Multi-Parameter Optimization of Stapled Peptides and Other Proteins Via Directed Evolution

by

Marshall Case

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctorate of Philosophy
(Chemical Engineering)
in the University of Michigan
2023

Doctoral Committee:

Professor Greg Thurber, Chair
Professor Neil Marsh
Professor Peter Tessier
Professor Fei Wen

Marshall Case

marcase@umich.edu

ORCID iD: 0000-0002-1504-4802

Dedication

This dissertation is dedicated to my grandmother and late grandfather, Lorraine and Robert Salmer. Your generosity and support have made me who I am today and I am so grateful for all you've provided for our family. Love you always.

Acknowledgements

First, I'd like to thank my friends I've made here at Michigan who supported me from highs to lows along the way. I've been lucky to have some of the greatest friends by my side throughout: Alex, Zach, Anna, Sean, Rachel, Patrick, Emily, Mackenzie, Harrison, Mische, Ray, Jacques, Jordyn, and many others. We've shared so many great memories and these have truly been the highlight of all my years here. I'm going to miss regular hangouts at Depot and discussing the edge cases of rules under the thin guise of the "Elder Council". Very importantly, I want to acknowledge the fuzzy friends who provided support along the way: Apple, Relo, Emerson, and Berkeley.

Of course, I would not be writing this document without all the support from my family. My parents, Vicki and Steve, couldn't have set me up in a better place to focus on all the challenges getting a PhD entails. My roommate (and also, grandmother), Lorraine "Lala", who was nice enough to let me stay in her house and helped me become less of a mess – and has provided countless support and love throughout all my years. My sister, Rachel, for always being around for phone calls and support and making me laugh. And finally, my partner Ellen, who has been there for me at all moments: I am so excited for our future together. I can't wait for our next chapter in Boston where there will finally be enough Dunkin' stores and many new memories to share.

I'd also like to thank my friends and colleagues from Rackham Student Government. Serving on RSG was a tremendous opportunity and one of my favorite parts of my doctoral experience at Michigan. I'd specifically like to thank Lucca, who consistently supported me

grow as an advocate and leader. RSG could not have operated smoothly without those who served along with me: Claire, Olivia, Ashley, Brittany, Raz, Veronica, Aditya, Jason, Nick, Thomas, Austin, Jourdan, and many others. I am deeply appreciative of your service and I hope we cross paths again.

I'd next like to thank those who helped me along the academic journey. My advisor, Greg, who made me feel welcome from the beginning and always encouraged exploring new ideas (assuming of course that I could reasonably defend their scientific value – one of the key lessons learned as a PhD student). Next, I'd like to thank the many members of the Thurber Lab and collaborators who helped me along the way. I couldn't have done it without the guidance and support from Tejas and Lydia, who initially guided my project and helped steer me in the right direction. I had the pleasure of working with several undergraduate students, who not only helped me gain experience as a mentor and teacher, but frequently injected humor into otherwise boring situations: Jordan, Sophie, and Rahul. There were many friends and colleagues who served as council for scientific writing and putting ideas to papers: Matt, Emily, Patrick, Alex, Camille, and Jorge. There are countless individuals at the University who trained me and supported my scientific experiments: Luke, Andrea, Mark, Henriette, Mukesh, Vivek, Priya, Nicole, among many others. I'm also extremely grateful for the Cores that the University operates that has facilitated many components of my research: Advanced Genomics, Bio Nuclear Magnetic Resonance Spectroscopy, Mass Spectrometry, Proteomics and Peptide Synthesis, Flow Cytometry, CyTOF, and the Center for Structural Biology.

Table of Contents

List of Tables

List of Figures

Abstract

There is a wealth of disease-related proteins that are 'undruggable' by common therapeutic modalities, owing to their difficult location within cells and lacking specific structural motifs that facilitate specific targeting. Stapled peptides, a class of therapeutic that leverages synthetic biology and protein engineering, are a promising approach to overcome these barriers, but their development is rendered difficult by complex chemical synthesis and myriad design factors. In this thesis, Stabilized Peptide Engineering by *E. coli* Display (SPEED), a technique that can greatly accelerate stapled peptide development, is used to explore new design criteria for stapled peptides, such as staple location, amino acid hot spots, and stapling chemistry. In this technique, methionine auxotrophic bacteria are transformed with DNA that encodes for a peptide and azide-containing non-natural amino acids are incorporated. Then, copper catalyzed click chemistry (CuAAC) is performed on the cell surface before treatment with any combination of fluorescently- or magnetically- activated proteins for subsequent property measurement or cell sorting application.

To demonstrate how SPEED coupled with an expanded set of design criteria can yield therapeutic leads towards these challenging targets, high affinity and specific stapled peptides are developed towards two important targets: p53 and Bcl-2. The thesis describes the generation of highly focused protein variant libraries for multi-objective optimization. In **Chapter 2**, SPEED was used to confirm the hot spot analysis of p53-MDM2 with reduced affinity resulting from mutations to F19, W23, and L26. Likewise, it was used to show the importance of staple chemistry and location on binding affinity and specificity. With BIM peptides, for example, a

staple location at p1 and p5 showed significant preference for Mcl-1 binding, while p7 and p14 bound Bfl-1, Bcl-xL, Bcl-w, and Bcl-2 more than Mcl-1. This analysis establishes that both staple sequence and staple location are key determinants of peptide binding affinity and specificity. Then, in **Chapter 3**, I engineered stapled peptides targeting Bcl-xL (a protein in the B cell lymphoma 2 protein family that promotes cancer cell survival) with high specificity. Using an enriched library design and directed evolution campaign sorting for Bcl-xL specificity, I engineered Bcl-xL binding peptides with 10 nM affinity and 100-fold specificity with novel mutations that act in accordance with apoptosis biochemistry. Finally, in **Chapter 4**, I describe a machine learning approach that captures hidden information from simple binary sorting experiments by using next generation sequencing datasets. The trained model is able to predict fitness in unseen sequence space to expand discovery beyond experimentally measured sequences. To validate this method, I curate five protein directed evolution campaigns via cell surface display and find that across many protein families (single chain variable fragments, fragment antigen binding, globular proteins, among others) and objectives (fluorescence, specificity, binding affinity), this method consistently predicts continuous properties and identifies high functioning variants. We then prospectively design stapled peptides to identify high functioning Bcl-2 binders using sequence optimization when experimental techniques fail to yield consistent hits. Overall, this work presents novel peptide engineering strategies for stapled peptides, next-generation sequence analysis for selecting specific binders against homologous proteins, and machine learning methods to extract data and design novel peptides and proteins beyond experimentally measures space, which should find use in many protein engineering campaigns.

Chapter 1 Introduction

Proteins are a diverse class of biomolecules capable of catalyzing chemical reactions, binding diverse biomolecules, forming complex systems with emergent behavior, giving structure to cells, and much more. They are also responsible for many aspects of modern life, from the enzymes in laundry detergent to monoclonal antibodies used for treating cancer and immune disease. It is therefore unsurprising that a longstanding goal of biochemistry is to map the sequence of a protein to its structure and function.[1] However, the complex biophysics that govern the protein fitness landscape, including how a protein folds and how its structure influences function make the coupling of sequence to function an extremely difficult task. Protein engineers thus often focus on a much smaller subdomain of the protein fitness landscape, using the confined resources of experimental protein science to explore variants close to a known functional protein with the goal of incrementally improving function. In this doctoral thesis, new methods towards the design of proteins and peptides are described.

1.1 Challenging targets and the motivation for development of novel therapeutic modalities

There are a wealth of proteins involved in disease that cannot be targeted by current therapeutics **(Figure 1.1)**.[2] The two modalities of therapeutics that dominate the market are small molecule drugs and protein biologics. Small molecule drugs are characterized by their rapid distribution across biological barriers (such as cell membranes) and subsequent rapid entry into cells. However, their small size means that they are largely limited to targeting disease related proteins with a small hydrophobic binding pocket. Conversely, protein biologics, which are most commonly an immunoglobulin (IgG) or antibody, are characterized by their large size and ability

to interact with proteins of all shapes and sizes. However, their large size inhibits transport across multiple scales of biological barriers. One of the biggest limitations of protein biologics is their inability to penetrate cell membranes to access the intracellular compartment, where many disease related proteins are located. It is estimated that 66-90% of all disease related proteins are both inside cells and lack hydrophobic binding pockets. The inability of common therapeutic modalities to reach such a large class of drug targets highly motivates the further development of new modalities that are able to navigate these targeting and transport barriers.

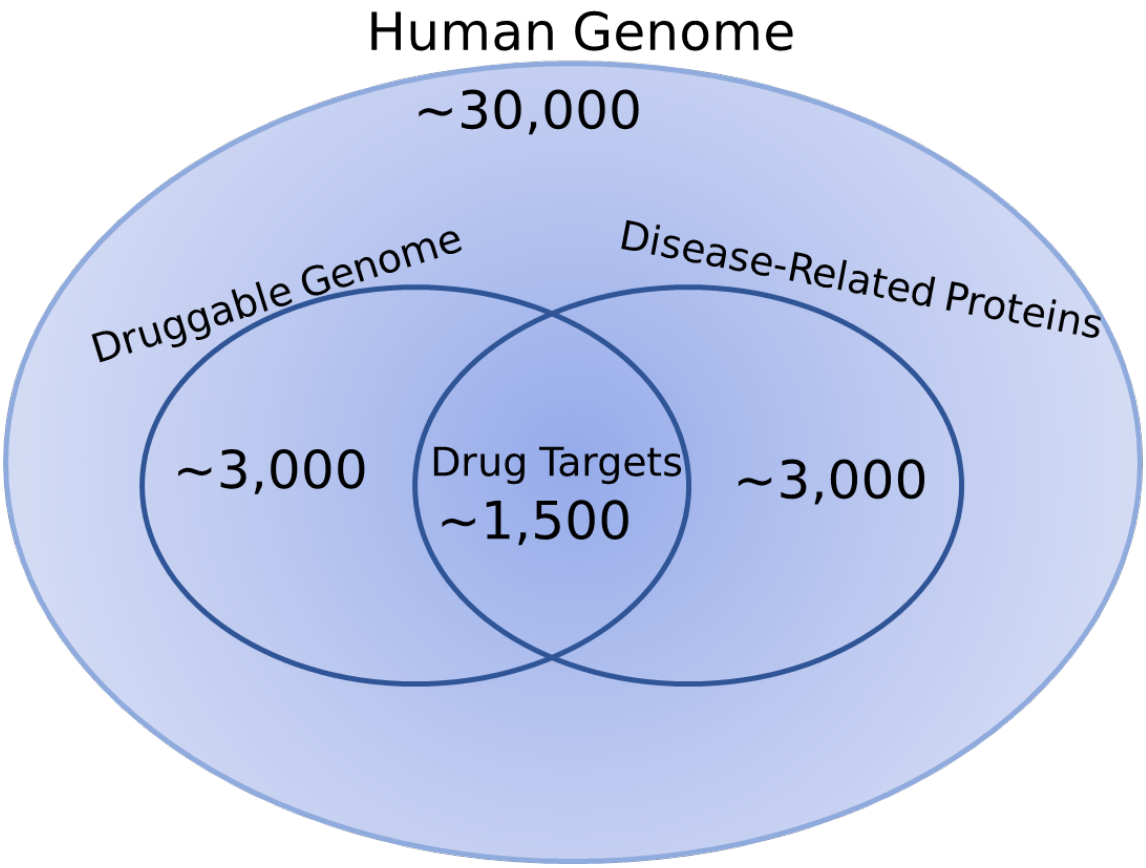


Figure 1.1: The space of druggable proteins is far smaller than the number of disease related proteins, which is itself far smaller than the number of proteins that are encoded by the human genome. The key limitation in drugging these proteins is that they are inaccessible to the most common modalities of therapeutics: small molecule drugs and biologics. Developing new modalities of drugs that can target the remaining disease related proteins may yield new treatments for diseases. Figure adapted from ref [3].

Peptide therapeutics attempt to overcome these barriers as their intermediate size allows entry into cells and permits the targeting of suitably large binding interfaces to inhibit most protein-

protein interactions.[4] Because many protein-protein interactions are mediated through alpha helical secondary structural motifs, and linear peptides dynamically conform to form alpha helical structures, peptide therapeutics are an obvious choice for the selective targeting and antagonism of protein-protein interfaces.[5] It is therefore favorable to have peptides that preferentially exist in their folded state. However, linear peptides rapidly exchange between their unfolded, non-helical state and their folded, helical state. This equilibrium leads to peptide therapeutics suffering in the clinic from their low target affinity, low proteolytic stability, and slow uptake into cells.[6] Because each of these limitations partially arises from a peptide's secondary structure, improvements to the alpha helicity and minimization of unfolded peptide states could evade proteases, improve binding affinity, and reduce unfavorable interactions between the hydrophilic peptide backbone and the hydrophobic cell membrane.

To improve the alpha helicity, researchers engineered 'stapled peptides', which covalently cross link two amino acids' side chains, locking the peptide into a specific alpha-helical conformation.[7] This single modification can improve target affinity, facilitate cell entry, and enhance proteolytic stability. The stapled peptide field has had modest success developing therapeutics towards HIV, cancer, and other diseases.[8,9,10] As outlined above, stapled peptides overcome many of the challenges associated with traditional peptide therapeutics. Primarily, the staple forces the peptide into a conformation with more secondary structure, decreasing the entropic penalty of binding and increasing affinity.[5] Secondly, since most proteases encountered by a peptide *in vivo* recognize linear epitopes, a peptide in a helical conformation tends to have a longer half-life.[8] Finally, the enhanced alpha helicity means that intramolecular hydrogen bonds shield the hydrophilic peptide backbone from the hydrophobic cell membrane, which has been shown to enhance cell permeability.[11]

These findings have motivated researchers to generate numerous chemistries for forming stapled peptides (**Figure 1.2**).[12–17] The choice of linker is a very important one as the linker influences many properties of the peptide, such as binding affinity, protease stability, cell accessibility, and more.[5] However, most previous research uses non-natural amino acid chemistry necessitating solid-phase peptide synthesis, which is costly, time-consuming, and limits the number of variants that can be screened. These restrictions limit exploration of design space and are typically limited to less than 100 different sequences.[18] Techniques that accelerate the design of such chemically modified peptides are highly needed.

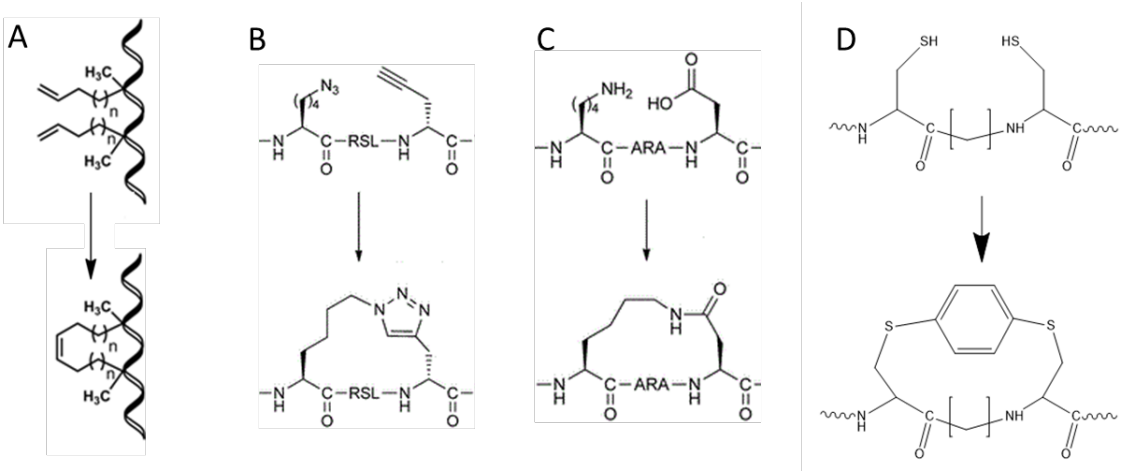


Figure 1.2: There are many strategies for the formation of stapled peptides. Some examples include hydrocarbon stapling through ruthenium catalysis **(A)**, triazole copper catalyzed click chemistry **(B)**, lactam bridge formation **(C)**, and thiol chemistry **(D)**.

1.1.1 Accelerated development of stapled peptides via bacterial surface display

To overcome the barrier of low-throughput protein engineering, scientists can take advantage of natural protein transcription and translation machinery to direct cells to make different protein variants and assay their function on an individual cell basis. Directed evolution is a powerful approach where researchers select protein variants according to their biological fitness analogously to how nature selects organisms that have higher fitness. This technology has been

widely applied to engineer proteins since its inception and has since been expanded to include multimeric proteins, proteins with non-natural amino acids, and even post translationally modified proteins. In brief, this technique works by designing a library of protein variants encoded by DNA, which are transcribed and translated into proteins, before selecting for the highest functioning variants. While there are many variations of how to perform directed evolution for proteins, cell surface display, where a protein variant is expressed as membrane fusion proteins and can easily be accessed in the extracellular compartment, has ushered its widespread adoption.[19–21] Depending on the category of protein, protein fitness via cell surface display is measured by intrinsic fluorescence (in the case of green fluorescent protein (GFP))[22], binding a fluorescently tagged protein (in the case of affinity maturation)[23], or a combination of fluorescently tagged proteins (when multiple binding events need to be measured, or two signals need to be compared).[24] Cells are sorted on the basis of the fluorescence (through incubation with a fluorescently tagged proteins) using fluorescent activated cell sorting (FACS) and the highest fitness variants are captured. Traditionally, several rounds of FACS are performed in series and the variants that emerge are assayed for their function in lower throughput experimental assays.

The design of peptides is well suited for directed evolution and cell surface display: peptides are small proteins that are expressed well by common cell surface display technologies (yeast, bacteria, and phage alike), require no post translational modification for folding, and are easily engineered using molecular cloning protocols. As such, there are many examples of linear peptide engineering in high-throughput using cell surface display, which can screen up to 10^9 peptides for activity in a single sorting campaign.[21,25–38] However, linear peptides have serious limitations as therapeutics and the process of optimizing a stapled peptide from its linear counterpart is difficult and requires significant quantity of rational design.[9,39,40] One reason for this

translational challenge is that the discovery of linear peptides does not account for the conformational effect of staple chemistry or staple location while screening, which can significantly affect drug-like properties such as binding affinity, protease stability, or specificity. [24] Towards the development of stapled peptides from high-throughput directed evolution experiments, researchers needed ways of assaying stapled peptides that could be performed using cell surface display.

One important challenge with the formation of stapled peptides on the cell surface is identifying chemistries that are easily accessible to natural protein machinery and chemists alike. One strategy is to use cysteine, the only naturally occurring amino acid that has a thiol, because of its unique reactivity that can be harnessed to form covalent, intramolecular bonds. The intramolecular reaction between two cysteines and a compound containing two bromines has been proven to be one effective approach as it includes the conformational effect of stapling while screening via phage display (as shown in **Figure 1.2D)**. [13] This improved binding affinities but precluded the use of cysteines elsewhere. This is a notable disadvantage as the formation of disulfide bonds on alpha-helical peptides is known to improve stability and affinity.[23,21] Reliance on cysteine stapling also presents a liability as cysteine chemistry is not bio-orthogonal, meaning that other proteins expressed on the cell surface containing free cysteines may cross react, muddling the ability to assay the fitness of peptides on the cell surface. Other approaches that are biologically compatible, such as lactam bridge formation, additionally suffer from reactions with native cell proteins. To further complicate matters, these reports of stapled peptide engineering via thiol chemistry use phage for surface display. As phage are too small for use in fluorescent activated cell sorting (FACS), the most powerful selection tools are inaccessible and therefore phage displays suffer from weaker and less interpretable fitness selections. On the other hand,

yeast are large enough for selection via FACS and have sufficient valency of expressed protein that the signal-to-noise fluorescence ratio is high. Furthermore, this valency ensures that the effective binding as measured via FACS is averaged over a large number of individual protein-protein interactions, minimizing the stochasticity of experimental binding interactions. However, because display platforms for yeast rely on cysteine for display (Aga1 and Aga2 are tethered via disulfide bonds), there is enhanced likelihood for cysteine containing peptides to disfavorably cross react with the expression proteins.[20] Furthermore, the incorporation of non-natural amino acids on yeast through stop codon manipulation is not efficient enough currently for stapled peptide formation.[41]

To overcome these limitations of stapled peptide selections via cell surface display, the Thurber Lab has developed an approach, **S**tabilized **P**eptide **E**ngineering by ***E***. *coli* **D**isplay (SPEED), using bacterial cells and uses bio-orthogonal chemistry to ensure minimal off-target reactions **(Figure 1.3)**.[23] In this approach, azide moieties are expressed on the cell surface through incorporation of azidohomoalanine (AHA) residues in place of methionine. Then, azide residues are reacted with a bisalkyne using copper catalyzed click chemistry (CuAAC) before being assayed for their fitness with magnetically or fluorescently activated cell sorting. In the first implementation of this technique, a library of p53-like peptides was engineered towards MDM2, an important protein that regulates p53, a tumor-suppressing transcription factor commonly known as the 'guardian of the genome'. The peptide discovered, SPD-V6-M1 had an eight-fold higher affinity (1.7nM vs 15nM) than the starting sequence, SPD-M0-E(-2). Additionally, this lead molecule had a unique disulfide bond in addition to its click chemistry

staple that contributed to its affinity and protease stability.

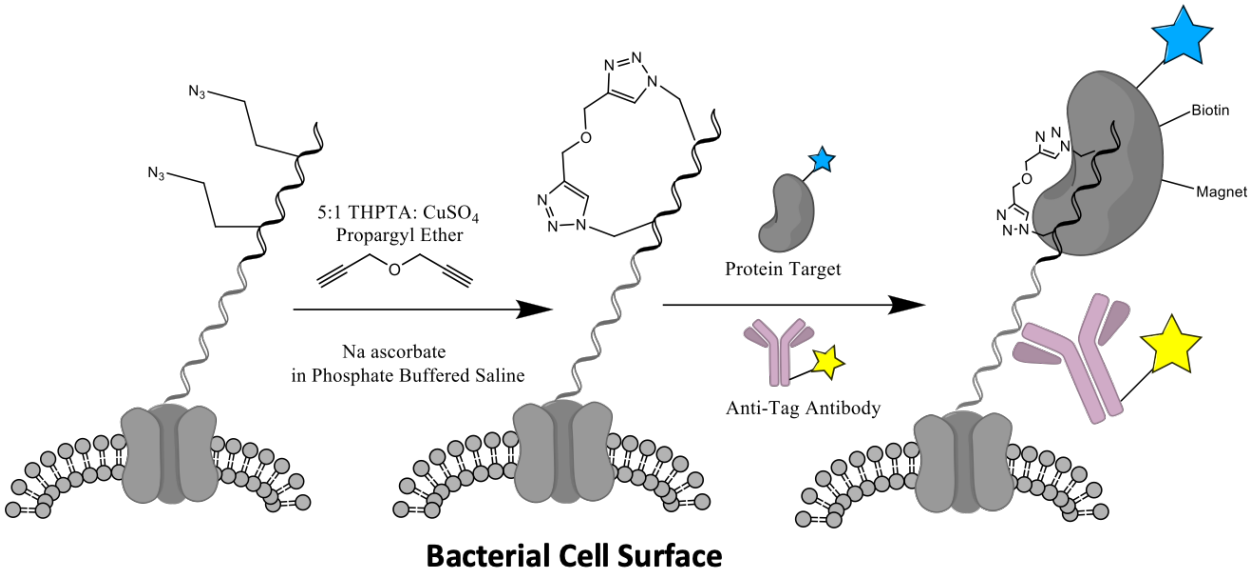


Figure 1.3: Stabilized Peptide Engineering by *E. coli* Display (SPEED). First, DNA encoding a peptide of interest is transformed into methionine auxotrophic *E. coli* bacteria. Then, the peptide is expressed, where azidohomoalanine (a methionine surrogate that contains an azide) is expressed and presented on the cell surface as a transmembrane protein. Copper catalyzed click chemistry (CuAAC) is performed, forming a stapled peptide on the cell surface. Finally, a combination of fluorescently- or magnetically- activated proteins is added to facilitate protein fitness measurements using flow cytometry.

1.1.2 Stapled peptide design parameters

As mentioned previously, there are several parameters that influence the drug-like properties of stapled peptide therapeutics **(Figure 1.4)**. With the powerful tool, **S**tabilized **P**eptide **E**ngineering by ***E. coli*** **D**isplay (SPEED), the design space of stapled peptides can be explored more easily. First, the peptide sequence is the most important variable as the sequence contributes most of the chemical diversity through the canonical 20 amino acids and their spatial orientation towards the binding target. One factor of sequence design is understanding 'hot spot residues', or amino acids that contribute large portions of binding enthalpy.[42,43] The loss of such residues is likely to result in nonfunctional sequences. One challenge is the selection of sequence variants to test: because SPEED can only assay ~10^9 variants, but the sequence space of all peptides is much large (~10^{30}), there is great need for selection of which amino acids and which

positions should be mutated. Because SPEED uses DNA-encoded peptide libraries, the mutation of sequence is easily accomplished with standard molecular biology protocols. Due to the unique structure of the DNA codon table, combinations of amino acids can be selected using degenerate codons (such as 'NNC', which samples all amino acids except E,K,M,Q, and W). These commercially available DNA primers are thus able to generate libraries of focused stapled peptide variants with ease. Therefore, SPEED is uniquely poised to understand how the sequence of a stapled peptide contributes to its function.

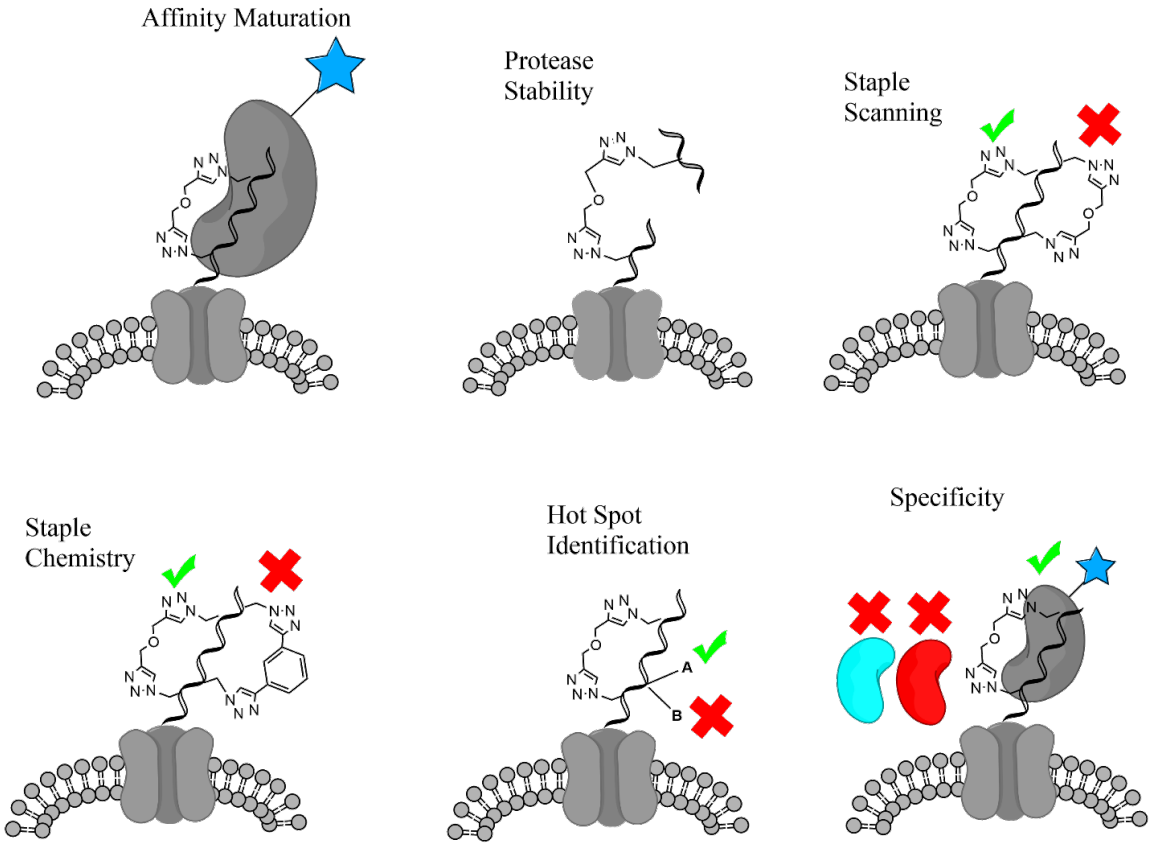


Figure 1.4: Parameters of stapled peptide. This includes binding affinity, protease stability, staple location, staple chemistry, hot spot residues and their contributions to binding affinity, and specificity.

Another important design criterion for stapled peptides is the location of the staple. Because these staples protrude from the peptide backbone, they can form new interactions that are favorable (or disfavorable) towards binding targets. Therefore, the selection of staple location

is an important factor for the binding affinity alone. However, the staple location also stabilizes secondary structure differently, replaces naturally occurring amino acids that might lend properties to peptides, and forms patches of hydrophobicity that can improve cell permeability. These considerations led us to use stapled peptide design to perform a staple scan, where each staple location is synthesized in the peptide context and its fitness is measured.[8,18,44–46] Because this step normally involves many rounds of chemical synthesis, staple scanning is very laborious. SPEED can easily change the staple location by altering the location of methionine residues (an 'ATG' codon) and thus is able to study the relationship of staple location with relative ease.

The final important consideration of stapled peptide design is the choice of staple itself. Because there are many chemistries that form stapled peptides, different chemical properties that arise from the staple influence the design of the peptide itself. One particularly successful design of stapled peptides involves the use of hydrocarbon staples, which are extremely hydrophobic and thus greatly contribute to cell permeability.[5,18,46,47] Other linkers impart different functionalities, such as enhanced protease stability, different charge, new hydrogen bonding motifs, among others.[24,48] While solid phase peptide synthesis builds peptides one amino acid at a time and is therefore amenable to changing staple chemistries with relative ease, SPEED presents azide residues which enables complete modularity with any bisalkyne motifs without any additional synthesis. There are many commercially available bisalkyne compounds, enabling the screening of many stapling chemistries with minimal effort.[49]

1.1.3 B cell lymphoma 2 targets

Another important design criteria for stapled peptides is the target itself. While stapled peptide therapeutics can inhibit many protein-protein interactions (PPI), not all PPI's are disease related and thus targeting proteins that are highly specific to disease states will minimize the off-

target activity.[3] One important family of disease-related proteins are the B cell lymphoma 2 (Bcl-2) proteins, which modulate directed cell death or apoptosis **(Figure 1.5)**. In healthy cells, there is a careful balance of anti-apoptotic Bcl-2 proteins and pro-apoptotic (Bad/Noxa/Bim, among others) proteins in dynamic equilibrium on the mitochondrial cell surface within cells.[50,51] Certain stimuli can activate this pathway, such as irreversible DNA damage or hypoxia, which signals the cell to undergo apoptosis and relinquish function to other cells. In this case, pro-apoptotic proteins sequester anti-apoptotic proteins, allowing pore-forming proteins Bak and Bax to hetero-oligomerize, destroying the proton gradient in mitochondria, which inhibits cellular metabolism and starts a biochemical cascade towards apoptosis. However, the dysregulation of this pathway is common among many cancers as cells can evade apoptosis through the over expression of Bcl-2 proteins and facilitate indefinite proliferation (among other factors). Many cancers are characterized by this overexpression of Bcl-2 proteins and thus the antagonism of the Bcl-2/ pro-apoptotic proteins is a powerful approach towards the selective killing of cancer cells.[46,50]

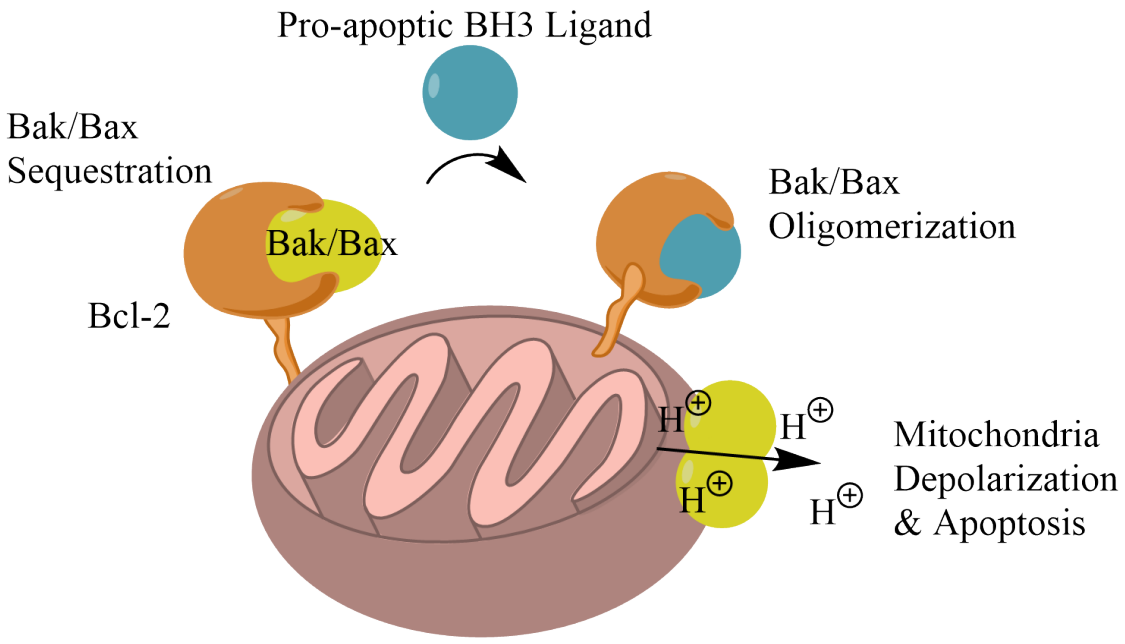


Figure 1.5: Simplified Bcl-2 apoptosis biochemistry. At the surface of mitochondria, cancer cells overexpress Bcl-2 apoptosis, which inhibits Bak/Bax from hetero-oligomerizing, depolarizing mitochondria and causing a biochemical

cascade that results in apoptosis. Through the displacement of Bak/Bax from Bcl-2 proteins with a pro-apoptotic BH3 ligand, such as naturally occurring Bim or Bad or synthetic alternatives, normal apoptotic function can occur.

Despite the promise of selective targeting of cancer cells via Bcl-2 protein antagonism, its targeting via common therapeutic modalities has had limited clinical success. Due to the location of Bcl-2 proteins on mitochondria, there are no protein-based therapeutics approaches towards their antagonism. Small molecule drugs, on the other hand, are difficult to engineer with high affinity owing to the structural motifs between the pro- and anti-apoptotic proteins, which is relatively flat and not hydrophobic.[52,53] An additional challenge facing small molecule drugs that is there are multiple members of the Bcl-2 protein family: Mcl-1, Bfl-1, Bcl-x_L, Bcl-w, and Bcl-2.[30,51,52,54–58] These proteins are highly homologous but play different roles in regulating apoptosis and resistance to chemotherapy. Thus, achieving both affinity and high specificity is necessary to produce a therapeutic molecule with minimal off-target activity.

To overcome these barriers, researchers have proposed peptide therapeutics, which can likely antagonize these interactions with high affinity and specificity as evidenced by naturally occurring peptides that target a subset of Bcl-2 proteins.[30,46,59] While there has been much effort to engineer high affinity and specificity linear pro-apoptotic peptides,[30–38,55,59–65] there are far fewer reports of analogous stapled peptides.[40,46,66] SPEED is uniquely poised to develop high affinity and specificity stapled peptides for these targets, owing to its ability to rapid evaluate import factors like peptide sequence, staple location, and staple chemistry.

1.1.4 Advancing information gained from binary sorting experiments to expand design space

Before SPEED can be applied to identify candidates with desired fitness, a library of protein variants needs to be designed, and a plan for measuring and selecting high fitness variants needs to be constructed. As mentioned earlier, the space of protein variants is vastly large compared to the number of sequences even a high-throughput approach like SPEED can

assay. Thus, determining which sequences are tested is an extremely important consideration. Another important design criteria is the ratio of design space and the experimental throughput. For example, if a library of 10^5 possible protein variants is designed, SPEED can measure each clone with confidence. However, if none of the 10^5 variants have the desired fitness, there is no ability to 'extrapolate' in the design space by combining mutations that individually contribute towards high fitness. Conversely, if a library is designed that has 10^9 variants, but only 10^5 points are measured, it is equally likely that functional clones are identified (assuming that there are equally many sequences that are worth testing) but data driven approaches can be applied to identify new candidates using the dataset of experimentally measured peptides **(Figure 1.6)**. This

is an emerging approach for the design of protein variants libraries and SPEED is well equipped to test its implementation for stapled peptides.

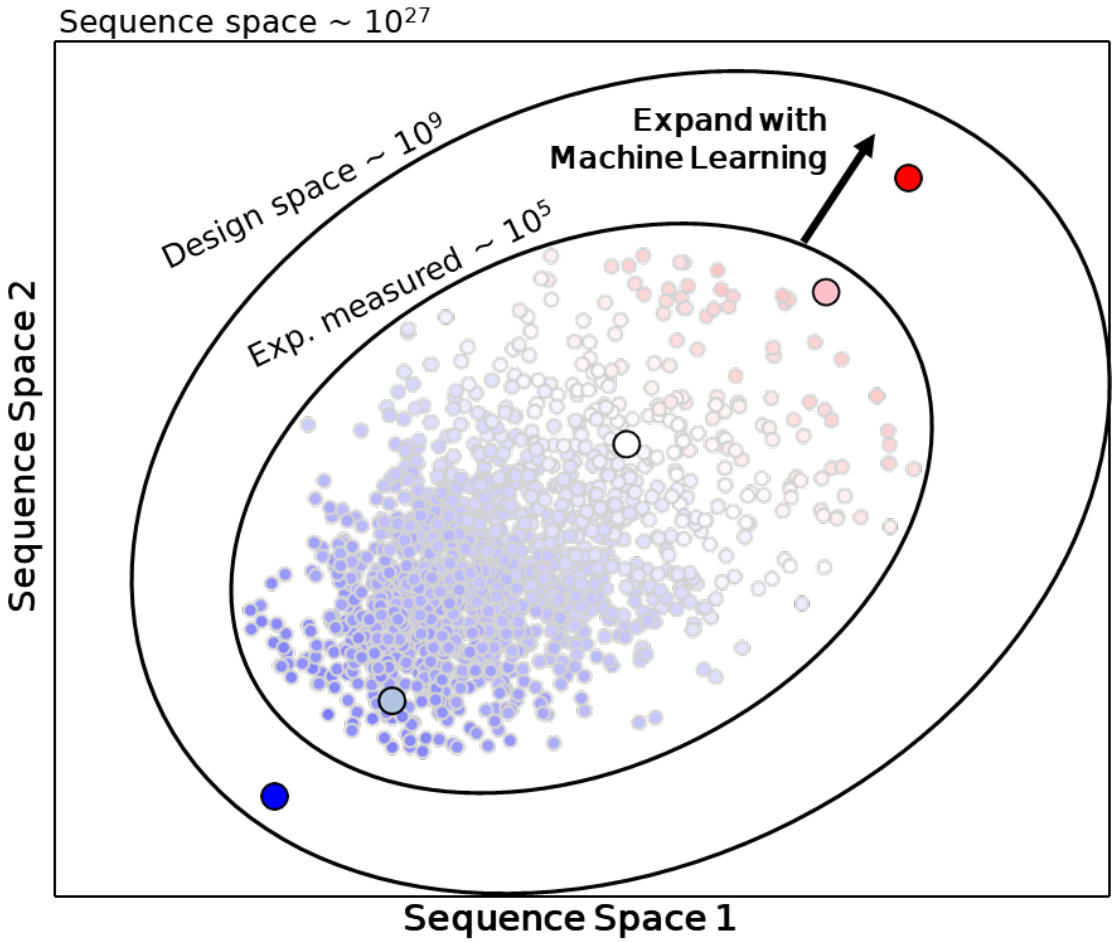


Figure 1.6: The space of all possible peptide sequences is far smaller than experimental capacity. However, by leveraging the data gained from a large number of experimentally measured sequences with techniques like machine learning, the function of unseen sequences can be predicted, which may be higher than ones experimentally observed.

Generating data in such a manner that is easily amenable to models that connect sequence to function is non-trivial. Because sorting experiments are dependent on a multitude of factors, such as cell surface display platform, protein target, function of interest, cell sorter used, library size, among many others, it is difficult to set heuristics for all sorting campaigns. Furthermore, sorting is usually done in a 'binary' manner: sort the highest fitness clones and leave the rest behind **(Figure 1.7)**. While protein properties exist on a continuous scale (such as binding

affinities, which range from 10^{-15} for streptavidin: biotin to 10^{-3} M for small molecule: enzyme interactions), data generated from sorting is not directly amenable to quantitative measures of protein fitness. A method capable of converting data from typical sorting experiments into quantitative measures of protein fitness would greatly accelerate not only the design of stapled peptides but additionally other proteins.

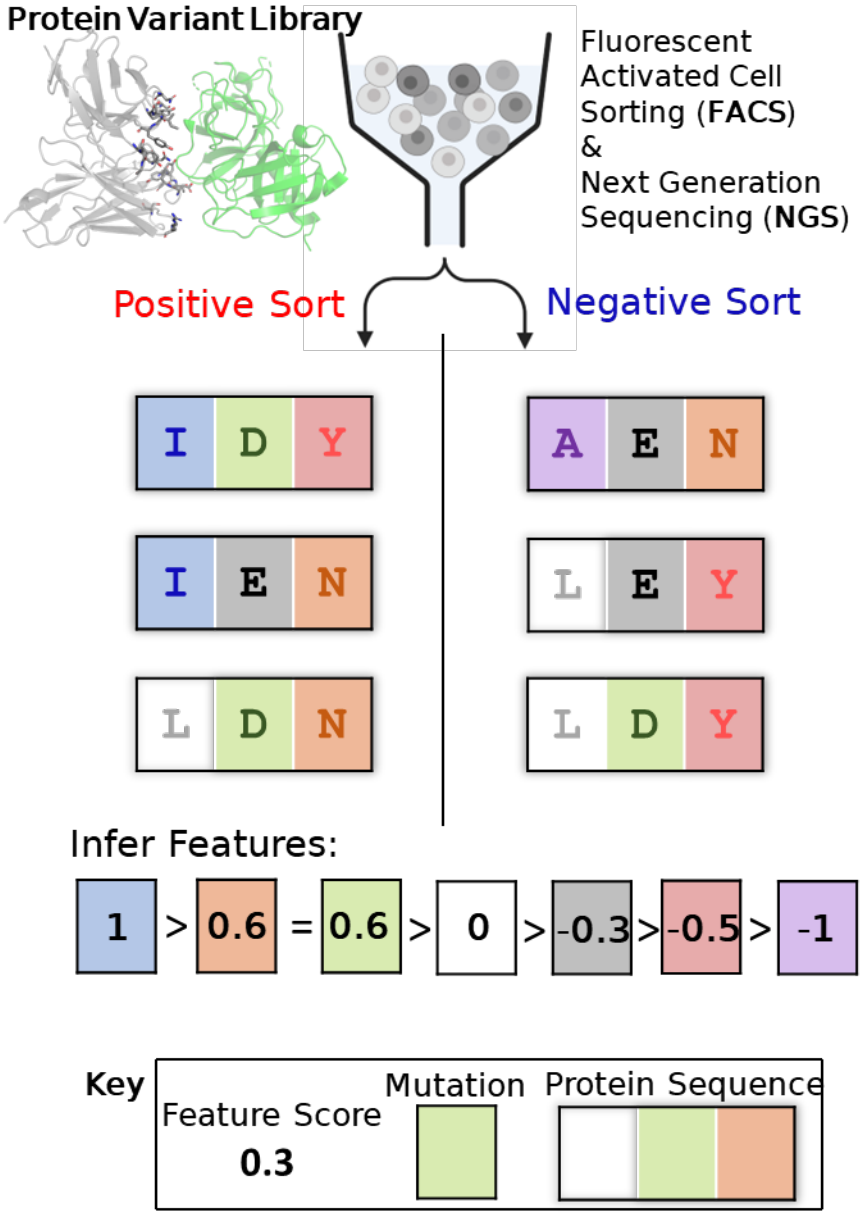


Figure 1.7: Protein variant libraries are typically sorted into two pools: ones denoted by high function and another with low function. While these bins do not explicitly map to quantitative measures of protein fitness, by interpreting the protein sequences and their frequencies from cell sorting and deep sequencing, quantitative features

of these mutations can be inferred. Then, these quantitative features can be used to score unseen sequences towards higher fitness variants.

1.1.5 Introduction Summary

In summary, this thesis explores new methods and results in the generation of stapled peptides and other proteins using directed evolution. In **Chapter 2**, this thesis explores new methods that accelerate the design of stapled peptides, including several key parameters such as hot spot amino acids, staple location, and staple chemistry. In **Chapter 3,** this thesis applies the design parameters from Chapter 2 to design highly specific peptides towards Bcl-x_L, an important drug-target in many cancers. Finally, in **Chapter 4**, this thesis describes a new method towards the measurement of continuous properties from simple binary sorting experiments using machine learning. By applying the weights from machine learning, we optimize peptide properties (binding affinity and specificity) by extrapolating beyond experimentally seen sequence space. We evaluate this method on several protein engineering tasks, such as measuring fluorescence, multi-objective optimization of antibodies, and predicting specificity of Bcl-2 linear peptides, and find that it works across all tasks evaluated.

1.1.6 References

1. Anfinsen, C. B. Principles that Govern the Folding of Protein Chains. *Science (1979)* **181**, 223–232 (1973).

2. Verdine, G. L. & Hilinski, G. J. All-hydrocarbon stapled peptides as synthetic cell-accessible mini-proteins. *Drug Discov Today Technol* **9**, e41–e47 (2012).

3. Hopkins, A. L. & Groom, C. R. The druggable genome. *Nat Rev Drug Discov* **1**, 727–730 (2002).

4. Walensky, L. D. & Bird, G. H. Hydrocarbon-Stapled Peptides: Principles, Practice, and Progress. (2015).

5. Walensky, L. D. & Bird, G. H. Hydrocarbon-stapled peptides: principles, practice, and progress. *J Med Chem* **57**, 6275–6288 (2014).

6. Walensky, L. D. & Bird, G. H. Hydrocarbon-Stapled Peptides: Principles, Practice, and Progress. (2015).

7. Atangcho, L., Navaratna, T. & Thurber, G. M. Hitting Undruggable Targets: Viewing Stabilized Peptide Development through the Lens of Quantitative Systems Pharmacology. *Trends Biochem Sci* 1–19 (2019).

8. Bird, G. H. *et al.* Hydrocarbon double-stapling remedies the proteolytic instability of a lengthy peptide therapeutic. *Proceedings of the National Academy of Sciences* **107**, 14093–14098 (2010).

9. Grossmann, T. N. *et al.* Inhibition of oncogenic Wnt signaling through direct targeting of B-catenin. *Proceedings of the National Academy of Sciences* **109**, 17942–17947 (2012).

10. 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).

11. Rossi Sebastiano, M. *et al.* Impact of Dynamically Exposed Polarity on Permeability and Solubility of Chameleonic Drugs beyond the Rule of 5. *J Med Chem* **61**, 4189–4202 (2018).

12. Mendive-Tapia, L. *et al.* New peptide architectures through C-H activation stapling between tryptophan-phenylalanine/tyrosine residues. *Nat Commun* **6**, 1–9 (2015).

13. Diderich, P. *et al.* Phage Selection of Chemically Stabilized α-Helical Peptide Ligands. *ACS Chem Biol* **11**, 1422–1427 (2016).

14. Kawamoto, S. *et al.* Design of Triazole-Stapled BCL9 α-Helical Peptides to Target the β-Catenin/B-Cell CLL/lymphoma 9 (BCL9) Protein−Protein Interaction. *J Med Chem* **55**, 1137–1146 (2011).

15. Zhang, G. *et al.* A Solid-Phase Approach to Accessing Bisthioether-Stapled Peptides Resulting in a Potent Inhibitor of PRC2 Catalytic Activity. 17073–17078 (2018) doi:10.1002/anie.201810007.

16. Pessi, A. *et al.* Cholesterol-conjugated stapled peptides inhibit Ebola and Marburg viruses in vitro and in vivo. *Antiviral Res* 104592 (2019) doi:10.1016/j.antiviral.2019.104592.

17. Meng, G. *et al.* Design and Biological Evaluation of m -Xylene Thioether-Stapled Short Helical Peptides Targeting the HIV-1 gp41 Hexameric Coiled–Coil Fusion Complex . *J Med Chem* (2019) doi:10.1021/acs.jmedchem.9b00882.

18. Bird, G. H. *et al.* Biophysical determinants for cellular uptake of hydrocarbon-stapled peptide helices. *Nat Chem Biol* **12**, 845–852 (2016).

19. Kenrick, S., Rice, J. & Daugherty, P. Flow Cytometric Sorting of Bacterial Surface-Displayed Libraries. *Curr Protoc Cytom* 1–27 (2007) doi:10.1002/0471142956.cy0406s42.

20. Boder, E. T. & Wittrup, K. D. Yeast surface display for screening combinatorial polypeptide libraries. *Nat Biotechnol* **15**, 553–557 (1997).

21. Bessette, P. H., Rice, J. J. & Daugherty, P. S. Rapid isolation of high-affinity protein binding peptides using bacterial display. *Protein Engineering, Design and Selection* **17**, 731–739 (2004).

22. Sarkisyan, K. S. *et al.* Local fitness landscape of the green fluorescent protein. *Nature* **533**, 397–401 (2016).

23. Navaratna, T. *et al.* Directed Evolution Using Stabilized Bacterial Peptide Display. *J Am Chem Soc* **142**, 1882–1894 (2020).

24. Case, M., Navaratna, T., Vinh, J. & Thurber, G. M. Rapid Evaluation of Staple Placement in Stabilized Alpha Helices using Bacterial Surface Display. *ACS Chem Biol* **18**, 905–914 (2023).

25. Derda, R. *et al.* Diversity of phage-displayed libraries of peptides during panning and amplification. *Molecules* **16**, 1776–1803 (2011).

26. Rice, J. J. & Daugherty, P. S. Directed evolution of a biterminal bacterial display scaffold enhances the display of diverse peptides. *Protein Engineering, Design and Selection* **21**, 435–442 (2008).

27. Getz, J. A., Schoep, T. D. & Daugherty, P. S. Peptide discovery using bacterial display and flow cytometry. *Methods Enzymol* **503**, 75–97 (2012).

28. Li, A. *et al.* High-throughput profiling of sequence recognition by tyrosine kinases and SH2 domains using bacterial peptide display. *bioRxiv* 1–41 (2022) doi:10.1101/2022.08.01.502334.

29. Liu, G. W. *et al.* Efficient Identification of Murine M2 Macrophage Peptide Targeting Ligands by Phage Display and Next-Generation Sequencing. *Bioconjug Chem* **26**, 1811–1817 (2015).

30. Foight, G. W. & Keating, A. E. Locating Herpesvirus Bcl-2 Homologs in the Specificity Landscape of Anti-Apoptotic Bcl-2 Proteins. *J Mol Biol* **427**, 2468–2490 (2015).

31. Dutta, S., Chen, T. S. & Keating, A. E. Peptide ligands for pro-survival protein Bfl-1 from computationally guided library screening. *ACS Chem Biol* **8**, 778–788 (2013).

32. Chen, T. S., Palacios, H. & Keating, A. E. Structure-based redesign of the binding specificity of anti-apoptotic Bcl-xL. *J Mol Biol* **425**, 171–185 (2013).

33. Stewart, M. L., Fire, E., Keating, A. E. & Walensky, L. D. The MCL-1 BH3 helix is an exclusive MCL-1 inhibitor and apoptosis sensitizer. *Nat Chem Biol* **6**, 595–601 (2010).

34. Jenson, J. M., Ryan, J. A., Grant, R. A., Letai, A. & Keating, A. E. Epistatic mutations in PUMA BH3 drive an alternate binding mode to potently and selectively inhibit anti-apoptotic Bfl-1. *Elife* **6**, 1–23 (2017).

35. Foight, G. W., Ryan, J. A., Gullá, S. v., Letai, A. & Keating, A. E. Designed BH3 peptides with high affinity and specificity for targeting Mcl-1 in cells. *ACS Chem Biol* **9**, 1962–1968 (2014).

36. Dutta, S. *et al.* Potent and specific peptide inhibitors of human pro-survival protein bcl-xl. *J Mol Biol* **427**, 1241–1253 (2015).

37. Jenson, J. M. *et al.* Peptide design by optimization on a data parameterized protein interaction landscape. *Proc Natl Acad Sci U S A* **115**, E10342–E10351 (2018).

38. Dutta, S. *et al.* Determinants of BH3 Binding Specificity for Mcl-1 versus Bcl-xL. *J Mol Biol* **398**, 747–762 (2010).

39. Bertoldo, D. *et al.* Phage Selection of Peptide Macrocycles against β-Catenin to Interfere with Wnt Signaling. *ChemMedChem* **11**, 834–839 (2016).

40. Araghi, R. R. *et al.* Iterative optimization yields Mcl-1–targeting stapled peptides with selective cytotoxicity to Mcl-1–dependent cancer cells. *Proc Natl Acad Sci U S A* **115**, E886–E895 (2018).

41. Stieglitz, J. T., Kehoe, H. P., Lei, M. & van Deventer, J. A. A Robust and Quantitative Reporter System to Evaluate Noncanonical Amino Acid Incorporation in Yeast. *ACS Synth Biol* **7**, 2256–2269 (2018).

42. Kortemme, T. & Baker, D. A simple physical model for binding energy hot spots in protein-protein complexes. *Proceedings of the National Academy of Sciences* **99**, (2002).

43. Clackson, T. & Wells, J. A. A Hot Spot of Binding Energy in a Hormone-Receptor Interface. *Science (1979)* **267**, 383–386 (1995).

44. Bernal, F. *et al.* A Stapled p53 Helix Overcomes HDMX-Mediated Suppression of p53. *Cancer Cell* **18**, 411–422 (2010).

45. Lau, Y. H. *et al.* Investigating peptide sequence variations for 'double-click' stapled p53 peptides. *Org Biomol Chem* **12**, 4074–4077 (2014).

46. Loren D. Walensky *et al.* Activation of Apoptosis in Vivo by a Hydrocarbon-Stapled BH3 Helix. *Science (1979)* **23**, 1–7 (2004).

47. Bird, G. H. *et al.* Hydrocarbon double-stapling remedies the proteolytic instability of a lengthy peptide therapeutic. *Proceedings of the National Academy of Sciences* **107**, 14093–14098 (2010).

48. Lau, Y. H. *et al.* Functionalised staple linkages for modulating the cellular activity of stapled peptides. *Chem Sci* **5**, 1804–1809 (2014).

49. Lau, Y. H. *et al.* Functionalised staple linkages for modulating the cellular activity of stapled peptides. *Chem Sci* **5**, 1804–1809 (2014).

50. Adams, J. M. & Cory, S. The Bcl-2 Protein Family: Arbiters of Cell Survival. *Science (1979)* **281**, 1322–1326 (1998).

51. Shamas-Din, A., Kale, J., Leber, B. & Andrews, D. W. Mechanisms of action of Bcl-2 family proteins. *Cold Spring Harb Perspect Biol* **5**, 1–21 (2013).

52. Chen, L. *et al.* Differential targeting of prosurvival Bcl-2 proteins by their BH3-only ligands allows complementary apoptotic function. *Mol Cell* **17**, 393–403 (2005).

53. Shin, Y. H. & Yang, H. Exploration of α/β/γ-peptidomimetics design for BH3 helical domains. *Chemical Communications* **58**, 945–948 (2022).

54. Kale, J., Osterlund, E. J. & Andrews, D. W. BCL-2 family proteins: Changing partners in the dance towards death. *Cell Death Differ* **25**, 65–80 (2018).

55. London, N., Gullá, S., Keating, A. E. & Schueler-Furman, O. In silico and in vitro elucidation of BH3 binding specificity toward Bcl-2. *Biochemistry* **51**, 5841–5850 (2012).

56. Opferman, J. T. Attacking cancer's Achilles heel: antagonism of anti-apoptotic BCL-2 family members. *FEBS Journal* **283**, 2661–2675 (2016).

57. Czabotar, P. E., Lessene, G., Strasser, A. & Adams, J. M. Control of apoptosis by the BCL-2 protein family: Implications for physiology and therapy. *Nat Rev Mol Cell Biol* **15**, 49–63 (2014).

58. Certo, M. *et al.* Mitochondria primed by death signals determine cellular addiction to antiapoptotic BCL-2 family members. *Cancer Cell* **9**, 351–365 (2006).

59. DeBartolo, J., Taipale, M. & Keating, A. E. Genome-Wide Prediction and Validation of Peptides That Bind Human Prosurvival Bcl-2 Proteins. *PLoS Comput Biol* **10**, 1–10 (2014).

60. Frappier, V. & Keating, A. E. Data-driven computational protein design. *Curr Opin Struct Biol* **69**, 63–69 (2021).

61. Reich, L., Dutta, S. & Keating, A. E. SORTCERY - A High-Throughput Method to Affinity Rank Peptide Ligands. *J Mol Biol* **427**, 2135–2150 (2015).

62. Fu, X., Apgar, J. R. & Keating, A. E. Modeling Backbone Flexibility to Achieve Sequence Diversity: The Design of Novel α-Helical Ligands for Bcl-xL. *J Mol Biol* **371**, 1099–1117 (2007).

63. Debartolo, J., Dutta, S., Reich, L. & Keating, A. E. Predictive Bcl-2 family binding models rooted in experiment or structure. *J Mol Biol* **422**, 124–144 (2012).

64. Fire, E., Gullá, S. v., Grant, R. A. & Keating, A. E. Mcl-1-Bim complexes accommodate surprising point mutations via minor structural changes. *Protein Science* **19**, 507–519 (2010).

65. Frappier, V., Jenson, J. M., Zhou, J., Grigoryan, G. & Keating, A. E. Tertiary Structural Motif Sequence Statistics Enable Facile Prediction and Design of Peptides that Bind Anti-apoptotic Bfl-1 and Mcl-1. *Structure* **27**, 606-617.e5 (2019).

66. Rezaei Araghi, R., Ryan, J. A., Letai, A. & Keating, A. E. Rapid Optimization of Mcl-1 Inhibitors using Stapled Peptide Libraries Including Non-Natural Side Chains. *ACS Chem Biol* **11**, 1238–1244 (2016).

Chapter 2 Rapid Evaluation of Staple Placement in Stabilized Alpha Helices using Bacterial Surface Display

This chapter is derived from the following publication:

Marshall Case, Tejas Navaratna, Jordan Vinh, and Greg Thurber. "Rapid Evaluation of Staple Placement in Stabilized α Helices Using Bacterial Surface Display." *ACS Chemical Biology* **18** (4), 905-914 (2023). DOI: 10.1021/acschembio.3c00048

Abstract

There are a wealth of proteins involved in disease that cannot be targeted by current therapeutics because they are inside cells, inaccessible to most macromolecules, and lack small-molecule binding pockets. Stapled peptides, where two amino acid side chains are covalently linked, form a class of macrocycles that have the potential to penetrate cell membranes and disrupt intracellular protein-protein interactions. However, their discovery relies on solid phase synthesis, greatly limiting queries into their complex design space involving amino acid sequence, staple location, and staple chemistry. Here, we use Stabilized Peptide Engineering by *E. coli* Display (SPEED), which utilizes non-canonical amino acids and click-chemistry for stabilization, to rapidly screen staple location and linker structure to accelerate peptide design. After using SPEED to confirm hot spots in the mdm2-p53 interaction, we evaluated different staple locations and staple chemistry to identify several novel nanomolar and sub-nanomolar antagonists. Next, we evaluated SPEED in the B cell lymphoma 2 (Bcl-2) protein family, which is responsible for regulating apoptosis. We report that novel staple locations modified in the

context of BIM, a high affinity but non-specific naturally occurring peptide, improve its specificity against the highly homologous proteins in the Bcl-2 family. These compounds demonstrate the importance of screening linker location and chemistry in identifying high affinity and specific peptide antagonists. Therefore, SPEED can be used as a versatile platform to evaluate multiple design criteria for stabilized peptide engineering.

Introduction

It is estimated that ~85% of disease-associated proteins are "undruggable": inside the cell and inaccessible to large biologics but lacking small molecule binding sites [3]. Stapled peptides, short chains of amino acids where two residues are covalently crosslinked, have been proposed as one type of therapeutic framework to fill this gap.[46,67,68] Covalent sidechain crosslinking has the potential to improve target affinity, facilitate cell entry, and enhance proteolytic stability.[5] Since most protein-protein interactions are mediated through alpha-helical secondary structure, there is a natural precedent for binding peptides to be locked in this conformation. The staple forces the peptide into a conformation with enhanced alpha helicity, decreasing the entropic penalty of binding and increasing affinity.[69] The larger size of the peptide enables a greater binding surface area, thereby creating the potential for high affinity interactions without the need for a deep hydrophobic binding pocket. The intramolecular hydrogen bonding in the peptide backbone can reduce the energy barrier to diffuse across the lipid bilayer by shedding solvating water molecules.[18] Additionally, since most proteases encountered by a peptide *in vivo* recognize linear conformations, a peptide in a helical structure tends to have a longer pharmacological half-life.[8]

Despite this promise, stapled peptides have several challenges that must be overcome. These agents must be engineered with high enough membrane permeability and intracellular

stability to engage sufficient levels of target based on the binding affinity and specificity.[70] Engineering these properties is challenging, and the discovery and development of stapled peptides typically relies on solid-phase synthesis, where peptides are chemically synthesized one amino acid at a time, limiting throughput and evaluation. A peptide's sequence and staple location have 10^{26} possibilities even for a simple peptide of length 20, which presents both a design opportunity but also an enormous challenge. Therefore, rational design has been the common approach, with the number of evaluated sequences often on the order of dozens.[16,18,40,45,71–78] Previously, we have developed Stabilized Peptide Engineering by *E. coli* Display (SPEED) to discover high affinity mdm2 binders.[23] In this work, we extend this approach and demonstrate its ability to accelerate stapled peptide design by identifying hotspot residues, functional staple locations, and diverse chemical linkers.

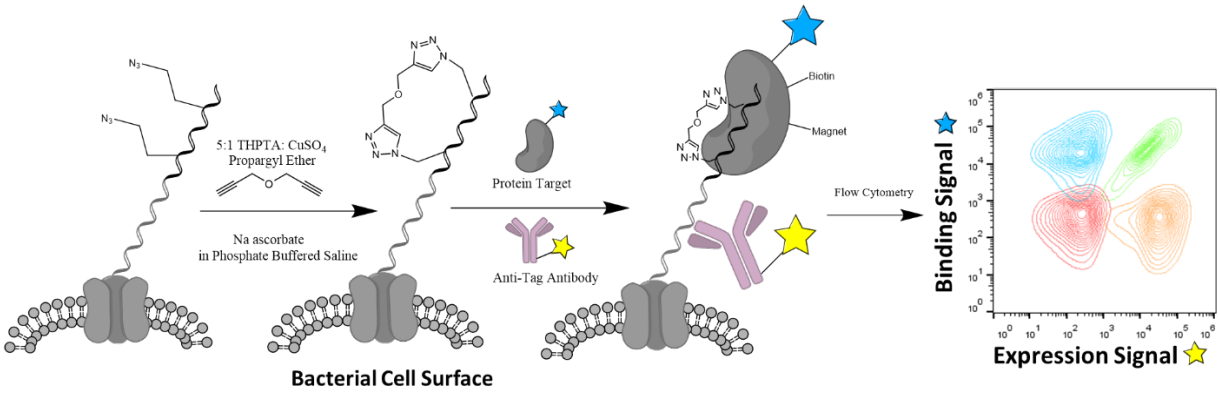


Figure 2.1: Stabilized Peptide Engineering by E. coli Display (SPEED). DNA encoding peptide is transformed into E. coli and expressed on the cell surface by incubating bacteria in an azide containing methionine analog. After click chemistry is performed directly on the cell surface, bacteria are incubated with fluorescent epitope tag antibody and protein target. Finally, bacterial cells are analyzed via flow cytometry.

Bacteria possess several unique abilities suitable for tackling the challenges of stabilized peptide design. First, most stapling chemistries are not compatible with the 20 canonical amino acids, and other surface-based presentation approaches such as phage- and yeast surface display do not currently have high enough non-natural amino acid incorporation efficiencies to staple on

the cell surface, although progress is being made.[5,41,79–81] Meanwhile, bacteria are able to incorporate many types of non-natural residues, whether from methionine substitution, stop codon read through, or other genetic code manipulation.[82–85] Of particular importance is azidohomoalanine residues which contain copper catalyzed click chemistry (CuAAC) suitable azides, and demonstrate exemplary efficiencies of >95% incorporation.[28,86,87] SPEED leverages this high incorporation efficiency to display two azides directly on the cell surface to form a stapled peptide with an intramolecularly reacted bisalkyne (**Figure 2.1**).[23] The modularity of this reaction scheme enables the use of any bisalkyne for reaction, meaning that SPEED is equipped to engineer both the peptide sequence and staple. The bio-orthogonality of click chemistry gives bacteria an additional advantage over chemistries that use canonical amino acids, such as lactam bridge formation or cysteine alkylation, which can interfere with other proteins present on the cell surface and may impact peptide property measurement. Ribosome-based display can incorporate many types of non-natural amino acids that facilitate peptide stapling, but like phage, their small size makes it challenging to use in assays that rapidly measure stapled peptide properties like flow cytometry and fluorescent activated cell sorting (FACS).[88] Similarly, one-bead-one-compound approaches have enabled the measurement of ~10^3 stapled peptides in parallel but rely on mass spectrometric based methods and imaging that render property measurement difficult.[66] This approach also relies on library members having unique masses, meaning that residues with the same mass, like leucine and isoleucine, cannot be distinguished. Furthermore, this approach is less well-suited to evaluate staple location and peptide sequence in tandem as staple location cannot be randomized efficiently using solid phase peptide synthesis. In summary, *E. coli* possess several attractive traits: facile genetic manipulation, efficient non-

natural acid incorporation, modular bisalkyne linker chemistry, and compatibility with high-throughput screening methods.

In this work, we demonstrate that bacteria enable rapid measurement of affinity and specificity of peptides with different staple locations, staple types, and sequence mutations in the context of two systems, murine double minute-2 (mdm2) and B cell lymphoma 2 (Bcl-2) targeted peptides. In the first model system, the critical p53 tumor suppressor transcription factor is rendered unstable by mdm2 overexpression.[89] Inhibition of mdm2 by a stabilized p53-like peptide (PLP) reduces the viability of cancer cells.[76] In the second system, overexpression of Bcl-2 proteins leads to inhibition of apoptosis factors that prevent cancer cells from dying. Inhibition of Bcl-2 proteins by stabilized BH3 peptides regenerates cells' ability to undergo apoptosis.[57] We use SPEED to design novel stapled peptides in both these systems in the pursuit of higher affinity, greater specificity, and more structurally diverse molecules. We then translate these peptides from the cell surface to solution phase binding to confirm that the bacterial surface display captures soluble peptide properties. The results demonstrate that bacterial surface display can be used to accelerate stapled peptide engineering.

Methods

2.1.1 Purification of Mdm2 and Bcl-2 protein

Mdm2-GST was expressed and purified as described previously.[23,38] Briefly, mdm2-GSTexpressing plasmid was ordered from AddGene (plasmid #16237). (. Protein was harvested from pLysS BL21 DE3 *E.* coli cells and subsequently purified with agarose glutathione beads (Thermo Fisher) in phosphate buffered saline (PBS) at pH 7.4. After washing with PBS, mdm2-GST was eluted with 50mM Tris pH 8.0 and 10mM reduced glutathione before purification via size exclusion chromatography in PBS with 1% glycerol and 1mM DTT. Protein was

concentrated using a 10kDa centrifugal concentrator and labeled with NHS-EzLink-Biotin (Thermo Fisher) and excess biotin was removed via dialysis in PBS with 3.5K molecular weight tubing (Thermo Fisher) with 10% glycerol and 1mM DTT.

Bcl-2 genes were ordered from IDT and cloned into the pqe80L vector using BamHI and HindIII with an N-terminal His tag. Briefly, LB with ampicillin was inoculated with cells from an overnight grow-out until reaching OD_{600} ~1.0 and induced with IPTG at 1mM for 5 hours at 37^{o}C. Cells were resuspended in resuspension buffer (comprised of Tris-HCl buffer at pH 8.0 with 0.5mM NaCl, 5mM imidazole, protease inhibitor, and 2mM DTT), sonicated, and centrifuged at 35,000g x 30 min at 4^{o}C. The supernatant was then loaded onto 2mL of prewashed Ni-NTA resin and washed with 10-20mL of resuspension buffer. For protein intended for subsequent labeling with NHS-biotin or NHS-fluorophore, it was buffer exchanged on resin into PBS with 2mM DTT and 5mM imidazole. Protein was eluted in the resuspension buffer supplemented with 500mM imidazole. Proteins were assayed for purity and yield using SDS-page gel and spectroscopy. Proteins were labeled in 0.1M $NaHCO_3$ and NHS-biotin or NHS-fluorophore was added at a 10:1 NHS:protein ratio. Proteins were then purified using size exclusion chromatography on a S200 10/300 increase GL or S75 10/300 increase GL column in PBS with 1mM DTT and 1% glycerol. Fractions corresponding to Bcl-2 proteins were concentrated using a 10kDa molecular weight cut-off filter in PBS with 2mM DTT supplemented with 10% glycerol at 4^{o}C. Degree of labeling was quantified using a fluorescent biotin quantification kit (Thermo Fisher) or quantified directly using spectroscopy. Typically, a DoL greater than 0.3 gave sufficient separation between flow cytometry signal and noise.

2.1.2 Bacterial surface display and on-cell click chemistry

Primers encoding peptides were purchased from IDT and incorporated into the eCPX2-pqe80L plasmid[21,26] with a 2 step PCR protocol using Q5 Hot Start Polymerase, SfiI restriction enzyme, and T4 ligase (all from NEB). An extended peptide linker containing an HA tag (final sequence YPYDVPDYAAGGGSGGGS) was incorporated into the bacterial cell surface display scaffold to normalize binding to display level with two-color labeling.[30] The peptide sequence was confirmed via Sanger sequencing (Michigan Advanced Genomics Core or Eurofins Genomics). Methionine auxotrophic *E. coli* cells (TYJV2 strain) were used in all surface display experiments and were grown overnight in M9 media containing 4mg/mL methionine and 100ug/mL ampicillin. Then, media was inoculated with the overnight culture at a 1:20 ratio for 150 minutes at 37^{o}C. Cells were then switched to M9-amp with no methionine for metabolic depletion for 30 minutes at 37^{o}C, followed by a 4 hour induction in M9-amp with 4mg/mL azidohomoalanine and 1mM IPTG at 22^{o}C. At this point, bis-azide containing peptides on the surface of bacteria were reacted to form stapled peptides in 50uM $CuSO_4$, 250uM THPTA, 500uM (p53-like-peptides) or 100uM (Bcl-2) propargyl ether for 4 hours at 4^{o}C. The extent of reaction was determined as previously reported and shown in **Figure 2.2**.[23]To measure the affinity of a peptide, the peptide-displaying bacteria was incubated for at least 4 hours on ice with 6-8 concentrations of biotinylated or fluorescently labeled mdm2 or Bcl-2 protein. Cells were washed once with PBS/ 0.1% BSA, labeled anti-HA-Alexa Fluor 488 for display level measurement and streptavidin-Alexa Fluor 647 for biotinylated protein detection for 15 minutes on ice, washed again, and then

analyzed via flow cytometry (Attune NXT or BioRad Ze5).

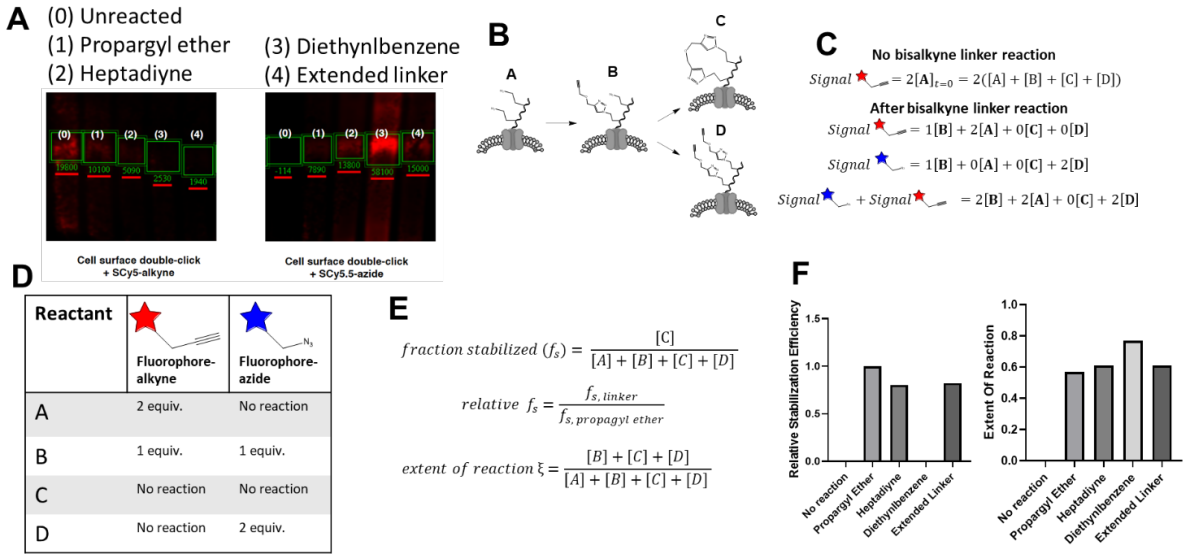


Figure 2.2: Efficiency of diverse bisalkyne reactions on bacterial cell surface. Bacteria displaying p53-like peptide are reacted with bisalkyne staple before treatment with either fluorophore-azide or fluorophore-alkyne. Bacteria with peptide but no reaction with bisalkyne staple are also prepared as a display level control. All samples are run on a SDS-PAGE gel (Thermo Fisher) and the bands according to the molecular weight of eCPX are quantified **(A)**. A mass balance is done on the azides or alkynes **(B,C,D)** on the cell surface and the signals are converted into fraction stabilized or extent of reaction **(E,F)**. Stabilization efficiency is reported as relative to propargyl ether.

2.1.3 Mdm2 library generation and sorting

A methionine codon-free version of eCPX (except for the start codon) was used as a PCR template for generating the NNC library as described previously.[23] After digestion and ligation, the library was transformed into electrocompetent methionine auxotrophic TYJV2 *E. coli* (a generous gift from J. van Deventer) achieving a library size of approximately 3 x 10^8 members.[27] TYJV2 cells were used for all sorting experiments. Bis-alkyne reacted cells, in tenfold excess of the library diversity, were first labelled with 18 nM mdm2-GST-biotin in 0.2% PBS/BSA. Cells were then washed once with PBS/BSA and incubated with 500 uL MyOne C1 beads (Thermo Fisher) for 25 min at 4 °C with gentle rotation using MACSmix (Miltenyi Biotec). Magnetic beads were pulled down by a DynaMag-5 magnet (Thermo Fisher) and

washed gently with 5 mL PBS/BSA. DNA was isolated from bead-bound cells using a Qiagen miniprep kit according to Ramesh et al.[90] Resulting DNA, generally 100-500 ng total, was transformed into fresh TYJV2 cells for additional sorting and analysis. For each linker library, 5 mL of cells were grown out, induced, and reacted as described above. Serial rounds of FACS were carried out with increasing stringency. For the first round of sorting, cells were incubated with 4 nM mdm2-GST-AF647 (collecting 2% brightest cells), second round 1 nM (collecting 0.5% brightest cells) , and the third and fourth rounds incubated with 1 nM of mdm2-GST-AF647 first and then 30 nM mdm2-GST-AF488 as described previously to select for tight binders regardless of display level (roughly 1% of cells collected).[23] Sorting was carried out in a MoFlo Astrios FACS instrument, and plasmids extracted and re-transformed into TYJV2 cells for further analysis and sorting if needed.

2.1.4 Mdm2 deep sequencing

Plasmids were isolated from bacterial pellets by miniprep (Qiagen). Illumina sequencing regions were added to either side of the eCPX-peptide gene by PCR amplification using Q5 DNA polymerase (New England Biolabs) following the manufacturer's protocol and primers 1-F and 1-R (see **Figure 2.13**) PCR products were cleaned by gel extraction and re-concentrated using a ZymoClean Clean & Concentrate kit. Another PCR amplification was performed also using Q5 to add the P5 and P7 Illumina sequences for flow cell annealing as well as a unique 8 letter barcode on each end of the amplicon for demultiplexing using primers 5-(0-7) and 7-(0-7).[91] The second round of PCR was cleaned and re-concentrated identically. DNA concentrations were quantified using a QuBit fluorimeter, pooled, and submitted to the University of Michigan DNA Advanced Genomics core for analysis. Samples were demultiplexed by filtering for samples with perfectly matched barcodes and ones that differed by up to one base pair. Filtered

reads were then analyzed with FastQC and samples with a PHRED score of less than 36 were discarded. Fastq files were then analyzed using custom Python scripts with Biopython and SeqIO packages. Forward and reverse reads were pairwise analyzed, discarding any sequences with differences in base pairs. The portion of the read that corresponds to the p53-based peptide was translated.

2.1.5 Synthesis and preparation of peptides

Bcl-2 and mdm2 peptides were synthesized using Fmoc chemistry on a CEM Liberty Blue microwave peptide synthesis instrument as described previously or obtained through the University of Michigan Proteomics and Peptide Synthesis Core.[23] The bis-alkyne stapling reaction was performed as described previously. Briefly, peptide in 0.1M $NaHCO_3$ was added to a 1:1:6:1.2 ratio of $CuSO_4$: THPTA: Sodium Ascorbate: bis-alkyne (propargyl ether, heptadiyne, or other) and reacted for 16 hours under gentle mixing at room temperature. Peptides were purified using reverse phase liquid chromatography on a C18 column using 0.1% TFA H_2O / acetonitrile gradient. Fractions were collected and lyophilized before analysis by mass spectrometry using ESI or MALDI through the University of Michigan Mass Spectrometry core. HPLC chromatograms, mass spectra, and tabulation of masses can be found in **Figure 2.14**, **Figure 2.15**, **Figure 2.16**, **Figure 2.17**, and **Table 2.1** respectively. Chemical structures for all compounds in the study are tabulated in **Figure 2.18**. Circular dichroism measurements and extinction coefficient calculations are available in **Figure 2.19**.

2.1.6 Circular Dichroism

Mdm2 or Bcl2 peptides were dissolved in 1:1 (v/v) H_2O: acetonitrile at approximately 0.1mg/mL. Peptides were added to a 3mL quartz cuvette and analyzed on a Jasco J-815 CD

Spectrometer at 100nm min^{-1} at 22°C. Data is reported as baseline-corrected with solvent blank. We used the BeStSel webserver to calculate alpha helicity.[92]

2.1.7 Biolayer Interferometry

Peptides and Bcl-2-biotin proteins were quantified using A_{280} measurements and added to 0.3% BSA in PBS pH 7.4. One well was prepared for each Bcl-2-biotin protein at 100-500nM and no dependence on concentration for sensor loading was observed. To obtain multiple binding curves for each peptide-protein interaction, 5 wells with concentrations varying from 10-1000nM of peptide were prepared along with 6 wells with 0.3% BSA in PBS for dissociation. An OctetRED 96 instrument with super streptavidin tips was used for all BLI experiments. The following times were used: load, 900 seconds; wash, 900 seconds; baseline, 60 seconds; association, 1200 seconds; dissociation, 3600 seconds. Data was analyzed using GraphPad Prism v10.0 using a single-phase association and dissociation for all data. Representative biolayer interferometry data

can be found in **Figure 2.20** and **Figure 2.3**.

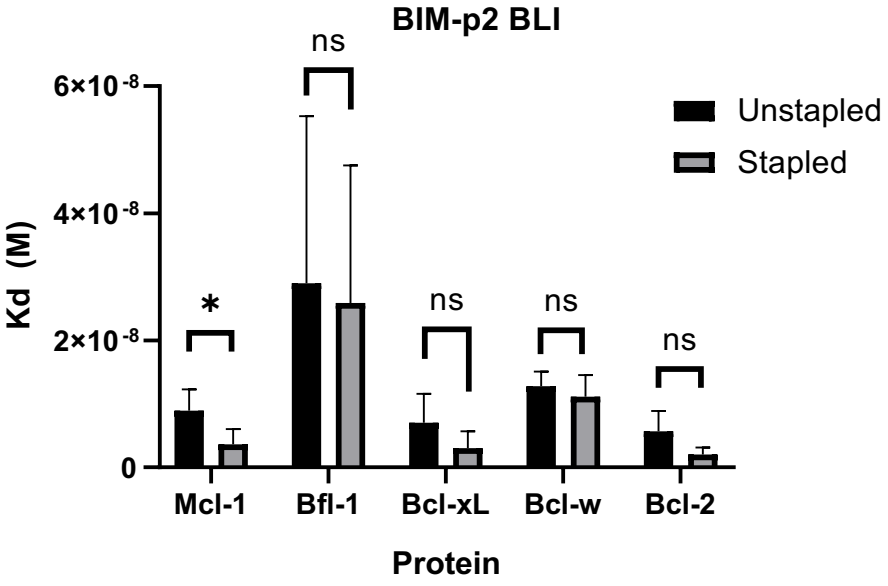


Figure 2.3: Fit K_d from biolayer interferometry. Error bars are standard deviations and significance was determined by paired t-test. *: $p<0.05$.

Results

2.1.8 The bacterial surface confirms hotspot residues via alanine scanning mutagenesis

Protein-protein interactions are known to be driven by a select subset of surface exposed residues known as 'hot-spot' residues, contributing up to 80% of the interaction strength.[43] Molecular recognition of p53-like-peptides towards mdm2 has long been known to be dominated by three hotspot residues: Phe^{19}, Trp^{23}, and Leu^{26}.[93] We tested the ability of the bacterial surface

to identify hot spot residues via alanine scanning, where each wild type amino acid was replaced by alanine (**Figure 2.4**).

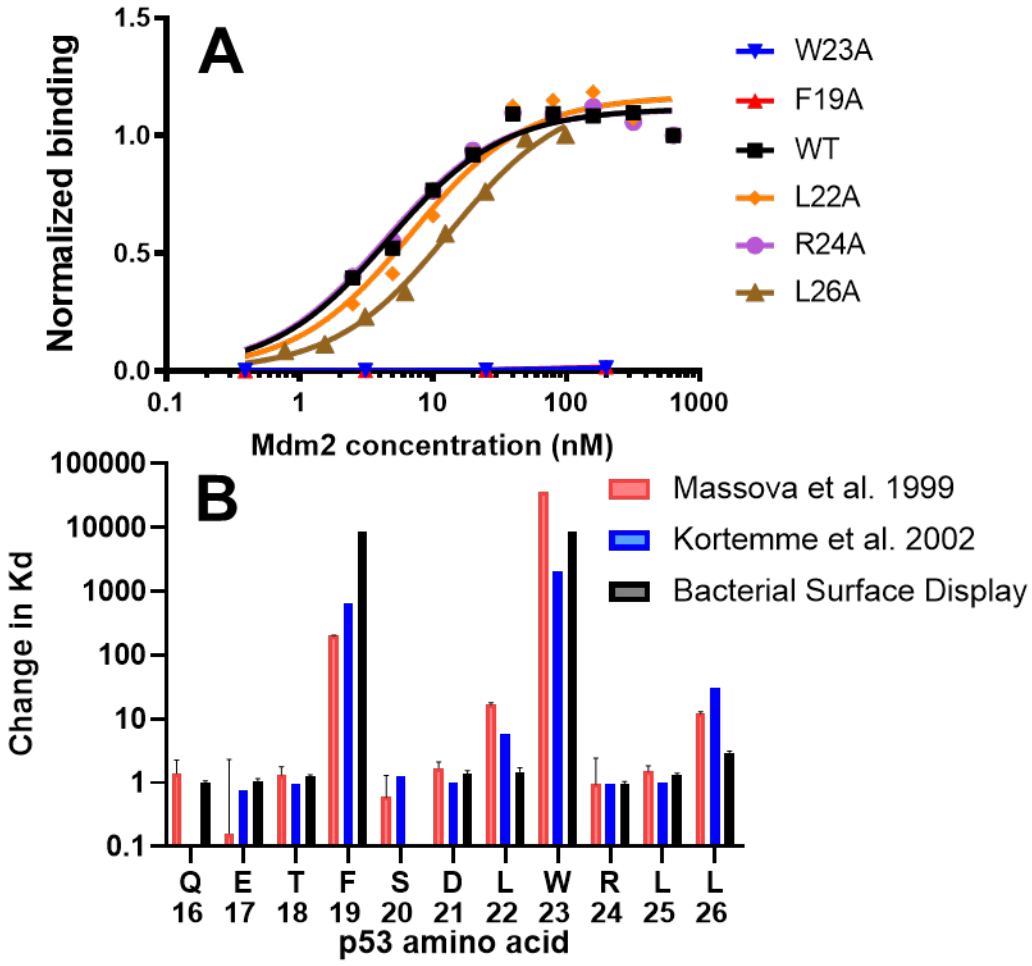


Figure 2.4: Hotspot identification via alanine scanning of the mdm2-p53 interaction on bacteria cell surface. **(A):** Select titrations of p53-like peptide alanine mutants on the bacterial surface. **(B):** Change in Kd from wild type p53-like peptide for each alanine mutant.

The complete set of titration curves for the alanine scanning mutagenesis data is located in **Figure 2.5**. These results were compared with molecular mechanics approaches[94] and statistical approaches[42] as validation for surface display of p53-like peptides. We generally found strong agreement between all three approaches; non-hotspot residues did not affect the affinity while F19A and W23A mutants did not bind at any measured concentration (where we display

>10,000 relative K_d). L26A yielded a smaller decrease in binding affinity, agreeing with the fact that Leu^{26} contributes less binding free energy than Phe^{19} or Trp^{23}. Hotspot residue identification is an important tool in the design of protein-protein interaction inhibitors as it can inform which sites are more amenable to optimization. The magnitude of binding affinity decrease can aid in the decision to preserve hotspot residues or mutate them with structurally similar groups.

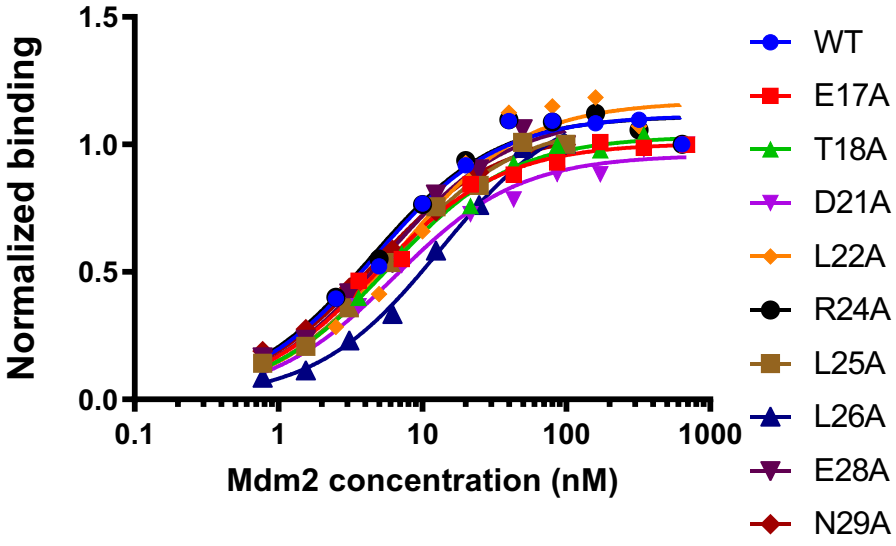


Figure 2.5: **Alanine scanning mutagenesis titration curves.**

2.1.9 Steric hindrance governs p53-like-peptide staple location

Beyond avoiding the disruption of hotspot residues, the design of a stapled peptide requires identification of an optimal staple location. This is an important design criterion, because an inappropriate staple location can cause steric clashes with the target protein, resulting in decreased affinity, and can influence the physicochemical properties (e.g. amphiphilicity) which impact the membrane permeability.[18] In contrast, forming new target interactions (e.g hydrogen bonds) and structurally stabilizing peptide residues are desired properties of a staple since they typically increase target affinity.[8,33,95,96] It is difficult to predict *a priori* whether a

given staple location will improve or decrease binding strength beyond high-level observations with crystal structures, if available. We hypothesized that peptide display on the bacterial surface could both re-confirm the importance of stapling location from previous work and identify steric factors that might play a role in abrogating binding by comparing the unstapled and stapled p53-like peptides. We started by modifying the p53-like peptide (PLP) with alternative stapling locations: PLP(1-8) and PLP(6-13), compared to the previously published 4-11 location (**Figure 2.6**).[76] Competitive inhibition curves can be found in **Error! Reference source not found.**. We selected these positions based on the alanine scanning mutagenesis data, as neither of these staple locations replace residues that are responsible for the core interaction of p53 and mdm2. These locations are additionally of interest as staples near the protein interface that have greater potential to form novel contacts than those solely exposed to the solvent.[95] Because the mutated residues do not form key contacts with mdm2, we hypothesized that these minimally invasive substitutions in the unstapled form would have little effect on binding. In contrast, the stapled form could exhibit steric hindrance and/or new contacts that may cause changes in affinity.

Indeed, affinity determination on the bacterial cell surface found that the 1-8 location was more staple-permissive than the 6-13 location, which weakened binding. Peptide alpha helicities calculated from circular dichroism spectrophotometry show that PLP(4-11) is considerably more alpha helical than the other PLP's (60% versus 5%). However, CD measurements also showed that for all PLP's, there were minimal changes in alpha helicity upon stapling. This suggests that the decrease in affinity for PLP(1-8) and PLP(6-13) as a result of stapling is not related to staple mediated secondary structure stabilization, further supporting the hypothesis that new steric hindrance effects drive weakened binding. Equilibrium constants from solution phase agreed

qualitatively with equilibrium measurements on the bacterial cell surface, confirming that SPEED is well equipped to perform staple scanning.

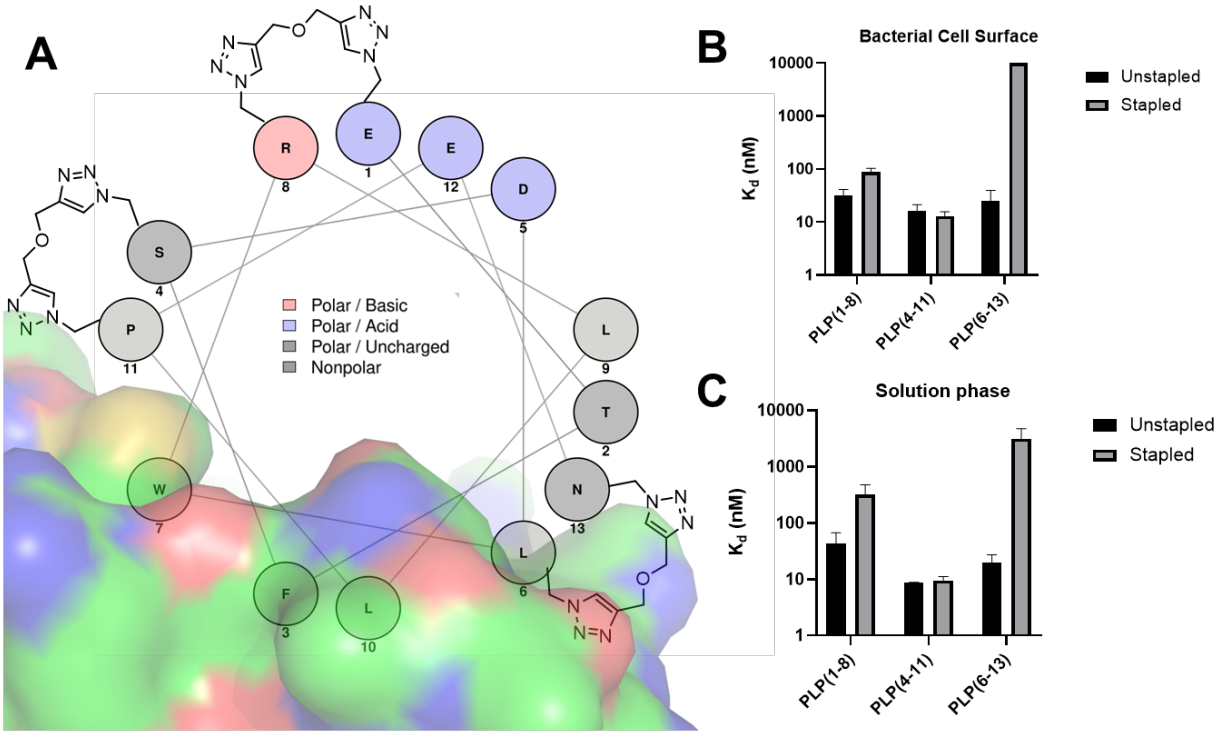


Figure 2.6: Staple scanning p53-like peptides using the bacterial cell surface. (A): Helix-wheel diagram of the p53-mdm2 interaction overlaid with the mdm2 crystal structure (PDB: 1YCR). K_d's of p53-like peptide staple scan mutants on bacterial cell surface **(B)** or solution phase **(C)**. PLP(4-11) data was adapted from ref. 24. Helix wheel diagrams were made with NetWheels.[47]

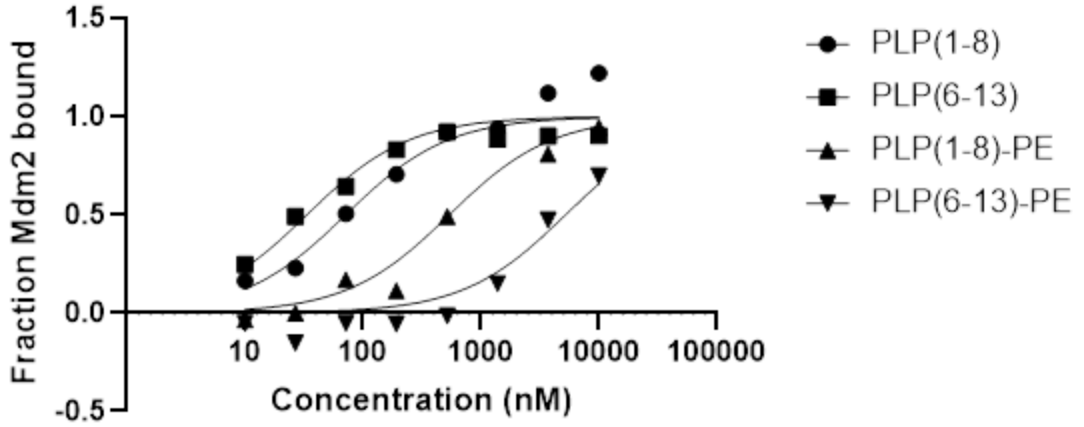


Figure 2.7 : Competitive inhibition experiments of p53-like peptides. Soluble p53-like peptides at various concentrations are incubated with 10nM biotinylated mdm2 until equilibration (> 4hr on ice). Then, bacteria expressing a high affinity peptide with known affinity are added for a short duration to capture unbound mdm2. Fluorescent signal arising from bacteria is inversely proportional to binding of soluble peptide. K_i are converted to K_d via the Cheng-Prusoff equation.

2.1.10 Engineering potent mdm2 binders with diverse bisalkyne linkers

We sought to explore the capabilities of SPEED to engineer structurally diverse peptides by varying the bisalkynes used in the stapling reaction. In our original work[23], we used propargyl ether to sort a randomized library, where three critical mdm2-binding residues (Phe19, Trp23, Leu26) as well as the two sites for stabilization (Aha20, Aha27) were kept fixed (**Figure 2.10**). All other sites were randomized by an NNC codon scheme permitting 15 possible amino acids and no stop codons at each position, for a theoretical diversity of $3x10^{10}$. We hypothesized that by repeating this process with different bisalkynes, we would obtain potent stapled peptides with diverse sequences and linkers. We selected three new bisalkyne staples - heptadiyne, a purely aliphatic staple; a PEG2 linker with a primary amine for functionalization; and (1,3)-diethynlbenzene, a non-flexible aromatic linker; and an unreacted control. [49,98] Randomized libraries were stabilized and then sorted by one round of magnetic sorting and four rounds of

fluorescent sorting with increasing stringency for mdm2 binding. After sorting, we deep sequenced the peptides from each of the bisalkyne libraries and identified sequence patterns that emerged. Data from the (1,3)-diethynlbenzene library indicated that the sequences failed to yield major consensus groups, likely arising from an incomplete reaction due to linker inflexibility (**Figure 2.8**).

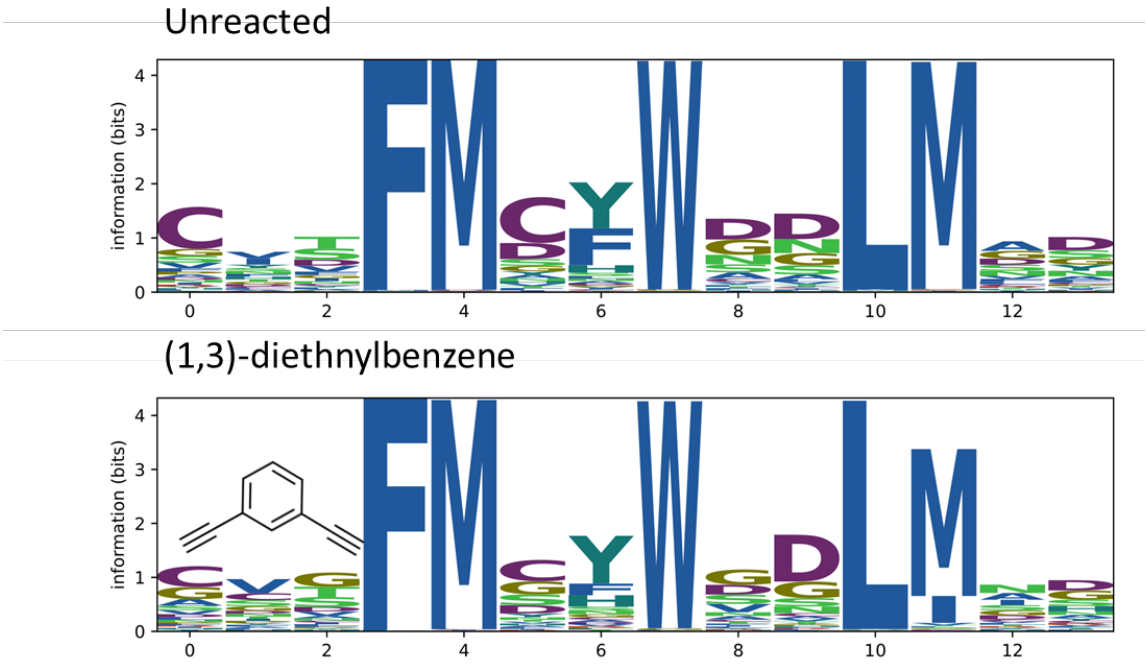


Figure 2.8: Logoplots of unreacted and (1,3)-diethynlbenzene p53-like peptides from fluorescent activated cell sorting.

In the other libraries, deep sequencing revealed a number of potential new dual cysteine motifs, potentially forming new topologies of i,i+1 and i,i+5 disulfide bonds expanding from the i,i+4 disulfide we confirmed in our previous lead molecule via nuclear magnetic resonance.[23] Deep sequencing data additionally yields insights into the proportion and enrichment of potential disulfide motifs; if disulfides are forming, frequencies of potential disulfide motifs should increase compared to those with single cysteine residues. When we calculate the frequency of

disulfide versus single cysteine residues, we see specific enrichment of potential disulfide motifs (**Figure 2.9**).

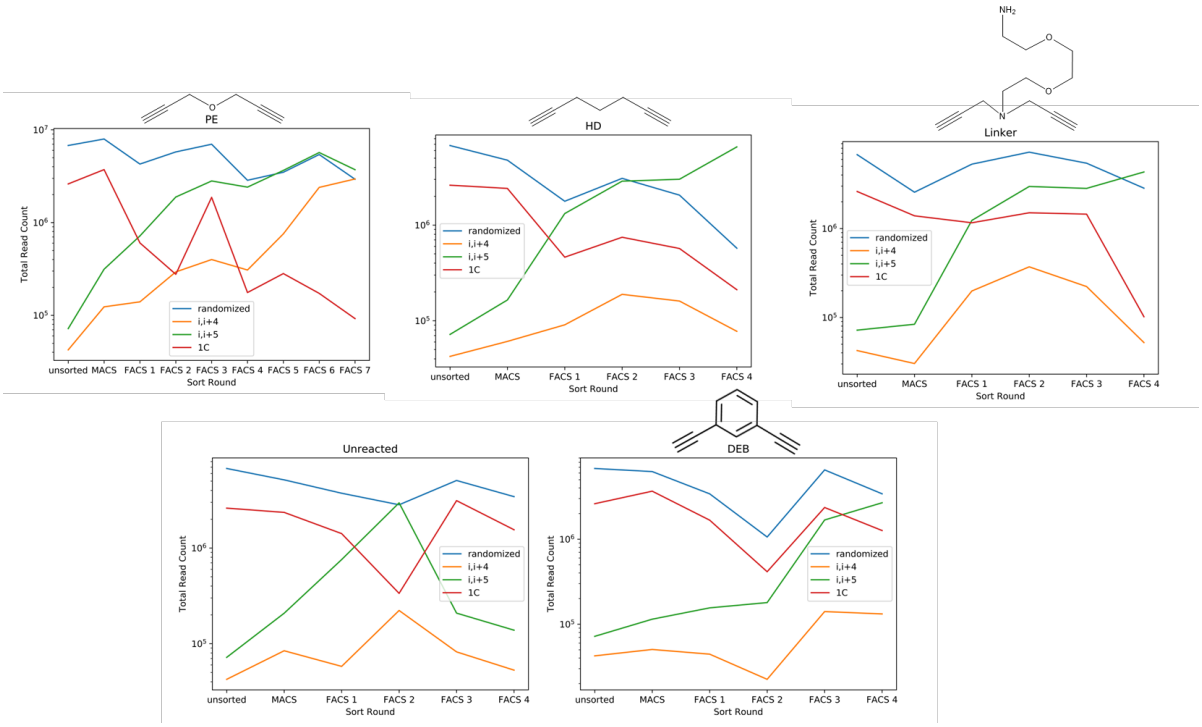
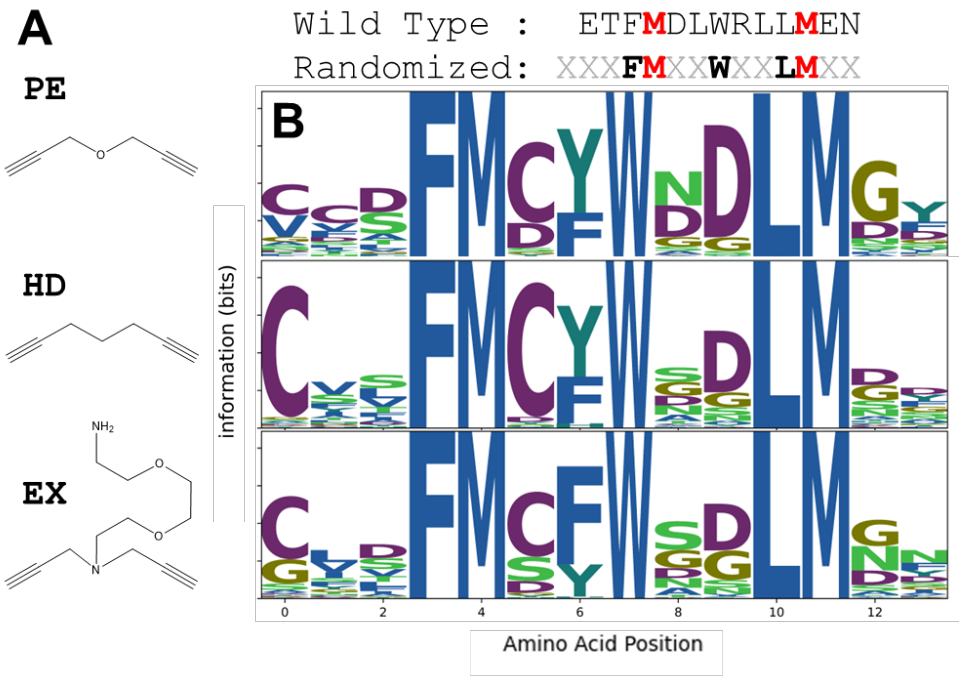


Figure 2.9: Enrichment of potential disulfide motifs compared to single cysteine peptides. Peptides sequenced from Mdm2 linker libraries were grouped according to their cysteine residues. 'Randomized' refers to any peptide matching our library design scheme, 'i,i+4' refers to any sequence with two cysteines separated by 3 residues, 'i,i+5' refers to any sequence with two cysteines separated by 4 residues, and '1C' refers to any sequence with a single cysteine.

This phenomenon was not observed as strongly in libraries that yielded worse enrichment ((1,3)-diethynlbenzene and the unreacted control). Next, we surveyed some of the most highly enriched sequences and performed low-throughput titrations of mdm2 to measure their binding affinity. Enrichment trajectories and all measured affinities are reported in **Figure 2.21** and

Figure 2.22 respectively. We observed that in each of the bisalkyne libraries, there were multiple peptides that demonstrated improved affinity compared to the wild type sequence.



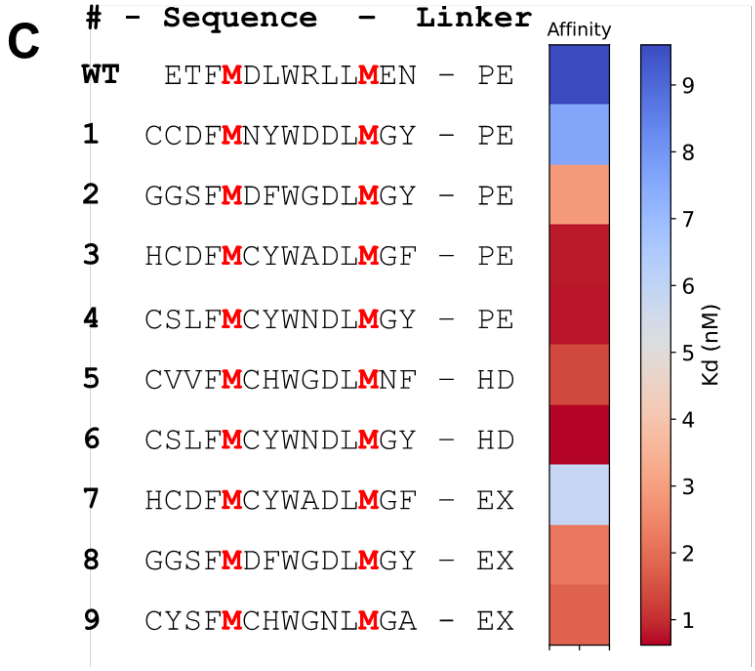


Figure 2.10: Engineering diverse bisalkyne stapled peptides. (A) A randomized library of p53-like peptides was displayed on the surface of bacteria, reacted with one of three bisalkynes, and subjected to one round of magnetic and four rounds of fluorescent sorting. After sorting, cells from each library were analyzed via next generation

sequencing **(B).** Select clones were picked based on frequency in the final library and their affinities were measured by titrating mdm2 on the bacterial cell surface **(C)**. Logoplots were made using Logomaker.[99]

2.1.11 Identification of optimal staple location in BH3 domains

To utilize bacterial surface display for the selection of optimal linker location, we varied the staple location rather than the staple structure while targeting the Bcl-2 family of proteins. The B cell lymphoma 2 class of proteins was chosen for multiple reasons. First, the alpha helical domain where Bcl-2 antagonists and Bcl-2 interact is much larger than that of p53-mdm2 and is therefore more expensive to screen staple locations using solid-phase peptide synthesis.[51] Second, there are several hotspot residues, like Leu^{3a} and Asp^{3f} that would provide convenient controls for non-functional mutants.[59] Finally, the Bcl-2 family is comprised of 5 highly homologous proteins that have varying levels of cellular expression and play different roles in apoptosis and resistance to chemotherapy.[35] A standing goal of Bcl-2-targeted therapeutics development is therefore to generate highly specific inhibitors.[37,40] While previous work has evaluated the impact of amino acid mutations on specificity,[33,38,47] we sought to investigate if and how the staple location can be used to improve specificity among Bcl-2 family proteins, which stabilized bacterial surface display is well suited to answer.

BIM, a naturally occurring BH3 domain with high promiscuity to all 5 Bcl-2 proteins, was selected as a scaffold for the staple scan to ensure the generalizability of staple location to different Bcl-2 proteins.[38] To investigate the effect of how staple location might impact binding affinity, we tested every potential location in a 23-length BH3 domain its effect on affinity to the Mcl-1 protein, the protein for which BIM has the highest affinity. (**Figure 2.11**) Select titration curves can be found in **Figure 2.23**. The bacterial cell surface identified seven different staple locations that did not completely abrogate binding. These results were consistent with known hotspot residues: Leu^{3a} , Gly^{3b}, and Asp^{3f}. Mutation of these residues resulted in a complete loss

of function (p8, p9, and p11).[33] We then measured the affinity of the BIM staple scan mutants to each of the other 4 Bcl-2 targets: Bfl-1, Bcl-x_L, Bcl-w, and Bcl-2. Interestingly, we found that the specificity trends of the wild type BIM peptide, which has a small degree of specificity (higher affinity) for Mcl-1, were not the same as its stapled variants. This suggests that the staple

location is playing a role in binding specificity and can serve as an additional handle for optimizing specificity.

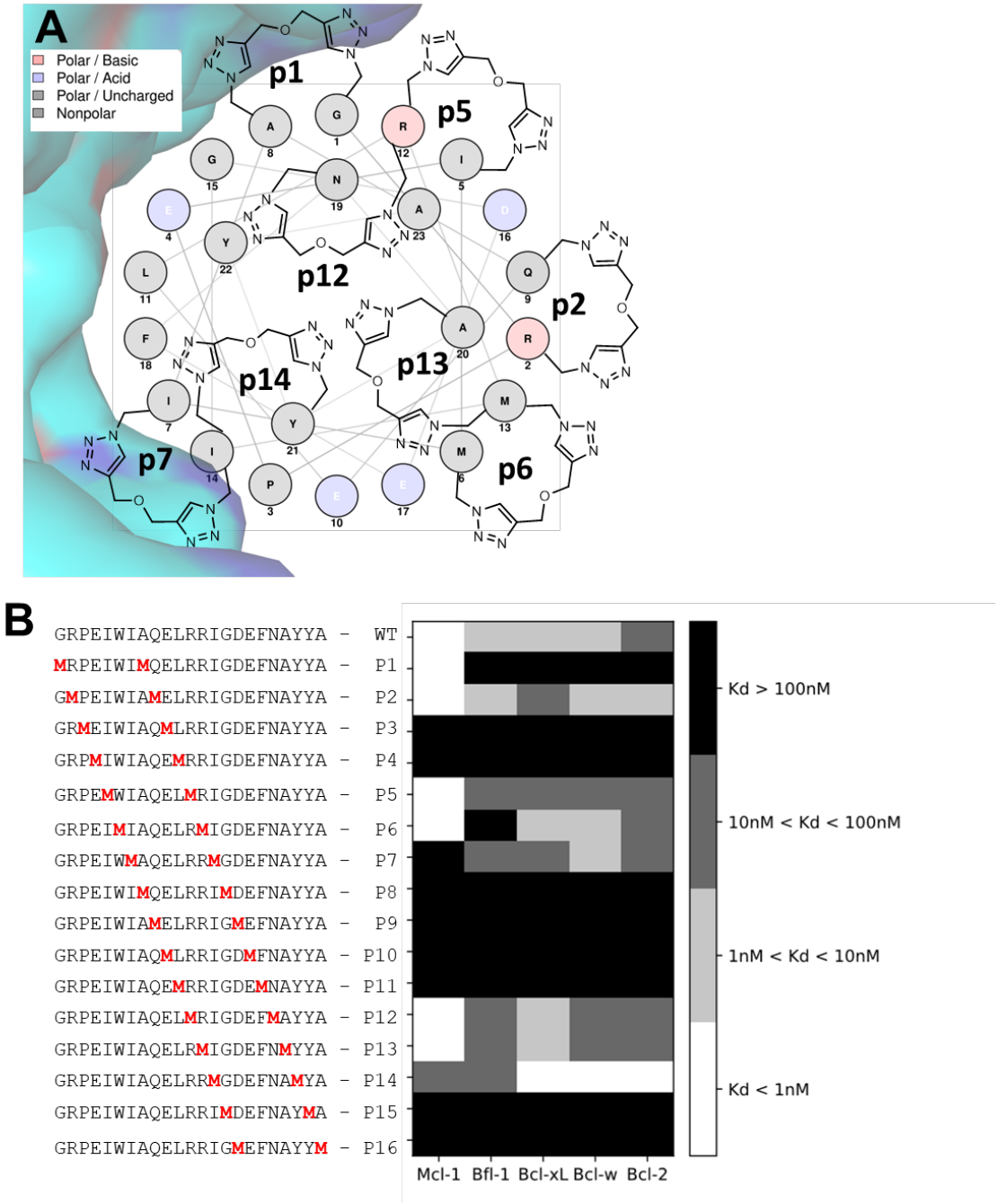


Figure 2.11: Modulating affinity and specificity of B cell lymphoma 2 peptide antagonists by staple location. **(A)** A helix-wheel diagram of the Bcl-2:BIM interaction overlaid with the Mcl-1 crystal structure (PDB: 2NL9) shows were the higher affinity variants were stapled. **(B)** The affinity and specificity of all BIM staple mutants was evaluated for all 5 Bcl-2 proteins.

2.1.12 Biolayer interferometry confirms bacterial surface display trends

We translated select p53-like peptides and stapled variants off the bacterial cell surface and measured their binding affinities to each of the Bcl-2 proteins to evaluate whether bacterial surface equilibrium association measurements matched kinetic rate constants from biolayer interferometry. First, we observe that both PLP(1-8) and PLP(6-13) in their unstapled form bind mdm2 as strongly as PLP(4-11), confirming our hypothesis that substitution to azidohomoalanine doesn't result in loss of function. We then measured the binding affinity of synthesized peptides in their unstapled and stapled forms and confirmed that stapling in either location weakened binding compared to their unstapled versions, to similar extents as with bacterial surface display. Overall, there was a strong correlation between BSD and BLI (pearson R^2 0.82 and $p < 0.0001$) measurements. The slope from linear regression does not significantly differ from 1 (p=0.38) nor does the y-intercept significantly differ from 0 (p=0.13) although we observed that generally the bacterial cell surface overestimated binding affinities 3-10 fold **(Figure 2.12).**

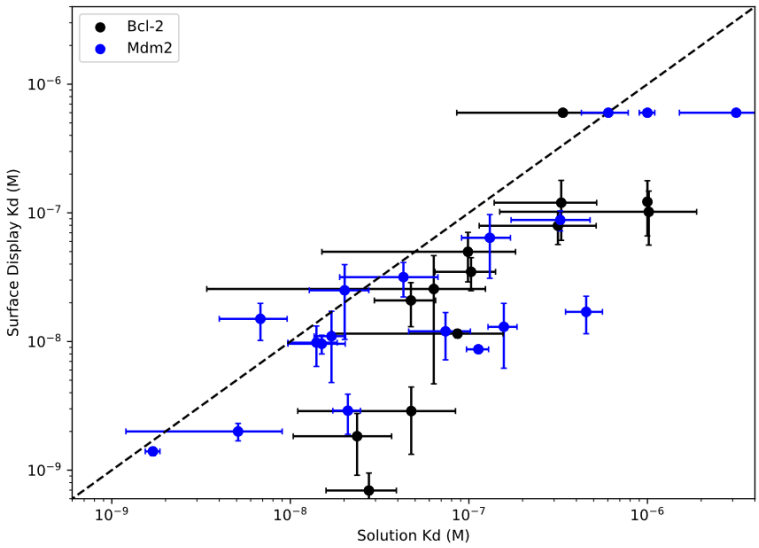


Figure 2.12: Solution phase peptide affinity measurement correlates with bacterial surface. BSD clones with measurable binding but not saturable binding are plotted as 600 nM.

Discussion

In this work, we used SPEED, which utilizes cell-surface stabilized peptides using non-natural amino acid incorporation and click chemistry (**Figure 2.1**), to measure the impact of staple chemistry and staple location on affinity peptides for mdm2 and the Bcl-2 family of proteins. We first confirmed that alanine scanning mutagenesis via bacterial surface display closely agrees with both experimental and computational approaches for hot spot identification.[42,94] The three major hotspot residues closely agreed with experimental and computational work (**Figure 2.4**). Next, we explored the landscape of stapled p53-like peptides (PLPs) with modified linker structure and location. We hypothesized there may be potent PLPs that have different staple locations and staple chemistries that maintain high binding affinity but allow different linker properties, such as greater charge/lipophilicity or a functional handle.[18,98] Previous work has demonstrated the importance of the linker properties and location.[5,8,18,33,40] Initial work in the optimization of stapled PLPs tested five locations of hydrocarbon staples that did not interfere with hot spot residues.[76]The binding affinity of these five variants spanned multiple orders of magnitude, and ultimately the authors focused on one staple location, PLP(4-11). This staple location improved affinity one hundred-fold over wild type, but it required additional mutations based on adding positive charge to have sufficient cytotoxicity. These mutations had the negative side effect of reducing the affinity fifty-fold. Chang et al. developed ATSP-7041 through rational design, resulting in a more potent version of PLP(4-11) with several mutations informed by linear phage display.[96] Aileron Therapeutics modified this molecule into ALRN-6924 which is currently in Phase 1b clinical trials for chemoprotection in breast cancer chemotherapy.[100] Using stabilized bacterial surface display, we tested PLP(1-8), PLP(4-11), and PLP(6-13) with triazole-based stapled peptides (**Figure 2.6**) and found that they had comparable

affinities to molecules engineered with hydrocarbon staples. This approach can accelerate the measurements of mutations effect on affinity and can also serve as a tool to probe protease stability by treating the cells with a protease that simulates *in vitro* or *in vivo* conditions. In previous work, we showed that surface display experiments with protease treatment correlated with those from solution, reducing the burden of experimental measurement by synthesis and evaluation *in vitro*.[21]

Aside from linker location, we investigated how bacterial surface display could be used to probe the importance of the staple's chemical properties. In many systems, it is difficult to simultaneously evaluate the contributions arising from multiple design criteria in high throughput, such as staple location, stapling chemistry, or amino acid mutations. Lau et al. investigated the effects of different triazole-based chemical linkers, and found PLPs with an aromatic linker, 1,3-diethynlbenzene, only had K_d values of greater than 1000nM with the 4-11 staple location, highlighting the complex trade-off between staple location and its chemical properties.[45] Because bacterial surface display has modularity for these components, we investigated its ability to design stapled peptides with variable sequence and staple chemistry. After generating a randomized library, reacting it with diverse bisalkyne linkers, and sorting for binding mdm2 with increasing stringency, we identified several new sequences that bind to mdm2 with high affinity (**Figure 2.10**). Importantly, these peptides each have unique sequences and staples with varying physicochemical properties such as isoelectric point, hydrophobicity, and potential intramolecular disulfide bond topology. Finally, the incorporation of the functionalizable linker **(3)** into high-affinity molecules could accelerate related tasks for peptide development such as inclusion of a fluorophore containing linker for imaging applications, a polyarginine motif for cellular penetration, or a ubiquitin ligase recruiting modality for formation

of a protease targeting chimera (PROTAC).[49,98,101] These results highlight the important trade-offs between affinity, cytosolic access, and cytotoxicity as a function of sequence and staple location and the need for a method that can easily explore sequence and staple design space. SPEED is a method that can quickly assay staple location, evaluate amino acid mutations, and translate to solution phase measurements for greater coverage of design space for potent peptides such as p53-like binders.

To establish the generalizability of SPEED, we expanded the system to design stapled variants of BIM, a high affinity but non-specific inhibitor of B cell lymphoma 2 (Bcl-2) proteins that regulate apoptosis. Recent work shows that yeast surface display coupled with machine learning and sequence optimization are efficient at generating highly specific linear peptides, but rational design was necessary to generate highly specific stapled peptides.[10,55] We sought to address this challenge by measuring the affinity and specificity of stapled BIM variants (**Figure 2.11**). We identified several staple locations that dramatically change the specificity profile of BIM. These results recapitulate many factors that have previously been identified about BIM-based peptides as well as identify new impacts of linker location. For example, staple locations that disrupt key hotspots abrogate binding as expected. We find that Bfl-1 has the lowest affinity across all variants evaluated, which agrees with Jenson et al. where the authors had to use a PUMA-based (**p**53 **u**pregulated **m**odulator of **a**poptosis) library rather than BIM to find potent inhibitors of Bfl-1 since BIM-based peptides had low affinities.[34] Similarly, we find that these variants have very high affinities towards Mcl-1 which is consistent with the sub-nanomolar affinity of BIM and the lack of interference between the staple and those high affinity interactions.[38] Finally, we see a high degree of correlation of affinity between Bcl-xL, Bcl-w, and Bcl-2, likely resulting from high structural homology between these three proteins.[30] We also

discovered variable specificity with different double-click linker locations. These BIM staple variants displaying altered specificity could serve as starting points for applications that rely on specific members within Bcl-2. Future work includes screening randomized libraries of Bcl-2 antagonists via SPEED using these specificity-driving staple location towards the discovery of high affinity, specificity, and efficacious Bcl-2 stapled peptide inhibitors.

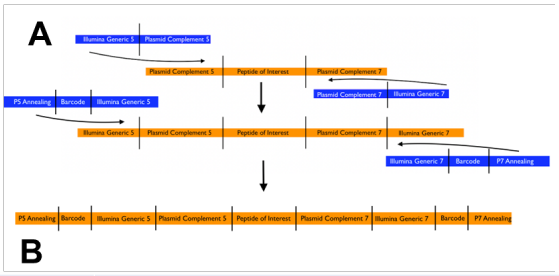
Finally, we translated PLP and BIM variants off the bacterial cell surface and evaluated their binding affinities in solution using biolayer interferometry and competitive inhibition experiments (**Figure 2.12**). While equilibrium affinity measurements from bacterial surface display highly correlate with those from solution phase, the bacterial surface tends to overestimate solution-phase binding affinities by 3 to 10-fold in our system. The exact cause of this discrepancy is unclear. We hypothesize that the molecular crowding on the bacterial cell surface improves the conformational stability of displayed peptides and results in lower measured K_d values relative to the same sequence in solution regardless of peptide stapling. This feature is more pronounced in PLPs than BIM variants as PLPs tend to be less structured in phosphate buffers or trifluoroethanol (a helix inducing solvent), likely due to their shorter length (13 AA vs 23).[53,102] Therefore, synthesis and evaluation of peptides in their soluble form remains an important step in the design of new stapled peptide inhibitors. Likewise, bacterial surface display may not be able to resolve small differences in affinity, which could be a limitation if molecules need to be tuned with high precision. However, in the design of Bcl-2 inhibitors, specificities on the order of 100-1000x are needed, which is well within the abilities of surface display to measure.[37] In conclusion, we have established the discovery of stapled peptides via bacterial surface display is a powerful method that can optimize sequence, staple location, and staple chemistry with respect to binding affinity and specificity.

Appendices

Table 2.1 contains calculated masses, observed masses, and extinction coefficients for all soluble peptides used in the study. **Figure 2.2** and other reference figures include information on bacterial surface display reaction efficiency; next generation sequencing primer and their reaction scheme and subsequent analysis; reaction, structures, and purification of all soluble peptides; circular dichroism; and biolayer interferometry and titration curves.

Table 2.1: Mass spectrometry calculated masses, observed masses, and molar extinction coefficients

		Unstapled			Stapled		
Name	Sequence	Predicted Exact Mass	Observed Exact Mass	Extinction Coefficient (1/M/cm)	Predicted Exact Mass	Observed Exact Mass	Extinction Coefficient (1/M/cm)
BIM	GRPEIWIAQELRRIGDEFNAYYA	2807.4	2807.4	8480	N/A	N/A	8480
BIM-p2	GXPEIWIAXELRRIGDEFNAYYA	2736.3	2736.3	8480	2828.4	2828.4	8480
BIM-p5	GRPEXWIAQELXRIGDEFNAYYA	2790.3	2790.3	8480	2884.4	2884.4	8480
BIM-p6	GRPEIXIAQELRXIGDEFNAYYA	2718.3	2718.3	2980	2812.4	2812.4	2980
BIM-p7	GRPEIWXAQELRRXGDEFNAYYA	2834.3	2834.4	8480	2928.4	2928.4	8480
BIM-p14	GRPEIWIAQELRRXGDEFNAXYA	2784.4	2784.4	6990	2878.4	2878.4	6990
PLP(4-11)	ETFXDLWRLLXEN	1728.9	1728.9	5500	1821.9	1821.9	5500
PLP(1-8)	ETFSDXWRLLPEX	1626.8	1626.8	5500	1720.8	1720.8	5500
PLP(6-13)	XTFSDLWXLLPEN	1684.8	1684.8	5500	1778.9	1778.7	5500



1st round forward	TCG TCG GCA GCG TCA GAT GTG TAT AAG AGA CAG CAG GTA CTT CCG TAG CTG GCC AGT CT
1st reverse	GTC TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GCA CCG TAG ATG CTT GCC CAG TCG TTA
NGS 5-0	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT AGA TCG CTC GTC GGC AGC GTC
NGS 5-1	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACC TCT CTA TTT CGT CGG CAG CGT C
NGS 5-2	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT ATC CTC TGT TCG TCG GCA GCG TC
NGS 5-3	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACA GAG TAG ACG ATC GTC GGC AGC GTC
NGS 5-4	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACG TAA GGA GAT GAT CGT CGG CAG CGT C
NGS 5-5	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACA CTG CAT ATG CGA TCG TCG GCA GCG TC
NGS 5-6	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACA AGG AGT AGA GTG GTC GTC GGC AGC GTC
NGS 5-7	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACC TAA GCC TCC TGT GGT CGT CGG CAG CGT C
NGS 7-0	CAA GCA GAA GAC GGC ATA CGA GAT TAA GGC GAG TCT CGT GGG CTC GG
NGS 7-1	CAA GCA GAA GAC GGC ATA CGA GAT CGT ACT AGA GTC TCG TGG GCT CGG
NGS 7-2	CAA GCA GAA GAC GGC ATA CGA GAT AGG CAG AAT CGT CTC GTG GGC TCG G
NGS 7-3	CAA GCA GAA GAC GGC ATA CGA GAT TCC TGA GCC TAG TCT CGT GGG CTC GG
NGS 7-4	CAA GCA GAA GAC GGC ATA CGA GAT GGA CTC CTG ATA GTC TCG TGG GCT CGG
NGS 7-5	CAA GCA GAA GAC GGC ATA CGA GAT TAG GCA TGA CTC AGT CTC GTG GGC TCG G
NGS 7-6	CAA GCA GAA GAC GGC ATA CGA GAT CTC TCT ACT TCT CTG TCT CGT GGG CTC GG
NGS 7-7	CAA GCA GAA GAC GGC ATA CGA GAT CAG AGA GGC ACT TCT GTC TCG TGG GCT CGG

C

BIM	
p1-1	CAGGAATTACGCCGCATTGGTGATGAATTTAATGCGTACTACGCCGGCGGTGGCAGCGGA
p1-2	GCTGGCCAGTCTGGCCAGATGCGTCCCGAGATTTGGATCATGCAGGAATTACGCCGCATT
p2-1	ATGGAACTGCGCCGCATTGGCGATGAATTTAACGCGTATTATGCGGGCGGTGGCAGCGGA
p2-2	GCTGGCCAGTCTGGCCAGGGCATGCCGGAAATTTGGATTGCGATGGAACTGCGCCGCATT
p3-1	CAAATGCTGCGCCGCATTGGCGATGAGTTTAACGCATATTATGCTGGCGGTGGCAGCGGA
p3-2	GCTGGCCAGTCTGGCCAGGGTCGCATGGAAATTTGGATTGCCCAAATGCTGCGCCGCATT
p4-1	CAGGAGATGCGCCGCATCGGTGACGAATTTAACGCATACTATGCGGGCGGTGGCAGCGGA
p4-2	GCTGGCCAGTCTGGCCAGGGTCGTCCAATGATTTGGATCGCGCAGGAGATGCGCCGCATC
p5-1	CAGGAACTGATGCGCATTGGCGATGAATTTAACGCGTATTATGCGGGCGGTGGCAGCGGA
p5-2	GCTGGCCAGTCTGGCCAGGGCCGCCCGGAAAATGTGGATTGCGCAGGAACTGATGCGCATT
p6-1	CAGGAGCTGCGTATGATTGGGGATGAATTTAACGCATACTATGCGGGCGGTGGCAGCGGA
p6-2	GCTGGCCAGTCTGGCCAGGGCCGTCCTGAAATCATGATTGCCCAGGAGCTGCGTATGATT
p7-1	CAGGAATTACGTCGTATGGGCGACGAGTTCAATGCCTACTACGCCGGCGGTGGCAGCGGA
p7-2	GCTGGCCAGTCTGGCCAGGGACGCCCCGAAATCTGGATGGCACAGGAATTACGTCGTATG
p8-1	CAGGAACTGCGCCGCATTATGGATGAATTTAACGCGTATTATGCGGGCGGTGGCAGCGGA
p8-2	GCTGGCCAGTCTGGCCAGGGCCGCCCGGAAATTTGGATTATGCAGGAACTGCGCCGCATT
p9-1	ATGGAGTTACGTCGTATCGGAATGGAATTTAATGCTTACTATGCAGGCGGTGGCAGCGGA
p9-2	GCTGGCCAGTCTGGCCAGGGTCGTCCGGAAATTTGGATCGCGATGGAGTTACGTCGTATC
p10-1	CAGGAACTGCGCCGCATTGGCATGGAATTTAACGCGTATTATGCGGGCGGTGGCAGCGGA
p10-2	GCTGGCCAGTCTGGCCAGGGCCGCCCGGAAATTTGGATTATGCAGGAACTGCGCCGCATT
p11-1	CAGGAGATGCGCCGTATTGGCGATGAGATGAACGCATACTATGCGGGCGGTGGCAGCGGA
p11-2	GCTGGCCAGTCTGGCCAGGGACGTCCCGAAATTTGGATCGCACAGGAGATGCGCCGTATT
p12-1	CAAGAATTGATGCGTATTGGCGACGAATTTATGGCTTATTACGCAGGCGGTGGCAGCGGA
p12-2	GCTGGCCAGTCTGGCCAGGGCCGTCCAGAAATCTGGATTGCGCAAGAATTGATGCGTATT
p13-1	CAGGAACTTCGTATGATCGGCGACGAATTTAATATGTACTACGCTGGCGGTGGCAGCGGA
p13-2	GCTGGCCAGTCTGGCCAGGGGCGTCCGGAGATCTGGATCGCTCAGGAACTTCGTATGATC
p14-1	CAGGAACTGCGCCGCATGGGCGATGAATTTAACGCGATGTATGCGGGCGGTGGCAGCGGA
p14-2	GCTGGCCAGTCTGGCCAGGGCCGCCCGGAAATTTGGATTGCGCAGGAACTGCGCCGCATG
p15-1	CAGGAGCTGCGCCGTATTATGGACGAATTCAATGCTTACATGGCTGGCGGTGGCAGCGGA
p15-2	GCTGGCCAGTCTGGCCAGGGACGCCCCGAGATCTGGATCGCGCAGGAGCTGCGCCGTATT
p16-1	CAGGAATTGCGCCGCATTGGAATGGAGTTTAACGCCTATTATATGGGCGGTGGCAGCGGA
p16-2	GCTGGCCAGTCTGGCCAGGGACGTCCCGAAATTTGGATCGCACAGGAATTGCGCCGCATT
WT-1	CAGGAATTACGCCGCATTGGTGATGAATTTAATGCGTACTACGCCGGCGGTGGCAGCGGA
WT-2	GCTGGCCAGTCTGGCCAGGGTCGTCCCGAGATTTGGATCGCCCAGGAATTACGCCGCATT
pqeB0L reverse primer	GAGGTCATTACTGGATCTATCAACAGGAGTCCAAGCTCAGC
PLP alanine scan	
E17A	TCCGTAGCTGGCCAGTCTGGCCAGGCAACCTTTATGGATCTGTGGCGCCTGC
T18A	GTAGCTGGCCAGTCTGGCCAGGAAGCCTTTATGGATCTGTGGCGCCTGCTGATG
F19A	GCTGGCCAGTCTGGCCAGGAAACCGCTATGGATCTGTGGCGCCTGCTGAT
D21A	GGCCAGTCTGGCCAGGAAACCTTTATGGCTCTGTGGCGCCTGCTGATGGAAAAC
L22A	GGCCAGTCTGGCCAGGAAACCTTTATGGATCGTGGCGCCTGCTGATGGAAAAC
W23A	GCCAGTCTGGCCAGGAAACCTTTATGGATCTGGCGCGCCTGCTGATGGAAAACGGCGGT
R24A	AGTCTGGCCAGGAAACCTTTATGGATCTGTGGGCCCTGCTGATGGAAAACGGCGGTGGCA
L25A	GGCCAGGAAACCTTTATGGATCTGTGGCGCGCGCTGATGGAAAACGGCGGTGGCA
L26A	CAGGAAACCTTTATGGATCTGTGGCGCCTGGCGATGGAAAACGGCGGTGGCAGCGGA
E28A	ACCTTTATGGATCTGTGGCGCCTGCTGATGGCAAACGGCGGTGGCAGCGGAGGGCAGT
N29A	GGATCTGTGGCGCCTGCTGATGGAAGCCGGCGGTGGCAGCGGAGGGCAGTCT
PLP(1-8)	ACTTCCGTAGCTGGCCAGTCTGGCCAGATGACCTTTAGCGATCTGTGGATGCTGCTGCCGGAAAACGGCGGTGGCAGCG GAGGGCAGTCTGGGCAGTCTG
PLP(4-11)	ACTTCCGTAGCTGGCCAGTCTGGCCAGGAAACCTTTATGGATCTGTGGCGCCTGCTGATGGAAAACGGCGGTGGCAGCG GAGGGCAGTCTGGGCAGTCTG
PLP(6-13)	ACTTCCGTAGCTGGCCAGTCTGGCCAGGAAACCTTTAGCGATATGTGGCGCCTGCTGCCGGAAATGGGCGGTGGCAGCG GAGGGCAGTCTGGGCAGTCTG

Figure 2.13: Next generation sequencing (Illumina NovaSeq) reaction scheme and primers. Next generation sequence **(A),** its primers **(B),** and all primers used for generation of bacterial cell surface constructs **(C)**.

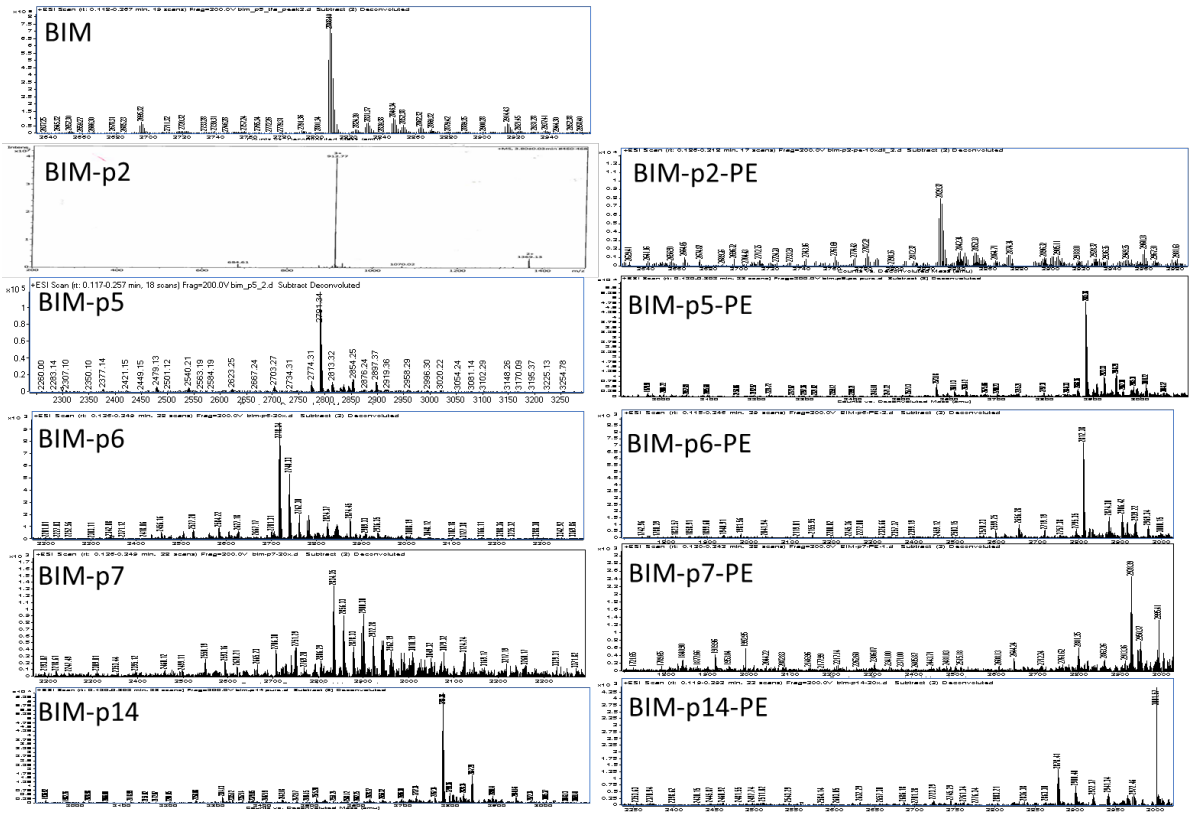


Figure 2.14: B cell lymphoma 2 peptides mass spectra

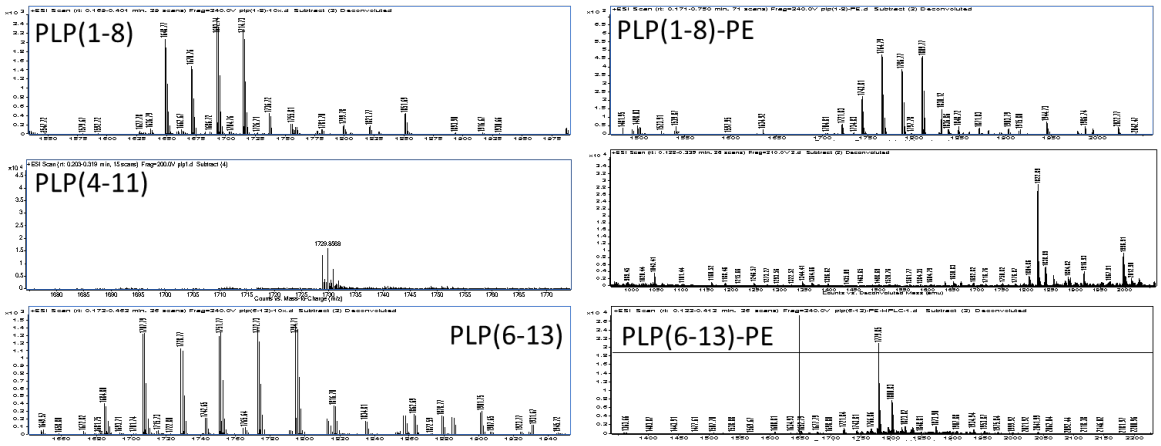


Figure 2.15: p53-like peptide mass spectra

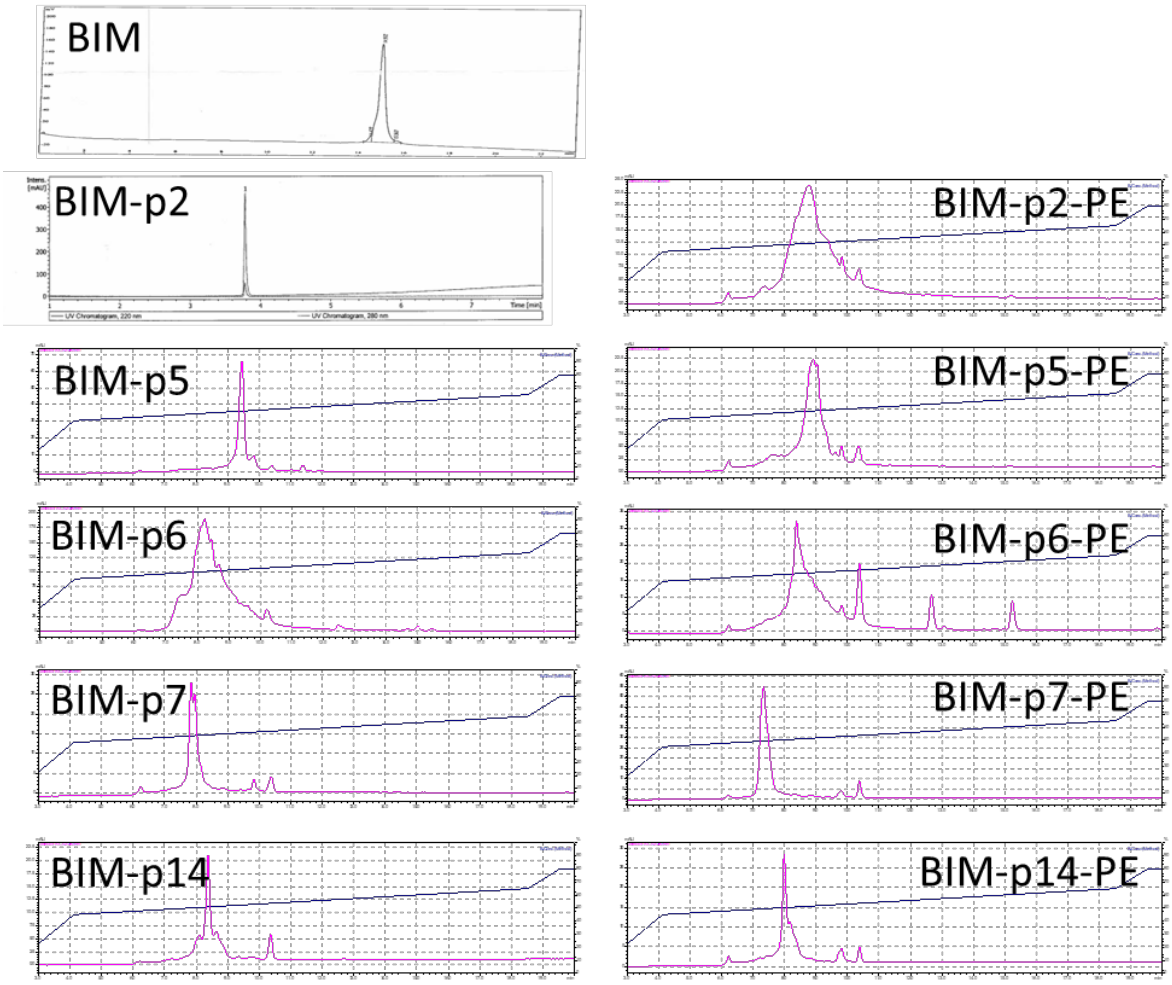


Figure 2.16: B cell lymphoma 2 chromatograms

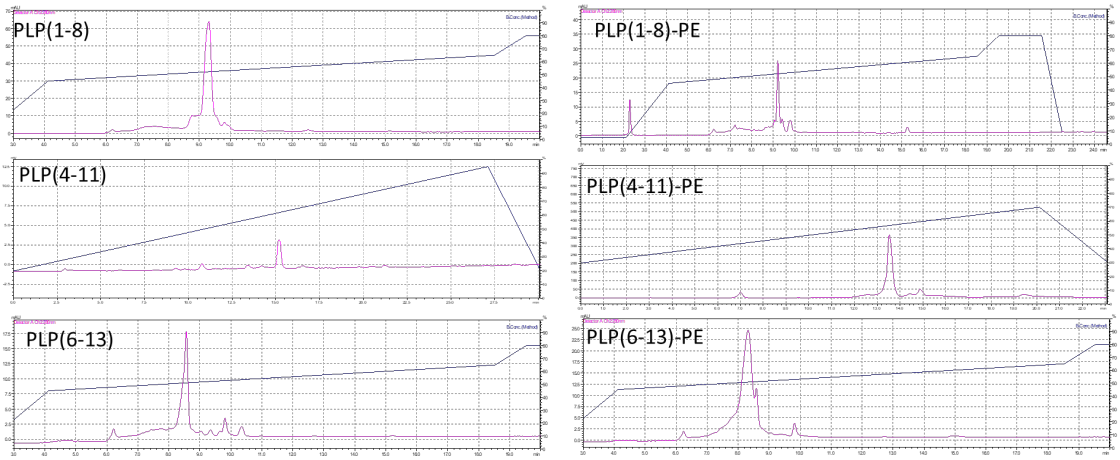


Figure 2.17: p53-like peptide chromatograms

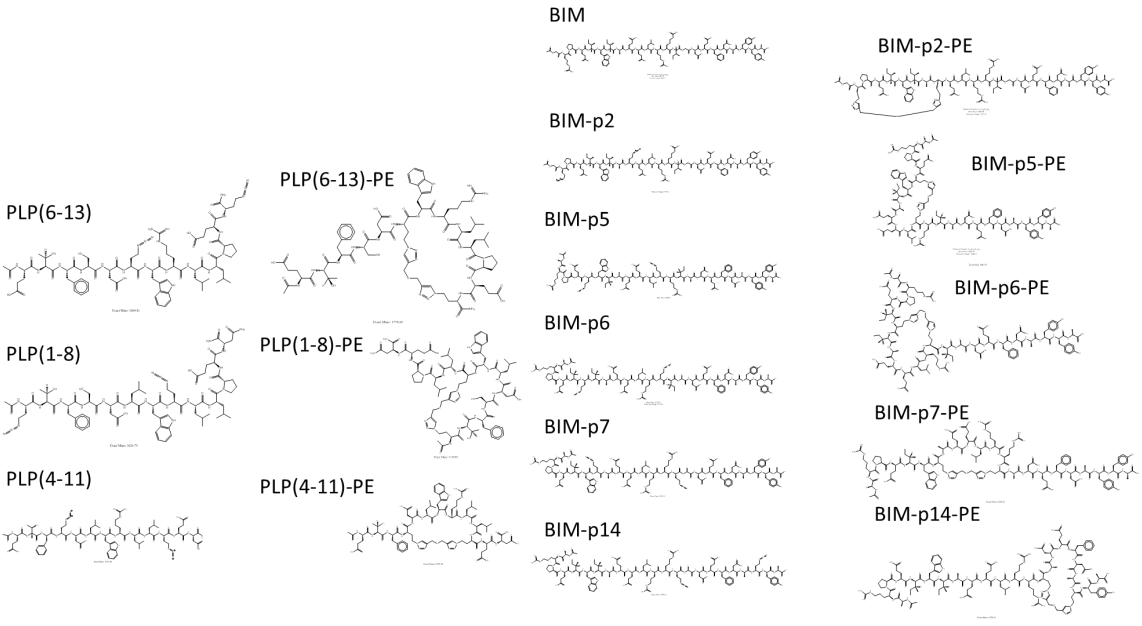
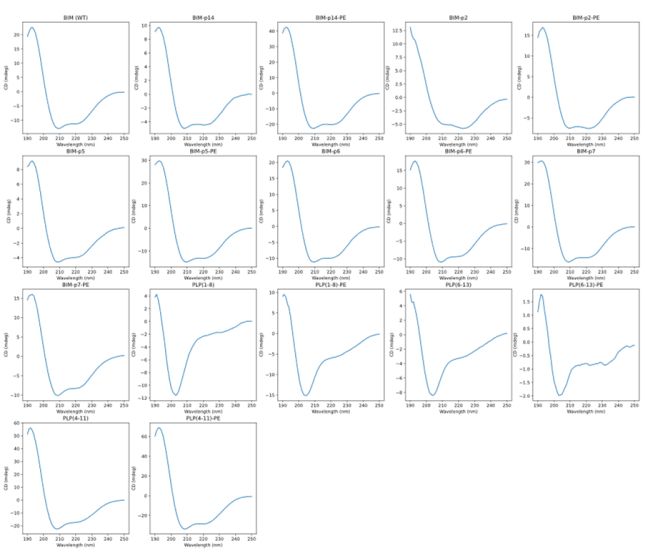


Figure 2.18: Chemical structures of peptides used in this study



	Extinction Coefficient (1/M/cm)			
Peptide	**205nm**	**214nm**	**280nm**	**Alpha Helicity**
BIM	105820	69118	8480	31.0
BIM-p2	109080	70891	8480	3.5
BIM-p2-PE	109080	70891	8480	45.2
BIM-p5	109480	70988	8480	26.9
BIM-p5-PE	109481	70988	8480	40.4
BIM-p6	89080	41983	2980	54.7
BIM-p6-PE	89080	41983	2980	33.8
BIM-p7	110830	71045	8480	23.1
BIM-p7-PE	110830	71045	8480	17.7
BIM-p14	104750	65715	6990	18.8
BIM-p14-PE	104750	65715	6990	36.2
PLP(1-8)	66420	50443	5500	1.9
PLP(1-8)-PE	66420	50443	5500	4.8
PLP(4-11)	67770	47914	5500	59.5
PLP(4-11)-PE	67770	47914	5500	55.0
PLP(6-13)	58150	46636	5500	2.0
PLP(6-13)-PE	58150	46636	5500	0.0

Figure 2.19: Circular dichroism, alpha helicity, and calculated extinction coefficients measurements for all compounds in this study

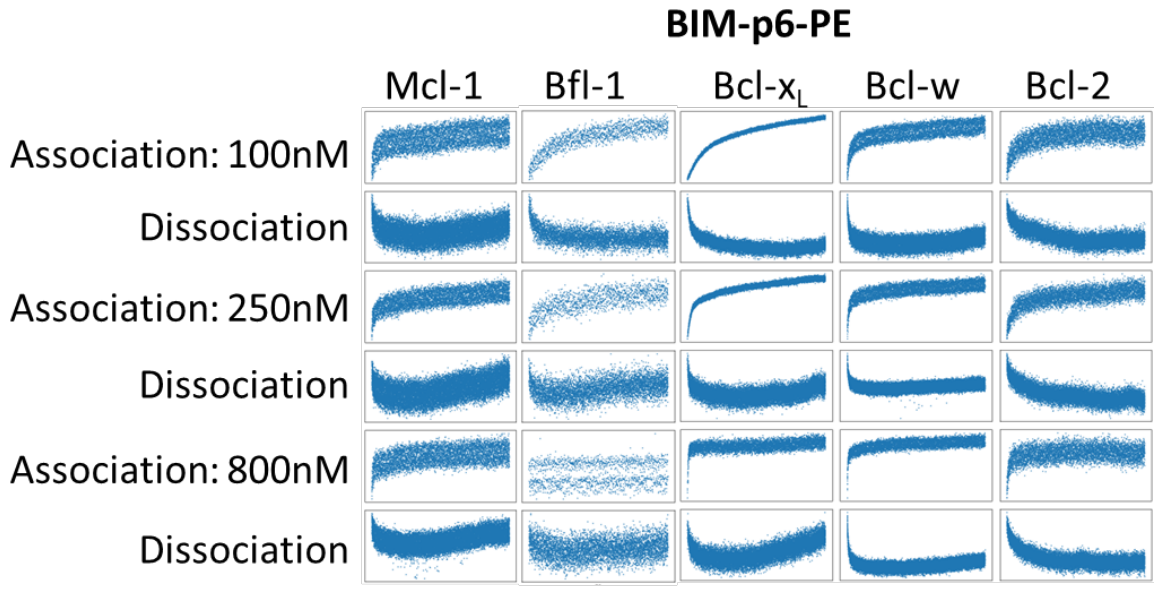


Figure 2.20: Representative biolayer interferometry data

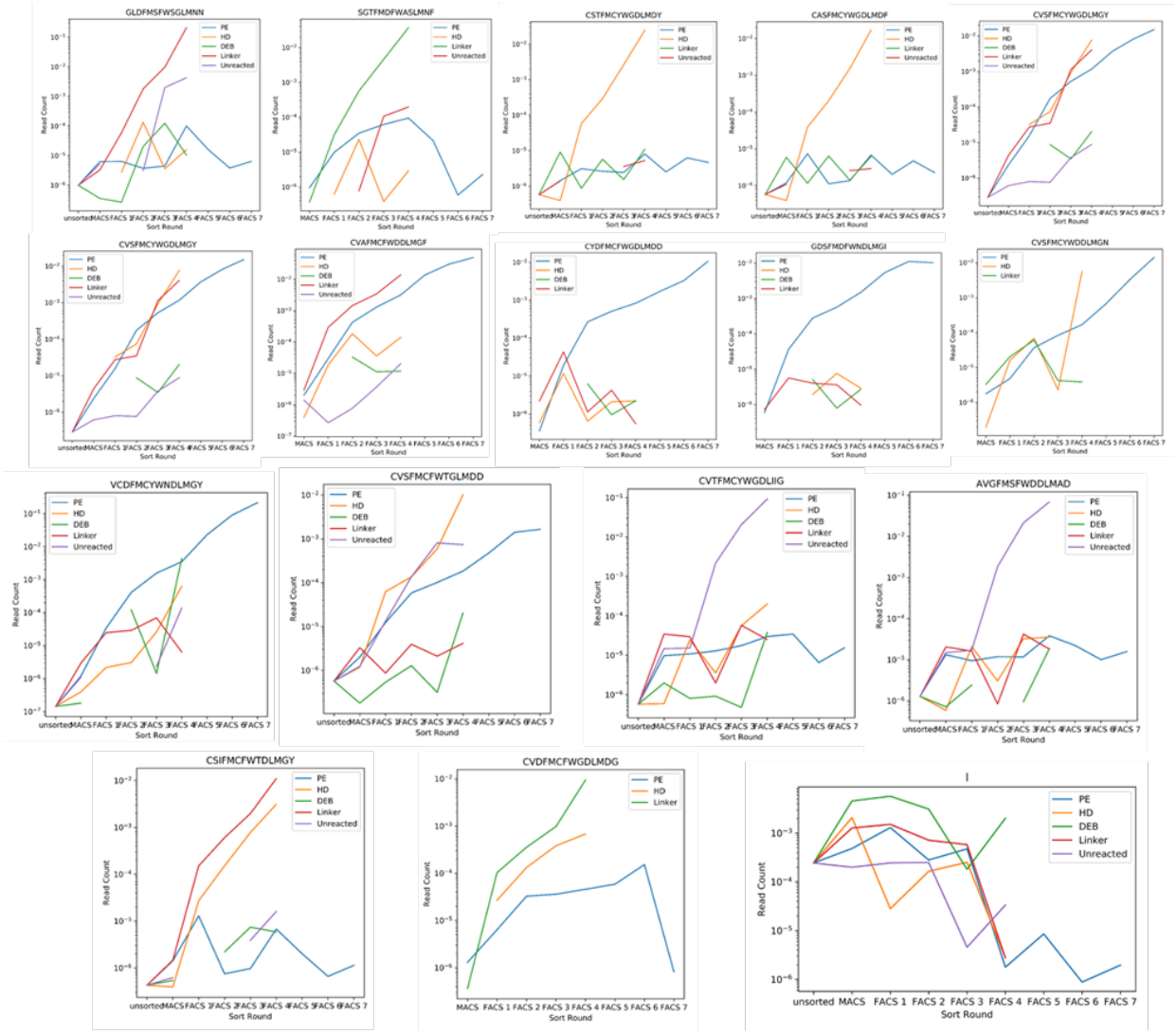


Figure 2.21: Enrichment trajectories of select p53-like peptides from fluorescent activated cell sorting. Individual peptides chosen for analysis and their enrichment over rounds of deep sequencing. Gaps in the plotted frequencies are places where no sequences were observed (either due to a lack of read depth or a complete loss of enrichment).

#	Sequence	Kd PE	Kd HD	Kd ExLinker
1	GCDFMCYWDDLMGI	nd	nd	2.62
2	CCDFMNYWDDLMGY	7.6	9.8	2.24
3	GASFMNFWDDLMGY	6.5	10.5	3.23
4	CIVFMCFWTDLMAH	4.56	1.06	2.32
5	CVVFMCHWGDLMNF	1.51	1.36	4.62
6	RGFFMDYWSGLMAD	0.85	1.6	40.1
7	CYSFMCHWGNLMGA	nd	nd	1.8
8	CSLFMCYWAGLMGG	1.77	1.86	13.7
9	CSLFMCYWNDLMGY	0.78	0.62	9.42
10	HCDFMCYWADLMGF	0.806	1.56	5.88
11	GGSFMDFWGDLMGY	2.9	0.71	2.18
12	CLLFMCFWNDLMGV	3.7	4.3	3.29
13	CSIFMCFWTDLMGA	6	1.2	3.6
14	CSIFMCFWTDLMGN	1.34	1.2	5.47
15	CVDFMCFWGDLMGV	0.55	0.57	3.84
18	CVVFMCFWGDLMGD	3.1	0.55	4.42

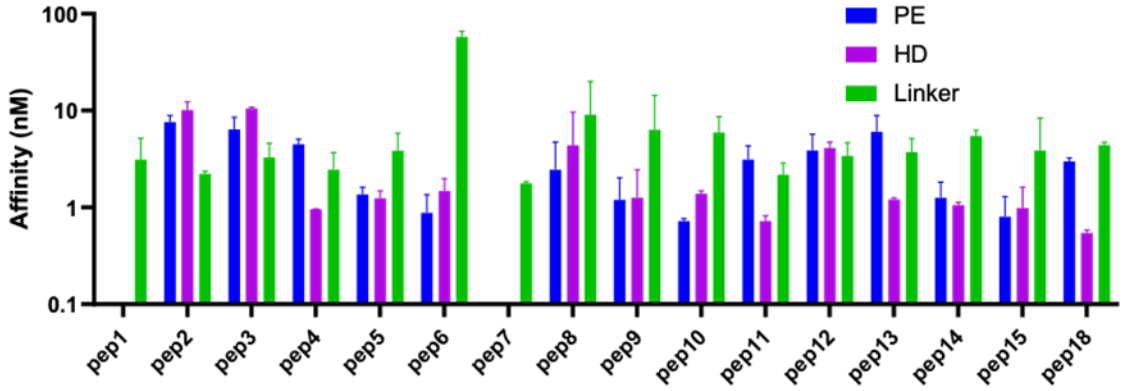


Figure 2.22: Affinities of select p53-like peptides from fluorescent activated cell sorting to mdm2. All flow experiments were performed with 8 concentrations with 3 replicates. K_d's were fit using GraphPad Prism v8.0.

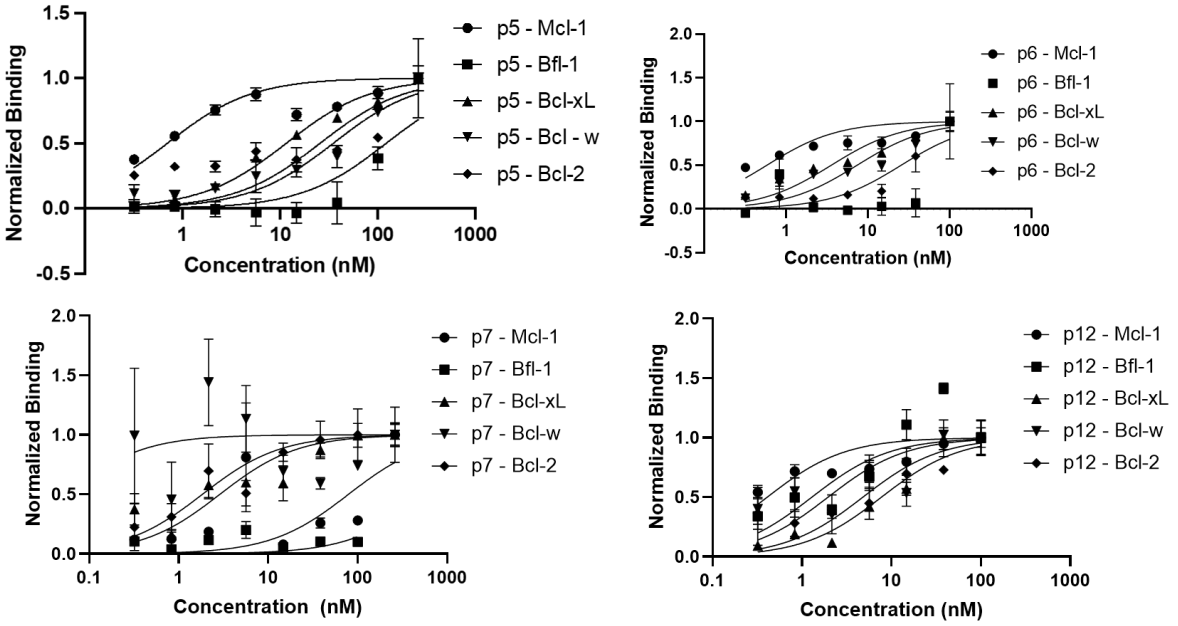


Figure 2.23: Titration of select bcl-2 peptides with all 5 Bcl-2 proteins via bacterial cell surface. All flow experiments were performed with 7 or 8 concentrations with 3 replicates. K_d's were fit using GraphPad Prism v8.0.

References

1.

1. Hopkins, A. L. & Groom, C. R. The druggable genome. *Nat Rev Drug Discov* **1**, 727–730 (2002).
2. Loren D. Walensky *et al.* Activation of Apoptosis in Vivo by a Hydrocarbon-Stapled BH3 Helix. *Science (1979)* **23**, 1–7 (2004).
3. Schafmeister, C. E., Po, J. & Verdine, G. L. An all-hydrocarbon cross-linking system for enhancing the helicity and metabolic stability of peptides. *J Am Chem Soc* **122**, 5891–5892 (2000).
4. Bluntzer, M. T. J., O'Connell, J., Baker, T. S., Michel, J. & Hulme, A. N. Designing stapled peptides to inhibit protein-protein interactions: An analysis of successes in a rapidly changing field. *Peptide Science* vol. 113 Preprint at https://doi.org/10.1002/pep2.24191 (2021).
5. Walensky, L. D. & Bird, G. H. Hydrocarbon-stapled peptides: principles, practice, and progress. *J Med Chem* **57**, 6275–6288 (2014).
6. Unarta, I. C. *et al.* Entropy of stapled peptide inhibitors in free state is the major contributor to the improvement of binding affinity with the GK domain. *RSC Chem Biol* **2**, 1274–1284 (2021).
7. Bird, G. H. *et al.* Biophysical determinants for cellular uptake of hydrocarbon-stapled peptide helices. *Nat Chem Biol* **12**, 845–852 (2016).
8. Bird, G. H. *et al.* Hydrocarbon double-stapling remedies the proteolytic instability of a lengthy peptide therapeutic. *Proceedings of the National Academy of Sciences* **107**, 14093–14098 (2010).

9. Atangcho, L., Navaratna, T. & Thurber, G. M. Hitting Undruggable Targets: Viewing Stabilized Peptide Development through the Lens of Quantitative Systems Pharmacology. *Trends Biochem Sci* **44**, 241–257 (2019).

10. Araghi, R. R. *et al.* Iterative optimization yields Mcl-1–targeting stapled peptides with selective cytotoxicity to Mcl-1–dependent cancer cells. *Proc Natl Acad Sci U S A* **115**, E886–E895 (2018).

11. Lama, D. *et al.* Structural insights reveal a recognition feature for tailoring hydrocarbon stapled-peptides against the eukaryotic translation initiation factor 4E protein. *Chem Sci* **10**, 2489–2500 (2019).

12. Lau, Y. H. *et al.* Investigating peptide sequence variations for 'double-click' stapled p53 peptides. *Org Biomol Chem* **12**, 4074–4077 (2014).

13. Dougherty, P. G. *et al.* Enhancing the Cell-Permeability of Stapled Peptides with a Cyclic Cell-Penetrating Peptide. *J Med Chem* (2019) doi:10.1021/acs.jmedchem.9b00456.

14. Phillips, C. *et al.* Design and structure of stapled peptides binding to estrogen receptors. *J Am Chem Soc* **133**, 9696–9699 (2011).

15. Lai, Y. *et al.* Inhibition of calcium-triggered secretion by hydrocarbon-stapled peptides. *Nature* **603**, 1–65 (2022).

16. Wu, Y. *et al.* Toolbox of Diverse Linkers for Navigating the Cellular Efficacy Landscape of Stapled Peptides. *ACS Chem Biol* **14**, 526–533 (2019).

17. 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).

18. Cromm, P. M. *et al.* Protease-Resistant and Cell-Permeable Double-Stapled Peptides Targeting the Rab8a GTPase. *ACS Chem Biol* **11**, 2375–2382 (2016).

19. Pessi, A. *et al.* Cholesterol-conjugated stapled peptides inhibit Ebola and Marburg viruses in vitro and in vivo. *Antiviral Res* 104592 (2019) doi:10.1016/j.antiviral.2019.104592.

20. Gallagher, E. E. *et al.* A cell-penetrant lactam-stapled peptide for targeting eIF4E protein-protein interactions. *Eur J Med Chem* **205**, 112655 (2020).

21. Navaratna, T. *et al.* Directed Evolution Using Stabilized Bacterial Peptide Display. *J Am Chem Soc* **142**, 1882–1894 (2020).

22. Stieglitz, J. T., Kehoe, H. P., Lei, M. & van Deventer, J. A. A Robust and Quantitative Reporter System to Evaluate Noncanonical Amino Acid Incorporation in Yeast. *ACS Synth Biol* **7**, 2256–2269 (2018).

23. Urquhart, T., Daub, E. & Honek, J. F. Bioorthogonal Modification of the Major Sheath Protein of Bacteriophage M13: Extending the Versatility of Bionanomaterial Scaffolds. *Bioconjug Chem* **27**, 2276–2280 (2016).

24. Oller-Salvia, B. & Chin, J. W. Efficient Phage Display with Multiple Distinct Non-Canonical Amino Acids Using Orthogonal Ribosome-Mediated Genetic Code Expansion. *Angewandte Chemie* **131**, 10960–10964 (2019).

25. Stieglitz, J. T., Lahiri, P., Stout, M. I. & Van Deventer, J. A. Exploration of Methanomethylophilus alvus Pyrrolysyl-tRNA Synthetase Activity in Yeast. *ACS Synth Biol* (2022) doi:10.1021/acssynbio.2c00001.

26. Wals, K. & Ovaa, H. Unnatural amino acid incorporation in E. coli: Current and future applications in the design of therapeutic proteins. *Frontiers in Chemistry* vol. 2 Preprint at https://doi.org/10.3389/fchem.2014.00015 (2014).

27. Wang, L., Brock, A., Herberich, B. & Schultz, P. G. Expanding the Genetic Code of Escherichia coli. *Science (1979)* **292**, 498–500 (2001).

28. Young, T. S. *et al.* Evolution of cyclic peptide protease inhibitors. *Proceedings of the National Academy of Science* **108**, 11052–11056 (2011).

29. Xie, J. & Schultz, P. G. An expanding genetic code. *Methods* **36**, 227–238 (2005).

30. Kiick, K. L., Saxon, E., Tirrell, D. A. & Bertozzi, C. R. Incorporation of azides into recombinant proteins for chemoselective modification by the Staudinger ligation. *Proc Natl Acad Sci U S A* (2001).

31. Strable, E. *et al.* Unnatural amino acid incorporation into virus-like particles. *Bioconjug Chem* **19**, 866–875 (2008).

32. Li, A. *et al.* High-throughput profiling of sequence recognition by tyrosine kinases and SH2 domains using bacterial peptide display. *bioRxiv* 1–41 (2022) doi:10.1101/2022.08.01.502334.

33. Wang, X. S. *et al.* A Genetically Encoded, Phage-Displayed Cyclic-Peptide Library. *Angewandte Chemie* **131**, 16051–16056 (2019).

34. Rezaei Araghi, R., Ryan, J. A., Letai, A. & Keating, A. E. Rapid Optimization of Mcl-1 Inhibitors using Stapled Peptide Libraries Including Non-Natural Side Chains. *ACS Chem Biol* **11**, 1238–1244 (2016).

35. Chène, P. Inhibiting the p53-MDM2 interaction: An important target for cancer therapy. *Nature Reviews Cancer* vol. 3 102–109 Preprint at https://doi.org/10.1038/nrc991 (2003).

36. Czabotar, P. E., Lessene, G., Strasser, A. & Adams, J. M. Control of apoptosis by the BCL-2 protein family: Implications for physiology and therapy. *Nat Rev Mol Cell Biol* **15**, 49–63 (2014).

37. Dutta, S. *et al.* Determinants of BH3 Binding Specificity for Mcl-1 versus Bcl-xL. *J Mol Biol* **398**, 747–762 (2010).

38. Rice, J. J. & Daugherty, P. S. Directed evolution of a biterminal bacterial display scaffold enhances the display of diverse peptides. *Protein Engineering, Design and Selection* **21**, 435–442 (2008).

39. Foight, G. W. & Keating, A. E. Locating Herpesvirus Bcl-2 Homologs in the Specificity Landscape of Anti-Apoptotic Bcl-2 Proteins. *J Mol Biol* **427**, 2468–2490 (2015).

40. Getz, J. A., Schoep, T. D. & Daugherty, P. S. Peptide discovery using bacterial display and flow cytometry. *Methods Enzymol* **503**, 75–97 (2012).

41. Ramesh, B., Frei, C. S., Cirino, P. C. & Varadarajan, N. Functional enrichment by direct plasmid recovery after fluorescence activated cell sorting. *Biotechniques* **59**, 157–161 (2015).

42. Fadrosh, D. W. *et al.* An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. *Microbiome* **2**, 1–7 (2014).

43. Micsonai, A. *et al.* BeStSel: Webserver for secondary structure and fold prediction for protein CD spectroscopy. *Nucleic Acids Res* **50**, W90–W98 (2022).

44. Clackson, T. & Wells, J. A. A Hot Spot of Binding Energy in a Hormone-Receptor Interface. *Science (1979)* **267**, 383–386 (1995).

45. Böttger, A. *et al.* Molecular Characterization of the hdm2 ± p53 Interaction. *J Mol Biol* (1997).

46. Massova, I. & Kollman, P. A. Computational alanine scanning to probe protein-protein interactions: A novel approach to evaluate binding free energies. *J Am Chem Soc* **121**, 8133–8143 (1999).

47. Kortemme, T. & Baker, D. A simple physical model for binding energy hot spots in protein-protein complexes. *Proceedings of the National Academy of Sciences* **99**, (2002).

48. Stewart, M. L., Fire, E., Keating, A. E. & Walensky, L. D. The MCL-1 BH3 helix is an exclusive MCL-1 inhibitor and apoptosis sensitizer. *Nat Chem Biol* **6**, 595–601 (2010).

49. Baek, S. *et al.* Structure of the stapled p53 peptide bound to Mdm2. *J Am Chem Soc* **134**, 103–6 (2012).

50. Chang, Y. S. *et al.* Stapled α-helical peptide drug development: A potent dual inhibitor of MDM2 and MDMX for p53-dependent cancer therapy. *Proc Natl Acad Sci U S A* **110**, (2013).

51. Mól, A. R., Castro, M. S. & Fontes, W. NetWheels: A web application to create high quality peptide helical wheel and net projections. *bioRxiv* (2018) doi:10.1101/416347.

52. Zhang, L., Navaratna, T., Liao, J. & Thurber, G. M. Dual-purpose linker for alpha helix stabilization and imaging agent conjugation to glucagon-like peptide-1 receptor ligands. *Bioconjug Chem* **26**, 329–337 (2015).

53. Lau, Y. H. *et al.* Functionalised staple linkages for modulating the cellular activity of stapled peptides. *Chem Sci* **5**, 1804–1809 (2014).

54. Tareen, A. & Kinney, J. B. Logomaker: Beautiful sequence logos in Python. *Bioinformatics* **36**, 2272–2274 (2020).

55. Shamas-Din, A., Kale, J., Leber, B. & Andrews, D. W. Mechanisms of action of Bcl-2 family proteins. *Cold Spring Harb Perspect Biol* **5**, 1–21 (2013).

56. DeBartolo, J., Taipale, M. & Keating, A. E. Genome-Wide Prediction and Validation of Peptides That Bind Human Prosurvival Bcl-2 Proteins. *PLoS Comput Biol* **10**, 1–10 (2014).

57. Jenson, J. M. *et al.* Peptide design by optimization on a data parameterized protein interaction landscape. *Proc Natl Acad Sci U S A* **115**, E10342–E10351 (2018).

58. Saleh, M. N. *et al.* Phase 1 trial of ALRN-6924, a dual inhibitor of MDMX and MDM2, in patients with solid tumors and lymphomas bearing wild-type TP53. *Clinical Cancer Research* **27**, 5236–5247 (2021).

59. Sakamoto, K. M. *et al.* Protacs: Chimeric molecules that target proteins to the Skp1-Cullin-F box complex for ubiquitination and degradation. *Proc Natl Acad Sci U S A* **98**, 8554–8559 (2001).

60. Jenson, J. M., Ryan, J. A., Grant, R. A., Letai, A. & Keating, A. E. Epistatic mutations in PUMA BH3 drive an alternate binding mode to potently and selectively inhibit anti-apoptotic Bfl-1. *Elife* **6**, 1–23 (2017).

61. Shin, Y. H. & Yang, H. Exploration of α/β/γ-peptidomimetics design for BH3 helical domains. *Chemical Communications* **58**, 945–948 (2022).

62. Atzori, A., Baker, A. E., Chiu, M., Bryce, R. A. & Bonnet, P. Effect of Sequence and Stereochemistry Reversal on p53 Peptide Mimicry. *PLoS One* **8**, 1–10 (2013).

Chapter 3 Discovery of High Affinity and Specificity Stapled Peptide Bcl-x_L Inhibitors using Bacterial Surface Display

This chapter is derived from the following publication:

Marshall Case, Jordan Vinh, Mukesh Mahajan, Vivek Subramanian, Anna Kopp, Matthew Smith, and Greg Thurber. "Discovery of high affinity and specificity stapled peptide Bcl-xL inhibitors using bacterial surface display." *Manuscript to be submitted.*

Abstract

There is great need for therapeutic modalities that are able to target intracellular protein-protein interactions involved in disease. One such interaction is the dysregulation of apoptosis, or programmed cell death, which is co-opted by cancer to evade cell death and enable proliferation. Several diseases are characterized by their overexpression of Bcl-x_L, an anti-apoptotic B cell lymphoma 2 (Bcl-2) protein expressed on mitochondrial membranes. Bcl-x_L overexpression inhibits a biochemical cascade ultimately leading to apoptosis; selective inhibition of Bcl-x_L has the potential to increase cancer cell apoptosis while leaving healthy cells relatively unaffected. However, high homology between Bcl-x_L and other Bcl-2 proteins has resulted in the difficulty of its selective inhibition by small molecule drugs. In this chapter, we engineer stapled peptides, a chemical modification that can improve cell penetration, protease stability, and conformational stability, towards the selective inhibition of Bcl-xL. To accomplish this task, we built a focused combinatorial mutagenesis library of peptide variants on the bacterial cell surface, used copper catalyzed click chemistry to form stapled peptides, and sorted the library for high binding to Bcl-

x_L and minimal binding towards other Bcl-2 proteins. We then characterized the sequence and staple placement trends that governed specificity and characterized highly selective molecules on and off the cell surface for affinity and specificity. Finally, we confirmed the mechanism of action of these peptides is consistent with apoptosis biology. The molecules generated in this chapter represent improvements to the specificity for stapled peptides against Bcl-xL.

Introduction

Bcl-2 family members are involved with the regulation of apoptosis, or programmed cell death, via their transient or constitutive interactions with mitochondria. Mitochondria are primarily involved in oxidative phosphorylation and energy production, but these organelles also play an essential role in governing programmed apoptosis.[58,103] When limits on homeostasis of the cell are exceeded, the cell initiates a complex series of biochemical pathways to initiate programmed cell death. As one of the hallmarks of cancer, cells can evade the protective mechanism of apoptosis and allow the continued survival of the cancer cell. Central to cancer's ability to escape apoptosis is blocking the 'intrinsic' apoptotic pathway, a carefully regulated signaling pathway of cytosolic and mitochondrial proteins.[104] On the outer membrane of mitochondria, a signaling network regulates whether mitochondrial membranes remain intact or start large pore formation, releasing cytochrome c and initiating the formation of the apoptosome and eventual cell death.[54,57] At the center of this pathway, B cell lymphoma 2 (Bcl-2) proteins are interacting with a balance of pro- and anti-apoptotic factors under healthy conditions. However, many cancers overexpress these proteins, which dysregulates mitochondrial function and inhibits apoptosome formation.[105] The inhibition of Bcl-2 proteins via small molecules or peptides is therefore a direct way to re-establish apoptosis controls for dysregulated cells, providing a powerful approach for the treatment of cancer.[106]

While the Bcl-2 proteins are highly homologous and share many functions, structural differences lead to profound differences in how each of the 5 Bcl-2 proteins (Mcl-1, Bfl-1, Bcl-$x_{L,}$, Bcl-w, and Bcl-2) contribute to apoptosis among other biochemical phenomena.[50,52] Pathological inquiries have shown that cancers typically overexpress a subset of these 5 proteins. Therefore, selective inhibition of anti-apoptotic proteins has been a longstanding goal of the drug discovery field towards the minimization of off-target toxicity.[105] However, the design of selective Bcl-2 inhibitors is impactful beyond therapeutic molecules: highly specific inhibitors can be used as 'tool' molecules to probe Bcl-2 dependency for novel cancers or used in other biochemical assays. Therefore, discovery of highly specific novel agents may accelerate pathological analyses.[107] In particular, Bcl-x_L remains one of the most important targets among the Bcl-2 family owing to its link with drug resistance, angiogenesis, and cancer cell stemness.[36,108] Bcl-x_L upregulation in breast, glioblastoma, melanoma, among many others is correlated with cancer cell invasion and metastasis.[109–111] Specific inhibition of Bcl-x_L is suggested to address the acquired resistance to poly-specific small molecule drugs like venetoclax.[105,112–116] However, currently no small molecule drug targeting Bcl-x_L has passed clinical trials at any stage.[105]

Despite the promise of cancer treatment via Bcl-x_L inhibition, the discovery of high affinity and selectivity drugs is challenging for several reasons. First, these proteins rely on alpha helical binding "BH" motifs for recognition of apoptotic proteins, which are shallow and hydrophilic and therefore challenging to target with small molecule drugs.[54,117] The lack of a traditional hydrophobic binding pocket has led to generation of peptidomimetic drugs which possess many shared characteristics with the proteins the cell naturally uses.[118] However, the small size of these peptidomimetic drugs often limits their specificity between Bcl-2 members,

leading to excessive off-target toxicity.[105] ABT-737 and its orally bioavailable analog ABT-263 (Navitoclax) were among the most specific small molecule drugs, having high affinity for only Bcl-$x_{L,}$ Bcl-w, and Bcl-2.[117] However, these drugs resulted in dose limiting thrombocytopenia due to Bcl-x_L related toxicity in circulating platelets. Small molecule drugs targeted at Bcl-x_L specifically have thus far failed in the clinic due to high *in vivo* toxicity. More recently, A-1155463 and A-1331852, small molecules engineered through structure-based drug discovery, have been engineered for high specificity for Bcl-xL over Bcl-2, though suffer from nanomolar binding to Bcl-w (or were not characterized for Bcl-w binding).[119,120] To overcome the challenge of targeting large hydrophilic protein-protein interaction domains while mitigating off-target toxicity, scientists have proposed peptides as therapeutics, resulting in high affinity and in some cases high specificity.[30,31,34–38,55,59,62–64] . The larger size of peptides enables high affinity and specificity interactions with otherwise difficult to target disease related proteins.[5] Keating and co-authors sorted a library of linear peptides using yeast surface display, trained machine learning models, and optimized sequences with integer linear programming to achieve high specificities between Bfl-1, Mcl-1, and Bcl-x_L.[37,61] Dutta et al. sorted a library of linear peptide variants based on a non-specific but high affinity wild type sequence (BIM), which yielded peptides with ~1000x specificities between Bcl-xL, Bcl-w, and Bcl-2.[36] These developments represent important milestones in the development of drug-like Bcl-2 inhibitory peptides. However, additional modifications are needed, since linear peptides suffer in the clinic from short *in vivo* half-lives and the inability to penetrate cell membranes.[46,67,121] Stapled peptide therapeutics, formed by crosslinking two amino acids, can help address some of the limitations of linear peptides by improving stability, affinity, and membrane permeability.[5,14,33,40,45,71,73,75,96,122,123] However, the development of stapled peptides is currently

constrained by low-throughput solid phase peptide synthesis, a bottleneck that limits evaluation of stapled peptides on the order of dozens.[40]

In this work, we use stabilized peptide engineering by *E. coli* display (SPEED) to rapidly evaluate high affinity and specificity Bcl-x_L stapled peptides **(Figure 3.1)**.[23,24] SPEED enables high-throughput screening of fully stapled peptides in a quantitative format by displaying genetically encoded stapled peptides on the surface of bacteria using non-natural amino acids, enabling evaluation of up to 10^9 peptides. To generate specific Bcl-x_L inhibitors, we designed a library of BIM mutants, a naturally occurring peptide that has high affinity for all Bcl-2 proteins but no specificity towards Bcl-x_L.[50] SPEED was used to simultaneously vary both peptide sequence and staple location to determine how staple location governs specificity in the context of a BIM-based library.[24] We sorted the library using a combination of magnetically activated- and fluorescently activated- cell sorting towards highly specific mutations that would target Bcl-x_L with both high affinity and specificity. By analyzing the final peptides in our library using next generation sequencing, we identified sequence trends and staple locations that govern specificity. We then translated select lead compounds off the bacterial surface and show that the peptides discovered retain their properties from surface display in both biolayer interferometry and competitive inhibition experiments.

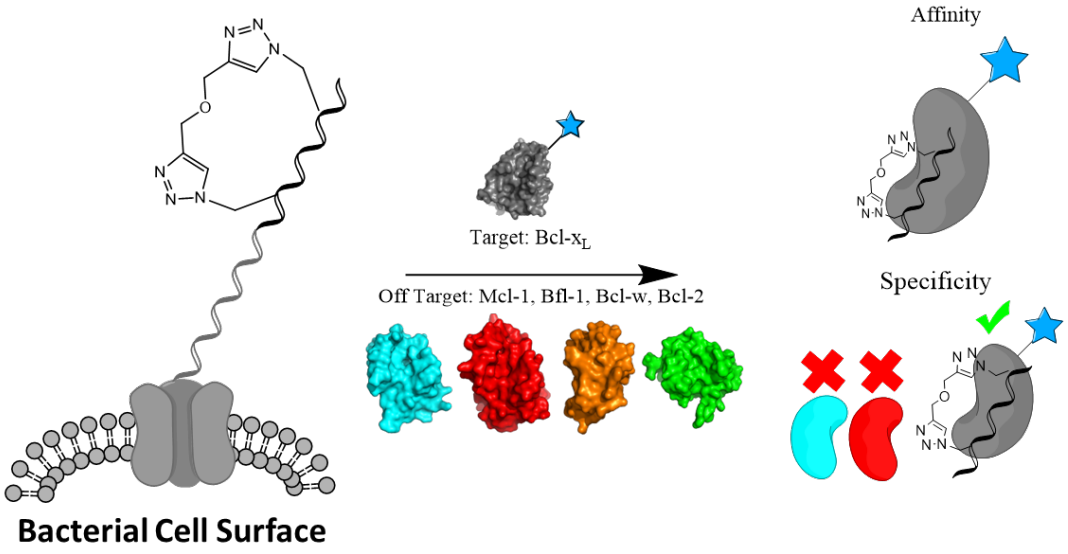


Figure 3.1: Engineering of high affinity and specificity pro-apoptotic Bcl-xL antagonistic stapled peptides using Stabilized Peptide Engineering by *E. coli* Display (SPEED). In this work, a library of focused stapled peptides variants are presented on the bacterial surface and engineered for high affinity and high specificity towards Bcl-x_L.

Methods

3.1.1 Purification of Bcl-2 protein

Bcl-2 proteins were purified as described previously.[24]

3.1.2 Library Design

To design the library of BIM variants, BH3 sequences and their affinities were collected from literature.[30–32,35,37,59] Sequences were split into 5 bins according to their affinities: <1nM, 1-10nM, 10-100nM, 1000nM. A position specific scoring matrix (PSSM) was generated for each of the 5 Bcl-2 proteins (Bcl-x_L, Bfl-1, Mcl-1, Bcl-2, or Bcl-w) based on the subset of sequences that had reported affinities for that target. To bias the library design towards high affinity clones, a weighting was applied based on the discretized bins where sequences were counted 1-10,000X in logarithmically spaced bins, depending on which bin it appeared in (high affinity sequences were counted more).

Another PSSM for each Bcl-2 protein was generated by extracting the weights from BIM variant SPOT arrays.[55,63] In brief, a SPOT array is a solid support with chemically synthesized, anchored peptides that are used to measure dozens to hundreds of peptide variants. First, a binary mask was generated based on the SPOT arrays to separate the location of peptide signal from background. Then, the average intensity of each SPOT was extracted using Fiji v2.0. For each Bcl-2 protein, the two PSSM's were averaged to capture information from both entire peptide sequences and BIM mutations. Then, mutations were selected to maximize the amount of specificity-driving residues (ones that were highly scored in one library and very weakly scored in others). We excluded Bcl-w and Bcl-2 mutations as they were qualitatively similar to Bcl-x_L and we wanted to maintain the balance between Mcl-1, Bfl-1, and Bcl-x_L residues. Because bacterial surface display libraries can only be generated on available equipment with ~10^9 unique sequences, we constrained the design by first locking amino acids in positions that had the highest absolute magnitude, such as Leu[3a] and Asp[3f]. Next, we fixed amino acids that didn't contribute to specificity based on their low weights in the PSSM, such as Gly[1e], Arg[1f], Tyr[4d,4e], and Ala[4f]. Still, the library had more mutations than were possible to display on bacteria. We next eliminated all Cys and Met residues from analysis because we wanted to initially minimize potential disulfide bond formation and out-of-position stapling residues respectively. To put a hard constraint on the size of the library, we merged the processed PSSM's for Mcl-1, Bfl-1, and Bcl-xL into one by applying the following equation:

$$PSSM_{p,AA} = \sum_{i=1}^{3} \left| 2 * PSSM_{p,AA,on-target} - PSSM_{p,AA,off-target} - PSSM_{p,AA,off-target} \right|$$

Where p is the position of the peptide sequence, AA is the amino acid, the on-target refers to the PSSM corresponding to the protein target, and the two off-target terms are for the other two Bcl-2 proteins. Finally, degenerate codons were optimized using SwiftLib for a given

protein sequence diversity of 10^8.[124] The primers and degenerate codon are tabulated in **Table 3.1** and **Table 3.2**. The design overview is available in **Figure 3.2**.

3.1.3 Library construction

Primers for focused BIM mutants with degenerate codons from the library design step were ordered from IDT whose sequences can be found in **Table 3.3**. Libraries were displayed using the pqe80L-eCPX2 plasmid as described previously.[23,24] We added an N-terminal HA tag and a peptide linker to mitigate steric clashes between display and binding measurements as described previously.[30]

3.1.4 Synthesis and preparation of peptides

Bcl-2 peptides were synthesized, stapled, and purified as described previously.[23,24] Peptide structures, mass spectra, and chromatograms are located in **Figure 3.14**, **Figure 3.15**, and **Figure 3.9** respectively.

3.1.5 Circular Dichroism

Peptides were dissolved in acetonitrile: water at 1:1 v/v at ~ 0.1 mg/mL and analyzed on a Jasco J-815 Circular Dichroism (CD) Spectrometer at 100nm/min at 25^{o}C. Spectra were baseline corrected and averaged over 3 runs. Alpha helicity was calculated using the BeStSel web server.[92] Circular dichroism spectra and alpha helicities are located in **Figure 3.9**.

3.1.6 Bacterial Surface Display, Flow Cytometry, and Competitive Inhibition Experiments

The preparation of bacterial cells for flow cytometry and FACS was performed as described previously.[23,24] Briefly, bacteria expressing the eCPX gene were grown in 1mL M9-methionine-ampicillin overnight. Fresh media was inoculated at a 1:20 ratio and grown for 150

minutes at 37°C. Next, the cells were metabolically depleted for 30 min, then induced with 0.5mM IPTG at 22°C for 4hr. Peptides on the surface of cells were clicked using 100μM propargyl ether at 4°C for 4hrs. Cells were washed once in PBS before incubation in Bcl-2 protein and expression markers overnight on ice. Cells were washed once in PBS before 15 min incubation with secondary antibody if necessary, then resuspended in PBS before analysis or sorting. Flow cytometry was done using an Attune NXT or BioRad Ze5.

Affinities of peptides on bacterial cell surface were measured using 8 logarithmically spaced concentrations of protein in triplicate. Competitive binding inhibition experiments were performed by incubating a fixed concentration of Bcl-2-biotin protein (100nM for Bcl-xL, Bcl-w, and Bcl-2, and 10nM for Bfl-1 and Mcl-1) with BIM-p5, a known binder for all 5 Bcl-2 proteins, with 8 logarithmically spaced concentrations of peptide in triplicate. After several hours of incubation at 0°C, 1uL of BIM-p5 displaying bacteria were added for 15 min to capture any unbound Bcl-2 protein. Then the media containing protein-peptide complexes was removed by centrifugation and bacterial cells were prepared for flow cytometry as described above.

3.1.7 Magnetic Activated Cell Sorting (MACS)

The naïve library was subjected to three total rounds of magnetic sorting (see **Table 4.4** for sorting details). One round of anti-HA MACS was done using anti-HA magnetic beads (Thermo Fisher). Then, cells were subjected to two rounds of sequential binding-based MACS with 100nM Bcl-x_L with the goal of selecting clones below the maximum number of cells that can be analyzed via FACS. The sorting progression is described in **Figure 3.16**.

3.1.8 Fluorescent Activated Cell Sorting (FACS)

Libraries were subjected to four rounds of sequential fluorescent sorting (see **Table 3.4** for sorting details and **Figure 3.17** and **Figure 3.18** for affinity-based and specificity-based representative FACS plots respectively). The first and second rounds were purely based on affinity and used 100nM and 10nM Bcl-x_L respectively. The highest 5% of cells expressing and binding were collected. The third and fourth round were done to improve the specificity of the library. The third round used 100nM Bcl-x_L and 25nM of each of the four other Bcl-2 proteins in competition (Mcl-1, Bfl-1, Bcl-w, and Bcl-2). The fourth used 10nM Bcl-x_L and 25nM of each of the four other Bcl-2 proteins in competition. In the third and fourth round, the top 5% of cells that were positive for Bcl-x_L binding and expression but negative for the competitive binding (in a third fluorescent channel) were selected. In parallel, another third round and fourth round were performed that was a negative sort (25nM each Mcl-1, Bfl-1, Bcl-w, and Bcl-2) and positive (1nM Bcl-x_L). Ultimately we found that the competitive screens yielded more enrichment based on NGS analysis and focused on these sorts for downstream analysis. The sorting progression is described in **Table 3.4** and **Figure 3.16**.

3.1.9 Biolayer Interferometry

Biolayer interferometry was performed as described previously.[23,24] Representative BLI traces are found in **Figure 3.11**.

3.1.10 Illumina Sequencing and Data Processing

Libraries identified for deep sequencing analysis were prepared as described previously.[24] Sequencing primers and PCR scheme is shown in **Figure 3.19**. Forward and reverse reads DNA were merged using NGmerge and aligned using in-house python scripts. [125] We filtered out sequences that didn't match the framework region of the eCPX2 protein and condensed identical

peptide sequences into read counts. To account for the differences in total read counts per library, we converted read counts into frequencies by normalizing to the number of reads. Sequence conservation plots were made using the Logomaker Python package.[99]

3.1.11 Mitochondrial Membrane Depolarization

Mitochondrial outer membrane polarization assays were performed as described previously.[126,127] In brief, mammalian cells were added to a mixture of MEB buffer, digitonin, voltage dependent fluorophore (JC-1), oligomycin, beta-mercaptoethanol, with either peptide treatment, FCCP (reversible decoupler of oxidative phosphorylation), alamethicin (a peptide that induces irreversible mitochondrial depolarization), or DMSO (negative control). Fluorescence was measured at 545nm excitation/ 590nm emission using a BioTek Synergy H1 plate reader at 32^{o}C at 5 min intervals for 180 minutes. MDA-MB-231 and Mcf7 cells were used to test Bcl-xL dependence in natural cancer cell lines and the B-ALL leukemia cells were used to test Bcl-2 family dependence.[128] By normalizing peptide data to the negative control (maximum mitochondrial polarization) and the positive control (complete mitochondrial depolarization), it is possible to measure peptides' ability to drive cellular apoptosis.

3.1.12 Crystallography

Bcl-xL was purified by the University of Michigan Life Sciences Institute Center for Structural Biology. Purified and stapled peptide was supplied as a lyophilized powder.

3.1.13 Nuclear Magnetic Resonance Spectroscopy

Nuclear magnetic resonance spectroscopy was performed as described previously.[23] Briefly, ^{1}H TOCSY and NOESY data were collected on a 600MHz NMR instrument.

Results

3.1.14 Library Design and Cell Sorting

Because the sequence space of BH3-like peptides is much larger than the experimental throughput to measure them, the primary objective of library design was to identify mutations that are likely to impact specificity. To design such peptides, we noted several design criteria: 1) the library must sample both sequence space and staple location simultaneously to evaluate the impact on affinity and specificity, 2) critical binding residues must be preserved to minimize non-functional variants, and 3) select residues that were predicted to improve affinity and specificity should be mutated to sample the library more efficiently versus random mutagenesis.

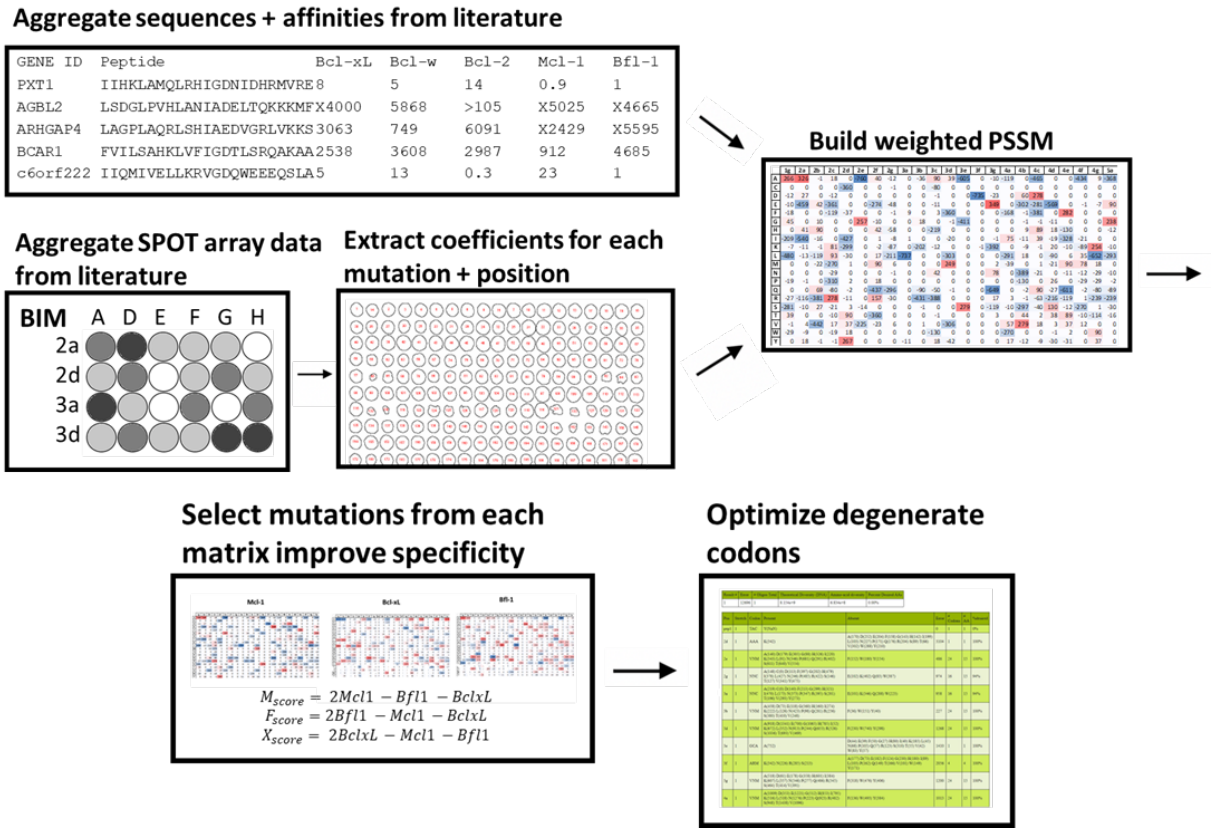


Figure 3.2: Technique used to construct library computationally. First, sequences with labeled affinities for Bcl-2 proteins are aligned and converted into a weighted position specific scoring matrix (PSSM) (see Methods for details). Separately, SPOT array data is aggregated and extracted and added to the position specific scoring matrix. One PSSM is constructed for Mcl-1, Bfl-1, and Bcl-xL, and then mutations are sampled until a desired library size

of ~10^8 is achieved. Finally, degenerate codons are generated for each staple position to efficiently sample random amino acids.

First, we chose to simultaneously evaluate staple location and sequence. We hypothesized that epistatic interactions between the staple and the peptide sequence might enable high affinities and/or specificities that would be lost if the staple location was fixed. We chose several staple locations that were previously validated to bind Bcl-2 proteins with varying specificities in the context of BIM.[24] We chose a peptide length of 23 because peripheral residues generally strengthen binding but do not significantly affect specificity.[36] Next, we aggregated sequence and affinity data from literature to design a library of variants predicted to improve affinity or specificity towards Bcl-2 members while minimizing the number of non-functional variants. [30–32,35,37,55,59,63] We included all naturally occurring and engineered BH3 sequences that had been assayed for binding affinities. Data from SPOT arrays, where the change in binding was measured for BIM single mutants, was combined with sequence data to generate a position specific scoring matrix (PSSM) for each Bcl-2 protein. We weighted mutations based on their affinity to each target and then sampled mutations according to their magnitude of specificity (mutations that were predicted to be highly specific towards one or multiple Bcl-2 proteins) until the desired design space was achieved (~10^8 sequences). To maximize the number of sequences from our PSSM and minimize the number of suboptimal residues from degeneracy in the codon table, the final DNA sequences were optimized using SwiftLib.[124] More details about the computational library design are available in the methods section and **Figure 3.2** and **Table 3.1**, **Table 3.2**, and **Table 3.3**. After transformation into bacteria, the library had 5.5*10^8 unique peptides **(Figure 3.3)**.

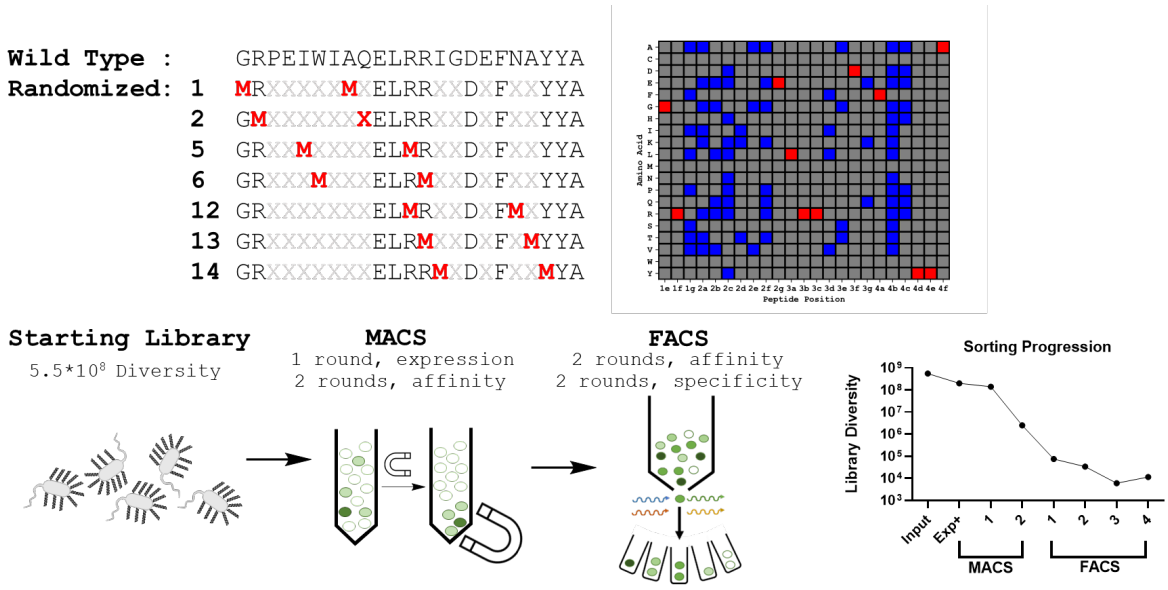


Figure 3.3: A library of stapled peptides is designed to be enriched with residues and staple positions that govern Bcl-xL affinity and specificity. Staple location and sequence were determined to be two key determinants of affinity and specificity and thus a library where they are simultaneously evaluated was generated. Then, the library is transformed into bacteria, sorted 3 times magnetically and four times fluorescently (in series) before deep sequencing each sort round.

Initial flow cytometry experiments with the naïve library showed only 10-30% of all cells were displaying peptide, indicating that many cells either had ampicillin resistance but did not contain functional plasmid or that peptide-eCPX2 was being inefficiently shuttled to the cell surface. To enrich the libraries towards functional peptides, we performed 3 rounds of magnetic activated cell sorting (MACS) followed by 4 rounds of fluorescent activated sorting (FACS) (see **Table 3.4** for sorting details, **Figure 3.17** and **Figure 3.18** for representative FACS plots, and **Figure 3.20** for logoplots for FACS 1-4). MACS experiments were split into two phases: one round to improve the expression of the library and two to improve binding while simultaneously shrinking the library to a size that could be sorted by FACS, where more specific boundaries can be chosen for desired peptide fitness. FACS experiments were also split into two phases: two rounds to improve the affinity of the library towards Bcl-x_L by screening with decreasing concentrations of protein (100nM and 10nM respectively), and the last two rounds were done to improve the specificity by performing competition experiments and selecting towards highly

specific binders. We hypothesized that this new library would be optimally sorted via FACS by first finding the high affinity Bcl-x_L binders and then identifying the subset that were highly specific as previously reported.[31] After sorting, we performed low throughput flow cytometry experiments, which suggested that the library had been highly enriched towards specific binding, as evidenced by nearly saturable binding at low concentrations (<10nM) of Bcl-x_L but minimal binding towards other Bcl-2 proteins even at high concentrations (>100nM). In addition to sorting for highly specific peptides with competition sorting experiments, we also tested whether a round of negative sorting (the lack of binding to off-target proteins) followed by a positive round of sorting (target binding) would yield highly specific clones. Both competitive and non-competitive sorts yielded a similar set of enriched sequences (**Figure 3.21**), suggesting the library was well suited to finding specific Bcl-x_L inhibitory stapled peptides.

3.1.15 Next Generation Sequencing

After magnetic and fluorescent sorting, we analyzed the set of enriched peptides along the sorting progression using Illumina NovaSeq next generation sequencing (NGS) (**Figure 3.4**). We first investigated how sorting influenced the enrichment and proportions of the library composition; all rounds of sorting resulted in an enrichment of sequences and depletion of others. We next investigated the relationship between the staple position and the peptide sequence among highly functional Bcl-x_L peptides. Despite staple scanning in the context of BIM suggesting that many positions had high affinity or specificity for Bcl-x_L (the 2nd, 6th and 12-14th staple position BIM mutants had high affinity toward Bcl-x_L), surprisingly, all stapled positions except the 6th were nearly elimination after the first round of FACS.[24] While most sequences observed had the 6th staple position, peptide sequence is highly dependent on the location of staple (**Figure 3.22**). For example, while Lys^{2d} is the dominant mutation for the 6th

position, Ile is more prevalent for the 5th and 14th position. This suggests that there exists a complex relationship between the staple and the sequence that justifies the simultaneous screening of staple location and peptide sequence.

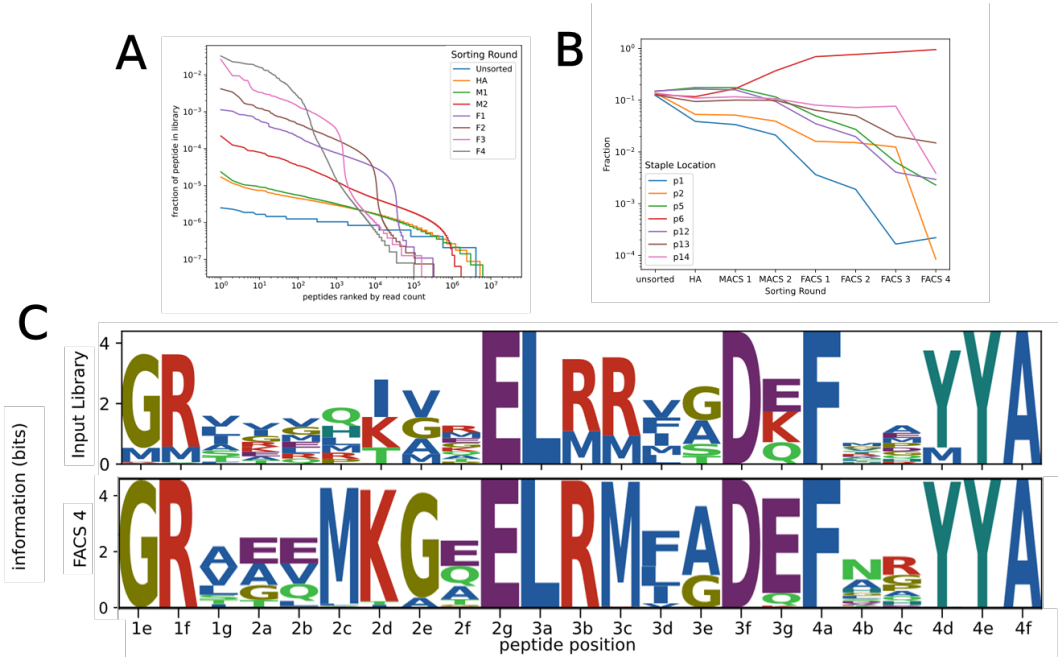


Figure 3.4: Next generation sequencing of sorted Bcl-xL peptides yields insights into the staple location and sequence patterns that govern specificity. (A) First, the distribution of peptide sequences' enrichments was calculated across sorting rounds. Whereas there is little bias in the library composition in the naïve library, as indicated by its nearly uniform distribution, each subsequent round of sorting biases the library towards a subset of sequences. By the final rounds of sorting, ~10^3 sequences represent nearly 100% of all sequences remaining. **(B)** By the second round of magnetic sorting, the 6th staple position nearly dominates the library and continues to displace other staple positions, suggesting that this staple position contributes to affinity and specificity. **(C)** Compared to the naïve library (top), which displays no bias towards certain residues in randomized positions, the Bcl-xL sorted library has clear sequence conservation patterns.

Finally, we investigated whether the sequence trends that emerged from sorting resulted in conserved patterns and whether those patterns matched other Bcl-x_L peptides previously described. Compared to the naïve library, where any position not fixed by design displays nearly uniform distribution between selected mutations, the Bcl-x_L library has clearly conserved sequence patterns. The most dominant mutations, Lys^{2d}, Gly^{2e}, and Asp^{3g}, appeared in nearly every peptide that remained in the library. Positions 2a, 2b, and 2f did not display the same level of conservation but generally have enriched negative glutamic residues, whereas the naïve

library had mostly small hydrophobic residues. Positions 1g, 4b, and 4c had a smaller magnitude of enrichment, consistent with their position peripheral to the main alpha helical interface.[108]

The naïve library was predicted to contain mutations in the context of BIM that would be highly specific towards Mcl-1, Bfl-1, or Bcl-x_L based on computational design. Comparing the set of mutations that emerged post sorting to the naïve library could lend insight into the utility of leveraging prior data in library design. First, we analyzed the mutational space of Bcl-xL specific peptides with the final scores from library design (as detailed in **Figure 3.2**). These scores reflect the total importance of that mutation for overall specificity between these three Bcl-2 members. We observed that generally, the mutations that were predicted to significantly impact specificity were not exclusive to the set of mutations that were enriched. We found that while there was not an obvious correlation between the importance score as predicted in library design with the enrichment value post sorting, the mutations predicted to be important were increasingly more present in highly enriched sequences.(**Figure 3.5**). There were several residues that were not predicted to be important by overall library design but were highly enriched, representing important residues for further analysis, such as Glu^{2c}, Glu^{2d}, and Lys^{2f}. This indicates that while library design that incorporates prior information (such as homologous sequences or the effect of mutations such as SPOT arrays) can focus on residues with a higher

likelihood of increased fitness, it is important to include a large array of mutations for maximum specificity.

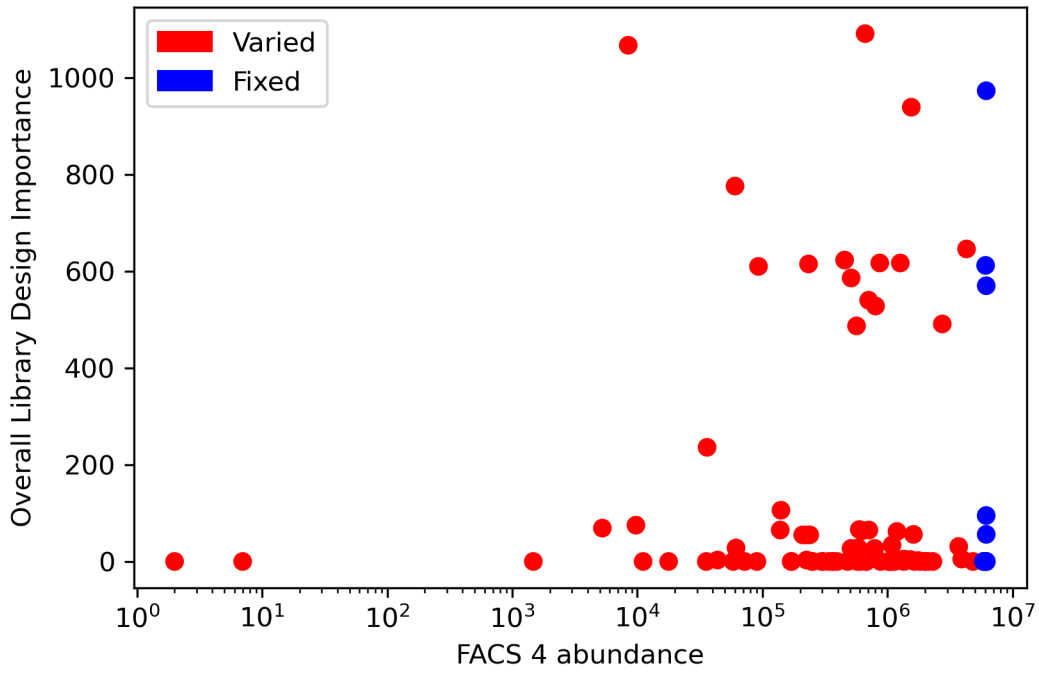


Figure 3.5: Comparison of mutational importance values from library design versus enrichment in experimental sorting. Overall library design importance was calculated as the average of weighted PSSM values for each of Bcl-xL, Bfl-1, and Mcl-1. See **Figure 3.2** for more information about library design. A significant number of residues predicted to be important for specificity were highly abundant in later rounds of FACs, as seen by the ~20% of residues in the upper right quadrant. However, 80% of residues had low or no predicted importance. This could be due to the use of information from linear peptides sequences for this stapled peptide sorting campaign.

We then evaluated whether the mutations enriched for Bcl-xL were among those predicted to be specific for Bcl-xL specificity specifically (**Figure 3.6**). Any residues that were abundant post sorting but not predicted in library design represent unexpected findings that were not present in natural homology or SPOT array data and may yield further improvements to Bcl-xL specificity compared to previous sorting campaigns. Many mutations predicted to govern specificity for Bcl-xL were more enriched than those predicted to not be beneficial. However, we were surprised that many mutations predicted to affect specificity minimally played a larger role than anticipated. This includes mutations such as His^{4e}, Gly^{2d} and Gly^{3a} (circled residues in

Figure 3.6), suggesting there were motifs beneficial for Bcl-xL specificity not predicted from the starting dataset.

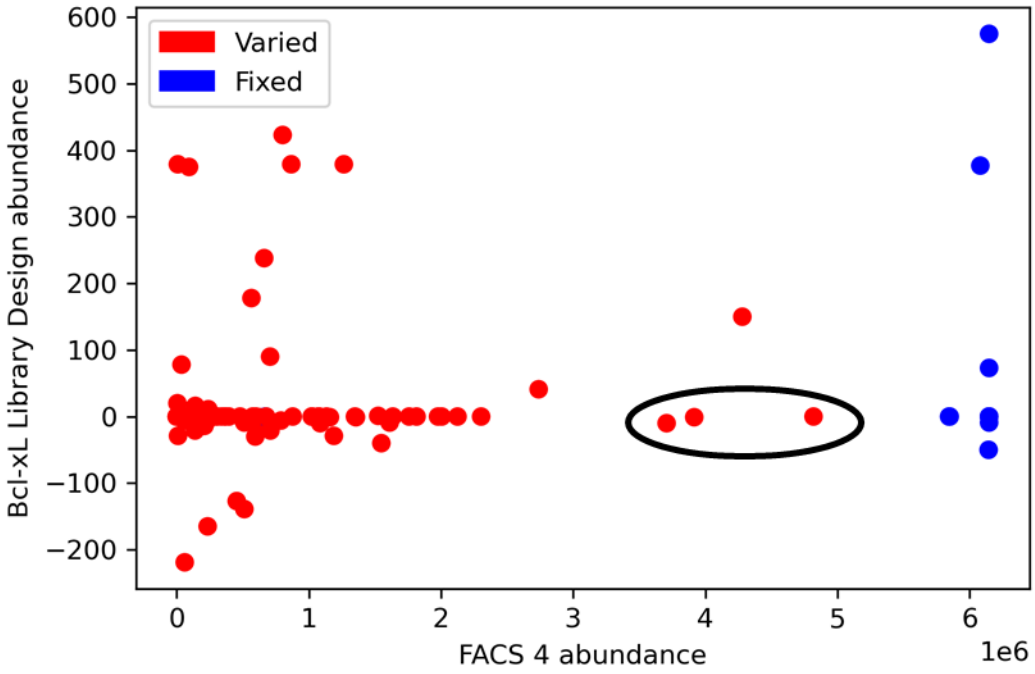


Figure 3.6: Comparison of mutations predicted to govern specificity for Bcl-xL by library design versus preference for those mutations from sorting. Mutations that have high predicted values for Bcl-xL binding (y value > 0). Bcl-xL design importance is the weighted PSSM values for Bcl-xL alone. See **Figure 3.2** for more information about library design. Mutations circled are three selected mutations, His^{4e}, Gly^{2d} and Gly^{3a}, which were not previously predicted to significantly contribute to specificity but were highly enriched during cell sorting.

While we generally observed that the mutations predicted to impact specificity were not correlated with their enrichment, this analysis treats mutations as if they were non-interacting; mutations could have varying levels of epistasis, and thus deviate from their predicted values irrespective of library design strategy. To evaluate this hypothesis, we used position specific enrichment ratio matrices (PSERM) to measure the epistasis present within the dataset (**Figure 3.7**). This analysis shows the connection between stapled positions (such as methionines 1e and 2e, which are required to be mutated together by design), which show an epistasis value of 1 (maximum epistasis).With the exception of the positions not varied (2g, 3a, 3f, 4a, 4e, and 4f),

which have zero epistasis by definition, the extent of epistasis across the sequence space is consistently higher than zero (which denotes the lack of epistasis). For example, the mutations sampled at position 2c, which is central to the binding interface and peripheral to other varied positions on the helix (such as 1g and 2f), has epistasis scores consistently around one (denoting maximal epistasis). Moreover, the epistasis present in these positions strongly depends on the identity of the mutation; at position 2c, residues D, E, H, and Y have higher epistasis than other sampled residues, suggesting that these residues are influenced by other proximal mutations. This analysis suggests that the mutated residues are interacting epistatically, which could explain why the unique sequences identified via the stapled library directed evolution campaign did not align with the library design predictions.

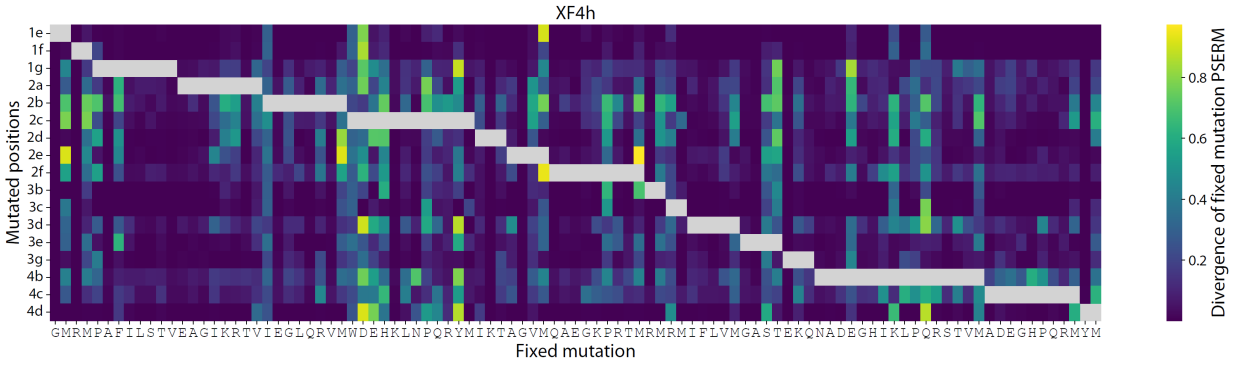


Figure 3.7: Position specific enrichment ratio matrix reveals the extent of epistasis within the dataset. For each position within the peptide, the dependence of each mutation against each other mutation at every position is calculated. The higher the context dependence (as indicated by color), the more dependent these mutations are.

3.1.16 Evaluation of peptide hits

The sequencing results and analyses at the library level suggested that the peptides had been selected towards high affinity and specificity Bcl-x_L peptides. We next investigated the affinity and specificity for individual sequences. We randomly selected thirteen of the highest frequency clones from the final round of FACS and evaluated their affinity and specificity

towards the Bcl-2 proteins (**Figure 3.8**). First, we measured binding at two concentrations of target, significantly above (100nM) and below (1nM) the median library binding affinity, to obtain a crude estimate of affinity. First, we measured the binding of each of these peptides on the bacterial cell surface towards Bcl-x_L at 1nM in triplicate. This concentration was chosen because it does not saturate BIM-p5 for Bcl-x_L, which is a double-digit nanomolar binder. Thus, increases in binding when normalized to display level are indicative of an improvement in binding affinity. All sequences evaluated demonstrated significantly improved binding, suggesting that peptide hits had K_d's in the low double digit nanomolar range. We similarly evaluated the binding of the other Bcl-2 proteins and observed significantly decreased binding towards Mcl-1 and Bfl-1 (~10,000 fold weaker binding). However, the magnitude of binding towards Bcl-w and Bcl-2 was not reduced to the same extent (10-100 fold weaker binding), though most clones had statistically significantly reduced binding for all four off-target proteins (**Figure 3.23**). We then tested binding at 100nM, which should saturate all but the weakest binders. This analysis suggested that most peptides were highly specific for Bcl-x_L, though fewer peptides had statistically significant specificity for Bcl-w and Bcl-2.[36]

Based on the promising crude estimates from the library, we selected two clones for further analysis. These variants were selected based on their diminished binding towards all 4 off-target Bcl-2 proteins while maintaining high affinity towards Bcl-x_L. We titrated two of these clones, denoted **12** and **13** with various concentrations of proteins on the cell surface. These peptides had high affinities towards Bcl-x_L (~10 nM K_d) and greatly weakened affinity towards the other targets, except for **13** binding to Bcl-2 with ~80 nM K_d. The identification of these specific stapled peptides for Bcl-x_L further demonstrated the ability of SPEED to find diverse sequences that have desirable activity.

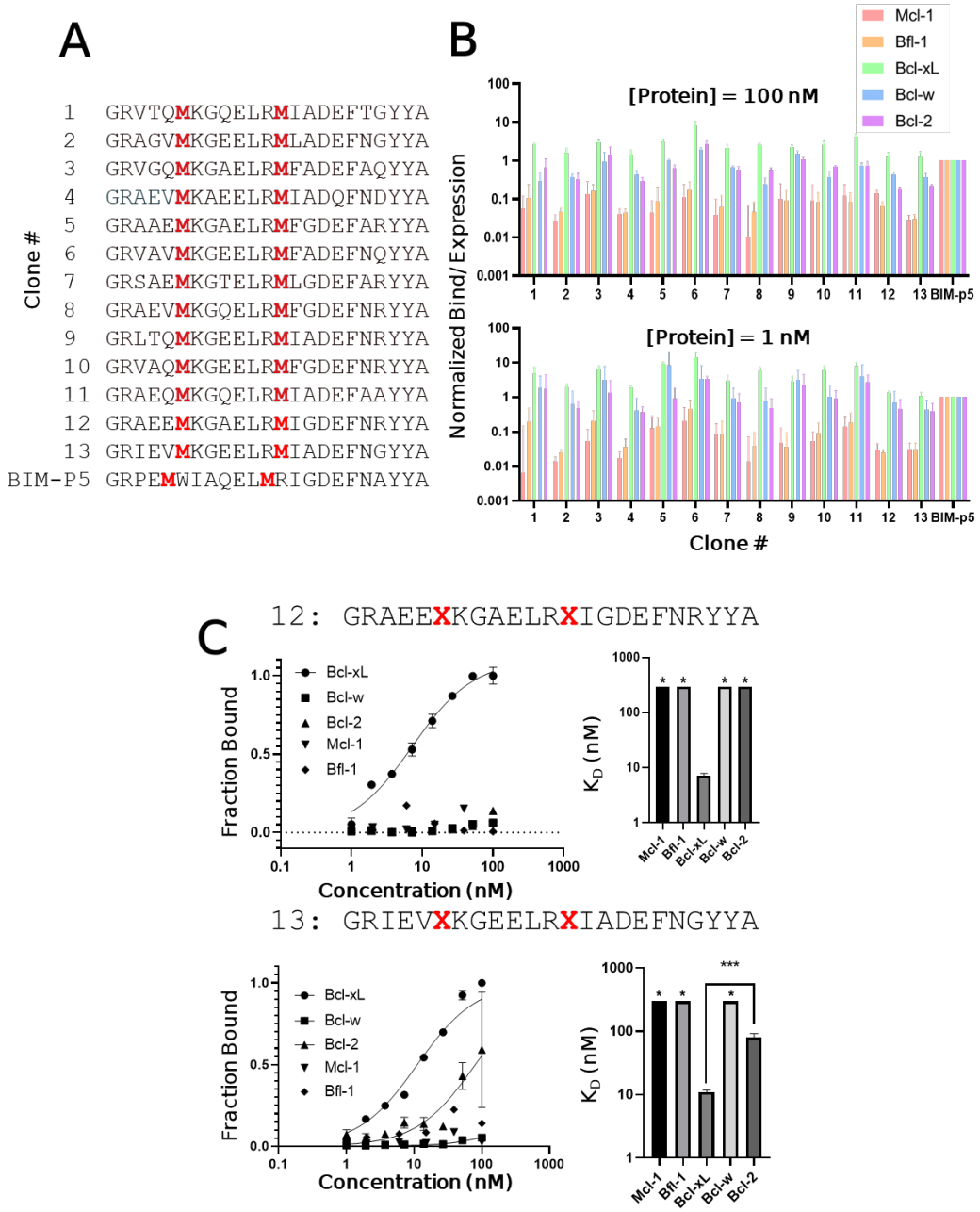


Figure 3.8: Individual peptides from the library are highly specific towards Bcl-xL. (A) Thirteen peptides were randomly chosen from the final round of cell sorting and their sequences were identified via Sanger sequencing. Each of these peptides were measured for their binding against all 5 members of the Bcl-2 protein family at **(B, bottom)** 1nM, which is significantly below the K_d of the wild type sequence for the main target, Bcl-xL, or **(B, top)** 100nM, which is a high enough concentration to saturate all but the weakest of binders. **(C)** Two of the thirteen peptides were chosen for more thorough analysis and their binding affinities were measured on the cell surface by flow cytometry.

3.1.17 Solution phase measurements

SPEED has been previously validated to predict binding properties of stapled peptides with comparable accuracy to solution phase measurements.[24] However, to ensure the binding affinities measured via bacterial surface agreed with solution phase peptide binding, we synthesized the two peptides using solid phase peptide synthesis. After stapling the peptides (as detailed in Methods), we characterized the peptides' secondary structures using circular dichroism (**Figure 3.9**). Both peptides had moderate alpha helicity before stapling (18% and 47%) but had significantly enhanced helicity when stapled (44% and 67%, respectively). Then, we used two techniques to measure the binding affinity and specificity: competitive inhibition and biolayer interferometry (**Figure 3.11**).

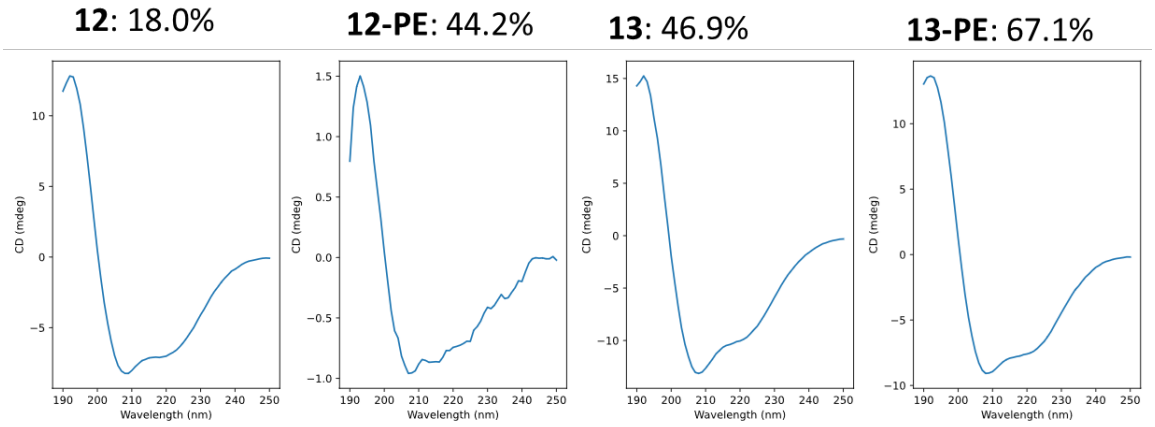


Figure 3.9: Circular dichroism and alpha helicities for compounds generated in this study

We next wondered whether bacterial surface display favors enrichment of helical peptides, given the helical nature of the Bcl-2 binding interface. When compared to BIM (unstapled, wild type), whose alpha helicity according to circular dichroism is 31% (**Figure 2.19**), these peptides (**12** and **13)** have significantly enhanced helicity in their stapled forms. However, in their unstapled forms, they do not demonstrate enhanced helicity across both lead Bcl-xL antagonists (**12** and **13**), suggesting that this sorting campaign may not have selected peptides with enhanced helicity. To further query this hypothesis, we used sequence to secondary

structure prediction tools to computationally predict peptide secondary structure at higher throughput than would be possible chemically. While these tools have limitations – they do not account for the chemical differences for non-natural amino acids or the conformational dynamics of a stapled peptide - the unstapled and unsubstituted peptide variants' secondary structure can be predicted with modest accuracy. Using the webserver PEP2D (http://crdd.osdd.net/raghava/pep2d/index.html), we found that the peptides identified post sorting were predicted to have helicities comparable to BIM (between 30 and 70% helicity, see **Figure 3.10**). We also found that peptides randomly selected from the input rounds of sequencing (naïve or expression positive) had comparable predicted helicities (between 30 and 70%). This analysis supports the conclusion from circular dichroism analysis (see **Figure 3.9**): sorting for Bcl-xL did not select significantly more helical peptides than our wild type starting molecule.

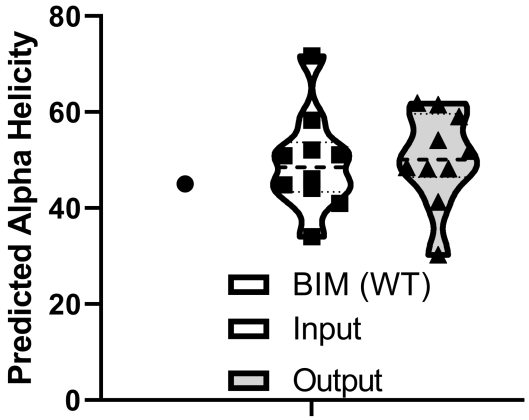


Figure 3.10: Predicted alpha helicities of the wild type molecule (BIM), before, and after sorting the designed library. Using the webserver PEP2D (http://crdd.osdd.net/raghava/pep2d/index.html), the predicted alpha helicities of the wild type peptide, BIM, was compared with ten randomly selected peptides from the naïve library and the final FACS round for Bcl-xL. Because the webserver does not account for contributions from non-natural amino acids or the stapling reaction, the sequences were submitted using methionine in place of azidohomoalanine.

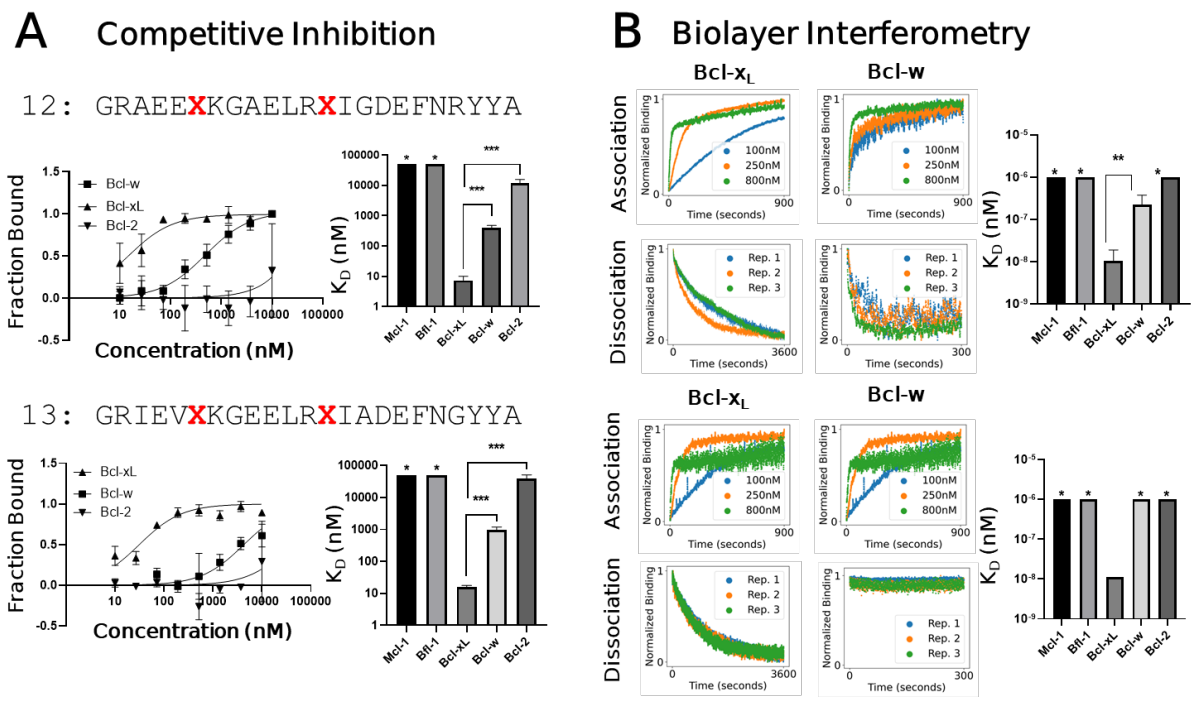


Figure 3.11: Solution phase characterization of select Bcl-xL stapled peptides. (A) Both hits' binding affinities are assayed by competitive inhibition. In this experiment, soluble peptide (in excess) and protein are allowed to equilibrate before a high-affinity peptide displayed on bacteria is added, which sequesters any unbound protein. The bacteria are then analyzed via flow cytometry and the fraction bound is inversely proportional to soluble peptide binding. The K_i is calculated and converted to a K_d using the Cheng-Prusoff equation. **(B)** The binding affinity is also measured using biolayer interferometry, where a sensor is loaded with biotinylated protein and the binding of peptide at various concentrations is measured in real time. The on and off rate are fit using GraphPad Prism v10.0. The K_d is calculated as the ratio of k_{off} to k_{on}. *: no binding detected. **: $p < 0.01$ ***: $p < 0.005$.

In the competitive inhibition experiments, peptides of various concentrations are equilibrated with select soluble Bcl-2 protein before the addition of BIM-p5 displayed on the surface of *E. coli* bacteria. After a short incubation of equilibrated peptide and protein with bacteria, any unbound protein is rapidly sequestered by bacteria, which are analyzed via flow cytometry. The fraction of protein bound is inversely proportional to the fraction of protein blocked by soluble peptide. The K_i is calculated by fitting the fraction bound as the peptide concentration is titrated before converting to a K_d based on the affinity of BIM-p5 to soluble protein using the Cheng-Prusoff equation. In the biolayer interferometry experiments, a streptavidin sensor is loaded with biotinylated Bcl-2 protein and incubated with various

concentrations of soluble peptide. Real time measurements allow the determination of equilibrium K_d and kinetic rates of k_{on} and k_{off} (kinetic parameters and statistics are available in **Figure 3.12**). Both solution phase techniques gave similar results to bacterial surface display measurements, in agreement with previous reports.[23,24] As previous reports confirm that bacterial surface measurements of affinity may overestimate affinity, disagreement between the value of **13** binding Bcl-2 via SPEED versus those from solution phase peptides confirm that solution phase measurements are still recommended. This analysis further demonstrated that both evaluated peptides had specificities towards Bcl-x_L over Bcl-w by an order of 100, Bcl-2 by 1,000, and all others by >10,000.

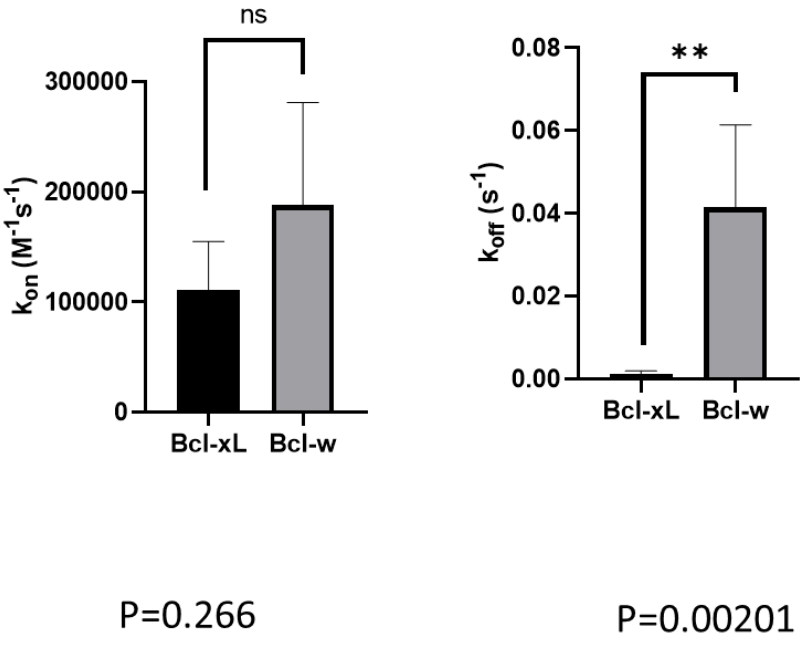


Figure 3.12: Comparison of fit kinetic rate parameters from biolayer interferometry data. The p-value is calculated as a unpaired t-test.

3.1.18 In vitro characterization

We next sought to characterize whether peptides functionally induced apoptosis in human cancer cell lines **Figure 3.13**. On the surface of mitochondria, Bcl-2 proteins sequester Bak and Bax,

which hetero-oligomerize to form pores and depolarize mitochondria. This phenomenon can be measured using a voltage sensitive fluorophore (JC-1) through a mitochondrial outer membrane polarization (MOMP) assay.[126,127] First, we incubated MDA-MB-231 and MCF7 (human derived cancer cells overexpressing Bcl-xL) with various concentrations of **13** and found that the peptide was able to depolarize mitochondria with nanomolar concentrations. Next, we tested a peptide known to not bind Bcl-xL (**F2**, binding affinities are shown in **Figure 3.24**) and demonstrated that binding affinity was the determinant of activity. Finally, we measured the B-ALL cell lines which are engineered to overexpress specific Bcl-2 proteins and showed that **13** specifically depolarized cell lines with Bcl-xL overexpression.[128]

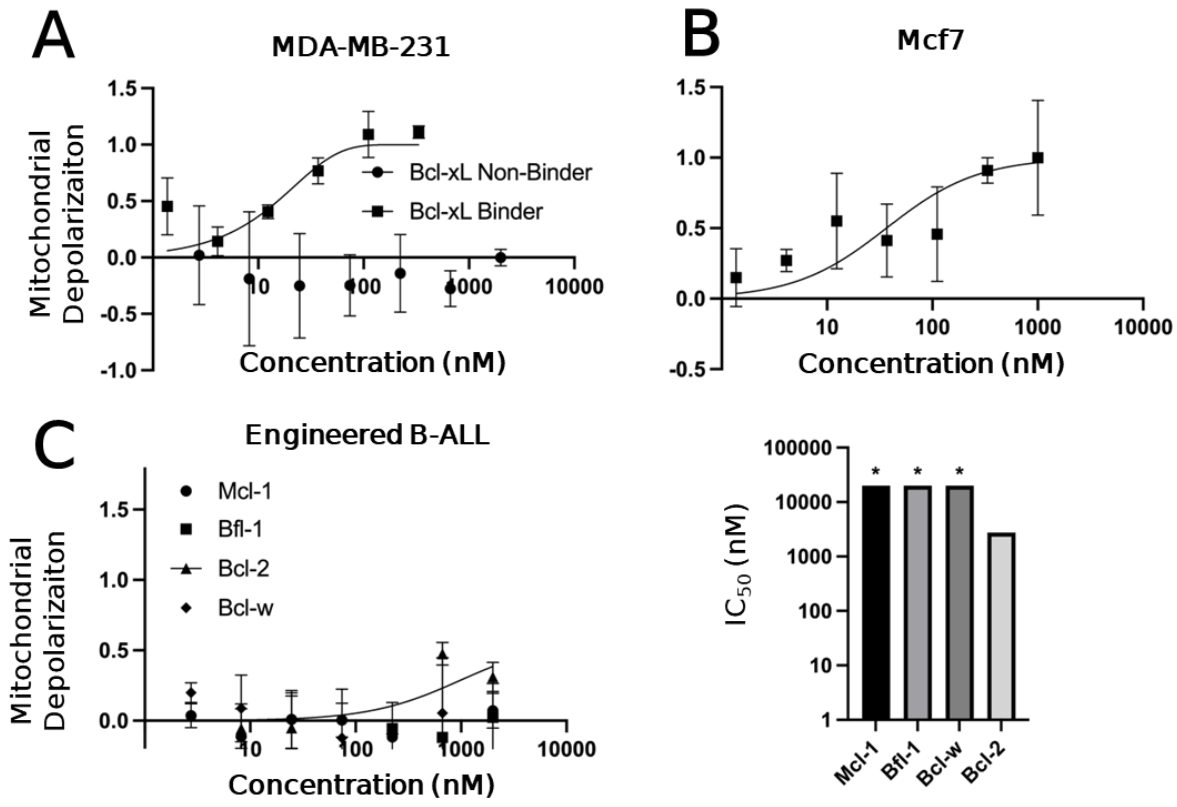


Figure 3.13: Mitochondrial depolarization is measured for W10 peptide at various concentrations in two Bcl-xL dependent cell lines: MCF7 and MDA-MB-231.

3.1.19 Structural Biology

To study the molecular basis of specificity, we used nuclear magnetic resonance spectroscopy to measure the structure of the peptide and the protein-peptide complex. To study the conformation of the peptide, we first used Total Correlation Spectrscopy (TOSCY) which evaluates bond connectivity via hydrogens within the peptide. Then, we used Rotating-frame Overhauser Enhancement Spectroscopy (ROESY), a variant of Nuclear Overhauser Effect Spectroscopy (NOESY), to study both bond connectivity and spatial proximity of hydrogens. See **Table 3.5** for NMR table and **Figure 3.25** for full NMR results.

Discussion

Direct targeting of Bcl-2 proteins is a highly effective approach to restore apoptosis in cancerous cells. Therefore, high affinity and specificity molecules targeting the B cell lymphoma 2 (Bcl-2) family of proteins have important therapeutic potential. In this work, we used SPEED to screen stapled peptides with varying linker location and sequence to select specific and high affinity compounds towards Bcl-x_L **Figure 3.1**. Given the large theoretical sequence space, we designed a focused computational library that we hypothesized contained mutations that would drive specificity between Bcl-xL and the other Bcl-2 proteins (**Figure 3.3**). To accomplish this design task, we leveraged two sources of data: sequence-affinity databases and SPOT arrays of BIM mutants. [30–32,35,37,55,59,63] First, we pooled the body of BH3 peptides' sequences and affinities for Bcl-2 proteins. While it has been shown that higher order epistatic effects are present in BH3 peptides, many more mutations act non-epistatically, and thus we assumed independent mutational effects on affinity as a reasonable simplification.[37] We aggregated these data and used them to predict which mutations were most likely to govern affinity and/or specificity (see

Methods). After optimizing degenerate codons that sampled these critical residues and constraining the library size to ~10^8, we transformed the library into bacteria, yielding a library with $5.5*10^8$ diversity.

A combination of magnetic (MACS) and fluorescent (FACS) sorting, focusing on affinity or specificity, were used to enrich the library towards specific Bcl-x_L peptides (**Figure 3.4**). While MACS was used to improve the expression of the library and then enrich the library to a diversity smaller than 10^7, FACS was the primary tool used to achieve affinity and specificity in the peptide libraries due to its ability to directly select from the binding/ expression landscape. We previously established that the staple location is a major driver of specificity.[24] Because SPEED enables simultaneous evaluation of sequence mutants and staple locations, we were able to sample this complex relationship across the sort progression. Sequence patterns that emerged among highly specific Bcl-x_L peptides generally agree with previous reported sequences. For example, we observed that Phe^{3d} was the most highly enriched amino acid and has been suggested shown to destabilize binding towards Mcl-1.[36,129] Both Glu^{2f} and Glu^{3g} were previously shown to be specific for Bcl-xL.[59] Both Phe^{3d} and Leu^{3d} appear more abundantly in the final round than Ile^{3d}, which has been suggested to drive specificity towards Mcl-1. However, we also achieved specificity while not requiring many previously established mutations, such as Val^{4a} or Lys^{4e}, which drives specificity for Mcl-1 or Bcl-2 respectively but was not randomized in our library (instead, Phe^{4a} and Tyr^{4e}).[36,129] This suggests further improvements could be identified by further combining specificity-driving mutations discovered in this report with those reported elsewhere.

Data from both flow cytometry and next generation sequencing at the library level suggested that peptides were highly specific towards Bcl-x_L. However, to confirm that this was

true for individual peptides, we sampled sequences randomly from the final round of sorting and evaluated their affinity and specificity compared to a known binder towards all 5 proteins, BIM-p5 (**Figure 3.8**).[24] With significantly more specificity towards Mcl-1 and Bfl-1 than Bcl-2 or Bcl-w, these peptide libraries are consistent with reports that specificity between the three Bcl- proteins is a more challenging task and that further improvements to sorting strategy or new mutations are necessary to achieve higher margins of specificity.[36] Two peptides with maximum specificity were chosen for a more thorough analysis beyond the two-concentration binding estimates as predictive of affinity. These peptides showed ~10 nM affinity and > 10-fold specificity for all 4 family members on the bacterial cell surface, demonstrating a successful selection of high affinity and specificity peptides for Bcl-x_L. The ability of SPEED to evaluate the fitness of a peptide with few flow cytometry samples makes it amenable to analyzing hits quickly before more thorough downstream analysis.

While bacterial cell surface experiments suggested that molecules discovered from sorting were high affinity and strongly specific, it may weakly overestimate affinities (i.e. the measured affinity via SPEED may be 1-5 fold higher than that reported via solution phase, see **Figure 2.12**) , and we therefore sought to translate those molecules into solution phase to confirm their binding properties (**Figure 3.11**).[24] After synthesizing, stapling, and purifying the peptides using standard Fmoc chemistry (see Methods), we measured the binding properties of the peptides using two methods: competitive inhibition and biolayer interferometry. Both methods closely agreed and confirmed SPEED measured affinities; both peptides analyzed were extremely specific, having ~10nM affinity for Bcl-x_L but >200nM K_d for Bcl-w and >10,000nM K_d for all others. Biolayer interferometry measurements additionally allow for measure of kinetic parameters, yielding insight into whether the on- or off-rate dominates the specificity differences.

Comparing the binding between Bcl-x_L and Bcl-w suggests that off-rates drive specificity (**Figure 3.12**); k_{on} for Bcl-w and Bcl-x_L are not significantly different (p=0.266) while k_{off} for Bcl-w is more than 40 times faster than that of Bcl-x_L (p = 0.00201). These results confirm previous reports that bacterial surface display measured affinities highly correlate with those from solution phase, whether from competitive inhibition experiments or biolayer interferometry.[24]

We further sought to confirm that the peptides were acting with mechanisms consistent with apoptosis biology. To confirm that the affinities and specificities demonstrated by peptide on or off the bacterial cell surface were not artifacts of soluble versions of the Bcl-2 proteins, and to confirm they act consistently with known apoptosis mechanisms, we measured the permeabilization of mitochondrial outer membranes when titrated with peptides (**Figure 3.13**).[126,127] First, we confirmed that in Bcl-xL overexpressing cell lines (MDA-MB-231 and MCF7), depolarization was the result of Bcl-xL specificity by testing the highly specific peptide (**13**) and a Bcl-xL non-binder (**F2**). Next, we tested whether the peptide depolarized cell lines not driven by Bcl-xL (B-ALL engineered for Mcl-1 or Bfl-1 overexpression. These results confirmed that peptides discovered via SPEED act in accordance with apoptosis biology and are minimally affected by the presentation of soluble ectodomains.

Finally, we applied two structural biology techniques to assay the molecular basis of specificity by solving the structure of the peptide-protein complex. We initially tried to solve the structure of the peptide: protein complex through crystallography, but ultimately they failed to co-crystallize and we then applied nuclear magnetic resonance spectroscopy (NMR). First, we used total correlation spectroscopy (TOCSY) and nuclear overhauser effect spectroscopy (NOESY) to solve the structure of the peptide in isolation. We then used saturation-transfer

difference (STD) to identify the interactions between the peptide and the Bcl-x_L structure. It is the subject of ongoing work to fit a structure to NMR data and use AutoDock Vina to identify the binding confirmation of the **12** peptide using the PDB 1PQ1 template (**Chapter 5**) .

While SPEED yielded high affinity and specificity Bcl-x_L stapled peptides, there are some limitations and future directions for this work. While flow cytometry experiments confirmed improvements in fitness as sorting progressed, we speculate that our campaign could have been improved by incorporating specificity-based sorting earlier on. High affinity and non-specific peptides were more abundant than highly specific but weakly binding peptides; the elimination of non-specific peptides earlier may have yielded specific hits with fewer rounds of sorting. This would likely be more important in future experiments that target Bcl-w and Bcl-2, where there are fewer defined mutations that improve specificity compared to Bcl-x_L, Mcl-1, or Bfl-1.

In conclusion, we used SPEED to engineer high affinity and highly specific Bcl-x_L stapled peptide antagonists. We demonstrated they are highly specific when presented on the cell surface or synthesized in soluble form, act in accordance with apoptosis biology, and have unique structural motifs that enable their high specificity. Future work includes incorporating design rules towards cell permeability and protease stability; the diverse set of peptides that SPEED generated with desired fitness yield numerous molecules that can be translated for these important drug-like properties.[8,18] Additionally, the rich set of sequence information generated towards Bcl-xL could be used to design sequences that are predicted to have high activity in other areas (such as cell permeability). This work suggests that SPEED is a versatile platform towards the generation of potent stapled peptide therapeutics.

Appendices

Table 3.1: Predicted and calculated masses for compounds in this study

Compound	Predicted Unstapled Mass	Observed Unstapled Mass	Predicted Stapled Mass	Observed Stapled Mass
X3	2736.3	2736.3	2830.4	2830.4
W10	2681.3	2681.3	2776.4	2776.4

Table 3.2: Degenerate codons sampled for the bacterial cell surface stapled peptide variant library

	1e	1f	1g	2a	2b	2c	2d	2e	2f	2g	3a	3b	3c	3d	3e	3f	3g	4a	4b	4c	4d	4e	4f
wt	G	R	P	E	I	W	I	A	Q	E	L	R	R	I	G	D	E	F	N	A	Y	Y	A
	ggt	cgc	ccg	gaa	att	tgg	att	gcg	caa	gaa	ttg	cgc	cgc	att	ggt	gac	gaa	ttt	aac	gcg	tat	tat	gcg
p1	M	R	X	X	X	X	X	M	X	E	L	R	R	X	X	D	X	F	X	X	Y	Y	A
	atg	cgc	dya	rna	sda	can	aha	atg	vva	gaa	ttg	cgc	cgc	ntc	gsa	gac	vaa	ttt	vnc	svm	tat	tat	gcg
p2	G	M	X	X	X	X	X	X	M	E	L	R	R	X	X	D	X	F	X	X	Y	Y	A
	ggt	atg	dya	rna	sda	can	aha	gba	atg	gaa	ttg	cgc	cgc	ntc	rsc	gac	vaa	ttt	vnc	smm	tat	tat	gcg
p5	G	R	X	X	M	X	X	X	X	E	L	M	R	X	X	D	X	F	X	X	Y	Y	A
	ggt	cgc	dya	rna	atg	can	awa	gba	vva	gaa	ttg	atg	cgc	ntc	rsc	gac	vaa	ttt	vnc	smm	tat	tat	gcg
p6	G	R	X	X	X	M	X	X	X	E	L	R	M	X	X	D	X	F	X	X	Y	Y	A
	ggt	cgc	dya	rna	sda	atg	aha	gba	vva	gaa	ttg	cgc	atg	ntc	gsa	gac	vaa	ttt	vnc	svm	tat	tat	gcg
p12	G	R	X	X	X	X	X	X	X	E	L	M	R	X	X	D	X	F	M	X	Y	Y	A
	ggt	cgc	myc	rna	sda	cna	aha	gba	vva	gaa	ttg	atg	cgc	ntc	rsc	gac	vaa	ttt	atg	snm	tat	tat	gcg
p13	G	R	X	X	X	X	X	X	X	E	L	R	M	X	X	D	X	F	X	M	Y	Y	A
	ggt	cgc	dya	rna	sda	cna	awa	gba	sva	gaa	ttg	cgc	atg	ntc	rsc	gac	vaa	ttt	vnm	atg	tat	tat	gcg
p14	G	R	X	X	X	X	X	X	X	E	L	R	R	M	X	D	X	F	X	X	M	Y	A
	ggt	cgc	dya	rna	sda	cna	aha	gba	vva	gaa	ttg	cgc	cgc	atg	gsa	gac	vaa	ttt	vnc	svm	atg	tat	gcg

Table 3.3: Primers used to generate the stapled peptide library in bacteria

Primer Name	DNA Sequence
P1	GCTGGCCAGTCTGGCCAGTATCCGTATGATGTGCCGGATTATGCGGGCGGGCAGCGGCGGCCAGAGCATGCGCDYARNASDAVNGAHAATGVNGGAATTGCGCCGCDTSRSCGACVAATTTVNC
	GMCTATTATGCGGGAGGGCAGTCTGGGCAG
P2	GCTGGCCAGTCTGGCCAGTATCCGTATGATGTGCCGGATTATGCGGCGGGCGGCAGCGGCGGCCAGAGCGGTATGDYARNASDAVNGAHAGBAATGGAATTGCGCCGCDTSRSCGACVAATTTVNC
	SMMTATTATGCGGGAGGGCAGTCTGGGCAG
P5	GCTGGCCAGTCTGGCCAGTATCCGTATGATGTGCCGGATTATGCGGCGGGCGGCAGCGGCGGCCAGAGCGGTCGCDYARNAATGVNGAHAGBAVVAGAATTGATGCGCDTSGSAGACVAATTTVNC
	SVMTATTATGCGGGAGGGCAGTCTGGGCAG
P6	GCTGGCCAGTCTGGCCAGTATCCGTATGATGTGCCGGATTATGCGGGCGGGCGGCAGCGGCGGCCAGAAGCGGTCGCDYARNASDAATGAHAGBAVNGGAATTGCGCATGDTSGSAGACVAATTTVNC
	GCGTATMNSGCGGGAGGGCAGTCTGGGCAG
P12	GCTGGCCAGTCTGGCCAGTATCCGTATGATGTGCCGGATTATGCGGCGGGCGGCAGCGGCGGCCAGAGCGGTCGCDYARNASDAVNGAHAGBAVNGGAATTGATGCGCDTSGSAGACVAATTATG
	GCGTATMNSGCGGGAGGGCAGTCTGGGCAG
P13	GCTGGCCAGTCTGGCCAGTATCCGTATGATGTGCCGGATTATGCGGGCGGGCGGCAGCGGCGGCCAGAGCGGTCGCDYARNASDAVNGAHAGBAVNGGAATTGCGCATGDTSGSAGACVAATTTVNC
	ATGTATTATGCGGGAGGGCAGTCTGGGCAG
P14	GCTGGCCAGTCTGGCCAGTATCCGTATGATGTGCCGGATTATGCGGGCGGGCGGCAGCGGCGGCCAGAGCGGTCGCDYARNASDAVNGAHAGBAVNGGAATTGCGCCGCATGRSCGACVAATTTVN
	MGMCATGTATGCGGGAGGGCAGTCTGGGCAG
eCPX reverse	GAGGTCATTACTGGATCTATCAACAGGAGTCCAAGCTCAGC

Table 3.4: Details for magnetic and fluorescent cell sorting

Sort Round	Binding	Expression	Off-Target
MACS expression	N/A	1st: anti-HA magnetic beads	N/A
MACS 1	1st: 100nM biotin-target + streptavidin magnetic beads	N/A	N/A
MACS 2	1st: 100nM biotin-target + streptavidin magnetic beads	N/A	N/A
FACS 1	1st: 100nM biotin-target 2nd: 1:100 neutravidin-DL488	1st: 1:100 anti-HA mouse 2nd: 1:100 chicken anti-mouse-AF647	N/A
FACS 2	1st: 10nM biotin-target 2nd: 1:100 neutravidin-DL488	1st: 1:100 anti-HA mouse 2nd: 1:100 chicken anti-mouse-AF647	N/A
FACS 3 competitive	1st: 100nM AF647-target	1st: 1:100 anti-HA mouse 2nd: 1:100 goat anti-mouse-AF488	1st: 25nM biotin-competitors 2nd: 1:100 neutravidin-DL405
FACS 4 competitive	1st: 10nM AF647-target	1st: 1:100 anti-HA mouse 2nd: 1:100 goat anti-mouse-AF405	1st: 25nM x 4 biotin-competitors 2nd: 1:100 neutravidin-DL488
FACS 3 negative	N/A	1st: 1:100 anti-HA mouse 2nd: 1:100 goat anti-mouse-AF405	1st: 25nM x 4 biotin-competitors 2nd: 1:100 neutravidin-DL488
FACS 4 positive	1st: 10nM AF647-target	1st: 1:100 anti-HA mouse 2nd: 1:100 goat anti-mouse-AF405	N/A

Table 3.5: Nuclear magnetic resonance spectroscopy parameters

NMR Instrument	Bruker Avance II 600 MHz
Buffer	Phosphate Buffered Saline, pH 7.4 + 10% D2O
Peptide Concentration	1.52mM
Stoichiometric STD Ratio	3:1 Peptide: Protein
TOCSY Mixing Time	80ms
NOESY Mixing Time	150ms
TOCSY Scan #	32
NOESY Scan #	64

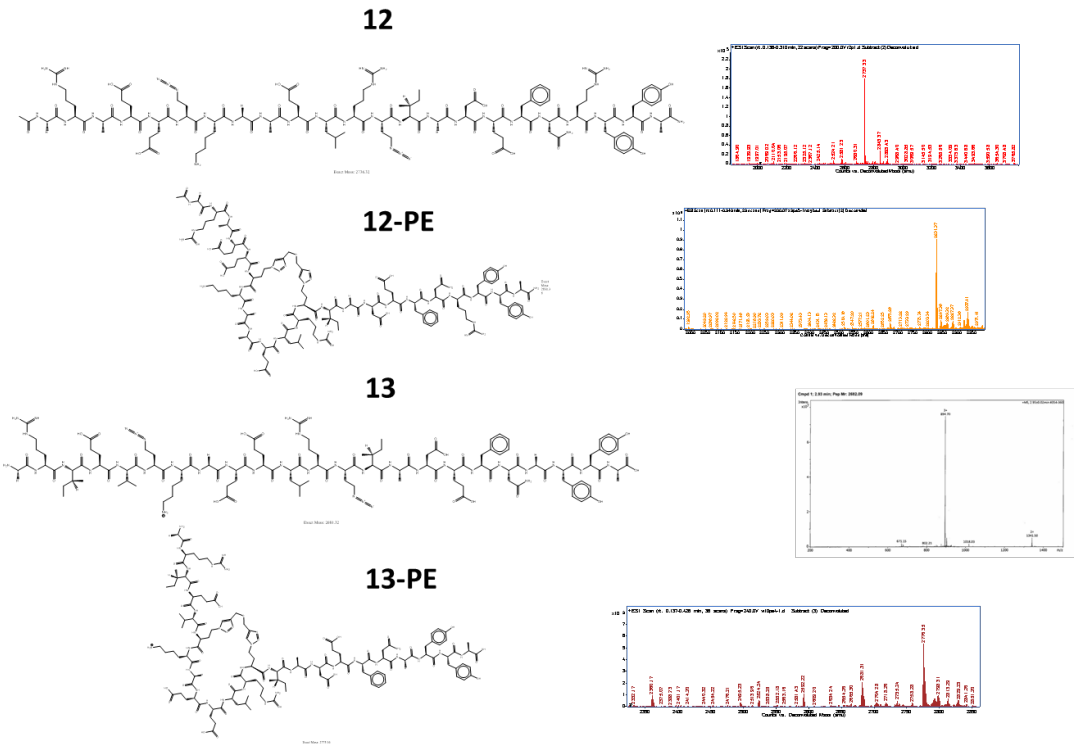


Figure 3.14: Structures and mass spectra from compounds generated in this study

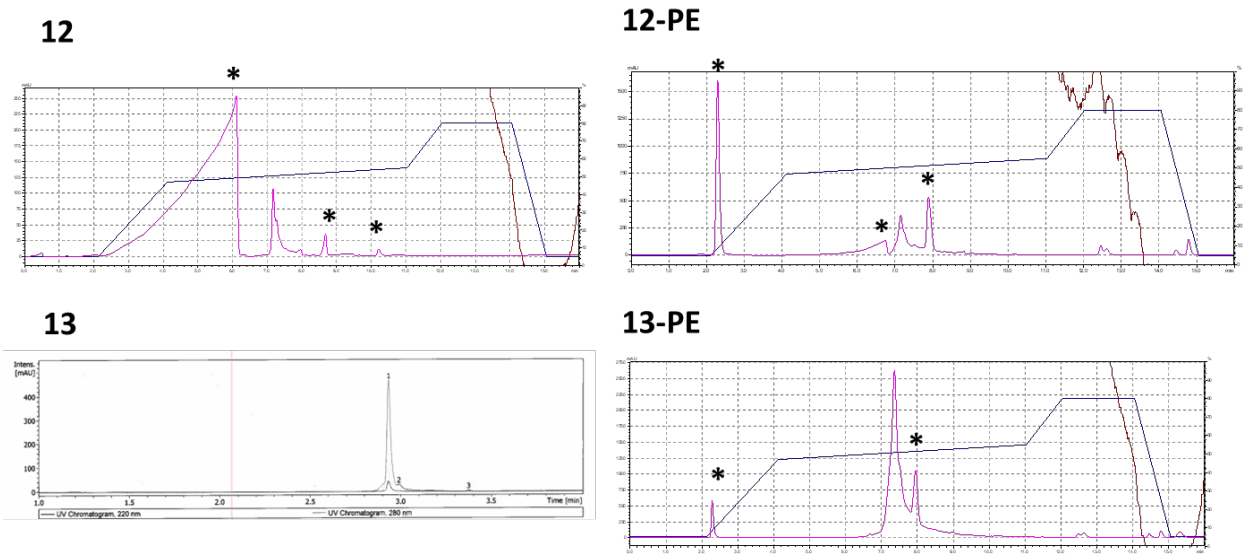


Figure 3.15: Reverse phase high performance liquid chromatography traces for compounds generated in this study

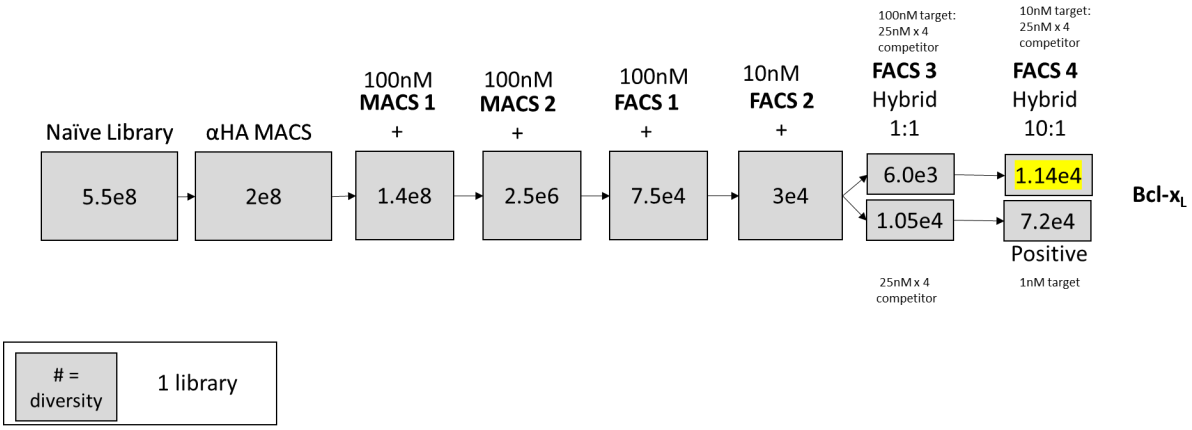


Figure 3.16: Sorting progression of bacterial surface library

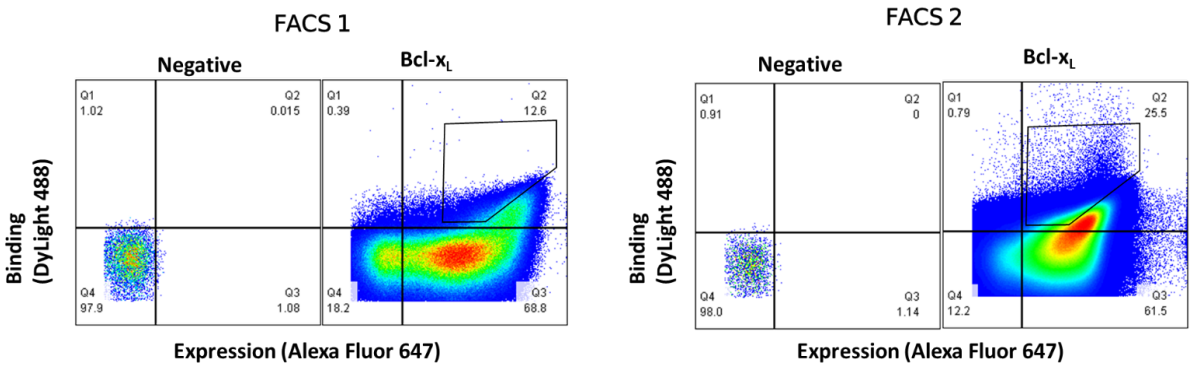


Figure 3.17: Representative fluorescent activated cell sorting (FACS) diagrams for FACS 1 and 2

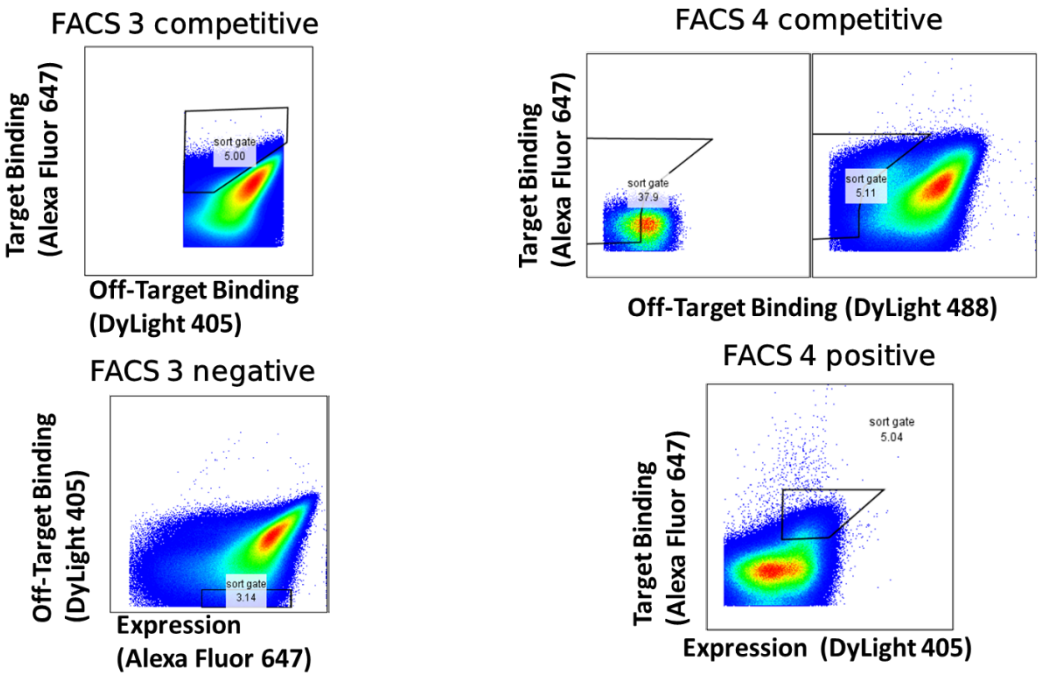
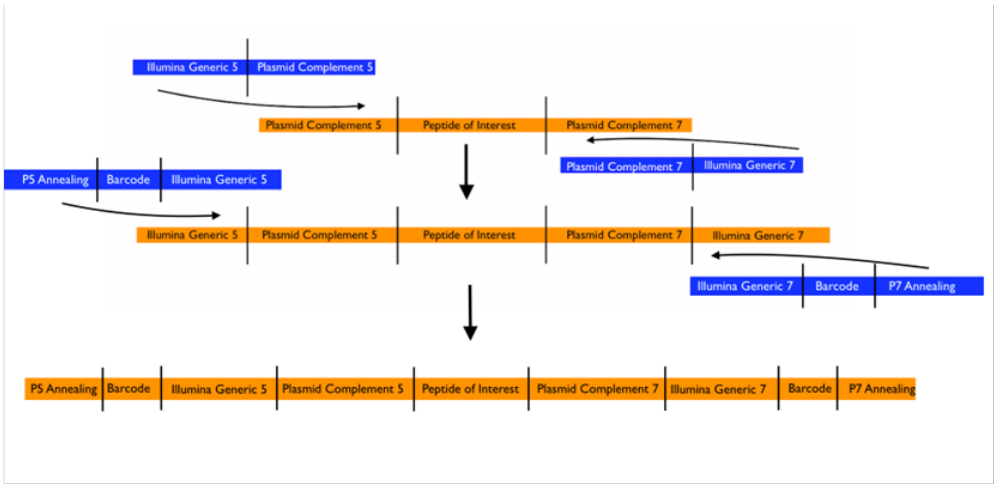


Figure 3.18 Representative fluorescent activated cell sorting (FACS) diagrams for FACS 3 and 4



1st round forward	TCG TCG GCA GCG TCA GAT GTG TAT AAG AGA CAG CAG GTA CTT CCG TAG CTG GCC AGT CT
1st reverse	GTC TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GCA CCG TAG ATG CTT GCC CAG TCG TTA
NGS 5-0	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT AGA TCG CTC GTC GGC AGC GTC
NGS 5-1	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACC TCT CTA TTT CGT CGG CAG CGT C
NGS 5-2	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT ATC CTC TGT TCG TCG GCA GCG TC
NGS 5-3	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACA GAG TAG ACG ATC GTC GGC AGC GTC
NGS 5-4	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACG TAA GGA GAT GAT CGT CGG CAG CGT C
NGS 5-5	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACA CTG CAT ATG CGA TCG TCG GCA GCG TC
NGS 5-6	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACA AGG AGT AGA GTG GTC GTC GGC AGC GTC
NGS 5-7	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACC TAA GCC TCC TGT GGT CGT CGG CAG CGT C
NGS 7-0	CAA GCA GAA GAC GGC ATA CGA GAT TAA GGC GAG TCT CGT GGG CTC GG
NGS 7-1	CAA GCA GAA GAC GGC ATA CGA GAT CGT ACT AGA GTC TCG TGG GCT CGG
NGS 7-2	CAA GCA GAA GAC GGC ATA CGA GAT AGG CAG AAT CGT CTC GTG GGC TCG G
NGS 7-3	CAA GCA GAA GAC GGC ATA CGA GAT TCC TGA GCC TAG TCT CGT GGG CTC GG
NGS 7-4	CAA GCA GAA GAC GGC ATA CGA GAT GGA CTC CTG ATA GTC TCG TGG GCT CGG
NGS 7-5	CAA GCA GAA GAC GGC ATA CGA GAT TAG GCA TGA CTC AGT CTC GTG GGC TCG G
NGS 7-6	CAA GCA GAA GAC GGC ATA CGA GAT CTC TCT ACT TCT CTG TCT CGT GGG CTC GG
NGS 7-7	CAA GCA GAA GAC GGC ATA CGA GAT CAG AGA GGC ACT TCT GTC TCG TGG GCT CGG

Figure 3.19: Next generation sequencing scheme and primers used in this study

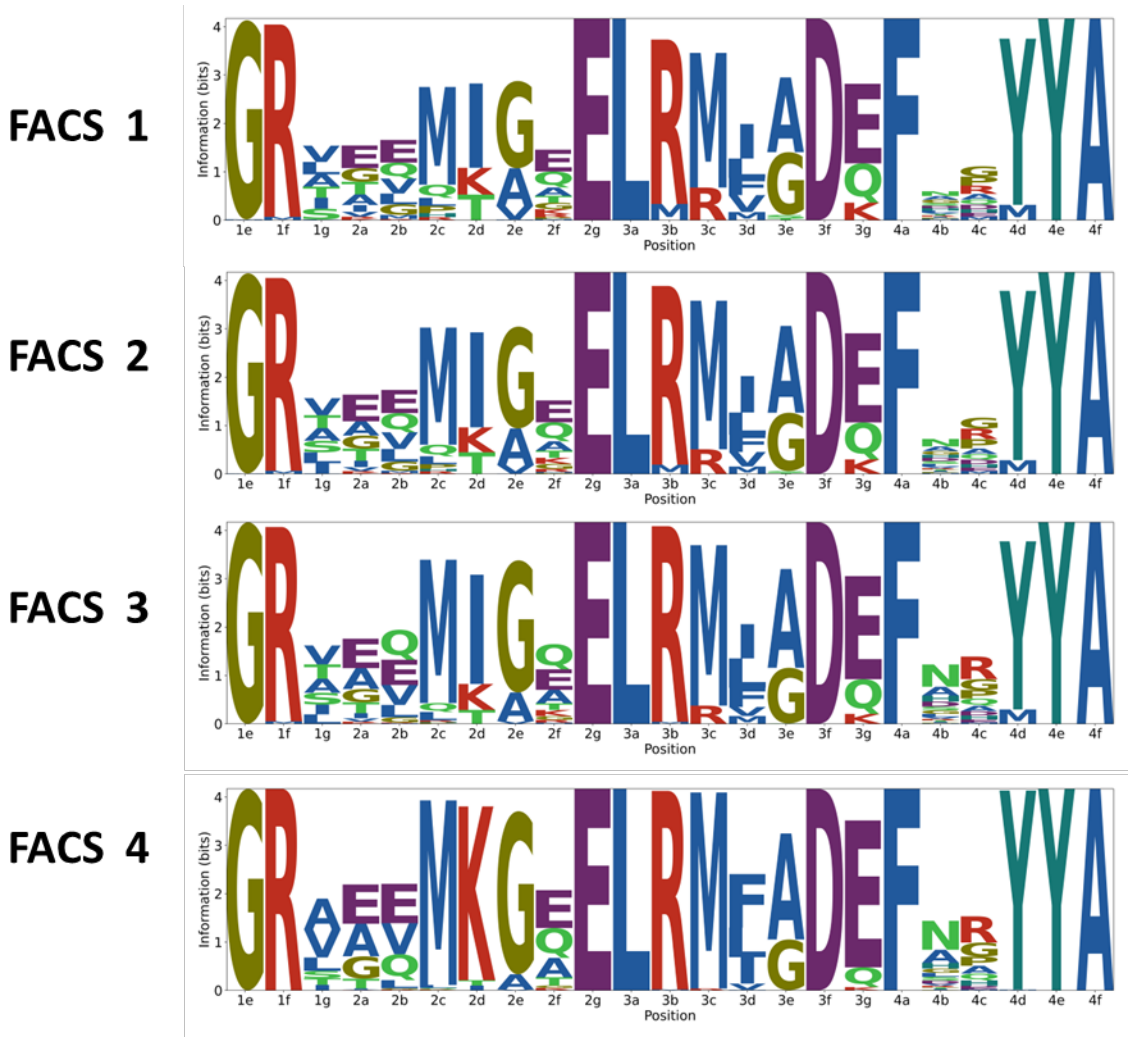


Figure 3.20: Logoplots for FACS rounds 1-4 from NGS analyses

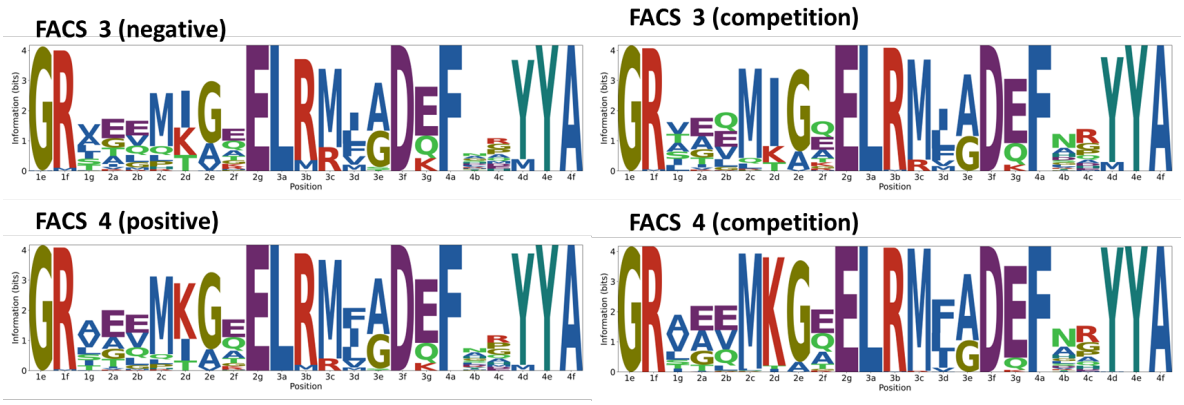


Figure 3.21: Comparison of logoplots for negative/ positive FACS versus competitive binding FACS

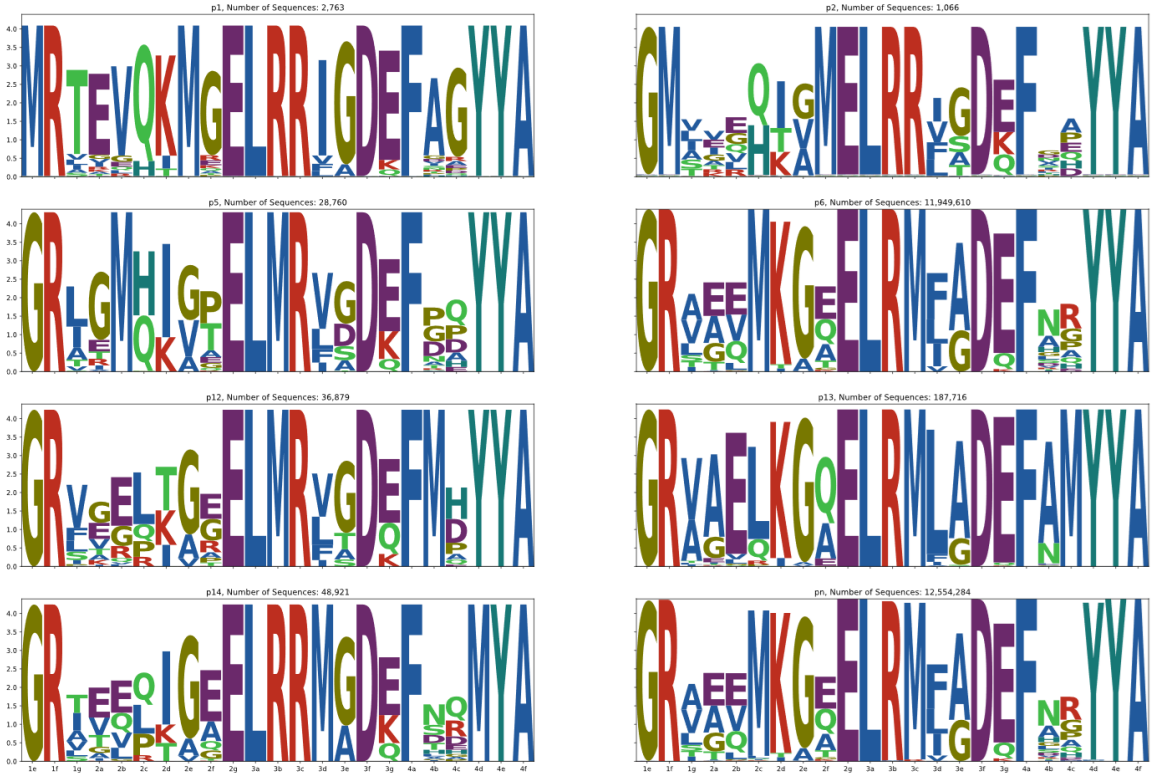


Figure 3.22: Logoplots for FACS 4 binders when restricted to particular staple locations

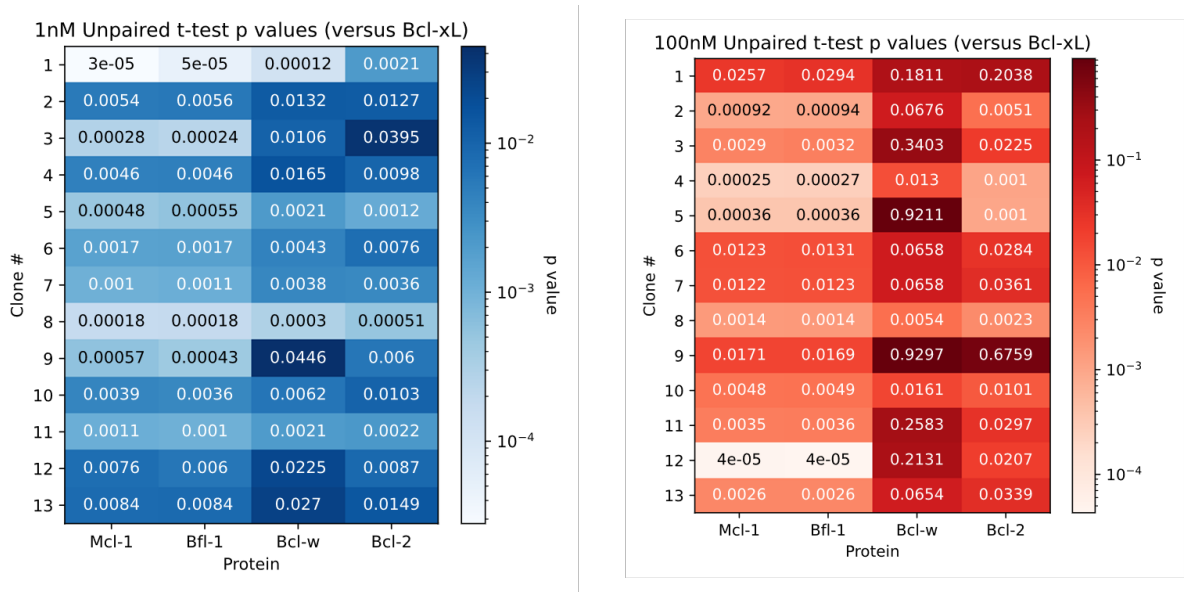


Figure 3.23: P-values for all single concentration binding assays for FACS 4 stapled peptides. The p-value is calculated as a unpaired t-test.

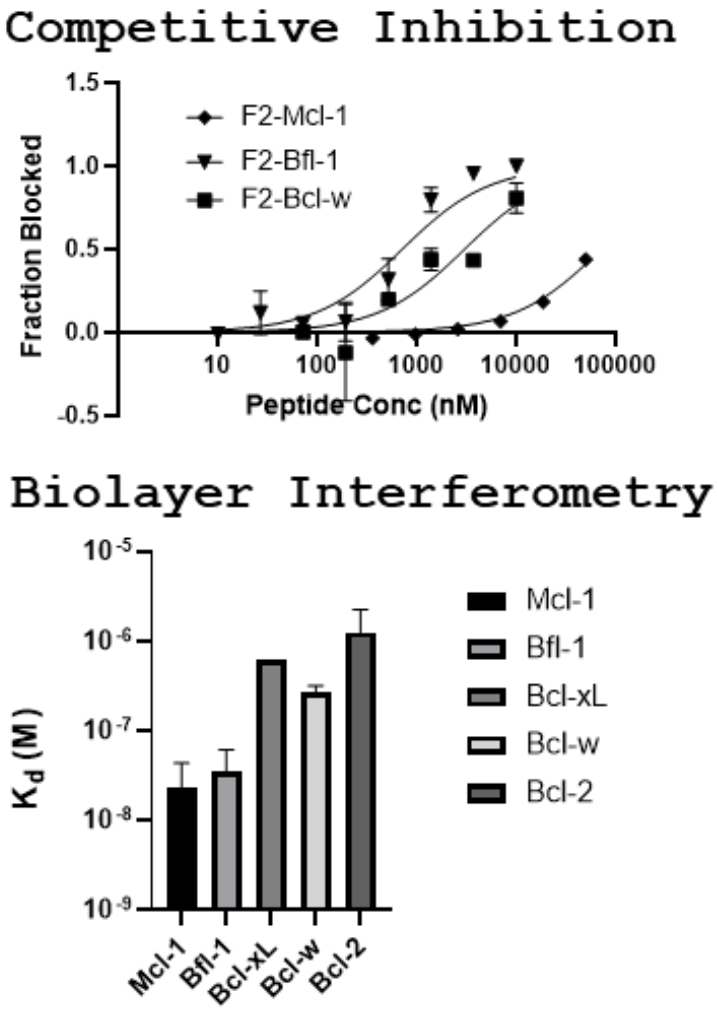


Figure 3.24: Competitive inhibition and biolayer interferometry data for F2, a Bcl-xL non binding stapled peptide used for comparison in Figure 6.

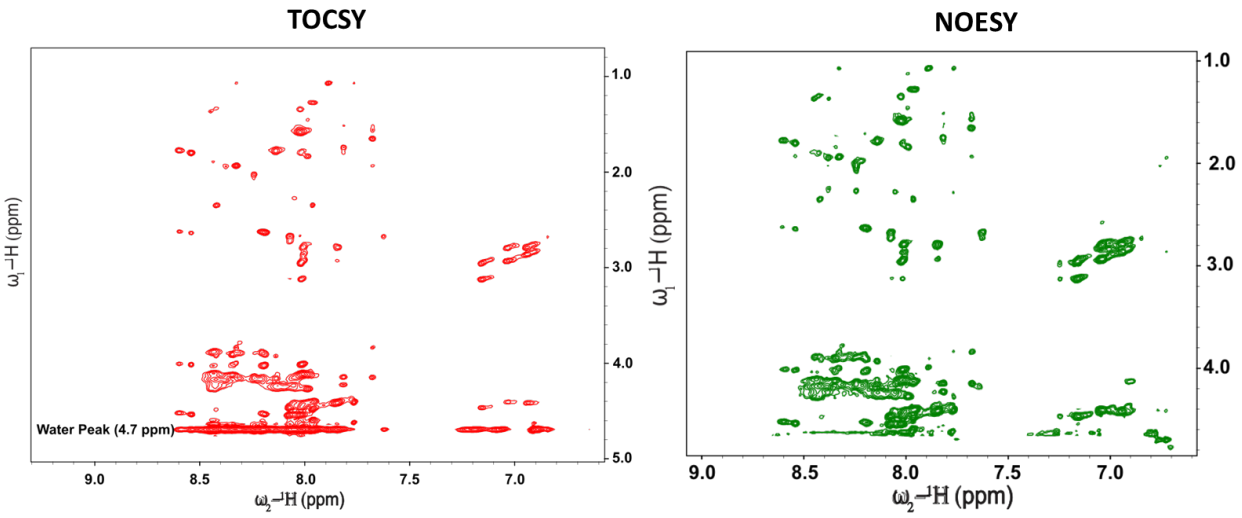


Figure 3.25: TOCSY and NOESY nuclear magnetic resonance spectroscopy results.

References

1. Danial, N. N. & Korsmeyer, S. J. Cell Death: Critical Control Points. *Cell* **116**, 205–219 (2004).
2. Certo, M. *et al.* Mitochondria primed by death signals determine cellular addiction to antiapoptotic BCL-2 family members. *Cancer Cell* **9**, 351–365 (2006).
3. Debatin, K. M. Apoptosis pathways in cancer and cancer therapy. *Cancer Immunology, Immunotherapy* **53**, 153–159 (2004).
4. Kale, J., Osterlund, E. J. & Andrews, D. W. BCL-2 family proteins: Changing partners in the dance towards death. *Cell Death Differ* **25**, 65–80 (2018).
5. Czabotar, P. E., Lessene, G., Strasser, A. & Adams, J. M. Control of apoptosis by the BCL-2 protein family: Implications for physiology and therapy. *Nat Rev Mol Cell Biol* **15**, 49–63 (2014).
6. D'Aguanno, S. & Del Bufalo, D. Inhibition of Anti-Apoptotic Bcl-2 Proteins in Preclinical and Clinical Studies: Current Overview in Cancer. *Cells* **9**, (2020).
7. Ukrainskaya, V. M. *et al.* Death Receptors: New Opportunities in Cancer Therapy. *Acta Naturae* **9**, 55–64 (2017).
8. Adams, J. M. & Cory, S. The Bcl-2 Protein Family: Arbiters of Cell Survival. *Science (1979)* **281**, 1322–1326 (1998).
9. Chen, L. *et al.* Differential targeting of prosurvival Bcl-2 proteins by their BH3-only ligands allows complementary apoptotic function. *Mol Cell* **17**, 393–403 (2005).
10. Deng, J. *et al.* BH3 Profiling Identifies Three Distinct Classes of Apoptotic Blocks to Predict Response to ABT-737 and Conventional Chemotherapeutic Agents. *Cancer Cell* **12**, 171–185 (2007).

11. Liu, X., Dai, S., Zhu, Y., Marrack, P. & Kappler, J. W. The structure of a Bcl-xL/Bim fragment complex: Implications for Bim function. *Immunity* **19**, 341–352 (2003).

12. Dutta, S. *et al.* Potent and specific peptide inhibitors of human pro-survival protein bcl-xl. *J Mol Biol* **427**, 1241–1253 (2015).

13. Lucianò, A. M., Pérez-Oliva, A. B., Mulero, V. & Bufalo, D. Del. Bcl-xl: A focus on melanoma pathobiology. *Int J Mol Sci* **22**, 1–17 (2021).

14. Trisciuoglio, D. *et al.* BCL-XL overexpression promotes tumor progression-associated properties article. *Cell Death Dis* **8**, (2017).

15. Um, H.-D. Bcl-2 family proteins as regulators of cancer cell invasion and metastasis: a review focusing on mitochondrial respiration and reactive oxygen species. *Oncotarget* **7**, 5193–5203 (2015).

16. Moore, V. D. G. *et al.* Chronic lymphocytic leukemia requires BCL2 to sequester prodeath BIM, explaining sensitivity to BCL2 antagonist ABT-737. *Journal of Clinical Investigation* **117**, 112–121 (2007).

17. Chonghaile, T. N. *et al.* Pretreatment mitochondrial priming correlates with clinical response to cytotoxic chemotherapy. *Science (1979)* **334**, 1129–1133 (2011).

18. Leverson, J. D. *et al.* Exploiting selective BCL-2 family inhibitors to dissect cell survival dependencies and define improved strategies for cancer therapy. *Sci Transl Med* **7**, 1–12 (2015).

19. Bose, P., Gandhi, V. & Konopleva, M. Pathways and mechanisms of venetoclax resistance. *Leuk Lymphoma* **58**, 2026–2039 (2017).

20. Tahir, S. K. *et al.* Potential mechanisms of resistance to venetoclax and strategies to circumvent it. *BMC Cancer* **17**, 1–10 (2017).

21. Leverson, J. D. *et al.* Found in translation: How preclinical research is guiding the clinical development of the BCL2-selective inhibitor venetoclax. *Cancer Discov* **7**, 1376–1393 (2017).

22. Oltersdorf, T. *et al.* An inhibitor of Bcl-2 family proteins induces regression of solid tumours. *Nature* **435**, 677–681 (2005).

23. Tao, Z. F. *et al.* Discovery of a potent and selective BCL-XL inhibitor with in vivo activity. *ACS Med Chem Lett* **5**, 1088–1093 (2014).

24. Wang, L. *et al.* Discovery of A-1331852, a First-in-Class, Potent, and Orally-Bioavailable BCL-XL Inhibitor. *ACS Med Chem Lett* **11**, 1829–1836 (2020).

25. Foight, G. W. & Keating, A. E. Locating Herpesvirus Bcl-2 Homologs in the Specificity Landscape of Anti-Apoptotic Bcl-2 Proteins. *J Mol Biol* **427**, 2468–2490 (2015).

26. DeBartolo, J., Taipale, M. & Keating, A. E. Genome-Wide Prediction and Validation of Peptides That Bind Human Prosurvival Bcl-2 Proteins. *PLoS Comput Biol* **10**, 1–10 (2014).

27. Dutta, S., Chen, T. S. & Keating, A. E. Peptide ligands for pro-survival protein Bfl-1 from computationally guided library screening. *ACS Chem Biol* **8**, 778–788 (2013).

28. Fu, X., Apgar, J. R. & Keating, A. E. Modeling Backbone Flexibility to Achieve Sequence Diversity: The Design of Novel α-Helical Ligands for Bcl-xL. *J Mol Biol* **371**, 1099–1117 (2007).

29. Debartolo, J., Dutta, S., Reich, L. & Keating, A. E. Predictive Bcl-2 family binding models rooted in experiment or structure. *J Mol Biol* **422**, 124–144 (2012).

30. London, N., Gullá, S., Keating, A. E. & Schueler-Furman, O. In silico and in vitro elucidation of BH3 binding specificity toward Bcl-2. *Biochemistry* **51**, 5841–5850 (2012).

31. Fire, E., Gullá, S. v., Grant, R. A. & Keating, A. E. Mcl-1-Bim complexes accommodate surprising point mutations via minor structural changes. *Protein Science* **19**, 507–519 (2010).

32. Jenson, J. M., Ryan, J. A., Grant, R. A., Letai, A. & Keating, A. E. Epistatic mutations in PUMA BH3 drive an alternate binding mode to potently and selectively inhibit anti-apoptotic Bfl-1. *Elife* **6**, 1–23 (2017).

33. Foight, G. W., Ryan, J. A., Gullá, S. v., Letai, A. & Keating, A. E. Designed BH3 peptides with high affinity and specificity for targeting Mcl-1 in cells. *ACS Chem Biol* **9**, 1962–1968 (2014).

34. Jenson, J. M. *et al.* Peptide design by optimization on a data parameterized protein interaction landscape. *Proc Natl Acad Sci U S A* **115**, E10342–E10351 (2018).

35. Dutta, S. *et al.* Determinants of BH3 Binding Specificity for Mcl-1 versus Bcl-xL. *J Mol Biol* **398**, 747–762 (2010).

36. Walensky, L. D. & Bird, G. H. Hydrocarbon-stapled peptides: principles, practice, and progress. *J Med Chem* **57**, 6275–6288 (2014).

37. Reich, L., Dutta, S. & Keating, A. E. SORTCERY - A High-Throughput Method to Affinity Rank Peptide Ligands. *J Mol Biol* **427**, 2135–2150 (2015).

38. Loren D. Walensky *et al.* Activation of Apoptosis in Vivo by a Hydrocarbon-Stapled BH3 Helix. *Science (1979)* **23**, 1–7 (2004).

39. Schafmeister, C. E., Po, J. & Verdine, G. L. An all-hydrocarbon cross-linking system for enhancing the helicity and metabolic stability of peptides. *J Am Chem Soc* **122**, 5891–5892 (2000).

40. Bluntzer, M. T. J., O'Connell, J., Baker, T. S., Michel, J. & Hulme, A. N. Designing stapled peptides to inhibit protein-protein interactions: An analysis of successes in a rapidly changing field. *Peptide Science* **113**, 1–17 (2020).

41. Araghi, R. R. *et al.* Iterative optimization yields Mcl-1–targeting stapled peptides with selective cytotoxicity to Mcl-1–dependent cancer cells. *Proc Natl Acad Sci U S A* **115**, E886–E895 (2018).

42. Lama, D. *et al.* Structural insights reveal a recognition feature for tailoring hydrocarbon stapled-peptides against the eukaryotic translation initiation factor 4E protein. *Chem Sci* **10**, 2489–2500 (2019).

43. Chu, Q. *et al.* Towards understanding cell penetration by stapled peptides. *Medchemcomm* **6**, 111–119 (2015).

44. Phillips, C. *et al.* Design and structure of stapled peptides binding to estrogen receptors. *J Am Chem Soc* **133**, 9696–9699 (2011).

45. Lau, Y. H. *et al.* Investigating peptide sequence variations for 'double-click' stapled p53 peptides. *Org Biomol Chem* **12**, 4074–4077 (2014).

46. Wu, Y. *et al.* Toolbox of Diverse Linkers for Navigating the Cellular Efficacy Landscape of Stapled Peptides. *ACS Chem Biol* **14**, 526–533 (2019).

47. Song, J. M., Gallagher, E. E., Menon, A., Mishra, L. D. & Garner, A. L. The role of olefin geometry in the activity of hydrocarbon stapled peptides targeting eukaryotic translation initiation factor 4E (eIF4E). *Org Biomol Chem* **17**, 6414–6419 (2019).

48. Kawamoto, S. *et al.* Design of Triazole-Stapled BCL9 α-Helical Peptides to Target the β-Catenin/B-Cell CLL/lymphoma 9 (BCL9) Protein−Protein Interaction. *J Med Chem* **55**, 1137–1146 (2011).

49. Stewart, M. L., Fire, E., Keating, A. E. & Walensky, L. D. The MCL-1 BH3 helix is an exclusive MCL-1 inhibitor and apoptosis sensitizer. *Nat Chem Biol* **6**, 595–601 (2010).

50. Chang, Y. S. *et al.* Stapled α-helical peptide drug development: A potent dual inhibitor of MDM2 and MDMX for p53-dependent cancer therapy. *Proc Natl Acad Sci U S A* **110**, (2013).

51. Case, M., Navaratna, T., Vinh, J. & Thurber, G. M. Rapid Evaluation of Staple Placement in Stabilized Alpha Helices using Bacterial Surface Display. *ACS Chem Biol* **18**, 905–914 (2023).

52. Navaratna, T. *et al.* Directed Evolution Using Stabilized Bacterial Peptide Display. *J Am Chem Soc* **142**, 1882–1894 (2020).

53. Chen, T. S., Palacios, H. & Keating, A. E. Structure-based redesign of the binding specificity of anti-apoptotic Bcl-xL. *J Mol Biol* **425**, 171–185 (2013).

54. Jacobs, T. M., Yumerefendi, H., Kuhlman, B. & Leaver-Fay, A. SwiftLib: Rapid degenerate-codon-library optimization through dynamic programming. *Nucleic Acids Res* **43**, 1–10 (2015).

55. Micsonai, A. *et al.* BeStSel: Webserver for secondary structure and fold prediction for protein CD spectroscopy. *Nucleic Acids Res* **50**, W90–W98 (2022).

56. Gaspar, J. M. NGmerge: Merging paired-end reads via novel empirically-derived models of sequencing errors. *BMC Bioinformatics* **19**, 1–9 (2018).

57. Tareen, A. & Kinney, J. B. Logomaker: Beautiful sequence logos in Python. *Bioinformatics* **36**, 2272–2274 (2020).

58. Ryan, J. & Letai, A. BH3 profiling in whole cells by fluorimeter or FACS. *Methods* **61**, 156–164 (2013).

59. Fraser, C., Ryan, J. & Sarosiek, K. BH3 Profiling: A Functional Assay to Measure Apoptotic Priming and Dependencies. *Methods in Molecular Biology* **1877**, 61–76 (2019).

60. Koss, B. *et al.* Defining specificity and on-target activity of BH3-mimetics using engineered B-ALL cell lines. *Oncotarget* **7**, 11500–11511 (2016).

61. Smith, M., Case, M., Makowski, E. & Tessier, P. Position specific enrichment ratio matrix predicts antibody variant properties. *Bioinformatics* (2023).

62. Dutta, S. Determinants of BH3 binding specificity for Mcl-1 vs. Bcl-xL. *J Mol Biol* **398**, 747–762 (2011).

63. Bird, G. H. *et al.* Biophysical determinants for cellular uptake of hydrocarbon-stapled peptide helices. *Nat Chem Biol* **12**, 845–852 (2016).

64. Bird, G. H. *et al.* Hydrocarbon double-stapling remedies the proteolytic instability of a lengthy peptide therapeutic. *Proceedings of the National Academy of Sciences* **107**, 14093–14098 (2010).

Chapter 4 Machine Learning to Predict Continuous Protein Properties from Simple Binary Sorting and Deep Sequencing Data

This chapter is derived from the following publication:

Marshall Case, Matthew Smith, Jordan Vinh, and Greg Thurber. "Machine Learning to Predict Continuous Protein Properties from Binary Cell Sorting Data and Map Unseen Sequence Space." *Manuscript Submitted*.

Abstract

Proteins are a diverse class of biomolecules responsible for wide-ranging cellular functions, from catalyzing reactions and recognizing pathogens to forming dynamic cellular structure. The ability to evolve proteins rapidly and inexpensively towards improved properties is a common objective for protein engineers. Powerful high-throughput methods like fluorescent activated cell sorting (FACS) and next-generation sequencing (NGS) have dramatically improved directed evolution experiments. However, it is unclear how to best leverage this data to characterize protein fitness landscapes more completely and identify lead candidates. In this work, we develop a simple yet powerful framework to improve protein optimization by predicting continuous protein properties from simple directed evolution experiments using interpretable machine learning. Evaluated across five diverse protein engineering tasks, continuous properties are consistently predicted from readily available deep sequencing data. To prospectively test the utility of this approach, we generated a library of stapled peptides and applied the framework to predict and optimize both affinity and specificity. We coupled integer

linear programming with interpretable machine learning model coefficients to identify new variants from experimentally unseen sequence space that have desired properties. This approach represents a versatile tool for improved analysis and identification of protein variants across many domains of protein engineering.

Introduction

A longstanding goal of biochemistry has been to map the sequence of a protein to its structure and function.[1] However, the complex biophysics that govern the protein fitness landscape, including how a protein folds and how its structure influences function, make the coupling of sequence to function an extremely difficult task. Protein engineers thus often focus on a much smaller subdomain of the protein fitness landscape, using the confined resources of experimental protein science to explore variants close to a known functional protein with the goal of incrementally improving function. A common and extremely powerful approach is directed evolution, where a protein is encoded by DNA, expressed by cells, and assayed by magnetic or fluorescent activated cell sorting (MACS or FACS) and, more recently, next generation sequencing (NGS) to identify variants with improved fitness. While these techniques represent powerful tools in the protein engineering arsenal, it is unclear how to best leverage information from deep sequencing towards the optimization of protein variants. A method capable of generating both fitness estimates from directed evolution experiments and predictions of sequences with higher activity would greatly expand the power and efficiency of directed evolution experiments.

The combination of directed evolution and next generation sequencing (NGS) has enabled protein engineers to rapidly evaluate millions to billions of protein variants in a highly focused manner. With maintenance of the genotype-phenotype connection, any technique that manipulates

DNA in a high-throughput manner can be applied to design focused protein variant libraries and assay protein function.[130,131] Techniques like mRNA display and phage display can evaluate the largest libraries, although their small size precludes them from sorting approaches such as FACS.[132] Cell surface display techniques, which use bacteria or yeast, enable facile measurement of the interaction between protein variants with soluble proteins which can be used for assaying binding affinity in high-throughput sorting and sequencing technologies.[133] Coupling FACS with cell surface display technologies allows for the selection of rare protein variants among a large library with extreme selectivity.[20,134] These techniques have enabled a wide range of protein engineering campaigns, from affinity maturation of protein-protein interactions to highly enantioselective enzymes.[135,136] However, one challenge with these large libraries is how to identify the best lead molecules from the hundreds to thousands of observed sequences in the final sorted population. Traditional approaches for lead molecule identification select variants according to their abundance in the enriched library under the assumption that higher enrichment is indicative of higher function.[137–139] One downside to this approach is that optimal rare variants are excluded from selection and more complex descriptions of how mutations contribute to protein function are difficult to ascertain.[138] Application of NGS to the output pool of a protein variant sort improves the accuracy of clone frequency, but frequency rarely correlates with protein properties directly.[140–142] These challenges arise from sources of error that are difficult to eliminate: variation in cell-to-cell growth, PCR/cloning biases, sequencing errors, and FACS instrument noise.[25,143] With additional sequencing of the input library, enrichment ratios can be calculated, which improves the accuracy of protein property prediction.[144,145] Despite these improvements, there is still little consensus on the best experimental design and analysis of these directed evolution experiments.

Several approaches have been proposed to mitigate these sources of error and enable the prediction of quantitative protein properties from high-throughput sorting experiments. Deep mutational scanning (DMS) measures the enrichment of many variants. However, several challenges exist; their accuracy in resolving affinity is often limited to a narrow linear region (~10X dynamic range), the results are sensitive to the sorting conditions, stability, and expression effects, and the outcomes can differ from true quantitative measurements of binding affinity (equilibrium dissociation constants or K_D's).[146,147] Sort-seq aims to address noise from sorting by using multiple bins across the entire fluorescent channel, followed by deep sequencing, to infer the distribution of each sequence in fluorescent space.[148] These techniques, while often successful, require more sorting time and 8-12 fold increased deep sequencing throughput and still have a narrow range of resolution. Several more sophisticated sorting techniques address these issues: SORTCERY creates a rank ordering of affinities by sorting cells according to their binding and expression at a single concentration;[61] amped SORTCERY further improves this technique by converting rank order to free energy changes by adding titration standards;[37] TiteSeq sorts protein variants at multiple ligand concentrations and fits the affinity to the fraction bound.[147] These methods leverage additional sorting and sequencing to improve the predicted outcomes. In this work, we seek to utilize deep sequencing with interpretable machine learning approaches to determine if we can predict continuous protein properties (like affinity) from binary sorting data (positive versus negative sorting).

Methods

4.1.1 Curation of NGS Data for Validation

Five datasets were used to test the simple method of using binary labels to predict continuous properties (**Table 4.1**). The datasets and brief descriptions are given below.

1. ***Adams et al. 2016***[147]

   NGS data was downloaded from their GitHub repository: https://github.com/jbkinney/16_titeseq. The read counts and CDR^{1H} and CDR^{3H} sequences for each clone were extracted and aligned using in-house python scripts. Read counts were converted to frequencies.

2. ***Starr et al. 2022 and Greaney et al. 2021***[149,150]

   NGS data was downloaded from their GitHub repository: https://github.com/jbloomlab/SARS-CoV-2-RBD_DMS_variants. The data for the Delta mutation is stored in a different repository: https://github.com/jbloomlab/SARS-CoV-2-RBD_Delta. Due to limits in Illumina paired end reading length, each sequence was given a unique molecular barcode, which was sequenced in high depth, but each full-length sequence was sequenced with its unique barcode separately. The sequences and their TiteSeq profiles were associated with their corresponding barcodes and read counts were converted to frequencies. In the current method, sequences with more than one mutation were not discarded.

3. ***Makowski et al. 2022***[151]

   Processed data was downloaded from their GitHub repository: https://github.com/Tessier-Lab-UMich/Emi_Pareto_Opt_ML. Raw data was available from their repository.

4. ***Sarkisyan et al. 2016***[22]

Like Starr et al. 2022, the GFP sequence is too long for high-depth Illumina sequencing, and therefore the authors gave each sequence a unique molecular barcode. We downloaded the accurate full length protein sequences, their matching unique barcodes, and the high-depth sequencing of Sort-seq data from their repository: https://figshare.com/articles/dataset/Local_fitness_landscape_of_the_green_fluorescent_protein/3102154. The read accuracy on the barcodes was low and the authors used a Levenshtein distance of <= 1 to connect barcodes that were close but not identical to the full protein sequence. We used the Levenshtein module with in-house python scripts to cluster sequences to their barcodes, which were available at https://pypi.org/project/python-Levenshtein/. After clustering, sequences and their barcodes were merged with their Sort-Seq distributions like Starr et al. Read counts were converted to frequencies.

5. ***Jenson et al. 2018***[37]

   NGS data was obtained from their GitHub repository: https://github.com/KeatingLab/sortcery_design. The peptides' short lengths permitted high depth deep sequencing and thus counts were directly converted to frequencies without further preprocessing.

4.1.2 Binarization of FACS/NGS Data

The variety of factors that influence the design of an experiment makes it challenging to generalize a sorting and sequencing workflow for any given protein engineering campaign. Each of these projects were analyzed by a different group, using different cell sorters, expression platforms, sequencing instruments, and protein types among other parameters (see **Table 4.1** for dataset property summaries). Thus, controlling each of those parameters in our data processing

workflow was an important consideration towards the application of this approach to existing datasets and new targets alike. Many of the experiments use sophisticated sorting techniques to infer quantitative protein properties. We simulated a simple binary sort experiment by truncating the dataset such that it only includes the top or bottom 20% of sorted sequences (or as close as possible). This subsample of sequencing data approximates a simple sorting campaign from these quantitative sorting techniques. For example, **Sarkisyan *et al.*** contains sequencing data of GFP variants that were sorted into 8 bins; to simulate a simple binary sort, we aggregated the top two bins as positive and the bottom two bins as negative. For TiteSeq experiments (Starr 2022, Greaney 2021, and Adams 2016) we only included data from sorts that used ligand concentrations near the average K_D of the library (10^{-9}, 10^{-9} , and 10^{-8} M respectively). Because the K_D of a library can be readily obtained from low-throughput flow cytometry experiments, sorting at the K_D of the library is a feasible approach to yield the largest separation between high and low affinity variants.[143] This was 10^{-8} M for the COVID datasets, this was 10^{-8} M and 10^{-9} M, respectively. For the **Makowski dataset**, data was provided as a positive and negative dataset with varying cutoffs for each selection.

4.1.3 Machine Learning Method

In all cases, in-house python scripts were used to perform the data preparation and modeling on each of the datasets. Scikit-learn (https://scikit-learn.org/stable/) was used for linear discriminant analysis (LDA), one-hot encoding, scaling label vectors, and other pre-processing steps. Pandas (https://pandas.pydata.org/) and NumPy (https://numpy.org/) were used to handle sequencing and numerical data. PyTorch (https://pytorch.org/) was used to train neural network models.

First, sequences were one-hot encoded, eliminating positions that were not randomized in the study or appeared with very low abundance. Then, we calculated the frequencies of each sequence for the high- and low- protein property, and a multi-sequence alignment (MSA) was performed to ensure every vector was the same length and columns corresponded to the correct residues. The data was split into positive and negative groups by computing the ratio of high- and low- frequency of each clone and selecting a percentile cutoff. Initial percentiles were chosen as the top or bottom 20% of sequences, setting any sequences that contained zero frequency in the low property pool to the maximum ratio observed and any sequences that contained zero frequency in the high property pool to the minimum ratio observed. Positive ('1') and negative ('0') labels were assigned accordingly. The one-hot encoded protein sequences and their labels were then split into an 80:20 training:test split. The test set was held aside until all analyses were complete and used to validate the model training process. In later analyses, to explore the hyperparameter space of these cutoff parameters, we tested all combinations of the read count, replicate count, and ratio percentile and measured the change in modeling performance. Sensitivity to training:test splitting and the ratio of positive negative labels was tested by performing five-fold cross validation using SciKit Learn's ShuffleSplit function.

We selected linear discriminant analysis for several reasons. First, this method has previously been shown to predict continuous properties from binary sorting data.[151] Next, hyperparameter optimization for this model was straightforward, as the Sci-Kit Learn implementation of LDA has very few parameters, including the solver ('svd' was the only one to converge consistently), n_components (which is fixed to 1 for projection to a single dimension to correlate with protein properties), and tol (which did not change the outcome). Another benefit of using LDA is its simplicity; the linear nature of the model allows for the direct interpretation of

how certain residues contribute to function. While we also evaluated several other models that can create an internal continuous representation for classification (such as support vector classifiers, with the option of using different kernels), we found that LDA models trained much faster. The transform method was used to project data into the 1-dimensional LDA projection after training. Because LDA is a classification model and does not have a regression analog, we used ridge regression, a modified version of linear regression that penalizes large weights, to compare LDA projections to models trained on continuous data. Furthermore, ridge regression did not result in extreme overfitting that was observed by regular linear regression. Finally, ridge regression has been shown to be a powerful modeling technique for protein engineering tasks.[152]

Neural network models were used to evaluate whether non-linear models would capture additional useful information that linear models are unable of modeling, as proposed previously.[151] Standard fully connected, feed forward networks were used with dropout $p = 0.5$ as shown to be effective in the literature.[153] The hidden size (32-256) and number of layers (1-3) did not dramatically affect the results and we ultimately chose the midpoint for both, 128 and 2 respectively (data not shown). We used 700 epochs and a batch size of 32 was for all datasets. Binary Cross Categorical Entropy Loss was used as the loss function, and Stochastic Gradient Descent optimizer with a learning rate of 0.01 was used for all datasets. Training was done on a Nvidia Tesla V100 and typically took between 5 minutes and 2 hours depending on the size and complexity of the dataset.

4.1.4 Stapled peptide cell sorting, sequencing, and flow cytometry

Experimental stapled peptide libraries targeting B cell lymphoma 2 (Bcl-2) proteins were used to evaluate the computational methods on novel datasets. These libraries were sorted and sequenced as described previously in **Chapter 3**. In brief, a computational library of BIM mutants

(a non-specific anti-apoptotic peptide) was designed and transformed into bacteria that displays stapled peptides (see **Table 4.2** for mutagenesis codons, **Table 4.3** for sampled amino acids, and **Table 4.4** for library primers).[23,24] This library was sorted using a combination of MACS and FACS as follows: one round of expression MACS, two rounds of affinity MACS, two rounds of affinity FACS, and two rounds of specificity FACS. Two of such libraries were sorted in parallel: one towards Bcl-x_L and another towards Mcl-1. These libraries were deep sequenced using Illumina NovaSeq S4, demultiplexed, merged using NGMerge, and analyzed using in-house python scripts (see **Table 4.5** for NGS primers).[125] Each peptide sequence was identified by aligning the DNA with the scaffold eCPX protein and then translating the peptides in the corresponding open reading frame. Peptide sequences and their frequencies were aligned across all rounds of sorting, and sequences that had mutations not specified by the original library design were removed (~10% of all sequences). Sequences from the four rounds of FACS were denoted as 'hits' and sequences from the expression sort were denoted as 'not hits' (see **Table 4.1** for dataset summary). Then, the ratio of each round of FACS to the expression was computed and fed into the machine learning pipeline. LDA models were trained identically to the other datasets.

A smaller number of peptide sequences were expressed on the surface of bacteria and measured in low-throughput flow cytometry experiments. To evaluate whether LDA projections were predictive of continuous properties, we expressed 57 stapled peptides on the surface of bacteria from various rounds of sorting (Mcl-1 FACS 2, 3, or 4, and Bcl-x_L FACS 2 or 4) to capture a wider distribution of specificities: peptides from later in the rounds of sorting should have more specificity while those from earlier rounds should be less specific if sorting enriched towards higher performing sequences. We then measured their binding at the approximate K_D of the wild type sequence in triplicate (1nM and 10nM for Mcl-1 and Bcl-x_L respectively). Fraction bound

was calculated by normalizing to expression and dividing by a saturated binder (BIM-p5 at 250nM).[24] LDA projections were calculated and compared to continuous values identically to the other datasets.

4.1.5 SORTCERY

To get continuous estimates of binding properties from cell sorting, peptides from the final round of FACS for both Mcl-1 and Bcl-x_L were evaluated using SORTCERY. Peptides were incubated with either Mcl-1 or Bcl-x_L at 1nM and sorted into twelve bins following the protocol from Reich et al.[61] Briefly, cells labeled with target Bcl-2 protein and anti-HA display tag were sorted into twelve bins along the 'axis of affinity', the diagonal gates that resolves the fraction bound. To compare the SORTCERY value with those measured from binary sorting, we computed the gate score of each sequence as described in the original work.[61] Each of these gates were collected individually and processed for deep sequencing as described previously. The deep sequencing data from these experiments was processed identically to the stapled peptide libraries as above.

4.1.6 Sequence Optimization via Integer Linear Programming

To optimize protein sequences, we applied integer linear programming (ILP), an approach that solves an objective problem given discrete input variables and constraints. Compared to other techniques that maximize an objective given an input, ILP scales more efficiently with a large number of samples and does not rely on iterative predict and test loops that require additional experimental resources.[154–156] Furthermore, ILP is directly amenable to multi-objective optimization through the addition of inequality requirements.[37] We set up this problem using the PuLP python module.[157] First, we defined the objective as maximizing the dot product of the model

coefficient vector and the positions and amino acid constraints as defined by the library design. This objective is the maximization of the confidence of binding for a given sequence. Next, we constrained the optimization by only allowing one amino acid at each position, requiring that each peptide had two azidohomoalanine residues (responsible for peptide stapling), and that the two stapled residues were at a distance as specified by the library design (i,i+7). Finally, we formulated the problem as a multi-objective problem by adding the additional constraint that the dot product of the off-target coefficients and peptide sequence was in the non-binding regime.

Results

4.1.7 Overview of Method

Despite significant efforts to gather quantitative data from high throughput sorting, most directed evolution campaigns rely on basic metrics of protein fitness. We utilized a simple workflow to extract continuous protein properties from NGS datasets while keeping the experimental design simple and affordable (**Figure 4.1**). To accomplish this task, we generated binary labels from enrichment ratios, trained machine learning models using these binary labels to infer continuous protein properties,[151] and optimized protein sequence and function beyond experimentally sampled space into unseen sequence space.[37] We hypothesized that continuous protein properties can be obtained from simple sorting and sequencing analyses for three primary reasons. First, because cell sorting is a stochastic process, cells sorted into discrete bins are sampled from an underlying continuous distribution. Thus, cells sorted in a binary manner may allow inference of this distribution.[158] Second, biased sampling towards the most and least functional variants may allow models to 'interpolate' function of intermediate fitness. Finally, sampling many epistatically interacting motifs may allow inference between them.[159] We also hypothesized this

approach would work across multiple protein engineering objectives, including affinity maturation, fluorescence, deep mutational scanning, and specificity.

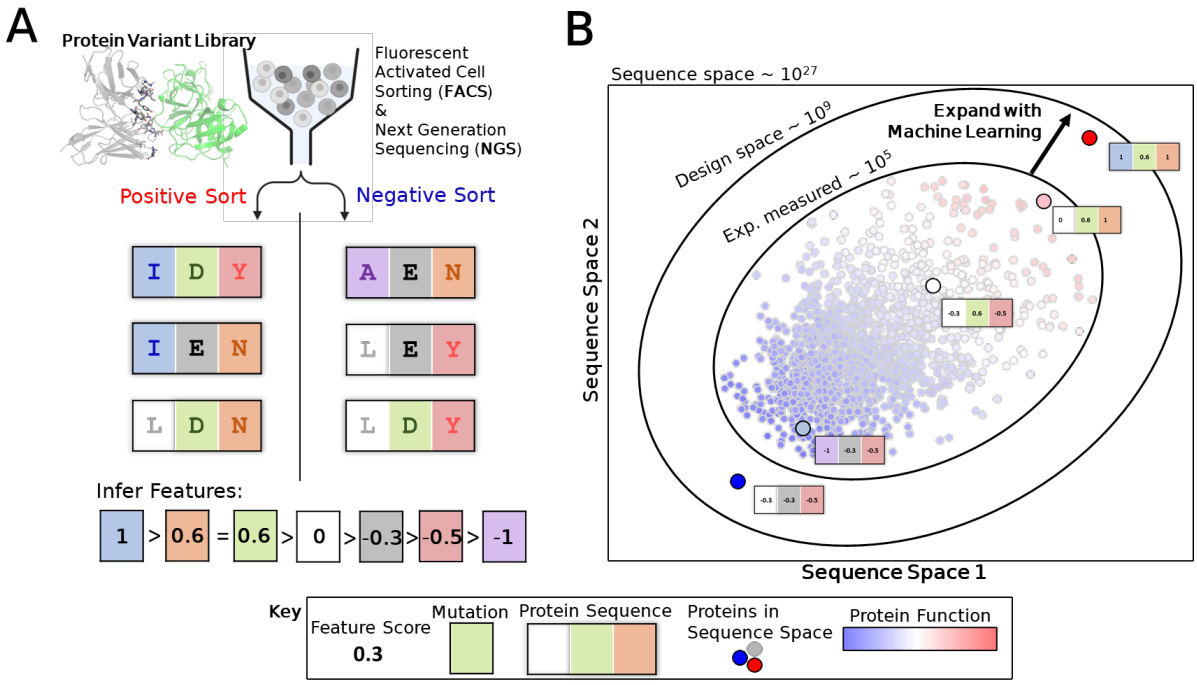


Figure 4.1: Extraction of quantitative protein fitness data from simple binary sorting and sequencing experiments and extrapolation into unseen sequence space towards higher fitness variants. (A) A library of protein variants is expressed on the cell surface and sorted on the basis of fitness (fluorescence, binding affinity, etc.). The library is sorted into two pools: one that denotes function (positive sort) and one that denotes lack thereof (negative sort). While these pools don't map to quantitative fitness, they are indicative of qualitative fitness. Through analysis of the sequence patterns in each pool, quantitative features can be inferred and used to extrapolate beyond experimentally measured sequences **(B)**. Because cell surface display can assay far fewer sequences than are possible, extrapolating beyond experimentally seen sequences may identify higher fitness variants.

To validate the approach, we aggregated data from multiple protein engineering campaigns that fulfilled two criteria: 1. they had many data points of multi-mutant proteins from a sorting campaign and 2. they had measured many continuous protein properties among these variants. These datasets were the fitness landscape of GFP,[22] the directed evolution of a fluorescein-binding scFv,[147] and the fitness landscape of SARS-COV-2 Spike protein .[149,150] Because the co-optimization of multiple properties is often needed, we also gathered datasets that design high-affinity and high-specificity monoclonal antibodies[151] and highly specific peptides between three B cell lymphoma 2 (Bcl-2) proteins.[37]

4.1.8 Data processing pipeline for varying protein variant libraries and sorting schemes

The modular data processing and machine learning pipeline to analyze multiple protein variant libraries consists of multiple steps **Figure 4.2**. First, a library of protein variants is sorted, and the ratio of the positive to negative gate frequencies is calculated for all sequences based on the deep sequencing data. If a sequence found in the positive gate but was unobserved in the negative gate, the ratio was set to the maximum observed; conversely, if a sequence found in the negative gate was unobserved in the positive gate, the ratio was set to 0. We hypothesized this ratio scheme balances the information gained from enrichment ratios while still including clones that were overwhelmingly enriched or depleted. Labels ('1' for high performing variants and '0' for low performing variants) were assigned by determining a cutoff based on the average ratio (percentile $\geq$0.8 and $\leq$0.2 respectively) across how many replicates they appeared in ($\geq$2). We hypothesized that while splitting the positive and negative labels at the 50th percentile would increase the data size, sorting noise around the midpoint would confound information gained from binary ratios (**Figure 4.11**, **Figure 4.12**, **Figure 4.13**, **Figure 4.14**, and **Figure 4.15**). We also hypothesized that removing sequences with 1 replicate would further reduce noise from sorting. Initial estimates of these parameters were chosen to balance the size of the dataset, the strictness of inclusion, and the confidence of the sequencing data. Having easily modifiable parameters for label assignment serves as both a tool for sequencing quality processing and a powerful hyperparameter in the subsequent machine learning steps (see **Figure 4.11**, **Figure 4.12**, **Figure 4.13**, **Figure 4.14**, and **Figure 4.15** for hyperparameter effect on dataset size for each of five datasets).

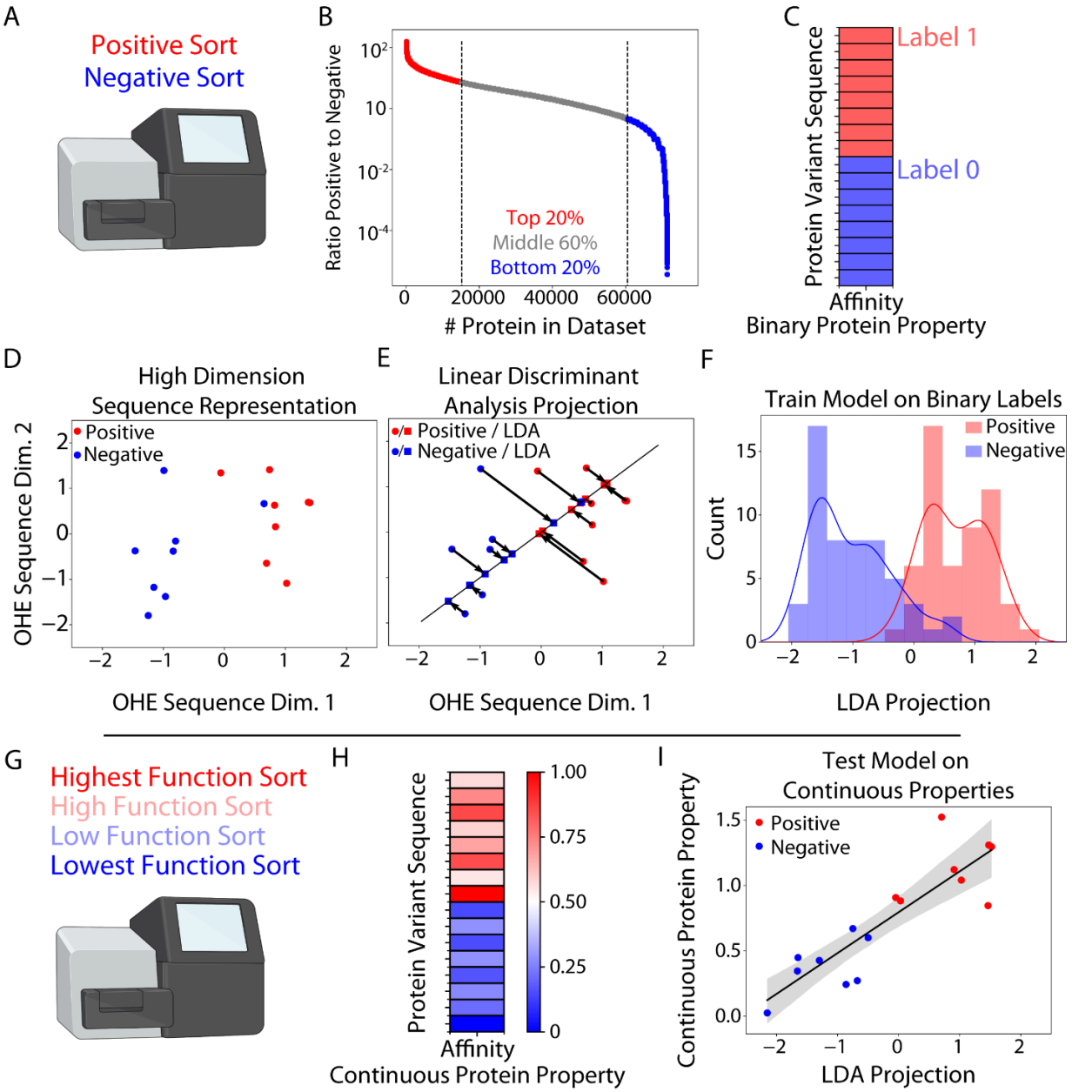


Figure 4.2: Deep sequencing, data pre-processing, and machine learning overview. The most and least functional protein variants from the binary sort are sequenced **(A)** and the ratio of sequence reads in the positive versus negative gate is calculated **(B)**. Binary labels are assigned to each sequence according to its ratio **(C)**; the label thresholds are easily modified depending on the library construction, sorting strategy, and sequencing data quality. Protein sequences are one hot encoded for machine interpretability **(D)** before being used to train a Linear Discriminant Analysis (LDA) model **(E)**, which is evaluated on a hold-out test set **(F)**. Then to calibrate the LDA model, continuous protein properties are obtained either from a quantitative sort (SORTCERY, Sort-Seq, or TiteSeq) or from low throughput measurements (flow cytometry titrations, ELISA, etc.) **(G,H)**. Finally, the projections from the LDA model are used to predict continuous protein properties **(I)**.

Armed with a dataset of sequences and binary function labels, a linear discriminant analysis (LDA) machine learning model was trained because it fulfilled two criteria: it could perform

classification of sequence with its function label, and it had an internal continuous measurement that could be used to correlate with continuous properties. Because LDA models project high dimensional sequence data to maximize class separation, the final projection is a continuous representation that has been previously shown to correlate with continuous properties.[151] The model was trained and tested by splitting the sequencing data into train and test sets randomly (80:20 train:test). To evaluate whether the weights learned by the LDA model correlated to meaningful continuous properties, a subset of the sequences were assayed for their property from a lower throughput but more accurate technique. For all but the Makowski dataset, this was a quantitative cell sorting experiment, and otherwise a low throughput measurement of affinity or specificity via flow cytometry with individual sequences. We then predicted the continuous properties of proteins by comparing the projections from LDA with actual continuous measurements.

4.1.9 Binary labels predict protein properties with equal correlation power

To evaluate whether the LDA models trained on binary sorting data inferred meaningful features of the protein properties, we curated five datasets as described in the methods (see **Table 4.1** for dataset summaries). Using data from each of the these, we compared the measured continuous properties of protein variants to their predicted values from LDA models trained on binary sorting data, as shown in **Figure 4.3** (A, left) for the **Sarkisyan et al.** dataset. We next sought to determine the performance of a comparable model trained on continuous data. Continuous data is more expensive and/or complicated to obtain but presumably is more information rich. Therefore, we hypothesized models trained on continuous data would have stronger correlative power. To evaluate this hypothesis, we trained Ridge regression models, which have been previously shown to be powerful linear models that are not prone to over-fitting.[152] We

then compared the ability for both LDA and Ridge models to predict continuous properties (**Figure 4.3** A, right). Surprisingly, for the **Sarkisyan 2016 dataset**, the LDA models performed similarly to the Ridge regression models as evidenced by a similar Spearman's ρ (0.846 for the LDA model and 0.855 for the Ridge regression model).

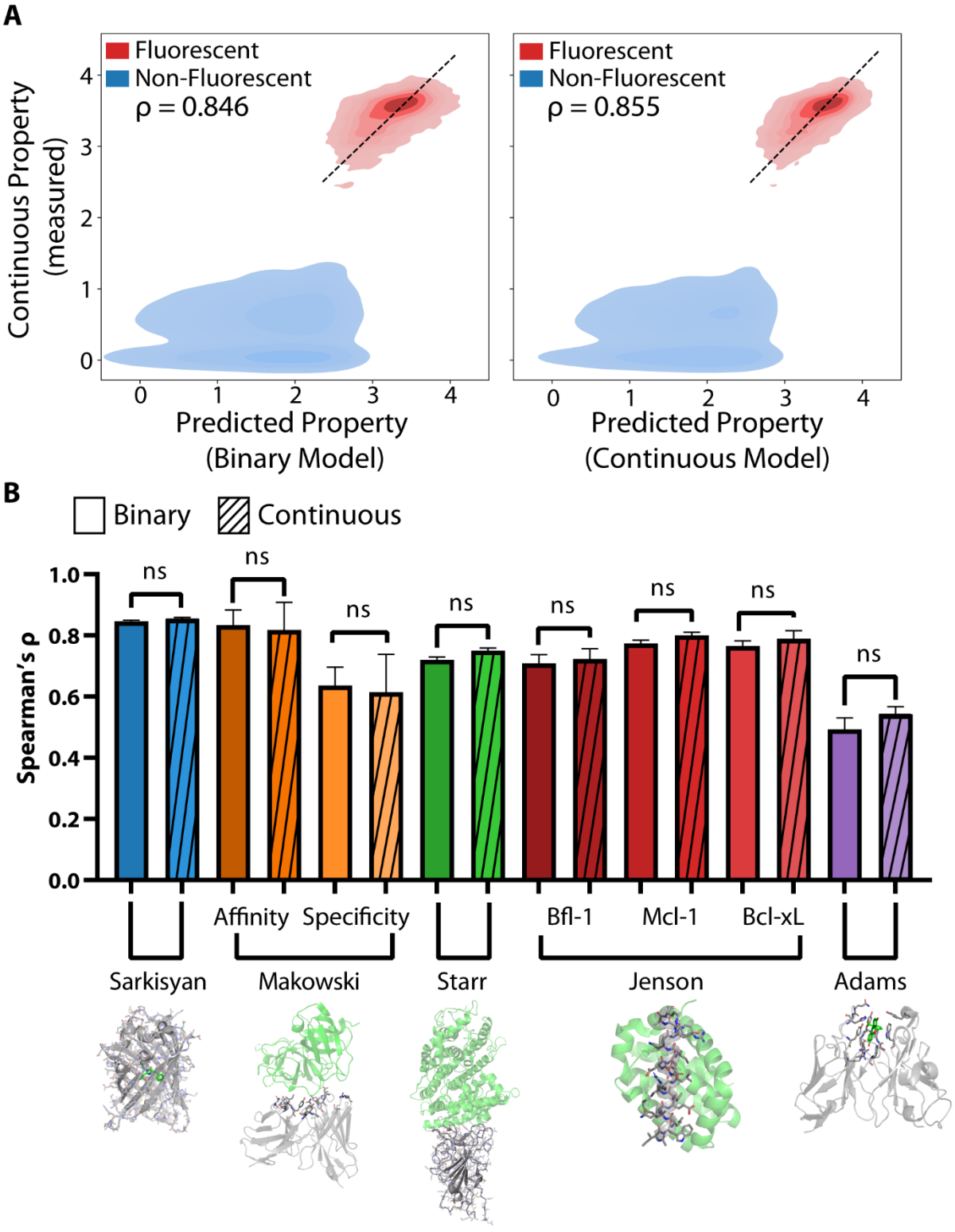


Figure 4.3: Predictions from models trained on binary data are highly correlated with continuous protein properties and equally powerful as models trained on continuous data. Evaluated on the Sarkisyan data, LDA models trained on binary data **(A, left)** or Ridge models trained on continuous data **(A, right)** are correlated with

fluorescence. Across five protein engineering datasets, models trained on binary data are equally predictive of continuous properties **(B).**

We then tested whether the other four datasets had similar performance. First, we observed that LDA models achieved high classification performance on the held-out test set for all datasets (see **Table 4.6** for accuracy, precision, recall and F_1 score) and were not overfit as evidenced by similar performance on the training and test sets. Next, we observed that LDA projections were highly correlated with continuous measurements, as evidenced by Spearman's ρ between 0.5 and 0.85 (**Figure 4.3b**, additionally see **Figure 4.11**, **Figure 4.12**, **Figure 4.13**, **Figure 4.14**, and **Figure 4.15**for hyperparameter effect on performance on each of five datasets). To get an estimate of model sensitivity to dataset splitting, we performed 5-fold cross validation (see Methods) on each training dataset (**Figure 4.16**). Strikingly, for each of the datasets, we observed no significant difference in the correlation (significance was measured as a t-test on the unbounded Z transform of the Spearman ρ).[160] Encouraged by the success of correlation, we also sought to explain the magnitude of correlation, which was consistently high but had two outliers. **Adams 2016** dataset had a significantly lower predictive value of ρ~0.5. We suspect this decrease in performance has two sources: noise in the dataset due to an abundance of unresolvable low affinity variants (see **Figure 4.17** for correlation plots for each dataset), and the lack of discrimination between binding affinity and expression level in the experimental sorting design, which can attribute higher affinity to sequences with higher display and vice versa.[143] The **Makowski 2022** specificity dataset also had lower than average performance; we hypothesize this model suffered due to the difficult nature of measuring antibody off-target binding.[151,161,162]

To test whether linear models were limiting the predictive capabilities of continuous properties, we also tested fully connected, feed forward neural networks, which have been shown to similarly identify continuous values from binary data.[151] While non-linear models may capture

higher order epistatic behavior, these models generally performed as strongly as LDA models (**Figure 4.18**). Over this wide range of protein engineering objectives, this approach consistently predicts continuous properties and has comparable accuracy to models trained on state-of-the-art sequencing and sorting data.

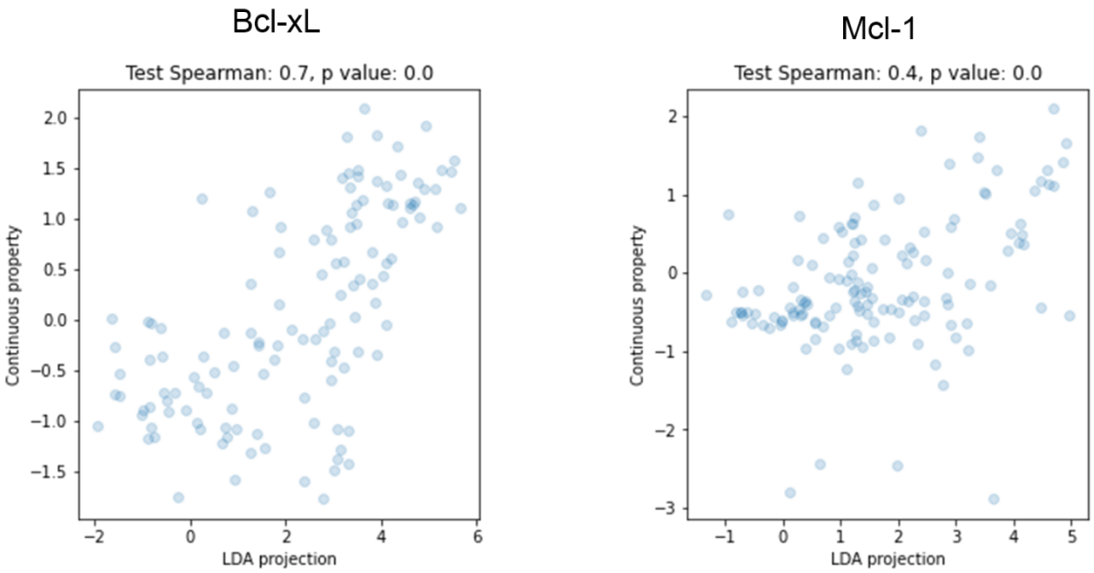


Figure 4.4: LDA Projections from binary sorting versus multi-gate predicted continuous affinity (see Jenson et al. 2018 *PNAS* for more details).

4.1.10 Prediction of stapled peptide affinity and specificity from binary labels

To apply this method prospectively to a new dataset following the promising retrospective analysis, we chose B cell lymphoma 2 (Bcl-2) stapled peptide antagonists as our design case. In addition to requiring non-natural amino acids, making it incompatible with modeling approaches based on naturally evolved proteins, these peptides are well suited for this approach because we can evaluate not just a single property but the tradeoff between affinity and specificity. We generated a dataset of B cell lymphoma 2 (Bcl-2) stapled peptide variants that were sorted over several rounds (**Figure 4.5a**) using the bacterial cell surface display.[23,24] This library was designed based on naturally occurring peptide sequences, SPOT arrays of BIM mutants, and previously

designed high-affinity or specificity BH3 variants (Case 2023, manuscript in progress) (see **Table 4.2**, **Table 4.3**, and **Table 4.4**)[36,37,129] Because bacterial surface display libraries are highly limited by size compared to the theoretical diversity of BH3 peptides (~10^{30}), mutations were prioritized that were predicted to govern specificity between Mcl-1 and Bcl-x_L. The final library of ~10^9 was transformed into bacteria (**Figure 4.5b**) and sorted against either Mcl-1 or Bcl-x_L with a combination of three magnetic and four fluorescent activated cell sorting (MACS/FACS) (**Figure 4.5c**). The magnetic cell sorting was performed until the library was sufficiently reduced in diversity for analysis with FACS, which offers more precise control over property selection. We deep sequenced these pools to isolate highly active peptides (**Figure 4.5d**), which enabled an understanding of sequence trends that governed high affinity and specificity (see **Figure 4.19** for sequence trends) and provided a source of data to train and evaluate the capabilities of LDA models

to predict peptide function (**Figure 4.5 e and f**). We observed high correlation between for both Mcl-1 and Bcl-x_L LDA models (Spearman's ρ of 0.893 for Mcl-1 and 0.708 for Bcl-x_L).

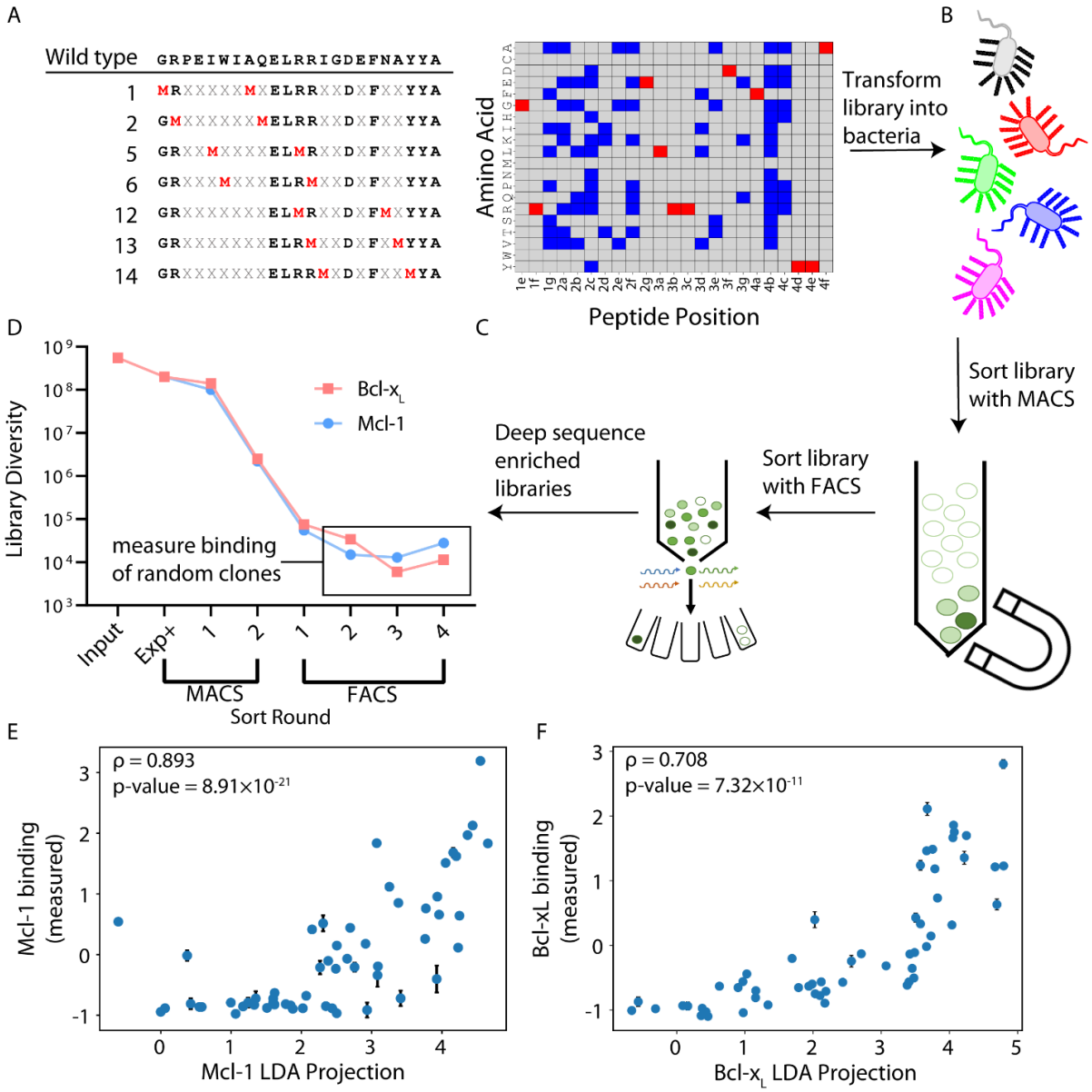


Figure 4.5: **Prospective analysis of B cell lymphoma 2 (Bcl-2) pro-apoptotic stapled peptides via bacterial surface display, deep sequencing, and machine learning.** A combinatorial mutagenesis library of stapled BIM variants was designed including staple locations (left) and sequence (red positions fixed, blue positions variable, right) **(A)**, transformed into bacteria **(B)**, sorted using a combination of magnetic activated cell sorting (MACS) **(C)** and fluorescent activated cell sorting (FACS) towards Bcl-x_L and Mcl-1 (two members of the Bcl-2 family) in parallel. The library was next generation sequenced (NGS) to calculate frequencies of each unique sequence along the sorting progression **(D)**. Finally, a LDA model was trained on the binary labels from NGS and used to predict the continuous binding of 57 peptide variants, which were selected randomly from FACS 2-4 for both Mcl-1 **(E)** and Bcl-x_L **(F)**.

To generate training data, we aggregated all four rounds of FACS and the expression positive MACS sorts, hypothesizing that would provide additional confidence for 'hits' and expressing but non-binding sequences. The ratio of these counts was computed as described above and used to generate labels for LDA training and testing (see **Figure 4.19** for logoplots of negative and positive sequences). First, we observed that LDA models had high classification performance and were not overfit (see **Table 4.7** for performance statistics and **Figure 4.20** for hyperparameter effect). We then tested the performance of LDA to predict continuous properties by randomly sampling 57 sequences among the FACS sorts, measuring their continuous binding via flow cytometry, and measuring the correlation between predicted LDA binding and the sequences' actual binding **(Figure 4E-F)** (see **Figure 4.21** for sequences and data). We observed a strong correlative power between LDA projections and continuous measurements of peptide affinity: spearman ρ of ~0.7 and ~0.8 for Mcl-1 and Bcl-x_L respectively ($p < 0.00001$). Finally, we sorted the final round of sorted cells via SORTCERY for a comparison with high-throughput, semi-quantitative measurements of binding affinity. Surprisingly, the binary sorting data coupled with an LDA model trained with NGS data had better performance than selecting clones from the final 2 rounds of sorting for Mcl-1 specificity (**Figure 4.4**), suggesting that the information contained from simple sorting experiments provides a powerful method to predict continuous protein properties.

4.1.11 Optimization of stapled peptides using machine learning and integer linear programming

While directed evolution campaigns may yield the desired properties after sorting, sequencing, and modeling, it is also possible that further optimization is necessary. In such cases, protein engineers rely on a combination of manual and automated approaches to further optimize lead

candidates.[37,154–156] We sought to explore how our modeling workflow could not only score entire sequences, but how the contributions of individual amino acids contributed, potentially enabling the generation of new, unsampled sequences. Because linear models have associated weights for each amino acid and sequence position, the same scoring tools to find the best measured clones can also be used to score sequences that have never been evaluated experimentally. We therefore applied an optimization approach that can optimize discrete inputs for continuous properties and explore unseen sequence space: integer linear programming (ILP) (**Figure 4.6**), which has previously been applied to design specific linear peptides towards the Bcl-2 proteins.[37] To establish the baseline of specificity from sorting, we further characterized variants from the final round of sorting that were predicted to be specific for Bcl-x_L and Mcl-1. Interestingly, most peptides from the Bcl-x_L library were highly specific (**Figure 4.21**), while fewer from Mcl-1 performed favorably (~80% had significant off-target binding, **Figure 4.6a**). We hypothesized we could recover specific Mcl-1 clones by optimizing sequences from sorting and sequencing data that otherwise yielded mixed results. We solved the ILP model three times, once for Bcl-x_L specific peptides, once for Mcl-1 specific peptides, and once more for bispecific peptides (see Methods for more details). Out of many sequences predicted to have high activity for Mcl-1 (**Figure 4.22**), we randomly selected several sequences for low-throughput flow cytometry analysis (**Table 4.8**). Strikingly, we observed that the optimized Mcl-1 sequences displayed similar or improved specificity compared to the highest activity clones assayed experimentally.

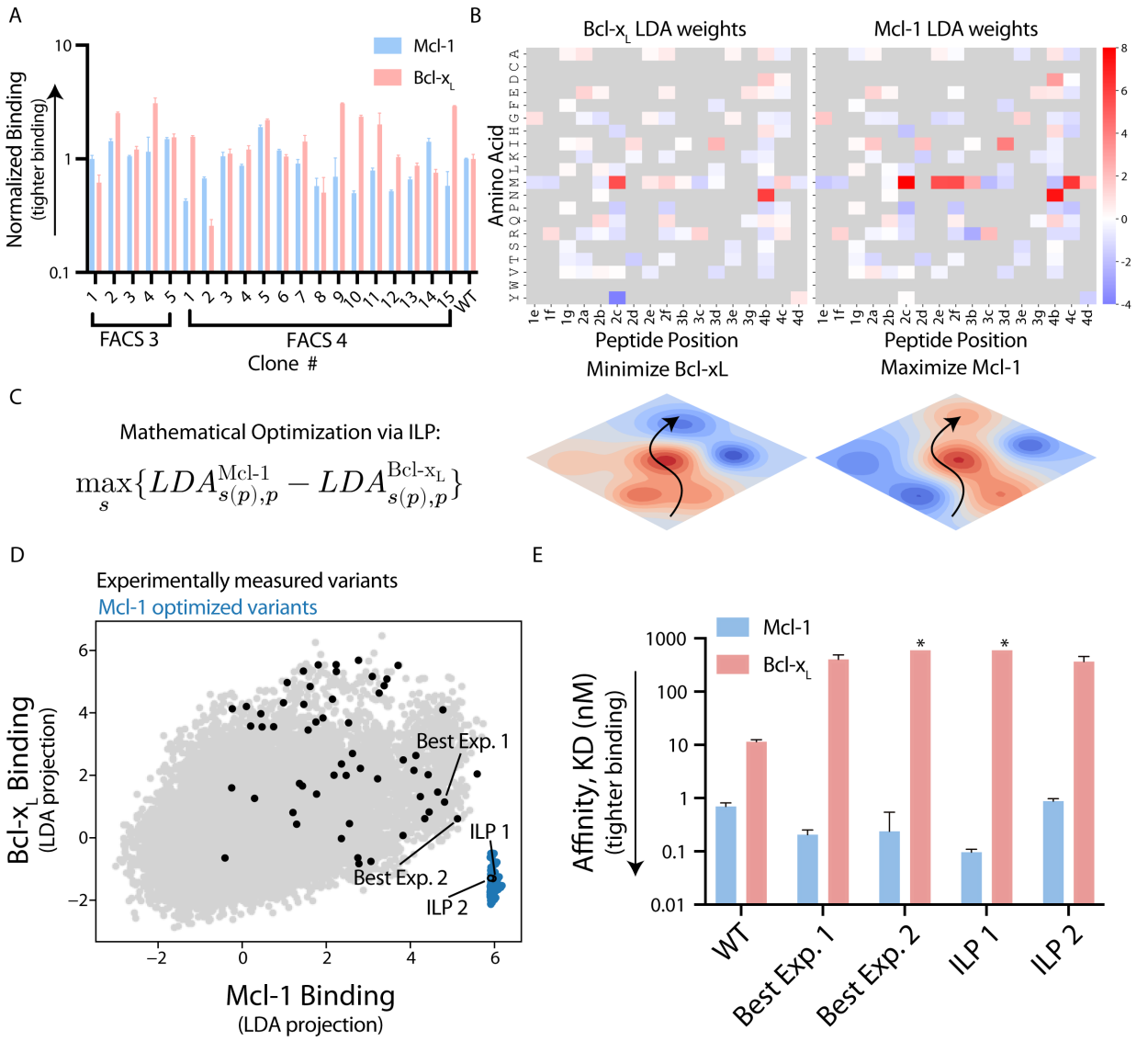


Figure 4.6: Extrapolation of interpretable ML model weights to generate novel, highly specific Mcl-1 inhibitors. Of 20 sequences randomly selected from the final two rounds of sorting towards Mcl-1, many did not display high levels of specificity towards Mcl-1 when measured in low throughput binding assays **(A)**. We hypothesized the weights from linear discriminant analysis (LDA) machine learning could be used to design peptides with high affinity to Mcl-1 **(B)** or Bcl-x_L **(C)**. To optimize the sequences, we applied integer linear programming (ILP) **(C)** to maximize the likelihood a peptide binds Mcl-1 while minimizing its binding to Bcl-x_L **(D)**. ILP identified numerous sequences that were predicted to be highly specific **(E)** that were not among the 10^5 sequences assayed experimentally. Two variants were randomly chosen among this set and were found to be as specific as the best clones identified from sorting **(F)**.

Sequences initially identified by minimizing Mcl-1 binding while maximizing Bcl-x_L binding resulted in peptides that did not bind either Mcl-1 or Bcl-x_L (**Figure 4.7** and **Table 4.8**);

it has been previously shown that subtle differences in ILP set up can affect the efficiency of outcome.[37]

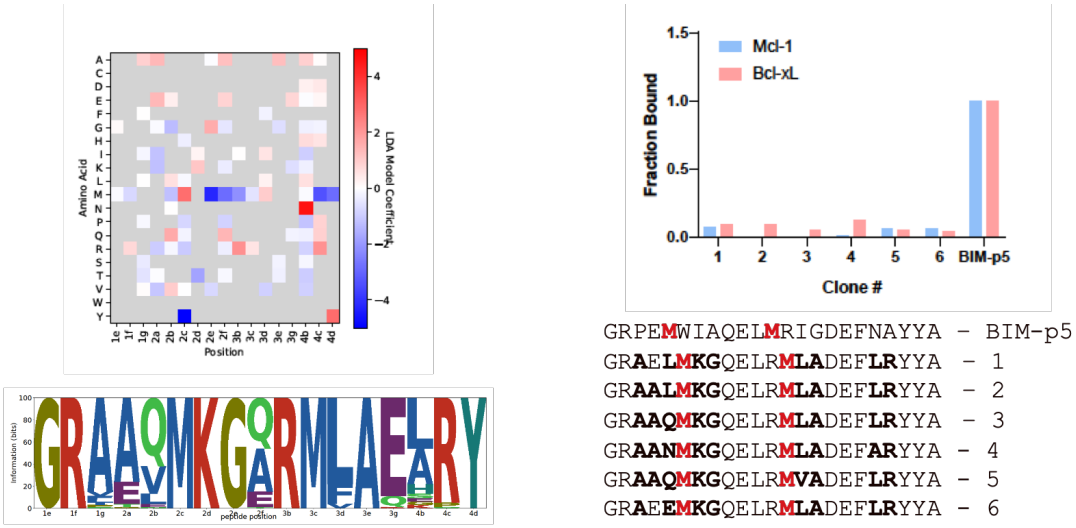


Figure 4.7: Initial designs for Bcl-xL using ILP yielded non-binding sequences for both Bcl-xL and Mcl-1. The optimization problem was set up as a maximization of Bcl-xL affinity subject to a low cutoff of Mcl-1 binding.

We suspect this failure was due to the model being overly sensitive to mutation at Asp at position 4b, which was the only mutation consistently sampled that had a high score for both Bcl-x_L and Mcl-1 but was slightly higher for Mcl-1. To address this issue, we maximized Bcl-x_L

binding then chose the sequences which had the lowest Mcl-1 scores, which preserved Bcl-x_L binding and resulted in highly specific peptides (**Table 4.9** and **Figure 4.8**).

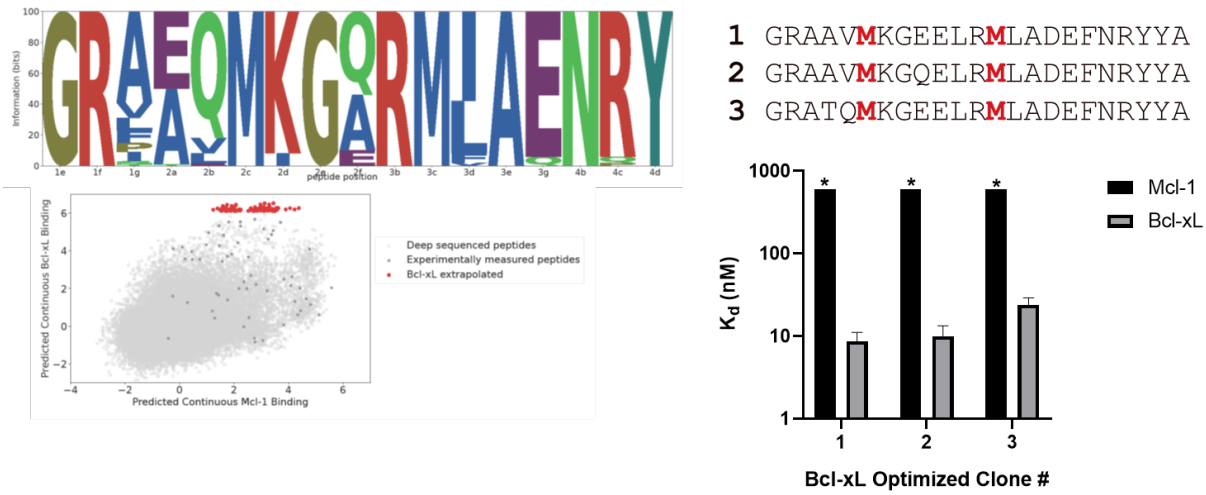


Figure 4.8: Second iteration for Bcl-xL specific stapled peptides using ILP. *: no binding detected up to 250nM.

While our sorting campaign was originally designed to identify highly specific peptides, we also pursued bispecific peptides, which serve as proof that the model can interpolate in sequence-function space but could also serve as therapeutics in diseases driven by both Bcl-2 proteins. Sequences were identified by maximizing both Mcl-1 and Bcl-x_L binding, yielding

peptides with relatively high affinity for both targets that had significant sequence difference from wild type (BIM) (**Figure 4.9**).

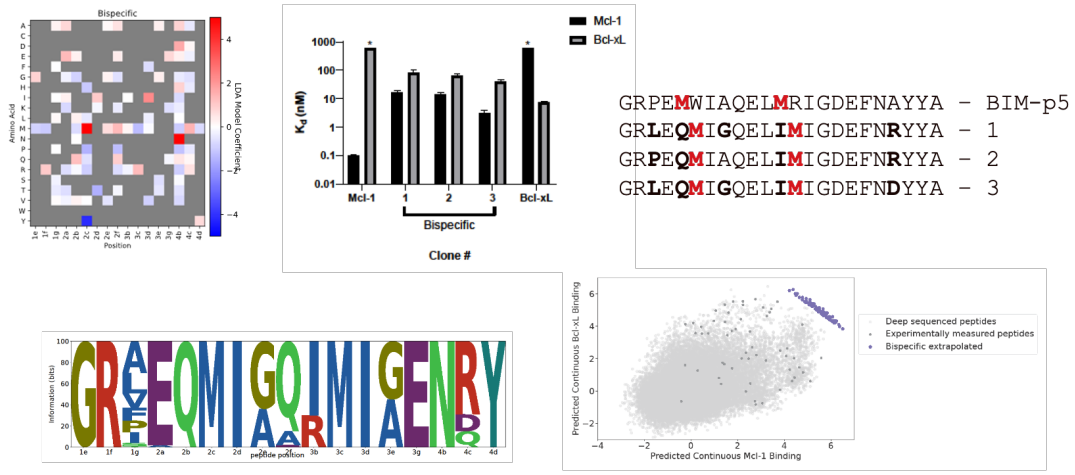


Figure 4.9: Bispecific peptides designed via ILP. The ILP objective was set as the maximization of both Mcl-1 and Bcl-xL score.

To show generalizability of ILP to generate functional protein variants, we additionally set up the optimization problem using the **Makowski dataset** (**Figure 4.10**). We defined the objective of this optimization as the minimization of off-target binding, subject to the maintenance of affinity. We solved the model and compared the highest functional sequences according to our predictions to those described in the original manuscript. We found that the predicted sequences were extremely close to those identified as co-optimal by Makowski and co-authors.

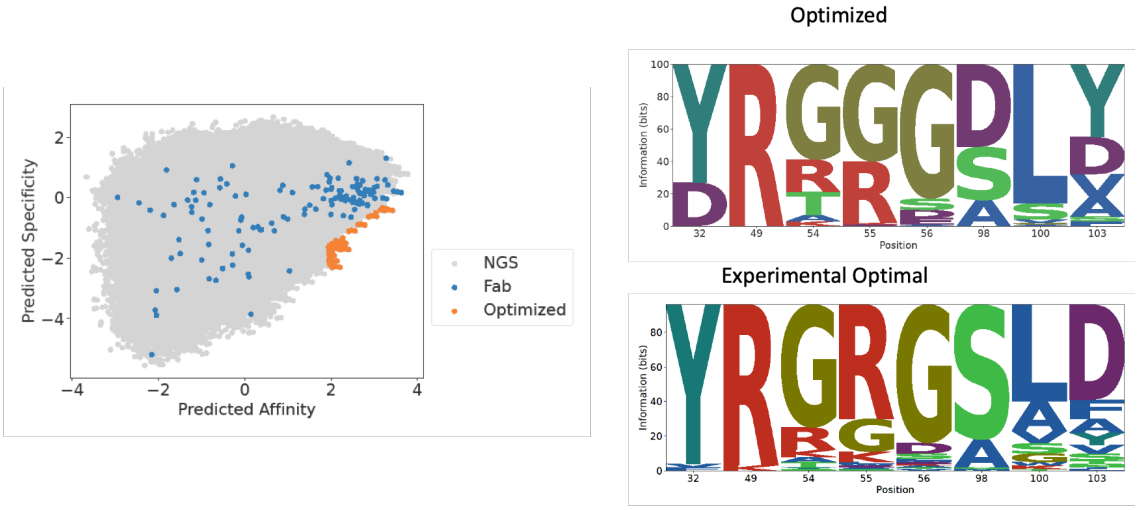


Figure 4.10: High affinity and specificity antibodies from Makowski et al. (2022) *Nature Communications* via ILP. The ILP objective was set as the maximization of affinity and minimization of specificity, subject to a minimum affinity threshold.

Discussion

In this work, we developed a method to utilize NGS data from simple binary sorting results with machine learning to infer continuous protein properties. These results can also be utilized to extend the sequence space beyond sequences directly observed in the library (**Figure 4.1**). The workflow consists of two important parts: the label assignment process from deep sequencing data, and the use of linear machine learning models to predict continuous protein properties from binary data (**Figure 4.2**). Currently, there is a lack of consensus on how to best analyze directed evolution data for lead molecule selection and protein optimization. This lack of consensus likely arises from variations in how experiments are set up, which depends on surface display platform, sequencing instrumentation, FACS instrumentation, the design of sort gates, sequencing depth, among other factors. This technique provides a practical but powerful method compared to typical enrichment ratio analysis through a simple binary classification from any sorting experiment. By defining a ratio of frequencies based on any two gates (positive/negative sort, input/output sort, etc.) and

binarizing the ratios into '1' and '0', any directed evolution experiment can be transformed into a dataset for downstream analysis. The transformation to binary labels is important because the next component of the workflow is the use of linear machine learning models that can be used to predict continuous properties from directed evolution data (linear discriminant analysis, LDA).[151] The noise in enrichment ratios is likely mitigated by binarization, and the information contained from labels and sorted protein sequences facilitates the continuous transformation yielded by machine learning models.

To test our method, we curated data from five large protein engineering campaigns: the fluorescent landscape of avGFP,[22] the directed evolution of a fluorescein-binding scFv,[147] the RBD affinity landscape towards SARS-COV-2 Spike protein,[149,150] high-affinity and high-specificity Fabs,[151] and the design of highly specific peptides against B cell lymphoma 2 (Bcl-2) proteins (**Figure 4.3**).[37] Proteins in these data vary in complexity from short alpha helical peptides to large globular proteins and in objective from protein fluorescence to multi-objective affinity and specificity optimization. Furthermore, each of these datasets varied in both sorting strategy and complexity: **Makowski *et al.*** sorted for the top ~5% of antibody variants while **Adams *et al.*** quantified the binding of an entire family of fluorescein binders. While many of the projects relied on complex sorting techniques to obtain quantitative protein labels, we simulated simple binary sorting experiments by limiting the sequencing data (see Methods). We then evaluated the predictive power of LDA models trained on these simple sorting experiments and observed both impressive classification performance and strong prediction of continuous properties from LDA binary projections. Interestingly, models trained on binary data were highly correlated with continuous data (Spearman correlation coefficients ranged from 0.5-0.9). Furthermore, when we compared the predictive power of LDA models trained on binary data to regression models trained

on continuous data, we observed no increase in rank order performance, suggesting that models trained on simple sorting experiments yield comparable information to models trained on data from experiments that generate hundreds to thousands of continuous measurements.[37,61,147]

Next, we sought to explore how this workflow could be used for prospective analysis in addition to retrospective analysis (**Figure 4.5**). We hypothesized that because the workflow is agnostic to protein type and display platform, any directed evolution campaign with sufficient sorting and sequencing data is a suitable environment for testing. As such, we chose to analyze libraries of stapled peptides, an important class of protein formed by a covalent crosslinking of two amino acids.[46] Stapled peptides are being explored as therapeutics for previous 'undruggable' disease related proteins, owing to their location inside the cell and untargetable by small molecule drugs.[5] Stabilized Peptide Engineering by *E. coli* Display (SPEED) has previously been demonstrated to accelerate the development of stapled peptides by displaying them on the surface of bacteria, where libraries of peptides varying in sequence and staple location simultaneously can be optimized for protein-peptide interactions.[23,24] One additional challenge in the optimization of stapled peptides is their reliance on non-natural amino acids, which generally results in the incompatibility of models trained on naturally occurring sequences.[163–166] We built on previous work by generating a library of randomized stapled peptides towards two B cell lymphoma 2 proteins (Bcl-2), an important class of apoptosis regulatory proteins that is responsible for cancer cells immortality.[50] We sorted this library against two important members: Mcl-1 and Bcl-x_L,[51] each of which drives immortality in different diseases.[57] Selective targeting among Bcl-2 proteins is an outstanding goal in drug targeting but is difficult due to the highly homologous nature of these proteins. After several rounds of cell sorting and subsequent deep sequencing, we trained LDA models on a subset of the binary sequencing data, evaluated the model on both the hold-out

test set, and generally observed high classification performance. We then measured the binding of 57 sequences from various rounds of sorting with low throughput flow cytometry experiments and observed that many of the clones did not demonstrate favorable affinity or specificity properties when sampling from these enriched libraries. However, we did observe a high degree of correlative power between LDA projections and continuous peptide binding. Finally, these models were able to identify molecules within the set of experimentally observed sequences that were highly specific but may not have been selected for lead compounds due to their rarity.[138] Several sequences along the Pareto frontier, or the boundary of co-optimality where an increase in one property leads to a decrease in the other, were translated into bacteria and assayed via flow cytometry. We also characterized several clones that were bispecific, which could have applications in specific diseases, but also serves as a test case if the model can interpolate function where it wasn't directly engineered via cell sorting. The specificities of these peptides agreed with model predictions, indicating the model was able to identify functional and rare peptides from across the specificity landscape.

Finally, we sought to use the interpretive nature of the linear machine learning models to explore unseen sequence space and generate highly diverse and novel sequences **(Figure 4.6)**. To accomplish this task, we used integer linear programming and the coefficients from machine learning to mathematically optimize peptide sequences beyond the properties that were experimentally observed (from deep sequencing or flow cytometry).[37] We hypothesized that such an approach could recover functional peptides with consistency where sorting did not; while the final round of Bcl-x_L sorting yielded consistently high affinity and specificity variants (**Figure 4.21**), the Mcl-1 sort had a small fraction of sequence variants with desired properties. We thus prioritized the design of Mcl-1 binders and identified a new peptide sequence that improved

peptide properties beyond the experimentally measured Pareto front. Importantly, this variant demonstrated specificity at least as potent as the most specific clone identified from experimental work.

To test whether our sequence optimization workflow generalizes beyond small alpha helices, we also applied ILP to the **Makowski dataset** (**Figure 4.10**). While antibodies have been the subject of optimization using highly sophisticated models,[167–170] we hypothesized that the high performance from linear ML models would make it amenable to ILP optimization. Like Bcl-2 inhibitors, antibodies need to demonstrate properties beyond high affinity to be considered therapeutic, and ILP is uniquely suited to tackle co-optimization.[171] We observed that the set of sequences predicted to be co-optimal by ILP are similar to their most optimal clones identified experimentally. Furthermore, their lead antibody identified as co-optimal (EM1) was among the set of antibodies predicted by ILP. Makowski and co-authors designed a comparatively small library (~10^6) for their experimentally measured sequences (~10^4), resulting in a more confident sampling of mutated amino acids experimentally. In contrast, the library of stapled peptides we designed had a much larger ratio of design space (~10^9) to experimentally measured sequences (~10^5), making this library suitable for extrapolation beyond experimentally measured space using machine learning. For protein variant libraries where mutations are sufficiently independent (minimal higher order epistatic interactions), a strategic subsampling of design space can be advantageous for subsequent protein optimization with linear models[172,173] and help to de-risk sorting campaigns, as exploration through the full design space can improve function beyond those originally assayed.

The use of ML with NGS data from binary sorting campaigns has many advantages, but the approach also has a few limitations. It is important to note that LDA projections are correlated

with, but not predictive of, continuous measurements. Therefore, LDA-informed properties may not match 1:1 with continuous properties. However, because many protein engineering campaigns do not seek to quantify the exact magnitude of fitness, but rather seek to maximize or minimize a property or trade-off between properties, this correlation can still provide direct insight into protein fitness and accelerate optimization efforts. We also found that ILP optimization was sensitive to model weights as evidenced by the initial failure of generating highly specific Bcl-x_L peptides. Two approaches to address this are incorporating uncertainty into model predictions that could yield more confident extrapolation into unseen sequence space,[154] or selecting a range of sequences from multiple modes of optimization simultaneously.[37] Despite identifying peptides with high specificity towards Mcl-1 and Bcl-x_L, more work is needed to yield effective peptide therapeutics: it is equally important to show these peptides do not bind the other 3 Bcl-2 members.[57] Lacking knowledge of the sequence space of high affinity binders, we were unable to explore this aspect of peptide design; future work includes designing stapled peptides against the entire Bcl-2 family, which play additional roles in off-target toxicity and are responsible for immortality in other cancers. Because this approach is amenable to higher dimension multi-objective optimization, we expect that optimizing specificity for five proteins with this approach is possible.

Despite these limitations, the ability to score sequences beyond those observed experimentally is important because drug-like properties not easily assayable by high-throughput techniques (immunogenicity, stability, cell permeability, etc.) are often highly dependent on sequence and may need further optimization.[18,122,171,174] For example, minimization of positive charge in CDR regions of antibodies has been shown to minimize off-target binding,[162] while selective placement of hydrophobicity and positive charge has been shown to improve cell penetration for stapled peptides.[18,175] This combined machine learning and optimization approach

provides a powerful method to identify highly functional protein variants if experimentally measured clones did not meet fitness criteria or further sequence optimization is necessary.

In summary, the data processing and modeling workflow designed in this work is a versatile tool towards the improved analysis and identification of protein variants across many domains of protein engineering by utilizing machine learning and NGS data to predict continuous properties from binary sorting data.

Appendices

Table 4.1 contains metadata about datasets used in this study; **Table 4.2** is the library codon table and **Table 4.3**, **Table 4.4**, and **Table 4.5** are DNA primers for library design and sequencing; **Table 4.6** and **Table 4.7** are LDA performance statistics; **Table 4.8** and **Table 4.9** is DNA primers for ILP design. **Figure 4.11**, **Figure 4.12**, **Figure 4.13**, **Figure 4.14**, and **Figure 4.15** are dataset hyperparameter effect on performance; **Figure 4.16** is training and test correlation statistics; **Figure 4.17** is correlation plots for all datasets; **Figure 4.18** is neural network modeling statistics; **Figure 4.19** is stapled peptide input and output logoplots; **Figure 4.20** is hyperparameter performance for stapled peptide modeling; **Figure 4.21** is sequences and binding specificities of random stapled peptides; **Figure 4.4**, **Figure 4.22**, **Figure 4.7**, **Figure 4.8**, **Figure 4.9**, and **Figure 4.10** are sequences and information about ILP sequence optimization for Mcl-1, Bcl-xL, Bcl-xL round 2, bispecific, and Makowski et al. 2022 datasets respectively.

Table 4.1: Parameters of each dataset used in this study

First Author	Publication Year	Protein	Property	# of sequences	Total Sequence Length	Varied Sequence Length
Sarkisyan	2016	GFP	Fluorescence	281131	238	238
Starr/Greaney	2022/2021	SARS-COV-2 Spike	ACE2 Binding	75971	201	201
Makowski	2022	Emibetuzumab	HGFR Binding	1157969	115	7
Makowski	2022	Emibetuzumab	HGFR Specificity	1157969	115	7
Jenson	2018	BIM	Bcl-xL Binding	3457	22	18
Jenson	2018	BIM	Bfl-1 Binding	3489	22	18
Jenson	2018	BIM	Mcl-1 Binding	3480	22	18
Adams	2016	Anti-fluorescein scFv	Fluorescein Binding	3682	266	20
Case	N/A	BIM	Mcl-1 Binding	90117	23	17
Case	N/A	BIM	Bcl-xL Binding	75870	23	17

Table 4.2: Degenerate codon design for pro-apoptotic anti-Bcl-2 bacterial surface display library

	AA diversity
wt	1
p1	6.37E+07
p2	7.17E+07
p5	6.37E+07
p6	4.78E+07
p12	6.37E+07
p13	5.97E+07
p14	4.78E+07
Total	4.18E+08

	1e	1f	1g	2a	2b	2c	2d	2e	2f	2g	3a	3b	3c	3d	3e	3f	3g	4a	4b	4c	4d	4e	4f
wt	G	R	P	E	I	W	I	A	Q	E	L	R	R	I	G	D	E	F	N	A	Y	Y	A
	ggt	cgc	ccg	gaa	att	tgg	att	gcg	caa	gaa	ttg	cgc	cgc	att	ggt	gac	gaa	ttt	aac	gcg	tat	tat	gcg
p1	M	R	X	X	X	X	X	M	X	E	L	R	R	X	X	D	X	F	X	X	Y	Y	A
	atg	cgc	dya	rna	sda	can	aha	atg	vva	gaa	ttg	cgc	cgc	ntc	gsa	gac	vaa	ttt	vnc	svm	tat	tat	gcg
p2	G	M	X	X	X	X	X	X	M	E	L	R	R	X	X	D	X	F	X	X	Y	Y	A
	ggt	atg	dya	rna	sda	can	aha	gba	atg	gaa	ttg	cgc	cgc	ntc	rsc	gac	vaa	ttt	vnc	smm	tat	tat	gcg
p5	G	R	X	X	M	X	X	X	X	E	L	M	R	X	X	D	X	F	X	X	Y	Y	A
	ggt	cgc	dya	rna	atg	can	awa	gba	vva	gaa	ttg	atg	cgc	ntc	rsc	gac	vaa	ttt	vnc	smm	tat	tat	gcg
p6	G	R	X	X	X	M	X	X	X	E	L	R	M	X	X	D	X	F	X	X	Y	Y	A
	ggt	cgc	dya	rna	sda	atg	aha	gba	vva	gaa	ttg	cgc	atg	ntc	gsa	gac	vaa	ttt	vnc	svm	tat	tat	gcg
p12	G	R	X	X	X	X	X	X	X	E	L	M	R	X	X	D	X	F	M	X	Y	Y	A
	ggt	cgc	nyc	rna	sda	cna	aha	gba	vva	gaa	ttg	atg	cgc	ntc	rsc	gac	vaa	ttt	atg	smm	tat	tat	gcg
p13	G	R	X	X	X	X	X	X	X	E	L	R	M	X	X	D	X	F	X	M	Y	Y	A
	ggt	cgc	dya	rna	sda	cna	awa	gba	sva	gaa	ttg	cgc	atg	ntc	rsc	gac	vaa	ttt	vnm	atg	tat	tat	gcg
p14	G	R	X	X	X	X	X	X	X	E	L	R	R	M	X	D	X	F	X	X	M	Y	A
	ggt	cgc	dya	rna	sda	cna	aha	gba	vva	gaa	ttg	cgc	cgc	atg	gsa	gac	vaa	ttt	vnc	svm	atg	tat	gcg

Table 4.3: Sampled amino acids for pro-apoptotic anti-Bcl-2 bacterial surface display library

Position	1e	1f		1g		2a	2b	2c	2d	2e	2f	2g		3a	3b		3c	3d		3e		3f	3g		4a	4b		4c		4d	4e		4f
BIM WT	G	R		P		E	I	W	I	A	Q	E		L	R		R	I		G		D	E		F	N		A		Y	Y		A
p1		M		AILSTV		AEGIKRTV	EGLQRV	DEHKNQY*	IKT	M	AEGKPQRT							FILV		AG		EKQ				ADGHILNPRSTV		ADEGHPQR					
p2			M	AILSTV		AEGIKRTV	EGLQRV	DEHKNQY*	IKT	AGV	M							FILV		AGST		EKQ				ADGHILNPRSTV		ADEHPQ					
p5				AILSTV		AEGIKRTV	M	DEHKNQY*	IK	AGV	AEGKPQRT				M			FILV		AGST		EKQ				ADGHILNPRSTV		ADEHPQ					
p6				AILSTV		AEGIKRTV	EGLQRV	M	IKT	AGV	AEGKPQRT						M	FILV		AG		EKQ				ADGHILNPRSTV		ADEGHPQR					
p12				AFILPSTV		AEGIKRTV	EGLQRV	LPQR	IKT	AGV	AEGKPQRT				M			FILV		AGST		EKQ				M		ADEHPQ					
p13				AILSTV		AEGIKRTV	EGLQRV	LPQR	IK	AGV	AEGPQR						M	FILV		AGST		EKQ				ADEGHIKLNPQRSTV		M					
p14				AILSTV		AEGIKRTV	EGLQRV	LPQR	IKT	AGV	AEGKPQRT							M		AG		EKQ				ADGHILNPRSTV		ADEGHPQR		M			

Table 4.4: Library design primers for pro-apoptotic anti-Bcl-2 bacterial surface display library

TEMPLATE	GCTGGCCAGTCTGGCCAGTATCCGTATGATGTGCCGGATTATGCGGCGGGCGGCAGCGGCGGCAGCGGCGGCCAGAGCATGCCGGAAATTTGGNNCGCAATGNDCCTGVNMAGANNCGGAGACGAANHCVNMG CATATNNCGCACGCGGAGGGCAGTCTGGGCAG
P1	GCTGGCCAGTCTGGCCAGTATCCGTATGATGTGCCGGATTATGCGGCGGGCGGCAGCGGCGGCAGCGGCGGCCAGAGCATGCGCDYARNASDACANAHAATGVVAGAATTGCGCCGCNTCGSAGACVAATTTVNCS VMTATTATGCGGGAGGGCAGTCTGGGCAG
P2	GCTGGCCAGTCTGGCCAGTATCCGTATGATGTGCCGGATTATGCGGCGGGCGGCAGCGGCGGCAGCGGCGGCCAGAGCGGTATGDYARNASDACANAHAGBAATGGAATTGCGCCGCNTCRSCGACVAATTTVNCS MMTATTATGCGGGAGGGCAGTCTGGGCAG
P5	GCTGGCCAGTCTGGCCAGTATCCGTATGATGTGCCGGATTATGCGGCGGGCGGCAGCGGCGGCAGCGGCGGCCAGAGCGGTCGCDYARNAATGCANAWAGBAVVAGAATTGATGCGCNTCRSCGACVAATTTVNC SMMTATTATGCGGGAGGGCAGTCTGGGCAG
P6	GCTGGCCAGTCTGGCCAGTATCCGTATGATGTGCCGGATTATGCGGCGGGCGGCAGCGGCGGCAGCGGCGGCCAGAGCGGTCGCDYARNASDAATGAHAGBAVVAGAATTGCGCATGNTCGSAGACVAATTTVNCS VMTATTATGCGGGAGGGCAGTCTGGGCAG
P12	GCTGGCCAGTCTGGCCAGTATCCGTATGATGTGCCGGATTATGCGGCGGGCGGCAGCGGCGGCAGCGGCGGCCAGAGCGGTCGCNYCRNASDACNAAHAGBAVVAGAATTGATGCGCNTCRSCGACVAATTTATGS MMTATTATGCGGGAGGGCAGTCTGGGCAG
P13	GCTGGCCAGTCTGGCCAGTATCCGTATGATGTGCCGGATTATGCGGCGGGCGGCAGCGGCGGCAGCGGCGGCCAGAGCGGTCGCDYARNASDACNAAWAGBASVAGAATTGCGCATGNTCRSCGACVAATTTVNM ATGTATTATGCGGGAGGGCAGTCTGGGCAG
P14	GCTGGCCAGTCTGGCCAGTATCCGTATGATGTGCCGGATTATGCGGCGGGCGGCAGCGGCGGCAGCGGCGGCCAGAGCGGTCGCDYARNASDACNAAHAGBAVVAGAATTGCGCCGCATGGSAGACVAATTTVNCS VMATGTATGCGGGAGGGCAGTCTGGGCAG

Table 4.5: Next generation sequencing primers for bacterial cell surface display

Name	Sequence
Fwd1	AATGATACGGCGACCACCGAGATCTACAC TAGATCGC TCG TCG GCA GCG TC
Fwd2	AATGATACGGCGACCACCGAGATCTACAC CTCTCTAT TCG TCG GCA GCG TC
Fwd3	AATGATACGGCGACCACCGAGATCTACAC TATCCTCT TCG TCG GCA GCG TC
Fwd4	AATGATACGGCGACCACCGAGATCTACAC AGAGTAGA TCG TCG GCA GCG TC
Fwd5	AATGATACGGCGACCACCGAGATCTACAC GTAAGGAG TCG TCG GCA GCG TC
Fwd6	AATGATACGGCGACCACCGAGATCTACAC ACTGCATA TCG TCG GCA GCG TC
Fwd7	AATGATACGGCGACCACCGAGATCTACAC AAGGAGTA TCG TCG GCA GCG TC
Fwd8	AATGATACGGCGACCACCGAGATCTACAC CTAAGCCT TCG TCG GCA GCG TC
Fwd9	AATGATACGGCGACCACCGAGATCTACAC CGTCTAAT TCG TCG GCA GCG TC
Fwd10	AATGATACGGCGACCACCGAGATCTACAC TCTCTCCG TCG TCG GCA GCG TC
Fwd11	AATGATACGGCGACCACCGAGATCTACAC TCGACTAG TCG TCG GCA GCG TC
Fwd12	AATGATACGGCGACCACCGAGATCTACAC TTCTAGCT TCG TCG GCA GCG TC
Rev1	CAAGCAGAAGACGGCATACGAGAT TAAGGCGA GTC TCG TGG GCT CGG
Rev2	CAAGCAGAAGACGGCATACGAGAT CGTACTAG GTC TCG TGG GCT CGG
Rev3	CAAGCAGAAGACGGCATACGAGAT AGGCAGAA GTC TCG TGG GCT CGG
Rev4	CAAGCAGAAGACGGCATACGAGAT TCCTGAGC GTC TCG TGG GCT CGG
Rev5	CAAGCAGAAGACGGCATACGAGAT GGACTCCT GTC TCG TGG GCT CGG
Rev6	CAAGCAGAAGACGGCATACGAGAT TAGGCATG GTC TCG TGG GCT CGG
Rev7	CAAGCAGAAGACGGCATACGAGAT CTCTCTAC GTC TCG TGG GCT CGG
Rev8	CAAGCAGAAGACGGCATACGAGAT CAGAGAGG GTC TCG TGG GCT CGG
Rev9	CAAGCAGAAGACGGCATACGAGAT GCTACGCT GTC TCG TGG GCT CGG
Rev10	CAAGCAGAAGACGGCATACGAGAT CGAGGCTG GTC TCG TGG GCT CGG
Rev11	CAAGCAGAAGACGGCATACGAGAT AAGAGGCA GTC TCG TGG GCT CGG
Rev12	CAAGCAGAAGACGGCATACGAGAT GTAGAGGA GTC TCG TGG GCT CGG
NGS first Fwd	TCG TCG GCA GCG TCA GAT GTG TAT AAG AGA CAG CAG GTA CTT CCG TAG CTG GCC AGT CT
NGS first Rev	GTC TCG TGG GCT CGG AGA TGT GTA TAA GAG ACA GCA CCG TAG ATG CTT GCC CAG TCG TTA

Table 4.6: Linear discriminant analysis classification performance for previously reported datasets

Dataset	Accuracy	Precision	Recall	F1
Bloom	0.88294	0.846908	0.85922	0.853019
Adams	0.626943	0.614035	0.714286	0.660377
Sarkisyan	0.837917	0.895413	0.769792	0.827864
Makowski - Affinity	0.915	0.85044	0.944625	0.895062
Makowski - Specificity	0.838554	0.616284	0.35281	0.448731
Jenson - Mcl-1	0.900783	0.863946	0.875862	0.869863
Jenson - Bfl-1	0.864035	0.878431	0.903226	0.878431
Jenson - Bcl-xL	0.865052	0.851648	0.928144	0.888252

Table 4.7: Linear discriminant analysis classification performance for stapled peptide library

Dataset	Accuracy	Precision	Recall	F1
Case – Bcl-xL	0.986	0.704	0.460	0.556
Case – Mcl-1	0.977	0.839	0.780	0.808

Table 4.8: Integer linear programming designed sequence and primers

Name	Sequence	full sequence
M1	GRLIMYIAQELMRIGDEFNDYYA	gctggccagtctggccagggccgcctgattatgtatattgcgcaggaactgatgcgcattggcgatgaatttaacgattattatgcggcgggcggcggcagcggcggtggcagc
M2	GRIIMYIAQELMRIGDEFNDYYA	gctggccagtctggccagggccgcattattatgtatattgcgcaggaactgatgcgcattggcgatgaatttaacgattattatgcggcgggcggcggcagcggcggtggcagc
M3	GRPIMYIAQELMRIGDEFNDYYA	gctggccagtctggccagggccgcccgattatgtatattgcgcaggaactgatgcgcattggcgatgaatttaacgattattatgcggcgggcggcggcagcggcggtggcagc
M4	GRVIMYIAQELMRIGDEFNDYYA	gctggccagtctggccagggccgcgtgattatgtatattgcgcaggaactgatgcgcattggcgatgaatttaacgattattatgcggcgggcggcggcagcggcggtggcagc
M5	GMLIQYIAMELIRIGDEFNDYYA	gctggccagtctggccagggcatgctgattcagtatattgcgatggaactgattcgcattggcgatgaatttaacgattattatgcggcgggcggcggcagcggcggtggcagc
M6	GRLIQYIAQELIMIGDEFNMYYA	gctggccagtctggccagggccgcctgattcagtatattgcgcaggaactgattatgattggcgatgaatttaacatgtattatgcggcgggcggcggcagcggcggtggcagc
M7	GRIIQYIAQELIMIGDEFNMYYA	gctggccagtctggccagggccgcattattcagtatattgcgcaggaactgattatgattggcgatgaatttaacatgtattatgcggcgggcggcggcagcggcggtggcagc
M8	GMLEQYIAMELIRIGDEFNDYYA	gctggccagtctggccagggcatgctggaacagtatattgcgatggaactgattcgcattggcgatgaatttaacgattattatgcggcgggcggcggcagcggcggtggcagc
M9	GRFIQYIAQELIMIGDEFNMYYA	gctggccagtctggccagggccgctttattcagtatattgcgcaggaactgattatgattggcgatgaatttaacatgtattatgcggcgggcggcggcagcggcggtggcagc
M10	MRLIQYIMQELIRIGDEFNDYYA	gctggccagtctggccagatgcgcctgattcagtatattatgcaggaactgattcgcattggcgatgaatttaacgattattatgcggcgggcggcggcagcggcggtggcagc
XM1	GRAEQMIGQELIMIGDEFNRYYA	gctggccagtctggccagggccgcgcgcggaacagatgattggccaggaactgattatgattggcgatgaatttaaccgctattatgcggcgggcggcggcagcggcggtggcagc
XM2	GRAEQMIGQELIMIADEFNRYYA	gctggccagtctggccagggccgcgcgcggaacagatgattggccaggaactgattatgattgcggatgaatttaaccgctattatgcggcgggcggcggcagcggcggtggcagc
XM3	GRLEQMIGQELIMIGDEFNRYYA	gctggccagtctggccagggccgcctggaacagatgattggccaggaactgattatgattggcgatgaatttaaccgctattatgcggcgggcggcggcagcggcggtggcagc
XM4	GRVEQMIGQELIMIGDEFNDYYA	gctggccagtctggccagggccgcgtggaacagatgattggccaggaactgattatgattggcgatgaatttaacgattattatgcggcgggcggcggcagcggcggtggcagc
XM5	GRAAQMIGQELIMIADEFNRYYA	gctggccagtctggccagggccgcgcggcgcagatgattggccaggaactgattatgattgcggatgaatttaaccgctattatgcggcgggcggcggcagcggcggtggcagc
XM6	GRFEQMIGQELIMIGDEFNDYYA	gctggccagtctggccagggccgctttgaacagatgattggccaggaactgattatgattggcgatgaatttaacgattattatgcggcgggcggcggcagcggcggtggcagc
XM7	GRSEQMIAQELIMIGDEFNRYYA	gctggccagtctggccagggccgcagcgaacagatgattgcgcaggaactgattatgattggcgatgaatttaaccgctattatgcggcgggcggcggcagcggcggtggcagc
XM8	GRPEQMIAQELIMIGDEFNRYYA	gctggccagtctggccagggccgcccggaacagatgattgcgcaggaactgattatgattggcgatgaatttaaccgctattatgcggcgggcggcggcagcggcggtggcagc
XM9	GRFEQMIGAELIMIGDEFNRYYA	gctggccagtctggccagggccgctttgaacagatgattggcgcggaactgattatgattggcgatgaatttaaccgctattatgcggcgggcggcggcagcggcggtggcagc
XM10	GRTEQMIGQELIMIGDEFNRYYA	gctggccagtctggccagggccgcaccgaacagatgattggccaggaactgattatgattggcgatgaatttaaccgctattatgcggcgggcggcggcagcggcggtggcagc
X1	GRAAQMKGAELRMLADEFARYYA	gctggccagtctggccagggccgcgcgcggcgcagatgaaaggcgcggaactgcgcatgctggcggatgaatttgcgcgctattatgcggcgggcggcggcagcggcggtggcagc
X2	GRAAQMKGQELRMLADEFLRYYA	gctggccagtctggccagggccgcgcgcggcgcagatgaaaggccaggaactgcgcatgctggcggatgaatttctgcgctattatgcggcgggcggcggcagcggcggtggcagc
X3	GRAEQMKGAELRMLADEFLRYYA	gctggccagtctggccagggccgcgcgcggaacagatgaaaggcgcggaactgcgcatgctggcggatgaatttctgcgctattatgcggcgggcggcggcagcggcggtggcagc
X4	GRAEVMKGAELRMLADEFARYYA	gctggccagtctggccagggccgcgcgcggaagtgatgaaaggcgcggaactgcgcatgctggcggatgaatttgcgcgctattatgcggcgggcggcggcagcggcggtggcagc
X5	GRAALMKGQELRMLADEFARYYA	gctggccagtctggccagggccgcgcgcggcgctgatgaaaggccaggaactgcgcatgctggcggatgaatttgcgcgctattatgcggcgggcggcggcagcggcggtggcagc
X6	GRAELMKGQELRMLADEFLRYYA	gctggccagtctggccagggccgcgcgcggaactgatgaaaggccaggaactgcgcatgctggcggatgaatttctgcgctattatgcggcgggcggcggcagcggcggtggcagc
X7	GRAAQMKGQELRMVADEFLRYYA	gctggccagtctggccagggccgcgcgcggcgcagatgaaaggccaggaactgcgcatggtggcggatgaatttctgcgctattatgcggcgggcggcggcagcggcggtggcagc
X8	GRAAVMKGQELRMFADEFHRYYA	gctggccagtctggccagggccgcgcgcggcggtgatgaaaggccaggaactgcgcatgtttgcggatgaatttcatcgctattatgcggcgggcggcggcagcggcggtggcagc
X9	GRAEEMKGQELRMLADEFLRYYA	gctggccagtctggccagggccgcgcgcggaagaaatgaaaggccaggaactgcgcatgctggcggatgaatttctgcgctattatgcggcgggcggcggcagcggcggtggcagc
X10	GRAANMKGQELRMLADEFARYYA	gctggccagtctggccagggccgcgcggcgaacatgaaaggccaggaactgcgcatgctggcggatgaatttgcgcgctattatgcggcgggcggcggcagcggcggtggcagc

Table 4.9: Integer linear programming designed sequences for Bcl-xL design 2

Name	Sequence	primer sequence
1	GRAAVMKGAELRMLADEFNRYYA	gctggccagtctggccagggccgcgcggcggtgatgaaaggcgcggaactgcgcatgctggcggatgaatttaaccgctattatgcggcgggcggcggcagcggcggtggcagc
2	GRAAVMKGAELRMLADQFNRYYA	gctggccagtctggccagggccgcgcggcggtgatgaaaggcgcggaactgcgcatgctggcggatcagtttaaccgctattatgcggcgggcggcggcagcggcggtggcagc
3	GRAAVMKGQELRMLADEFNRYYA	gctggccagtctggccagggccgcgcggcggtgatgaaaggccaggaactgcgcatgctggcggatgaatttaaccgctattatgcggcgggcggcggcagcggcggtggcagc
4	GRAAVMKGAELRMFADEFNRYYA	gctggccagtctggccagggccgcgcggcggtgatgaaaggcgcggaactgcgcatgtttgcggatgaatttaaccgctattatgcggcgggcggcggcagcggcggtggcagc
5	GRATVMKGQELRMLADEFNRYYA	gctggccagtctggccagggccgcgcgaccgtgatgaaaggccaggaactgcgcatgctggcggatgaatttaaccgctattatgcggcgggcggcggcagcggcggtggcagc
6	GRAAVMKGEELRMLADEFNRYYA	gctggccagtctggccagggccgcgcggcggtgatgaaaggcgaagaactgcgcatgctggcggatgaatttaaccgctattatgcggcgggcggcggcagcggcggtggcagc
7	GRAAVMKGAELRMLADEFNPYYA	gctggccagtctggccagggccgcgcggcggtgatgaaaggcgcggaactgcgcatgctggcggatgaatttaacccgtattatgcggcgggcggcggcagcggcggtggcagc
8	GRAAVMKGQELRMLADKFNRYYA	gctggccagtctggccagggccgcgcggcggtgatgaaaggccaggaactgcgcatgctggcggataaatttaaccgctattatgcggcgggcggcggcagcggcggtggcagc
9	GRAAVMKGQELRMLADQFNRYYA	gctggccagtctggccagggccgcgcggcggtgatgaaaggccaggaactgcgcatgctggcggatcagtttaaccgctattatgcggcgggcggcggcagcggcggtggcagc
10	GRAAQMKGAELRMLADEFNRYYA	gctggccagtctggccagggccgcgcggcgcagatgaaaggcgcggaactgcgcatgctggcggatgaatttaaccgctattatgcggcgggcggcggcagcggcggtggcagc
11	GRATQMKGAELRMLADEFNRYYA	gctggccagtctggccagggccgcgcgacccagatgaaaggcgcggaactgcgcatgctggcggatgaatttaaccgctattatgcggcgggcggcggcagcggcggtggcagc
12	GRAAQMKGAELRMLADKFNRYYA	gctggccagtctggccagggccgcgcggcgcagatgaaaggcgcggaactgcgcatgctggcggataaatttaaccgctattatgcggcgggcggcggcagcggcggtggcagc
13	GRAAVMKGQELRMFADEFNRYYA	gctggccagtctggccagggccgcgcggcggtgatgaaaggccaggaactgcgcatgtttgcggatgaatttaaccgctattatgcggcgggcggcggcagcggcggtggcagc
14	GRAAQMKGAELRMLADQFNRYYA	gctggccagtctggccagggccgcgcggcgcagatgaaaggcgcggaactgcgcatgctggcggatcagtttaaccgctattatgcggcgggcggcggcagcggcggtggcagc
15	GRAAVMKGQELRMLADEFNPYYA	gctggccagtctggccagggccgcgcggcggtgatgaaaggccaggaactgcgcatgctggcggatgaatttaacccgtattatgcggcgggcggcggcagcggcggtggcagc
16	GRAAVMKGEELRMFADEFNRYYA	gctggccagtctggccagggccgcgcggcggtgatgaaaggcgaagaactgcgcatgtttgcggatgaatttaaccgctattatgcggcgggcggcggcagcggcggtggcagc
17	GRAALMKGAELRMLADEFNRYYA	gctggccagtctggccagggccgcgcggcgctgatgaaaggcgcggaactgcgcatgctggcggatgaatttaaccgctattatgcggcgggcggcggcagcggcggtggcagc
18	GRAAQMKGQELRMLADEFNRYYA	gctggccagtctggccagggccgcgcggcgcagatgaaaggccaggaactgcgcatgctggcggatgaatttaaccgctattatgcggcgggcggcggcagcggcggtggcagc
19	GRAAVMKGQELRMFADQFNRYYA	gctggccagtctggccagggccgcgcggcggtgatgaaaggccaggaactgcgcatgtttgcggatcagtttaaccgctattatgcggcgggcggcggcagcggcggtggcagc
20	GRAAQMKGAELRMFADEFNRYYA	gctggccagtctggccagggccgcgcggcgcagatgaaaggcgcggaactgcgcatgtttgcggatgaatttaaccgctattatgcggcgggcggcggcagcggcggtggcagc
21	GRATQMKGQELRMLADEFNRYYA	gctggccagtctggccagggccgcgcgacccagatgaaaggccaggaactgcgcatgctggcggatgaatttaaccgctattatgcggcgggcggcggcagcggcggtggcagc
22	GRATQMKGAELRMFADEFNRYYA	gctggccagtctggccagggccgcgcgacccagatgaaaggcgcggaactgcgcatgtttgcggatgaatttaaccgctattatgcggcgggcggcggcagcggcggtggcagc
23	GRAAQMKGEELRMLADEFNRYYA	gctggccagtctggccagggccgcgcggcgcagatgaaaggcgaagaactgcgcatgctggcggatgaatttaaccgctattatgcggcgggcggcggcagcggcggtggcagc
24	GRAAQMKGAELRMLADEFNPYYA	gctggccagtctggccagggccgcgcggcgcagatgaaaggcgcggaactgcgcatgctggcggatgaatttaacccgtattatgcggcgggcggcggcagcggcggtggcagc
25	GRAAQMKGQELRMLADKFNRYYA	gctggccagtctggccagggccgcgcggcgcagatgaaaggccaggaactgcgcatgctggcggataaatttaaccgctattatgcggcgggcggcggcagcggcggtggcagc
26	GRAAQMKGAELRMFADKFNRYYA	gctggccagtctggccagggccgcgcggcgcagatgaaaggcgcggaactgcgcatgtttgcggataaatttaaccgctattatgcggcgggcggcggcagcggcggtggcagc
27	GRATQMKGEELRMLADEFNRYYA	gctggccagtctggccagggccgcgcgacccagatgaaaggcgaagaactgcgcatgctggcggatgaatttaaccgctattatgcggcgggcggcggcagcggcggtggcagc
28	GRAAQMKGQELRMLADQFNRYYA	gctggccagtctggccagggccgcgcggcgcagatgaaaggccaggaactgcgcatgctggcggatcagtttaaccgctattatgcggcgggcggcggcagcggcggtggcagc
29	GRAAQMKGAELRMFADQFNRYYA	gctggccagtctggccagggccgcgcggcgcagatgaaaggcgcggaactgcgcatgtttgcggatcagtttaaccgctattatgcggcgggcggcggcagcggcggtggcagc
30	GRAAQMKGEELRMLADKFNRYYA	gctggccagtctggccagggccgcgcggcgcagatgaaaggcgaagaactgcgcatgctggcggataaatttaaccgctattatgcggcgggcggcggcagcggcggtggcagc

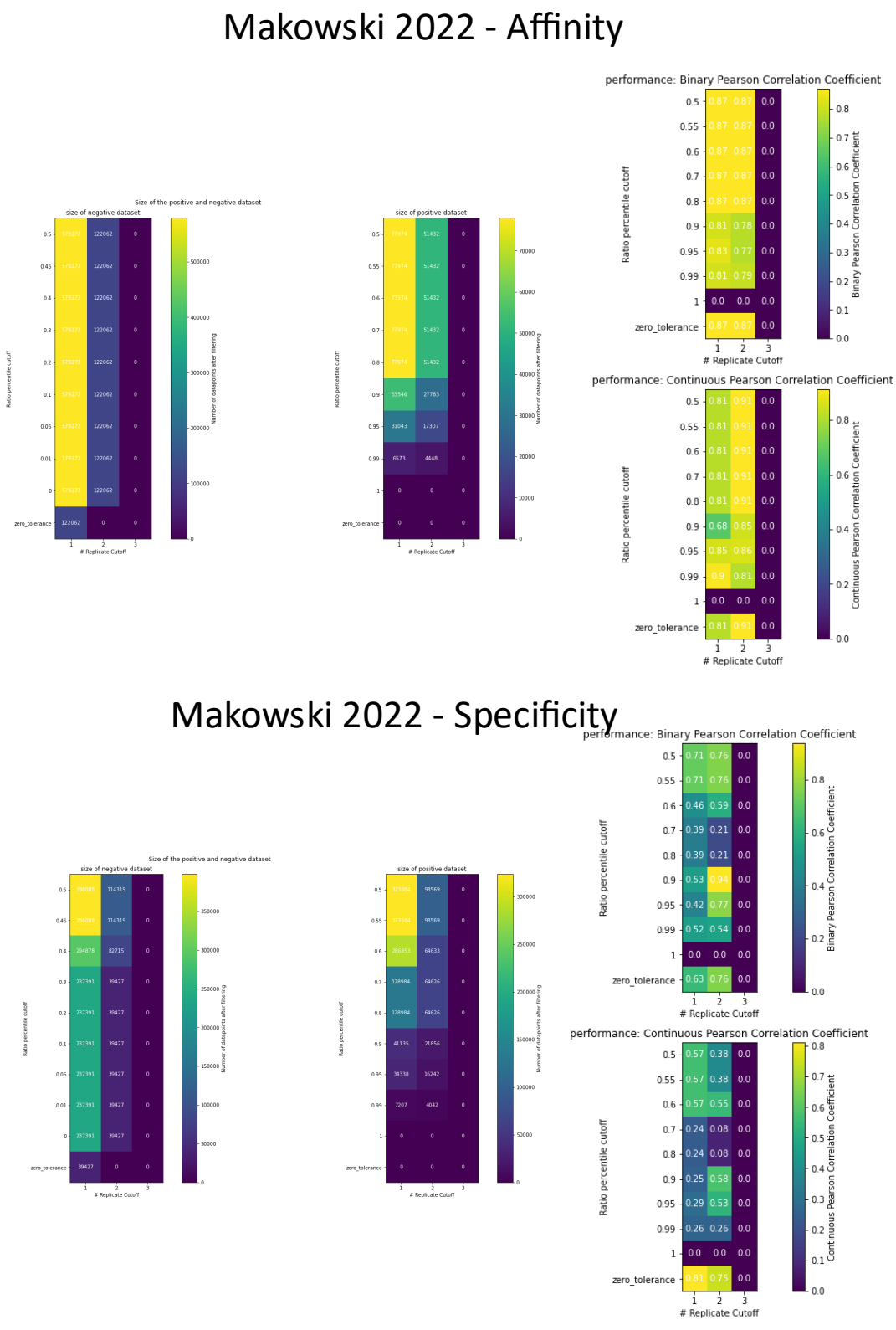


Figure 4.11: Dataset hyperparameters for Makowski et al. (2022) Nature Communications.

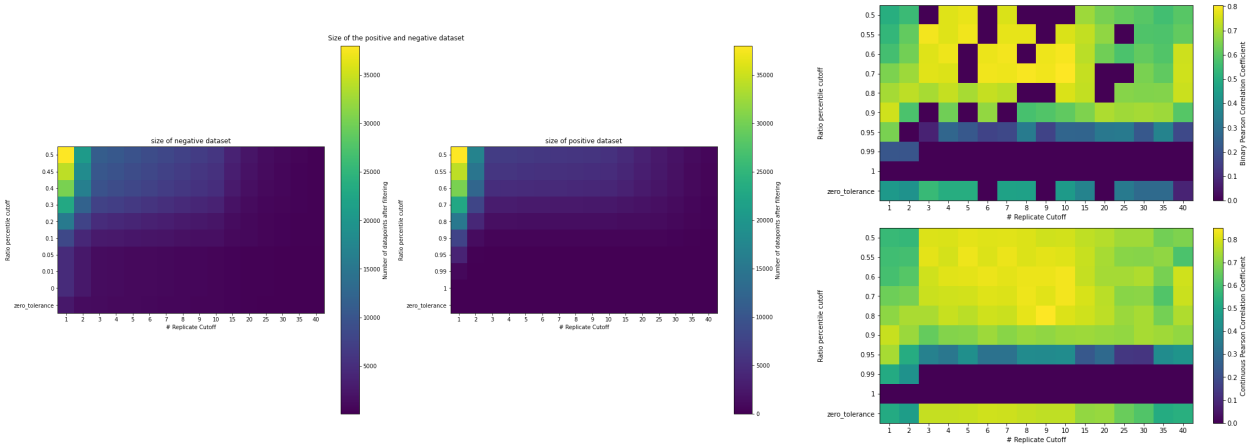


Figure 4.12: Dataset hyperparameters for Starr et al. (2022) Science.

Sarkisyan 2016

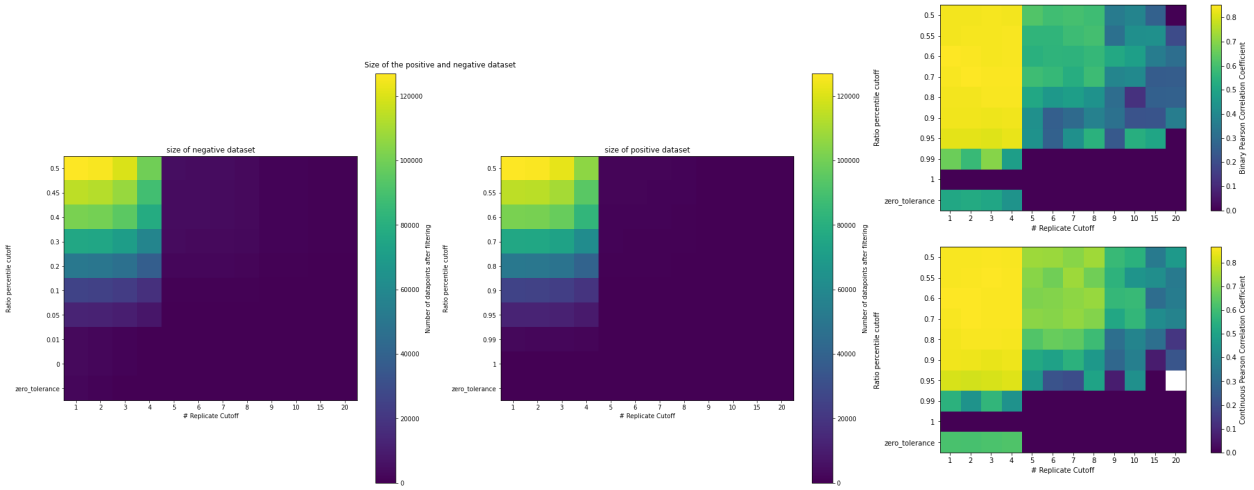


Figure 4.13: Dataset hyperparameters for Sarkisyan et al. (2016) Nature.

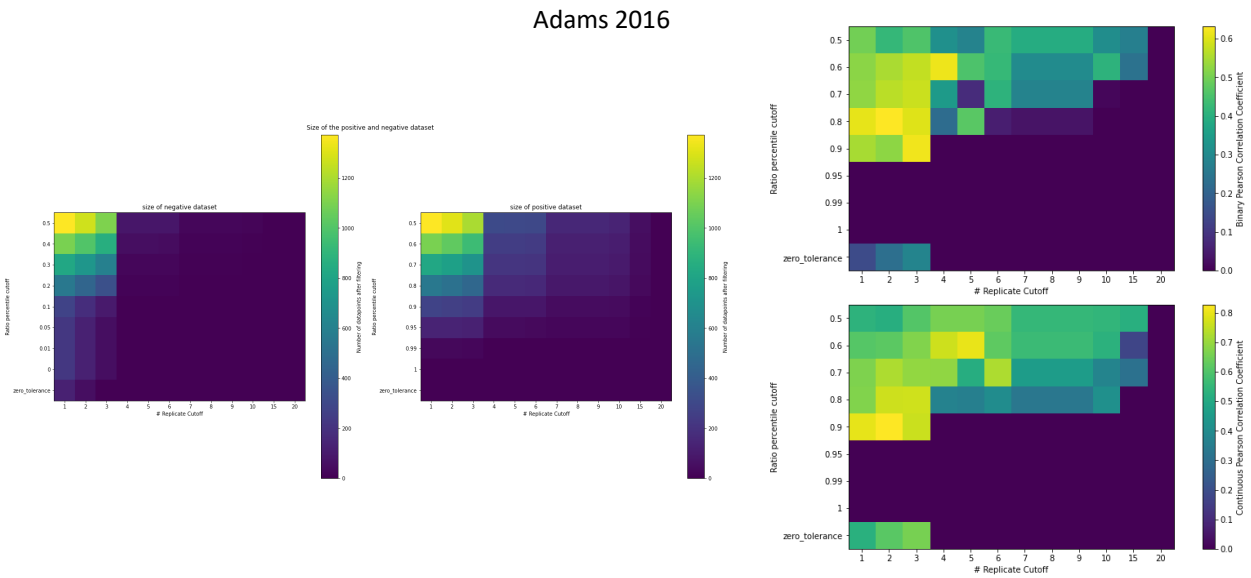


Figure 4.14: Dataset hyperparameters for Adams et al. (2016) eLife.

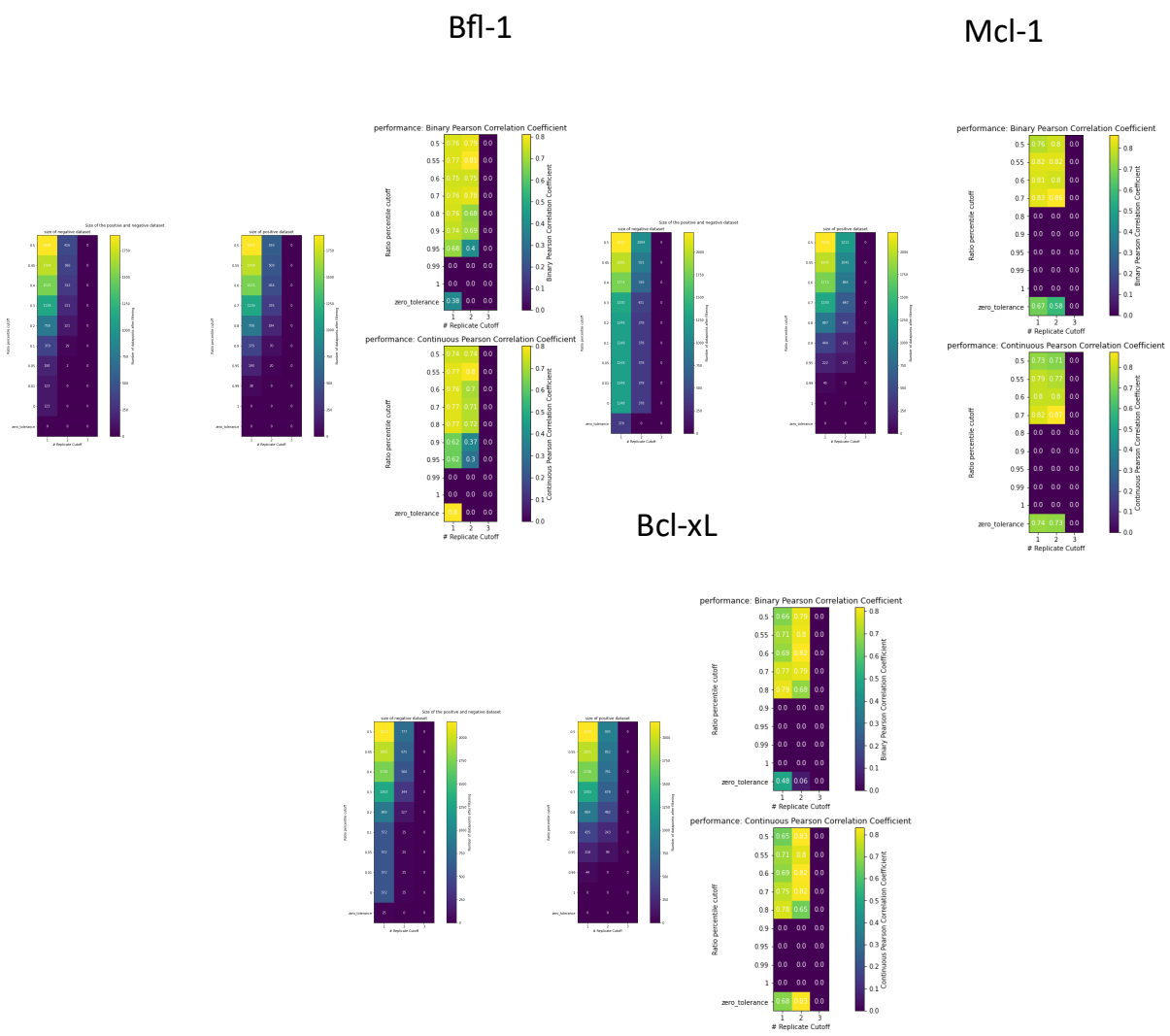


Figure 4.15: Dataset hyperparameters for Jenson et al. (2018) PNAS.

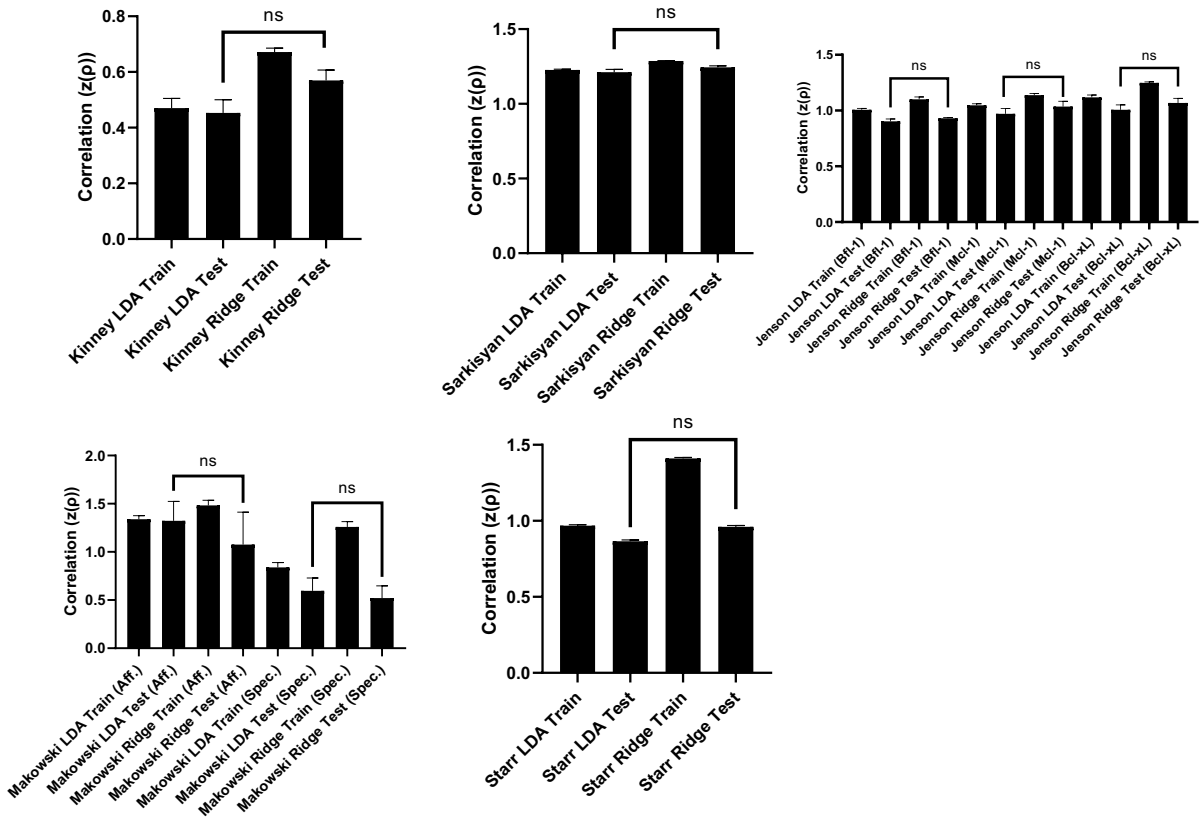


Figure 4.16: Training and Test Set Performance statistics

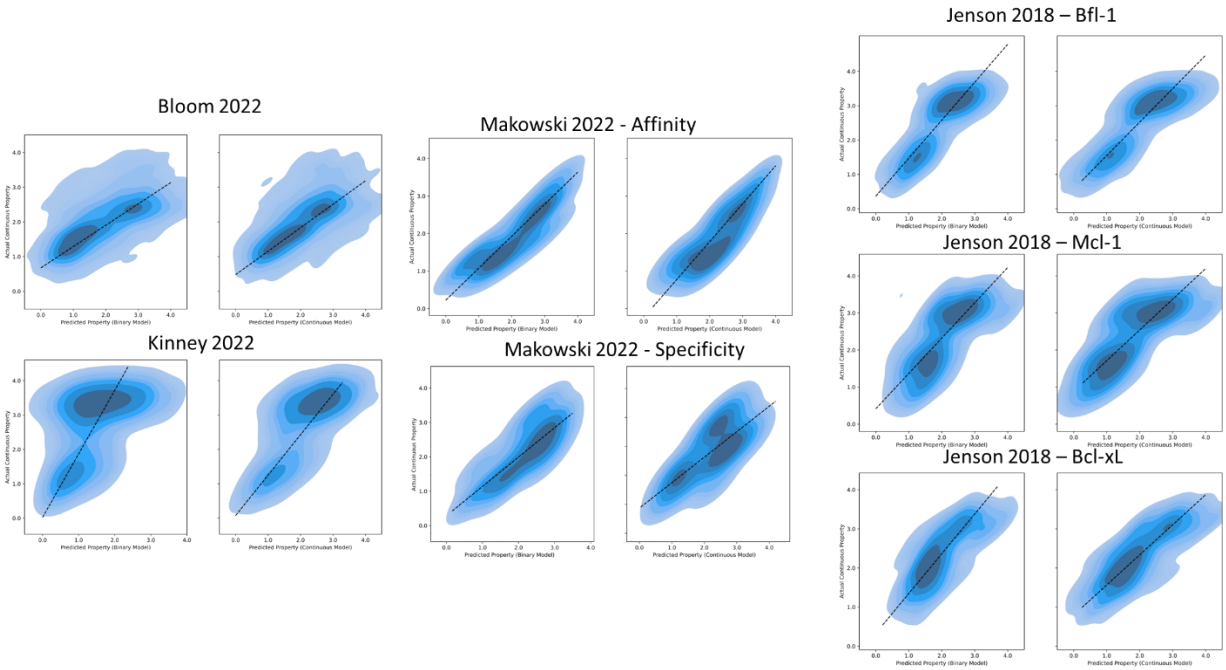


Figure 4.17: Kernel density estimates for linear discriminant models projects' correlations with continuous protein property values.

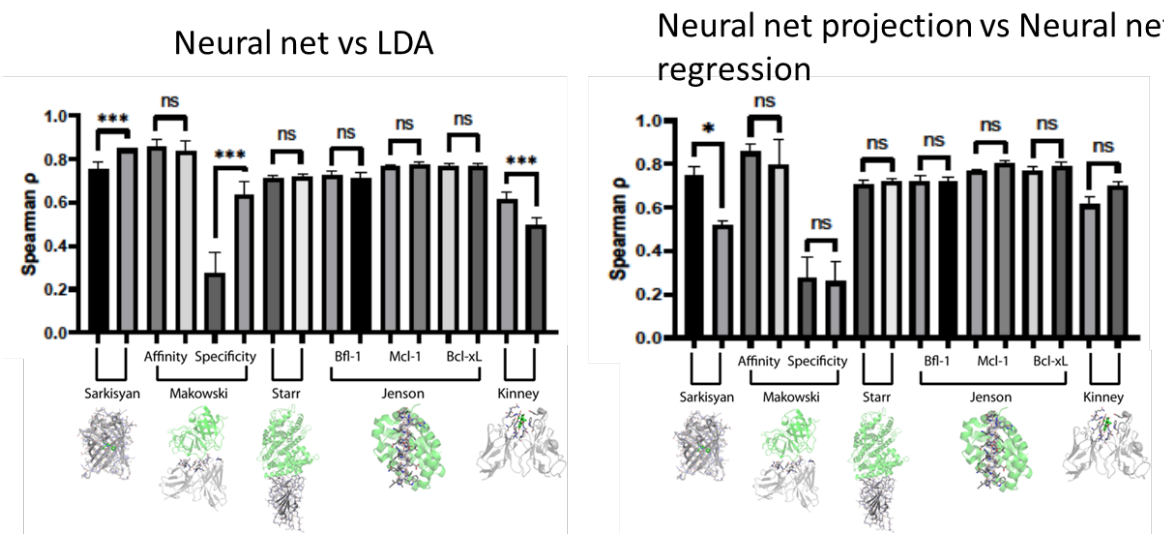


Figure 4.18: Neural net performance statistics. (Left) Neural net performance statistics versus linear discriminant analysis for test set. For all datasets, neural nets are on the left and LDA is on the right. **(Right)** Neural net classifier prediction of continuous mode versus neural net regressor for test set. For all datasets, neural net classifiers are on the left and the regressors are on the right. (*: $p < 0.05$, ***: $p<0.0001$).

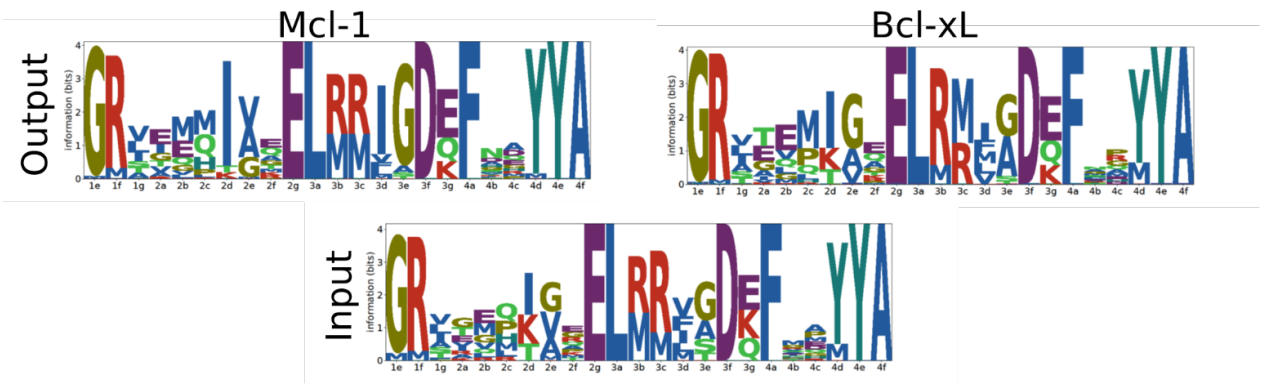


Figure 4.19: **Logoplots of input and output (pre- and post- sorting, respectively) for Bcl-xL and Mcl-1 stapled peptide libraries**

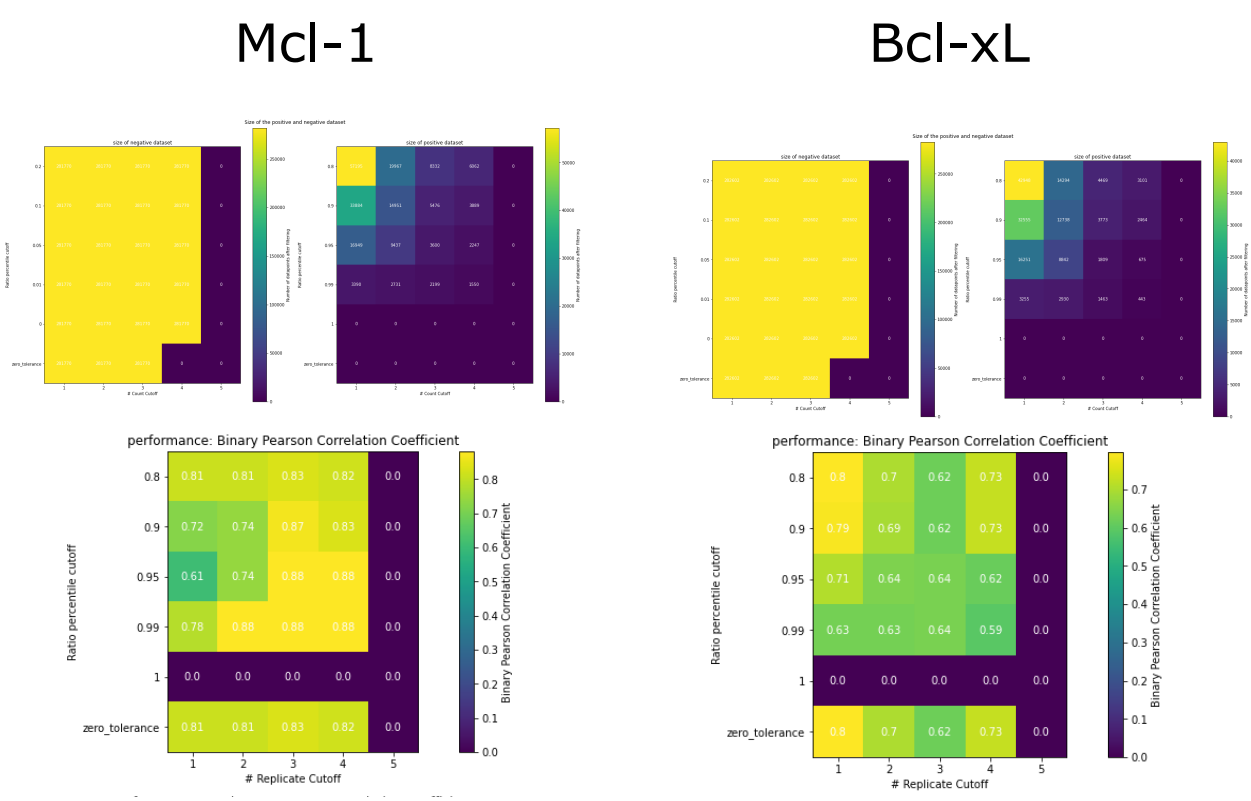


Figure 4.20: Dataset hyperparameter data size and performance for pro-apoptotic anti-Bcl-2 stapled peptide libraries.

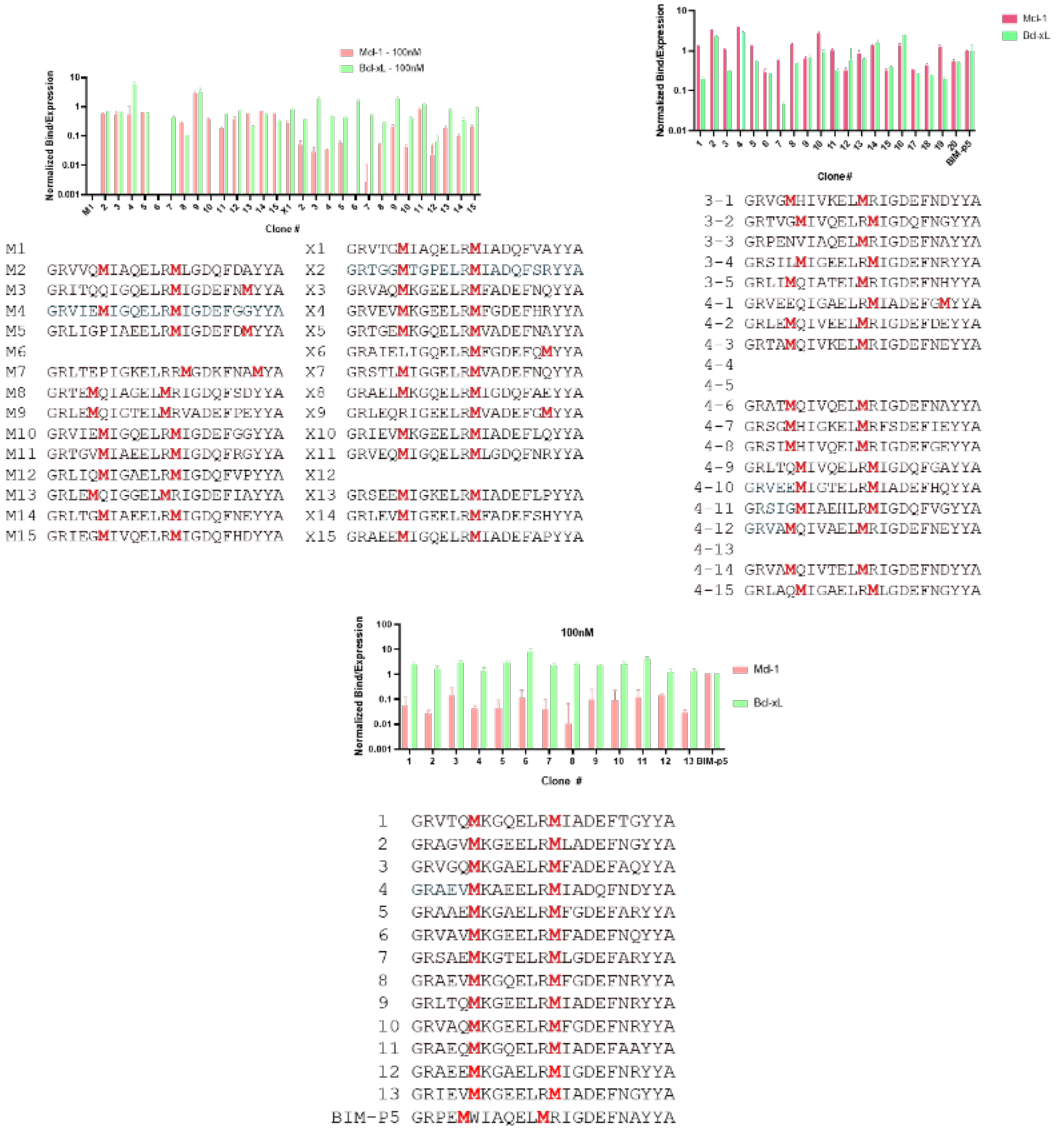
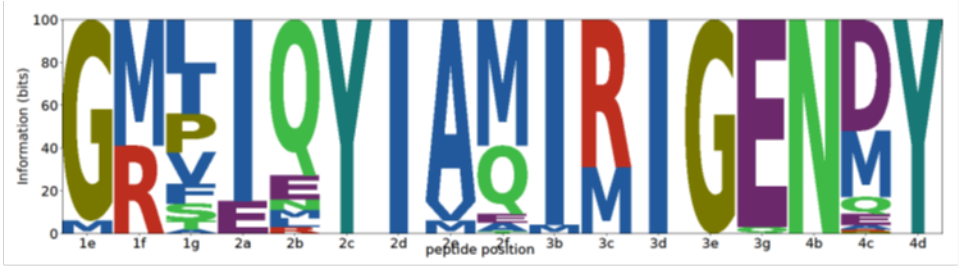


Figure 4.21: **Random variants from Mcl-1 and Bcl-xL FACS 2-4 for low-throughput continuous binding measurement via bacterial cell surface and flow cytometry.**




```
GRPEMWIAQELMRIGDEFNAYYA - BIM-p5
GRVAMQIVTELMRIGDEFNDYYA - Sanger 1
GRTAMQIVKELMRIGDEFNEYYA - Sanger 2
GRVIMYIAQELMRIGDEFNDYYA - Opt. 1
GRLIMYIAQELMRIGDEFNDYYA - Opt. 2
```

Figure 4.22: Sequences for select variants from Figure 4.6.

References

1. Anfinsen, C. B. Principles that Govern the Folding of Protein Chains. *Science (1979)* **181**, 223–232 (1973).
2. Cobb, R. E., Chao, R. & Zhao, H. Directed evolution: Past, present, and future. *AIChE Journal* **59**, 1432–1440 (2013).
3. Lerner, S. A., Wu, T. T. & Lin, E. C. C. Evolution of a catabolic pathway in bacteria. *Science (1979)* **146**, 1313–1315 (1964).
4. Roberts, R. W. & Szostak, J. W. RNA-peptide fusions for the in vitro selection of peptides and proteins. *Biochemistry* **94**, 12297–12302 (1997).
5. Smith, G. P. Filamentous Fusion Phage: Novel Expression Vectors That Display Cloned Antigens on the Virion Surface. *Science (1979)* **228**, 1315–1317 (1984).
6. Freudl, R., MacIntyre, S., Degen, M. & Henning, U. Cell Surface Exposure of the Outer Membrane Protein OmpA of Escherichia coli K-12. *J Mol Biol* **188**, 491–494 (1985).
7. Boder, E. T. & Wittrup, K. D. Yeast surface display for screening combinatorial polypeptide libraries. *Nat Biotechnol* **15**, 553–557 (1997).
8. Yang, K. K., Wu, Z. & Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nat Methods* **16**, 687–694 (2019).
9. Liu, B. *Yeast surface display: Methods, protocols, and applications*. *Yeast Surface Display: Methods, Protocols, and Applications* vol. 1319 (2015).
10. Barreto, K. *et al.* Next-generation sequencing-guided identification and reconstruction of antibody CDR combinations from phage selection outputs. *Nucleic Acids Res* **47**, (2019).
11. D'Angelo, S. *et al.* From deep sequencing to actual clones. *Protein Engineering, Design and Selection* **27**, 301–307 (2014).

12. Ravn, U. *et al.* By-passing in vitro screening - Next generation sequencing technologies applied to antibody display and in silico candidate selection. *Nucleic Acids Res* **38**, 1–11 (2010).

13. Rubin, A. F. *et al.* A statistical framework for analyzing deep mutational scanning data. *Genome Biol* **18**, 1–15 (2017).

14. Kowalsky, C. A. *et al.* Rapid fine conformational epitope mapping using comprehensive mutagenesis and deep sequencing. *Journal of Biological Chemistry* **290**, 26457–26470 (2015).

15. Fowler, D. M., Araya, C. L., Gerard, W. & Fields, S. Enrich: Software for analysis of protein function by enrichment and depletion of variants. *Bioinformatics* **27**, 3430–3431 (2011).

16. Derda, R. *et al.* Diversity of phage-displayed libraries of peptides during panning and amplification. *Molecules* **16**, 1776–1803 (2011).

17. Chao, G. *et al.* Isolating and engineering human antibodies using yeast surface display. *Nat Protoc* **1**, 755–768 (2006).

18. Kelil, A., Gallo, E., Banerjee, S., Adams, J. J. & Sidhu, S. S. CellectSeq: In silico discovery of antibodies targeting integral membrane proteins combining in situ selections and next-generation sequencing. *Commun Biol* **4**, (2021).

19. Maranhão, A. Q. *et al.* Discovering Selected Antibodies From Deep-Sequenced Phage-Display Antibody Library Using ATTILA. *Bioinform Biol Insights* **14**, 1–8 (2020).

20. Fowler, D. M. & Fields, S. Deep mutational scanning: A new style of protein science. *Nat Methods* **11**, 801–807 (2014).

21. Adams, R. M., Mora, T., Walczak, A. M. & Kinney, J. B. Measuring the sequence-affinity landscape of antibodies with massively parallel titration curves. *Elife* **5**, 1–27 (2016).

22. Kinney, J. B., Murugan, A., Callan, C. G. & Cox, E. C. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. **107**, 9158–9163 (2010).

23. Reich, L., Dutta, S. & Keating, A. E. SORTCERY - A High-Throughput Method to Affinity Rank Peptide Ligands. *J Mol Biol* **427**, 2135–2150 (2015).

24. Jenson, J. M. *et al.* Peptide design by optimization on a data parameterized protein interaction landscape. *Proc Natl Acad Sci U S A* **115**, E10342–E10351 (2018).

25. Greaney, A. J. *et al.* Comprehensive mapping of mutations in the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human plasma antibodies. *Cell Host Microbe* **29**, 463-476.e6 (2021).

26. Starr, T. N. *et al.* Shifting mutational constraints in the SARS-CoV-2 receptor-binding domain during viral evolution. *Science (1979)* **377**, 420–424 (2022).

27. Makowski, E. K. *et al.* Co-optimization of therapeutic antibody affinity and specificity using machine learning models that generalize to novel mutational space. *Nat Commun* **13**, 1–14 (2022).

28. Sarkisyan, K. S. *et al.* Local fitness landscape of the green fluorescent protein. *Nature* **533**, 397–401 (2016).

29. Hsu, C., Nisonoff, H., Fannjiang, C. & Listgarten, J. Learning protein fitness models from evolutionary and assay-labeled data. *Nat Biotechnol* **40**, 1114–1122 (2022).

30. Srivastava, N., Hinton, G., Krizhevsky, A. & Salakhutdinov, R. *Dropout: A Simple Way to Prevent Neural Networks from Overfitting*. *Journal of Machine Learning Research* vol. 15 (2014).

31. Navaratna, T. *et al.* Directed Evolution Using Stabilized Bacterial Peptide Display. *J Am Chem Soc* **142**, 1882–1894 (2020).

32. Case, M., Navaratna, T., Vinh, J. & Thurber, G. M. Rapid Evaluation of Staple Placement in Stabilized Alpha Helices using Bacterial Surface Display. *ACS Chem Biol* **18**, 905–914 (2023).

33. Gaspar, J. M. NGmerge: Merging paired-end reads via novel empirically-derived models of sequencing errors. *BMC Bioinformatics* **19**, 1–9 (2018).

34. Romero, P. A., Krause, A. & Arnold, F. H. Navigating the protein fitness landscape with Gaussian processes. *Proc Natl Acad Sci U S A* **110**, E291–E201 (2013).

35. Yang, K. K., Wu, Z. & Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nat Methods* **16**, 687–694 (2019).

36. Wu, Z., Jennifer Kan, S. B., Lewis, R. D., Wittmann, B. J. & Arnold, F. H. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc Natl Acad Sci U S A* **116**, 8852–8858 (2019).

37. Mitchell, S. *et al.* Optimization with PuLP. Preprint at (2009).

38. Trippe, B. L. *et al.* Randomized gates eliminate bias in sort-seq assays. *Protein Science* **31**, 1–8 (2022).

39. Somermeyer, L. G. *et al.* Heterogeneity of the GFP fitness landscape and data-driven protein design. *bioRxiv* 1–54 (2021).

40. Raghunathan, T. E., Rosenthal, R. & Rubin, D. B. *Comparing Correlated but Nonoverlapping Correlations*. *Psychological Methods* vol. 1 (1996).

41. Makowski, E. K., Wu, L., Desai, A. A. & Tessier, P. M. Highly sensitive detection of antibody nonspecific interactions using flow cytometry. *MAbs* **13**, 1–11 (2021).

42. Makowski, E. K. *et al.* Reduction of therapeutic antibody self-association using yeast-display selections and machine learning. *MAbs* **14**, 1–15 (2022).

43. Dutta, S. Determinants of BH3 binding specificity for Mcl-1 vs. Bcl-xL. *J Mol Biol* **398**, 747–762 (2011).

44. Dutta, S. *et al.* Potent and specific peptide inhibitors of human pro-survival protein bcl-xl. *J Mol Biol* **427**, 1241–1253 (2015).

45. Loren D. Walensky *et al.* Activation of Apoptosis in Vivo by a Hydrocarbon-Stapled BH3 Helix. *Science (1979)* **23**, 1–7 (2004).

46. Walensky, L. D. & Bird, G. H. Hydrocarbon-stapled peptides: principles, practice, and progress. *J Med Chem* **57**, 6275–6288 (2014).

47. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118**, e2016239118 (2021).

48. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat Methods* **16**, 1315–1322 (2019).

49. Riesselman, A. J., Ingraham, J. B. & Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nat Methods* **15**, 816–822 (2018).

50. Shin, J. E. *et al.* Protein design and variant prediction using autoregressive generative models. *Nat Commun* **12**, 1–11 (2021).

51. Adams, J. M. & Cory, S. The Bcl-2 Protein Family: Arbiters of Cell Survival. *Science (1979)* **281**, 1322–1326 (1998).

52. Shamas-Din, A., Kale, J., Leber, B. & Andrews, D. W. Mechanisms of action of Bcl-2 family proteins. *Cold Spring Harb Perspect Biol* **5**, 1–21 (2013).

53. Czabotar, P. E., Lessene, G., Strasser, A. & Adams, J. M. Control of apoptosis by the BCL-2 protein family: Implications for physiology and therapy. *Nat Rev Mol Cell Biol* **15**, 49–63 (2014).

54. Hie, B. L. *et al.* Efficient evolution of human antibodies from general protein language models. *Nat Biotechnol* (2023) doi:10.1038/s41587-023-01763-2.

55. Kang, Y., Leng, D., Guo, J. & Pan, L. Sequence-based deep learning antibody design for in silico antibody affinity maturation. *ArXiv* 1–9 (2021).

56. Ruffolo, J. A., Gray, J. J. & Sulam, J. Deciphering antibody affinity maturation with language models and weakly supervised learning. (2021).

57. Amimeur, T. *et al.* Designing Feature-Controlled Humanoid Antibody Discovery Libraries Using Generative Adversarial Networks. *BioRxiv* 1–34 (2020) doi:10.1101/2020.04.12.024844.

58. Jain, T. *et al.* Biophysical properties of the clinical-stage antibody landscape. *Proc Natl Acad Sci U S A* **114**, 944–949 (2017).

59. Taguchi, A. T. *et al.* Comprehensive Prediction of Molecular Recognition in a Combinatorial Chemical Space Using Machine Learning. *ACS Comb Sci* **22**, 500–508 (2020).

60. Mason, D. M. *et al.* Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nat Biomed Eng* **5**, 600–612 (2021).

61. Chu, Q. *et al.* Towards understanding cell penetration by stapled peptides. *Medchemcomm* **6**, 111–119 (2015).

62. Bird, G. H. *et al.* Biophysical determinants for cellular uptake of hydrocarbon-stapled peptide helices. *Nat Chem Biol* **12**, 845–852 (2016).

63. Bird, G. H. *et al.* Hydrocarbon double-stapling remedies the proteolytic instability of a lengthy peptide therapeutic. *Proceedings of the National Academy of Sciences* **107**, 14093–14098 (2010).

64. Chandramohan, A. *et al.* Design-Rules for Stapled Alpha-Helical Peptides with On-Target In Vivo Activity: Application to Mdm2/X dual antagonists. *Biorxiv* 1–55 (2023) doi:10.1101/2023.02.25.530030.

Chapter 5 Conclusion

This thesis has advanced protein engineering methods towards the design of stapled peptide therapeutics, the extraction of hidden information from simple binary sorting and sequencing data, and a powerful optimization tool to extrapolate beyond assayed sequence space.

Summary

In **Chapter 2**, this thesis describes how stabilized peptide engineering by *E. coli* display (SPEED) can be expanded to assay more design parameters than previously measurable: binding specificity, staple location, staple chemistry, and hot spot analysis. To establish these design parameters as assayable via SPEED, we generated numerous novel stapled peptide constructs on the bacterial cell surface and measured their binding affinities and specificities. We start by performing a hot spot analysis and analyzing the sequence space of mdm2 peptide variants stapled with different bis-alkynes and report many new high affinity sequences with diverse chemical properties. Interestingly, many of these sequences include new disulfide motifs, including i,i+1, i,i+4, and i,i+5 motifs, each of which likely results in a different peptide structure that may contribute differently to important drug-like properties such as protease susceptibility and cell permeability. Next, this chapter establishes that both staple sequence and staple location are key determinants of peptide binding affinity and specificity. Among the two classes of protein-protein interactions studied, mdm2/p53 and Bcl-2, both the staple location and sequence were key determinants of binding affinity, suggesting that methods that can simultaneously evaluate both criteria may yield better therapeutics. We also show that there is a complex

relationship between the staple position and specificity among Bcl-2 proteins in the context of BIM, which suggests that approaches able to simultaneously evaluate sequence and staple position are best poised for Bcl-2 antagonist design owing to the importance of specificity. To demonstrate SPEED's accuracy for peptide property prediction, we also synthesize many mdm2- and bcl-2-targeted stapled peptides and compare the binding properties in solution with cell surface measurements and show they are highly correlated. With these new design criteria evaluated for high-throughput analysis via SPEED, staple peptide antagonists of protein-protein interactions can be developed with improved speed and cost. This work also confirms that SPEED is a reliable tool in a peptide engineer's arsenal towards the accurate measurements of binding affinities through the comparison of dozens of bacterial surface displayed- peptides with their soluble counterparts. We generated a large dataset of peptides measured via SPEED and via established solution phase assays such as biolayer interferometry and confirmed that these measurements were in close agreement.

To apply these new design principles and to generalize SPEED to new protein-protein interactions, **Chapter 3** describes the design of highly specific Bcl-x_L stapled peptides. This chapter details high-throughput cell sorting strategies to identify high fitness proteins or proteins that demonstrate favorable trade-offs between properties (affinity and specificity). Bcl-x_L is a protein within the B cell lymphoma 2 (Bcl-2) family whose role in regulating (and resistance of) apoptosis makes it a highly important drug target. Its structure and location has rendered it a difficult target for common therapeutic modalities; the stapled peptides generated in this chapter are an advance in the binding affinities and specificities and may yield improved therapeutics with subsequent analysis and optimization *in vitro* and *in vivo* Stapled peptides targeting Bcl-xL (a protein in the B cell lymphoma 2 protein family) are engineered, which promote apoptotic

pathways in dysregulated cancer cells with high binding affinity and specificity. We describe an approach for building libraries of protein variants using multiple sequence alignments (MSAs) based on natural and engineered proteins in tandem with SPOT array data. By analyzing the peptides that emerge from this library via deep sequencing, we report many new mutations that were previously not reported nor predicted to govern specificity among Bcl-2 proteins. We then confirm the affinity and specificity properties of these Bcl-xL antagonists at the library level analyses both via surface display and in solution (following chemical synthesis), which further supports SPEED's robust predictive capabilities from **Chapter 2**. We describe the discover of two novel compounds, denoted **12** and **13**, which antagonize Bcl-xL with 100X specificity over Mcl-1 and 10nM affinity. Finally, we evaluate the mechanism of action of these peptides via BH3 profiling assays and confirm they act in accordance with apoptosis biology.

Towards the generation of stapled peptides specific for each of the 5 Bcl-2 proteins, **Chapter 4** describes a new approach involving machine learning to predict highly specific peptides. Because experimental data is laborious and costly to obtain, computational methods that accelerate and improve protein design are highly needed for protein engineering. They are particularly needed for multi-objective engineering, where complex trade-offs between important properties (such as binding affinity and specificity) limit the optimization of lead candidates. This chapter describes the generation of highly specific peptides for Mcl-1 over Bcl-x_L and vice versa, both of which are important drug targets in the resistance to apoptosis. We design Mcl-1 specific peptides with >10,000X specificity and Bcl-xL specific peptides with >100X specificity. To test the generalizability of this method, we also design Mcl-1 and Bcl-xL bispecific peptides, which could serve as potent therapeutics for personalized medicine, where a disease is characterized for its Bcl-2 dependency and specific drugs are administered for each patient.

Using this optimization method, we report several bispecific peptides that bind both targets with high affinity. Furthermore, it sheds light on the value of quantitative labels in sequence-fitness landscape and protein design. Surprisingly, binary information obtained from simple sorting experiments provides similar value to complex quantitative experimental methods. This chapter sets a precedent of design, where sequencing data with qualitative labels from sequencing can be transformed to high quality data for the design of high fitness stapled peptides for multiple parameters. Finally, it shows that with careful sequence optimization and adequate sampling of sequence-fitness space, an under-sampling of experimental design space coupled with machine learning can allow extrapolation into a much larger design space.

Future Work

There are several areas of this thesis that are subject to future work. From **Chapter 2** and **Chapter 3** and the work involving the design of pro-apoptotic stapled peptides, future work includes designing inhibitory stapled peptides for each of the Bcl-2 proteins: Bfl-1, Mcl-1, Bcl-xL, Bcl-2, and Bcl-w. Because the throughput of experimental work is far below the sequence space of all peptide variants, future work that expands the scope of design are likely to yield more potent peptide therapeutics. In this chapter, we showed that a library design informed by thousands of experimentally sequences yielded many mutations that improved specificity that were not predicted to; this suggests that expanding the design space of peptides measured (potentially coupled with techniques that allow interpolation among sequences, such as that detailed in **Chapter 4**) are promising future directions for protein design. With such focused computational library, sorting strategies for Bcl-2 proteins, and high-quality sequencing data for several of the targets, the addition of new members and fitness objectives should be relatively straightforward.

Furthermore, using integer linear programming (or other optimization tools), other drug-like properties such as protease stability and cell permeability can be explored while maintaining favorable affinity and specificity. These peptides would be similarly tested for mechanism of action in various cancer and engineered cell line models for activity before being translated for *in vivo* testing. Finally, the structure of the protein-peptide complex would yield key structural biology insights into how specificity was achieved. A molecular understanding of specificity may yield further insights into mutations or peptide motifs that improve drug-like properties (such as how a staple location and chemistry contributes to alpha helicity and stability). This structure could be explored using crystallography or nuclear magnetic resonance spectroscopy.

Another direction for future work is the more complete characterization and comparison of binary sorting and sequencing versus Sort-seq and Tite-seq experiments. In **Chapter 4**, machine learning predicted protein fitness with equal predictive capabilities when presented with inexpensive and readily available binary data versus more expensive/resource intensive continuous data. However, there may be certain protein engineering tasks that may be better suited for high-throughput quantitative cell sorting. At the trade-off between predictive power and model interpretability, datasets containing more epistasis (higher order interactions between mutations and resulting function) can be modeled with simple linear models but with each mutation acting non-linearly, the predicted fitness deviates further from actual fitness. Therefore, proteins displaying high levels of epistasis increasingly benefit from tools that leverage non-linear information, like neural network based models.

The field of protein engineering has seen a rapid adoption of techniques that leverage information from well folded proteins (such as those from organism genome databases); transfer learning of the motifs that govern stability and function is an exciting approach to *de novo*

protein design. In these cases, more accurate labels (as obtained via TiteSeq, for example) for protein fitness may be necessary to adequately sample sequence-fitness space. A complete description of when these more complex techniques yield improve modeling outcomes would save resources and accelerate protein engineering efforts. For example, binding affinity among a library of variants can span several orders of magnitude. If a specific affinity is needed, which is the case for some enzymatic or pharmacokinetic challenges, then having high resolution on intermediate values could yield improved variants. To test this, a library of variants spanning fitness of several orders of magnitude (such as an error prone PCR library of a monoclonal antibody) would be sorted quantitatively (TiteSeq) and compared with a standard sort (positive/ negative). Then, these data would be used to train models and measure the affinity of promising variants.

One question of particular importance is regarding the exploit-explore trade-off. Given the limitations of experimental protein science (such as 96 well plate assays and ~10^9 member directed evolution experiments), how sparse can a design space be sampled before machine learning/ deep learning approaches are unable to accurately interpolate towards variants with higher function? One simple experiment to probe this would be to design a large library with a significantly larger design space than what is measurable experimentally, sort the variants for increased fitness, then compare modeling techniques' ability to extrapolate successfully as you increasingly withhold data from model training. This analysis not only shows what modeling can do in the best case (maximal experimental information) but also shows how effective they are when data quantity is limited. To demonstrate generalizability, repeating this experiment for multiple unique proteins would be critical. In **Chapter 4**, we show that a 10,000X under sampling of design space provides sufficient accuracy for interpolation with relatively simple

machine learning techniques. It will be interesting to see how this ratio improves with increasingly powerful models and high-resolution protein property measurements. Because these tasks' performance is highly dependent on the class of protein, modeling details, among many other factors, another exciting aspect is how lessons about library design are translated between distant classes of proteins. Answers to these questions will dictate the scope of protein design experiments with available modeling tools.

In conclusion, the space of proteins that are undruggable by current therapeutic modalities continue to be challenging to target. Improvements described here represent powerful tools that can accelerate the development of novel therapeutic modalities. With continued investment in methods that improve 1) the ability to rapidly assay binding affinity and specificity 2) guide design towards drug-like properties that are heavily dependent on sequence such as protease susceptibility and cell penetration and 3) the ability to easily assay function in representative environments (such as reporter assays), these modalities will become more and more promising approaches towards the targeting of difficult-to-target disease-related proteins.