage from solid support were carried out using solution of 94% (v/v) TFA, 2.5% (v/v) ethanedithiol, 2.5% (v/v) water, and 1.0% (v/v) triisopropylsilane, for 1 hour minimum at ambient temperature (~2 mL of deprotection solution / 100 mg of resin). Upon which, the crude peptide and cleavage solution was isolated from the syringe into a 15 mL Falcon tube and triturated with cold diethyl ether (~12 mL, chilled on dry ice). The peptide was then suspended in 50% acetonitrile in water (0.1% TFA) and lyophilized.

Peptide purification was completed using reverse-phase flash purification or with preparative high performance liquid chromatography purification (HPLC). For flash purification, a Biotage Selekt was used with a Biotage® Sfär C18 D - Duo 100 Å 30 µm 12 g column. One-third of the cleaved, lyophilized peptide mass (< 10 mg) was suspended in 0.9 to 1.8 mL of 20% MeCN in Water (0.1% TFA), centrifuged at 3.4k rcf for 10 minutes, and the supernatant was loaded onto the column and separated using using a gradient of 10% to 55% MeCN in Water (0.1% TFA) over 12-15 column volumes (CVs) and observed by UV absorption at 210 and 280 nm and fraction collected with 3 mL maximum fraction sizes. Peptides that exhibited close elution to deletion products or poor elution profiles were purified by preparative HPLC. Preparative HPLC was performed on an Agilent 1260 Infinity LC equipped with a 6130 single quadrupole mass spectrometer. Samples were prepared as described above, filtered using a 0.2 µm filter, and loaded onto a Zorbax 300SB C18 column (9.4 x 150 mm, 5 µm, 8 mL/min) with a C8 guard column using a automated injector and separated using 5% to 55% MeCN in Water

(0.1% TFA) over 30 minutes with fractionation over the entire run using 62 fractions. Fractions were analyzed by LCMS and UPLC to assess purity.

5.5.13. Liquid chromatography-mass spectrometry (LC-MS) analysis

LC-MS analysis was acquired using an Agilent 6550 MS Q-TOF mass spectrometer with Dual Agilent Jet Stream (AJS) ESI ion source in extended dynamic mode in mass range 100 - 3000 m/z with scan rate of 1.00 spectra/sec. An isopump delivered a reference ion mass (922.0098 m/z). The following instrument parameters were used: gas temperature 200 ºC, gas flow 14 L/min, nebulizer pressure 55 psig, sheath gas temperature 350 ºC, sheath gas flow 11 L/min. The following scan source parameters were used: VCap: 3500, nozzle voltage 1000 V, fragmentor 175, and Octopole RF Vpp 750. Column was a Zorbax 300SB C3, 2.1 × 150 mm, 5 µm kept at 40 ºC. The gradient utilized 0.1% formic acid in water (solvent A) and 0.1% formic acid in acetonitrile (solvent B), flow rate 0.5 mL/min, starting at 1% B in A running to 91% B in A over 7 minutes with 1 minute at 91% B in A and 1 minute post-time re-equilibration at 1% B in A. Data were analyzed in Agilent MassHunter Qualitative Analysis B.06.00.

5.5.14. Purity analysis by ultra performance liquid chromatography (UPLC)

LC analysis was performed with an Agilent 1260 LC system controlled by ChemStation software, using an Agilent Zorbax RRHD 300SB-C18, 2.1 x 50 mm, 1.8 µm (Cat: 857750-902) column at 40 ºC. The gradient utilized 0.1% trifluoroacetic acid (TFA) in water (solvent A) and 0.1% TFA in acetonitrile (solvent B). The flow rate was 0.5 mL/min, starting at 5% B in A running to 65% B in A over 11 minutes, moving to 90% B in A in 0.25 minute, holding for 1 minute, moving to 5% B in A in 0.05 minute, and re-equilibrating for 1.5 minutes. Approximately 1-10 ug of each peptide was injected for analysis for a target response of <1000 mAU. The absorbance at 214 nm was recorded and integrated using ChemStation software to report the purity relative to an equal volume injection of 50% acetonitrile in water.

5.5.15. Biolayer interferometry (BLI) measurements

Ideally, proteins including 12ca5 would be immobilized and dipped into solutions of the peptides to test their binding activity. This immobilization orientation is preferred because it would use the same biotinylated 12ca5 used in AS-MS in the same orientation and avoid potential avidity affects. However, when immobilizing 12ca5 onto the BLI tip, insufficient signal was observed when dipping into solutions of known peptide binders. This lack of signal was attributed to the relatively small size of these peptides (e.g., ~2 kDa HA tag) to the size of the immobilized 12ca5 (~150 kDa). Thus, biotinylated peptides were prepared using a resin preloaded with GGSK(Biotin). To avoid avidity effects and use a 1:1 model, the ligand density of the immobilized biotinylated peptide or peptidomimetic on the BLI tip was immobilized slowly (over 300 s) up to ≤ 60% of saturation level.

BLI was carried out using the GatorBio GatorPlus Label-Free Analysis system using Greiner Bio-One 96-well Non-treated Black Polypropylene Microplates (FisherSci Cat 07-000-110) using Streptavidin (SA) Probes (GatorBio Cat 160002). All well solution conditions were prepared using kinetics buffer (K Buffer, 0.02% BSA and 0.02% Tween20 in 1x PBS pH 7.4, 0.2 μm filtered). SA tips were equilibrated in K Buffer for 15 minutes prior to analysis. Plate temperature was set to 30 °C with agitation speed at 1000 rpm during measurement and 200 μL well volumes were used.

During each run, sensor tips were equilibrated K buffer (120 seconds), then dipped into of 50–500 nM biotinylated peptide solution for peptides immobilization (300 seconds), with an additional well with no peptide as a control. Concentrations of the peptide immobilization solutions were surveyed beforehand and adjusted such that the peptide response signal (nm) arrived at 60% or less of its saturation level during 300 seconds of immobilization. This extra step was done to appropriately load the tip to minimize avidity effects during downstream association per manufacturer recommendation. Once loaded with peptides, the tips were then moved into wells containing various concentrations of 12ca5 (nonbiotinylated) for association measurement, with an additional well corresponding to a sensor tip with immobilized peptide with no protein as a control. After association (300 seconds), the tips were moved to a well with K buffer to obtain the dissociation (600 seconds). Peptide-only and protein-only conditions (concentration at 1000 nM) were used as references for background subtraction. The association and dissociation curves were fitted with the GatorOne Software (v 2.7.3.1013) using a 1:1 binding model (n ≥ 3 fit curves accepted with Full $R^2 > 0.8$ and $X^2 < 32$, see Table 5.19) to calculate the apparent dissociation constant (K_D, reported as the average of the fits ± standard deviation of the fits).

5.6. Code availability

Data supporting the findings of this work are available within following appendices and precending materials and methods sections, which provides additional information on the preparation of synthetic split-pool peptide and peptidomimetic libraries; AS-MS and nLC-MS/MS experiment protocols; details on the encoding and dimensionality reduction methods; report of all consensus, centroid, and logo plots for all clusters; comparison of our clustering method to perform motif detection versus the MEME suite; as well as peptidomimetic synthesis, purification, and verification. All data utilized in this work is available at https://github.com/josephsbrown1/Peptide-Map/

5.7. Appendix I: Clustering Information by Dimensionality Reduction Methods

5.7.1. Characteristics of the peptides sampled from the original peptide libraries (presumed to be nonbinders)

From the library validation analysis of the canonical library, 5,047 peptides were identified by sampling the original library before AS-MS. In all cases except the sensitivity analysis in Figure 5.3, these peptides were added to PCA- and UMAP-constructed maps without re-learning. MDS is unable to add additional data to its sequence map without re-learning.

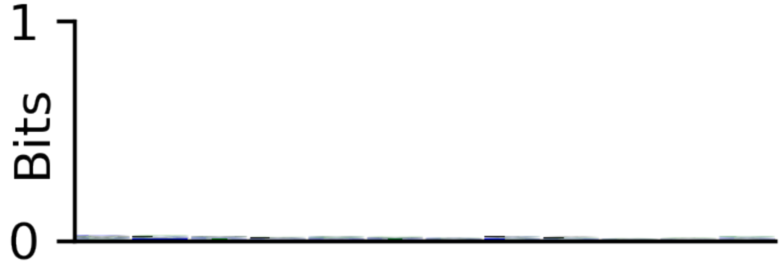


Figure 5.11: *Logo plot of the peptides sampled from the X12K library, presumed to be nonbinders. Essentially no residues are shown, even at this zoomed y-scale, meaning that the peptides are largely random. This is corroborated by the unsupervised clustering seen during the motif detection testing in Section 5.7.5, where the library peptides largely show a diffuse sequence space when the AS-MS ligand dataset is not added.*

5.7.2. Label definitions for 12ca5-specific and nonspecific binders

From the curated AS-MS data,12ca5-specific peptides are defined as *D..DYA* or *D..DYS* from the motif known in literature.[47,59] Note that "*" is a variable length wildcard, while "." is a single amino acid length wildcard.

Care was taken in defining nonspecific binders. From the full dataset, all *D..DYA* or *D..DYS* sequences were removed. Also, all possible mis-sequenced isobaric dipeptides based on of the D**DYA or D**DYS motif were removed. Isobaric was defined as within 10 ppm to match the *de novo* sequencing error tolerance. Sequences containing *DYA*, *DYS* and the commonly observed *PDY*, and *EDY* motifs, gapped isomers (e.g., *D.YA* and *D..YA*), and their dipeptide sequence isomers were removed for consideration as nonspecific binders. Lastly, sequence containing *D.D*, *D..D*, and *D…D* were also removed for consideration as nonspecific binders

All other sequences that were not considered 12ca5-specific or nonspecific were labeled as unknown.

Table 5.6: *Number of peptides manually assigned in each class as defined in Label definitions for 12ca5-specific and nonspecific binders.*

12ca5-specific	Nonspecific	Unknown	Total
3512	139	453	4014

5.7.3. All dimensionality reduction results with manually added common motif labels

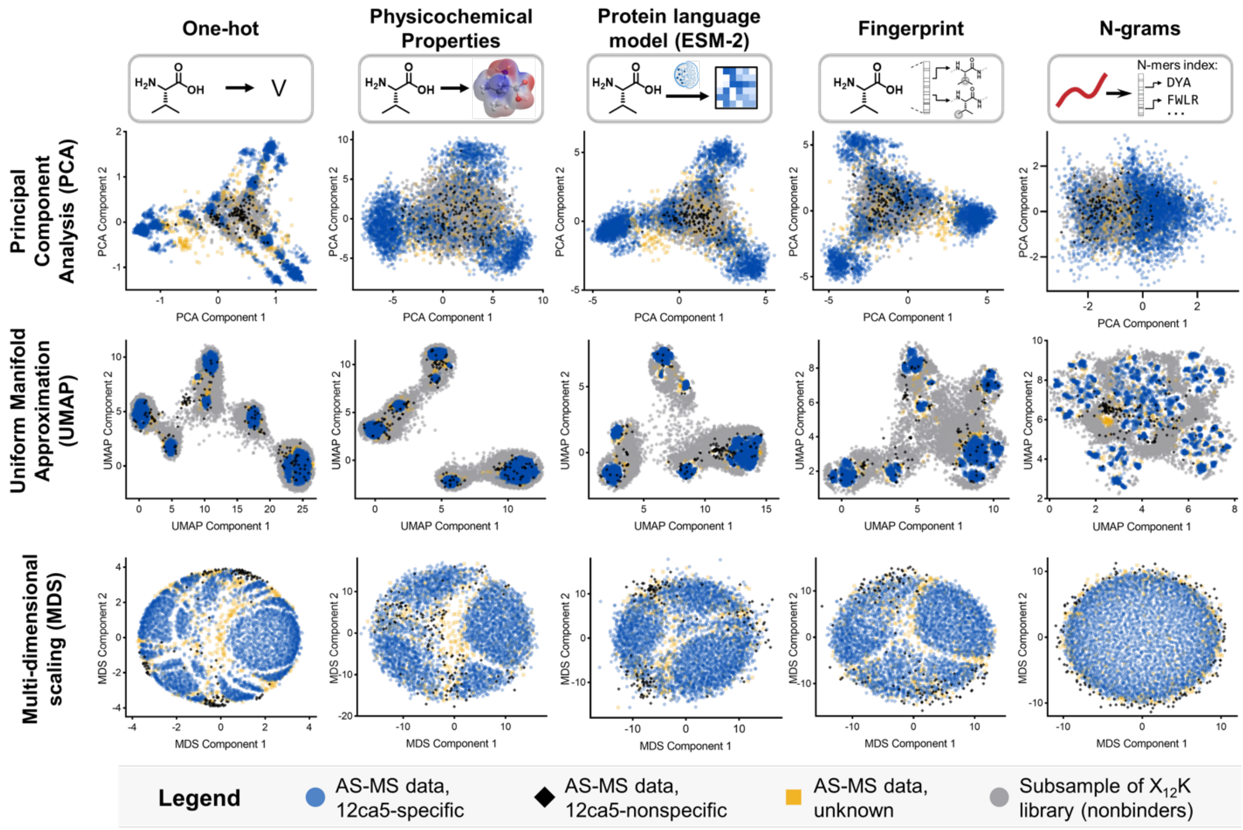


Figure 5.12: *All dimensionality reduction results using all representation encodings with manually added common motif labels as described in Section 5.7.2: Label definitions for 12ca5-specific and nonspecific binders.*

5.7.4. Information about all clusters from dimensionality reduction

Every report here on each combination of encoding and dimensionality reduction technique has the following:

1. The sequence map shown in the Main Text, with the manually categorized color-coded labels:

    a. **Common Motif in blue**, defined as D**DYA or D**DYS, where * is a single-character wildcard at any frameshift within a peptide,
    b. **Expanded Motif in orange**, defined as any reported motif that expands, deviates, or adds additional definition to the Common Motif, or
    c. **Weak in gray**, displays a weak signal, no clear motif.
2. The same sequence map with its respective automatous labels.
3. If any expanded motifs are observed in the analysis, a large plot reporting the centroid peptide from each cluster. While a single centroid peptide is reported here, the option is available to report more centroid peptides spread throughout the cluster.
4. A table of all information about each cluster including:
    a. Main text cluster number, if applicable
    b. Autonomously assigned cluster number
    c. The number of peptides in each cluster
    d. One centroid sequence. More centroids can be reported interspersed within each cluster.
    e. Consensus sequence, determined from each cluster with the requirement that the amino acid position shown must be present 33% or more in all of the peptides in the cluster, otherwise X.
    f. Logo of the cluster to infer Consensus sequence and Motif class, prepared using Logomaker.[60]
    g. Motif Class, assigned manually by inspecting the Logo.

Table 5.7: *Report of automated cluster detection algorithm and parameters used from scikit-learn with either Agglomerative Clustering (AggCl) or Density-Based Spatial Clustering of Applications with Noise (DBSCAN).[51] The parameters used and reported here were found by scanning the parameters and inspecting the results.*

Dimensionality Reduction Method	Encoding Method	Algorithm	eps	min_samples	# clusters observed
PCA	One-hot	AggCl		31	31
	Physicochemical	AggCl		5	5
	ESM-2	AggCl		6	6
	Fingerprint	AggCl		6	6
	N-grams	AggCl		2	2
UMAP	One-hot	DBSCAN	0.21	10	8
	Physicochemical	DBSCAN	0.21	10	7
	ESM-2	DBSCAN	0.1446	23	16
	Fingerprint	DBSCAN	0.1125	15	19
	N-grams	DBSCAN	0.1022	16	67

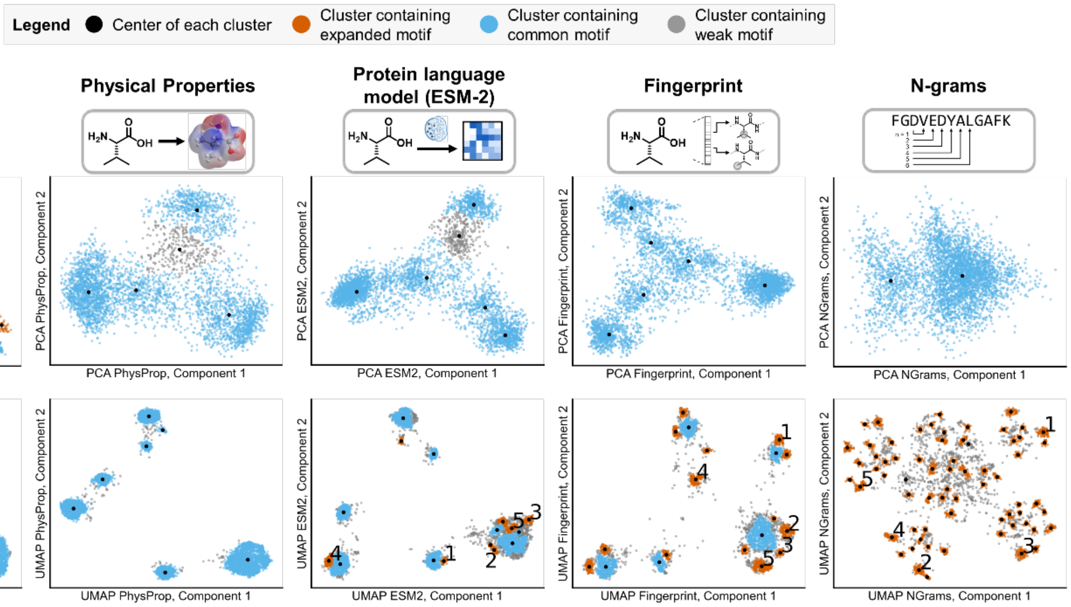


Figure 5.13: *Summary of analyzing the motif of each cluster across all encoding and dimensionality reduction techniques. All sequence maps are shown, with the color-coded labels based on motif class. Motif class was manually categorized as Common Motif in blue, defined as D**DYA or D**DYS, where * is a single-character wildcard at any frameshift within a peptide Expanded Motif in orange, defined as any reported motif that expands, deviates, or adds additional definition to the Common Motif, or Weak in gray, displays a weak signal, no clear motif. Note that no cluster information is available to multi-dimensional scaling as the clusters had little-to-no definition and could not be detected well with DBSCAN or Agglomerative clustering.*

5.7.4.1. PCA, one-hot encoding cluster information

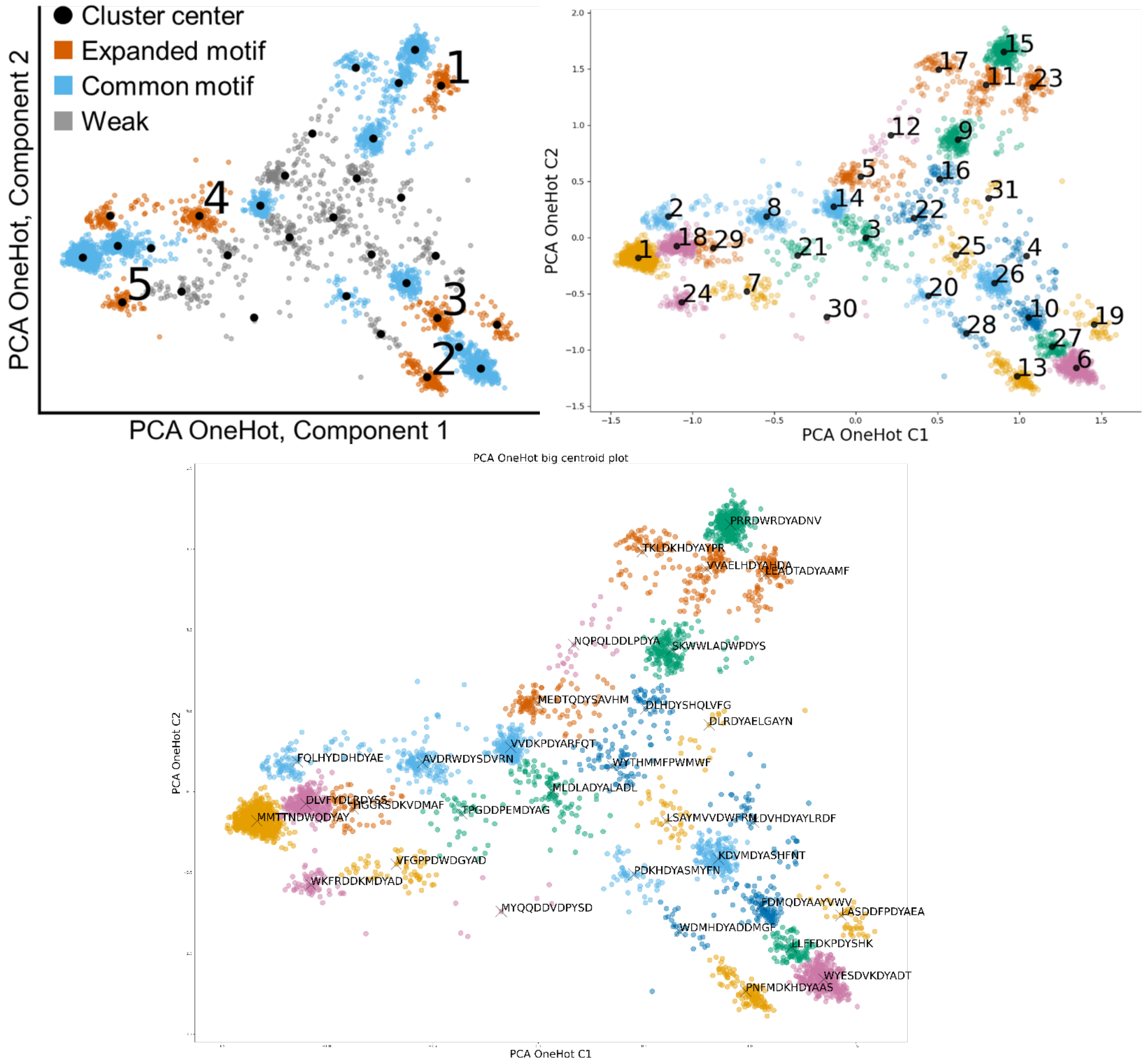
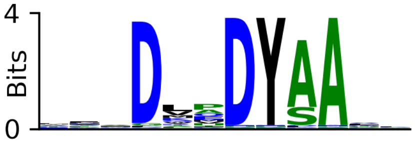
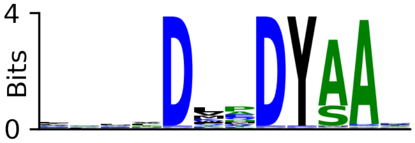
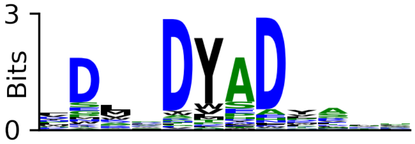
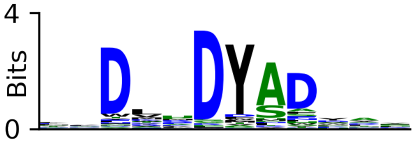
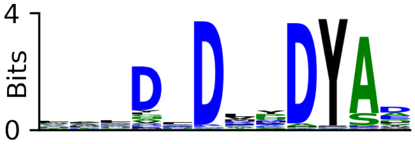
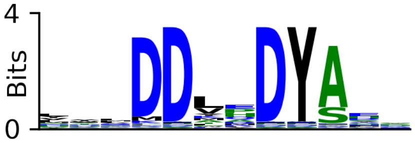
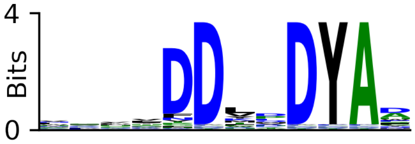
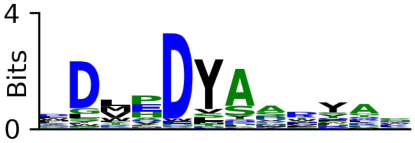
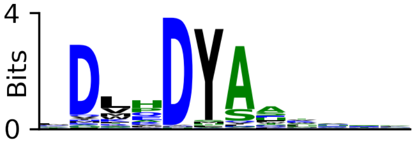
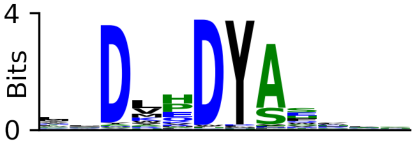
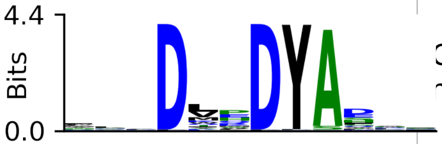
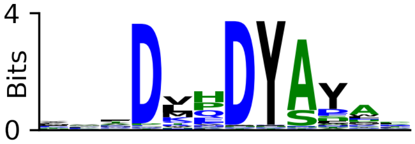
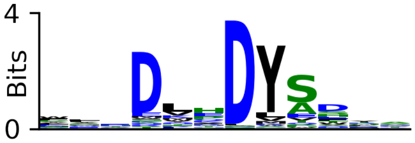
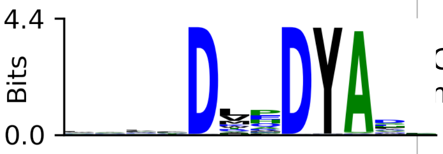
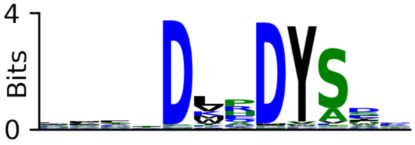
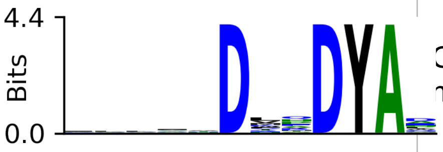
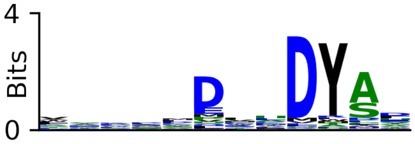
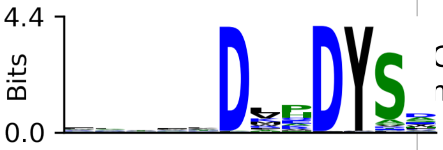
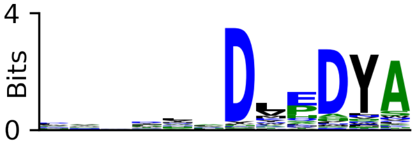
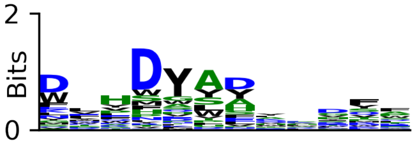


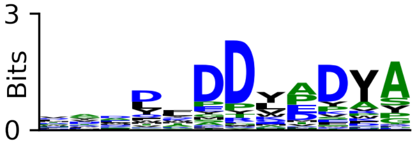
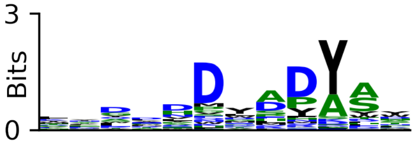
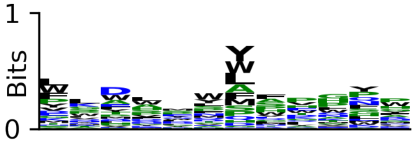
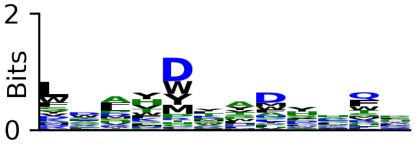
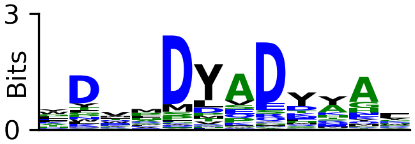
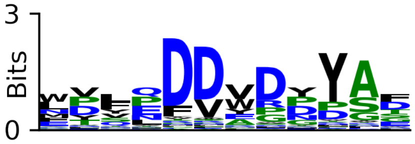
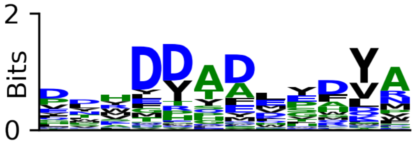
Figure 5.14: *PCA decomposition of all AS-MS data encoded by one-hot encoding with automated cluster detection as described in Table 5.8.* ***Top Left****: Figure as labeled in the main text.* ***Top Right****: The same data fully with its automatous labels. Note that in Main Text, Clusters {1,2,3,4,5,6,7} correspond to automatously labeled clusters {23,13,10,8,2,19,24}, respectively. Each cluster is colorblind color coded and labeled with a central point.* ***Bottom****: A single centroid peptide is reported for each cluster, with the option available to report more centroid peptides spread throughout the cluster.*

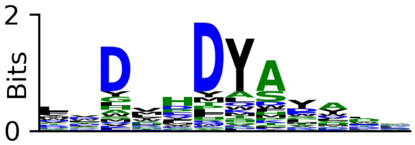
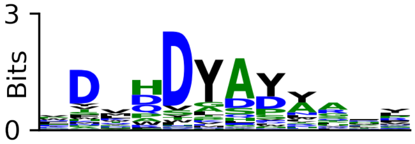
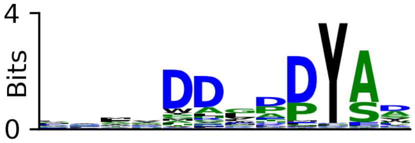
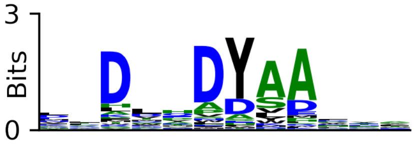
Table 5.8: *Sequence logo report of all clusters detected from PCA dimensionality reduction using one-hot encoding. Both cluster number labels in the Main Text and as autonomously labeled are reported in the table for clarity. Also reported are the number of peptides in each cluster, a single centroid sequence, consensus sequence, logos, and motif class. Details are described in Information about all clusters from dimensionality reduction.*

Main Text Cluster #	Auto-assigned cluster #	# of peptides	Centroid sequence, Consensus sequence	Sequence Logo	Motif Class
1	23	100	LEADTADYAAMF , XXXDXXDYAAX		Expanded motif
2	13	121	PNFMDKHDYAAS , XXXXDXXDYAA		Expanded motif
3	10	103	FDMQDYAAYVWV , XDXXDYADXXX		Expanded motif
4	8	139	AVDRWDYSDVRN , XXDXXDYADXX		Expanded motif
5	2	89	FQLHYDDHDYAE , XXXDXDXXDYA		Expanded motif
	19	44	LASDDFPDYAEA , XXXDDXXDYAX		Expanded motif

	24	65	WKFRDDKMDYAD, XXXXDDXXDYA		Expanded motif
	20	47	PDKHDYASMYFN, XDXXDYAXXXX		Common motif
	26	183	KDVMDYASHFNT, XDXXDYAXXXX		Common motif
	14	194	VVDKPDYARFQT, XXDXXDYAXXX		Common motif
	15	303	PRRDWRDYADNV, XXXDXXDYAXX		Common motif
	17	72	TKLDKHDYAYPR, XXXDXXDYAYX		Common motif
	11	94	VVAELHDYAHDA, XXXDXXDYSXX		Common motif

	6	429	WYESDVKDYADT, XXXXDXXDYAX		Common motif
	27	96	LLFFDKPDYSHK, XXXXDXXDYSX		Common motif
	1	921	MMTTNDWQDYAY, XXXXXDXXDYA		Common motif
	29	55	HGGKSDKVDMAF, XXXXXDXXDYA		Common motif
	18	302	DLVFYDLRDYSS, XXXXXDXXDYS		Common motif
	9	187	SKWWLADWPDYS, XXXXXXDXXDY		Common motif
	16	56	DLHDYSHQLVFG, XXXDXXXXXXX		Weak

	12	25	NQPQLDDLPDYA, XXXXXDDXXDY		Weak
	21	38	TPGDDPEMDYAG, XXXXXDXXDYX		Weak
	22	62	WYTHMMFPWMWF, XXXXXXXXXXX		Weak
	25	37	LSAYMVVDWFRM, XXXXXXXXXXX		Weak
	28	24	WDMHDYADDMGF, XDXXDYADXXA		Weak
	30	8	MYQQDDVDPYSD, XXXXDDXDXYA		Weak
	31	18	DLRDYAELGAYN, XXXDXXXXXXX		Weak

	3	91	MLDLADYALADL , XXDXXDYXXXX		Weak
	4	43	LDVHDYAYLRDF , XDXXDYAXXXX		Weak
	7	59	VFGPPDWDGYAD , XXXXDDXXDYA		Weak
	5	99	MEDTQDYSAVHM , XXDXXDYAAXX		Weak

5.7.4.2. PCA, physicochemical encoding cluster information

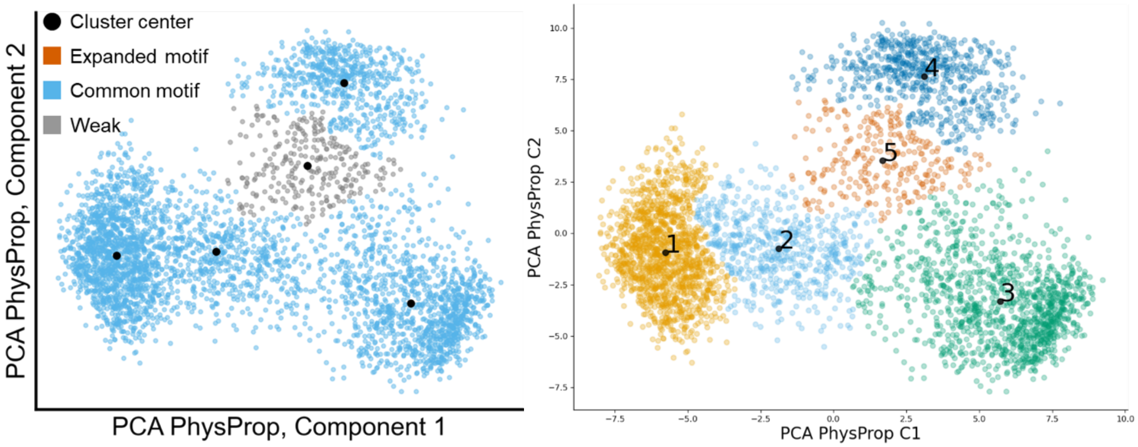
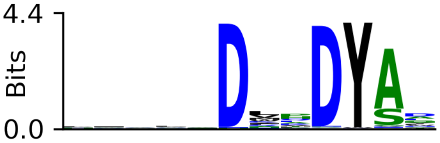
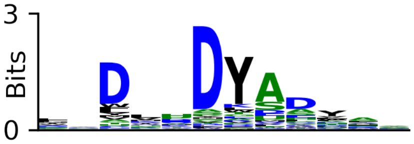
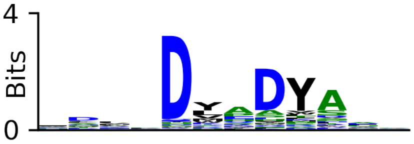
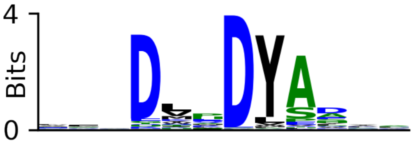
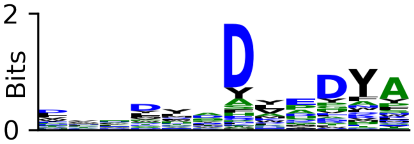


Figure 5.15: *PCA decomposition of all AS-MS data encoded by Physicochemical encoding with automated cluster as described in Table 5.9.* ***Left****: Figure as labeled in the main text.* ***Right****: The same data fully with its automatous labels. No clusters are labeled or reported in the main text because all clusters contain the common or a weak motif. Each cluster is colorblind color coded and labeled with a central point. No centroid plot is reported as no expanded motifs were observed.*

Table 5.9: *Sequence logo report of all clusters detected from PCA dimensionality reduction using Physicochemical encoding. In the table, automatously numbered clusters are reported with the number of peptides in each cluster, a single centroid sequence, consensus sequence, logos, and motif class. Details are described in Information about all clusters from dimensionality reduction.*

Auto-assigned cluster #	**# of peptides**	**Centroid sequence, Consensus sequence**	**Sequence Logo**	**Motif Class**
1	1403	LLQTQDYPDYSQ , XXXXXDXXDYA		Common motif
2	609	VFDLEDYAGRAP , XXDXXDYAXXX		Common motif
3	1197	YFNEDAPDYASP , XXXXDXXDYAX		Common motif
4	625	MPLDVGDYAAQN , XXXDXXDYAXX		Common motif
5	270	SPAVHHDVEDYA , XXXXXXDXXXX		Weak

5.7.4.3. PCA, ESM-2 encoding cluster information

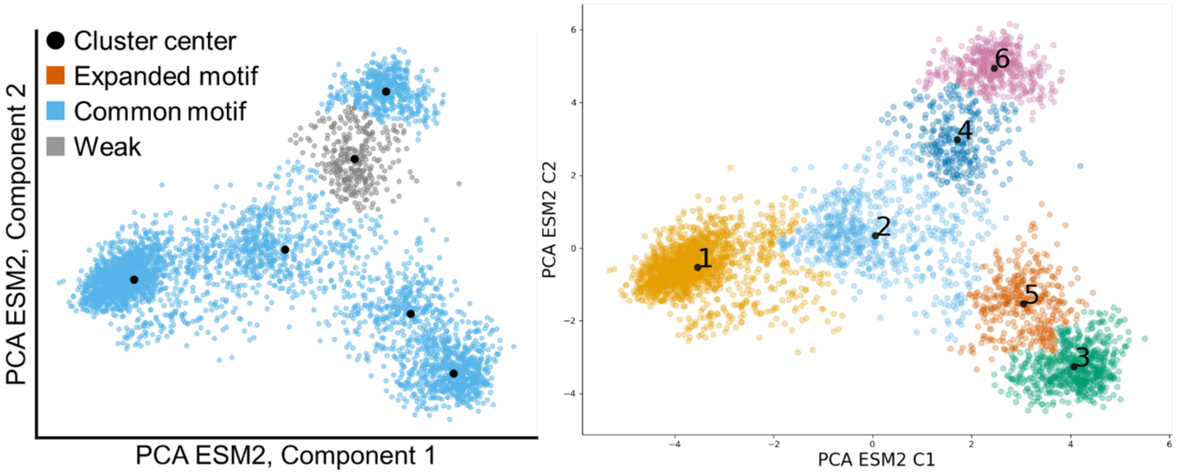
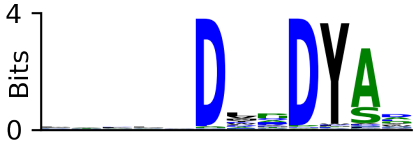
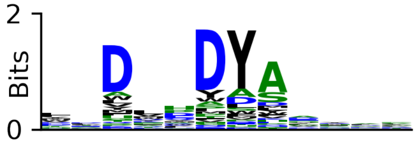


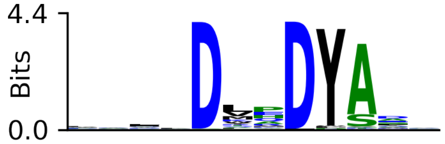
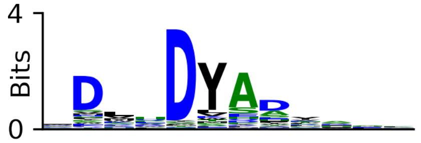
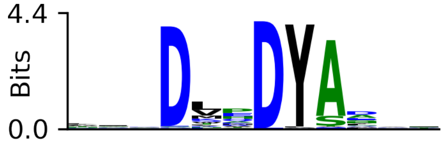
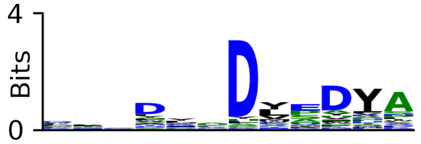
***Figure 5.16:** PCA decomposition of all AS-MS data encoded by ESM2 encoding with automated cluster detection as described in Table 5.10. **Left**: Figure as labeled in the main text. **Right**: The same data fully with its automatous labels. No clusters are labeled or reported in the main text because all clusters contain the common or a weak motif. Each cluster is colorblind color coded and labeled with a central point. No centroid plot is reported as no expanded motifs were observed.*

***Table 5.10:** Sequence logo report of all clusters detected from PCA dimensionality reduction using ESM2 encoding. In the table, automatously numbered clusters are reported with the number of peptides in each cluster, a single centroid sequence, consensus sequence, logos, and motif class. Details are described in Information about all clusters from dimensionality reduction.*

Auto-assigned cluster #	**# of peptides**	**Centroid sequence, Consensus sequence**	**Sequence Logo**	**Motif Class**

1	1599	`WFRAFDMEDYSD`
		`,`
		`XXXXXDXXDYA`
				Common motif

2	648	`LDDPADYAVGTK`
		`,`
		`XXDXXDYXXXX`
				Common motif

3	663	HHTYDLPDYSFY, XXXXDXXDYAX		Common motif
5	389	LDVQDYANVSES, XDXXDYAXXXX		Common motif
6	495	YLMDLFDYAHKT, XXXDXXDYAXX		Common motif
4	310	WDVFFPDYSHRP, XXXXXXDXXDY		Weak

5.7.4.4. PCA, fingerprint encoding cluster information

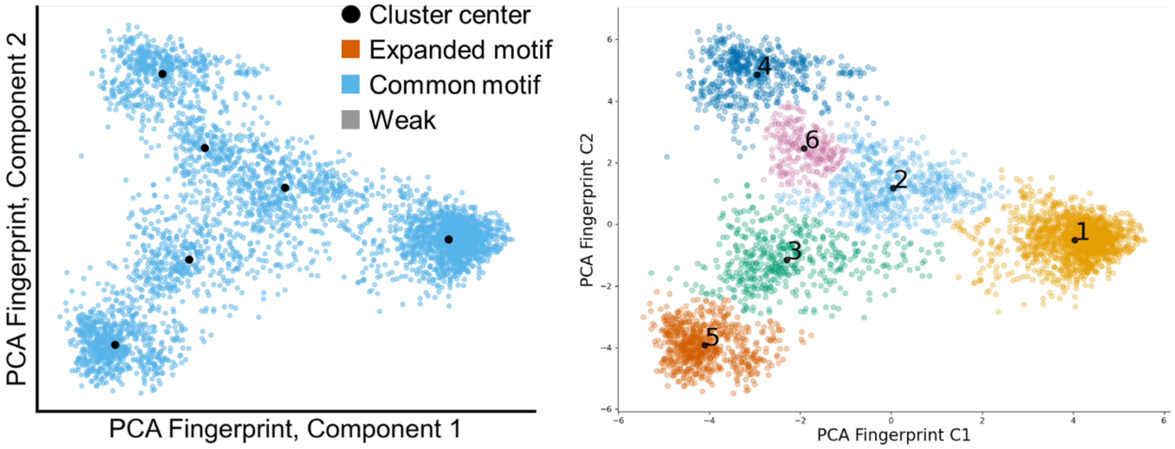
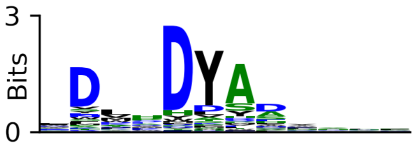
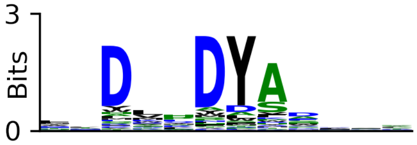
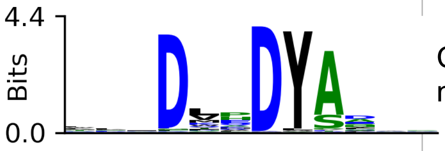
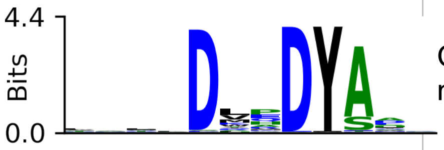
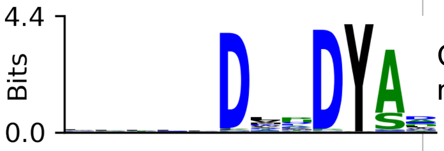
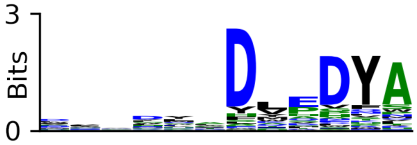


Figure 5.17: *PCA decomposition of all AS-MS data encoded by Fingerprint encoding with automated cluster detection as described in Table 5.11.* ***Left****: Figure as labeled in the main text.* ***Right****: The same data fully with its automatous labels. No clusters are labeled or reported in the main text because all clusters contain the common or a weak motif. Each cluster is colorblind color coded and labeled with a central point. No centroid plot is reported as no expanded motifs were observed.*

Table 5.11: *Sequence logo report of all clusters detected from PCA dimensionality reduction using Fingerprint encoding. In the table, automatously numbered clusters are reported with the number of peptides in each cluster, a single centroid sequence, consensus sequence, logos, and motif class. Details are described in Information about all clusters from dimensionality reduction.*

Auto-assigned cluster #	**# of peptides**	**Centroid sequence, Consensus sequence**	**Sequence Logo**	**Motif Class**
3	484	FDRLDYSDQFFK, XDXXDYAXXXX		Common motif
2	572	HADVQDYAFHYT, XXDXXDYAXXX		Common motif
4	575	LDGDLWDYADTY, XXXDXXDYAXX		Common motif
5	703	FFLMDLWDYARS, XXXXDXXDYAX		Common motif
1	1521	LLKWVDKHDYAY, XXXXXDXXDYA		Common motif
6	249	KDHDYAYFMETR, XXXXXXDXXDY		Common motif

5.7.4.5. PCA, N-grams encoding cluster information

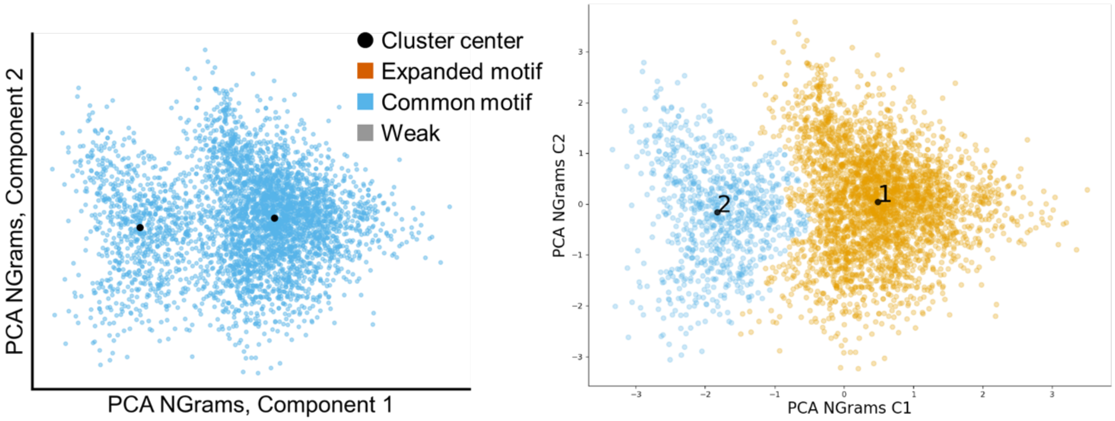
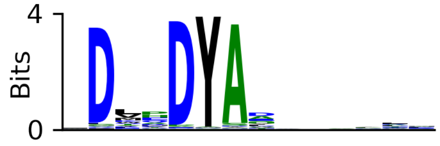
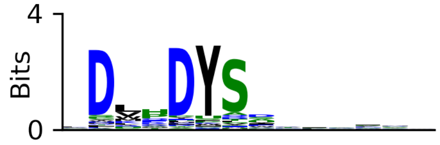


Figure 5.18: *PCA decomposition of all AS-MS data encoded by N-grams encoding with automated cluster detection as described in Table 5.12.* ***Left****: Figure as labeled in the main text.* ***Right****: The same data fully with its automatous labels. No clusters are labeled or reported in the main text because all clusters contain the common or a weak motif. Each cluster is colorblind color coded and labeled with a central point. No centroid plot is reported as no expanded motifs were observed.*

Table 5.12: *Sequence logo report of all clusters detected from PCA dimensionality reduction using N-grams encoding. In the table, automatously numbered clusters are reported with the number of peptides in each cluster, a single centroid sequence, consensus sequence, logos, and motif class. Details are described in Information about all clusters from dimensionality reduction.*

NOTE: Because N-grams encodes peptides by the presence of their motifs, irrespective of frameshift, the logo plot displays the sequences aligned by ClustalW to the second position to show the motif.

Auto-assigned cluster #	**# of peptides**	**Centroid sequence, Consensus sequence**	***<u>ALIGNED</u>*** **Sequence Logo**	**Motif Class**
1	3242	PSDLRDYAAGFF, XDXXDYAX-----		Common motif
2	862	QVDTRDYSDLYF, XDXXDYSX-----		Common motif

5.7.4.6. UMAP, One-hot encoding cluster information

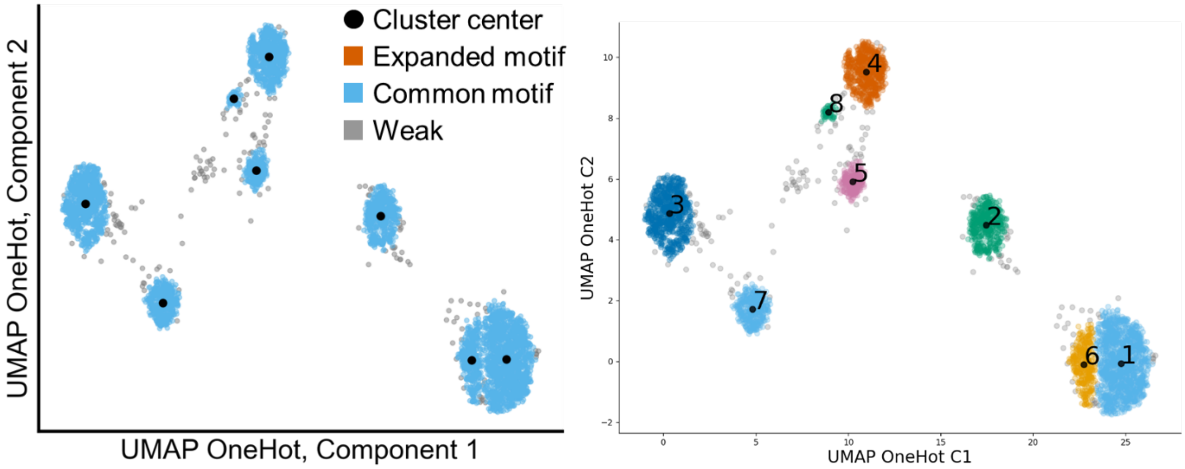
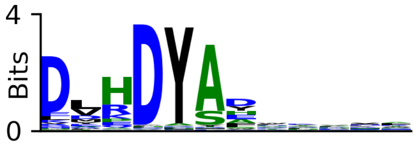
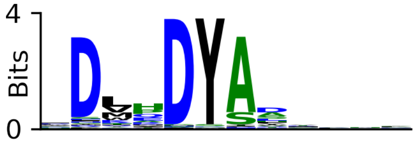
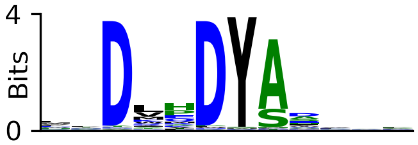
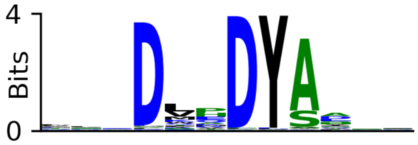
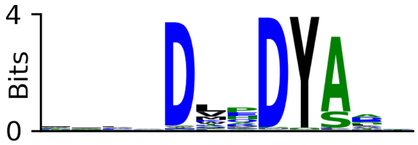
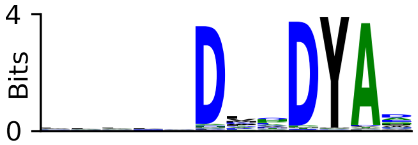
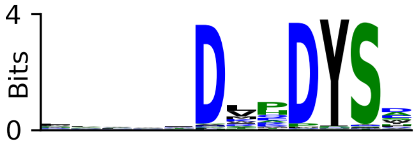
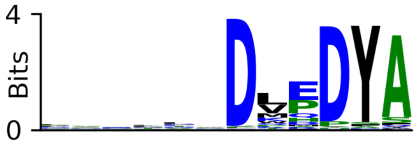
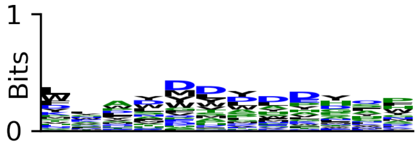


Figure 5.19: *UMAP decomposition of all AS-MS data encoded by one-hot encoding with automated cluster detection as described in Table 5.13.* ***Left****: Figure as labeled in the main text.* ***Right****: The same data fully with its automatous labels. No clusters are labeled or reported in the main text because all clusters contain the common or a weak motif. Each cluster is colorblind color coded and labeled with a central point. No centroid plot is reported as no expanded motifs were observed.*

Table 5.13: *Sequence logo report of all clusters detected from UMAP dimensionality reduction using one-hot encoding. In the table, automatously numbered clusters are reported with the number of peptides in each cluster, a single centroid sequence, consensus sequence, logos, and motif class. Details are described in Information about all clusters from dimensionality reduction.*

Auto-assigned cluster #	**# of peptides**	**Centroid sequence, Consensus sequence**	**Sequence Logo**	**Motif Class**
8	59	DVRDYAENDFLV, DXHDYAXXXXX		Common motif
7	354	LDMQDYAAGDWM, XDXXDYAXXXXX		Common motif

2	454	EGDAEDYAAFRG, XXDXXDYAXXX		Common motif
4	573	FNLDEQDYADTP, XXXDXXDYAXX		Common motif
3	739	FPVVDWEDYATW, XXXXDXXDYAX		Common motif
1	1230	SNEFSDMLDYAE, XXXXXDXXDYA		Common motif
6	323	FDLFLDVPDYSS, XXXXXDXXDYS		Common motif
5	209	LPGGFLDWEDYA, XXXXXXDXXDY		Common motif
0	163			Weak

5.7.4.7. UMAP, physicochemical encoding cluster information

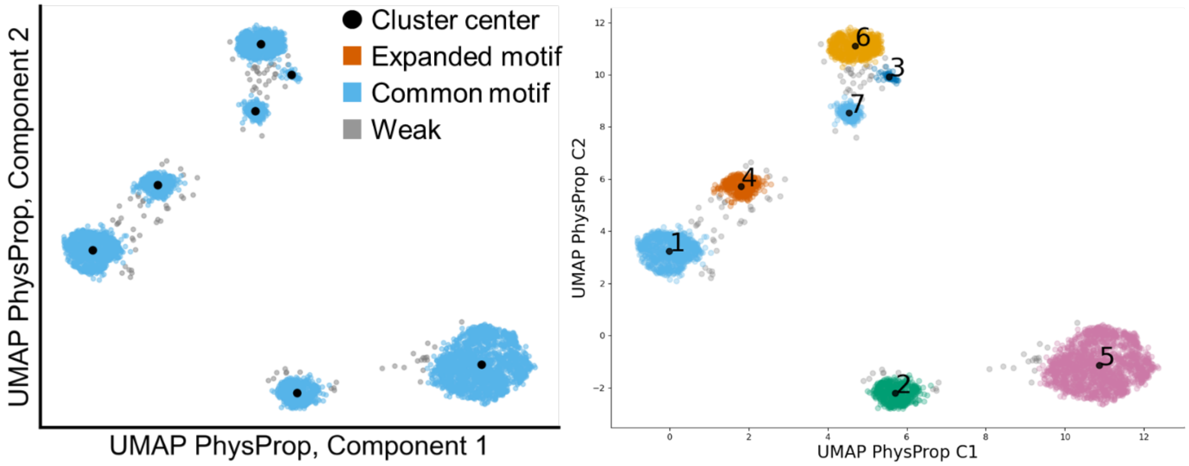
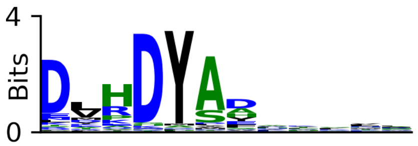
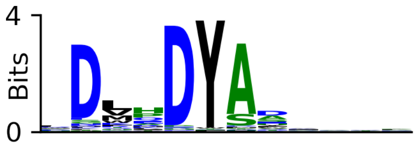


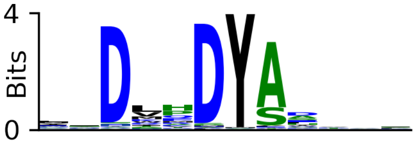
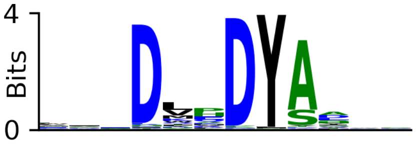
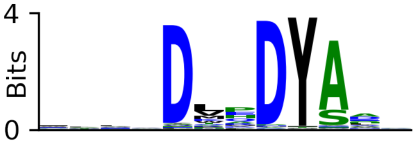
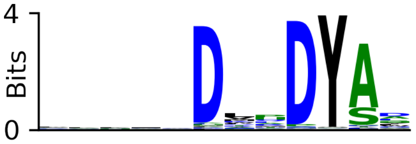
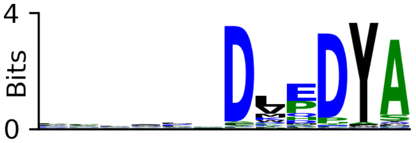
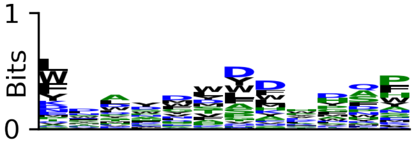
Figure 5.20: *UMAP decomposition of all AS-MS data encoded by Physicochemical encoding with automated cluster detection as described in Table 5.14.* ***Left****: Figure as labeled in the main text.* ***Right****: The same data fully with its automatous labels. No clusters are labeled or reported in the main text because all clusters contain the common or a weak motif. Each cluster is colorblind color coded and labeled with a central point. No centroid plot is reported as no expanded motifs were observed.*

Table 5.14: *Sequence logo report of all clusters detected from UMAP dimensionality reduction using Physicochemical encoding. In the table, automatously numbered clusters are reported with the number of peptides in each cluster, a single centroid sequence, consensus sequence, logos, and motif class. Details are described in Information about all clusters from dimensionality reduction.*

Auto-assigned cluster #	**# of peptides**	**Centroid sequence, Consensus sequence**	**Sequence Logo**	**Motif Class**

3	64	`DLKDYADNHWEA,`
		`DXXDYAXXXXXX`
				Common motif

4	358	`ADMEDYAQNYPL,`
		`XDXXDYAXXXXX`
				Common motif

2	465	FFDLPDYSVPKL, XXDXXDYAXXX		Common motif
6	578	PYLDMEDYAQLF, XXXDXXDYAXX		Common motif
1	756	LYWDDVEDYAEH, XXXXDXXDYAX		Common motif
5	1572	LDFGGDWPDYAH, XXXXXDXXDYA		Common motif
7	214	TPQMEADVDPYA, XXXXXXDXXDY		Common motif
0	97			Weak

5.7.4.8. UMAP, ESM-2 encoding cluster information

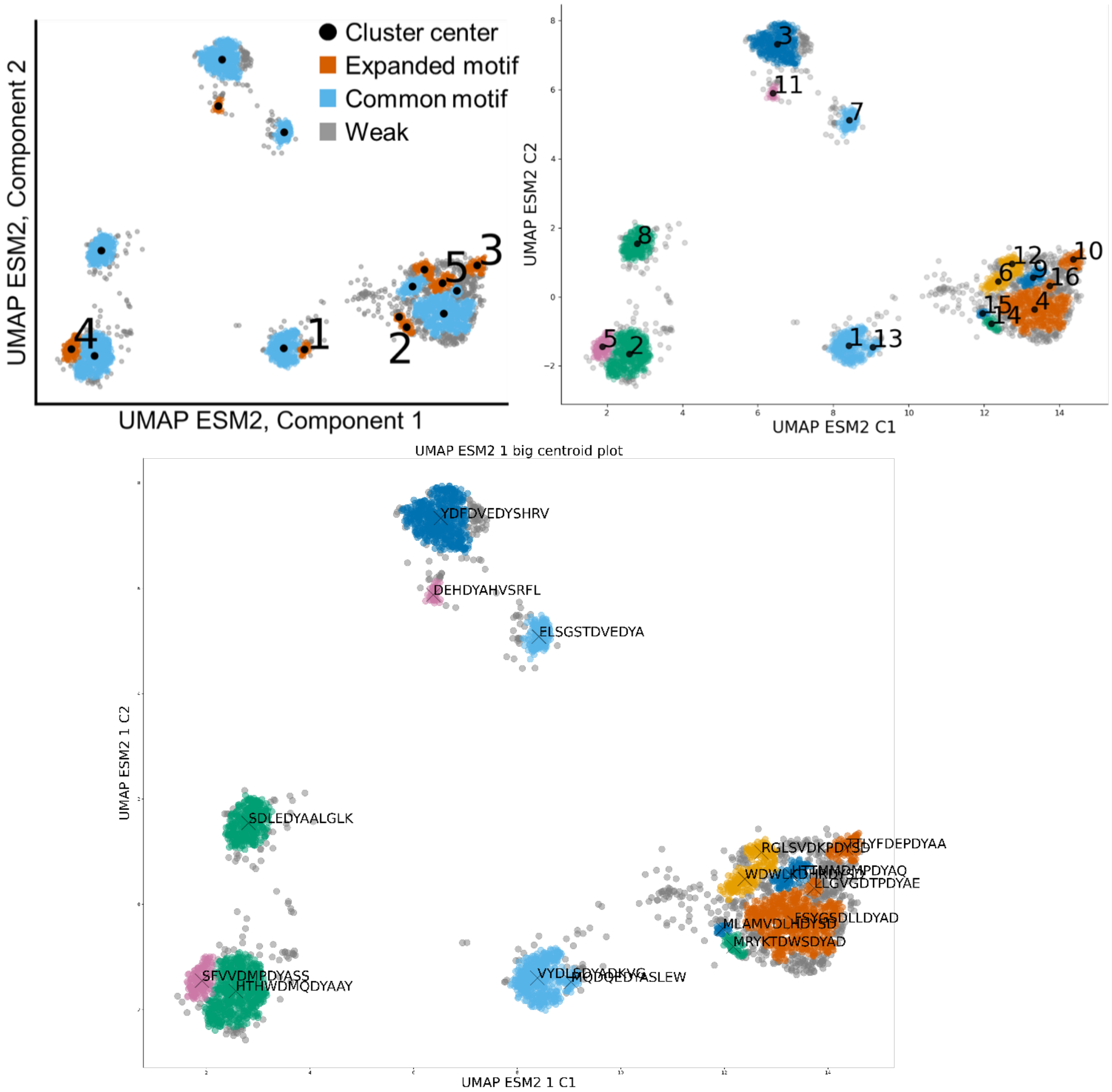
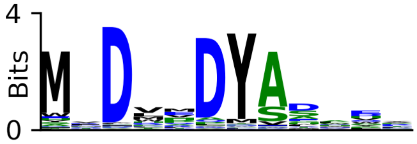
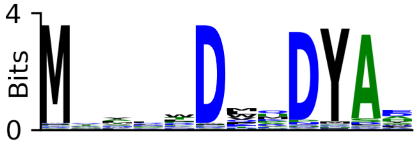
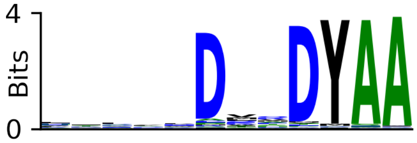
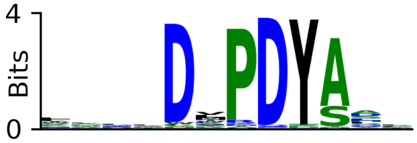
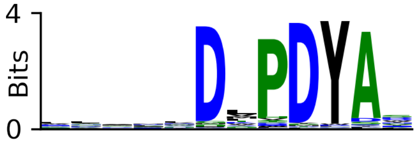
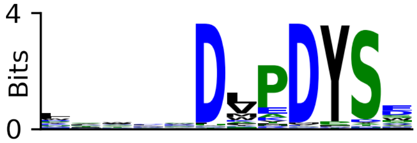
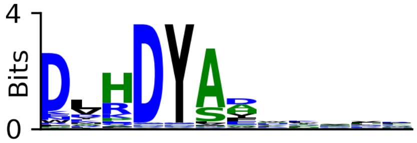
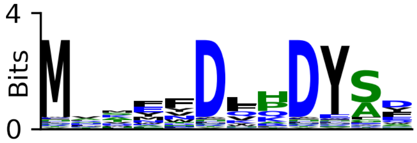
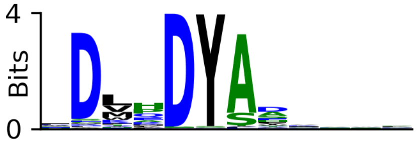
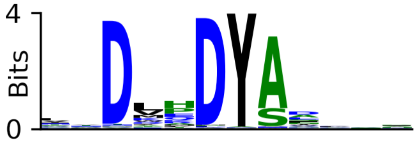
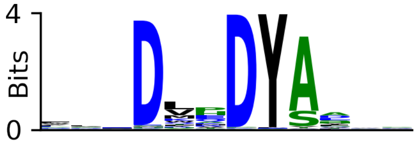
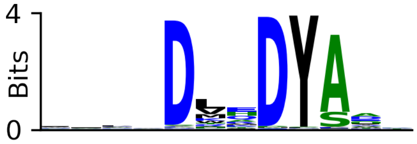
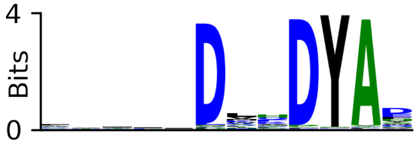
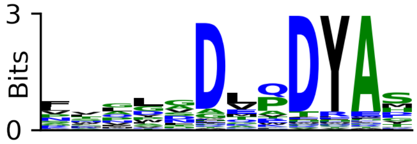
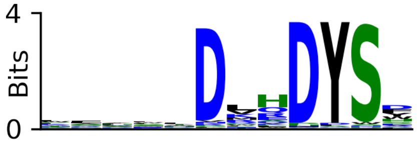
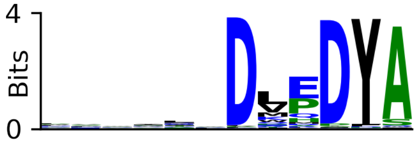
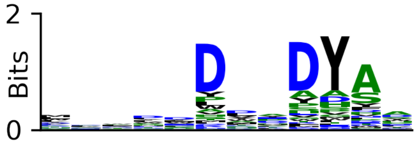


Figure 5.21: *UMAP decomposition of all AS-MS data encoded by ESM-2 encoding with automated cluster detection as described in Table 5.15.* ***Top Left****: Figure as labeled in the main text.* ***Top Right****: The same data fully with its automatous labels. Note that in Main Text, Clusters {1,2,3,4,5,6,7} correspond to automatously labeled clusters {13,14,15,10,5,9,12}, respectively. Each cluster is colorblind color coded and labeled with a central point.* ***Bottom****: A single centroid peptide is reported for each cluster, with the option available to report more centroid peptides spread throughout the cluster.*

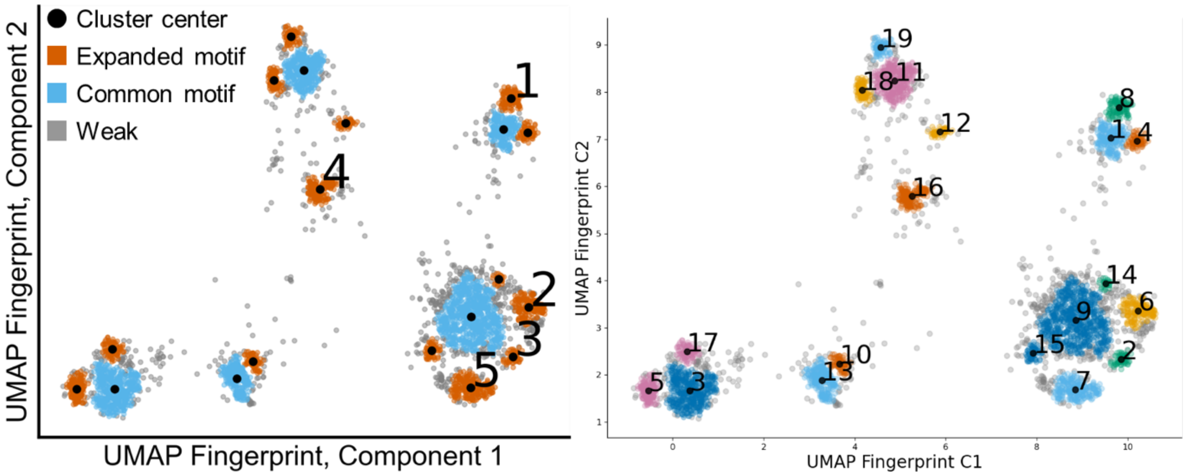
Table 5.15: *Sequence logo report of all clusters detected from UMAP dimensionality reduction using ESM-2 encoding. Both cluster number labels in the Main Text and as autonomously labeled are reported in the table for clarity. Also reported are the number of peptides in each cluster, a single centroid sequence, consensus sequence, logos, and motif class. Details are described in Information about all clusters from dimensionality reduction.*

Main Text Cluster #	**Auto-assigned cluster #**	**# of peptides**	**Centroid sequence, Consensus sequence**	**Sequence Logo**	**Motif Class**
1	13	45	MQDQEDYASLEW, MXDXXDYAXXX		Expanded motif
2	14	51	MRYKTDWSDYAD, MXXXXDXXDYA		Expanded motif
3	10	115	TTLYFDEPDYAA, XXXXXDXXDYA		Expanded motif
4	5	149	SFVVDMPDYASS, XXXXDXPDYAX		Expanded motif
5	9	109	HTTMMDMPDYAQ, XXXXXDXPDYA		Expanded motif
	12	87	RGLSVDKPDYSD, XXXXXDXPDYS		Expanded motif

	11	54	DEHDYAHVSRFL, DXHDYAXXXXX		Expanded motif
	15	22	MLAMVDLHDYSD, MXXXXDXXDYS		Expanded motif
	8	330	SDLEDYAALGLK, XDXXDYAXXXX		Common motif
	1	397	VYDLSDYADKVG, XXDXXDYAXXX		Common motif
	3	518	YDFDVEDYSHRV, XXXDXXDYAXX		Common motif
	2	564	HTHWDMQDYAAY, XXXXDXXDYAX		Common motif
	4	645	FSYGSDLLDYAD, XXXXXDXXDYA		Common motif
	16	24	LLGVGDTPDYAE, XXXXXDXXDYA		Common motif

	6	132	WDWLKDHRDYSD, XXXXXDXXDYS		Common motif
	7	191	ELSGSTDVEDYA, XXXXXXDXXDY		Common motif
	0	671			Weak

5.7.4.9. UMAP, fingerprint encoding cluster information



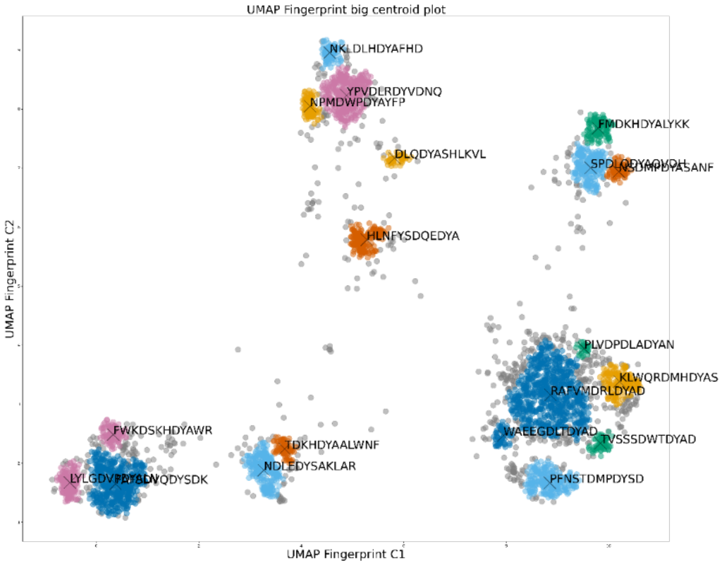
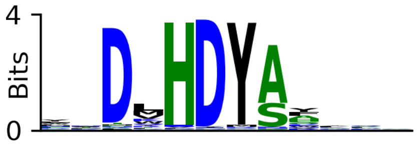
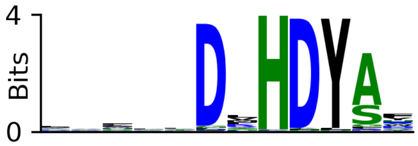
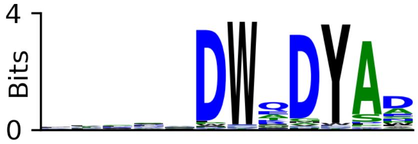
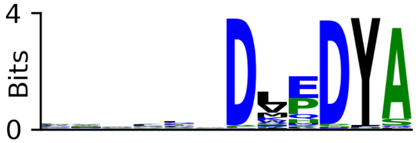
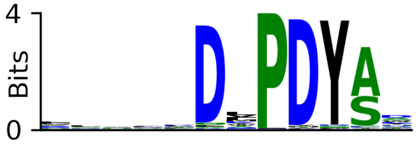
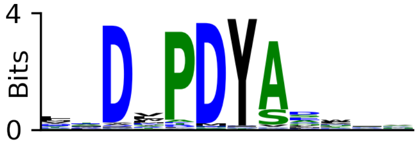
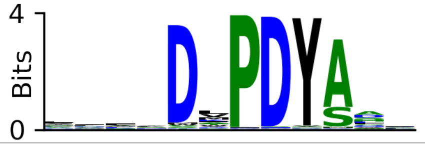
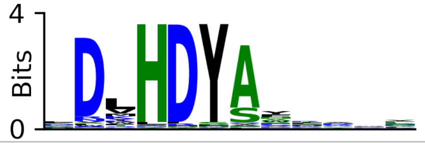
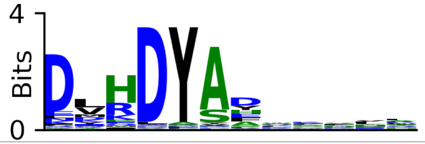
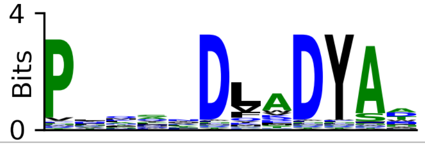
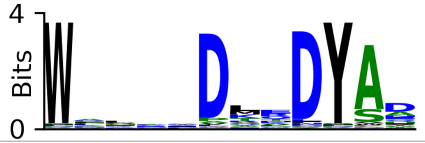
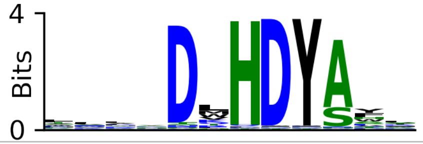
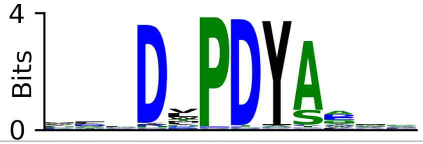
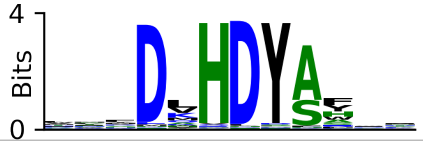
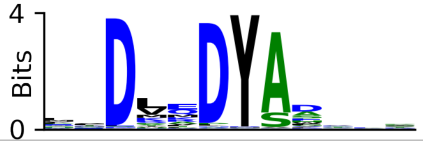
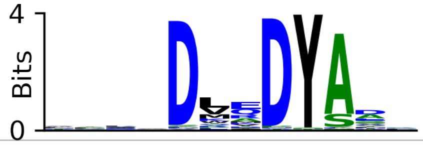
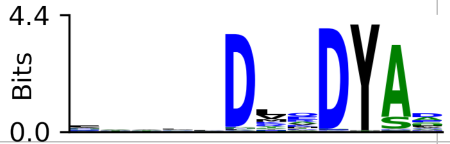
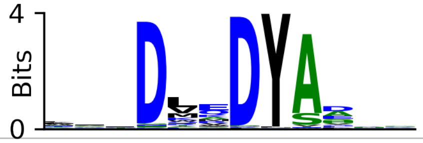
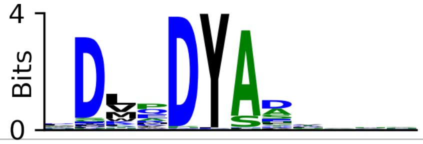
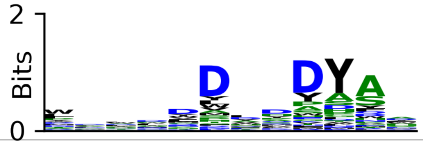


Figure 5.22: *UMAP decomposition of all AS-MS data encoded by Fingerprint encoding with automated cluster detection as described in Table 5.16.* ***Top Left****: Figure as labeled in the main text.* ***Top Right****: The same data fully with its automatous labels. Note that in Main Text, Clusters {1,2,3,4,5,6,7} correspond to automatously labeled clusters {8,6,15,2,16,5,7}, respectively. Each cluster is colorblind color coded and labeled with a central point.* ***Bottom****: A single centroid peptide is reported for each cluster, with the option available to report more centroid peptides spread throughout the cluster.*

Table 5.16: *Sequence logo report of all clusters detected from UMAP dimensionality reduction using Fingerprint encoding. Both cluster number labels in the Main Text and as autonomously labeled are reported in the table for clarity. Also reported are the number of peptides in each cluster, a single centroid sequence, consensus sequence, logos, and motif class. Details are described in Information about all clusters from dimensionality reduction.*

Main Text Cluster #	**Auto-assigned cluster #**	**# of peptides**	**Centroid sequence, Consensus sequence**	**Sequence Logo**	**Motif Class**
1	8	122	FMDKHDYALYKK, XXDXHDYAXXX		Expanded motif
2	6	172	KLWQRDMHDYAS, XXXXXDXHDYA		Expanded motif

3	2	68	TVSSSDWTDYAD, XXXXXDWXDYA		Expanded motif
4	16	193	HLNFYSDQEDYA, XXXXXXDXXDY		Expanded motif
5	7	223	PFNSTDMPDYSD, XXXXXDXPDYA		Expanded motif
	4	90	NSDMPDYASANF, XXDXPDYAXXX		Expanded motif
	5	149	LYLGDVPDYALN, XXXXDXPDYAX		Expanded motif
	10	81	TDKHDYAALWNF, XDXHDYAXXXX		Expanded motif
	12	56	DLQDYASHLKVL, DXHDYAXXXXX		Expanded motif
	14	29	PLVDPDLADYAN, PXXXXDLADYA		Expanded motif
	15	62	WAEEGDLTDYAD, WXXXXDXXDYA		Expanded motif

	17	90	FWKDSKHDYAWR, XXXXDXHDYAX		Expanded motif
	18	110	NPMDWPDYAYFP, XXXDXPDYAXX		Expanded motif
	19	92	NKLDLHDYAFHD, XXXDXHDYAXX		Expanded motif
	1	225	SPDLQDYAQVDH, XXDXXDYAXXX		Common motif
	3	435	FAFSDVQDYSDK, XXXXDXXDYAX		Common motif
	9	723	RAFVMDRLDYAD, XXXXXDXXDYA		Common motif
	11	336	YPVDLRDYVDNQ, XXXDXXDYAXX		Common motif
	13	243	NDLEDYSAKLAR, XDXXDYAXXXX		Common motif
	0	605			Weak

5.7.4.10. UMAP, N-grams encoding cluster information

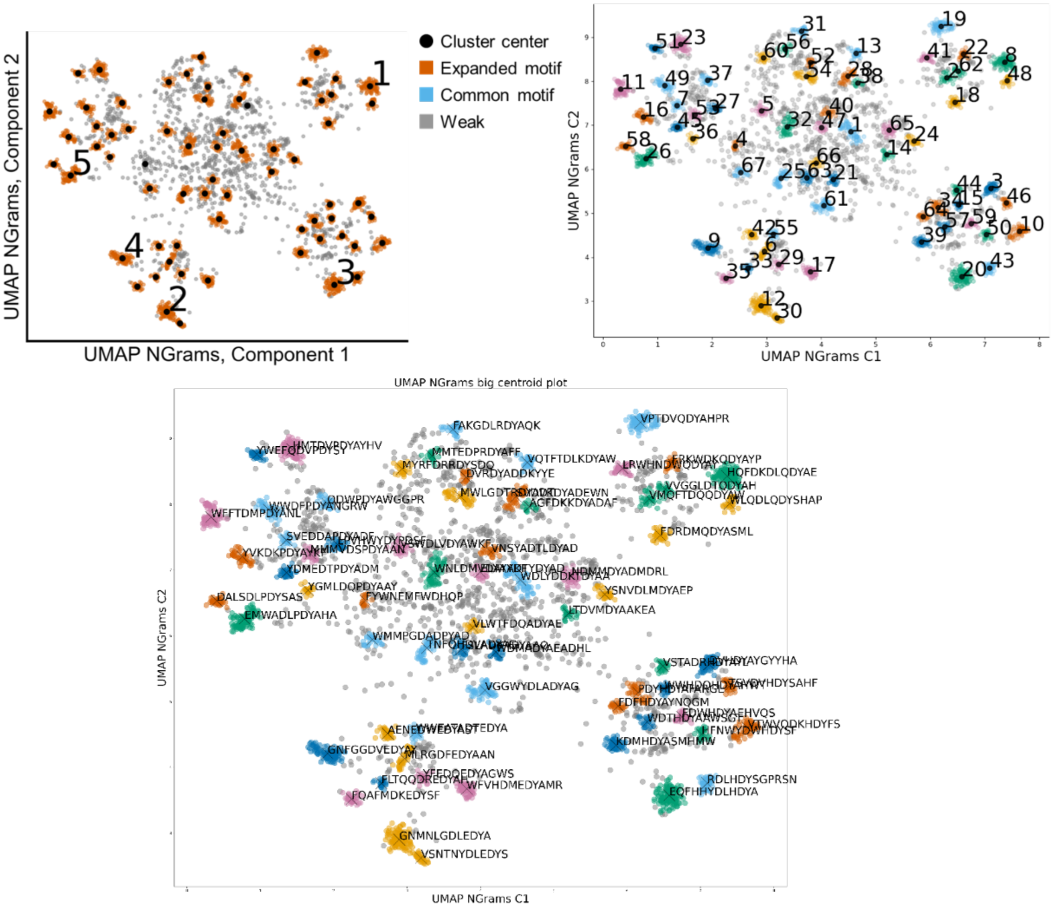
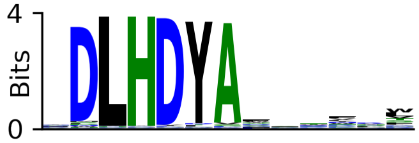
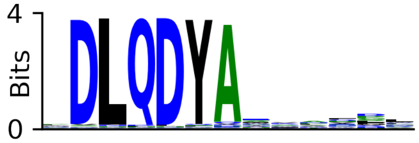
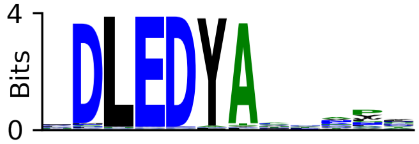
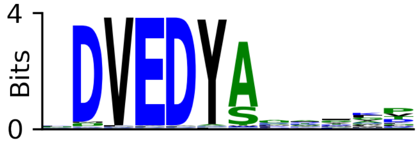
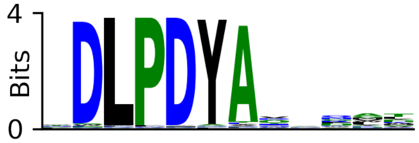
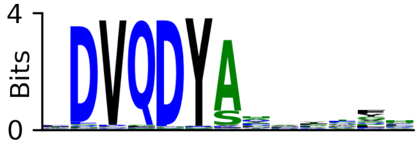
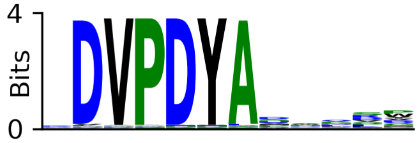
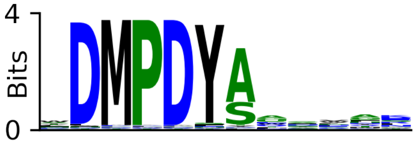
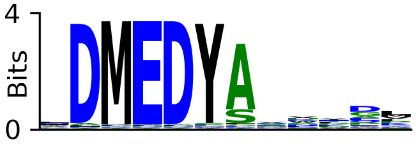
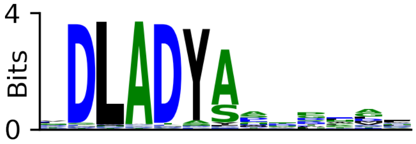
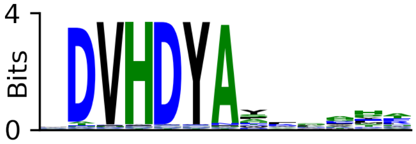
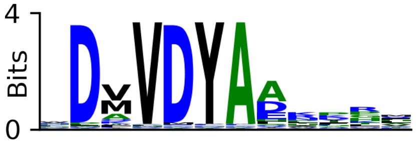
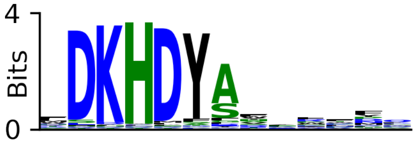
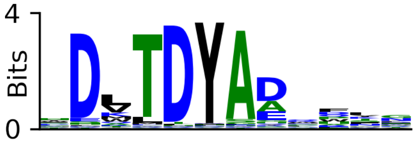


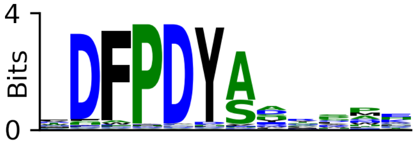
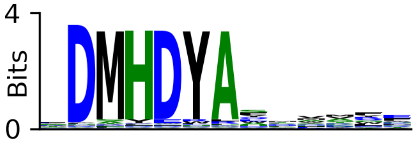
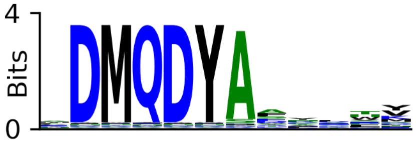
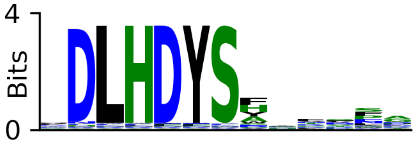
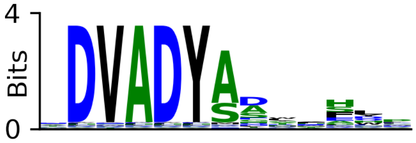
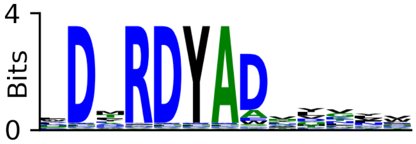
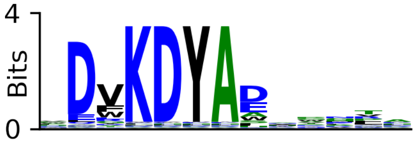
Figure 5.23: *UMAP decomposition of all AS-MS data encoded by N-grams encoding with automated cluster detection as described in Table 5.17.* ***Top Left****: Figure as labeled in the main text.* ***Top Right****: The same data fully with its automatous labels. Note that in Main Text, Clusters {1,2,3,4,5,6,7} correspond to automatously labeled clusters {8,12,20,9,26,19,23}, respectively. Each cluster is colorblind color coded and labeled with a central point.* ***Bottom****: A single centroid peptide is reported for each cluster, with the option available to report more centroid peptides spread throughout the cluster.*

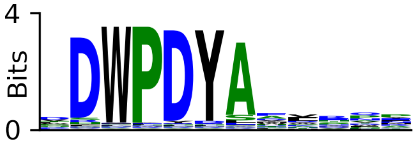
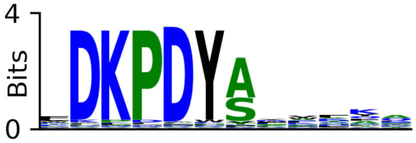
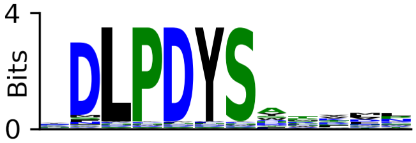
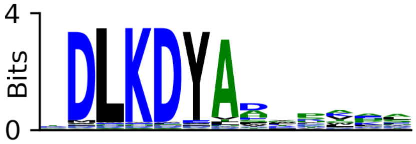
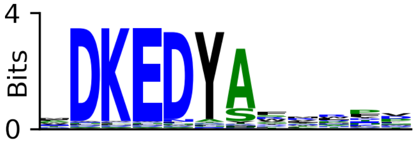
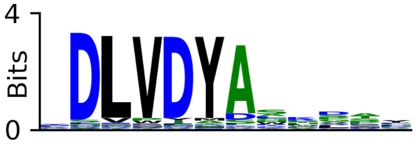
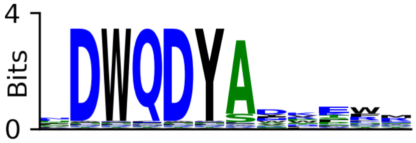
Table 5.17: *Sequence logo report of all clusters detected from UMAP dimensionality reduction using N-grams encoding. Both cluster number labels in the Main Text and as autonomously labeled are reported in the table for clarity. Also reported are the number of peptides in each cluster, a single centroid sequence, consensus sequence, logos, and motif class. Details are described in Information about all clusters from dimensionality reduction.*

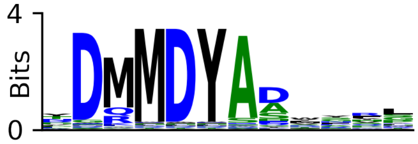
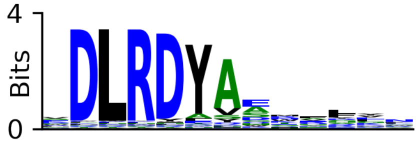
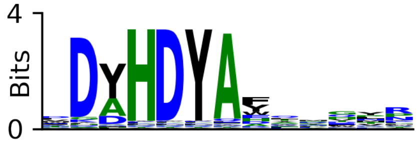
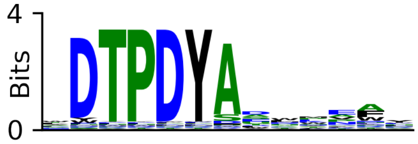
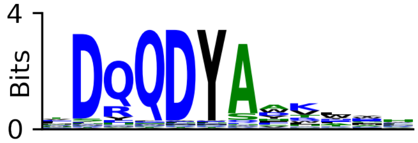
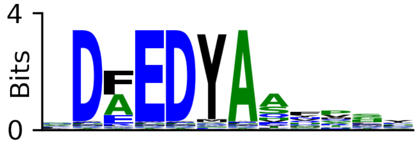
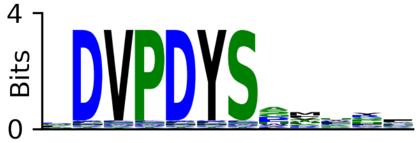
NOTE: Because N-grams encodes peptides by the presence of their motifs, irrespective of frameshift, the logo plot displays the sequences aligned by ClustalW to the second position to show the motif.

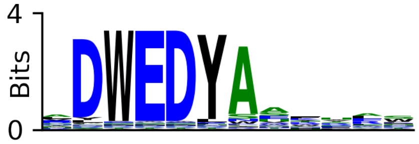
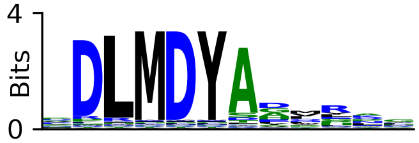
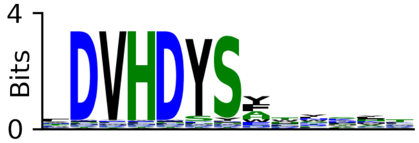
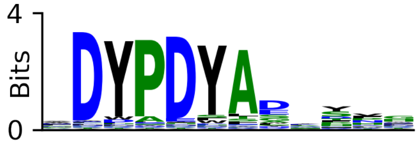
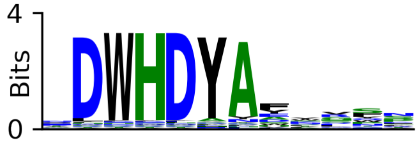
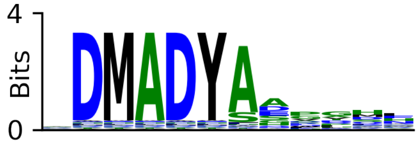
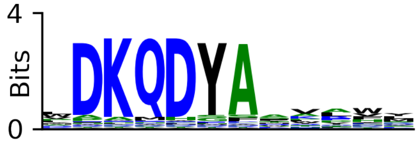
Main Text Cluster #	Auto-assigned cluster #	# of peptides	Centroid sequence, Consensus sequence	ALIGNED Sequence Logo	Motif Class
1	20	121	EQFHHYDLHDYA, -XXXXXDLHDYAXXXX-		Expanded motif
2	8	121	HQFDKDLQDYAE, -XXXXXDLQDYAXXX--		Expanded motif
3	12	121	GNMNLGDLEDYA, -XXXXXDLEDYAXXX-		Expanded motif
4	9	104	GNFGGDVEDYAY, -XXXXXDVEDYAXXX-		Expanded motif
5	26	102	EMWADLPDYAHA, -XXXXXDLPDYAXXX-		Expanded motif
	19	97	VPTDVQDYAHPR, -XXXXXDVQDYAXXX--		Expanded motif
	23	94	HMTDVPDYAYHV, -XXXXXDVPDYAXXX-		HA tag

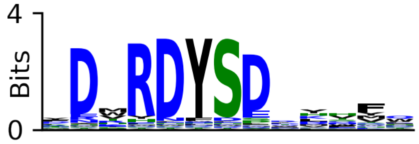
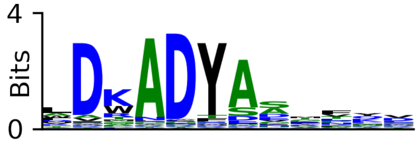
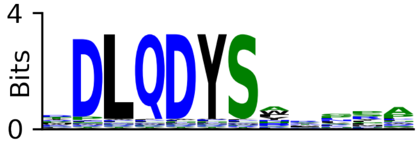
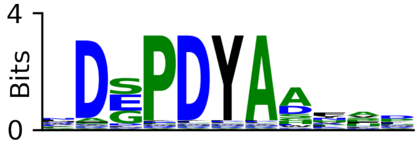
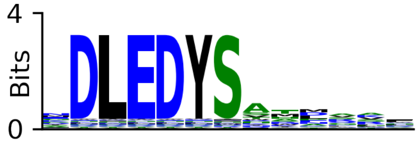
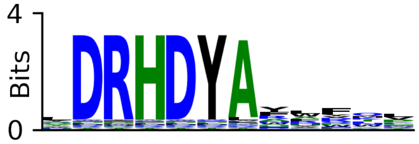
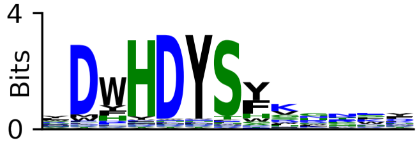
	11	80	WFFTDMPDYANL, -XXXXXDMPDYXXX--		Expanded motif
	17	71	WFVHDMEDYAMR, -XXXXXDMEDYAXX--		Expanded motif
	61	69	VGGWYDLADYAG, -XXXXXDLADYAXXX--		Expanded motif
	3	66	DVHDYAYGYYHA, --XXXXDVHDYAXXXX-		Expanded motif
	32	64	WNLDMVDYAAKF, -XXXXXDXVDYAXXX-		Expanded motif
	10	63	VTWVQDKHDYFS, -XXXXXDKHDYXXXX--		Expanded motif
	1	62	WDLYDDKTDYAA, XXXXXDXTDYAXX--		Expanded motif

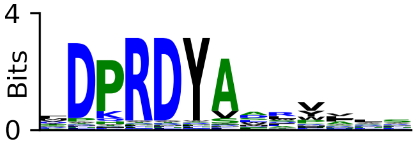
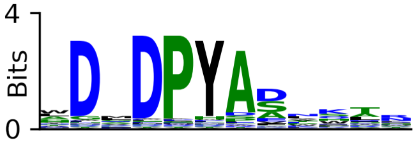
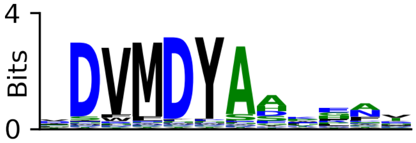
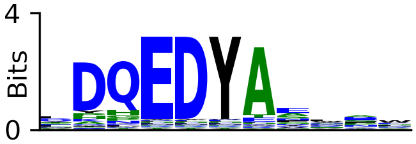
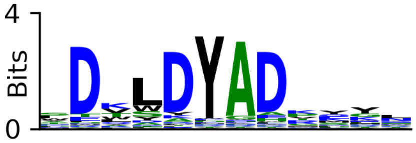
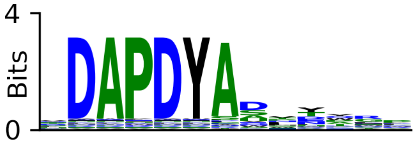
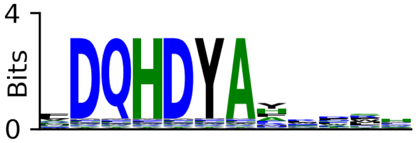
	49	55	WWDFPDYANGRW, XXXXXDFPDYXXXX-		Expanded motif
	39	52	KDMHDYASMHMW, -XXXXDMHDYAXXXX-		Expanded motif
	18	51	FDRDMQDYASML, -XXXXXDMQDYAXXXX-		Expanded motif
	43	50	RDLHDYSGPRSN, -XXXXDLHDYSXXXX-		Expanded motif
	25	49	TNFQHDVADYAG, XXXXXDVADYAXXX--		Expanded motif
	54	48	MWLGDTRDYADT, XXXXXDXRDYADX---		Expanded motif
	28	47	SVDVKDYADEWN, XXXXXDXKDYAXXX--		Expanded motif

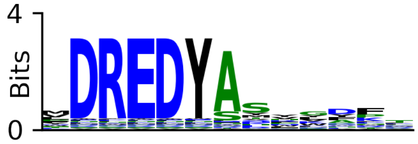
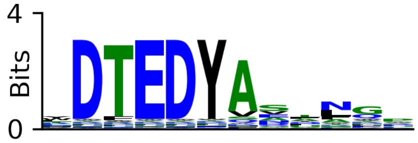
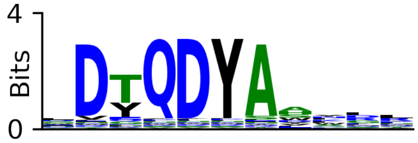
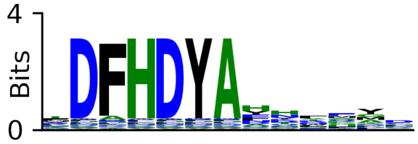
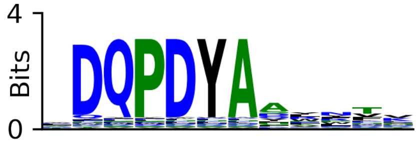
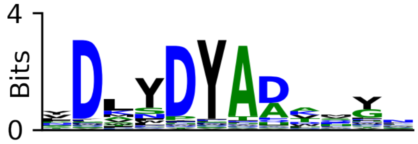
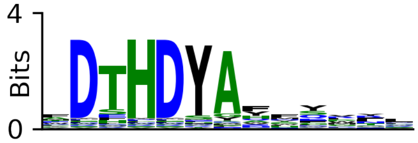
	37	46	QDWPDYAWGGPR, -XXXXXDWPDYAXXXX		Expanded motif
	16	45	YVKDKPDYAYKF, -XXXXXDKPDYXXX--		Expanded motif
	58	43	DALSDLPDYSAS, -XXXXXDLPDYSXXX-		Expanded motif
	13	42	VQTFTDLKDYAW, XXXXXDLKDYAXXX--		Expanded motif
	35	41	FQAFMDKEDYSF, -XXXXXDKEDYAXXX-		Expanded motif
	5	40	VSWDLVDYAWKF, -XXXXXDLVDYAXXX-		Expanded motif
	41	40	LRWHNDWQDYAY, XXXXXDWQDYAXX--		Expanded motif

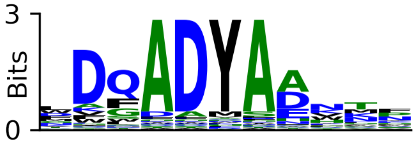
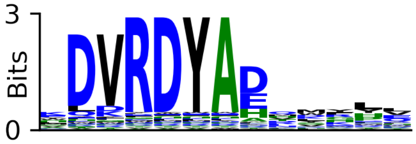
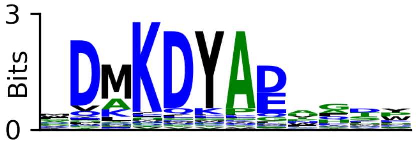
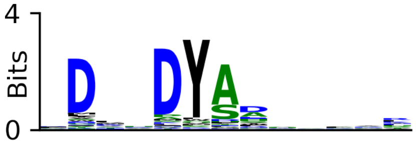
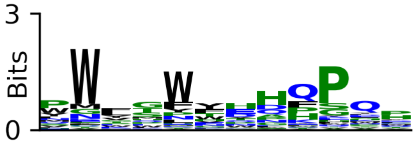
	65	40	NDMMDYADMDRL, XXXXXDMMDYAXXX-		Expanded motif
	31	38	FAKGDLRDYAQK, -XXXXXDLRDYAXXXXX-		Expanded motif
	34	38	PDYHDYAFARGL, XXXXXDXHDYAXXXXX-		Expanded motif
	45	38	YDMEDTPDYADM, XXXXXDTPDYAXXX--		Expanded motif
	2	37	VMQFTDQQDYAW, -XXXXXDQQDYAXX--		Expanded motif
	6	37	MLRGDFEDYAAN, -XXXXXDXEDYAXX--		Expanded motif
	51	37	YWEFQDVPDYSY, XXXXXDVPDYSXXX-		Expanded motif

	42	36	AENEDWEDYAST, -XXXXXDWEDYAXXX-		Expanded motif
	24	34	YSNVDLMDYAEP, -XXXXXDLMDYAXX--		Expanded motif
	46	34	TSVDVHDYSAHF, XXXXXDVHDYSXXXX-		Expanded motif
	27	33	LPVHWYDYPDSF, -XXXXXDYPDYAXXX-		Expanded motif
	59	33	FDWHDYAEHVQS, -XXXXXDWHDYAXXXX		Expanded motif
	21	32	WDMADYAEADHL, XXXXXDMADYAXXX-		Expanded motif
	22	30	FRKWDKQDYAYP, --XXXXXDKQDYAXX--		Expanded motif

	60	30	MYRFDRRDYSDQ, -XXXXDXRDYSDXXX-		Expanded motif
	63	30	FSLADKADYAAQ, XXXXXDXADYAXX--		Expanded motif
	48	29	WLQDLQDYSHAP, -XXXXDLQDYSXXXX		Expanded motif
	53	29	MMMVDSPDYAAN, XXXXXDXPDYAXX-		Expanded motif
	30	28	VSNTNYDLEDYS, -XXXXXDLEDYSXXX--		Expanded motif
	44	28	VSTADRHDYAYL, XXXXXDRHDYAXXX-		Expanded motif
	50	28	HFNWYDWHDYSF, XXXXXDXHDYSXXXX-		Expanded motif

	56	27	MMTEDPRDYAFF, -XXXXXDPRDYAXX---		Expanded motif
	67	27	WMMPGDADPYAD, -XXXXXDXDPYAXX--		Expanded motif
	14	26	LTDVMDYAAKEA, -XXXXXDVMDYAXXXX-		Expanded motif
	29	26	YFEDQEDYAGWS, -XXXXXDQEDYAXX-		Expanded motif
	40	26	VNSYADTLDYAD, XXXXXDXXDYADX--		Expanded motif
	7	25	SVEDDAPDYADF, -XXXXXDAPDYAXX---		Expanded motif
	15	25	WWHDQHDYAHWT, -XXXXDQHDYAXXXX-		Expanded motif

	33	24	FLTQQDREDYAH, -XXXXXDREDYAXX---		Expanded motif
	55	24	WWEATADTEDYA, -XXXXXDTEDYAXX--		Expanded motif
	62	24	VVGGLDTQDYAH, XXXXXDXQDYAX--		Expanded motif
	64	24	FDFHDYAYNQGM, XXXXXDFHDYAXXXX-		Expanded motif
	36	23	YGMLDQPDYAAY, -XXXXXDQPDYAXXX-		Expanded motif
	47	23	ELAYYDTYDYAD, XXXXXDXXDYAXX--		Expanded motif
	57	23	WDTHDYAAWSGT, XXXXXDTHDYAXXXX-		Expanded motif

	66	22	VLWTFDQADYAE, XXXXXDXADYAX--		Expanded motif
	52	18	DVRDYADDKYYE, XXXXXDVRDYAXXXX-		Expanded motif
	38	16	AGFDKKDYADAF, XXXXXDXKDYAXXX-		Expanded motif
	0	1102	, ----XXXXDXXDYXXXX---		Weak
	4	16	FYWNEMFWDHQP, ---XXXXWXXXXXXXXX-		Weak

5.7.5. Motif-based clustering sensitivity of UMAP dimensionality reduction

For this analysis, specific data were isolated from the AS-MS data. Specifically, a variable number of unaligned peptides containing the *DLHDYA* motif were added to random library peptides (which do not contain the motif) for 5000 total. The motif *DLHDYA* was used since it was discovered by clustering of the 12ca5 AS-MS data, most clearly seen in the UMAP + N-grams encoding analysis.

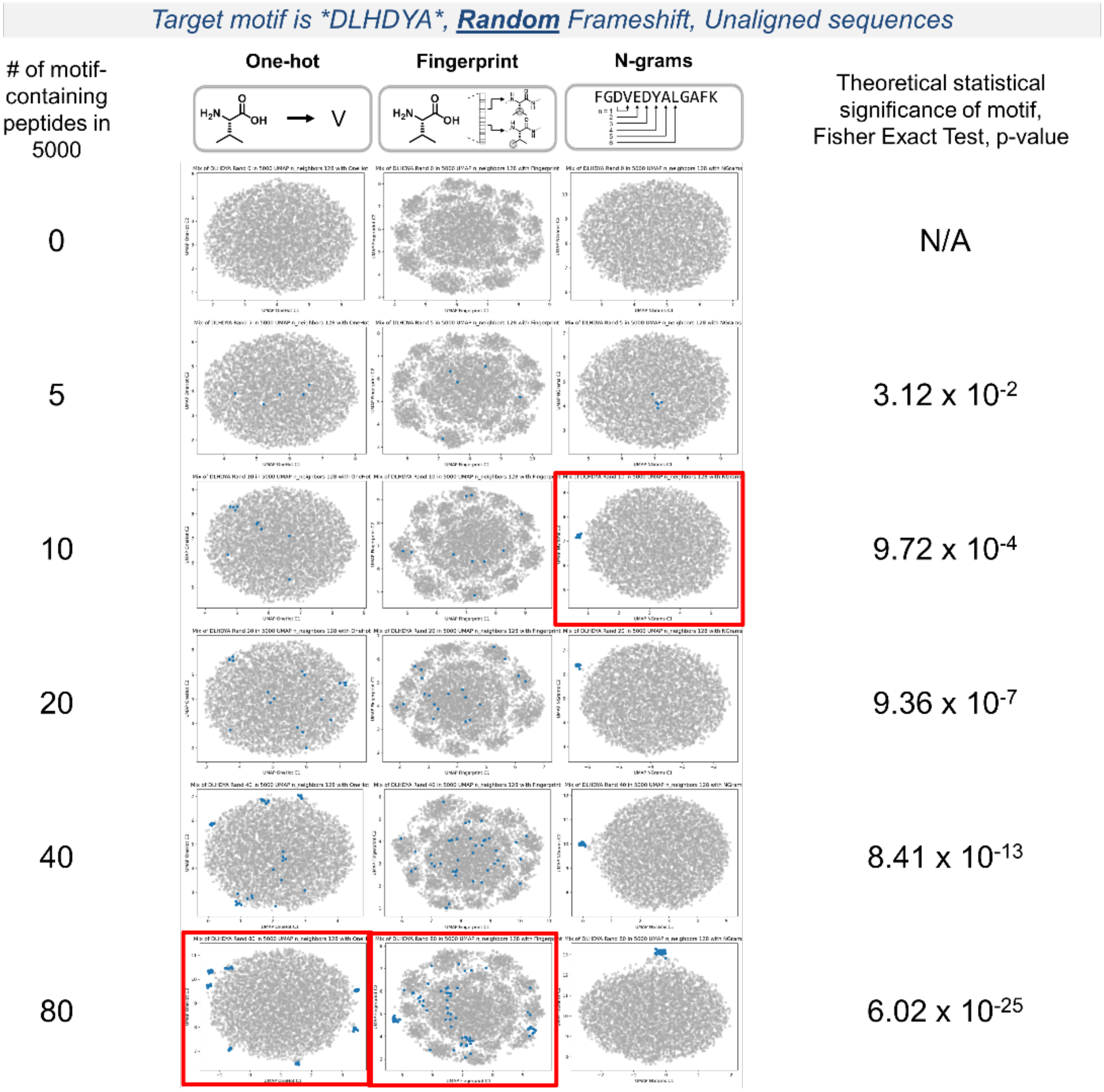


Figure 5.24: *UMAP sensitivity to cluster and enable the detection and isolation of target peptides in a 5000-peptide dataset. Unaligned target peptides contain the high-affinity binding motif of *DLHDYA* at random frameshifts. N-grams demonstrates the lowest sensitivity, with only 10 peptides required for a distinct cluster to appear. One-hot and Fingerprint encoding requires 80 and 160 peptides, respectively. This result is because N-grams encoding is performed irrespective of frameshift, whereas one-hot and Fingerprint encoding are frameshift sensitive. Thus, as the number of target peptides increases, one-hot and Fingerprint encoded UMAP sequence maps form seven clusters as the seven frameshifts of *DLHDYA* in a 12-mer variable region are populated to have at least 10 peptides in each cluster. A red box is placed to guide the readers eye to location in which clusters appear to form distinctly from the random library peptides. AS-MS peptides are*

*shown in blue with random library peptides in gray. The theoretical statistical significance via Fishers Exact Test of each condition is shown,[11–13] indicating that at only 5 sequences, the peptides with the *DLHDYA* motif could be theoretically distinguished from the background (randomized input dataset), though 10 are required for a clear cluster to form.*

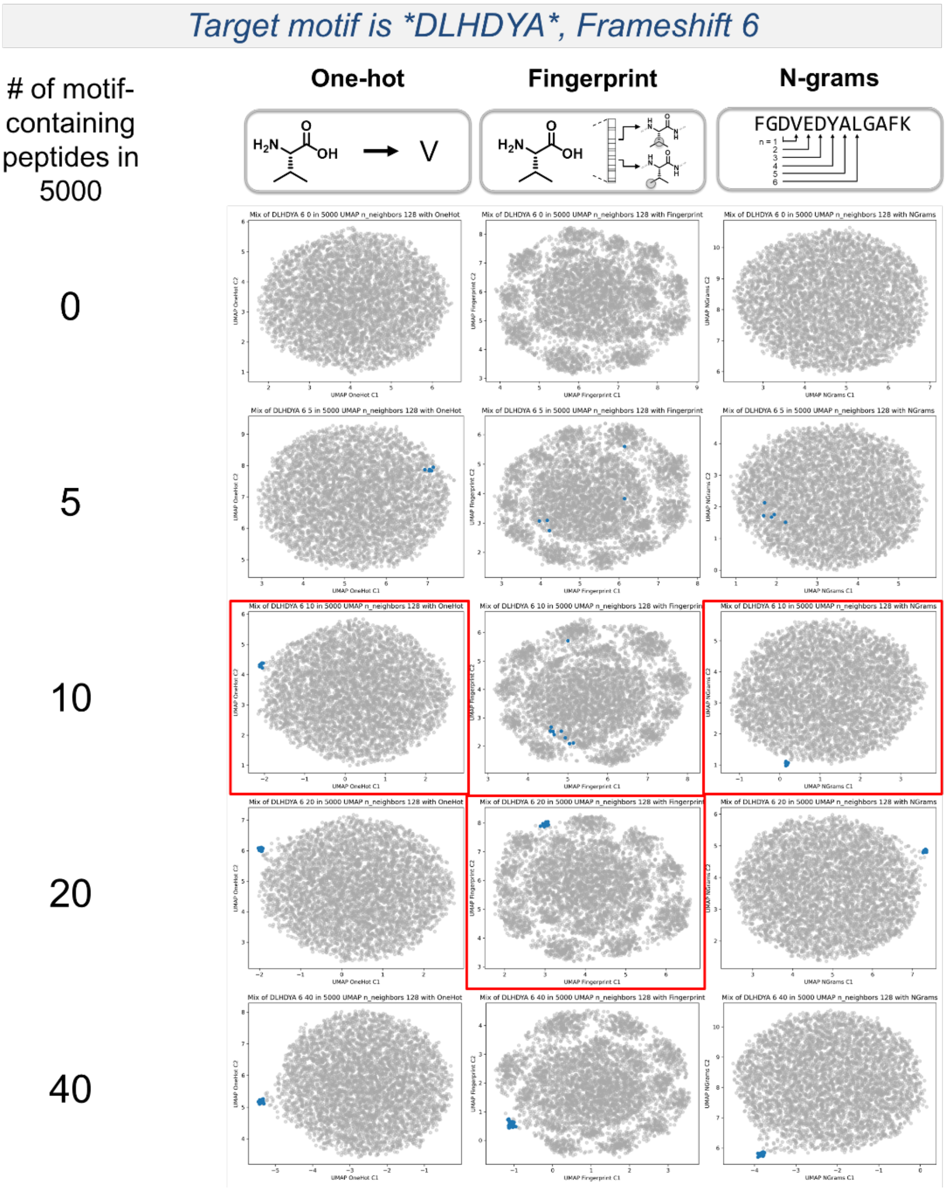


Figure 5.25: *N-grams, one-hot, and Fingerprint encoding provide similar clustering sensitivity with target peptides containing a motif at the same frameshift. See Figure 5.22 for further details. A red box is placed to guide the readers eye to location in which clusters appear to form distinctly from the random library peptides. AS-MS peptides are shown in blue with random library peptides in gray.*

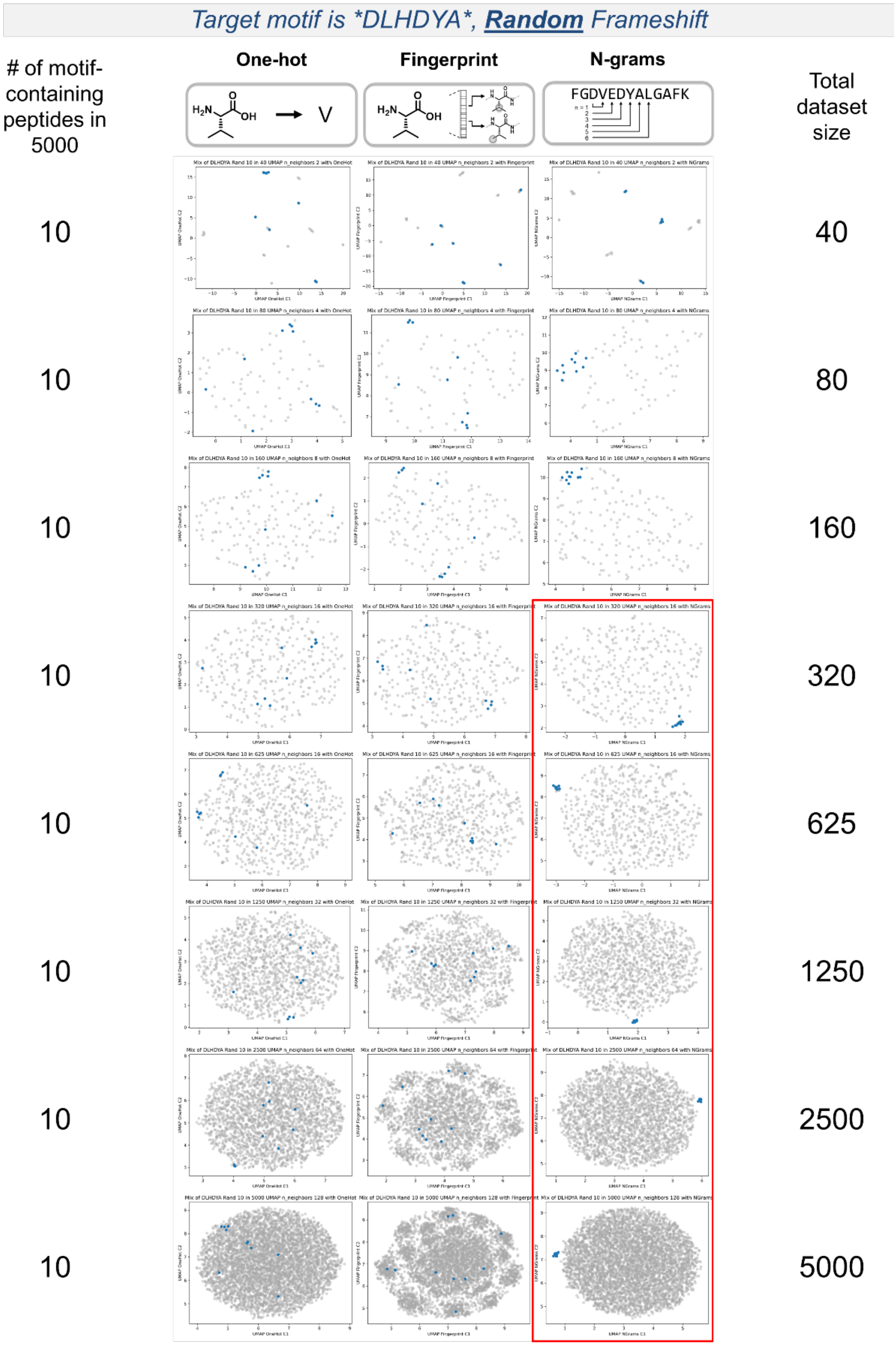


Figure 5.26: *The construction of UMAP sequence space is affected by the total dataset size. At low dataset sizes, highly similar peptides can be dispersed on the sequence space map. Thus, augmenting the total dataset size with random library peptides can sometimes improve clarity of the clusters of similar peptides.*

5.7.6. Comparison of motif-detection sensitivity with XSTREME

Motif discovery was performed using the XSTREME, part of the MEME Suite webserver.[61,62]

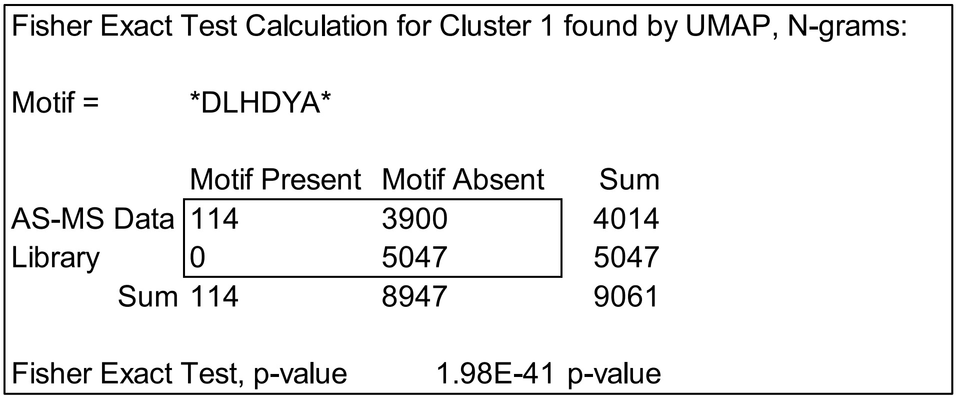
XSTREME combines:

- MEME, which discovers novel, ungapped motifs (recurring, fixed-length patterns) in sequences. MEME will split variable-length patterns into two or more separate motifs.
- STREME, which discovers ungapped motifs (recurring, fixed-length patterns) that are enriched in sequences or relatively enriched in comparison to a control dataset.

Two experiments were performed

1. The AS-MS data was input to XSTREME as the positive dataset with the randomly sampled library peptides as the negative dataset

The Fisher Exact Test can quantify the statistical significance of finding a specific motif, and is used by STREME when a background dataset is input. The motif *DLHDYA*, found in the clustering analysis using UMAP and N-grams encoding. The p-value is 1.98 x 10^{-41}, meaning it should be detected (see below)



Fisher Exact Test Calculation for Cluster 1 found by UMAP, N-grams:

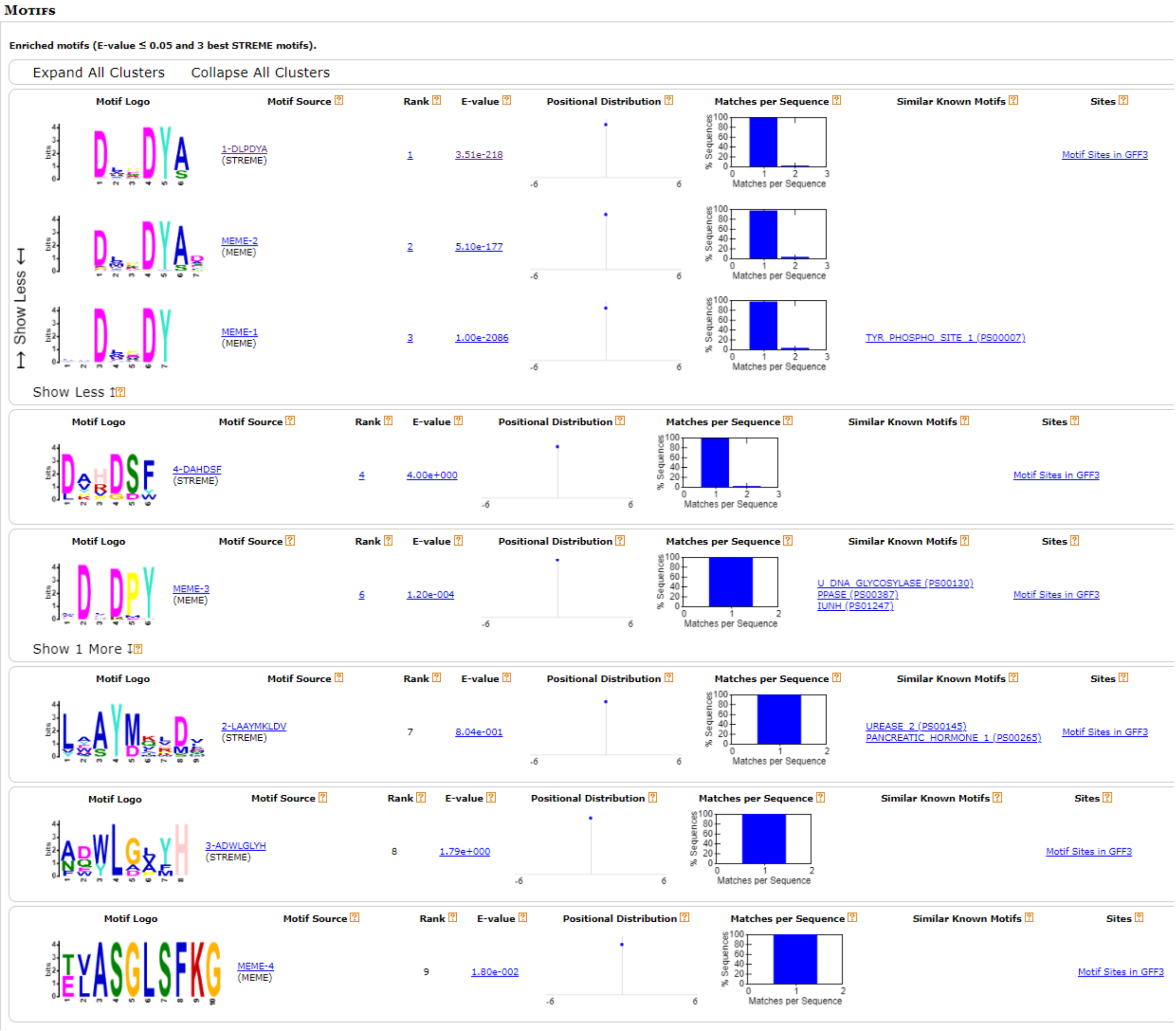
Motif = *DLHDYA*

	Motif Present	Motif Absent	Sum
AS-MS Data	114	3900	4014
Library	0	5047	5047
Sum	114	8947	9061

Fisher Exact Test, p-value 1.98E-41 p-value

2. The sensitivity of motif detection was determined using the same datasets in Figure 5.22, using either 5, 10, or 20 target peptides that contain a *DLHDYA* motif at random frameshifts.

5.7.6.1. XSTREME Experiment 1 (12ca5 AS-MS data vs library)



Motifs

Enriched motifs (E-value ≤ 0.05 and 3 best STREME motifs).

Expand All Clusters   Collapse All Clusters

Motif Source	Rank	E-value	Similar Known Motifs	Sites
1-DLPDYA (STREME)	1	3.51e-218		Motif Sites in GFF3
MEME-2 (MEME)	2	5.10e-177		
MEME-1 (MEME)	3	1.00e-2086	TYR_PHOSPHO_SITE_1 (PS00007)	
4-DAHDSF (STREME)	4	4.00e+000		Motif Sites in GFF3
MEME-3 (MEME)	6	1.20e-004	U_DNA_GLYCOSYLASE (PS00130) PPASE (PS00387) IUNH (PS01247)	Motif Sites in GFF3
2-LAAYMKLDV (STREME)	7	8.04e-001	UREASE_2 (PS00145) PANCREATIC_HORMONE_1 (PS00265)	Motif Sites in GFF3
3-ADWLGLYH (STREME)	8	1.79e+000		Motif Sites in GFF3
MEME-4 (MEME)	9	1.80e-002		Motif Sites in GFF3

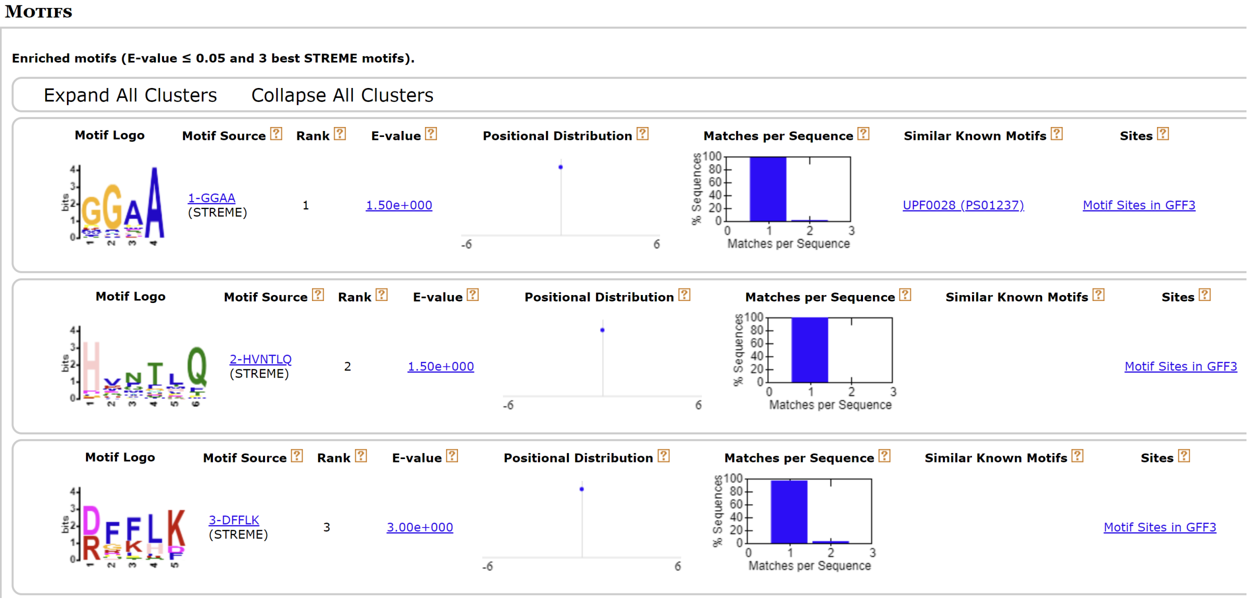
Show Less ↕   Show 1 More ↕

Figure 5.27: *XSTREME motif detection result of motifs enriched in the AS-MS dataset (positive) relative to the randomly sampled library peptides (negative). Boxed in red are the common motif D**DYA, as well as D**DYAD* and D*DPY* which were the only expanded motif discovered with statistical significance.*

5.7.6.2. XSTREME Experiment 2 (Analysis of detection sensitivity of unaligned, motif-containing peptides)

Next, the detection sensitivity was assessed using the same datasets as in Figure 5.22 with 5, 10, and 20 target peptides, containing an unaligned *DLHDYA* motif in dataset of 5000 random library peptides.

For our clustering approach, all 5,000 sequences were input, whereas for XSTREME analysis, the same 5,000 input sequences were compared against a background dataset constructed from the randomization of the input sequences.



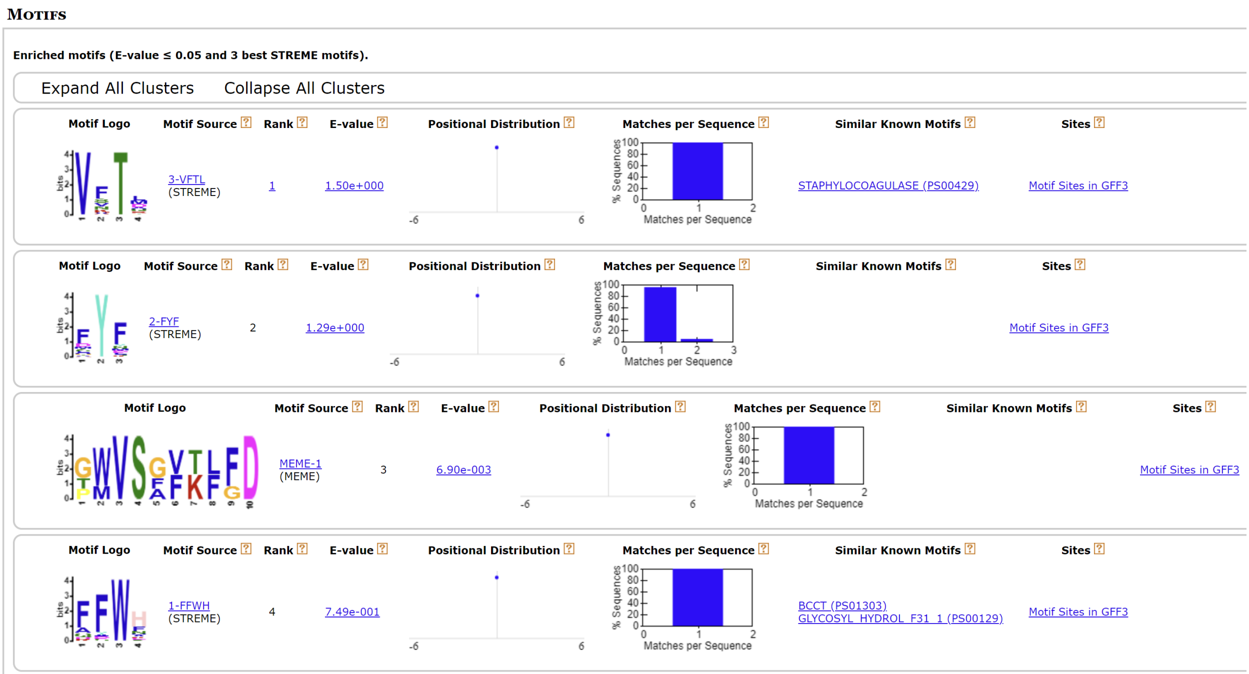
Motifs

Enriched motifs (E-value ≤ 0.05 and 3 best STREME motifs).

Expand All Clusters Collapse All Clusters

Motif Logo	Motif Source	Rank	E-value	Positional Distribution	Matches per Sequence	Similar Known Motifs	Sites
	1-GGAA (STREME)	1	1.50e+000			UPF0028 (PS01237)	Motif Sites in GFF3
	2-HVNTLQ (STREME)	2	1.50e+000				Motif Sites in GFF3
	3-DFFLK (STREME)	3	3.00e+000				Motif Sites in GFF3

Figure 5.28: *The XSTREME results for motif discovery and detection using the dataset of 5 target peptides in 5000 random library peptides. None of the motifs are statistically significant and the 5 *DLHDYA* peptides were not identified. STREME reported all these motifs, and motifs evaluated by the Binomial Test, providing the p-value reported.*



Motifs

Enriched motifs (E-value ≤ 0.05 and 3 best STREME motifs).

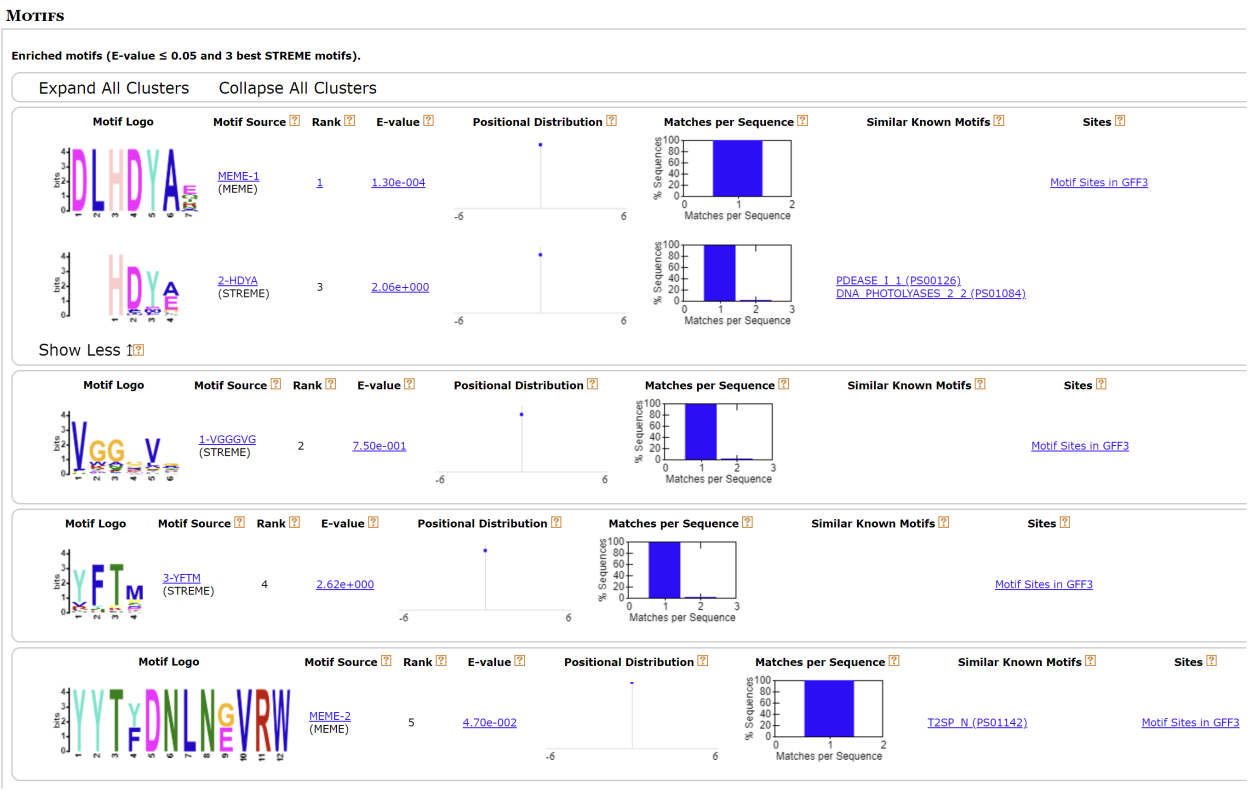
Motif Logo	Motif Source	Rank	E-value	Positional Distribution	Matches per Sequence	Similar Known Motifs	Sites
VFTL	3-VFTL (STREME)	1	1.50e+000			STAPHYLOCOAGULASE (PS00429)	Motif Sites in GFF3

Motif Logo	Motif Source	Rank	E-value	Positional Distribution	Matches per Sequence	Similar Known Motifs	Sites
FYF	2-FYF (STREME)	2	1.29e+000				Motif Sites in GFF3

Motif Logo	Motif Source	Rank	E-value	Positional Distribution	Matches per Sequence	Similar Known Motifs	Sites
GWVSGVTLFD	MEME-1 (MEME)	3	6.90e-003				Motif Sites in GFF3

Motif Logo	Motif Source	Rank	E-value	Positional Distribution	Matches per Sequence	Similar Known Motifs	Sites
FFWH	1-FFWH (STREME)	4	7.49e-001			BCCT (PS01303) GLYCOSYL_HYDROL_F31_1 (PS00129)	Motif Sites in GFF3

***Figure 5.29:** The XSTREME and STREME results for motif discovery and detection using the dataset of 10 target peptides in 5000 random library peptides. The 10 *DLHDYA* peptides were not identified. STREME reported motifs were evaluated by the Fisher Exact Test, providing the p-value (E-value * # of reported sequences) reported. MEME reported motifs were evaluated by E-value.*



MOTIFS

Enriched motifs (E-value ≤ 0.05 and 3 best STREME motifs).

Expand All Clusters   Collapse All Clusters

Motif Logo	Motif Source	Rank	E-value	Positional Distribution	Matches per Sequence	Similar Known Motifs	Sites
	MEME-1 (MEME)	1	1.30e-004				Motif Sites in GFF3
	2-HDYA (STREME)	3	2.06e+000			PDEASE_I_1 (PS00126) DNA_PHOTOLYASES_2_2 (PS01084)	

Show Less

Motif Logo	Motif Source	Rank	E-value	Positional Distribution	Matches per Sequence	Similar Known Motifs	Sites
	1-VGGGVG (STREME)	2	7.50e-001				Motif Sites in GFF3

Motif Logo	Motif Source	Rank	E-value	Positional Distribution	Matches per Sequence	Similar Known Motifs	Sites
	3-YFTM (STREME)	4	2.62e+000				Motif Sites in GFF3

Motif Logo	Motif Source	Rank	E-value	Positional Distribution	Matches per Sequence	Similar Known Motifs	Sites
	MEME-2 (MEME)	5	4.70e-002			T2SP_N (PS01142)	Motif Sites in GFF3

Figure 5.30: *The XSTREME and STREME results for motif discovery and detection using the dataset of 20 target peptides in 5000 random library peptides. The 20 *DLHDYA* peptides were identified, and were calculated to be statistically significant with an E-value of 1.3 x 10-4 from discovery by MEME analysis. This is twice as many as can be clearly seen by our clustering approach.*

5.7.7. Augmentation of sequence maps with noncanonical peptides discovered by AS-MS

Table 5.18: *Peptidomimetics discovered using AS-MS for purity and LCMS characterization see Analytical characterization of all synthesized noncanonical peptidomimetics discovered by AS-MS. For BLI characterization see Section 5.5.15 Biolayer interferometry (BLI) measurements, Table 5.19*

Peptide #	Sequence, 1-letter code	ALC	Sequence, 3-letter code for noncanonicals													Binder / Nonbinder	KD, nM (Ave ± SD)
1	HoiDueDYAoxPK	90	H	hArg	Tha	D	Nal	Hyp	D	Y	A	hArg	Psa	P	Lys	Binder	44 ± 29
2	duiDueDYAoxPK	98	Cpa	Nal	Tha	D	Nal	Hyp	D	Y	A	hArg	Psa	P	Lys	Binder	75 ± 56
3	giibmDpoDYAiK	99	Thp	Tha	Tha	Aib	3fF	D	hCit	hArg	D	Y	A	Tha	Lys	Binder	3.1 ± 0.67
5	tzwksnYVkuliK	93	Cxf	Dpf	Dph	4Af	Php	pSer	Y	V	4Af	Nal	Msn	Tha	Lys	Binder	77 ± 57
15	pgYDwDVADYADK	91	hCit	Thp	Y	D	Dph	D	V	A	D	Y	A	D	Lys	Binder	3.9 ± 0.68
16	jVVdDQPDYAtlK	99	Tic	V	V	Cpa	D	Q	P	D	Y	A	Cxf	Msn	Lys	Binder	0.21 ± 0.15
17	xPAGDTPDYADmK	93	Psa	P	A	G	D	T	P	D	Y	A	D	3fF	Lys	Binder	4.4 ± 2.7
4	ovuxjvVrbevGK	94	hArg	2F3F	Nal	Psa	Tic	2F3F	V	DfF	Aib	Hyp	2F3F	G	Lys	Nonbinder	
6	ktGwzTQwpptZK	91	4Af	Cxf	G	Dph	Dpf	T	Q	Dph	hCit	hCit	Cxf	Git	Lys	Nonbinder	
7	jmHVGwhYAQAHK	90	Tic	3fF	H	V	G	Dph	Amb	Y	A	Q	A	H	Lys	Nonbinder	
8	irhTAsjViDYAK	88	Tha	DfF	Amb	T	A	Php	Tic	V	Tha	D	Y	A	Lys	Nonbinder	
9	uTxpzdpmmjTzK	87	Nal	T	Psa	hCit	Dpf	Cpa	hCit	3fF	3fF	Tic	T	Dpf	Lys	Nonbinder	
10	TNXfQYvoTYifK	84	T	N	Agn	Pip	Q	Y	2F3F	hArg	T	Y	Tha	Pip	Lys	Nonbinder	
11	iiAldjwTtswzK	84	Tha	Tha	A	Msn	Cpa	Tic	Dph	T	Cxf	Php	Dph	Dpf	Lys	Nonbinder	
12	NfXlKDbutvzdK	83	N	Pip	Agn	Msn	K	D	Aib	Nal	Cxf	2F3F	Dpf	Cpa	Lys	Nonbinder	
13	swrYPzTmjGexK	81	Php	Dph	DfF	Y	P	Dpf	T	3fF	Tic	G	Hyp	Psa	Lys	Nonbinder	
14	NrTzzdkYmjzTK	81	N	DfF	T	Dpf	Dpf	Cpa	4Af	Y	3fF	Tic	Dpf	T	Lys	Nonbinder	

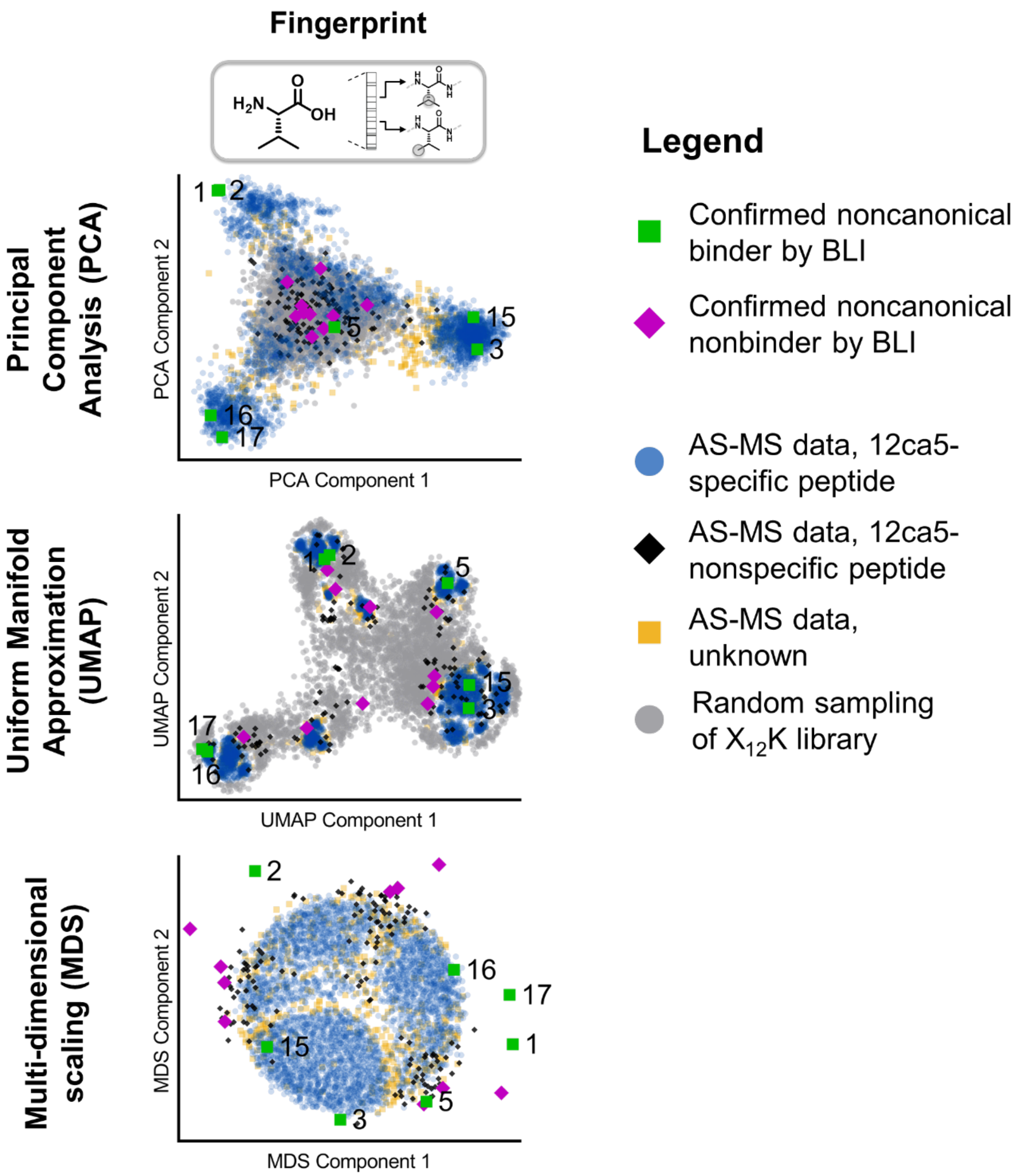


Figure 5.31: *Augmentation of canonical sequence maps with noncanonical peptides discovered from AS-MS and experimentally evaluated using BLI to distinguish binders from nonbinders (see Biolayer interferometry (BLI) measurements). Peptides are labeled with their respective numbers. Also included are the 12ca5-based labels as defined in Label definitions for 12ca5-specific and nonspecific binders. Seventeen noncanonical peptides were added to the dataset and the sequence space was relearned and then the randomly sampled peptides from the canonical $X_{12}K$ library were added to the PCA and UMAP maps. The randomly sampled peptides cannot be added to MDS without re-learning.*

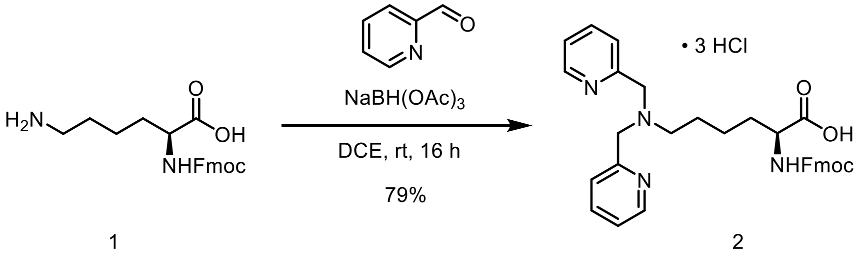
5.8. Appendix II: Synthesis, characterization, and biophysical measurements of discovered peptides containing unnatural amino acids

5.8.1. Synthesis of noncanonical momers

Reactions were monitored on glass-backed analytical thin-layer chromatography (TLC) plates (250 μm, 60 Å, SiliaPlate) containing a fluorescent indicator (254 nm). NMR spectra were recorded on a Bruker AVIII HD 400 MHz or Bruker Neo 500 MHz. ^{1}H NMR chemical shifts are reported in parts per million (ppm, δ scale) and are referenced to the residual protonated NMR solvent (DMSO-*d*6: δ 2.50). All ^{13}C spectra recorded are proton decoupled with chemical shifts reported in parts per million (ppm, δ scale) and are referenced to the carbon resonance of the NMR solvent (DMSO-*d*6: δ 39.5). ^{1}H NMR spectroscopic data are reported as follows: chemical shift in ppm (multiplicity, coupling constants J (Hz), assigned number of protons in molecule). The multiplicities are abbreviated with s (singlet), br. s (broad singlet), d (doublet), t (triplet), and m (multiplet). The chemical shift of all signals is reported as the center of the resonance range, except in the case of multiplets, which are reported as ranges in chemical shift. All raw fid files were processed, and the spectra analyzed using the program MestReNOVA 14.2 from Mestrelab Research S. L. High-resolution mass spectra were obtained on an Agilent Technologies 6550 Q-TOF LC/MS systems (see *Analysis methods with Liquid-Chromatography Mass Spectrometry (LC-MS)*).

5.8.1.1. Synthesis of Fmoc-Bpl-OH

N^{2}-(((9H-fluoren-9-yl)methoxy)carbonyl)-N^{6},N^{6}-bis(pyridin-2-ylmethyl)-L-lysine (Fmoc-Bpl-OH) (2)



To a 0°C suspension of Fmoc-Lys-OH (1.50 g, 4.07 mmol, 1.0 eq.) and $NaBH(OAc)_3$ (2.59 g, 12.2 mmol, 3.0 eq.) in dichloroethane (22.6 mL) under nitrogen atmosphere, 2-pyridinecarboxaldehyde (0.965 mL, 1.09 g, 10.2 mmol, 2.5 eq.) was added and the resulting suspension was stirred at rt for 16 h. After checking the completion of the reaction by LC-MS, the suspension was cooled to 0°C and quenched by addition of MeOH (25 mL). The resulting solution was concentrated under reduced pressure, the residue redissolved in 4:1 $MeCN/H_2O$ and purified by reverse phase column chromatography (Biotage® Sfär C18 D Duo 100 Å 30 µm 30 g, MeCN + 0.1% HCl : H_2O + 0.1% HCl = 1:9 → 4:1) to afford the title compound as dark yellow solid (2.12 g, 79%). 0.1% HCl was used rather than 0.1% trifluoroacetic acid to prevent against any possible trifluoracetylation during coupling.

ESI-HRMS: calc. $C_{33}H_{34}N_4O_4$ $[M+H]^+$ 551.2658 found 551.2714, 10.2 ppm error.

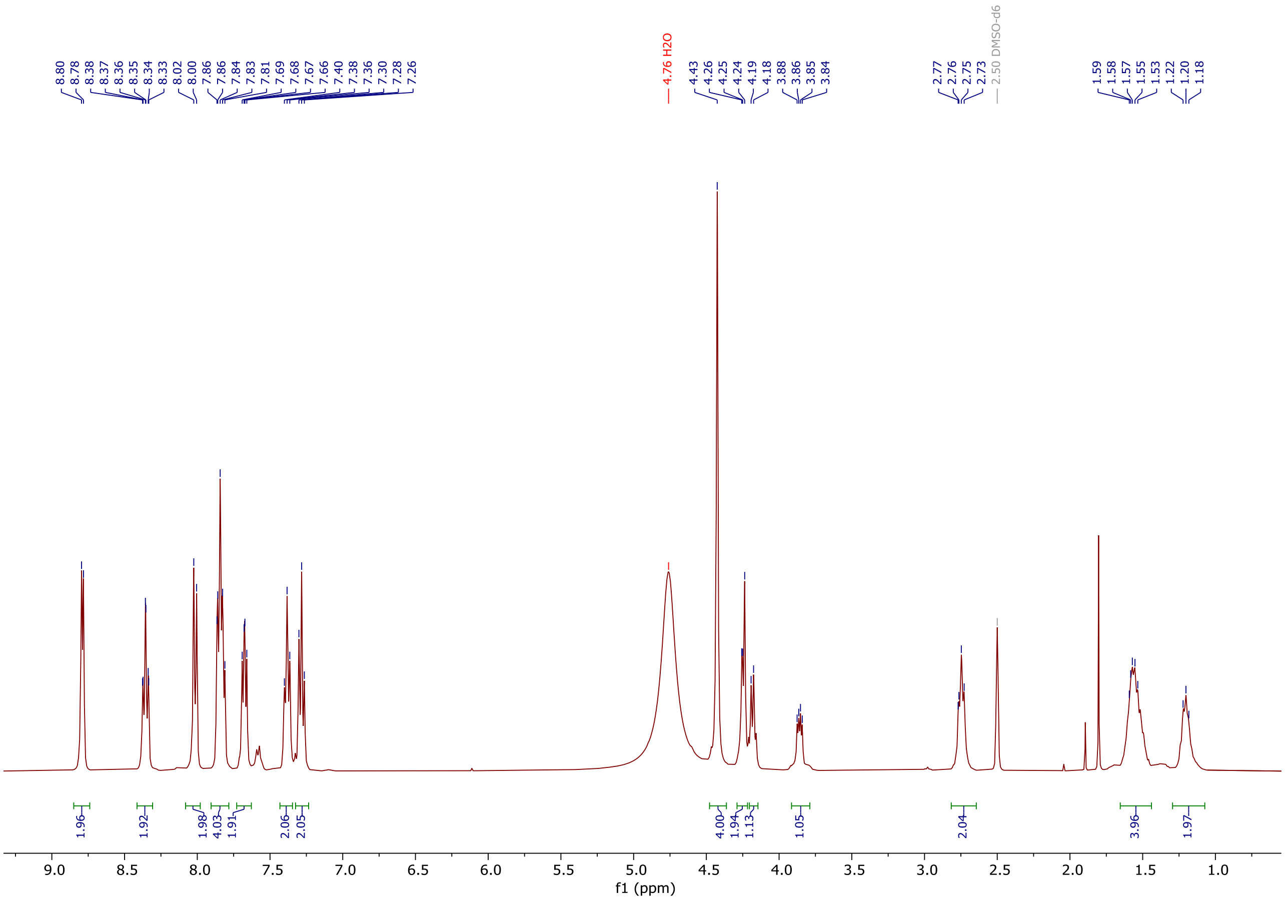


Figure 5.32: 1H NMR (400 MHz, DMSO-d6 + 1% D2O) of Fmoc-Bpl-OH: *δ 8.79 (d, J = 5.4 Hz, 2H), 8.40-8.31 (m, 2H), 8.01 (2H, J = 7.8 Hz, 2H), 7.89-7.79 (m, 4H), 7.72-7.64 (m, 2H), 7.38 (t, J = 7.4 Hz, 2H), 7.28 (t, J = 7.5 Hz, 2H), 4.43 (s, 2H), 4.29-4.14 (m, 3H), 3.90-3.82 (m, 1H), 2.75 (t, J = 8.3 Hz, 2H), 1.65-1.41 (m, 4H), 1.27-1.12 (m, 2H).*

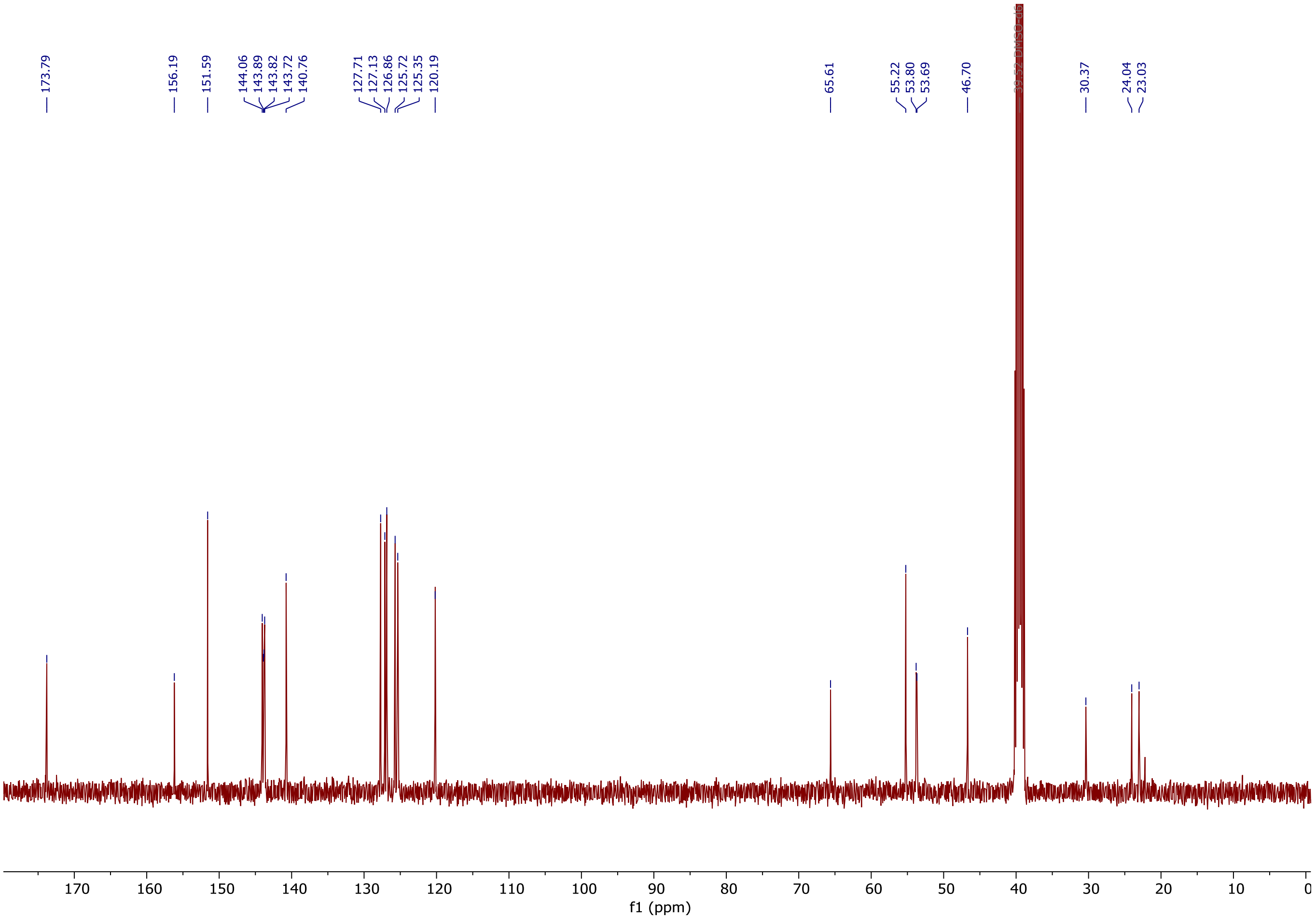
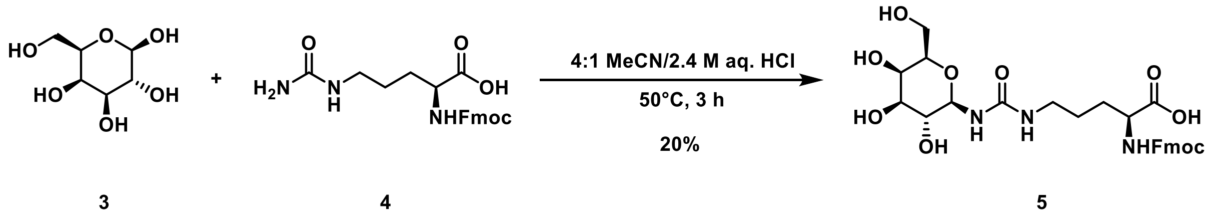


***Figure 5.33: ^{13}C NMR (101 MHz, DMSO-d6) of Fmoc-Bpl-OH:** δ 173.8, 156.2, 151.6, 144.1, 143.9, 143.7, 140.8, 127.7, 127.1, 126.9, 125.7, 125.4, 120.2, 65.61, 55.2, 53.8, 53.7, 46.7, 30.4, 24.0, 23.0.*

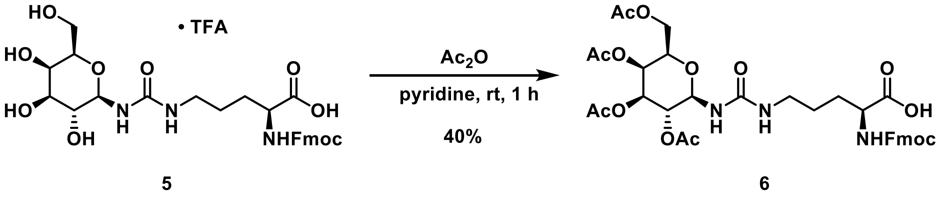
5.8.1.2. Synthesis of Fmoc-Git-OH

(S)-2-((((9H-fluoren-9-yl)methoxy)carbonyl)amino)-5-(3-((2R,3R,4S,5R,6R)-3,4,5-trihydroxy-6-(hydroxymethyl)tetrahydro-2H-pyran-2-yl)ureido)pentanoic acid (5)



A suspension of D-(+)-galactose (1.50 g, 8.33 mmol, 1.0 eq.) and Fmoc-Cit-OH (4.30 g, 10.8 mmol, 1.30 eq.) in 4:1 MeCN/2.4 M aq. HCl was heated to 50°C for 3 h. The mixture was concentrated and purified by reverse phase column chromatography (Biotage® Sfär C18 Duo 100 Å 30 µm 30 g, MeCN + 0.1% TFA/H_2O + 0.1% TFA = 1:9 → 1:1) to afford the title compound as white solid (1.11 g, 20%) that was used for the next step without further purification.

(S)-2-((((9H-fluoren-9-yl)methoxy)carbonyl)amino)-5-(3-((2R,3R,4S,5S,6R)-3,4,5-triacetoxy-6-(acetoxymethyl)tetrahydro-2H-pyran-2-yl)ureido)pentanoic acid (Fmoc-Git-OH) (6)



To a solution of (S)-2-((((9H-fluoren-9-yl)methoxy)carbonyl)amino)-5-(3-((2R,3R,4S,5R,6R)-3,4,5-trihydroxy-6-(hydroxymethyl)tetrahydro-2H-pyran-2-yl)ureido)pentanoic acid TFA salt (1.11 g, 1.65 mmol, 1.0 eq.) in pyridine (8.24 mL), acetic anhydride (7.79 mL, 8.41 g, 82.4 mmol, 50 eq.) was added and the resulting solution was stirred at rt for 1 h. After completion of the reaction, the mixture was cooled to 0°C and quenched with 2.4 M aq. HCl. The suspension was diluted with Et_2O (50 mL) and the aqueous phase was extracted with Et_2O (5x). The combined organic layers were dried over anhydrous $MgSO_4$, concentrated under reduced pressure and purified by reverse phase column chromatography (Biotage® Sfär C18 D Duo 100 Å 30 µm 30 g, MeCN + 0.1% HCl : H_2O + 0.1% HCl = 1:19 → 4:1) to yield the title compound as white solid (481 mg, 40%).

ESI-HRMS: calc. $C_{35}H_{42}N_3O_{14}$ $[M+H]^+$ 728.2667 found 728.2666. -0.1 ppm error.

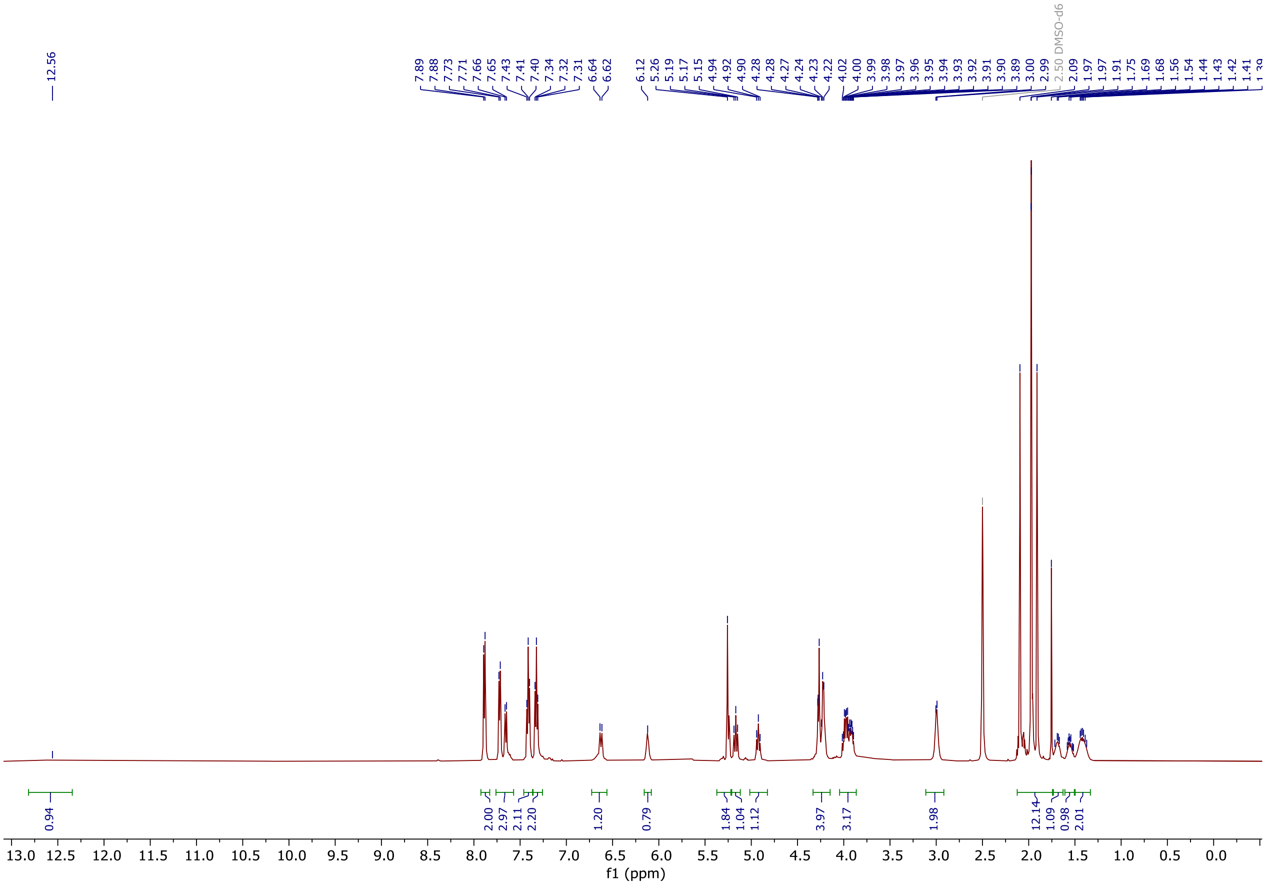


Figure 5.34: ^{1}H NMR (500 MHz, DMSO-d6) of Fmoc-Git-OH: *δ 12.56 (br s, 1H), 7.89 (d, J = 7.5 Hz, 2H), 7.72 (d, J = 7.4 Hz, 2H), 7.65 (d, J = 8.0 Hz, 1H), 7.41 (t, J = 7.4 Hz, 2H), 7.32 (t, J = 7.4 Hz, 2H), 6.63 (d, J = 10.2 Hz, 1H), 6.12 (s, 1H), 5.32 – 5.21 (m, 2H), 5.17 (t, J = 9.7 Hz, 1H), 4.92 (t, J = 9.4 Hz, 1H), 4.32 – 4.15 (m, 4H), 4.08 – 3.82 (m, 3H), 4.08 – 3.82 (m, 2H), 2.09 (s, 3H), 2.02 – 1.92 (m, 6H), 1.91 (s, 3H), 1.75 – 1.63 (m, 1H), 1.62 – 1.49 (m, 1H), 1.47 – 1.39 (m, 2H).*

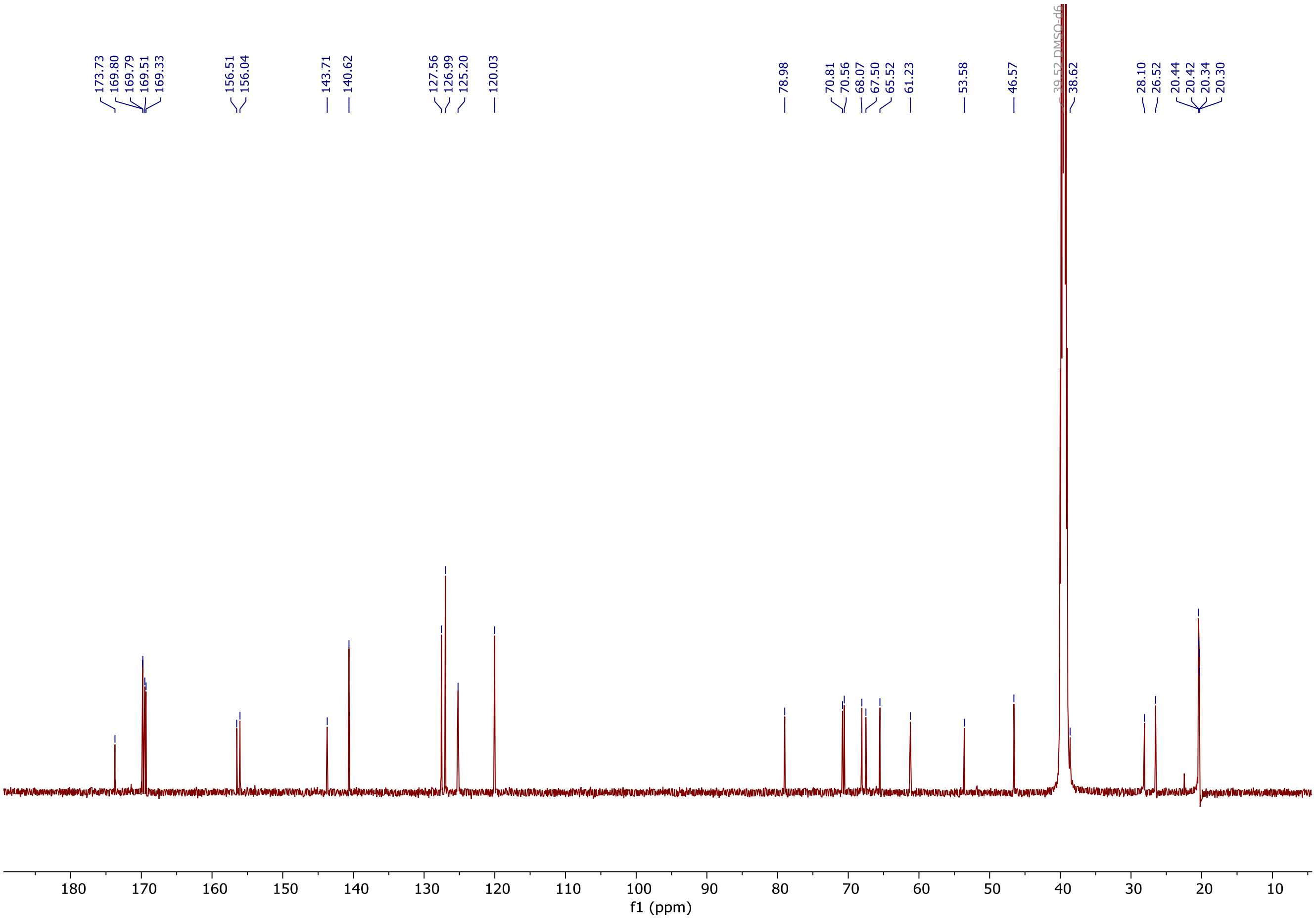
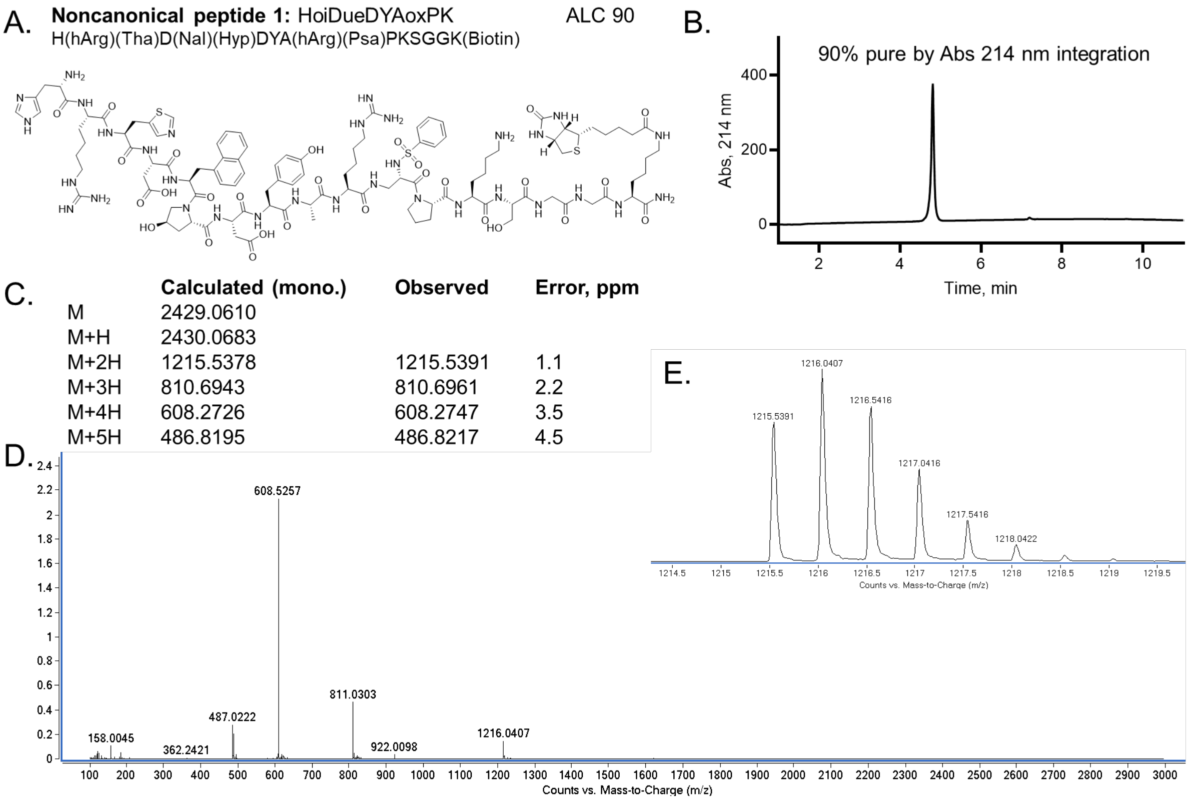


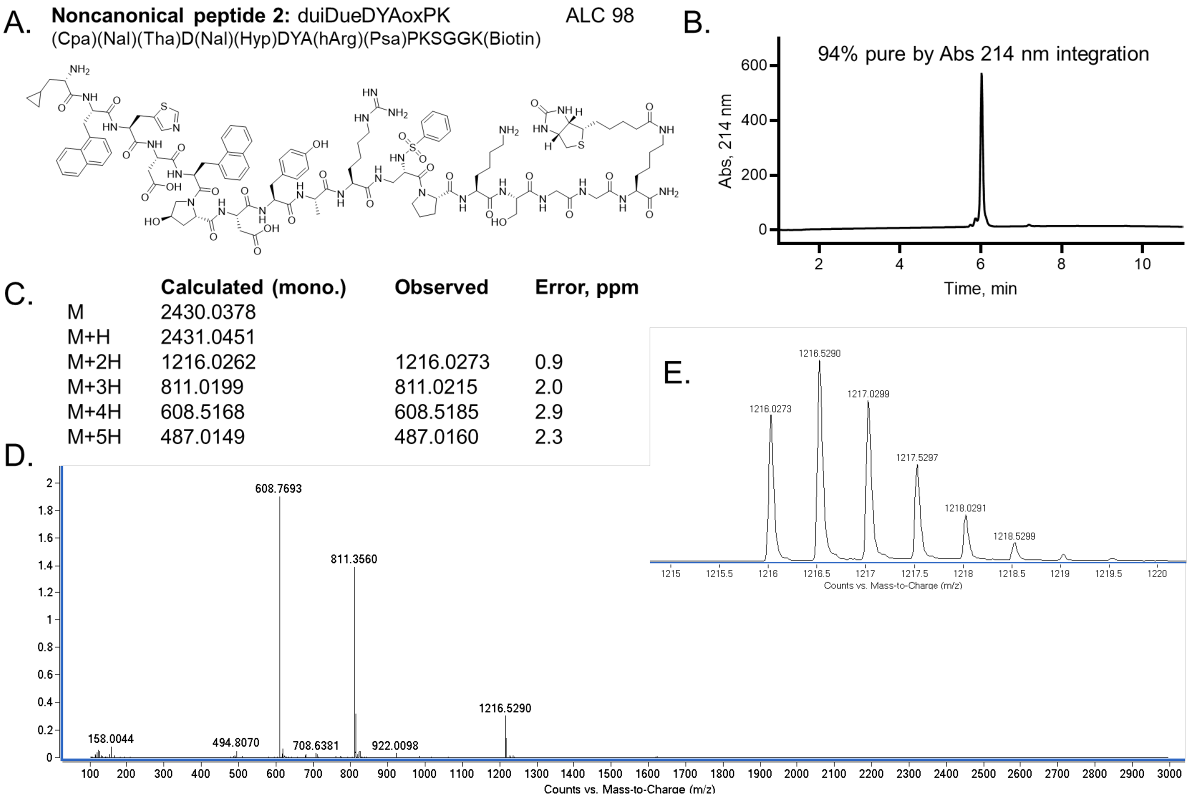
Figure 5.35: ^{13}C NMR (126 MHz, DMSO-d6) of Fmoc-Git-OH: *δ 173.7, 169.8, 169.8, 169.5, 169.3, 156.5, 156.0, 143.7, 140.6, 127.6, 127.0, 125.2, 120.0, 79.9, 70.8, 70.6, 68.0, 67.5, 65.5, 61.2, 53.6, 46.6 38.6, 28.1, 26.5, 20.4, 20.4, 20.3, 20.3.*

5.8.2. Analytical characterization of all synthesized noncanonical peptidomimetics discovered by AS-MS



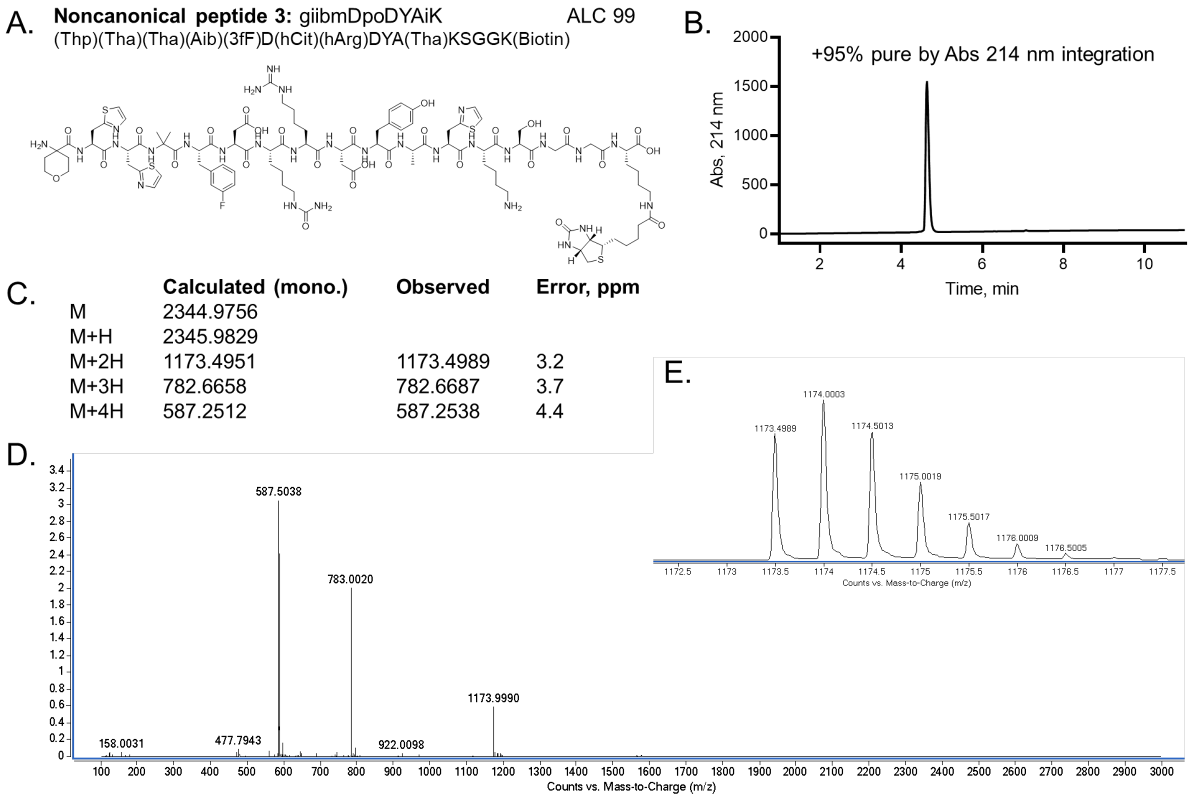
	Calculated (mono.)	Observed	Error, ppm
M	2429.0610		
M+H	2430.0683		
M+2H	1215.5378	1215.5391	1.1
M+3H	810.6943	810.6961	2.2
M+4H	608.2726	608.2747	3.5
M+5H	486.8195	486.8217	4.5

Figure 5.36: *Analytical characterization of purified Noncanonical Peptide 1. (A) Sequence information including 1-letter and 3-letter codes for the noncanonical amino acids and average local confidence (ALC) of each peptide. (B) Purity and UPLC chromatogram (C) Calculated and observed monoisotopic masses with ppm error reported. (D) Raw mass spectra of the peptide showing the charge state series (often z = 2,3,4 observed), and E. ~5 m/z zoom in on the lowest charge species observed (often z = 2).*



	Calculated (mono.)	Observed	Error, ppm
M	2430.0378		
M+H	2431.0451		
M+2H	1216.0262	1216.0273	0.9
M+3H	811.0199	811.0215	2.0
M+4H	608.5168	608.5185	2.9
M+5H	487.0149	487.0160	2.3

Figure 5.37: *Analytical characterization of purified Noncanonical Peptide 2. (A) Sequence information including 1-letter and 3-letter codes for the noncanonical amino acids and average local confidence (ALC) of each peptide. (B) Purity and UPLC chromatogram (C) Calculated and observed monoisotopic masses with ppm error reported. (D) Raw mass spectra of the peptide showing the charge state series (often z = 2,3,4 observed), and E. ~5 m/z zoom in on the lowest charge species observed (often z = 2).*

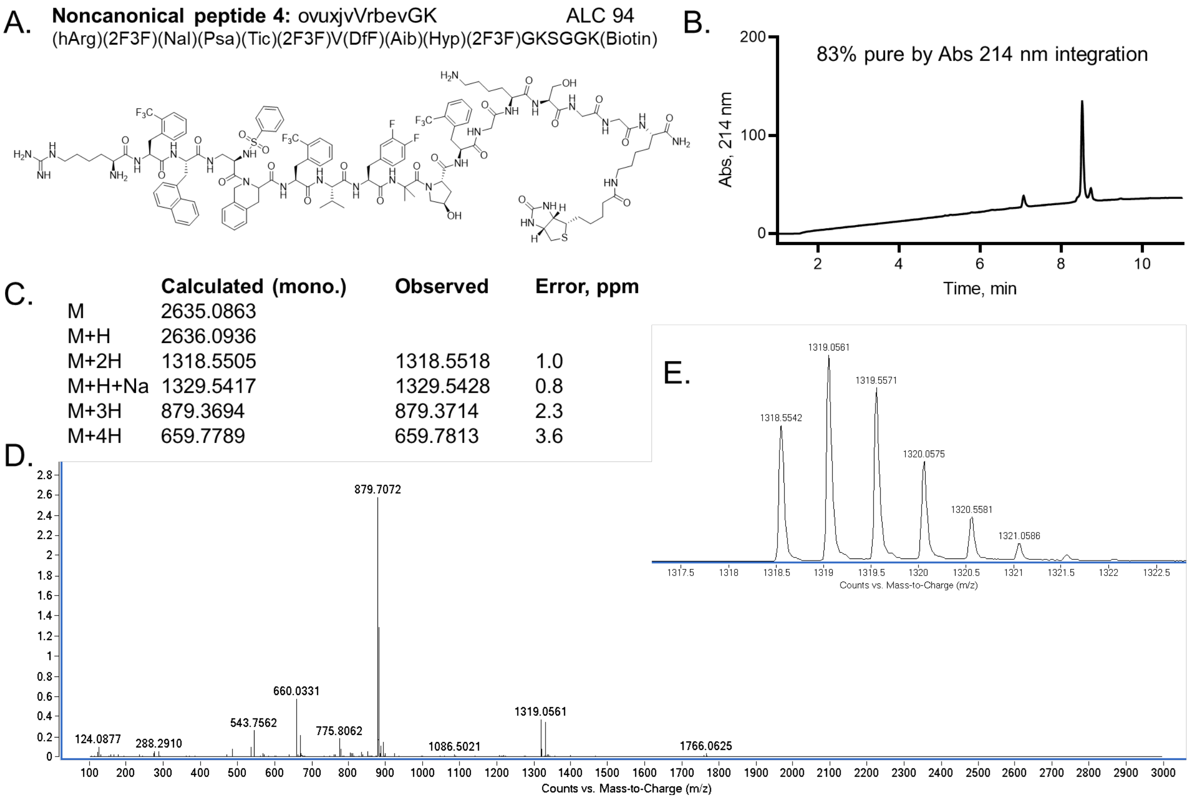


A. **Noncanonical peptide 3:** giibmDpoDYAiK ALC 99
(Thp)(Tha)(Tha)(Aib)(3fF)D(hCit)(hArg)DYA(Tha)KSGGK(Biotin)

C.

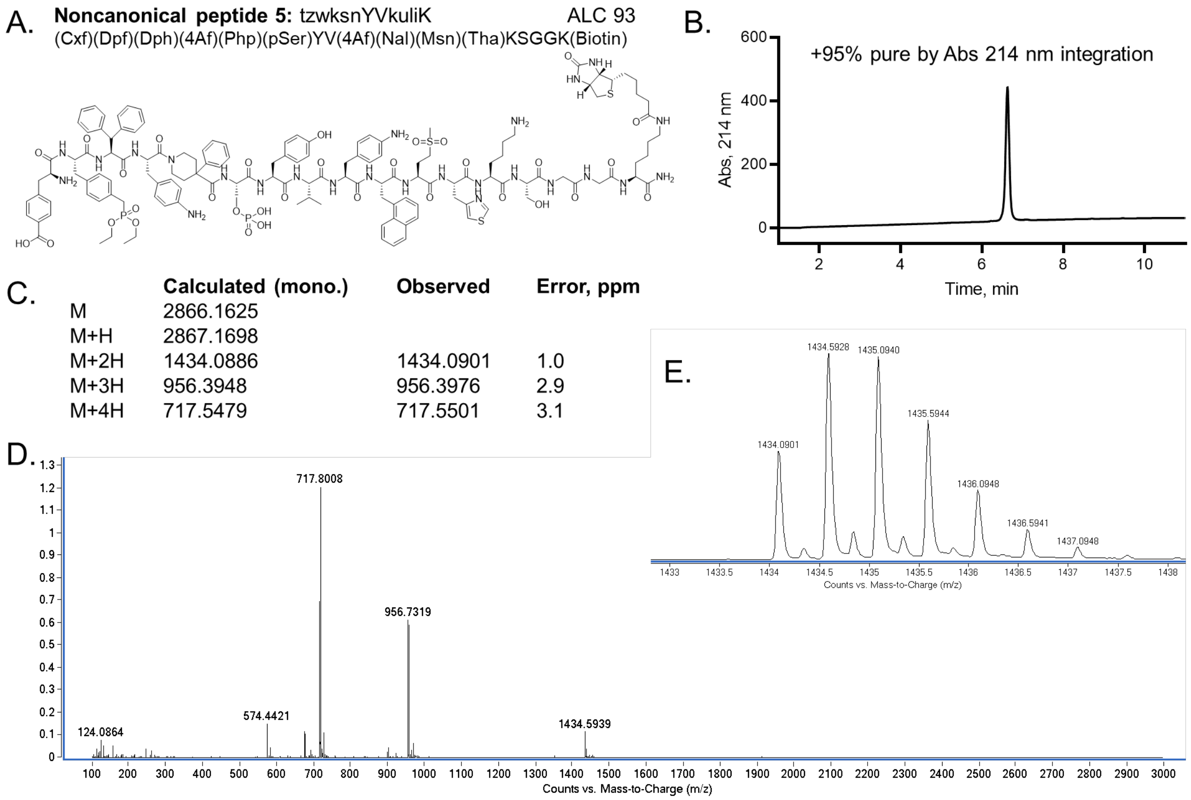
	Calculated (mono.)	Observed	Error, ppm
M	2344.9756		
M+H	2345.9829		
M+2H	1173.4951	1173.4989	3.2
M+3H	782.6658	782.6687	3.7
M+4H	587.2512	587.2538	4.4

Figure 5.38: *Analytical characterization of purified Noncanonical Peptide 3. (A) Sequence information including 1-letter and 3-letter codes for the noncanonical amino acids and average local confidence (ALC) of each peptide. (B) Purity and UPLC chromatogram (C) Calculated and observed monoisotopic masses with ppm error reported. (D) Raw mass spectra of the peptide showing the charge state series (often z = 2,3,4 observed), and E. ~5 m/z zoom in on the lowest charge species observed (often z = 2).*



	Calculated (mono.)	Observed	Error, ppm
M	2635.0863		
M+H	2636.0936		
M+2H	1318.5505	1318.5518	1.0
M+H+Na	1329.5417	1329.5428	0.8
M+3H	879.3694	879.3714	2.3
M+4H	659.7789	659.7813	3.6

Figure 5.39: *Analytical characterization of purified Noncanonical Peptide 4. (A) Sequence information including 1-letter and 3-letter codes for the noncanonical amino acids and average local confidence (ALC) of each peptide. (B) Purity and UPLC chromatogram (C) Calculated and observed monoisotopic masses with ppm error reported. (D) Raw mass spectra of the peptide showing the charge state series (often z = 2,3,4 observed), and E. ~5 m/z zoom in on the lowest charge species observed (often z = 2).*



Noncanonical peptide 5: tzwksnYVkuliK ALC 93

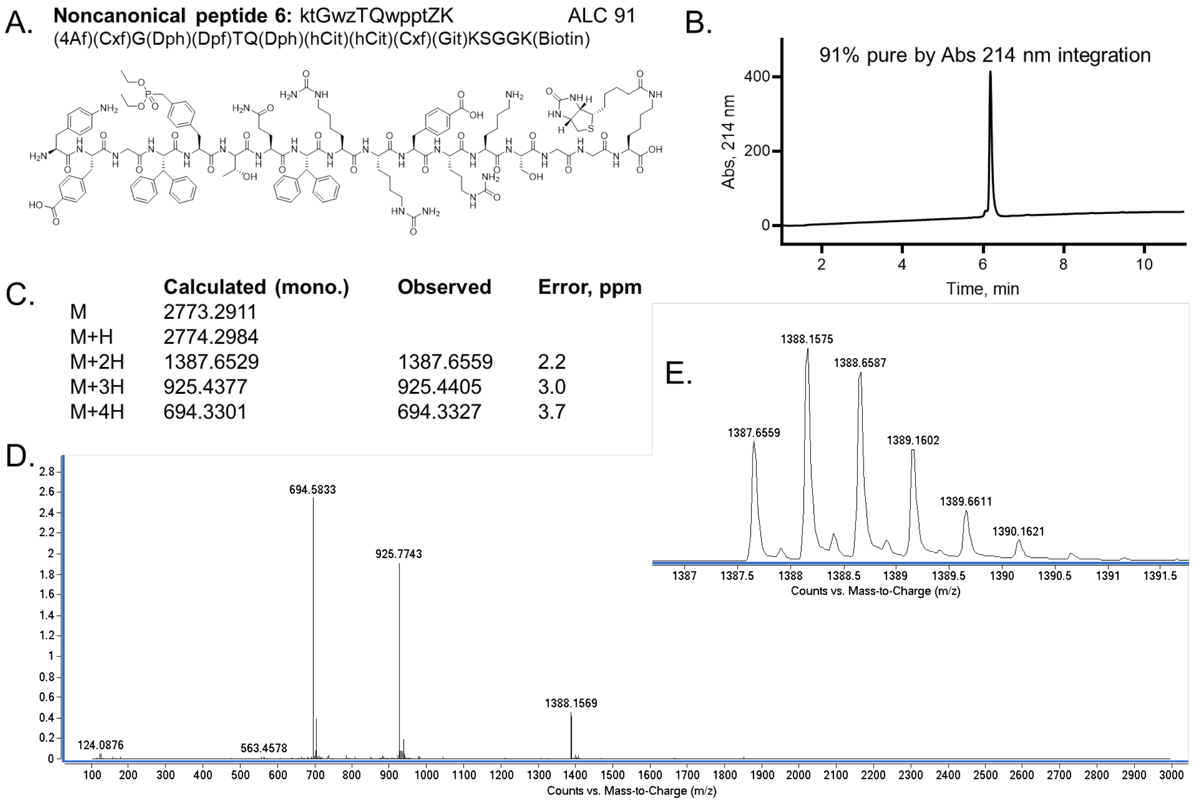
(Cxf)(Dpf)(Dph)(4Af)(Php)(pSer)YV(4Af)(Nal)(Msn)(Tha)KSGGK(Biotin)

+95% pure by Abs 214 nm integration

C.

	Calculated (mono.)	Observed	Error, ppm
M	2866.1625		
M+H	2867.1698		
M+2H	1434.0886	1434.0901	1.0
M+3H	956.3948	956.3976	2.9
M+4H	717.5479	717.5501	3.1

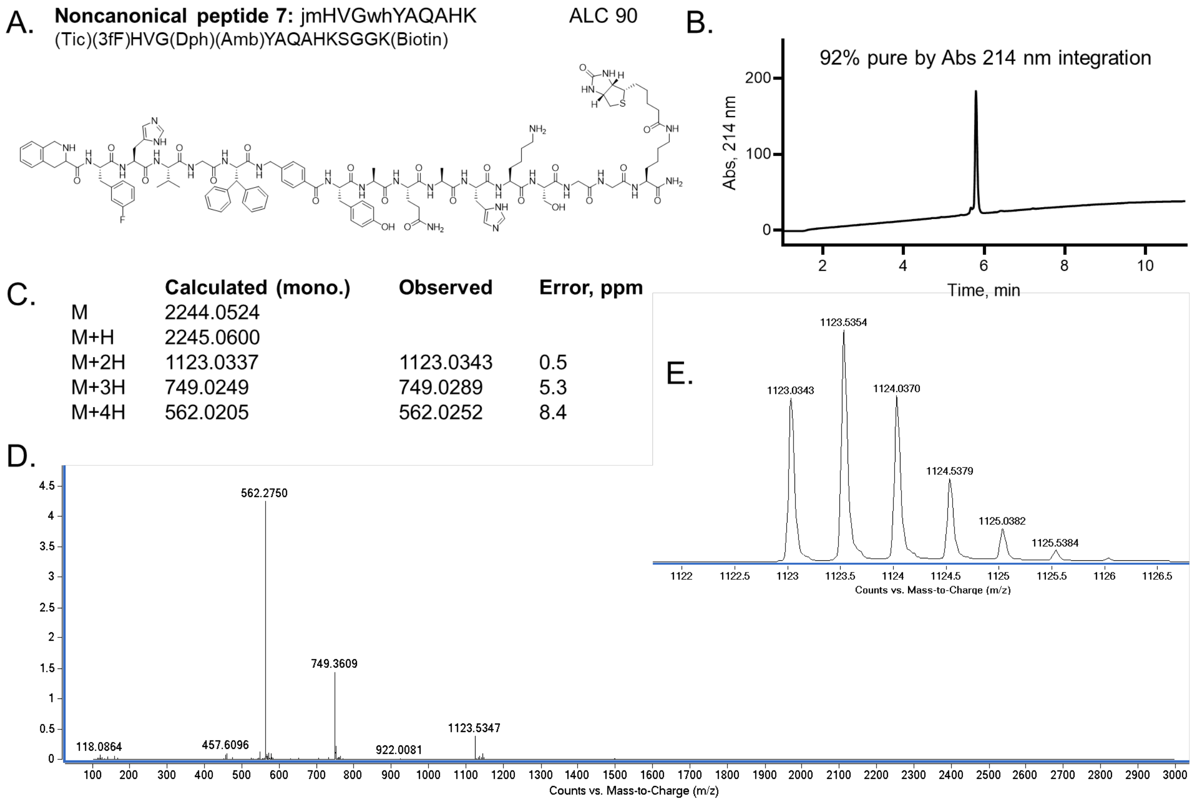
Figure 5.40: *Analytical characterization of purified Noncanonical Peptide 5. (A) Sequence information including 1-letter and 3-letter codes for the noncanonical amino acids and average local confidence (ALC) of each peptide. (B) Purity and UPLC chromatogram (C) Calculated and observed monoisotopic masses with ppm error reported. (D) Raw mass spectra of the peptide showing the charge state series (often z = 2,3,4 observed), and E. ~5 m/z zoom in on the lowest charge species observed (often z = 2).*



Noncanonical peptide 6: ktGwzTQwpptZK ALC 91
(4Af)(Cxf)G(Dph)(Dpf)TQ(Dph)(hCit)(hCit)(Cxf)(Git)KSGGK(Biotin)

	Calculated (mono.)	Observed	Error, ppm
M	2773.2911		
M+H	2774.2984		
M+2H	1387.6529	1387.6559	2.2
M+3H	925.4377	925.4405	3.0
M+4H	694.3301	694.3327	3.7

***Figure 5.41:** Analytical characterization of purified Noncanonical Peptide 6. (A) Sequence information including 1-letter and 3-letter codes for the noncanonical amino acids and average local confidence (ALC) of each peptide. (B) Purity and UPLC chromatogram (C) Calculated and observed monoisotopic masses with ppm error reported. (D) Raw mass spectra of the peptide showing the charge state series (often z = 2,3,4 observed), and E. ~5 m/z zoom in on the lowest charge species observed (often z = 2).*

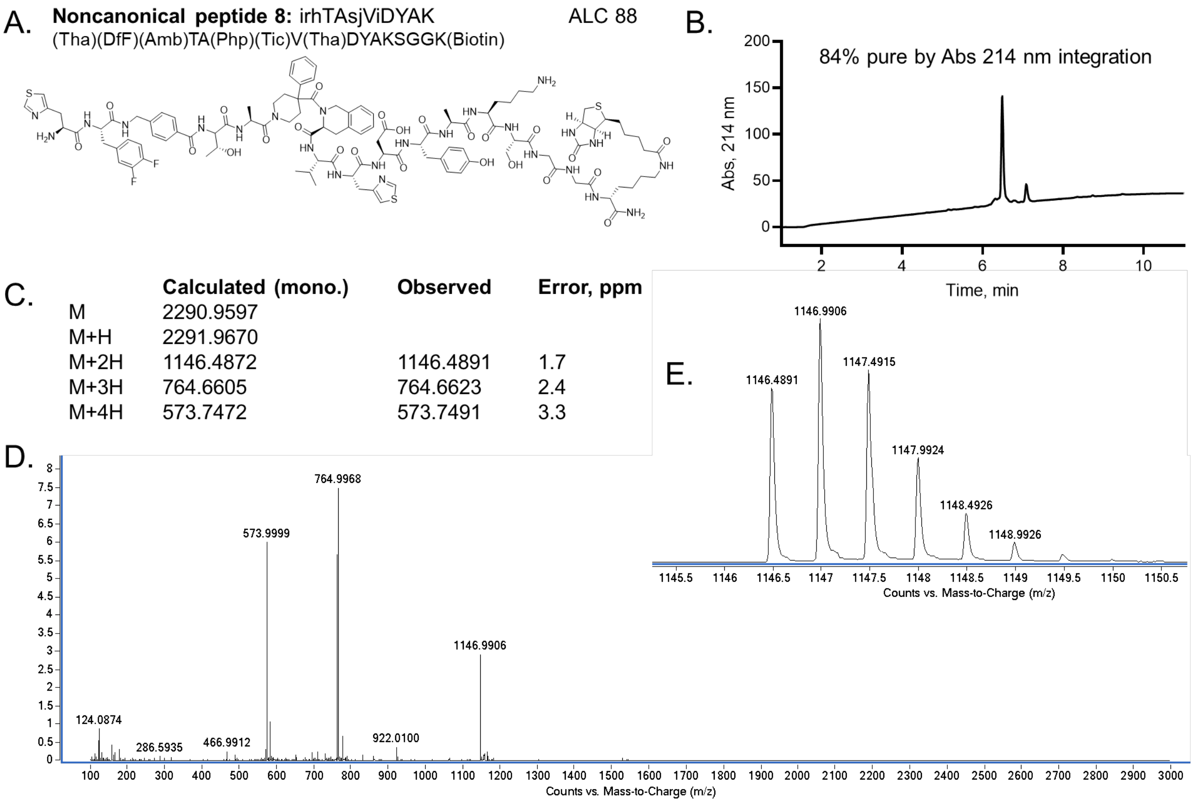


Noncanonical peptide 7: jmHVGwhYAQAHK ALC 90

(Tic)(3fF)HVG(Dph)(Amb)YAQAHKSGGK(Biotin)

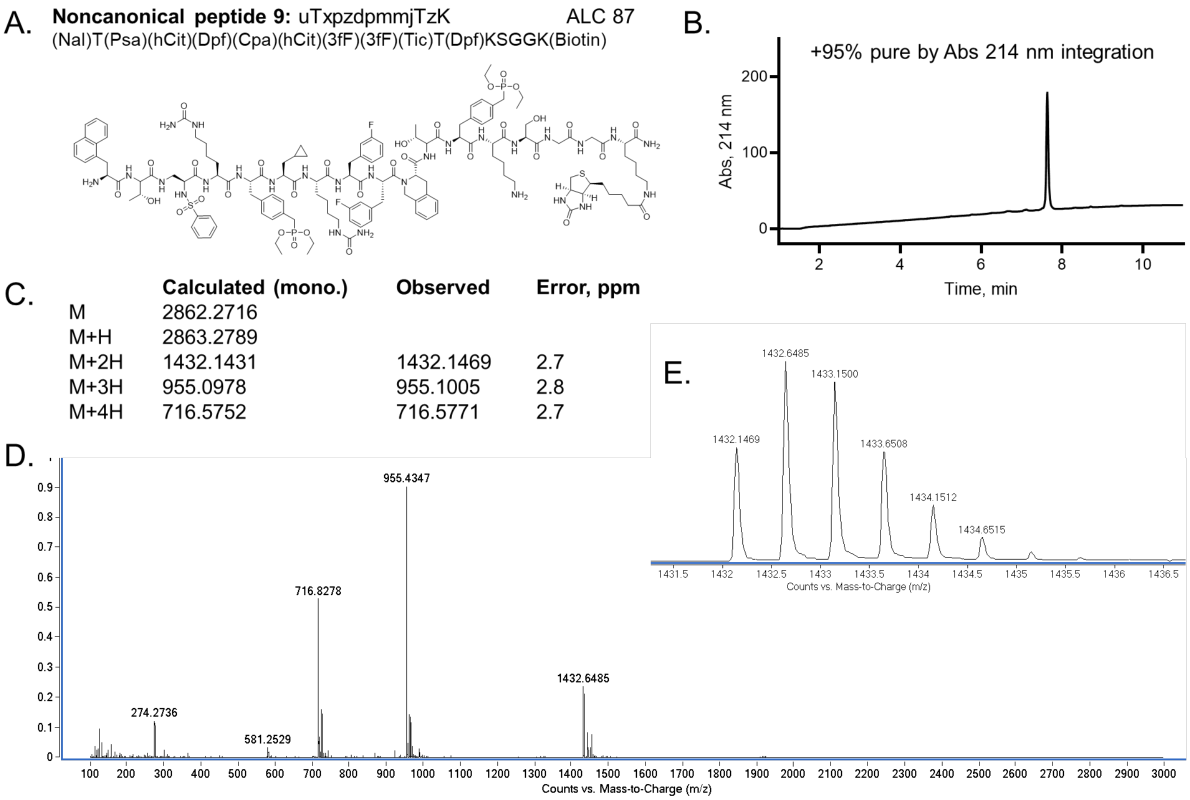
	Calculated (mono.)	**Observed**	**Error, ppm**
M	2244.0524		
M+H	2245.0600		
M+2H	1123.0337	1123.0343	0.5
M+3H	749.0249	749.0289	5.3
M+4H	562.0205	562.0252	8.4

Figure 5.42: *Analytical characterization of purified Noncanonical Peptide 7. (A) Sequence information including 1-letter and 3-letter codes for the noncanonical amino acids and average local confidence (ALC) of each peptide. (B) Purity and UPLC chromatogram (C) Calculated and observed monoisotopic masses with ppm error reported. (D) Raw mass spectra of the peptide showing the charge state series (often z = 2,3,4 observed), and E. ~5 m/z zoom in on the lowest charge species observed (often z = 2).*



	Calculated (mono.)	Observed	Error, ppm
M	2290.9597		
M+H	2291.9670		
M+2H	1146.4872	1146.4891	1.7
M+3H	764.6605	764.6623	2.4
M+4H	573.7472	573.7491	3.3

Figure 5.43: *Analytical characterization of purified Noncanonical Peptide 8. (A) Sequence information including 1-letter and 3-letter codes for the noncanonical amino acids and average local confidence (ALC) of each peptide. (B) Purity and UPLC chromatogram (C) Calculated and observed monoisotopic masses with ppm error reported. (D) Raw mass spectra of the peptide showing the charge state series (often z = 2,3,4 observed), and E. ~5 m/z zoom in on the lowest charge species observed (often z = 2).*



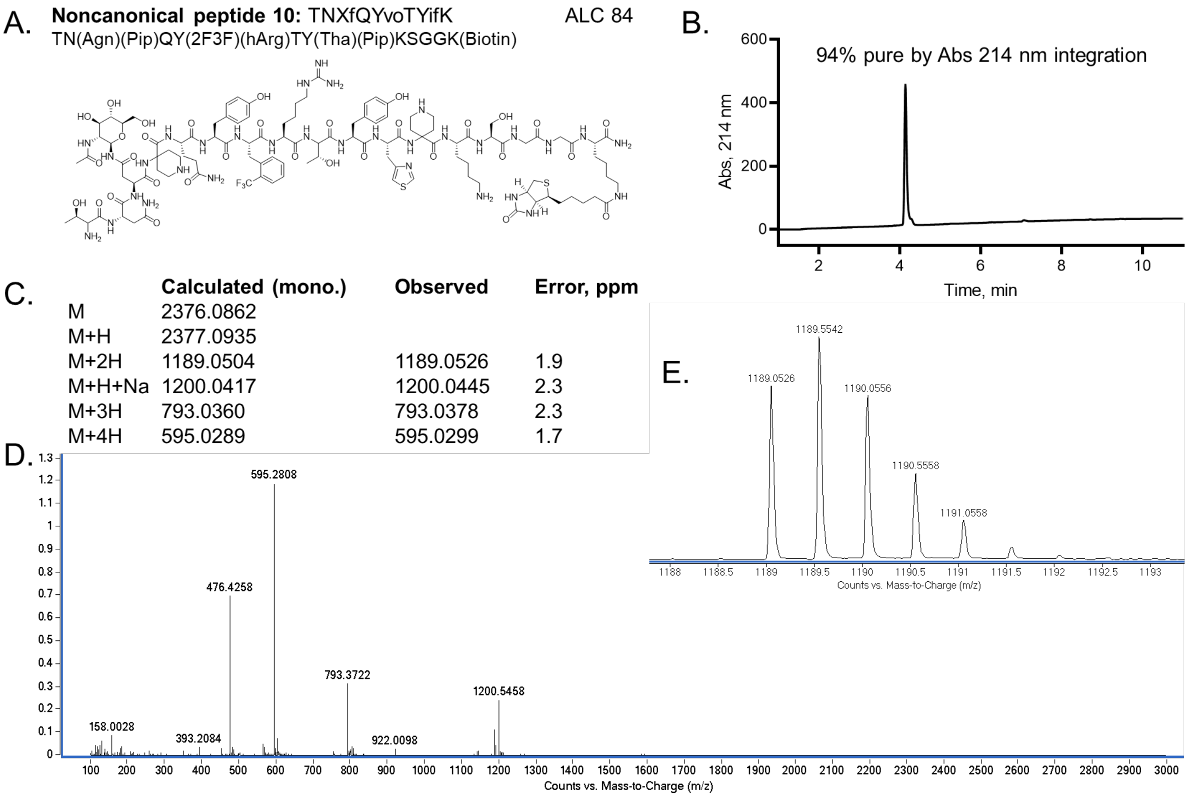
A. **Noncanonical peptide 9:** uTxpzdpmmjTzK ALC 87

(Nal)T(Psa)(hCit)(Dpf)(Cpa)(hCit)(3fF)(3fF)(Tic)T(Dpf)KSGGK(Biotin)

C.

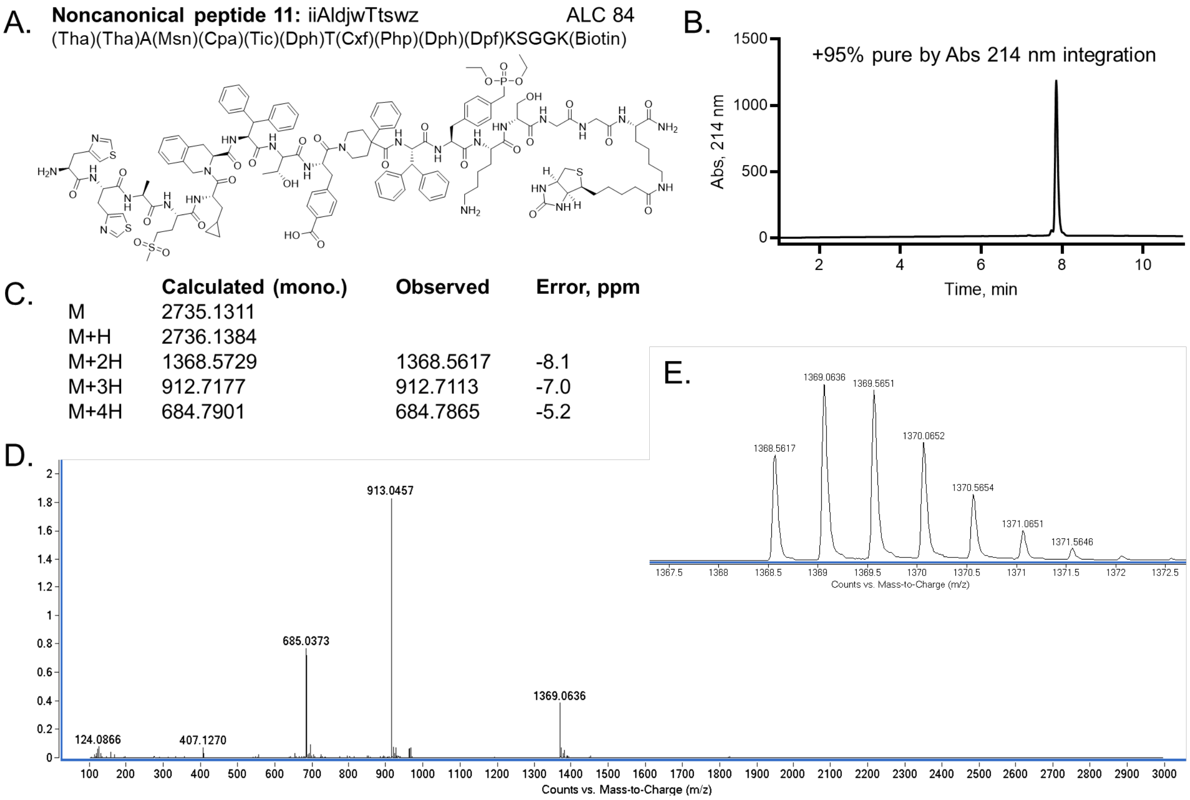
	Calculated (mono.)	Observed	Error, ppm
M	2862.2716		
M+H	2863.2789		
M+2H	1432.1431	1432.1469	2.7
M+3H	955.0978	955.1005	2.8
M+4H	716.5752	716.5771	2.7

Figure 5.44: *Analytical characterization of purified Noncanonical Peptide 9. (A) Sequence information including 1-letter and 3-letter codes for the noncanonical amino acids and average local confidence (ALC) of each peptide. (B) Purity and UPLC chromatogram (C) Calculated and observed monoisotopic masses with ppm error reported. (D) Raw mass spectra of the peptide showing the charge state series (often z = 2,3,4 observed), and E. ~5 m/z zoom in on the lowest charge species observed (often z = 2).*



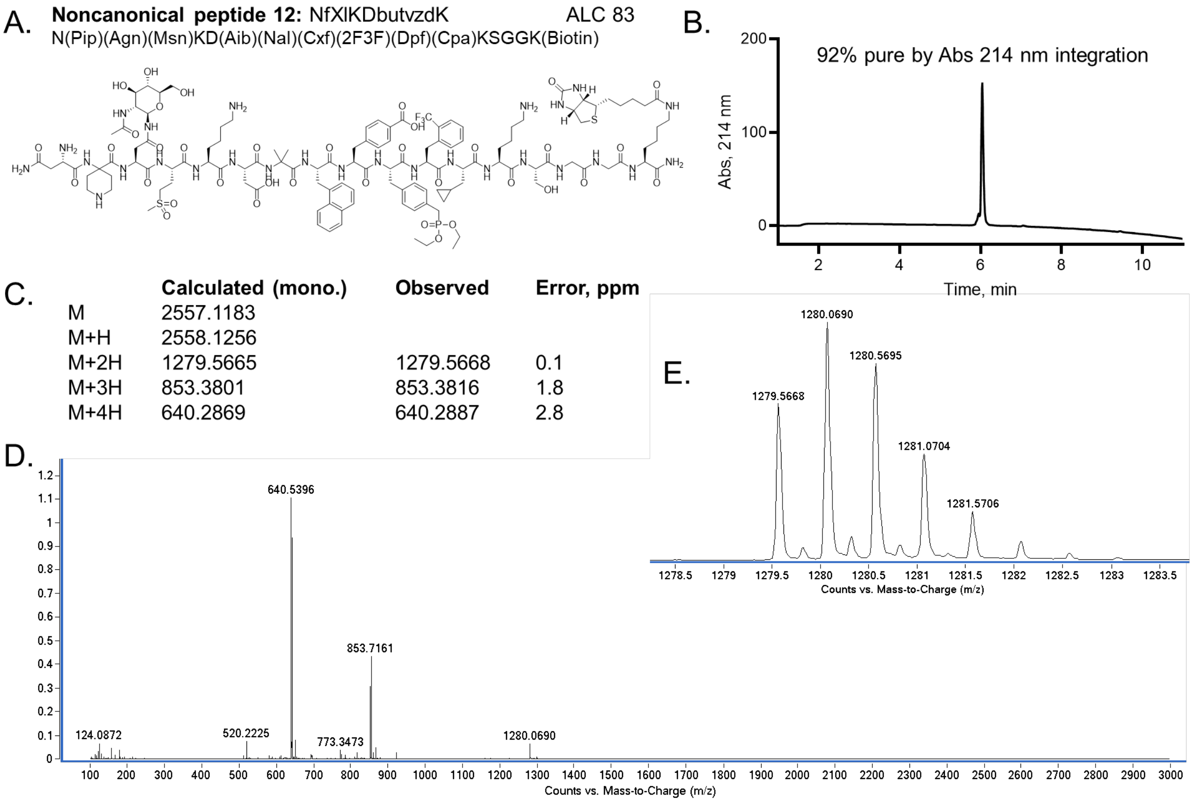
	Calculated (mono.)	Observed	Error, ppm
M	2376.0862		
M+H	2377.0935		
M+2H	1189.0504	1189.0526	1.9
M+H+Na	1200.0417	1200.0445	2.3
M+3H	793.0360	793.0378	2.3
M+4H	595.0289	595.0299	1.7

Figure 5.45: *Analytical characterization of purified Noncanonical Peptide 10. (A) Sequence information including 1-letter and 3-letter codes for the noncanonical amino acids and average local confidence (ALC) of each peptide. (B) Purity and UPLC chromatogram (C) Calculated and observed monoisotopic masses with ppm error reported. (D) Raw mass spectra of the peptide showing the charge state series (often z = 2,3,4 observed), and E. ~5 m/z zoom in on the lowest charge species observed (often z = 2).*



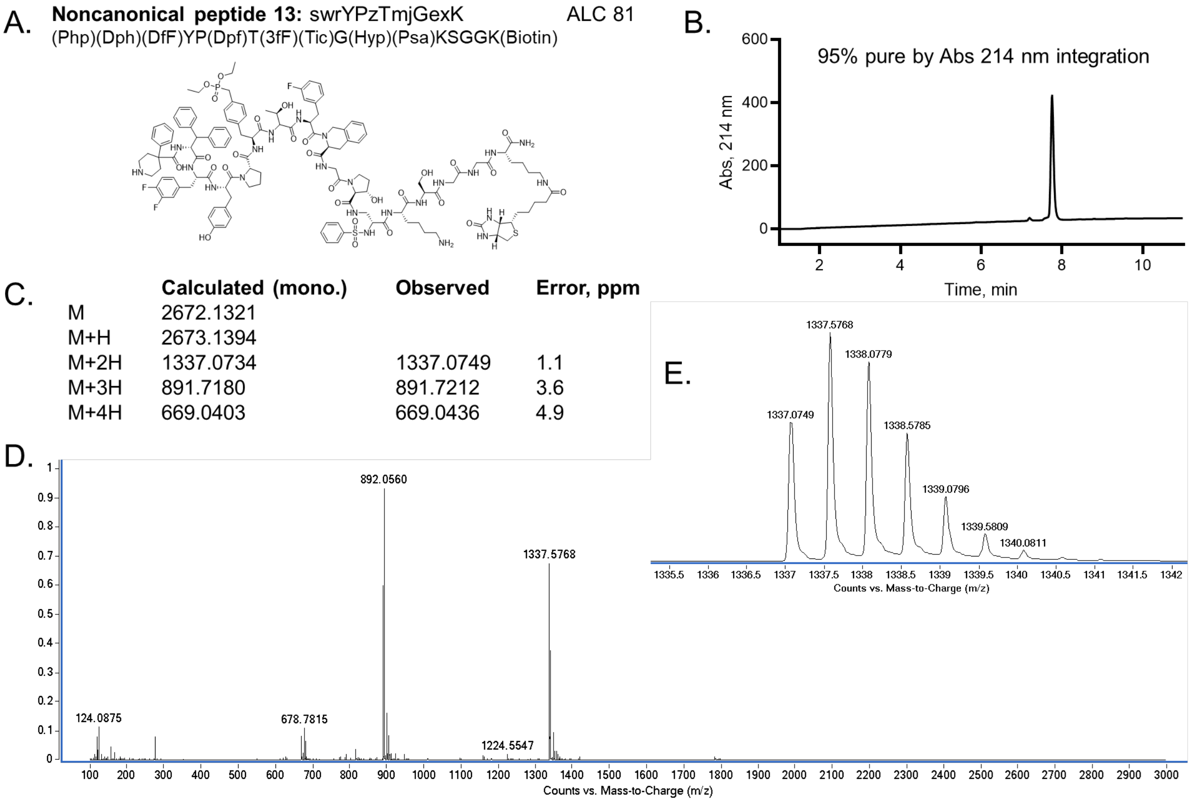
	Calculated (mono.)	Observed	Error, ppm
M	2735.1311		
M+H	2736.1384		
M+2H	1368.5729	1368.5617	-8.1
M+3H	912.7177	912.7113	-7.0
M+4H	684.7901	684.7865	-5.2

Figure 5.46: *Analytical characterization of purified Noncanonical Peptide 11. (A) Sequence information including 1-letter and 3-letter codes for the noncanonical amino acids and average local confidence (ALC) of each peptide. (B) Purity and UPLC chromatogram (C) Calculated and observed monoisotopic masses with ppm error reported. (D) Raw mass spectra of the peptide showing the charge state series (often z = 2,3,4 observed), and E. ~5 m/z zoom in on the lowest charge species observed (often z = 2).*



	Calculated (mono.)	Observed	Error, ppm
M	2557.1183		
M+H	2558.1256		
M+2H	1279.5665	1279.5668	0.1
M+3H	853.3801	853.3816	1.8
M+4H	640.2869	640.2887	2.8

Figure 5.47: *Analytical characterization of purified Noncanonical Peptide 12. (A) Sequence information including 1-letter and 3-letter codes for the noncanonical amino acids and average local confidence (ALC) of each peptide. (B) Purity and UPLC chromatogram (C) Calculated and observed monoisotopic masses with ppm error reported. (D) Raw mass spectra of the peptide showing the charge state series (often z = 2,3,4 observed), and E. ~5 m/z zoom in on the lowest charge species observed (often z = 2).*

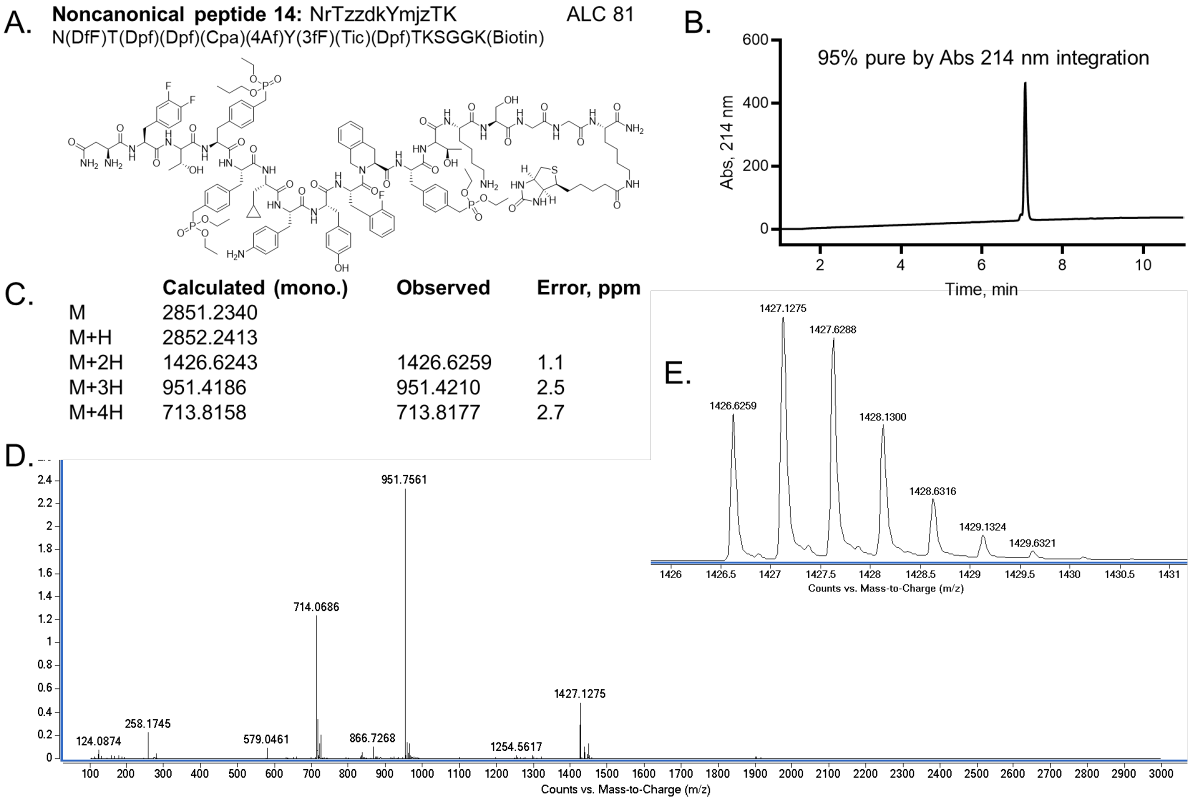


Noncanonical peptide 13: swrYPzTmjGexK ALC 81

(Php)(Dph)(DfF)YP(Dpf)T(3fF)(Tic)G(Hyp)(Psa)KSGGK(Biotin)

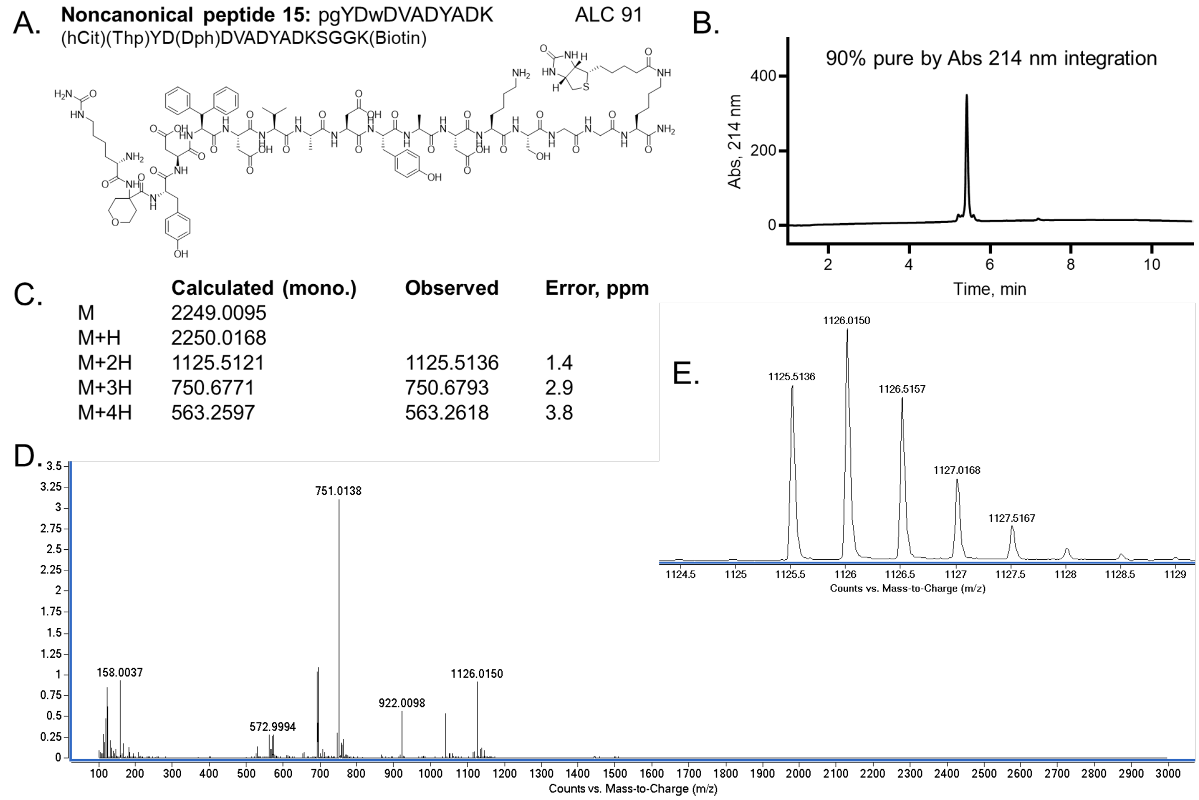
	Calculated (mono.)	Observed	Error, ppm
M	2672.1321		
M+H	2673.1394		
M+2H	1337.0734	1337.0749	1.1
M+3H	891.7180	891.7212	3.6
M+4H	669.0403	669.0436	4.9

Figure 5.48: *Analytical characterization of purified Noncanonical Peptide 13. (A) Sequence information including 1-letter and 3-letter codes for the noncanonical amino acids and average local confidence (ALC) of each peptide. (B) Purity and UPLC chromatogram (C) Calculated and observed monoisotopic masses with ppm error reported. (D) Raw mass spectra of the peptide showing the charge state series (often z = 2,3,4 observed), and E. ~5 m/z zoom in on the lowest charge species observed (often z = 2).*



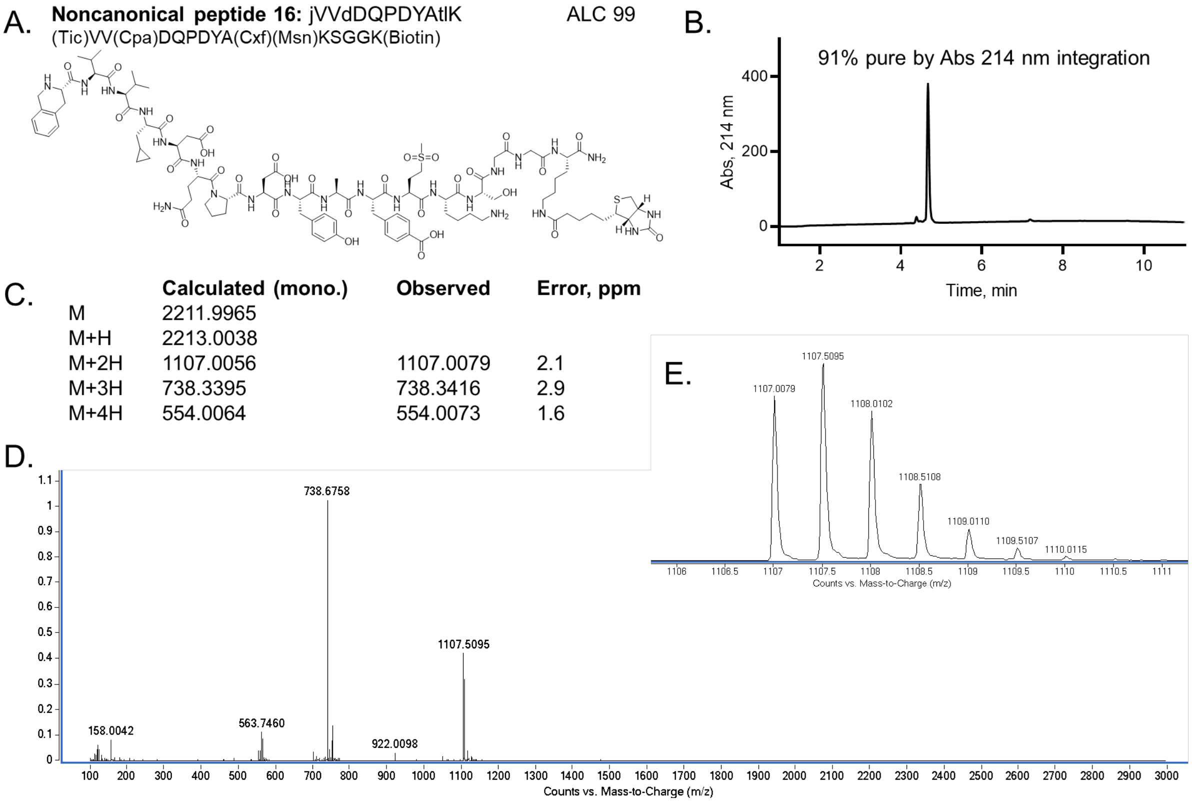
	Calculated (mono.)	Observed	Error, ppm
M	2851.2340		
M+H	2852.2413		
M+2H	1426.6243	1426.6259	1.1
M+3H	951.4186	951.4210	2.5
M+4H	713.8158	713.8177	2.7

Figure 5.49: *Analytical characterization of purified Noncanonical Peptide 14. (A) Sequence information including 1-letter and 3-letter codes for the noncanonical amino acids and average local confidence (ALC) of each peptide. (B) Purity and UPLC chromatogram (C) Calculated and observed monoisotopic masses with ppm error reported. (D) Raw mass spectra of the peptide showing the charge state series (often z = 2,3,4 observed), and E. ~5 m/z zoom in on the lowest charge species observed (often z = 2).*



	Calculated (mono.)	Observed	Error, ppm
M	2249.0095		
M+H	2250.0168		
M+2H	1125.5121	1125.5136	1.4
M+3H	750.6771	750.6793	2.9
M+4H	563.2597	563.2618	3.8

***Figure 5.50:** Analytical characterization of purified Noncanonical Peptide 15. (A) Sequence information including 1-letter and 3-letter codes for the noncanonical amino acids and average local confidence (ALC) of each peptide. (B) Purity and UPLC chromatogram (C) Calculated and observed monoisotopic masses with ppm error reported. (D) Raw mass spectra of the peptide showing the charge state series (often z = 2,3,4 observed), and E. ~5 m/z zoom in on the lowest charge species observed (often z = 2).*



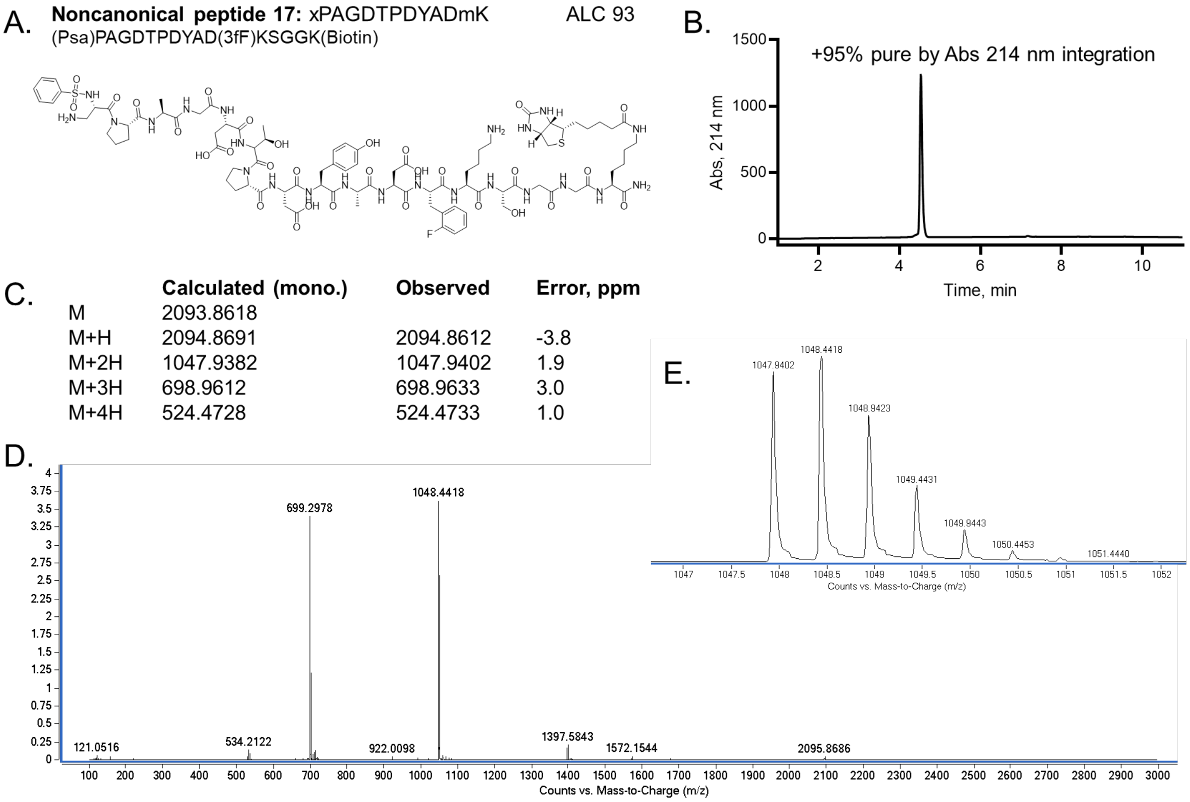
Noncanonical peptide 16: jVVdDQPDYAtlK ALC 99

(Tic)VV(Cpa)DQPDYA(Cxf)(Msn)KSGGK(Biotin)

C.

	Calculated (mono.)	Observed	Error, ppm
M	2211.9965		
M+H	2213.0038		
M+2H	1107.0056	1107.0079	2.1
M+3H	738.3395	738.3416	2.9
M+4H	554.0064	554.0073	1.6

Figure 5.51: *Analytical characterization of purified Noncanonical Peptide 16. (A) Sequence information including 1-letter and 3-letter codes for the noncanonical amino acids and average local confidence (ALC) of each peptide. (B) Purity and UPLC chromatogram (C) Calculated and observed monoisotopic masses with ppm error reported. (D) Raw mass spectra of the peptide showing the charge state series (often z = 2,3,4 observed), and E. ~5 m/z zoom in on the lowest charge species observed (often z = 2).*



Noncanonical peptide 17: xPAGDTPDYADmK ALC 93
(Psa)PAGDTPDYAD(3fF)KSGGK(Biotin)

+95% pure by Abs 214 nm integration

C.

	Calculated (mono.)	Observed	Error, ppm
M	2093.8618		
M+H	2094.8691	2094.8612	-3.8
M+2H	1047.9382	1047.9402	1.9
M+3H	698.9612	698.9633	3.0
M+4H	524.4728	524.4733	1.0

Figure 5.52: *Analytical characterization of purified Noncanonical Peptide 17. (A) Sequence information including 1-letter and 3-letter codes for the noncanonical amino acids and average local confidence (ALC) of each peptide. (B) Purity and UPLC chromatogram (C) Calculated and observed monoisotopic masses with ppm error reported. (D) Raw mass spectra of the peptide showing the charge state series (often z = 2,3,4 observed), and E. ~5 m/z zoom in on the lowest charge species observed (often z = 2).*

5.8.3. Biolayer interferometry (BLI) data of all AS-MS discovered noncanonical peptides

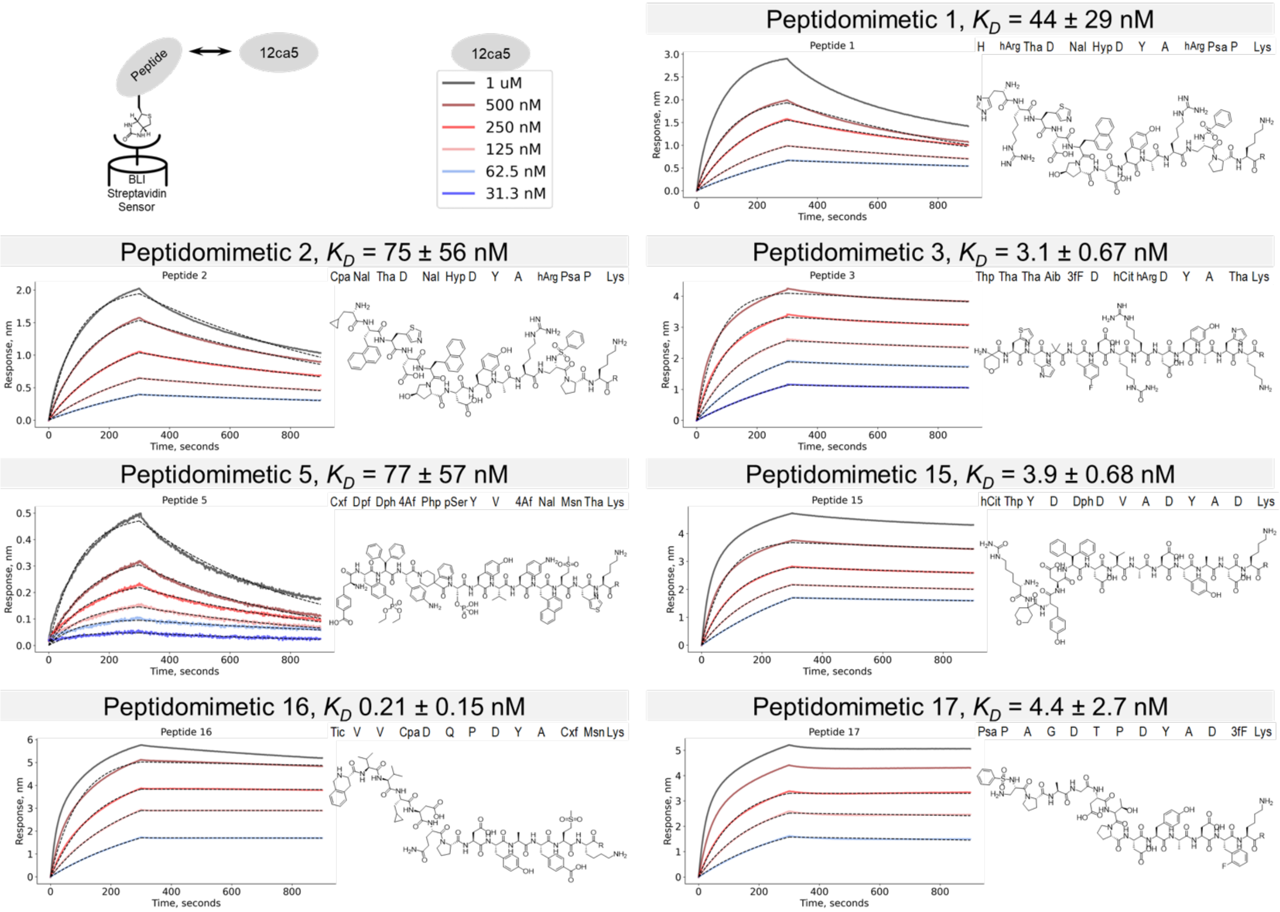


Figure 5.53: *BLI sensorgrams of all binding peptides and peptidomimietics with their monomers and structures shown. Peptides were labeled with a SGGLys(Biotin)-NH_2 (labeled as R) at the C-terminus. In the top left, the BLI assay format is shown, with biotinylated peptides immobilized and 12ca5 in solution at the concentrations shown. Note that Peptide 4, 6, 7, 8, 9, 10, 11, 12, 13, and 14 are nonbinders seen in Figure 5.31. The association and dissociation curves were fitted using a 1:1 binding model (n > 2 fit curves accepted shown as black dashed lines with Full $R^2 > 0.8$ and $X^2 < 32$, see Table 5.19) to calculate the apparent dissociation constant (K_D).*

Table 5.19: *BLI Data Summary of all binding peptides and peptidomimetics in this work. Note that Peptide 4, 6, 7, 8, 9, 10, 11, 12, 13, and 14 are nonbinders.*

Peptide 1	12ca5 Conc (nM)	koff(1/s)	kon(1/Ms)	KD(M)	Rmax	Req	Response	LoadingHeight	FullR2	Full X2
	2000	1.52E-03	1.05E+04	1.45E-07	3.13	2.92	3.06	0.624	0.978	38.30
	1000	1.27E-03	1.38E+04	9.21E-08	3.07	2.81	2.90	0.689	0.986	19.00
Dissociation Constant, K_D	500	1.08E-03	1.80E+04	6.03E-08	2.28	2.03	1.99	0.584	0.993	4.22
44 ± 29 nM	250	7.84E-04	2.16E+04	3.62E-08	2.10	1.84	1.58	0.674	0.998	0.98
Ave ± SD nM	125	5.71E-04	2.63E+04	2.17E-08	1.68	1.43	0.99	0.594	0.999	0.12
	62.5	3.44E-04	2.87E+04	1.20E-08	1.67	1.40	0.66	0.649	0.999	0.06

Peptide 2	12ca5 Conc (nM)	koff(1/s)	kon(1/Ms)	KD(M)	Rmax	Req	Response	LoadingHeight	FullR2	Full X2
	2000	1.35E-03	7.60E+03	1.77E-07	2.54	2.33	2.43	0.790	0.983	16.20
	1000	1.16E-03	1.01E+04	1.15E-07	2.24	2.01	2.02	0.780	0.991	5.86
Dissociation Constant, K_D	500	9.71E-04	1.31E+04	7.41E-08	1.97	1.71	1.57	0.794	0.996	1.57
75 ± 56 nM	250	7.37E-04	1.63E+04	4.52E-08	1.60	1.36	1.05	0.842	0.999	0.29
Ave ± SD nM	125	5.81E-04	2.13E+04	2.73E-08	1.26	1.03	0.65	0.770	0.999	0.08
	62.5	4.36E-04	3.23E+04	1.35E-08	0.92	0.75	0.40	0.743	0.998	0.06

Peptide 3	12ca5 Conc (nM)	koff(1/s)	kon(1/Ms)	KD(M)	Rmax	Req	Response	LoadingHeight	FullR2	Full X2
	1000	9.01E-05	1.90E+04	4.73E-09	4.54	4.52	4.67	0.507	0.963	73.80
	500	1.16E-04	2.81E+04	4.14E-09	4.19	4.16	4.21	0.513	0.986	30.10
Dissociation Constant, K_D	250	1.37E-04	3.91E+04	3.51E-09	3.55	3.50	3.38	0.496	0.996	6.57
3.1 ± 0.67 nM	125	1.54E-04	4.97E+04	3.10E-09	3.08	3.01	2.58	0.499	0.999	1.17
Ave ± SD nM	62.5	1.54E-04	6.15E+04	2.51E-09	2.80	2.69	1.89	0.496	1.000	0.43
	31.3	1.37E-04	6.01E+04	2.28E-09	2.69	2.51	1.13	0.505	1.000	0.15

Peptide 5	M Conc.(nM)	koff(1/s)	kon(1/Ms)	KD(M)	Rmax	Req	Response	LoadingHeight	FullR2	Full X2
	1000	1.85E-03	8.55E+03	2.16E-07	0.60	0.49	0.48	1.389	0.986	0.66
	500	1.82E-03	1.17E+04	1.56E-07	0.44	0.34	0.31	1.314	0.991	0.18
Dissociation Constant, K_D	250	1.50E-03	2.53E+04	5.92E-08	0.30	0.24	0.23	1.334	0.979	0.18
77 ± 57 nM	125	1.33E-03	5.63E+04	2.37E-08	0.19	0.16	0.15	1.345	0.955	0.16
Ave ± SD nM	62.5	8.12E-04	1.72E+05	4.72E-09	0.11	0.10	0.10	1.370	0.854	0.15
	31.3	1.21E-03	4.34E+05	2.79E-09	0.05	0.05	0.05	1.332	0.768	0.08

Peptide 15	12ca5 Conc (nM)	koff(1/s)	kon(1/Ms)	KD(M)	Rmax	Req	Response	LoadingHeight	FullR2	Full X2
	1000	9.35E-05	1.86E+04	5.01E-09	4.61	4.59	4.73	0.293	0.977	48.70
	500	1.07E-04	2.59E+04	4.15E-09	3.78	3.75	3.76	0.266	0.994	11.40
Dissociation Constant, K_D	250	1.28E-04	3.42E+04	3.73E-09	3.05	3.01	2.82	0.264	0.999	0.78
3.9 ± 0.68 nM	125	1.31E-04	4.20E+04	3.13E-09	2.76	2.70	2.17	0.256	1.000	0.17
Ave ± SD nM	62.5	1.08E-04	3.32E+04	3.26E-09	3.71	3.53	1.69	0.289	1.000	0.10

Peptide 16	12ca5 Conc (nM)	koff(1/s)	kon(1/Ms)	KD(M)	Rmax	Req	Response	LoadingHeight	FullR2	Full X2
	1000	1.08E-04	1.61E+04	6.71E-09	5.67	5.63	5.76	0.297	0.972	97.70
	500	5.15E-05	2.12E+04	2.43E-09	5.26	5.24	5.11	0.299	0.993	30.40
Dissociation Constant, K_D	250	8.47E-06	2.71E+04	3.12E-10	4.41	4.41	3.87	0.262	0.999	4.77
≤ 1 nM*	125	1.22E-07	3.33E+04	3.67E-12	4.06	4.06	2.93	0.303	1.000	0.72
*Measured 0.21 ± 0.15 nM (Ave ± SD), out of range for instrument	62.5	1.32E-05	4.15E+04	3.18E-10	3.15	3.14	1.73	0.286	1.000	0.27

Peptide 17	M Conc.(nM)	koff(1/s)	kon(1/Ms)	KD(M)	Rmax	Req	Response	LoadingHeight	FullR2	Assoc.X2
	1000	NA	1.86E+04	NA	4.99	4.99	5.20	0.254	0.966	67.90
	500	NA	2.54E+04	NA	4.37	4.37	4.41	0.257	0.991	19.50
Dissociation Constant, K_D	250	2.83E-05	3.31E+04	8.54E-10	3.63	3.61	3.38	0.221	0.999	1.78
4.4 ± 2.7 nM	125	2.02E-04	3.99E+04	5.05E-09	3.36	3.23	2.59	0.269	0.999	0.19
Ave ± SD nM	62.5	3.10E-04	4.17E+04	7.44E-09	3.07	2.74	1.62	0.267	0.999	0.11

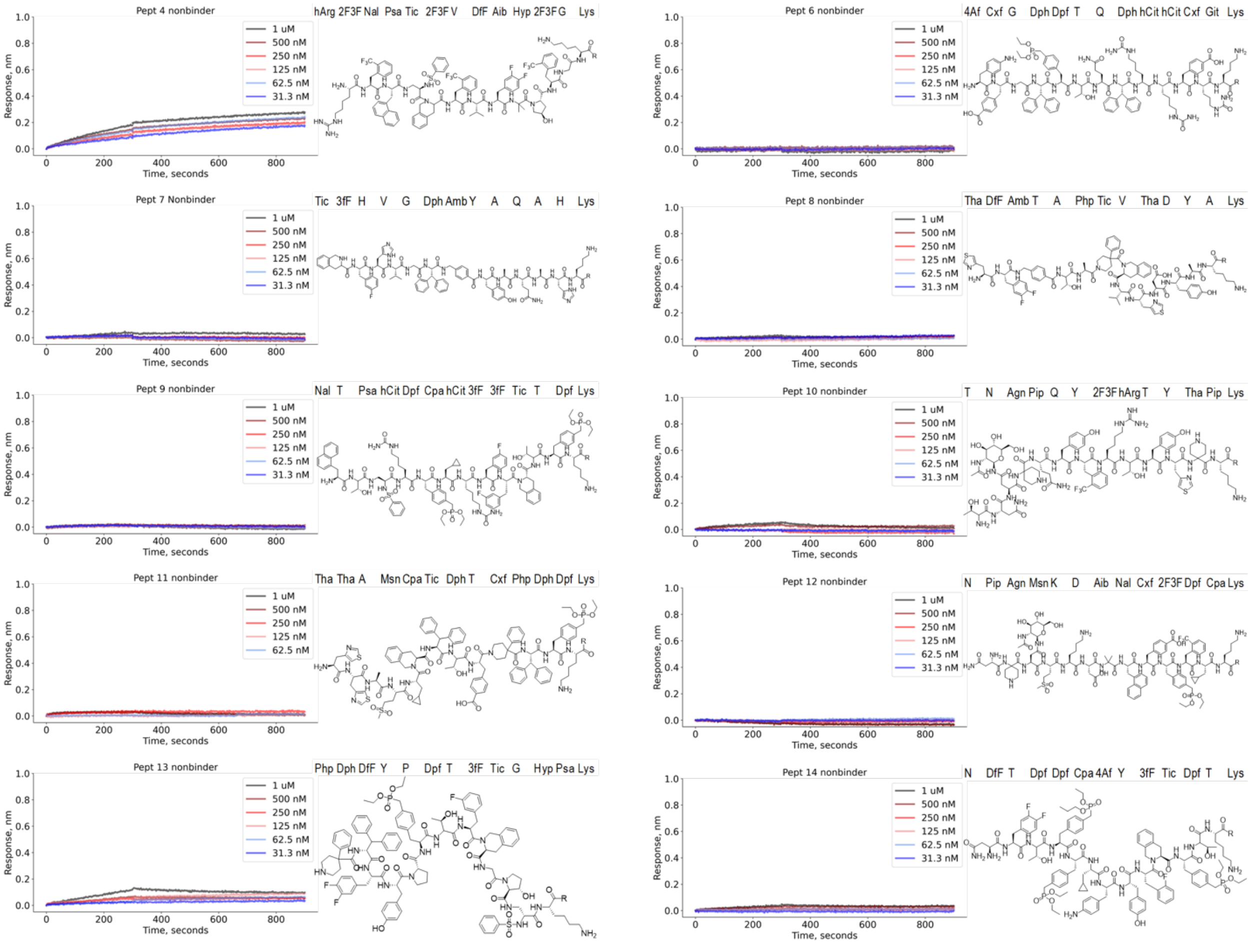


Figure 5.54: *BLI sensorgrams of all nonbinding peptides and peptidomimietics with their monomers and structures shown. Peptides were labeled with a SGGLys(Biotin)-NH$_2$ (labeled as R) at the C-terminus.*

5.9. References

1. Henninot, A., Collins, J. C. & Nuss, J. M. The Current State of Peptide Drug Discovery: Back to the Future? J Med Chem 61, 1382–1414 (2018).
2. Muttenthaler, M., King, G. F., Adams, D. J. & Alewood, P. F. Trends in peptide drug discovery. Nat Rev Drug Discov 20, 309–325 (2021).
3. Lau, J. L. & Dunn, M. K. Therapeutic peptides: Historical perspectives, current development trends, and future directions. Bioorg Med Chem 26, 2700–2707 (2018).
4. Wells, J. A. & McClendon, C. L. Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. Nature 450, 1001–1009 (2007).
5. Cunningham, A. D., Qvit, N. & Mochly-Rosen, D. Peptides and peptidomimetics as regulators of protein–protein interactions. Curr Opin Struct Biol 44, 59–66 (2017).
6. Lubell, W. D. Peptide-Based Drug Development. Biomedicines 2022, Vol. 10, Page 2037 10, 2037 (2022).
7. Clackson, T. & Wells, J. A. In vitro selection from protein and peptide libraries. Trends Biotechnol 12, 173–184 (1994).
8. Smith, G. P. & Petrenko, V. A. Phage display. Chem Rev 97, 391–410 (1997).
9. Vinogradov, A. A., Yin, Y. & Suga, H. Macrocyclic Peptides as Drug Candidates: Recent Progress and Remaining Challenges. J Am Chem Soc 141, 4167–4181 (2019).
10. 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).
11. Zhang, G. et al. Rapid de novo discovery of peptidomimetic affinity reagents for human angiotensin converting enzyme 2. Commun Chem 5, 1–10 (2022).
12. Quartararo, A. J. et al. Ultra-large chemical libraries for the discovery of high-affinity peptide binders. Nat Commun 11, 3183 (2020).
13. Bowling, J. J., Shadrick, W. R., Griffith, E. C. & Lee, R. E. Going Small: Using Biophysical Screening to Implement Fragment Based Drug Discovery. in Special Topics in Drug Discovery vol. i 13 (InTech, 2016).
14. Bergsdorf, C. & Ottl, J. Affinity-based screening techniques: their impact and benefit to increase the number of high quality leads. Expert Opin Drug Discov 5, 1095–1107 (2010).
15. Katsuno, K. et al. Hit and lead criteria in drug discovery for infectious diseases of the developing world. Nature Reviews Drug Discovery 2015 14:11 14, 751–758 (2015).
16. Touti, F., Gates, Z. P., Bandyopdhyay, A. & Lautrette, G. In-solution enrichment identifies peptide inhibitors of protein – protein interactions protein-protein interactions. Nat Chem Biol 318–339 (2019) doi:10.1038/s41589-019-0245-2.
17. Sheridan, R. P. et al. Experimental Error, Kurtosis, Activity Cliffs, and Methodology: What Limits the Predictivity of Quantitative Structure-Activity Relationship Models? J Chem Inf Model 60, 1969–1982 (2020).

18. Kusumoto, Y. et al. Highly Potent and Oral Macrocyclic Peptides as a HIV-1 Protease Inhibitor: mRNA Display-Derived Hit-to-Lead Optimization. ACS Med Chem Lett (2022) doi:10.1021/ACSMEDCHEMLETT.2C00310.
19. Iskandar, S. E. & Bowers, A. A. mRNA Display Reaches for the Clinic with New PCSK9 Inhibitor. ACS Med Chem Lett (2022) doi:10.1021/ACSMEDCHEMLETT.2C00319.
20. Rogers, J. M., Passioura, T. & Suga, H. Nonproteinogenic deep mutational scanning of linear and cyclic peptides. Proc Natl Acad Sci U S A 115, 10959–10964 (2018).
21. Naylor, M. R., Bockus, A. T., Blanco, M. J. & Lokey, R. S. Cyclic peptide natural products chart the frontier of oral bioavailability in the pursuit of undruggable targets. Curr Opin Chem Biol 38, 141–147 (2017).
22. Schneider, G. Automating drug discovery. Nat Rev Drug Discov 17, 97–113 (2018).
23. Vamathevan, J. et al. Applications of machine learning in drug discovery and development. Nature Reviews Drug Discovery 2019 18:6 18, 463–477 (2019).
24. Schroedl, S. Current methods and challenges for deep learning in drug discovery. Drug Discov Today Technol 32–33, 9–17 (2019).
25. Kobak, D. & Berens, P. The art of using t-SNE for single-cell transcriptomics. Nature Communications 2019 10:1 10, 1–14 (2019).
26. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci U S A 118, e2016239118 (2021).
27. Das, P., Moll, M., Stamati, H., Kavraki, L. E. & Clementi, C. Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. Proc Natl Acad Sci U S A 103, 9885–9890 (2006).
28. Bileschi, M. L. et al. Using deep learning to annotate the protein universe. Nat Biotechnol 40, 932–937 (2022).
29. Bannigan, P. et al. Machine learning models to accelerate the design of polymeric long-acting injectables. Nature Communications 2023 14:1 14, 1–12 (2023).
30. Mohapatra, S., An, J. & Gómez-Bombarelli, R. Chemistry-informed macromolecule graph representation for similarity computation, unsupervised and supervised learning. Mach Learn Sci Technol 3, 015028 (2022).
31. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv (2018) doi:10.48550/arxiv.1802.03426.
32. Schissel, C. K. et al. Deep learning to design nuclear-targeting abiotic miniproteins. Nat Chem 13, 992–1000 (2021).
33. Torres, M. D. T., Melo, M. C. R., Crescenzi, O., Notomista, E. & de la Fuente-Nunez, C. Mining for encrypted peptide antibiotics in the human proteome. Nature Biomedical Engineering 2021 6:1 6, 67–75 (2021).

34. Li, G., Iyer, B., Prasath, V. B. S., Ni, Y. & Salomonis, N. DeepImmuno: deep learning-empowered prediction and generation of immunogenic peptides for T-cell immunity. Brief Bioinform 22, 1–10 (2021).
35. Lei, Y. et al. A deep-learning framework for multi-level peptide–protein interaction prediction. Nature Communications 2021 12:1 12, 1–10 (2021).
36. Saka, K. et al. Antibody design using LSTM based deep generative model from phage display library for affinity maturation. Scientific Reports 2021 11:1 11, 1–13 (2021).
37. Philpott, D. N. et al. Rapid On-Cell Selection of High-Performance Human Antibodies. ACS Cent Sci 8, 102–109 (2022).
38. McCloskey, K. et al. Machine learning on DNA-encoded libraries: A new paradigm for hit finding. J Med Chem 63, 8857–8866 (2020).
39. 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 5, 62–71 (2021).
40. Ye, X. et al. Binary combinatorial scanning reveals potent poly-alanine-substituted inhibitors of protein-protein interactions. Communications Chemistry 2022 5:1 5, 1–10 (2022).
41. Chuang, K. v., Gunsalus, L. M. & Keiser, M. J. Learning Molecular Representations for Medicinal Chemistry. J Med Chem 63, 8705–8722 (2020).
42. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science (1979) 379, 1123–1130 (2023).
43. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. J Chem Inf Model 50, 742–754 (2010).
44. Fangping Wan, Daphne Kontogiorgos-Heintz & Fuente-Nunez, C. de la. Deep generative models for peptide design. Digital Discovery 1, 195–208 (2022).
45. Landrum, G. RDKit: Open-Source Cheminformatics Software. https://www.rdkit.org/ (2010).
46. Tipping, M. E. & Bishop, C. M. Mixtures of Probabilistic Principal Component Analyzers. Neural Comput 11, 443–482 (1999).
47. Rini, J. M., Schulze-Gahmen, U. & Wilson, I. A. Structural evidence for induced fit as a mechanism for antibody-antigen recognition. Science (1979) 255, 959–965 (1992).
48. Houghten, R. A. et al. Generation and use of synthetic peptide combinatorial libraries for basic research and drug discovery. Nature 1991 354:6348 354, 84–86 (1991).
49. Larkin, M. A. et al. Clustal W and Clustal X version 2.0. Bioinformatics 23, 2947–2948 (2007).
50. Tareen, A. & Kinney, J. B. Logomaker: beautiful sequence logos in Python. Bioinformatics 36, 2272–2274 (2020).
51. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011).
52. Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME Suite. Nucleic Acids Res 43, W39–W49 (2015).

53. Bailey, T. L. Discovering Sequence Motifs. in Comparative Genomics. Methods in Molecular Biology (ed. Bergman, N. H.) vol. 395 231–251 (Humana Press, 2007).
54. Chen, K. H. & Hu, Y. J. Residue–Residue Interaction Prediction via Stacked Meta-Learning. International Journal of Molecular Sciences 2021, Vol. 22, Page 6393 22, 6393 (2021).
55. Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? J Cheminform 7, 1–13 (2015).
56. Cox, M. A. A. & Cox, T. F. Multidimensional Scaling. in Handbook of Data Visualization vol. 4 315–347 (Springer Berlin Heidelberg, 2008).
57. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. Nature Protocols 2007 2:8 2, 1896-1906 (2007).
58. Vinogradov, A. A. et al. Library Design-Facilitated High-Throughput Sequencing of Synthetic Peptide Lirbaries. ACS Comb Sci 19, 694-701 (2017).
59. Pinilla, C., Appel, J. R. & Houghten, R. A. Investigation of antigen-antibody interactions using a soluble, non-support-bound synthetic decapeptide library composed of four trillion (4 x 1012) sequences. Biochemical Journal 301, 847-853 (1994).
60. Tareen, A. & Kinney, J. B. Logomaker: beautiful sequence logos in Python. Bioinformatics 36, 2272-2274 (2020).
61. Bailey, T.L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME Suite. Nucleic Acids Res 43, W39-W49 (2015).
62. Bailey, T. L. STREME: accurate and versatile sequence motif discovery. Bioinformatics 37, 2834-2840 (2021).

6. Appendix: Ligand discovery from phage display significantly improved by regularized proxy learning

The work presented in this chapter has been reproduced and adapted from the following publication:

Brown, J. S.*; Tseo, Y.*; **Lee, M. A.**; Wong, J.Y.-K.; Yang, S.; Kim, C.R.; Loas, A.; Gomez-Bombarelli, R.; Pentelute, B.L. Ligand discovery from phage display significantly improved by regularized proxy learning. *ChemRxiv Preprint*. *Manuscript submission for peer review in progress.* **2023**

A.1. Introduction

Phage display is a robust method to perform genetically encoded affinity selection against biomolecular targets.[1–7] As a growing therapeutic modality, peptides and peptidomimetics ligands can consolidate weak interactions across protein-protein interfaces, have recently experienced higher clinical trial success than small molecules, and remain less expensive to produce than biologics.[8–12] Phage display is easily accessible, inexpensive, and can serve as a first-line approach to perform peptide ligand discovery with several libraries available commercially. Furthermore, engineered phage libraries have enabled macrocyclization and covalent pharmacophore modification to improve discovery outcomes and pharmacokinetic properties. [4–7,13–15] Due to these advantages, phage display has been responsible for the discovery of clinically investigated peptidomimetics.[15–18] Phage display has found broad utility in a variety of contexts including in vivo,[2] ex vivo, and in vitro (on-cell) bio-panning.[3] Arguably, the robust use of phage display has been enabled by the phages' protection of its genetically-encoded tag and by the rise of sensitive next-generation sequencing (NGS). The sensitivity achieved by NGS has enabled several other ligand discovery technologies including DNA-encoded libraries,[19,20] mRNA display,[14,21] and yeast display.[22,23]

However, the discovery of peptide ligands to biomolecular targets using phage display faces several fundamental challenges, including amplification bias and the isolation of target-unrelated phage.[24–26] Traditionally, three to five rounds of bio-panning are required to enrich high-affinity peptides with specificity towards the target.[1] While designed to eliminate nonbinding phage, the biological bacterial-based amplification of phage bio-panning can introduce bias. A recurrent challenge is the propagation and isolation of target-unrelated phage (TUP) variants, which often have mutations that confer a growth advantage in biological amplification in host bacteria.[24–26] The isolation of target-specific phage can be especially challenging if the biomolecular target struggles to drive the affinity selection or "pulldown" of bio-panning (i.e., due to its disorder or allostery)[27] resulting in the predominant isolation of target-unrelated phage variants. In addition, the efficiency of bacterial amplification can be varied with high-affinity candidates being missed or in low representations.[26] Thus, the resulting NGS data readout from bio-panning has additional bias and background "noise" that convolute the straightforward ranking of peptides by their affinity toward the target.

Several descriptors attempt to overcome these challenges by parsing the bio-panning NGS data and ranking peptides based on their target-specificity. The bio-panning performance of each peptide can be understood based bioinformatic fitness descriptors of i) protein specificity, ii) enrichment, as well as iii) similarity to other identified peptide ligands. First, protein specificity has been quantified by fold change (FC), and with its associated statistical confidence (p-value) is usually represented as a volcano plot similar to other bioinformatic analyses.[4,5,28] However, utilizing FC with its p-value alone may struggle to identify high-affinity peptides that do not exhibit rapid increases in enrichment due to any detrimental amplification bias. Moreover, an identical subset of phage must be present across the proteins investigated in each replicate; otherwise, the fold change comparison between slightly different subsets would inherently lead to the false positive identification.[26] Second, peptide enrichment through bio-panning rounds has been quantified by enrichment ratio (ER).[19,26] However, enrichment does not include any measure of protein specificity and thus relies on the experiment design and/or post-analysis to prevent the isolation of nonspecific peptides. Third, clustering can group peptides by their chemical similarity and aid in the identification of frequently appearing peptide motifs,[29] but can be prone to over- or underfitting due to the difficulty in determining an optimal number of clusters. Additionally, the presence of nonspecific and parasitic peptides can further obfuscate clustering efforts if they share sequence similarity with putative hits. Many clustering methods will also make assumptions about the general shape of the data, such as *K*- means clustering assumes spherically shaped datasets. The high dimensionality of the input data along with the lack of data labeling can lead to issues with local versus global feature relevance.[30,31]

Machine learning (ML) is set to facilitate a paradigm shift in drug discovery and development for its ability to reveal underlying or nonobvious patterns beyond statistical analysis. Thus, it has been deployed for the discovery of antimicrobial, cell-penetrant, or immunogenic peptides.[32–34] For phage display, ML has improved discovery outcomes by being trained directly on the NGS data,[35] on curated sequences for classification,[28] or on peptide fitness descriptors (e.g., FC, ER).[36] However, these approaches using one type of data (e.g., ER) may be affected significantly by the isolation of target-unrelated phage and amplification biases. In contrast, the combination of all fitness descriptors and sequence-based information may provide a more complete set of information to elucidate high-affinity peptides from phage display. To our knowledge, a combined model has not been developed, likely because each piece of information (fitness descriptor or sequence) is more amenable to separate regression or classification tasks, respectively.

Herein, we describe the application of proxy learning for revealing underlying motif patterns within NGS phage display data to improve the discovery of high-affinity ligands (Figure A.1B). Specifically, we combine bioinformatic fitness (FC and ER) and peptide sequence information through a loss-metric mismatch and enforce proxy learning to model the bio-panning NGS data. We completed phage display against mouse double minute 2 (MDM2) with anti-hemagglutinin antibody (12ca5) as a control, given that the known binding MDM2 motif could facilitate model evaluation. This involves the utilization of a heavily regularized neural network bidirectional long short-term memory (BiLSTM), trained in a supervised manner upon multiple data (Sequence input, FC and ER output). By employing this transfer learning technique, we sought to clarify the underlying peptide motifs that could drive affinity from the biases present in the data (i.e., “noise”). Where previous ML approaches have proven successful using either metric (FC or ER) independently,[19,35] to our knowledge this is the first work to combine both sources of information for joint denoising and the first work to apply proxy transfer learning to the task of genetically-encoded affinity selection or bio-panning. Our proxy BiLSTM model demonstrates a remarkable improvement, with a greater than 300% increase in the prioritization of motif-containing high-affinity peptides (termed “hit rate”) compared to any experimental-only approach. (Figure A.1C). Furthermore, we examine the structure-activity relationship (SAR) and investigate the denoised peptide motif using UMAP and shapely additive analysis. From this framework, future work will evaluate the binding affinity of prioritized peptides across a wider range of biomolecular targets to assess generality of the approach.

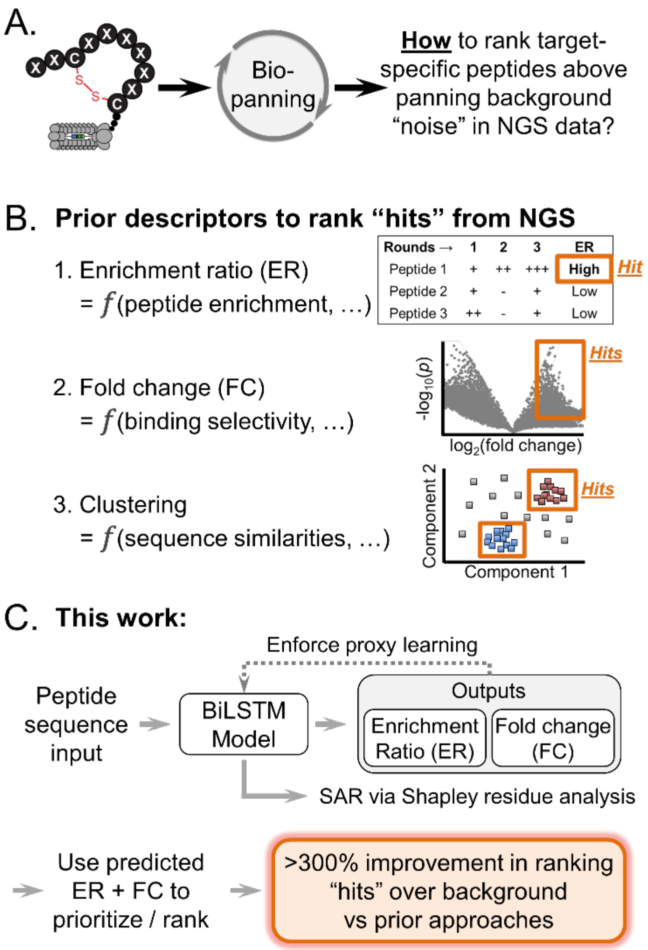

A.
X X C X X X X X X C
S S
Bio-panning
How to rank target-specific peptides above panning background "noise" in NGS data?
B. Prior descriptors to rank "hits" from NGS
1. Enrichment ratio (ER)
= ƒ(peptide enrichment, …)
Rounds → 1 2 3 ER
Peptide 1 + ++ +++ High
Peptide 2 + - + Low
Peptide 3 ++ - + Low
Hit
2. Fold change (FC)
= ƒ(binding selectivity, …)
-log10(p)
log2(fold change)
Hits
3. Clustering
= ƒ(sequence similarities, …)
Component 2
Component 1
Hits
C. This work:
Enforce proxy learning
Peptide sequence input
BiLSTM Model
Outputs
Enrichment Ratio (ER)
Fold change (FC)
SAR via Shapley residue analysis
Use predicted ER + FC to prioritize / rank
>300% improvement in ranking "hits" over background vs prior approaches

Figure A.1: *The approach presented in this work combines descriptors of peptide fitness in phage bio-panning along with sequence information to elucidate target-specific peptide hits from complex sequencing data. (A) A significant challenge of* de novo *discovery with phage display is the elucidation and ranking of target-specific peptide binders above the bio-panning background "noise," contributed by amplification bias and isolation target-unrelated phage, seen in next-generation sequencing data (NGS). (B) Several approaches process NGS data and rank peptide sequences for experimental synthesis and validation to improve research efficiency and maximize discovery success. Fitness descriptors of each individual peptide in bio-panning include enrichment ratio (ER), which quantifies the round-to-round change in individual peptide enrichment, and fold change (FC), which quantifies the peptide-protein selectivity. Clustering analysis has also been performed to parse and group the isolated peptides based on chemical similarity. (C) This work combines all descriptors (FC + ER) and sequence information within a bidirectional long short-term memory (BiLSTM) trained model that undergoes "proxy" training. Proxy training limits the learning of BiLSTM on pure accuracy of experimental FC and ER, which contain bias, and balances learning from the input sequence information as well. Overall, the resulting BiLSTM model provides a > 300% increase in accuracy of ranking peptide hits (motif-containing peptides) from the NGS background, termed "hit rate."*

A.2. Results and Discussion

Multiplexed phage display libraries with linear and macrocyclic peptides were selected against mouse double minute 2 (MDM2) and anti-hemagglutinin antibody (12ca5, see Figure A.2A). Briefly, three rounds of phage display panning were completed using an automated protocol, with the protein target pre-immobilized on magnetic beads. Bio-panning was completed simultaneously using multiple libraries linear peptide libraries (X_{12} and X_7) and macrocyclic peptide libraries ACX_7C and AX_MCX_NC (M+N=6) together, where the libraries were incubated first with unlabeled magnetic beads to remove high-affinity bead binders. Nonspecific binding was blocked with 2% non-fat milk for 1 hour before incubation of the protein target with depleted pooled phage library together for 1 hour. Overall, these measures to limit nonspecific binding were successful, with only 0.06% of peptide sequences containing the off-target streptavidin binding motif HPQ.

Within this work, the validation of our approach to prioritize high-affinity peptides will be judged by the presence of characteristic motif-containing sequences. Specifically, the known 12ca5 binding motif is D**DY(A/S)[37–39] and the known MDM2 binding motif from prior phage display work is F**ΦΦ, where Φ are the hydrophobic amino acids phenylalanine, tryptophan, leucine, isoleucine, valine and tyrosine.[40–44] As 12ca5 is a peptide-binding antibody, we placed more focus on validating our methodology to isolate motif-containing sequences against MDM2, with 12ca5 serving as a positive control.

In our data, FC and ER appeared partially orthogonal, though neither alone appeared able to easily isolate high-affinity, target-specific peptides, or “hits” from the data (shown in orange in Figure A.2). Both FC and ER are driven by the affinity selection process in bio-panning and each has led to the identification of high-affinity binders from genetically encoded libraries.[4,5,19,26] However, both fitness descriptors may be susceptible to amplification bias and target-unrelated phage contributing to the background “noise” from NGS data. Common to all affinity selections, a weaker signal may be observed if the biomolecular target does not strongly drive the selection through affinity-ranked interactions. In comparison to 12ca5, MDM2 appeared less able to drive a strong affinity selection as seen in the “volcano” plot of -ln(*p*-value) vs ln(FC) (Figure A.2B). The orange peptide hits would be poorly isolated from an FC-based approach. Relating FC to ER, we see that additional information emerges (Figure A.1C) with the appearance of three peptide groups. The orange peptide hits appear predominantly in the high FC and high-to-modest ER region, as expected. However, only 2.8 % of the peptides in this region (high FC (> 2.5)[4,5] and high-to-modest ER (> 0)) contain the desired, high-affinity MDM2 motif. Another high FC population has low ER and may represent weak target-selective binding peptides. The last population have low FC which are not target selective. The clearest localization of the desired orange motif-containing peptides can be seen in a three-dimensional overlay (Figure A.2D), where most of the hits appear to have a high FC and ER and a modest *p*-value. Overall, these data suggest that the combination of FC and ER may benefit any analysis approach toward the identification of specific high-affinity peptides from phage display data (Figure A.2B,C,D).

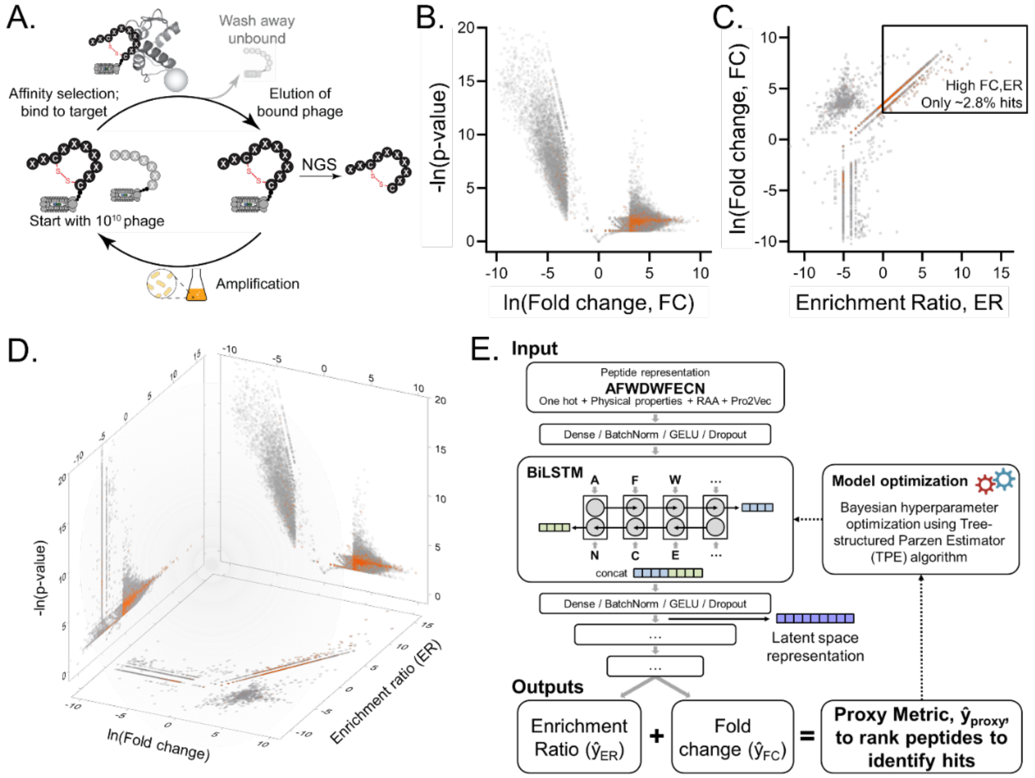
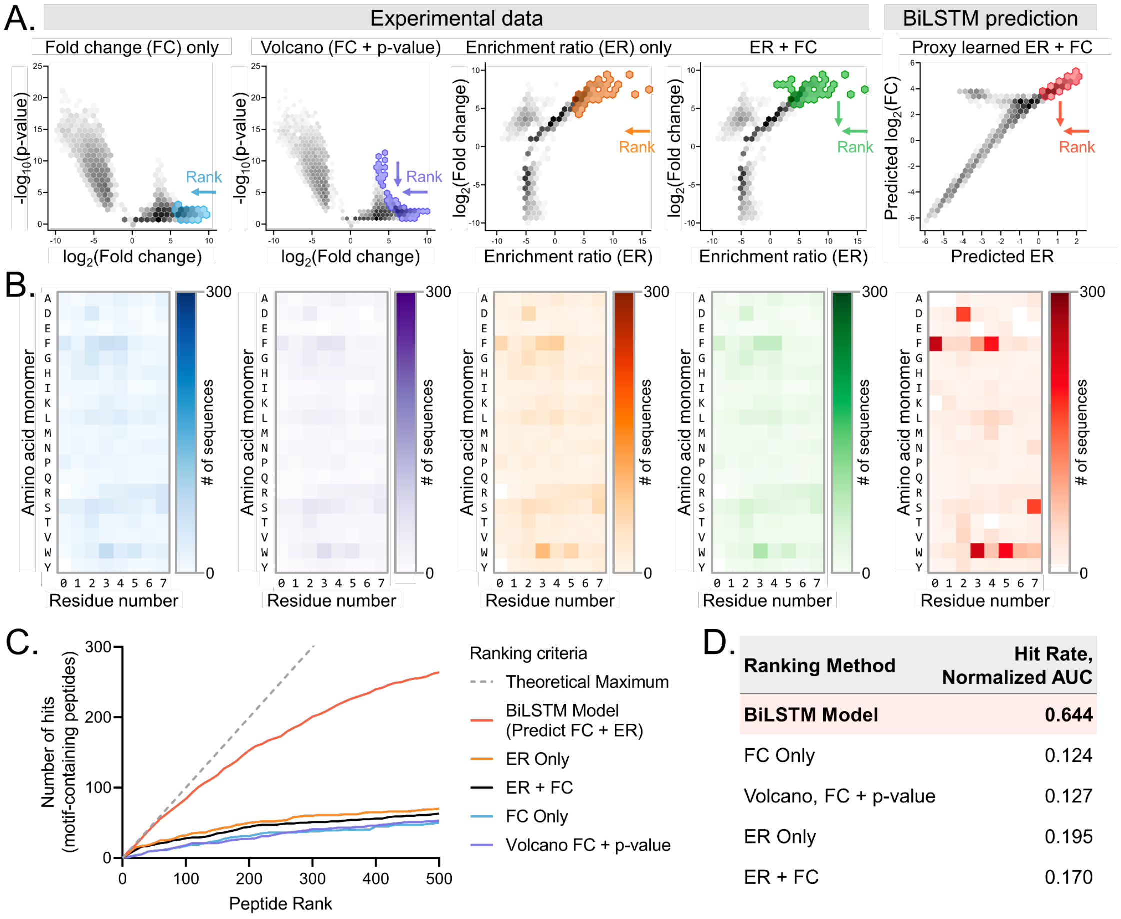


Figure A.2: *Fold change (FC) and enrichment ratio (ER) provide partially orthogonal information that when combined with peptide sequence in a proxy machine-learned approach may improve the isolation of high-affinity target-specific peptide ligands from phage display against MDM2 with 12ca5 as control. (A) Phage display bio-panning seeks to isolate peptide ligands by iterative affinity selection and biological amplification of bound phage which genetically-encode the displayed peptide. However, this process is plagued by the isolation of target-unrelated phage and is susceptible to a weakly-driven affinity selection. (B) Volcano plot of -ln(*p-value*) vs ln(FC), similar to other bioinformatic analyses. (C) The combination of FC and ER partially reveals motif-containing peptide hits as a population described by both high FC and high ER. However, only 2.8% of the peptide sequences within the high FC (> 2.5) and high-to-modest ER (> 0) region contain the high-affinity MDM2 motif. (D) The 3-dimensional projection of the data to include all three criteria (ER, FC and its associated p-value) aids in visualizing the location of the peptide hits. Orange points in plots B through D represent peptides that contain the MDM2 motif. € Our approach utilizes a bi-directional long short-term memory (BiLSTM) ML approach to direct the proxy learning of peptide sequence toward FC and ER.*

Before employing a more powerful model, we first determined that two common clustering methods, *K*-means clustering[30,45] and CD-HIT,[29] were insufficient to identify high-affinity peptide sequences from the MDM2 phage display data. For *K*-means clustering, the peptides were encoded by amino acid, each represented as a 36-length vector from one-hot encoding, relative propensity for binding score,[46] DELPHI predicted protein interaction score,[47] and 14 physicochemical property descriptors.[46,48] Residue-based encoding directly improves the ability to perform SAR analysis, ranging from amino acid-specific (one-hot) to generalizable (physical property) contributions toward binding affinity. Encoded peptides were decomposed using dimensionality reduction with Uniform Manifold Approximation and Projection (UMAP)[49] and principal component analysis (PCA). The data was then clustered using the *K*-means algorithm after optimization of the number of clusters *K* using the elbow method, and a logo plot was generated for each cluster (Figure A.9D). Additionally, we calculated the ER or FC of each cluster to attempt to guide the determination of target-selective clusters (Figure A.9B). However, the clustering primarily produced separated clusters for each library utilized (i.e., separating linear and macrocyclic libraries), affording no meaningful information about potential peptide hits or target-selective clusters within each library. Only a single cluster containing the 12ca5-based aspartic acid motif could be identified, with no clusters containing the MDM2 motif in any form. This result was made more apparent when the location of 12ca5- and MDM2-motifs were overlaid on the clusters, which were dispersed across all clusters and indicating that our clustering approach did not isolate any motif-containing hit peptides. Clustering using CD-HIT was also attempted across a range of similarity metrics but was unable to identify any cluster of peptides larger than three members. Overall, our efforts indicated that isolation of motif-containing peptides can be challenging with unsupervised clustering, warranting the use of more powerful tools to combine the partially orthogonal information from the peptide sequence, ER, FC and its associated p-value.

We proposed a proxy learning approach to enforce learned connections between sequence patterns and noisy experimental data of bio-panning efficiency (ER) and target selectivity (FC and its p-value) using a bidirectional long short-term memory (BiLSTM) model. Peptides that exhibit high FC and ER have a higher fraction of high-affinity, motif-containing peptide hits; but, for MDM2 here, only 2.8% of these peptides were hits. Thus, we hypothesized that multi-source learning could balance prioritization with respect to sequence patterns in the latent space, while learning and “denoising” the experimental surrogate metrics of FC and ER. Set-up as a supervised regression task, we employed a BiLSTM model with encoded peptide sequence inputs and predicted outputs of ER ($\hat{y}_{ER}$) and FC ($\hat{y}_{FC}$) can then be summed to provide a final proxy metric ($\hat{y}_{proxy}$, Figure A.2E). Thus, this proxy metric is a single number output that can be used to rank the peptides for their likelihood as high-affinity peptide ligands, and evaluated by whether they contain the MDM2-binding motif. From the input peptide sequence, we employed a BiLSTM model to predict the proxy metric (Figure A.2E).[50] The BiLSTM architecture was chosen for its capacity to preserve sequence order, represent peptide libraries of multiple lengths without padding, and handle cases of motif frame shift and macrocycle bidirectionality.

Strict regularization in our model demonstrated the greatest training performance to identify and highly rank MDM2 and 12ca5 peptide hits via the predicted proxy learned metric. Highly regularized, sparse learning has previously been shown to be important in DNA-encoded discovery.[20] Regularization limits overfitting during training and appeared well-suited for proxy learning since a highly accurate ML regression with experimental ER and FC would at most provide a mixed ranking with only 2.8% of the peptide containing the high-affinity motif (Figure A.2C). To arrive at this conclusion, we used Bayesian hyperparameter optimization with the Tree-Structured Parzen Estimator algorithm[51] to evaluate multiple hypotheses using Holm-Bonferroni corrected correlation of hyperparameter values against model performance (Figure A.10). Stringent regularization improved the model’s ability to highly rank peptide hits at the cost of reducing prediction accuracy (Figure A.11), following the classic variance-bias tradeoff.[52] Thus, we confirmed that strict regularization reduced overfitting and improved the BiLSTM model’s ability to perform multi-source learning on peptide sequence, FC, and ER to highly rank peptide hits from NGS data.



Ranking Method	Hit Rate, Normalized AUC
BiLSTM Model	**0.644**
FC Only	0.124
Volcano, FC + p-value	0.127
ER Only	0.195
ER + FC	0.170

Figure A.3: *BiLSTM model highly ranks MDM2 motif-containing peptide hits >300% better than any combination of experimental approaches. Ranking prioritizes the investment of synthesis and experimental binding validation toward peptides that have the highest predicted confidence to be hits. (A) Hexbin projections with highlighted zones corresponding to the top 500 peptides as determined by the different strategies to rank the peptides for their potential as peptide hits. Arrows shown in the bottom right display the direction of ranking (x-direction, y-direction, or both). (B) Positional frequency matrix of the top 500 identified peptides. The macrocyclic 9-mer library contained most of the motif-containing peptide hits, outperforming the other libraries. Thus, the positional frequency matrices of the top 500 show the 9-mer variable region of the 9-mer library (cysteines not shown). (C) A plot of the number of identified peptide hits versus their ranking shows that the BiLSTM model outperforms all other experimental methods to rank the peptides. The calculation of the normalized area under the curve reveals the BiLSTM model performs >300% better. D. Calculation of the area under the hit rate curve in C indicates that 64% of the top 500 BiLSTM ranked peptides contain the MDM2 motif.*

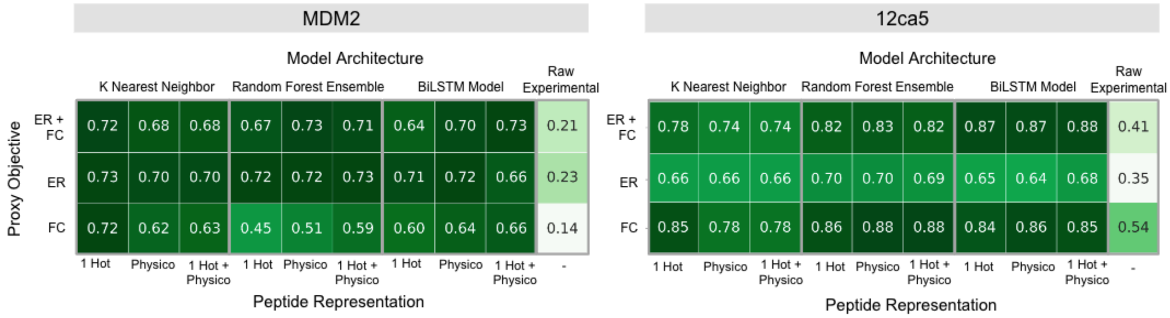
Our BiLSTM model was able to isolate peptides containing F**ΦΦ motif known to drive high-affinity peptide binding to MDM2 through proxy learning, where bioinformatic or statistical approaches failed (Figure A.3A,B). Additional confidence can be placed in the method used for the ligand discovery when a high number of compounds containing a similar set of critical features or residues are observed, called "motif."[8,9,53–56] In the context of affinity selection, these motifs are generally assumed to facilitate high-affinity interactions. Thus, our proposed method would be successful if it can isolate peptides containing the MDM2 motif from the NGS data. As such, we isolated the top 500 peptides ranked by FC-only, FC with its associated p-value (volcano plot),[4,5,28] ER-only, and FC+ER, and compared it to the 10-fold cross-validated prediction of the proxy metric ($\hat{y}_{proxy} = \hat{y}_{ER} + \hat{y}_{FC}$). The top 500 peptides from each of these approaches are highlighted in Figure A.3A along with the axis of the approach used to rank and isolate the 500 peptides (e.g., FC-only uses only the x-axis of the volcano plot). We assessed their motif pattern by using a positional frequency matrix seen in Figure A.3B. Only the BiLSTM model showed a clear motif pattern closely matching the F**ΦΦ motif known to drive high-affinity peptide binding to MDM2.[40–44] The other approaches showed no clear discernable pattern and appeared random.

Our BiLSTM was able to rank peptide hits from the NGS data >300% better hit rate than bioinformatic or statistical approaches, concentrating the likelihood of success for initial synthesis validation attempts (Figure A.3C). We sought to rank the peptides from bio-panning NGS data to efficiently prioritize the investment of experimental validation toward peptides with the highest likelihood of being high-affinity hits. A theoretically perfect ranking would rank all motif-containing peptides above all non-motif-containing peptides, meaning every peptide synthesized in the order of the rank would be a hit (gray dashed line in Figure A.3C). With our targets, we again used their known motifs to evaluate if our model correctly ranked peptide hits over other peptides within the NGS data. The proxy metric from the 10-fold cross-validated model can then be assessed for its accuracy throughout by ranking from the 1st to 500th peptide using a “hit rate.” The hit rate assesses the ranked peptides for their presence of the MDM2 motif as a function of their rank, as a normalized area under the curve (Equation 1).

Equation 1:

$$\text{Hit Rate} := \sum_{k=1}^{K} \frac{|\text{Hits}(\widehat{\text{Rank}}_{1:k})|}{|\text{Hits}(\text{Rank}_{1:k})|} \quad , \quad \text{Hits(P)} := \{p \in P | p \text{ matches the motif pattern}\}$$

Thus, per peptide invested in experimental validation, the hit rate assesses how quickly and how many hits are identified, with up to the rank of 500 peptides shown in Figure A.3C,D. Overall, the BiLSTM model hit rate significantly outperforms all approaches to analyze the experimental data by >300%, with ER only providing the next best performance.



MDM2

Proxy Objective	K Nearest Neighbor 1 Hot	K Nearest Neighbor Physico	K Nearest Neighbor 1 Hot + Physico	Random Forest Ensemble 1 Hot	Random Forest Ensemble Physico	Random Forest Ensemble 1 Hot + Physico	BiLSTM Model 1 Hot	BiLSTM Model Physico	BiLSTM Model 1 Hot + Physico	Raw Experimental -
ER + FC	0.72	0.68	0.68	0.67	0.73	0.71	0.64	0.70	0.73	0.21
ER	0.73	0.70	0.70	0.72	0.72	0.73	0.71	0.72	0.66	0.23
FC	0.72	0.62	0.63	0.45	0.51	0.59	0.60	0.64	0.66	0.14

12ca5

Proxy Objective	K Nearest Neighbor 1 Hot	K Nearest Neighbor Physico	K Nearest Neighbor 1 Hot + Physico	Random Forest Ensemble 1 Hot	Random Forest Ensemble Physico	Random Forest Ensemble 1 Hot + Physico	BiLSTM Model 1 Hot	BiLSTM Model Physico	BiLSTM Model 1 Hot + Physico	Raw Experimental -
ER + FC	0.78	0.74	0.74	0.82	0.83	0.82	0.87	0.87	0.88	0.41
ER	0.66	0.66	0.66	0.70	0.70	0.69	0.65	0.64	0.68	0.35
FC	0.85	0.78	0.78	0.86	0.88	0.88	0.84	0.86	0.85	0.54

Figure A.4: *Hit rate benchmark of model architectures, peptide representation, and proxy objective on both the MDM2 and 12ca5 target protein systems. All results reported on a 20% holdout set taken before hyperparameter tuning.*

Next, we benchmarked the performance of the BiLSTM model against random forests (RF) and K-Nearest Neighbor (KNN) models which are commonly employed in the analysis of NGS-based discovery data using the hit rate ranking metric.[19,57] Hyperparameter optimization of the RF model prioritized shallow tree depth (max depth of 10) and greater number of estimators (200) presumably to increase regularization by averaging multiple decision trees that individually suffer from high variance, resulting in the elucidation of underlying data patterns that are robust to noise in the dataset (Figure A.10). Similarly, during hyper parameter optimization, regularization in KNN models was implicitly controlled by adjusting for large averaged neighbor set sizes (35). Ultimately, all three model architectures achieve comparable performance demonstrating competitive amenability to the proxy learning task. Notably, all three model architectures across all peptide representations and proxy objectives achieved significantly higher hit rates relative to the raw experimental values. This result serves to further underscore the vital importance of framing hit discovery as a proxy learning task and applying highly regularized models to effectively parse the complex NGS data and discerning meaningful underlying patterns in the peptide sequences.

The BiLSTM model demonstrates greater robustness for combined peptide representations (1 Hot and Physicochemical) and multi-learning proxy objectives (ER + FC) relative to RF and KNN models (Figure A.4). In the two distinct protein systems examined, the experimental information sources (ER, FC, and P Value) encode varying levels of information pertaining to peptide fitness. Specifically, for MDM2, across the model benchmarks, ER provides the most salient information for peptide hit ranking, indicating round to round enrichment is the dominant indicator. Whereas for 12ca5, across the model benchmarks, FC and P-Value proxy objectives prove to be the most informative indicators of peptide binding, indicating intra-round replicates to be the dominant indicator. The BiLSTM model successfully incorporates the partially orthogonal information of FC + ER as well as multiple peptide encodings (one-hot + physicochemical parameters) to improve hit rate ranking whereas the RF and KNN models only learn to competitively rank peptides from single-task learning objectives and with individual peptide encodings (physicochemical for RF and one hot encoding for KNN). For peptide systems without known binding motifs and where RF and KNN cannot be benchmarked on different single-task objectives and individual encodings, the robustness of the BiLSTM model to multi-task learning and combined representations proves imperative.

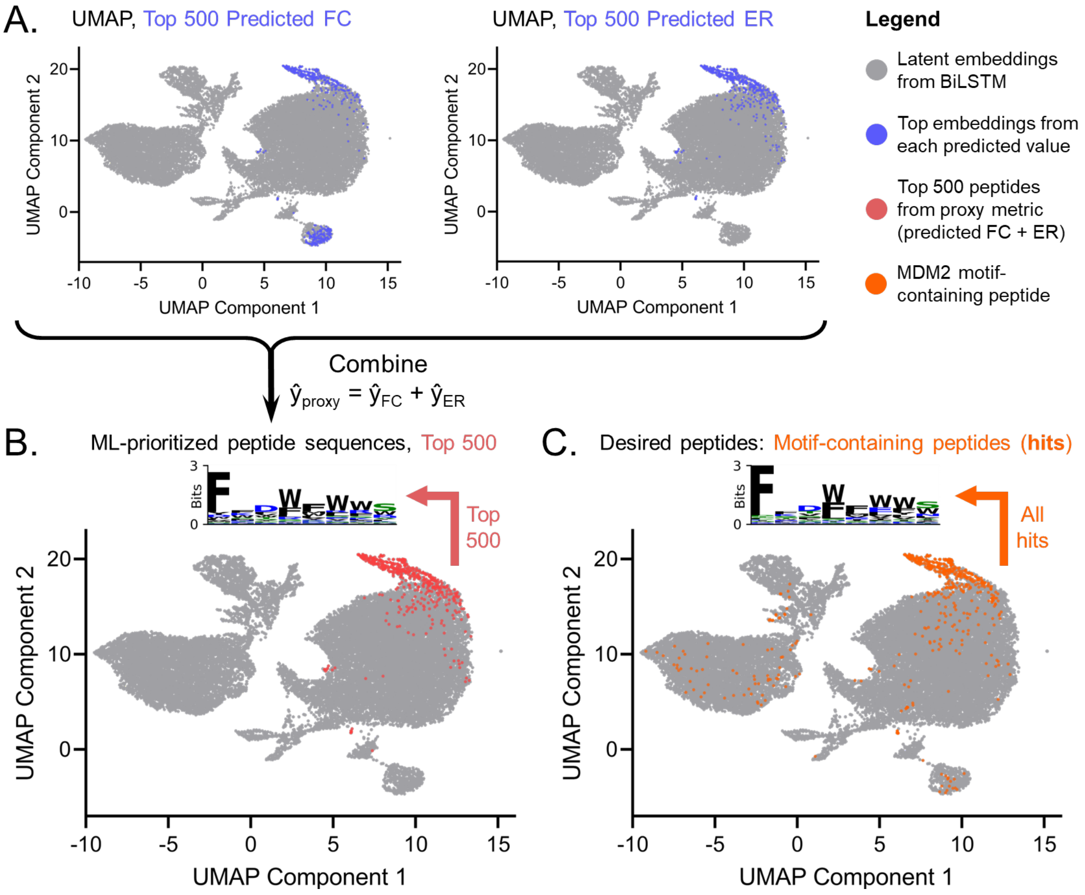


Figure A.5: *The UMAP decompositions of the learned latent features for each peptide indicate successful proxy learning, with strong prioritization toward clustering similar input peptide sequence features. (A) UMAP transformation obtained from the penultimate layer of the BiLSTM model with the top 500 predicted FC and ER peptides highlighted. (B) In combining FC and ER into the proxy metric ($\hat{y}_{proxy} = \hat{y}_{ER} + \hat{y}_{FC}$), the top 500 peptides are grouped together for MDM2 with a logo plot exhibiting multiple frameshifts of hydrophobic sequence features. (C) On the same UMAP plot, the motif-containing sequences (F**ΦΦ) are overlaid, indicating the nearly the same group of peptides prioritized in the top 500 by the BiLSTM model in B. A logo plot of the motif-containing peptides also shows nearly identical sequence patterns as prioritized by the BiLSTM model. Cysteine residues are excluded from the logo plots.*

To understand the BiLSTM model further, we confirmed proxy learning toward FC and ER as well as the cluster-like consolidation of motif-containing peptides in the BiLSTM latent space seen by UMAP (Figure A.5). Examining the latent space BiLSTM embeddings reveals the balance and influence of the proxy learning toward FC and ER (outputs) versus the model's ability to understand and consolidate similar peptide sequence input. The UMAP decomposition into two dimensions of all latent peptide embeddings from the penultimate layer of the BiLSTM is shown in Figure A.5. Additionally, the peptides with the top 500 predicted ER and FC values are highlighted to indicate the location of peptides that were highly ranked by the model. For MDM2, the top 500 predicted FC peptides are clustered in two distinctive islands, while the top 500 predicted ER peptides are clustered in a single island.

The clear consolidation of highly ER- and FC-ranked peptides indicated that the BiLSTM model placed significant weight toward understanding and grouping key motifs and denoising statistically significant structural patterns. The model has less prioritization toward the purely accurate matching of predicted ER and FC to experimental values, confirmed by parity plots (Figure A.11) as well as the lack of consolidation of peptides with high experimental FC or ER (Figure A.12). For MDM2 (Figure A.4), we see one region in the UMAP plot that contains motif-containing peptide sequences that strongly overlaps with those predicted to have high ER and FC, which are summed to give the proxy metric. For 12ca5, the top 500 peptides exhibiting the highest predicted 12ca5 ER and 12ca5 FC scores also exhibit a significant co-localization within the UMAP projection. In addition to successfully identifying 12ca5 motif-containing sequences (*DYA*), the BiLSTM model prioritized a similar sets of anionic peptides containing motifs including D**DY*, which likely drives high affinity binding, and LE*E, which has not been reported before. Overall, these findings underscore the nontrivial capability of our proxy learning approach to effectively group and denoise peptides from NGS data across these protein targets.

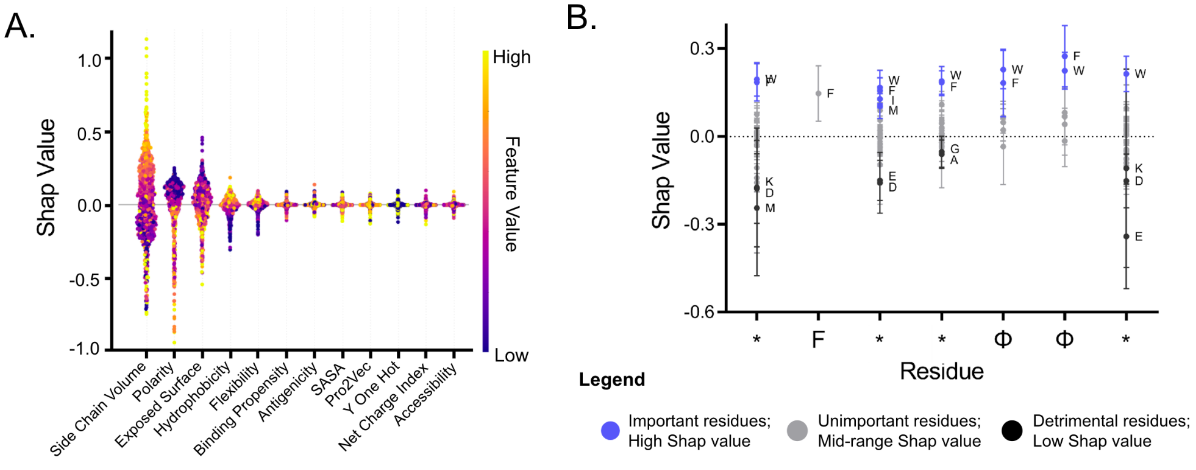


Figure A.6: *Built-in model interpretability using Shapley analysis provided amino-acid level SAR. (A) Shapely feature importance across representation features as calculated by the 10-model ensemble trained via cross validation splitting on a test set of 500 randomly sampled peptides. This result indicated that high volume, low polarity, high hydrophobicity, and high flexibility are predicted to improve MDM2 binding propensity. (B) Positional Shapely feature importance across residue identities as calculated by the 10-model ensemble on the set of all 558 MDM2 motifs containing hits within the dataset. Sequences are aligned by motif position, and error bars are calculated according to the standard deviation of Shapely values per residue across all peptides and all models. This result underlies the importance of hydrophobic amino acids to drive binding and the potential for small or polar amino acids to disrupt peptide binding to MDM2.*

In addition to potentiating discovery, we can analyze the BiLSTM model results and prioritizations at the individual amino acid level to gain structure activity relationship (SAR) information using Shapley Additive Explanation analysis.[58] While complex, ligand discovery data has a wealth of underexploited information from the affinity selection of many compounds. We have demonstrated that the proxy-trained BiLSTM model can identify and rank peptides for their likelihood as high-affinity binders with high accuracy. From the ranking action, the valuable sequence motifs, down to the specific amino acid level, can be inferred using Shapley Additive Explanation analysis of our 10-fold cross-validated BiLSTM model ensemble. Shapley analysis uses coalition game theory to calculate the contribution of each encoded feature (for this work amino acid and physicochemical property) to the final model prediction.[58] Thus, this analysis identifies the importance of each representation feature (Figure A.6A) for its influence in driving MDM2 binding.

For MDM2, high volume, low polarity (P_{12} polarizability), low exposed surface area, and high (H_{12}) hydrophobicity descriptors[46,48] were found to be the most indicative characteristics of residues to drive high-affinity binding (Figure A.6A). These parameters match well with the properties of canonical uncharged aromatic amino acids, including tyrosine, tryptophan, and phenylalanine, which are known to be a part of the MDM2 motif that drives high affinity binding. Other features such as low exposure, high flexibility, and median side chain net charge index according to the cross validated model ensemble correlate with the MDM2 binding likelihood. From the same analysis, favorable properties to drive 12ca5 can be inferred as well from the "low" Shapley values or the stand-alone analysis (Figure A.15). High polarity, high exposed surface area, low flexibility, and low hydrophobicity were seen to likely drive 12ca5 binding, consistent with the D**DYA motif. For both proteins, the two pretrained descriptors of relative binding propensity and DELPHI protein interaction scores were less connected to peptide binding activity. In addition, one-hot descriptors show relatively low shapely importance ranges, which suggested that the model ensemble eschewed specific categorical understanding in favor of deeper physicochemical understanding.

For additional SAR interpretability, we summed the Shapley values across the representation dimension to determine positional importance, also referred to as Positional Shapley (PoSHAP) (as illustrated in Figure A.6B).[59] Positional shapely analysis of peptide "hits" aligned by the theoretical MDM2 motif (Figure A.6B) allows us to quantitatively compare residue importance at different positions. Our findings revealed that uncharged aromatic amino acids had the most influence on the model's performance, with the highest contributions according to our proxy metric. Hydrophobe 1 (Φ_1 in F**$\Phi_1\Phi_2$) is often seen to be tyrosine in literature (MDM2 cite) but seen to be highly weighted as tryptophan by our model. The negatively charged residues aspartic and glutamic acid in position 3 or 7 (relative to the start of the motif) were recognized to significantly reduce propensity to bind by our model ensemble in addition to other small or polar amino acids. These polar amino acids likely prefer to be solvated rather than bound to the hydrophobic MDM2 patch surface. For 12ca5 (Figure A.15), the PoSHAP revealed significant importance of the aspartic acids within the motif (D**DYA) significantly more important for binding than the tryptophan and alanine residues. Overall, we hope that the integration of PoSHAP within the BiLSTM model improves the interpretability of the model and the SAR information gained from affinity selection and bio-panning discovery.

A.3. Conclusion

Multi-source proxy learning with a BiLSTM architecture effectively identified and ranked high-affinity peptide hits from NGS phage data based on the presence of their known binding motif to MDM2 and 12ca5. From phage display, bioinformatic statistical metrics including fold change (FC) and enrichment ratio (ER) provided partially orthogonal information to perform model training. Neither FC nor ER alone from the experiment was able to guide the identification and ranking of high affinity peptides from the NGS data, clearly indicating the "noise" present in the data, likely from favorable genetic bias of target-unrelated phage.[24,26] Because of this noise in the data, common clustering techniques including *K*-means clustering and CD-HIT were unable to robustly identify motif-containing peptides.

A proxy machine-learned approach was appropriate to combine information from FC and ER along with sequence level information to overcome the "noise" from target-unrelated phage and/or weak affinity-driven enrichment. Proxy learning was critically enforced by regularization, leading to the cluster-like consolidation of sequence feature information (Figure A.5) that was heavily weighted with FC and ER. This training led to the identification and ranking of peptides as hits to MDM2 >300% better than any combination of experimental approaches and stands to improve the efficiency and investment in experimental validation (Figure A.3). We benchmarked the BiLSTM model against other available supervised ML approaches to combine sequence, FC, and ER together including random forest (RF) and *K*-Nearest Neighbor (KNN) models. The BiLSTM model demonstrated similar hit rate benchmarks but was clearly more robust in its ability to combine the partially orthogonal FC and ER with the combined amino acid descriptors together for both MDM2 and 12ca5 (Figure A.4). In comparison, RF and KNN showed variability in its optimal encoding method and target objective (e.g., FC, FC+ER, etc), whereas the BiLSTM model improved with the more diverse information input. Lastly, the addition of Shapley Additive Explanation analysis allows for SAR-level information to be isolated from the ligand discovery experiment directly. From initial discovery experiments, Shapley analysis holds potential to guide the importance of peptide amino acid composition (Figure A.6A) as well as with respect to sequence (Figure A.6B), informing derivatization efforts.

Next, we will seek to apply this proxy-learning approach to phage display against novel targets, with a strong emphasis on experimental validation. This future direction will reveal the connection between predicted hit rate against these model protein targets (12ca5 and MDM2) as well as establish a true experimental hit rate against more challenging targets. We expect that this proxy-learned BiLSTM model will generally improve the experimental hit rate and discovery of high-affinity peptide ligands against biomolecular targets, all toward the generation of peptidomimetic therapeutics.

A.4. Materials

Mouse anti-hemagglutinin antibody (clone 12ca5) was purchased from Columbia Biosciences (Cat: 00-1722) and biotinylated according to a previously published protocol MDM2$^{25\text{-}109}$ was synthesized using automated flow synthesis as described below. Dynabeads™ His-Tag and Dynabeads™ MyOne™ Streptavidin T1 were obtained from Thermo Fisher

H-Rink Amide-ChemMatrix resin was obtained from PCAS BioMatrix Inc. *N,N*-diisopropylethylamine (DIEA; ReagentPlus ≥99%), piperidine (ACS reagent, ≥99.0%), trifluoroacetic acid (HPLC grade, ≥99.0%), triisopropylsilane (≥98.0%), acetonitrile (HPLC grade), formic acid (FA, ≥95.0%), dimethyl sulfoxide (DMSO, HPLC grade, ≥99.7%), 1,2- ethanedithiol (EDT, GC grade, ≥98.0%), and AldraAmine trapping agents (for 1000 – 4000 mL DMF, catalog number Z511706) were purchased from Sigma-Aldrich. Fmoc-protected amino acids (FmocAla-OHxH2O, Fmoc-Arg(Pbf)-OH; Fmoc-Asn(Trt)-OH; Fmoc-Asp-(Ot-Bu)-OH; FmocCys(Trt)-OH; Fmoc-Gln(Trt)-OH; Fmoc-Glu(Ot-Bu)-OH; Fmoc-Gly-OH; Fmoc-His(Trt)-OH; Fmoc-Ile-OH; Fmoc-Leu-OH; Fmoc-Lys(Boc)-OH; Fmoc-Met-OH; Fmoc-Phe-OH; Fmoc-ProOH; Fmoc-Ser(But)-OH; Fmoc-Thr(t-Bu)-OH; Fmoc-Trp(Boc)-OH; Fmoc-Tyr(t-Bu)-OH; Fmoc-Val-OH); Fmoc-Lys(Biotin)-OH were purchased from the Novabiochem product line of Millipore Sigma; FmocHis(Boc)-OH was purchased from ChemPep, Inc. O-(7-azabenzotriazol-1-yl)-N,N,N’,N’-tetramethyluronium hexafluorophosphate (HATU, ≥97.0%) and (7-azabenzotriazol-1- yloxy)tripyrrolidinophospho-nium hexa-fluorophosphate (PyAOP, ≥97.0%) were purchased from P3 Biosystems. Tween20 (Proteomics grade) was purchased from VWR. Bovine Serum Albumin (BSA; biotechnology grade) was obtained from VWR International. Water was deionized using a Milli-Q Reference water purification system (Millipore). Nylon 0.22 µm syringe filters were TISCH brand SPEC17984

A.5. Methods

A.5.1. Phage display

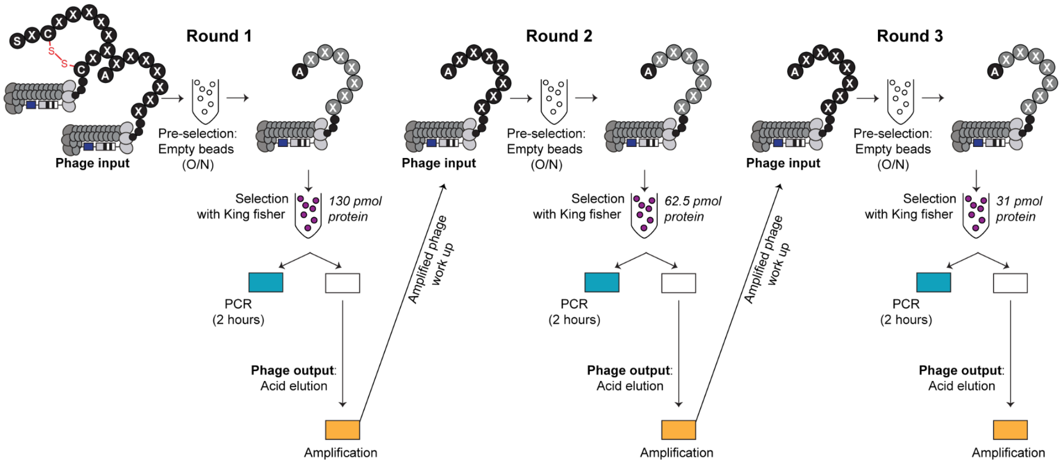


Figure A.7: *A typical phage display screening against immobilized proteins on magnetic beads.*

Phage display libraries (X_{12}, X_7 and ACX_7C) were purchased from New England Biolab (Cat# E8210S, E8211S and E8212S) The AX_MCX_NC (M+N=6) library was cloned and kindly donated from the Ratmir Derda lab (University of Alberta).

A.5.1.1. Phage amplification and precipitation

Phage were infected with 250 μL of E.coli K12 ER2738 for 5 mins. The infected E.coli were then transferred to 25 mL of pre-warmed LB broth for incubating for 5 hours at 37 °C with 220 rpm. After 5 hours, the solution was centrifuged at 5000×g for 10 mins. The supernatant was transferred to a 50 mL falcon tube with 5 mL of 6xPEG/NaCl to precipitate overnight at 4 °C. The next day, the solution was centrifuged at 14000×g and the supernatant was discarded. The pellet was redissolved in 1 mL 1×PBS. The dissolved phage solution was then transferred to a 1.7 mL centrifuge tube containing 200 μL of 6xPEG/NaCl to re-precipitate phage on ice for 1 hour. The precipitated phage solution was centrifuged at max speed for 30 mins, and the supernatant was discarded. The pellet was redissolved in 1 mL 1×PBS and centrifuged at max speed to remove large undissolved debris. The supernatant was transferred to a new 1.7 mL centrifuge and stored at 4 °C for next round selections.

A.5.1.2. Phage titering

10 μL of the phage solution was infected with 250 μL of E.coli K12 ER2738 for 5 mins. Then, the infected E.coli were combined with top agar and plated on an IPTG/Xgal LB plate. The plate was incubated overnight at 37 °C and calculated the phage concentration the next day.

A.5.1.3. Phage bio-panning, first three rounds of selection

In a 1.7 mL centrifuge tube, a mixture of phage library (10^{10} PFU/library or 10^{10} PFU amplified phage) was incubated with 20 μL of streptavidin magnetic beads overnight at 4 °C to deplete bead specific binders. On the next day, the depleted phage library mixture was transferred to a new 1.7 mL centrifuge tube for later use and the beads were discarded. 100 µL of streptavidin magnetic beads was washed with 1 mL of 1×PBS 3 times and was incubated with 130 pmol (1.2 eq) of biotinylated protein for 15 mins on ice. After protein immobilization, the protein-beads were washed with 1 mL of 1×PBS 3 times. The overall selection was performed on a KingFisher^{TM} Duo Prime Purification System. (See KingFisher protocol section A.5.1.5) The biotinylated protein immobilized beads, depleted phage library, blocking buffer and washing buffer were added into the KingFisher Plate in the corresponding wells. After selection, 100 µL of bead solution was transferred into a 1.7 mL centrifuge tube, the well was washed with 100 µL of PBS and transferred to the same centrifuge tube to a total of 200 µL of bead solution. 100 µL of the bead solution was added to 100 µL of glycine elution buffer (Glycine-HCl pH 2.2) for 9 min to elute phage from the beads. The elution supernatant was transferred into a new 1.7 mL microcentrifuge tube and neutralized with 15 µL of 1 M Tris-HCl (pH 9.1). The elution was amplified for the next round of bio panning. The remaining 100 µL of bead solution was amplified using PCR with an Illumina adapter sequence for Next Generation Sequencing.

A.5.1.4. Phage bio-panning, fourth round of selection

In Round 4, negative proteins were introduced to observe enrichment and specificity of the enriched library from Round 3.

In a 1.7 mL centrifuge tube, a mixture of phage library (10^{10} PFU/library or 10^{10} PFU amplified phage) was incubated with 20 µL of streptavidin magnetic beads overnight at 4 °C to deplete beads specific binders. On the next day, the depleted phage library mixture was transferred to a new 1.7 mL centrifuge tube for later use, and the beads were discarded. 100 µL of streptavidin magnetic beads was washed with 1 mL of 1×PBS 3 times and was incubated with 130 pmol (1.2 eq) of biotinylated protein for 15 mins on ice. After protein immobilization, the protein-beads were washed with 1×PBS 3 times. The overall selection was performed on a

KingFisher Instrument. (See KingFisher protocol Section A.5.1.5) The biotinylated protein immobilized beads, depleted phage library, blocking buffer and washing buffer were added into KingFisher Plate in the corresponding well. In negative control panning, the library against positive protein and against a negative protein was also performed in parallel. After selection with the KingFisher Instrument, 100 µL of bead solution was transferred into a 1.7 mL centrifuge tube, washed the well with 100 µL and transferred to the same centrifuge tube to a total of 200 µL of bead solution. 100 µL of the bead solution was buffered exchange with 100 µL of glycine elution buffer (Glycine-HCl pH 2.2) for 9 min to elute phage from the beads. The elution supernatant was transferred into a new 1.7 mL microcentrifuge tube and neutralized with 15 µL of 1 M Tris-HCl (pH 9.1). The elution was amplified for the next round of bio panning. The remaining 100 µL of bead solution was amplified by PCR with an Illumina adapter sequence for Next Gen Sequencing.

A.5.1.5. KingFisher^{TM} Duo Prime plate setup for selection

The protein immobilized beads suspension and other reagents were added to a 96 Deepwell Plate (Thermo Fisher, #95040450) as follows:

Table A.1: *Plate Setup for KingFisher^{TM} Duo Prime*

A	Protein coated magnetic beads (1×PBS Buffer)	300 µL
B	Reserved for 12-tip Deepwell magnetic comb (Thermo Fisher, #97003500)	1 mL
C	Wash Buffer (1×PBS buffer)	1 mL
D	Blocking Buffer (2% non-fat milk (w/v) in PBS Buffer)	1 mL
E	Solution of libraries (10^{10} PFU/mL 0.1 % Tween-20 (v/v), 0.2% non-fat milk (w/v) in PBS Buffer)	1 mL
F	Wash Buffer (0.1 % Tween-20 (v/v), 0.2% non-fat milk (w/v) in PBS Buffer)	1 mL
G	Wash Buffer (1×PBS Buffer)	1 mL
H	Wash Buffer (1×PBS Buffer)	1 mL

Elution strip:

1×PBS Buffer 100 µL/well

The following steps were performed using a KingFisher^{TM} Duo Prime Purification System with a magnetic comb to transfer the beads. The program is as follows:

A. Collect comb from row B
B. Collect beads from row A on comb,
C. Wash beads in row C – 30 s,
D. Block in row D – 1 h,
E. Phage binding in row E – 1 h,
F. Wash beads in row F – 1 min,
G. Wash beads in row G – 1 min,
H. Wash beads in row H – 1 min.

I. Transfer beads to elution strip from row H

At the end of the program, the protein coated beads with phage bound were left in Row H. The content of each well from row H was transferred to individual 1.7mL tube, and process for next round panning described in Section A.5.1.3 and for Illumina deep sequencing described in A.5.1.7.

A.5.1.6. Polymerase chain reaction (PCR)

Prior to PCR protocol, 25 μL of beads solution was buffer exchanged with nuclease free water. The mixture was boiled at 95 °C for 10 mins and the supernatant was used as a template for PCR with a total volume of 50 μL. A Typical 50 μL reaction mixture contained:

1. 5x Phusion buffer 10 μL
2. 10 mM dNTPs 1 μL
3. Phusion® High-Fidelity DNA Polymerase (NEB, cat#M0530S) 0.5 μL
4. Forward primer (3’-CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTXXXXCCTTTCTATTCTCACTCT-5’, 10 μM) 2.5 μL
5. Reverse primer (3’-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTXXXXACAGTTTCGGCCGA-5’, 10 μM) 2.5 μL
6. Template solution 25 μL
7. Nuclease free water 8.5 μL

Thermocycling was performed using the following steps:

a) 95 °C for 30 sec
b) 95 °C for 30 sec
c) 60.5 °C for 15 sec
d) 72 °C for 30 sec
e) Repeat step b) to d) 25 times
f) 72 °C for 5 min

g) Hold at 4°C

A.5.1.7. Next-generation sequencing (NGS)

45 µL of the PCR product was submitted for deep sequencing. The excess PCR primer was removed using a PureLink™ PCR Purification Kit from Thermo Fisher (cat# K310001). The concentration of the PCR product was quantified with an Advanced Analytical Fragment Analyzer. The PCR samples were pooled and submitted to NextSeq500 for deep sequencing.

A.5.2. Automated flow synthesis of MDM2$^{25-109}$

Sequence:

ETLVRPKPLL LKLLKSVGAQ KDTYTMKEVL FYLGQYIMTK RLYDEKQQHI VYCSNDLLGD
LFGVPSFSVK EHRKIYTMIY RNLVVK(Biotin)

The sequence was synthesized on 100 mg of pre-swollen LL ChemMatrix Rink Amide resin (0.17 mmol/g) using the published optimized protocol for automated flow peptide synthesis (AFPS) as described previously. Before synthesis, the resin was washed with DMF and coupled with Fmoc-Lys(Biotin)-OH (0.17 mmol, 10 equivalents) dissolved into 425 µL of PyAOP (0.38 M solution in DMF, 9.5 equivalents) with 89 µL of DIEA for activation (30 equivalents). The resin was stirred periodically over a coupling period of 3 hours, then washed with DMF (3 x 5 mL) before deprotection with 20% (v/v) piperidine in DMF (2 x 10 mL with 5 min each time). Briefly, utilizing an automated synthesizer, amine-free DMF washed the resin before coupling, after coupling, and after deprotection (40 strokes, ~25 mL). Coupling was performed with HATU (single-coupling, 8 strokes, ~5 mL) except S&A with HATU (double-coupling, 21 strokes, ~10 mL) and C, H, N, Q, R, V, T with PyAOP (double-coupling, 21 strokes, ~10 mL). Deprotection was completed with 20% piperidine in amine-free DMF with 2% formic acid (2 pumps, 40 mL/min). Amino acids were iteratively coupled and deprotected until finished. The resin was washed again with DMF (3 x 5 mL) and DCM (3 x 5 mL) then dried under reduced pressure. Reagent K solution (82.5% TFA, 5% water, 5% phenol, 5% thioanisole, 2.5% EDT) was used for global cleavage for 4 h at RT (15 mL + 5 mL washes), triturated with cold diethyl ether (3 x 45 mL), suspended in 50% AcN in Water (0.1% TFA), and lyophilized.

The lyophilized crude sample of MDM2$^{25\text{-}109}$ was weighed in batches of 25 mg, dissolved in 10 mL 6 M guanidinium chloride, 0.1 M dithiothreitol (DTT), in 50 mM sodium phosphate pH 7.5, vortexed briefly, 0.2 µm filtered, and subjected to RP-HPLC purification using a Agilent Zorbax 300SB-C18 PrepHT (9.4 × 250 mm, 5 µm) heated at 50 °C at 4.0 mL/min with the following gradient: isocratic 5% B from 0–5 min; linear gradient from 15–55% B from 5–65 min; linear gradient from 55-90% B from 65-66 min; isocratic 90% B from 66-71 min; isocratic 5% B from 71–76 min. Fractions showing high purity charge state series were combined and lyophilized, leading to the purification of 75 mg of crude to isolate 8.4 mg of HPLC-purified MDM2$^{25\text{-}109}$ (11% yield).

Purified MDM2$^{25\text{-}109}$ (1 mg, 96 nmol) was dissolved in phosphate buffered saline (PBS) containing 6 M Guanidine hydrochloride (vol) and 20 mM DTT at pH 7.2. MDM2$^{25\text{-}109}$ concentration was determined by UV^{280} and adjusted to 150 µM (extinction coefficient of MDM2$^{[25\text{-}109]}$: 10430 m-1 cm-1). The resulting solution was diluted six-fold using a folding buffer containing PBS and 20 mM DTT at pH 7.2 to a final MDM2$^{25\text{-}109}$ concentration of 25 µM. The solution was kept at room temperature for 1 hour before use in bio-panning or storage. Protein was concentrated using a 3 kDa molecular weight cutoff Amicon Ultra-15 centrifugal filter unit (Millipore Sigma) to isolate 0.43 mg of folded MDM2$^{25\text{-}109}$ (43% yield; 5% overall yield) as determined by UV^{280}.

A.5.2.1. Liquid chromatography-mass spectrometry (LC-MS) analysis

LC-MS chromatograms and associated high resolution mass spectra were acquired using an Agilent 1290 Infinity HPLC coupled to an Agilent 6545 LC/Q-TOF mass spectrometer using a Zorbax 300SB-C3 column (2.1 x 150 mm, 5 µm) heated to 40 °C. Solvent compositions used are 0.1% formic acid in H_2O (solvent A) and 0.1% formic acid in acetonitrile (solvent B). The following methods were used:

1. Gradient: isocratic 5% B from 0-1 min; linear gradient 5-65% B from 1-7 min; isocratic 91% B from 7-8 min; post time 5% B for 1 min. Flow rate: 0.5 mL/min. MS data was collected from 2-7 min; MS was run in positive ionization mode, extended dynamic range (2 GHz), and standard mass range (m/z in the range of 300 to 3000 a.m.u.).
2. Gradient: isocratic 5% B from 0-4 min; linear gradient 1-91% B from 4-11 min; isocratic 91% B from 11-12 min; post time 1% B for 3 min. Flow rate: 0.5 mL/min. MS data was collected from 2-7 min; MS was run in positive ionization mode, extended dynamic range (2 GHz), and standard mass range (m/z in the range of 300 to 3000 a.m.u.).

A.5.2.2. Analytical high-performance liquid chromatography (HPLC)

Analytical HPLC analysis was performed using an Agilent 1200 series system with UV detection at 214 nm on a Phenomenex Kinetex C18 LC column (2.1 x 100 mm, 2.6 µm, 100 Å). Solvent compositions used are 0.1% trifluoroacetic acid in H_2O (solvent A) and 0.08% trifluoroacetic acid in acetonitrile (solvent B). Gradient: isocratic 5% B from 0-3 min; linear gradient 5-65% B from 3-18 min; isocratic 65% B from 18-20 min; linear gradient 65-5% B from 20-21 min; isocratic 5% B from 21-26 min. Flow rate: 0.375 mL/min. Purity was determined through manual integration of all signals from 3 to 20 min.

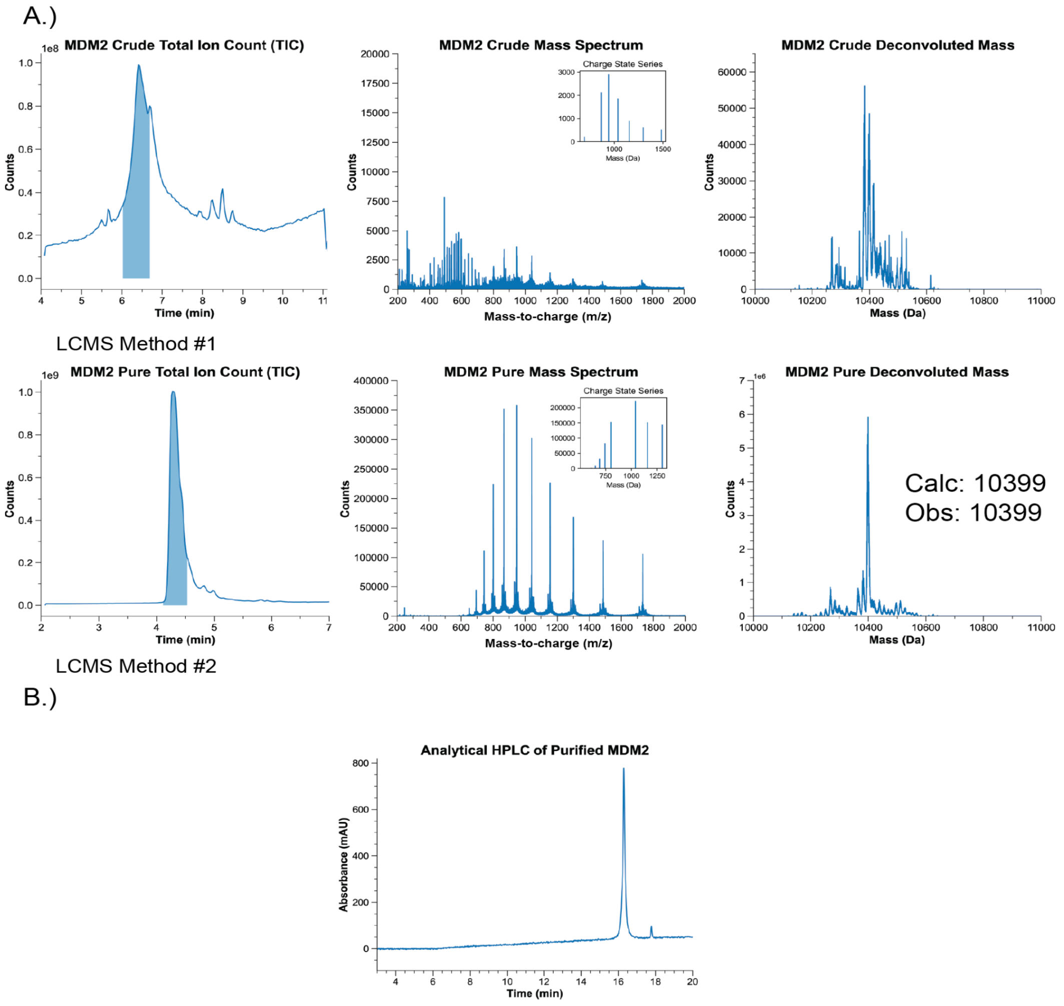


Figure A.8: *Analytical characterization of crude and preparative purified MDM2. (A) LC-MS characterization of crude (top) and reverse phase purified MDM2 (bottom). (B) Analytical reverse phase chromatogram demonstrating purity of MDM2.*

A.5.3. K-means clustering of raw data

Raw next-generation sequencing data was analyzed using the *K*-means clustering algorithm from the scikit-learn package in Python. Peptide sequences were filtered based on p-value ($p < 0.01$) and encoded using a 36 dimensional aggregate of one hot encodings (20 dimensional), relative propensity for binding score (1 dimensional), DELPHI predicted protein interaction score (1 dimensional), and physicochemical descriptors (14 dimensional). The data was then subjected to dimensionality reduction to prevent inflation of inertia via the uniform manifold approximation and projection, and *K*-means clustering was subsequently performed. The value of *k* was optimized per protein and encoding method using the elbow method. The clustering was run for a maximum of five hundred iterations, with the results reported being the output of the best fifty consecutive runs by inertia. Clusters of sequences were then visualized using logo plot analysis to reveal minimal significant motif analysis for 12ca5 and none for MDM2. Out of the 15 optimized clusters, one cluster was found to have a significant average fold change for each protein (Figure A.9B). Logo plot analysis of these clusters revealed only linear motif containing peptides for 12ca5 and no discernible motif for MDM2, meaning that almost all information from the macrocyclic libraries would be lost through an unsupervised analysis with *k* means clustering, as shown in the large amount of motif containing peptides outside of the two identified clusters (Figure A.9A)

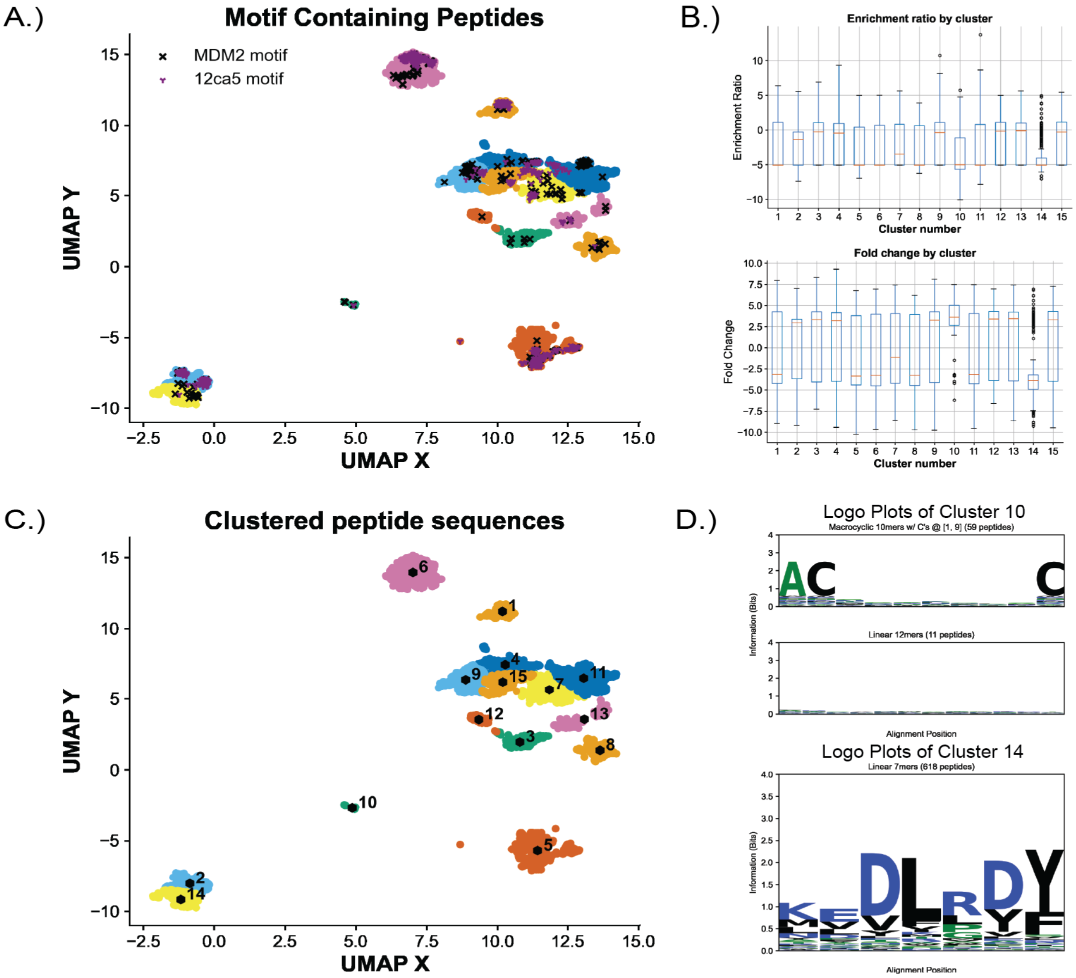


Figure A.9: *K-means clustering of raw NGS data from 12ca5 and MDM2 multiplex phage display panning. Both A and C are of the same clustering result using Uniform Manifold Approximation (UMAP) based clustering with K number of clusters. (A) Clustering result with the location of motif-containing peptides marked, showing no cluster cleanly isolates the motif-containing peptides. (B) Experimental enrichment ratio and fold change for each of the K clusters, showing that only Cluster 10 and 14 with differed from the average values across all clusters. (C) Cluster labeled with a center point. (D) Logo plot of Cluster 10 showed no clear sequence information beyond the library design of ACX$_7$C. Cluster 14 showed what appears to be part of the 12ca5-binding motif (D**DY(A/S)), though it had low experimental FC and low ER. No MDM2 containing motifs were observed.*

A.5.4. BiLSTM Hyperparameter Optimization

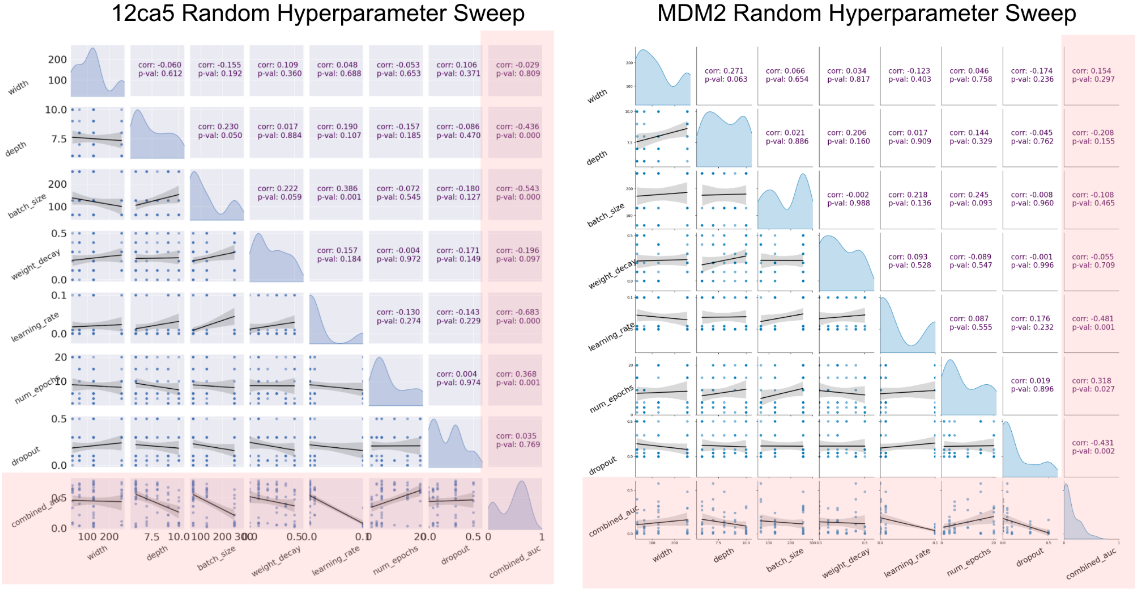


Figure A.10: *Parity plots of random hyperparameter sweeps across one hundred 12ca5 runs and one hundred MDM2 runs. The lower triangle of each parity plot displays scatterplots of the pairwise distribution of hyperparameters. The diagonal of each parity plot displays the distribution of individual hyperparameters. The upper triangle reports Holm-Bonferroni corrected p-value and pairwise Pearson correlation of hyperparameters. Focusing on the red highlighted row and column, certain hyperparameter instantiations are significantly correlated with improved hit rate ranking. Specifically, large regression penalty term (λ_{L2} = 0.5), high dropout ($\lambda_{Dropout}$ = 0.3), low learning rate (α = 0.001), shallow network depth (depth = 6), and few learning epochs (n = 5). We claim these hyperparameter choices effectively serve to regularize our model favoring sparser weight sets.*

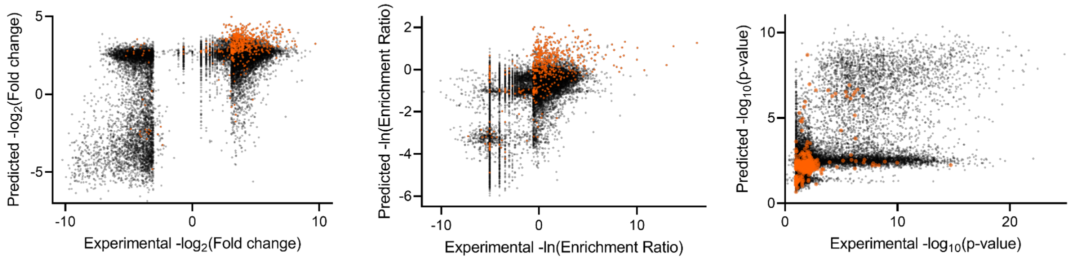


Figure A.11: *Parity plot of experimental versus predicted values including ER, as -ln(ER); FC, as -ln(FC) with its associated p-value as -log10(p-value). High-affinity motif-containing peptides are shown in orange.*

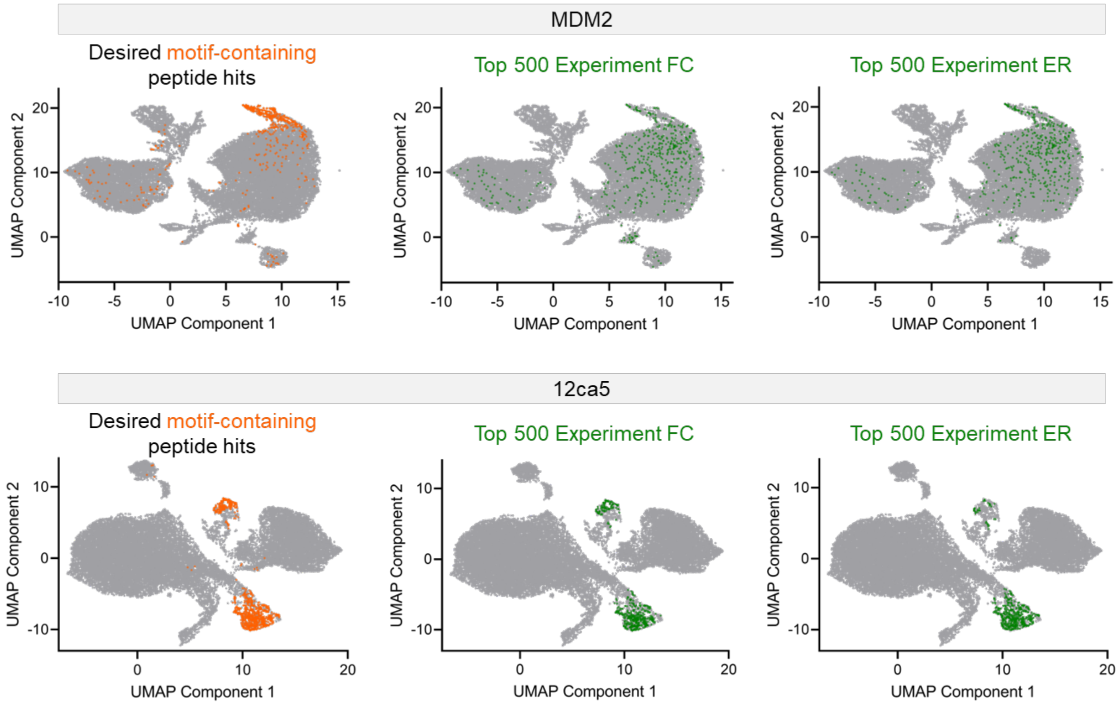


Figure A.12: *UMAP decompositions of the learned latent features for each peptide show that motif-containing sequences were somewhat clustered together by the BiLSTM proxy learning process. However, the peptides with the top experimental FC and ER (n = 500) were not clustered together for MDM2, indicating the that proxy learning process was not focused purely on accuracy. For 12ca5, the top experimental FC and ER (n = 500) were clustered together, but was used as a positive control.*

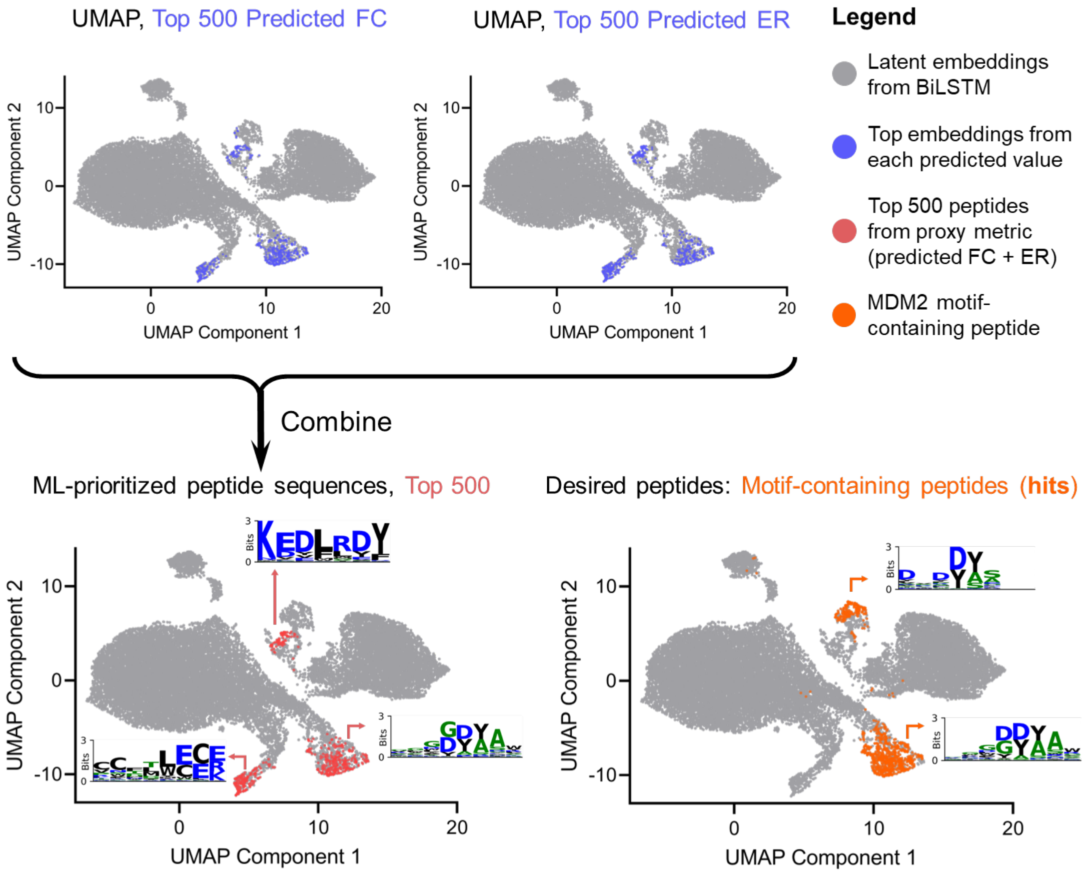
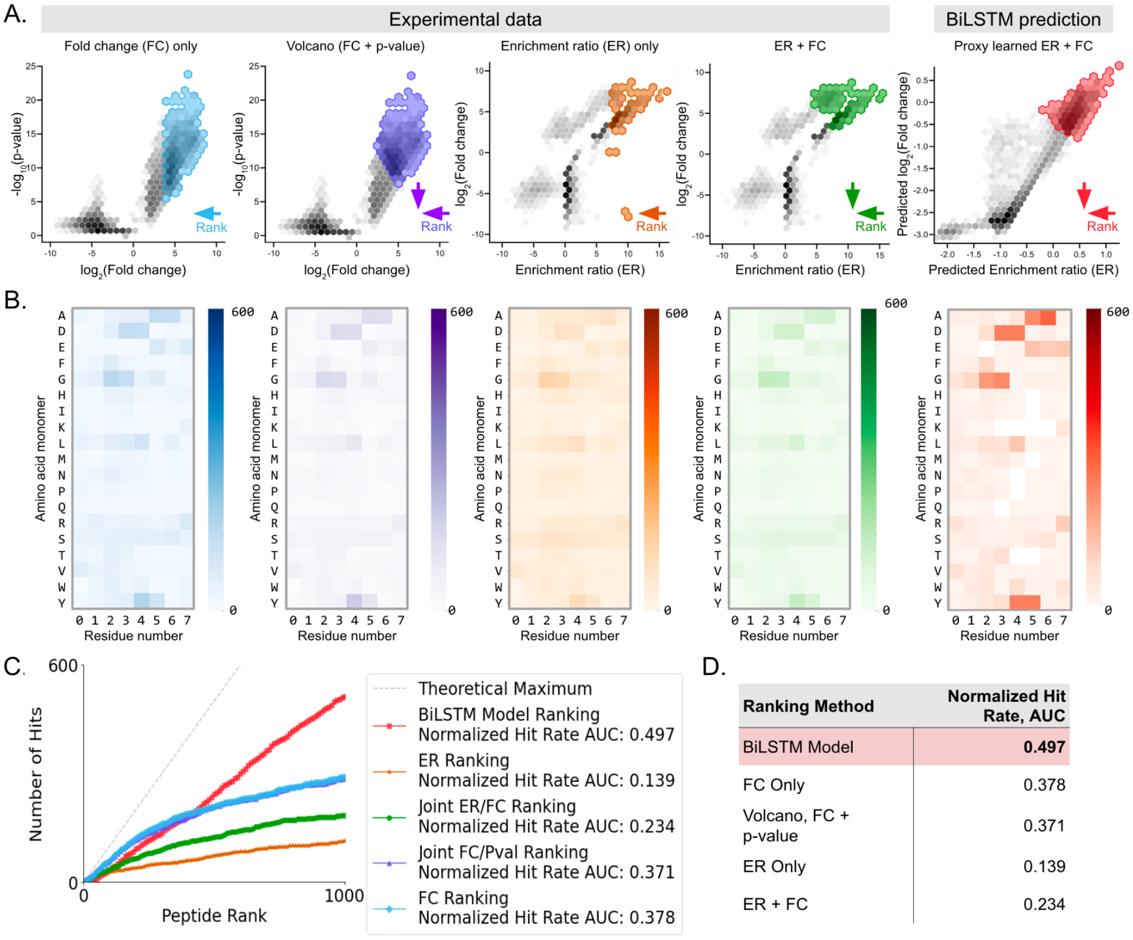


Figure A.13: *The UMAP projections of the learned latent features for each peptide from phage display with 12ca5 indicate successful proxy learning. UMAP transformation obtained from the penultimate layer of the BiLSTM model with the top 500 predicted FC and ER peptides highlighted. The smaller inlaid plots display blue highlights representing the top 500 peptides, ranked by predicted enrichment ratio and fold change. In the larger UMAP plot, a red overlay indicates the proxy metric combining predicted fold change and enrichment ratio. Additionally, the weblogo of the proxy metric learned space reveals a prominent 12ca5 binding motif (D**DY(A/S)).*

A.5.3. Hit rate prediction using 12ca5



Ranking Method	Normalized Hit Rate, AUC
BiLSTM Model	**0.497**
FC Only	0.378
Volcano, FC + p-value	0.371
ER Only	0.139
ER + FC	0.234

Figure A.14: *BiLSTM model highly ranks 12ca5 motif-containing peptide hits >30% better than any combination of experimental approaches. Ranking prioritizes the investment of synthesis and experimental binding validation toward peptides that have the highest predicted confidence to be hits. (A) Hexbin projections with highlighted zones corresponding to the top 500 peptides as determined by the different strategies to rank the peptides for their potential as peptide hits. Arrows shown in the bottom right display the direction of ranking (x-direction, y-direction, or both). (B) Positional frequency matrix of the top 500 identified peptides. The macrocyclic 9-mer library contained most of the motif-containing peptide hits, outperforming the other libraries. Thus, the positional frequency matrices of the top 500 show the 9-mer variable region of the 9-mer library (cysteines not shown). (C) A plot of the number of identified peptide hits versus their ranking shows that the BiLSTM model outperforms all other experimental methods to rank the peptides. The calculation of the normalized area under the curve reveals the BiLSTM model performs >30% better. (D) Calculation of the area under the hit rate curve in C indicates that 50% of the top 500 BiLSTM ranked peptides contain the 12ca5 motif.*

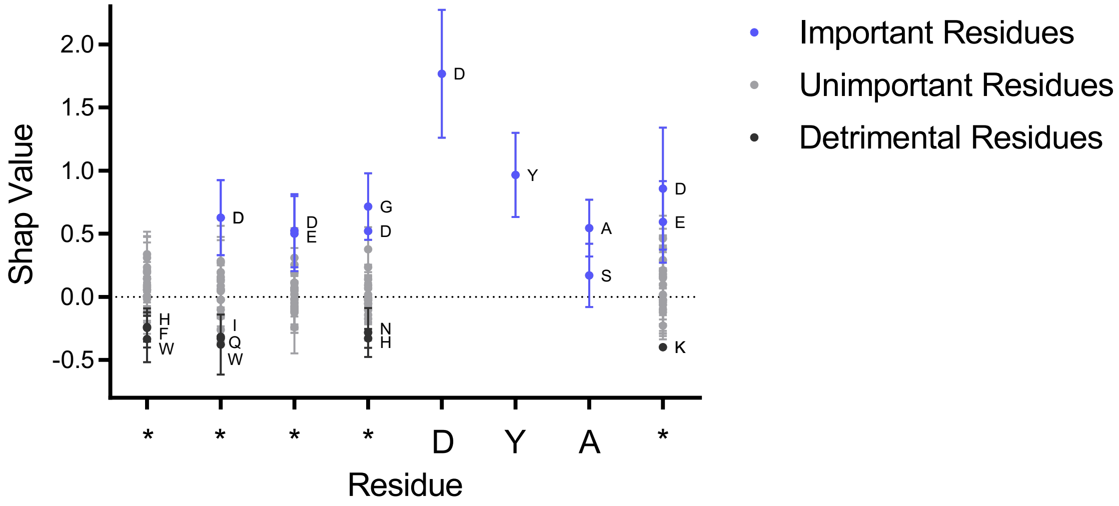


Figure A.15: *Positional Shapely feature importance across residue identities as calculated by the 10- model ensemble on the set of all 12ca5 motifs containing hits within the dataset. Sequences are aligned by motif position, and error bars are calculated according to the standard deviation of Shapely values per residue across all peptides and all models. This result underlies the importance of aspartic acid to drive binding and the potential for hydrophobic or positively-charged acids to disrupt peptide binding to 12ca5. From the 12ca5 motif (D**DY(A/S)1–3), the "DYA" was used for alignment because of concern the macrocyclic peptide structure could off-shift the first aspartic acid, which is seen to be D**DYA and D*DYA here. Also, the *DYAD* motif has been seen in other contexts.*

A.6. Acknowledgements

Funding for this work was provided by Novo Nordisk A/S. We thank Dr. Thomas E. Nielsen and Dr. Uli Stilz for their helpful discussions in support of our work. J.S.B. acknowledges support from Pharmaceutical Research and Manufacturers of America (PhRMA) Foundation through the Postdoctoral Fellowship in Drug Discovery. M.A.L. acknowledges support from the MIT Dean of Science Fellowship.

A.7. References

1. Smith, G. P. & Petrenko, V. A. Phage display. *Chem Rev* **97**, 391–410 (1997).
2. Pasqualini, R. & Ruoslahti, E. Organ targeting in vivo using phage display peptide libraries. *Nature* **380**, 364–366 (1996).
3. Philpott, D. N. *et al.* Rapid On-Cell Selection of High-Performance Human Antibodies. *ACS Cent Sci* **8**, 102–109 (2022).
4. Wong, J. Y. K. *et al.* Genetically-encoded discovery of proteolytically stable bicyclic inhibitors for morphogen NODAL. *Chem Sci* **12**, 9694–9703 (2021).
5. Ekanayake, A. I. *et al.* Genetically Encoded Fragment-Based Discovery from Phage-Displayed Macrocyclic Libraries with Genetically Encoded Unnatural Pharmacophores. *J Am Chem Soc* **143**, 5497–5507 (2021).
6. Oppewal, T. R., Jansen, I. D., Hekelaar, J. & Mayer, C. A Strategy to Select Macrocyclic Peptides Featuring Asymmetric Molecular Scaffolds as Cyclization Units by Phage Display. *J Am Chem Soc* **144**, 3644–3652 (2022).
7. Kong, X. D. *et al.* De novo development of proteolytically resistant therapeutic peptides for oral administration. *Nat Biomed Eng* **4**, 560–571 (2020).
8. Henninot, A., Collins, J. C. & Nuss, J. M. The Current State of Peptide Drug Discovery: Back to the Future? *J Med Chem* **61**, 1382–1414 (2018).
9. Muttenthaler, M., King, G. F., Adams, D. J. & Alewood, P. F. Trends in peptide drug discovery. *Nat Rev Drug Discov* **20**, 309–325 (2021).
10. Wells, J. A. & McClendon, C. L. Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. *Nature* **450**, 1001–1009 (2007).
11. Cunningham, A. D., Qvit, N. & Mochly-Rosen, D. Peptides and peptidomimetics as regulators of protein–protein interactions. *Curr Opin Struct Biol* **44**, 59–66 (2017).
12. Lubell, W. D. Peptide-Based Drug Development. *Biomedicines 2022, Vol. 10, Page 2037* **10**, 2037 (2022).
13. Heinis, C. Drug discovery: Tools and rules for macrocycles. *Nat Chem Biol* **10**, 696–698 (2014).
14. Vinogradov, A. A., Yin, Y. & Suga, H. Macrocyclic Peptides as Drug Candidates: Recent Progress and Remaining Challenges. *J Am Chem Soc* **141**, 4167–4181 (2019).
15. Giordanetto, F. & Kihlberg, J. Macrocyclic drugs and clinical candidates: What can medicinal chemists learn from their properties? *J Med Chem* **57**, 278–295 (2014).
16. Rigby, M. *et al.* BT8009; A Nectin-4 Targeting Bicycle Toxin Conjugate for Treatment of Solid Tumors. *Mol Cancer Ther* **21**, 1747–1756 (2022).
17. Rigby, M. *et al.* BT8009; A Nectin-4 Targeting Bicycle Toxin Conjugate for Treatment of Solid Tumors. *Mol Cancer Ther* **21**, 1747–1756 (2022).
18. Bendell, J. C. *et al.* BT5528-100 phase I/II study of the safety, pharmacokinetics, and preliminary clinical activity of BT5528 in patients with advanced malignancies associated with EphA2 expression. *https://doi.org/10.1200/JCO.2020.38.15_suppl.TPS3655* **38**, TPS3655–TPS3655 (2020).

19. McCloskey, K. *et al.* Machine learning on DNA-encoded libraries: A new paradigm for hit finding. *J Med Chem* **63**, 8857–8866 (2020).
20. Kómár, P. & Kalinić, M. Denoising DNA Encoded Library Screens with Sparse Learning. *ACS Comb Sci* **22**, 410–421 (2020).
21. 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 U S A* **98**, 3750–3755 (2001).
22. Gai, S. A. & Wittrup, K. D. Yeast surface display for protein engineering and characterization. *Curr Opin Struct Biol* **17**, 467–473 (2007).
23. Boder, E. T. & Wittrup, K. D. Yeast surface display for screening combinatorial polypeptide libraries. *Nature Biotechnology 1997 15:6* **15**, 553–557 (1997).
24. Thomas, W. D., Golomb, M. & Smith, G. P. Corruption of phage display libraries by target-unrelated clones: Diagnosis and countermeasures. *Anal Biochem* **407**, 237–240 (2010).
25. Matochko, W. L., Cory Li, S., Tang, S. K. Y. & Derda, R. Prospective identification of parasitic sequences in phage display screens. *Nucleic Acids Res* **42**, 1784–1798 (2014).
26. Ito, T. *et al.* Selection of target-binding proteins from the information of weakly enriched phage display libraries by deep sequencing and machine learning. *MAbs* **15**, 2168470 (2023).
27. Menendez, A. & Scott, J. K. The nature of target-unrelated peptides recovered in the screening of phage-displayed random peptide libraries with antibodies. *Anal Biochem* **336**, 145–157 (2005).
28. Tjhung, K. F. *et al.* Silent Encoding of Chemical Post-Translational Modifications in Phage-Displayed Libraries. *J Am Chem Soc* **138**, 32–35 (2016).
29. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
30. Coates, A. & Ng, A. Y. Learning feature representations with K-means. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **7700 LECTURE NO**, 561–580 (2012).
31. Saarela, M. & Jauhiainen, S. Comparison of feature importance measures as explanations for classification models. *SN Appl Sci* **3**, 1–12 (2021).
32. Schissel, C. K. *et al.* Deep learning to design nuclear-targeting abiotic miniproteins. *Nat Chem* **13**, 992–1000 (2021).
33. Torres, M. D. T., Melo, M. C. R., Crescenzi, O., Notomista, E. & de la Fuente-Nunez, C. Mining for encrypted peptide antibiotics in the human proteome. *Nature Biomedical Engineering 2021 6:1* **6**, 67–75 (2021).
34. Li, G., Iyer, B., Prasath, V. B. S., Ni, Y. & Salomonis, N. DeepImmuno: deep learning-empowered prediction and generation of immunogenic peptides for T-cell immunity. *Brief Bioinform* **22**, 1–10 (2021).
35. Saka, K. *et al.* Antibody design using LSTM based deep generative model from phage display library for affinity maturation. *Scientific Reports 2021 11:1* **11**, 1–13 (2021).

36. Mason, D. M. *et al.* Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nature Biomedical Engineering 2021 5:6* **5**, 600–612 (2021).
37. Quartararo, A. J. *et al.* Ultra-large chemical libraries for the discovery of high-affinity peptide binders. *Nat Commun* **11**, 3183 (2020).
38. Rini, J. M., Schulze-Gahmen, U. & Wilson, I. A. Structural evidence for induced fit as a mechanism for antibody-antigen recognition. *Science (1979)* **255**, 959–965 (1992).
39. Houghten, R. A. *et al.* Generation and use of synthetic peptide combinatorial libraries for basic research and drug discovery. *Nature 1991 354:6348* **354**, 84–86 (1991).
40. Chang, Y. S. *et al.* Stapled α−helical peptide drug development: A potent dual inhibitor of MDM2 and MDMX for p53-dependent cancer therapy. *Proceedings of the National Academy of Sciences* **110**, E3445–E3454 (2013).
41. Phan, J. *et al.* Structure-based design of high affinity peptides inhibiting the interaction of p53 with MDM2 and MDMX. *Journal of Biological Chemistry* **285**, 2174–2183 (2010).
42. Bernal, F., Tyler, A. F., Korsmeyer, S. J., Walensky, L. D. & Verdine, G. L. Reactivation of the p53 tumor suppressor pathway by a stapled p53 peptide. *J Am Chem Soc* **129**, 2456–2457 (2007).
43. Zondlo, S. C., Lee, A. E. & Zondlo, N. J. Determinants of specificity of MDM2 for the activation domains of p53 and p65: Proline27 disrupts the MDM2-binding motif of p53. *Biochemistry* **45**, 11945–11957 (2006).
44. Ye, X. *et al.* Binary combinatorial scanning reveals potent poly-alanine-substituted inhibitors of protein-protein interactions. *Communications Chemistry 2022 5:1* **5**, 1–10 (2022).
45. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
46. Chen, K.-H., Hu, Y.-J., Chen, K.-H. ; & Hu, Y.-J. Residue–Residue Interaction Prediction via Stacked Meta-Learning. *International Journal of Molecular Sciences 2021, Vol. 22, Page 6393* **22**, 6393 (2021).
47. Li, Y., Brian Golding, G. & Ilie, L. DELPHI: accurate deep ensemble model for protein interaction sites prediction. *Bioinformatics* **37**, 896–904 (2021).
48. Zamyatnin, A. A. & Anokhin, P. K. Amino Acid, Peptide, and Protein Volume in Solution. *https://doi.org/10.1146/annurev.bb.13.060184.001045* **13**, 145–165 (2003).
49. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv* (2018) doi:10.48550/arxiv.1802.03426.
50. Huang, Z., Research, B., Xu, W. & Baidu, K. Y. Bidirectional LSTM-CRF Models for Sequence Tagging. (2015).
51. Bergstra, J., Bardenet, R., Bengio, Y. & Kégl, B. Algorithms for Hyper-Parameter Optimization. *Adv Neural Inf Process Syst* **24**, (2011).

52. Luxburg, U. von & Schölkopf, B. Statistical Learning Theory: Models, Concepts, and Results. *Handbook of the History of Logic* **10**, 651–706 (2011).
53. Kusumoto, Y. *et al.* Highly Potent and Oral Macrocyclic Peptides as a HIV-1 Protease Inhibitor: mRNA Display-Derived Hit-to-Lead Optimization. *ACS Med Chem Lett* (2022) doi:10.1021/ACSMEDCHEMLETT.2C00310.
54. Iskandar, S. E. & Bowers, A. A. mRNA Display Reaches for the Clinic with New PCSK9 Inhibitor. *ACS Med Chem Lett* (2022) doi:10.1021/ACSMEDCHEMLETT.2C00319.
55. Lau, J. L. & Dunn, M. K. Therapeutic peptides: Historical perspectives, current development trends, and future directions. *Bioorg Med Chem* **26**, 2700–2707 (2018).
56. Rogers, J. M., Passioura, T. & Suga, H. Nonproteinogenic deep mutational scanning of linear and cyclic peptides. *Proc Natl Acad Sci U S A* **115**, 10959–10964 (2018).
57. Loh, W. Y. Fifty Years of Classification and Regression Trees. *International Statistical Review* **82**, 329–348 (2014).
58. Lundberg, S. M., Allen, P. G. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Adv Neural Inf Process Syst* **30**, (2017).
59. Dickinson, Q. & Meyer, J. G. Positional SHAP (PoSHAP) for Interpretation of machine learning models trained from biological sequences. *PLoS Comput Biol* **18**, e1009736 (2022).