Modeling and Designing Bcl-2 Family Protein Interactions Using High-Throughput Interaction Data

by

Vincent Xue

Submitted to the Program of Computational and Systems Biology in partial fulfillment of the requirements for the degree of Doctor of Philosophy at the MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2018

© Massachusetts Institute of Technology 2018. All rights reserved.

Signature redacted

Author ............................................................
Program of Computational and Systems Biology
March 30, 2018

Signature redacted

Certified by ..............
/ Amy Keating
Professor of Biology and Biological Engineering
Thesis Supervisor

Signature redacted

Accepted by ..............
Christopher B. Burge
Director, Computational and Systems Biology Graduate Program

MASSACHUSETTS INSTITUTE OF TECHNOLOGY JUN 28 2018
LIBRARIES ARCHIVES
Abstract

Protein-protein interactions (PPIs) play a major role in cellular function, mediating signal processing and regulating enzymatic activity. Understanding how proteins interact is essential for predicting new binding partners and engineering new functions. Mutational analysis is one way to study the determinants of protein interaction. Traditionally, the biophysical study of protein interactions has been limited by the number of mutants that could be made and analyzed, but advances in high-throughput sequencing have enabled rapid assessment of thousands of variants. The Keating lab has developed an experimental protocol that can rank peptides based on their binding affinity for a designated receptor. This technique, called SORTCERY, takes advantage of cell sorting and deep-sequencing technologies to provide more binding data at a higher resolution than has previously been achievable. New computational methods are needed to process and analyze the high-throughput datasets. In this thesis, I show how experimental data from SORTCERY experiments can be processed, modeled, and used to design novel peptides with select specificity characteristics. I describe the computational pipeline that I developed to curate the data and regression models that I constructed from the data to relate protein sequence to binding. I applied models trained on experimental data sets to study the peptide-binding specificity landscape of the Bcl-xL, Mcl-1, and Bfl-1 anti-apoptotic proteins, and designed novel peptides that selectively bind tightly to only one of these receptors, or to a pre-specified combination of receptors. My thesis illustrates how data-driven models combined with high-throughput binding assays provide new opportunities for rational design.

Thesis Supervisor: Amy Keating
Title: Professor of Biology and Biological Engineering
Acknowledgments

First, and most importantly, I would like to thank my advisor Amy Keating. Amy has always kept my interests in mind, and I am very appreciative of her supportive mentorship. Despite coming in with minimal biochemistry background, Amy helped me find a place in the lab. I have learned a lot from her about rigorous science and storytelling which are lifelong lessons for the future.

I would also like to thank my committee members, Chris Burge and Ernest Fraenkel. Their valuable feedback on my thesis throughout the years have provided much insight to help refocus my project. I would also like to thank Sandor Vajda for being on my defense committee.

I want to thank Mandana Sasanfar for introducing me to the research at MIT. After meeting Mandana at the quantitative biology workshop, she invited me to the MIT Summer Undergraduate Research Program which kickstarted my research career.

I would like to thank many of the people who have worked with me in the Keating lab. I want to especially thank Justin Jenson for his collaboration on the projects discussed in this thesis. His expertise and adept experimental skills facilitated much of the data collection, and I am very thankful for his help bridging the experimental and computational work. I also want to thank Luther Reich for his early contributions and mentorship. I would additionally like to thank Lindsey Stretz for her contributions.

I would also like to thank post-docs Dustin Whitney and Vincent Frappier. Dustin helped me get through some of my most difficult times in my Ph.D. career. Vincent Frappier has been a source for many intellectual conversations, ranging in topics from, science, to finance, and life in general. I would also like to thank Chris Negron for his early guidance in and out of the lab.

I want to further thank my classmates Nezar Abdennur, Rotem Gura, Colette Picard, Amanda Kedaigle, Peter Freese, and Mariana Guadalupe Matus Garcia for their moral support. We have had many amazing times during our time here, celebrating marriage, birthdays, and camping. I also want to thank Jacque Carota, for
helping me through the jungle of administrative work and for always keeping an open door for conversation.

Lastly, I would like to thank my family. I am grateful to Dongkui Ma and Liyan Wang for motivating me and sending supportive words throughout my Ph.D. I also want to thank my parents for being supportive of my career. Finally, I am very grateful to my wife Wenji Ma for being patient, for supporting me in the final months of writing, and for cheering me on in life.
Contents

1 Introduction ........................................... 17
   1.1 Domain - peptide interactions ....................... 17
   1.2 Study of protein-protein interactions .................. 19
      1.2.1 Experimental methods ........................ 19
      1.2.2 Computational approaches ................... 25
      1.2.3 Application of modeling to design PPIs .......... 29
   1.3 Bcl-2 family proteins as a model system ........... 31
      1.3.1 Epistasis .................................. 33
      1.3.2 Specificity ................................ 34
   1.4 SORTCERY .......................................... 35
   1.5 Thesis content .................................... 36

2 Computational processing and modeling of SORTCERY data .... 39
   2.1 Introduction ...................................... 40
   2.2 Results ........................................... 41
      2.2.1 Ten high-quality peptide binding affinity datasets were generated using SORTCERY .......... 41
      2.2.2 SORTCERY data were used to identify epistatic mutational effects on binding .................. 50
      2.2.3 Modeling the relationship between sequence and binding free energy with SORTCERY data ...... 59
   2.3 Methods ............................................ 75
      2.3.1 Experimental data collection .................. 75

7
### List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-1</td>
<td>Examples of protein-protein interaction domains families</td>
<td>18</td>
</tr>
<tr>
<td>1-2</td>
<td>Structure of a BH3 peptide</td>
<td>32</td>
</tr>
<tr>
<td>1-3</td>
<td>Overview of the SORTCERY protocol</td>
<td>35</td>
</tr>
<tr>
<td>2-1</td>
<td>Correlation of the SORTCERY mean affinity coordinate to binding energy for standards</td>
<td>45</td>
</tr>
<tr>
<td>2-2</td>
<td>Sequence logos for peptides curated from SORTCERY</td>
<td>48</td>
</tr>
<tr>
<td>2-3</td>
<td>Boxplots of $\Delta \Delta G$ measurements for single point mutations</td>
<td>51</td>
</tr>
<tr>
<td>2-4</td>
<td>Distribution of $</td>
<td>\Delta \Delta G</td>
</tr>
<tr>
<td>2-5</td>
<td>Double mutant cycle peptide profiles across SORTCERY gates</td>
<td>55</td>
</tr>
<tr>
<td>2-6</td>
<td>Double mutant cycle C</td>
<td>56</td>
</tr>
<tr>
<td>2-7</td>
<td>Schematic of the protocol used to train receptor-specific models</td>
<td>58</td>
</tr>
<tr>
<td>2-8</td>
<td>Nested cross-validation performance - amino acid encoding</td>
<td>60</td>
</tr>
<tr>
<td>2-9</td>
<td>Nested cross-validation performance - five-factor encoding</td>
<td>61</td>
</tr>
<tr>
<td>2-10</td>
<td>Nested cross-validation performance - eleven-factor encoding</td>
<td>62</td>
</tr>
<tr>
<td>2-11</td>
<td>Correlation of SORTCERY replicates</td>
<td>63</td>
</tr>
<tr>
<td>2-12</td>
<td>Alternative modeling benchmarks on SORTCERY data</td>
<td>67</td>
</tr>
<tr>
<td>2-13</td>
<td>Replicate holdout tests</td>
<td>69</td>
</tr>
<tr>
<td>2-14</td>
<td>Model extrapolation performance on Bcl-2 protein-protein interaction datasets</td>
<td>70</td>
</tr>
<tr>
<td>2-15</td>
<td>Extrapolation across affinities</td>
<td>76</td>
</tr>
<tr>
<td>2-16</td>
<td>Removing DNA artifacts</td>
<td>80</td>
</tr>
<tr>
<td>2-17</td>
<td>Nested Cross-Validation</td>
<td>85</td>
</tr>
</tbody>
</table>
Yeast titration of peptide standards used to calibrate Mcl-1 SORTCERY experiments
List of Tables

1.1 Peptide recognition motifs from the Eukaryotic Linear Motif resource 19
1.2 Datasets collected for peptide recognition domain interactions .... 25

2.1 Residues mutated in the pilot library. ................................. 42
2.2 Residues mutated in the specificity library .......................... 43
2.3 Sequences selected as standards for mapping SORTCERY to energy 44
2.4 Summary of datasets collected from SORTCERY .................. 49
2.5 Double mutant cycles identified as having large $|\Delta \Delta G|$ from SORTCERY 52
2.6 $K_d$ values for peptides selected for verification in the double mutant cycle analysis. 53
2.7 Evaluation of model performance to score sequences from a different experiment .............................. 68
2.8 Affinity extrapolation data ........................................ 71
2.9 Categorical extrapolation data ...................................... 72
2.10 SVR parameters ................................................. 86

3.1 Monospecific peptides designed ..................................... 113
3.2 Bispecific peptides designed ....................................... 114
Chapter 1

Introduction

Protein-protein interactions (PPIs) play an essential role in all biological processes, including metabolism, signaling, and establishing and controlling cellular and tissue structure. Knowledge of which proteins interact can reveal the underlying molecular pathways that lead to function and malfunction (disease) \[75\] \[13\]. By learning the patterns that dictate binding, we can predict new functions and apply rational engineering of protein interactions to develop therapeutics and diagnostics \[57\].

1.1 Domain - peptide interactions

Many PPIs occur between folded structured domains of two proteins that interlock with each other, forming a complex. However, an estimated 15-40% of PPIs occur between a structured domain of one protein and a short linear peptide region of another \[78\]. In most peptide-domain interactions, the short linear peptide lacks secondary structure when unbound, and structure is induced upon complex formation. In contrast, binding of a peptide to a peptide recognition domain (PRD) typically does not significantly alter the PRD structure \[69\]. The peptide binding preferences of a PRD are often represented as a motif summarizing the residue/site-specific patterns that are important for binding. A recent analysis showed that 22% of human disease mutations occur in unstructured protein regions, many of which disrupt important domain-peptide interactions \[106\].
Figure 1-1: Examples of protein-protein interaction domains families where the domain fold is conserved, but the binding specificity can vary among family members. A representative complex is shown for each. (A) A complex between PDZ domain (cyan) peptide (red) (PDB ID: 1MFG) (B) A complex between SH3 domain (cyan) and peptide (red) (PDB: 1ABO) (C) A complex between Bcl-2 domain (cyan) and peptide (red) (PDB ID: 3FDL) (D) A complex between SH2 domain (cyan) and peptide (red) (PDB ID: 1D4W).
Table 1.1: Peptide recognition motifs from the Eukaryotic Linear Motif resource

PRD genes, through mechanisms of duplication and divergence, can form copies that gain alternative functions. Evolutionarily related PRDs often bind to similar motifs but can have distinct and unique binding preferences for their interaction partners. Some well-studied PRD families and their binding motifs are presented in Figure 1-1 and Table 1.1. There are about 200 known PRD families [49].

The function of a PRD family member is determined by its binding profile. Identifying the peptide binding partners of a protein domain, as well as the affinities with which different complexes form, can help identify key regulators and/or therapeutic targets. Thus, elucidating key specificity determinants of binding is important for understanding the molecular logic of signaling. This motivates the need to identify new PPIs and to develop predictive models of how sequence determines protein-protein binding for different families of proteins.

Given that some protein families are large, and that different members can have distinct functions, understanding what defines domain specificity within a family is a key question, sometimes known as "the paralog problem".

1.2 Study of protein-protein interactions

1.2.1 Experimental methods

Decades of research have focused on answering the questions of which proteins bind one another, what features determine how tightly they interact ($\Delta G$ of binding), and how selective interactions are encoded. Much progress has been made using experimental assays. Many protein interactions have been discovered using low-throughput
cell biological techniques such as pull-down assays. High-throughput methods, discussed below, have also been developed for screening at the scale of small proteomes. To dissect the roles of different residues in modulating interaction affinity, it is necessary to perform mutational analyses. In such experiments, interactions are typically quantified by genetic construction, expression, purification, and physical analysis using low-throughput methods, and these one-at-a-time measurements are slow. Higher throughput methods provide greater amounts of less quantitative, and often less reliable, interaction data. Sections below summarize some common experimental approaches and the types of data and insights that can be obtained using them. This background provides context for our work on SORTCERY, a high-throughput technique for quantifying protein-protein interactions.

**Low throughput experiments**

**Site-directed mutagenesis** Mutagenesis approaches such as alanine scanning involve preparing and individually assaying a set of proteins that differ by just a single mutation. Such techniques have been used to identify important protein "hotspots" (residues that are especially important for binding) and to understand how well different sites tolerate mutation. For example, Cunningham and Wells applied alanine scanning to identify specific side chains in human growth hormone (hGH) that modulate binding to the hGH receptor [23]. Although site-directed mutagenesis can provide detailed information about the contributions of residues at specific sites, standard biochemical techniques have limited throughput (up to hundreds of interactions), and cannot systematically test the effects of multiple simultaneous mutations. Such methods are therefore not ideal for testing all possible substitutions, or for discovering epistatic effects, which have been shown to be important for many evolved systems. It is important to understand more than just the effect of single point mutations [9].

**High throughput experiments**

High-throughput experimental methods can be used to obtain information about peptide binding preferences and motifs. In contrast to one-change-at-a-time mutagenesis
methods, most high-throughput methods trade off quantitation of measurement in exchange for a boost in throughput. A few examples of high-throughput assays are array methods, yeast-two-hybrid, affinity purification-mass spectrometry, and chromatographic holdup assays.

**Array Methods** Peptide microarrays are made by printing small amounts of protein onto glass slides, where interactions are screened by probe molecules, typically labeled with a fluorescent dye. Peptide microarray technologies are advantageous when there is a list of candidate sequences to test. Ultra-dense peptide arrays can include up to $10^6$ synthesized peptides, and have been used for epitope mapping of antibodies [16]. However, peptide microarrays are generally expensive and, although some studies report correlation of signal with binding affinity, the data obtained is qualitative. Furthermore, synthesis of different peptides has variable yields that result in false positives and false negatives [8]. Some example datasets describe the selectivity of 157 mouse PDZ domains binding to 217 genome-encoded peptides, and the binding of three PDZ domains to peptide libraries comprising 6223 human protein C termini [100] [115]. Similar to microarray approaches are SPOT arrays. Peptides with a known sequence are chemically synthesized on a membrane and incubated with the target protein. Binding of the target protein is detected using specific antibodies, or fluorescent or radioactive labeled proteins that give a signal that is correlated to the binding affinity [112]. Site-saturation mutagenesis approaches via SPOT arrays have been applied to identify residues that are important for peptide binding to Bcl-2 family proteins [29].

**Yeast two-hybrid** Yeast two-hybrid (Y2H) assays are based on splitting the activation domain and the binding domain of a transcription factor. Each component is linked to a bait or prey protein, such that when bait and prey proteins interact, the reconstituted transcription factor activates a reporter signal. Y2H has been used to map ~14,000 distinct interacting protein pairs in the human proteome [85]. Although these assays can provide information about which proteins interact, they do not pro-
vide quantitative affinities. In addition, such assays have high false positive and false negative rates, as high as 70% [24]. Yeast two-hybrid also generates a positive signal for indirect interactions, and does not report on which part of the protein mediates the interaction. An example dataset generated by yeast two-hybrid screening is the work of Belotti et al., in which an array of 246 human PDZ domains was tested for interaction with C-terminal segments from many other proteins [6].

**Affinity Purification Mass Spectrometry (AP-MS)**  
Affinity Purification-Mass Spectrometry (AP-MS) identifies protein complexes through binding-based enrichment of interaction partners in the lysate. Members of purified complexes are identified by direct analysis via mass spectrometry. The target protein (or bait) is coupled to a solid support and exposed to potential binders. Non-specific binders are washed off, and the true binders are eluted, proteolyzed, and identified via comparison to theoretical fragmentation patterns. This assay reveals the binary interaction of proteins with their target. One example application of AP-MS is the work of Gavin et al., who characterized 232 distinct multiprotein complexes in the *Saccharomyces cerevisiae* proteome [42]. AP-MS is very efficient for high-throughput discovery. However, AP-MS is designed to detect protein complexes; the method does not assay for direct interactions between individual proteins. Also, many interactions are missed because complexes must survive the washing steps that are required to remove non-specific binders. In addition, a limiting step of AP-MS is the need to purify the protein complex prior to mass spectrometry.

**Chromatographic-holdup assay**  
A more recently developed high-throughput affinity assay is the chromatographic-holdup assay, which is capable of measuring thousands of peptide interaction affinities [111]. Soluble extracts of overexpressed domains are incubated with ligand saturated resin. Flow-throughs of unbound domain proteins are assessed via microfluidic capillary electrophoresis, where the signal of the unbound domain gives a readout of binding affinity. Although holdup assays potentially have high-throughput affinity readouts, the assay is costly, and the expression
and purification of the domains and peptides to be tested are slow. One application of this assay generated affinities for 209 PDZ domains binding to two human papillomavirus E6 PDZ-binding peptides [111].

Library screening experiments

In recent years, the increasing availability and decreasing costs of deep sequencing technologies have allowed for the development of new screens for assessing the activity of millions of mutant variants in a single experiment. Highly influenced by SELEX-seq, which assays ligand-oligonucleotide (single-stranded DNA or RNA) interactions through rounds of selection and amplification, new ligand-domain interaction assays apply similar selection steps to assay interactions by screening cell surface displayed proteins [41].

Display technologies comprise different methods for linking the phenotype of a protein to its sequence. DNA that encodes a library of designed peptides can be inserted into phage, bacteria, or yeast, such that the peptides are expressed and displayed on the vector's surface. To assay binding, the vectors are incubated with a target protein of interest. Yeast and bacteria displaying peptides can be assayed for binding using fluorescence activated cell sorting (FACS). Vectors that stick to the target are subjected to DNA sequencing to map binding phenotype to genotype. This approach can provide long lists of binding proteins, but not quantitative affinity information. It is inexpensive to test many variants in an unbiased manner using display technologies [87]. Yeast display can present libraries up to $10^8$ whereas phage display reaches libraries up to $10^{12}$ unique variants. One application of surface displayed selection is in the work of Tonikian et al., in which the binding specificity for over 330 PDZ domains were mapped via phage display [105]. Table 1.2 provides some example datasets that have been collected for domain-peptide interactions.

Enrichment of binders over multiple rounds

Binders of different affinities can be distinguished in library experiments if peptide-displaying vectors are subjected to a series of binding and wash steps to an immobi-
lized target. The frequency of high affinity binders that pass the selection will increase faster than that of low affinity binders in such an experiment, such that high-affinity binders are enriched in the population. Enrichment values can be calculated as a function of the frequency of each variant before and after selection. In theory, enrichment values derived from sorting can report on binding affinity, however, enrichment assays can be highly sensitive to sampling error when frequencies are low, and the range of affinities that can be resolved is limited [114] [39] [114] [72] [104].

Fowler et al. used enrichment methods to characterize the interaction of variants of the WW domain of the human Yes-associated protein 65 (hYAP65) with a peptide ligand, using phage display selection of a randomly mutated WW domain library [38]. More recently Foight and Keating applied bacterial-surface display and enrichment analysis to study peptide binders of TRAF2, TRAF3 and TRAF5 [36].

Precise affinity display based technologies

To provide a quantitative, high-throughput experimental assay to investigate protein-ligand binding, the Keating lab pioneered a high-resolution protocol called SORTCERY. This assay ranks thousands of peptide sequences according to their relative affinities for a given receptor [81]. Yeast-displayed peptide libraries are sorted into gates based on their binding affinity to a given receptor, computationally decoded, and mapped to sequence standards to yield binding free energies for thousands of peptide-receptor interactions (A more detailed description of SORTCERY is provided below). Similar to SORTCERY is Tite-Seq, which uses FACS sorting of library members at multiple protein concentrations to quantify interactions in high throughput. This method was used to investigate scFV variant binding affinities for fluorescein [1]. One disadvantage of these precise affinity-based assays is that the diversity of the input library is reduced to several thousand unique sequences, because of the requirement to have enough read coverage to generate sufficient statistics. In exchange, high-resolution affinity data is provided, which is useful for model building [65].
<table>
<thead>
<tr>
<th># Interactions</th>
<th>Domains/Description</th>
<th>Authors</th>
<th>Assay type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1301</td>
<td>Screened 217 peptides against 157 PDZ domains</td>
<td>Stiffler et al. [100]</td>
<td>Microarrays</td>
</tr>
<tr>
<td>&gt;1154</td>
<td>Screened 1930 ligands against 65 WW domains</td>
<td>Hu et al. [49]</td>
<td>Microarray</td>
</tr>
<tr>
<td>&gt;3,100</td>
<td>Screened 145 PDZ domain against random peptides libraries</td>
<td>Tonikian et al. [105]</td>
<td>Phage display</td>
</tr>
<tr>
<td>&gt;1871</td>
<td>Screened 25 SH3 domains against random libraries</td>
<td>Tonikian et al. [104]</td>
<td>Phage display</td>
</tr>
<tr>
<td>394</td>
<td>Screened 28 SH3 domains against random libraries</td>
<td>Tong et al. [103]</td>
<td>Phage display</td>
</tr>
<tr>
<td>1026</td>
<td>Ranked 1026 peptides for binding to Bcl-xL</td>
<td>Reich et al. [81]</td>
<td>Yeast Display</td>
</tr>
<tr>
<td>&gt;3000</td>
<td>Measured surface displayed scFv antibody for binding to fluorescein</td>
<td>Adams et al. [1]</td>
<td>Yeast Display</td>
</tr>
</tbody>
</table>

Table 1.2: Datasets collected for peptide recognition domain interactions

### 1.2.2 Computational approaches

Experimental observations have provided increasing amounts of data to help study protein interactions, but the sequence space of potential interactions is too large to be explored completely. For example, the sequence space of a peptide increases with its length such that for a peptide of length 22, there are more than $10^{28}$ unique sequences. To facilitate exploration of this space, computational methods are increasingly used to direct experimental searches to find or design new interactions [54] [66] [65]. Sections below discuss some of the computational approaches to scoring the interaction between proteins.

Methods can generally be classified as physics-based structural, statistics-based structural, or empirical modeling approaches.

#### Physics-based structural modeling

Physics-based models of PPIs are based on fundamental analysis of the forces between particles. Physical aspects of protein binding are computed from the structure, accounting for contributions from packing, electrostatics, hydrogen bonding, solvation, van der Waals interactions, and Coulomb electrostatics [49]. Physics-based models are potentially broadly generalizable and applicable to any protein, however, their accuracy is limited by the many approximations made in scoring. In addition, these
physically detailed models tend to be slow to evaluate, due to the complexity and number of calculations involved. The CHARMM energy function is an example of a physical model [12] that has been applied to redesign bacterial periplasmic binding proteins [70].

Statistics-based structural models

Statistics-based PPI models originate from quantitative analysis of protein structures from the Protein Data Bank (PDB). Early modeling methods proposed more than 40 years ago quantified the interaction buried surface area to predict binding affinity [21] [62] [48]. Other early approaches modeled properties such as solvent accessible surface area, hydrogen bond contacts, and hydrophobicity, that have been shown to affect affinity [53].

As more protein structure complexes were solved, models were developed that quantified the contributions of different structural elements and calculated statistics derived from the PDB. Popular structure-based models, such as ROSETTA and FoldX were developed that mostly apply empirically derived energy functions to score a given complex [61] [92]. As the PDB continues to grow, new knowledge-based approaches continue to be developed. Examples include STATIUM from the Keating lab and TERM scoring functions from the Grigoryan lab, which capture statistics on structural motifs in the PDB [26] [119].

Structure-based methods have the advantage of being generalizable. However, they require a starting structure, can be inaccurate, and can be very computationally intensive (ranging from minutes to hours to score each variant), which makes these methods unsuitable for large screens in the nearly infinite space of possible binders [79].

Empirical modeling and machine learning

Empirical models are built from experimental data. These models often require significant upfront experimental work to collect the data but can offer powerful domain specific models for scoring new sequences. Early approaches built position-specific scoring
matrices (PSSMs) or Hidden Markov Models (HMMs) from multiple sequence alignments of known interaction partners to score the fitness of unobserved motifs [101]. Empirical models are useful because they do not require a physical understanding of the binding behavior to achieve accurate prediction. In addition, the models are typically faster to evaluate and more accurate than other modeling approaches. However, empirically based models cannot predict new behavior and are specific to the system being studied. For example, models which learn sequence-to-function relationships have poor ability to generalize to residues unobserved in a given training set, unless assumptions about residue similarities are included in the model. Ideally, a general model is desirable to make accurate predictions for any protein complex. However, a more realistic alternative is to construct family-specific empirical binding models [71].

As quantitative PPI data for domains-peptide interactions become more available, the wealth of sequence-to-affinity data allows for the development of domain-specific computational models that provide insight into the determinants of binding affinity. Better statistics allow for novel analysis. One example is in the work of Gfeller et al. in which clustering analysis was applied on data from Tonikian et al. to identify modular peptide recognition motifs in the PDZ family [105][43].

Larger amounts of experimental data create a need for novel in-silico methods that can process and exploit the rich data sets. To respond to this need, machine learning (ML) models have been applied to process the larger amounts of data. I discuss several example models in the literature broadly separated into classification and regression models.

Classification methods for studying PPIs typically report on categorical binding vs. non-binding. A common ML classification method is the support vector machine (SVM), which categorizes an input into two populations via a dividing hyperplane [118]. Hui et al. applied SVMs to learn from PDZ domain interaction data and used their models to score the human proteome to identify new PDZ binding peptides [50]. Important to their analysis was the inclusion of pseudo-negative hits, which improved their model performance to discriminate true binders.
Regression models are powerful statistical tools used to analyze the relationship of a set of input parameters to a numerical response variable. Shao et al. modeled PDZ domain-peptide interactions by applying Semi-Support Vector Regression to learn from both resolvable and non-resolvable affinity data collected by Stiffler et al. [93]. Artificial neural networks (ANNs) are computing systems inspired by biological neural networks. Usually represented as an interconnected group of nodes, these self-learning methods carry high complexity and have the capability to learn from massively large datasets. Neural networks have been applied in data-rich domains such as that of predicting MHC class II-binding peptides [15]. Other quantitative models include Gaussian processes, which treat the input variables as a collection of normal distributions. In work by Romero et al., sequences were mapped to unique residue contact maps to model the thermostability of chimeric cytochrome P450s [86].

ML can be applied to many problems and benefits from large amounts of data. However, models trained on data are prone to overfitting. Furthermore, due to the flexibility that ML models have to represent a problem, these models tend to lack direct biological meaning, which is undesirable for design and interpretation.

Although novel experimental methods can assay thousands of PPI, these experiments also have higher false-positive and false-negative rates. Thus, it is necessary to develop methods to discern signal from experimental noise [11] [110]. Data-driven approaches have difficulty extrapolating - thus it is necessary to include additional priors. One solution to address this is demonstrated by Chen et al., who applied Bayesian methods to combine multiple sources of data into coherent models to achieve greater prediction accuracy [20]. Integration of other experimental data can also help provide priors for modeling. For example, the conservation of sequences in genomic context has been used to predict binding partners, and sequence coevolution has provided insight into 3D structure contact prediction [51] [5] [73].
1.2.3 Application of modeling to design PPIs

Challenges/objectives

Two common objectives in protein design are to improve interaction affinity or increase selectivity. Improving affinity or selectivity can provide benefits for therapeutic and engineering applications. Computationally, these objectives can be achieved by optimizing scoring functions to either minimize the binding energy of the binding complex or to maximize the difference in binding energy of a target and off-target. Guided by a scoring function, the next step in protein design to explore the large sequence space.

Due to the time and cost required to perform binding experiments, in-silico models are pursued to narrow down the sequence space of new mutations to evaluate [71]. Despite much progress, in-silico design remains a difficult task. Design with physical, structural, or empirical models each face similar challenges navigating sequence and structural space.

One of the difficulties of studying binding interactions with physical and structural models is the infinite complexity in conformational space. Evaluating a sequence requires solving for the optimal structure which can be very time-consuming to compute. Often we can reduce the search space to rotamers (side-chain torsion combinations corresponding to the local minima of potential energy) by assuming that a binding complex maintains a fixed-backbone. Searching mostly fixed backbone structures makes the design task more feasible, but these methods still tend to be slower and face the second challenge of navigating the sequence space.

Empirical models for design can ignore the structural search space but must explore the near infinite sequence space. Although empirical models are typically faster, scoring the entire search space is still infeasible. Monte Carlo sampling can explore a local space, but has difficulty finding global optima. Models that have parameterized interpretable features are easier to diagnose when models fail. In addition, the patterns learned from such models are more readily transferable to solve other problems. One limitation of design with empirical models is that they tend to explore small
sequence spaces and lack information about changes to the backbone. Even a small number of mutations can lead to significant variation in the backbone geometry [19].

**Existing computational design**

Recent advances in computational design have yielded successes such as altered binding preferences in bZips [45], DNase-inhibitor protein pairs [59] and PDZ domains [82]. Negron et al. designed antiparallel homodimeric coiled coils using Dfire combined with the multistate protein design framework CLASSY to search and score proteins [77] [33] [76]. Sammond et al. applied Rosetta to redesign native protein interfaces for altered affinity and specificity [89]. These computational approaches rely on having accurate scoring functions and have to overcome the complex challenge of exploring large sequence space. One approach is to use deterministic algorithms (e.g. integer linear programming) to optimize an objective function, such as in the work of Potapov et al. to design novel synthetic coiled coils [80]. Others groups have applied more stochastic approaches (e.g. Monte Carlo, FASTER) for optimization in structure and sequence space [3]. Although there has been some success, modeling still requires extensive manual input and tweaking expertise. Computational predictions still have high rates of failure, and some of the best efforts for design have a low success rate. For example, Tinberg et al. identified (2/17) successful binders to the steroid digoxigenin using Rosetta [102]. Nevertheless, computational modeling has the potential to narrow leads in the exponentially large search space.

An exciting future for computational modeling exists as new high-throughput experimental technologies emerge. The direct collection of sequence-to-function data via directed evolution and functional selection experiments gives rise to new datasets that can supplement existing computational models and improve precision. With this data, we can begin to answer questions that can only be explained by significant statistical evidence. However, with the new yield of high-throughput data come further questions about how to process the data and how this will change current modeling assumptions.
1.3 Bcl-2 family proteins as a model system

Similar to the highly studied PDZ/WW/SH2/SH3 family proteins, Bcl-2 family proteins mediate domain-peptide PPI. This family of proteins functions at the last step before intrinsic apoptosis, where interactions between structured anti-apoptotic family members and α-helical BH3 peptides mediate the downstream release of Cytochrome C. A few examples of anti-apoptotic Bcl-2 proteins include Bcl-2, Bcl-xL, Bcl-w, Bfl-1, and Mcl-1.

Typically found in unstructured regions of BH3-only proteins, BH3 peptides take on a helical structure when bound to the anti-apoptotic proteins (Figure 1-2). Two examples of BH3-only proteins are Bim and Puma. Mis-regulation of Bcl-2 family member interactions can lead to cancer, autoimmune, and neurodegenerative diseases, making this family of proteins an important target for study [117].

Anti-apoptotic protein members of the Bcl-2 family are important therapeutic targets due to their overexpression in cancer where they contribute to oncogenesis and resistance to chemotherapy. Accurate computational tools can aid the design of tight and specific peptidic binders to restore apoptotic function.

A number of past efforts to characterize Bcl-2 family binding preferences have resulted in several large datasets. Dutta et al. assayed Bim BH3 variants for binding to Mcl-1 and Bcl-xL via SPOT arrays [29]. Subsequent studies applied similar experimental assays to investigate the homologs Bfl-1, Bcl-2, and Bcl-w [25] [68]. Hydrophile scans from Boersma et al. systematically explored alanine, lysine, and glutamic acid substitutions to learn about the residues that contribute to selectivity [10].

Larger combinatorial library experiments were used to assay binding to Bfl-1, Bcl-xL, and Bfl-1. Lee et al. screened a phage display peptide library of 16mers for Mcl-1 binding [63]. Dutta et al. screened a library of BH3 peptides by yeast display for binding to Mcl-1 or Bcl-xL to investigate the determinants of specificity [29] [28]. Foight et al. screened peptide libraries for binding to viral Bcl-2 homolog Kaposi sarcoma herpesvirus Bcl-2 (KSBcl-2), and identified Mcl-1 specific inhibitors
Figure 1-2: (Top) Structure of a BH3 peptide (PDB ID: 3FDL) highlighting 6 of the positions mutated for one of the libraries used in this study. (Bottom) A multiple sequence alignment of 6 native BH3 sequences. The heptad convention, used to refer to positions in the BH3 peptide, is shown above the alignment. The conserved Leu and Asp at the highly conserved positions 3a and 3f are indicated in bold.
More recently, Reich et al. screened libraries of BH3 mutants of Bim and Puma to rank 1026 BH3 peptides for their affinity to the Bcl-xL receptor [81]. Jenson et al. screened pooled, targeted libraries to identify Bfl-1 specific peptides [52].

In the Keating lab, we developed an experimental protocol that can rank peptides for their binding affinity to a designated receptor. This assay, called SORTCERY, is applied to this work to provide more binding data at a higher resolution than has previously been achievable [81]. This thesis is a prototype case where we use these large sets of sequence-to-affinity data to understand binding affinity of ligands to the Bcl-2 family proteins. I investigate interpretable modeling approaches that capture cooperativity, and design novel peptides for specificity.

1.3.1 Epistasis

The question of whether multiple mutations make additive or non-additive contributions to affinity has important implications for modeling. Simple models like PSSMs assume that binding occurs in a fixed structure where each residue contribution is independent. Mutational experiments have demonstrated that in many cases, the energetic contributions of a combination of mutations exhibits simple additivity [113]. However, due to long-range interactions, spatial clashes, and other effects, naive models of independence are not sufficient to explain all of the variability in binding affinity. Studies on the effect of multiple mutations have identified that there generally exist non-additive small magnitude effects [44] [64].

Identifying when multiple mutations sum additively is important to understand protein function and for practical engineering. The identification of cooperativity is desirable to enhance our understanding of the mechanisms of binding. Given the new depth of quantitative data, we are interested as to what this new depth of data can tell us about the effect of multiple mutations.
1.3.2 Specificity

Although BH3-only proteins share similar sequence motifs, they do not all bind to the same targets. Of three native BH3 peptides (Bim, Bad and Noxa), Bim binds to Bcl-2, Bcl-w, Bcl-x<sub>L</sub>, Mcl-1, and Bfl-1; Bad preferentially binds to Bcl-2, Bcl-x<sub>L</sub> and Bcl-w; and Noxa binds Mcl-1 and Bfl-1. Developing specific binders for Bcl-2 family proteins has important therapeutic application and is an important topic to pursue. For example, understanding specificity differences is important for developing therapeutics. A small molecule inhibitor ABT-263 was found to function well to inhibit Bcl-2 as a potential therapeutic against acute lymphoblastic leukemia. However, Bcl-x<sub>L</sub> off-target effects led to dose-limiting thrombocytopenia [17].

We only have a handful of selective BH3s that has taken years to generate. Most of the identified selective BH3s were found by rational design and screening of combinatorial libraries. For example, Dutta et al. designed and screened a library enriched in Bcl-x<sub>L</sub> specific BH3s to identify a Bim variant with a 1000-fold preference for Bcl-x<sub>L</sub> over other pro-survival proteins [29]. More recently, Foight et al. and Jenson et al. designed and screened rationally designed libraries to identify tight and selective peptides for Mcl-1 and Bfl-1 specific peptides, respectively [37] [52]. Other approaches for designing specificity have applied Rosetta and multiple rounds of site-directed saturation mutagenesis to take advantage of receptor-specific contacts outside of the BH3 binding groove. Berger et al. designed de novo protein inhibitors specific for each of the Bcl-2 family proteins based on the BINDI three-helix bundle [7].

The increasing availability of new high throughput data for multiple receptors will help us design more accurate functions and identify receptor specific motifs. Proteins of paralogous families are particularly interesting because individual family members may have distinct binding specificities, which may be critical to understanding, or inhibiting, function. Most existing approaches that study specificity have focused on motif approaches [55]. Studies of domain-peptide binding specificity have generally been limited to a few binding observations per domain. Existing PDZ and SH2/SH3 domains datasets are limited where there exists a few interactions observed per family.
Figure 1-3: Overview of the SORTCERY protocol. (A) A Yeast displayed library of peptides is sorted by two-color FACS into 12 gates that correspond to different ranges of affinities. (B) DNA encoding the surface-displayed peptides from yeast cells in each gate are deep sequenced. (C) Frequency distributions over FACS gates are constructed using the sequencing data. (D) Standards are chosen and individually measured for affinity. (E) Affinities for peptide standards are used to calibrate the mean affinity coordinates into units of kcal/mol.

1.4 SORTCERY

SORTCERY is a high-throughput assay for measuring protein interaction affinity (see Figure 1-3). It combines Fluorescence Activated Cell Sorting (FACS), and high throughput sequencing of yeast displayed peptides to measure thousands of protein member. This limits the amount of data available for models to learn and extrapolate within a given domain [18] [100] [90] [34]. SORTCERY gives us thousands of hits for a single domain, which is the next step required for more accurate specificity models.
interaction affinities simultaneously in a single assay.

Yeast Surface display (YSD) can be constructed with combinatorial protein libraries to screen thousands of proteins at once. YSD utilizes a two-color fluorophore labeling system, where one fluorophore reports on the expression of peptides on the cell surface, and the second reports on the signal from surface bound receptors. Visualized on log-log plots of binding vs. expression, each clonal population falls in 2d Gaussian-like profiles where tight binding peptides have a higher binding signal (blue), and weaker peptides have a lower binding signal (green).

SORTCERY utilizes this shift in binding signal to sort libraries of cells into populations of similar affinity. Theory predicts that FACS profiles are linear and exhibit parallel shifts to each other along an axis of affinity [81]. 12 gates are drawn orthogonal to this axis of affinity to capture the distribution of cells. This places high-affinity binders in the upper gates, and weak binders in the lower gates. After collecting and binning cells, the sequences of the displayed peptides are decoded via deep sequencing, and computationally processed to estimate the probability distribution profile of the clonal population across the 12 gates.

Given that SORTCERY relies on accurately capturing the probability distribution of clones, the assay limits the maximum diversity to peptides observed in a ~10,000 cell presort to have enough coverage to reconstruct the PDFs. Given that the PDFs are constructed from the distribution of reads across the gates, as deep sequencing coverage improves with advances in technology, such growth will allow us to explore larger libraries.

1.5 Thesis content

In Chapter 2 I show how raw SORTCERY output can be processed, with the use of experimental standards, to generate quantitative measurements of affinities for thousands of cell-surface interactions. I use the resulting data to build regression models that relate protein sequence to binding, and I compare the performance of these models to other prediction methods. I present evidence supporting the existence of higher
order cooperativity between peptide residues, which helps explain why polynomial models show better performance than first-order models. As a test of the utility of the regression models, I show that even where we have limited information, we can extrapolate into a larger sequence space than was sampled directly. The simple kernels employed in my models allow me to determine which residues and residue combinations contribute to binding affinity. Such interpretable models can support rational design of specific and tight binding peptides. Computational models can also help accelerate experimental peptide screening, because accurate scoring functions can be applied to reduce the library search space for identification of tight and specific binders.

In Chapter 3, I describe how I applied the curated data and models developed in Chapter 2 to study the specificity landscape of the Bcl-2 family proteins Bcl-xL, Mcl-1, and Bfl-1. I compare peptide binding preferences for all three receptors and map the specificity landscape of peptides assayed by SORTCERY. Using the regression models described in Chapter 2, I further demonstrate the potential expansion of the specificity boundaries by extrapolating into a novel sequence space composed of a mixture of residues from 6 distinct libraries. To validate the model's utility for studying specificity in this novel sequence space, I designed peptides with specific receptor binding characteristics. I extracted residue-pair weights from the polynomial regression models and used them to design peptides with integer linear programming. I showed that by incorporating residues that were sampled in different libraries into a single model, it is possible to design novel sequences that demonstrate specificity for only one receptor or a combination of receptors.
Chapter 2

Computational processing and modeling of SORTCERY data

Collaborator notes:

L. Reich collected the SORTCERY data for the pilot library

J. Jenson collected the SORTCERY data for the specificity library

L. Stretz performed the titration measurements for the peptide standards and prepared the DNA for deep sequencing
2.1 Introduction

Characterizing protein-protein interactions has traditionally been a slow process because interactions are measured in low-throughput. Higher throughput assays like SPOT arrays and yeast-two-hybrid screens are used to test thousands of peptide interactions, but these approaches are often error-prone and provide semi-quantitative data [8] [95]. With the recent decreases in cost of next-generation DNA sequencing, and the consequent increases in accessibility, a new class of high-throughput methods is emerging for quantitatively and accurately characterizing thousands of protein interactions [87] [19].

One approach to investigating protein interactions in high-throughput is through fluorescence activated cell sorting (FACS) of cell surface displayed proteins. Combinatorial libraries of protein mutants displayed on cell surfaces can access up to $10^{10}$ variants. While it has long been possible to identify the best binders in such a library through rounds of screening with increasing stringency [40] [4], only recently has it become possible to adapt such screens to provide information about the binding behavior of thousands of library members [41]. Such screens have been used to understand well-characterized PDZ and SH3 domains in greater depth, and large datasets about the relative affinities and specificities of these domains for their binding partners have been generated [105] [103] [100] [72]. These datasets and assays are useful but mostly qualitative.

The Keating lab has pioneered a novel high-throughput combinatorial library assay to quantitatively measure the binding affinity of domain-peptide interactions. This assay called SORTCERY reports interaction energies for a library of yeast-displayed peptides binding to a given receptor (see Introduction). We applied SORTCERY to investigate the binding energies of short peptides to members of the Bcl-2 family of proteins. Bcl-2 family proteins regulate apoptosis. High-affinity binders for the Bcl-2 family are of interest because of their potential therapeutic application to diagnose or relieve apoptotic blockades in cancers (see Introduction) [27]. Learning the rules to design high-affinity peptides is a difficult task, but is made more feasible
when large amounts of experimental data are available to guide model development and testing. Now, as we enter into a new era of quantitative binding screens, we are provided with an unprecedented wealth of data. I am interested in what we can learn from these datasets and how I can use high-throughput binding data to engineer new protein-protein interactions.

This chapter describes my pipeline for processing data generated by the SORTCERY experimental protocol. I present the technical details of the methods I developed for curating the raw SORTCERY output for use in regression-based modeling. I provide ten high-quality datasets that describe the interaction affinities of thousands of peptides for binding to Bcl-2 family anti-apoptotic proteins Mcl-1, Bcl-xL, and Bfl-1. I analyze the Bcl-xL interaction data and demonstrate evidence for cooperative interactions between peptide residues; this has implications for which types of models are most appropriate for describing the relationship between sequence and binding. I apply regression modeling with second order terms to capture cooperative intra-peptide contributions. I demonstrate the benefit of second-order models and show that epistasis exists, and can be captured to improve modeling. I used a variety of benchmark tests to evaluate how well our models extrapolate beyond the sequence spaces that were measured using SORTCERY. The modeling described in this chapter is a critical first step for enabling engineering applications such as Bcl-2 family peptide design, which is described in Chapter 3.

2.2 Results

2.2.1 Ten high-quality peptide binding affinity datasets were generated using SORTCERY

Peptide libraries were assayed for binding to Bcl-2 family proteins

To understand the relationship between BH3 peptide sequences and their binding affinities for anti-apoptotic Bcl-2 family proteins, we applied SORTCERY to investigate the interactions of two libraries of yeast-displayed peptides (see Chapter 1).
We applied the first library — referred below as the pilot library — to investigate SORTCERY’s potential as a high-throughput affinity assay [81] [29]. This pilot library varied six positions in two native BH3 binding peptides, Bim and Puma, for a combination of 944,784 possible peptide variants (see Table 2.1). Reich et al. sampled \( \sim 10^4 \) cells from the pilot library, and assayed the subpopulation for binding to Bcl-x\(_L\) at 1 nM, in two replicate experiments.

<table>
<thead>
<tr>
<th>Position</th>
<th>WT Bim</th>
<th>WT Puma</th>
<th>Residues allowed</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Gl</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Rf</td>
<td>Q</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>lg</td>
<td>P</td>
<td>W</td>
</tr>
<tr>
<td>3</td>
<td>2a</td>
<td>E</td>
<td>A</td>
</tr>
<tr>
<td>4</td>
<td>2b</td>
<td>I</td>
<td>R</td>
</tr>
<tr>
<td>5</td>
<td>2c</td>
<td>W</td>
<td>E</td>
</tr>
<tr>
<td>6</td>
<td>2d</td>
<td>I</td>
<td>I</td>
</tr>
<tr>
<td>7</td>
<td>2e</td>
<td>A</td>
<td>G</td>
</tr>
<tr>
<td>8</td>
<td>2f</td>
<td>Q</td>
<td>A</td>
</tr>
<tr>
<td>9</td>
<td>2g</td>
<td>E</td>
<td>Q</td>
</tr>
<tr>
<td>10</td>
<td>3a</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>11</td>
<td>3b</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>12</td>
<td>3c</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>13</td>
<td>3d</td>
<td>I</td>
<td>M</td>
</tr>
<tr>
<td>14</td>
<td>3e</td>
<td>G</td>
<td>A</td>
</tr>
<tr>
<td>15</td>
<td>3f</td>
<td>D</td>
<td>D</td>
</tr>
<tr>
<td>16</td>
<td>3g</td>
<td>E</td>
<td>D</td>
</tr>
<tr>
<td>17</td>
<td>4a</td>
<td>F</td>
<td>L</td>
</tr>
<tr>
<td>18</td>
<td>4b</td>
<td>N</td>
<td>N</td>
</tr>
</tbody>
</table>

Table 2.1: Residues mutated in the pilot library.

We applied the second library — referred below as the specificity library — to investigate the peptide binding specificity landscape for three Bcl-2 family paralogs: Bcl-x\(_L\), Mcl-1, and Bfl-1 The specificity library was constructed by pooling six previously generated targeted libraries (see Table 2.2) [52]. Each targeted library was designed to include residues predicted to be selective for one of the three Bcl-2 family member proteins. For each library, we introduced variability at 7 or 8 positions of native Bim or Puma BH3 sequences, for a total sequence space of 27,696,384 peptide variants. We selected a sample of \( \sim 10^4 \) cells from this pooled library and assayed
the subpopulation against the three receptors at 1 nM and 100 nM concentrations (see Methods).

<table>
<thead>
<tr>
<th>Position</th>
<th>WT Bim</th>
<th>WT Puma</th>
<th>Bcl-xl Library</th>
<th>Mcl1 Library</th>
<th>Bfl1 Library</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1e</td>
<td>G</td>
<td>G</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1f</td>
<td>R</td>
<td>Q</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1g</td>
<td>P</td>
<td>W</td>
<td>AGPRSTW</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2a</td>
<td>E</td>
<td>A</td>
<td>DEHIKLMNQV</td>
<td>AEIKLPQTV</td>
</tr>
<tr>
<td>4</td>
<td>2b</td>
<td>I</td>
<td>R</td>
<td>ACDFGHLPRTSVY</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2c</td>
<td>W</td>
<td>E</td>
<td>DEHQ</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>2d</td>
<td>I</td>
<td>I</td>
<td>DFHILNVY</td>
<td>AITV</td>
</tr>
<tr>
<td>7</td>
<td>2e</td>
<td>A</td>
<td>G</td>
<td>AGISTV</td>
<td>ACDGSY</td>
</tr>
<tr>
<td>8</td>
<td>2f</td>
<td>Q</td>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>2g</td>
<td>E</td>
<td>Q</td>
<td></td>
<td>CDFGHILNRSVY</td>
</tr>
<tr>
<td>10</td>
<td>3a</td>
<td>L</td>
<td>L</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>3b</td>
<td>R</td>
<td>R</td>
<td>AGIRTV</td>
<td>AEGIKRTV</td>
</tr>
<tr>
<td>12</td>
<td>3c</td>
<td>R</td>
<td>R</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>3d</td>
<td>I</td>
<td>M</td>
<td>ACDFGHILNPRTSVY</td>
<td>ACFGILPRSTV</td>
</tr>
<tr>
<td>14</td>
<td>3e</td>
<td>G</td>
<td>A</td>
<td>AG</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>3f</td>
<td>D</td>
<td>D</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>3g</td>
<td>E</td>
<td>D</td>
<td>DEHIKLMNQV</td>
<td>ADFHILNPSTVY</td>
</tr>
<tr>
<td>17</td>
<td>4a</td>
<td>F</td>
<td>L</td>
<td>ADFHILNPSTVY</td>
<td>ADFHILNPSTVY</td>
</tr>
<tr>
<td>18</td>
<td>4b</td>
<td>N</td>
<td>N</td>
<td>ADHILNPTV</td>
<td>DEHKNQ</td>
</tr>
<tr>
<td>19</td>
<td>4c</td>
<td>A</td>
<td>A</td>
<td>AEKT</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>4d</td>
<td>Y</td>
<td>Q</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>4e</td>
<td>Y</td>
<td>Y</td>
<td>ACDGHNPRSTY</td>
<td>AFILPSTV</td>
</tr>
</tbody>
</table>

Table 2.2: Residues mutated in the specificity library

I subjected the sequence data collected from SORTCERY to processing that imposed quality control filters. I introduced new processing steps, relative to the published work of Reich et al., which included: a reimplementation of the deep sequencing processing pipeline to extract full-length sequences rather than only sequences for variable positions; filtering to remove deep sequencing artifacts, contaminated sequences, and outlier sequence profiles; and additional annotation to describe the confidence of any assigned affinity (see Methods) [81].
<table>
<thead>
<tr>
<th>Peptide standards</th>
<th>Name</th>
<th>Attempts made to titrate peptide</th>
<th>Average $K_d$ (nM)</th>
<th>std $K_d$ (nM)</th>
<th># of $K_d$ measurements</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Bcl-xL Bfl-1 Mcl-1</td>
<td>Bcl-xL Bfl-1 Mcl-1</td>
<td>Bcl-xL Bfl-1 Mcl-1</td>
</tr>
<tr>
<td>GQWMREFGAQLRLADDLDAQY</td>
<td>B16</td>
<td>4 3 5</td>
<td>2.4 6.8 9.7</td>
<td>0.7 3.6 6.3</td>
<td>4 3 5</td>
</tr>
<tr>
<td>GQAAHEIAAQRLRMADDVEAQY</td>
<td>M2</td>
<td>4 3 4</td>
<td>31.7 3.5 4.5</td>
<td>5.8 1.5 3.7</td>
<td>3 3 3</td>
</tr>
<tr>
<td>GQWRELGAQLVRLGDDLITQY</td>
<td>M23</td>
<td>6 3 3</td>
<td>454.6 109.8 1.4</td>
<td>258.8 20.4 1.4</td>
<td>2 3 3</td>
</tr>
<tr>
<td>GQSAFQIAQLVMADDVDAQY</td>
<td>M24</td>
<td>3 4 4</td>
<td>12.5 58.7 4.8</td>
<td>5.6 81.5 2.9</td>
<td>3 4 3</td>
</tr>
<tr>
<td>GQWREVGAQRLRYGDDLTEQS</td>
<td>M31</td>
<td>3 3 3</td>
<td>121.8 7.4 3.4</td>
<td>28.1 2.3 2.4</td>
<td>3 3 3</td>
</tr>
<tr>
<td>GQWIREFQGQLRLGDDLTAQT</td>
<td>M33</td>
<td>6 3 5</td>
<td>26.4 2.9 0.2</td>
<td>24.9 1.2 0.2</td>
<td>4 3 2</td>
</tr>
<tr>
<td>GRSEPDIVQELRIGDENEAYY</td>
<td>M34</td>
<td>6 3 5</td>
<td>466.8 34.6 0.2</td>
<td>NaN 19.2 NaN</td>
<td>1 3 1</td>
</tr>
<tr>
<td>GQWREVQGQLRLGDDLVAQN</td>
<td>M4</td>
<td>4 4 4</td>
<td>33.1 36.6 0.3</td>
<td>10.4 22.9 0</td>
<td>3 4 2</td>
</tr>
<tr>
<td>GQWMREFGAQLTRSADDLVEQT</td>
<td>X12</td>
<td>2 2 2</td>
<td>500 500 500</td>
<td>0 0 0</td>
<td>2 2 2</td>
</tr>
<tr>
<td>GPRAIWIAQGRLRIGDNVAYP</td>
<td>X16</td>
<td>4 5 6</td>
<td>19.8 5.7 1</td>
<td>8.9 3.9 0.8</td>
<td>3 5 5</td>
</tr>
<tr>
<td>GQWEREAIANLRRIADEFNAQL</td>
<td>X16_B</td>
<td>5 3 2</td>
<td>13.5 4 1</td>
<td>18.2 2.3 0</td>
<td>2 3 2</td>
</tr>
<tr>
<td>GQLRELGAQILIADDLDEQS</td>
<td>X17</td>
<td>4 4 4</td>
<td>61.9 148.9 84.8</td>
<td>15.9 136.3 70.7</td>
<td>3 4 3</td>
</tr>
<tr>
<td>GREAIEAQKLRIGDENSEAYY</td>
<td>X18</td>
<td>4 3 6</td>
<td>127.2 20.6 7.5</td>
<td>122.9 9.6 8.2</td>
<td>2 3 5</td>
</tr>
<tr>
<td>GQWFHIIAOQTRMADSNAYQ</td>
<td>X22</td>
<td>2 2 2</td>
<td>500 500 500</td>
<td>0 0 0</td>
<td>2 2 2</td>
</tr>
<tr>
<td>GRPQIWAQGLRLRFDQINAYI</td>
<td>X25</td>
<td>5 3 5</td>
<td>78.5 3.9 39.7</td>
<td>48.4 3.1 38.1</td>
<td>4 3 5</td>
</tr>
<tr>
<td>GQWIREISACLRRIADDYNAQF</td>
<td>X26</td>
<td>4 5 4</td>
<td>2.5 7 0.9</td>
<td>0.6 5 0.3</td>
<td>4 4 3</td>
</tr>
<tr>
<td>GRPWIAEQLRLGDFEFLAYX</td>
<td>X92</td>
<td>3 NaN NaN</td>
<td>500 NaN NaN</td>
<td>0 NaN NaN</td>
<td>3 NaN NaN</td>
</tr>
<tr>
<td>GQWREICAFLLRRSADDINAQI</td>
<td>X93</td>
<td>3 NaN NaN</td>
<td>35.6 NaN NaN</td>
<td>5.6 NaN NaN</td>
<td>3 NaN NaN</td>
</tr>
</tbody>
</table>

Table 2.3: Sequences selected as standards for mapping SORTCERY to energy. Peptide standards were titrated to all three receptors. A number of attempts were made to titrate each peptide, but some were more less well behaved than others. The average $K_d$ value reflects the average over the number of successful titrations. The standard deviation of the measurements is shown in the std columns. NaN indicates values that could not be determined.
Figure 2-1: Correlation of the SORTCERY mean affinity coordinate (Å) to binding energy in kcal/mol, for standards - The title of each plot describes the dataset and number of data points. The blue lines show the correlation for all of the data points. The red lines show the correlation excluding points that could not be assigned a numerical $K_d$ and points whose titration curves saturated at low binding signal.
Peptide standards map SORTCERY experimental values to binding free energies

In previous work, SORTCERY was used to generate a relative affinity ranking of sequences, without a numerical assignment of affinity values. In this work, I assigned each sequence a numerical value, \( (\tilde{A}) \), derived from the weighted average affinity coordinate of the peptide profile across the FACS gates (see Methods). Theory predicts that \( (\tilde{A}) \) will be linearly correlated with binding free energies, over a certain resolution range, under certain conditions [81]. To test the relationship of \( (\tilde{A}) \) to free energies of binding, 16-39 peptide standards spanning the SORTCERY dynamic range were individually titrated against all three receptors (see Table 2.3 and Table 2.6). We observed a high correlation of \( (\tilde{A}) \) to free energies of binding, which allowed us to map \( (\tilde{A}) \) values from different experiments to a common scale. (see Figure 2-1).

We applied SORTCERY multiple times to investigate the pilot and specificity libraries. After computational filtering, I compiled ten high-quality datasets based on metrics of reproducibility and agreement with the measurements made for the standards (see Methods). For the ten datasets provided, I uniquely reference each dataset using abbreviations that indicate the Bcl-2 family receptor, and the receptor concentration. For the receptor, I use x to indicate Bcl-xL, f to indicate Bfl-1, and m to indicate Mcl-1. The receptor concentration, of 1 or 100 nM, is indicated using 1 or 100, and an ‘r’ is added to a dataset name to distinguish a repeated experiment. The two pilot library sorts are prefixed with ‘pilot’ to distinguish the two datasets generated from experiments that tested binding to Bcl-xL. From the experiments that tested the pilot library for binding to Bcl-xL at 1 nM (pilot_x1, pilot_x1r), I curated 403 and 5108 high-quality, unique peptide measurements. For the experiments that assayed the specificity library, I curated two datasets for Bcl-xL binding at 1 nM (2679 and 3457 unique peptides); two datasets for Bcl-xL binding at 100 nM (1081 and 2782 peptides); two data sets for Mcl-1 binding at 1 nM (3326 and 3480 unique peptides); and two datasets for Bfl-1 binding at 100 nM (1292 and 3489 unique peptides). Affinities determined in replicate experiments gave Pearson r between 0.76
and 0.98. Table 2.4 summarizes the ten datasets and includes information about the number of peptides analyzed, the reproducibility of replicated experiments, and the error of SORTCERY measurements, assessed using the individual titrations of the standards. Figure 2-2 shows sequence logos for peptides in the ten datasets, illustrating the frequency of residues at each position among the peptides for which measurements were made [22].

After rigorous curation and analysis of these datasets, I found that the experimental replicate reproducibility can quickly indicate whether an experiment was successful or not. SORTCERY’s reproducibility has been observed as high as Pearson r ~ 0.98; thus it is reasonable to have less confidence in datasets which have a weak correlation. I also found that for analysis of Illumina deep sequencing datasets, it is necessary to assess the experimental error by comparing the input library space with the observed space. DNA contamination and Illumina deep sequencing errors are frequent, which can influence downstream analysis if not accounted for (see Methods).
Figure 2-2: Sequence logos for peptides curated from SORTCERY
<table>
<thead>
<tr>
<th>Dataset name</th>
<th>Receptor conc. (nM)</th>
<th># peptides</th>
<th># standards measured</th>
<th>Pearson R of mean affinity coord to standards</th>
<th>RMSE of energy mapped standards (kcal/mol)</th>
<th>Replicate overlap</th>
<th>Replicate RMSE (kcal/mol)</th>
<th>Replicate Pearson R of mapped binding free energies</th>
</tr>
</thead>
<tbody>
<tr>
<td>pilot_x1</td>
<td>1</td>
<td>403</td>
<td>25</td>
<td>0.7</td>
<td>0.64</td>
<td>355(345)*</td>
<td>0.35 (0.25)*</td>
<td>0.84 (0.95)*</td>
</tr>
<tr>
<td>pilot_xlr</td>
<td>1</td>
<td>5108</td>
<td>39</td>
<td>0.74</td>
<td>0.72</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x1</td>
<td>1</td>
<td>2679</td>
<td>17</td>
<td>0.82</td>
<td>0.56</td>
<td>1749</td>
<td>0.35</td>
<td>0.91</td>
</tr>
<tr>
<td>xlr</td>
<td>1</td>
<td>3457</td>
<td>18</td>
<td>0.85</td>
<td>0.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>f100</td>
<td>100</td>
<td>1292</td>
<td>16</td>
<td>0.88</td>
<td>0.33</td>
<td>975</td>
<td>0.21</td>
<td>0.95</td>
</tr>
<tr>
<td>f100r</td>
<td>100</td>
<td>3489</td>
<td>16</td>
<td>0.84</td>
<td>0.39</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>m1</td>
<td>1</td>
<td>3326</td>
<td>16</td>
<td>0.89</td>
<td>0.47</td>
<td>2315</td>
<td>0.21</td>
<td>0.98</td>
</tr>
<tr>
<td>mlr</td>
<td>1</td>
<td>3480</td>
<td>16</td>
<td>0.92</td>
<td>0.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x100</td>
<td>100</td>
<td>1081</td>
<td>17</td>
<td>0.6</td>
<td>0.7</td>
<td>707</td>
<td>0.43</td>
<td>0.76</td>
</tr>
<tr>
<td>x100r</td>
<td>100</td>
<td>2782</td>
<td>18</td>
<td>0.67</td>
<td>0.71</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Extreme outliers were removed (< 3%) of data

Table 2.4: Summary of datasets collected from SORTCERY
2.2.2 SORTCERY data were used to identify epistatic mutational effects on binding

Detecting epistasis using single-point mutation analysis

Some changes to a peptide sequence make additive energy contributions to binding, but for others, the coupling between sites leads to non-additive effects. This has been observed previously for BH3 peptide interactions [31] [37]. To investigate epistatic effects in the new datasets, I calculated $\Delta\Delta G$ values for all single-point mutants in the pilot_xlr dataset. For mutations that are additively independent, e.g., those that do not change the peptide binding mode and do not make intra-peptide contacts that influence binding, $\Delta\Delta G$ values are expected to be constant with Gaussian measurement noise, regardless of the context in which the mutation occurs. Significant deviation of $\Delta\Delta G$ measurements from a normal distribution would suggest context-dependent effects.

$\Delta\Delta G$ values for all single-point mutations with greater than 100 unique contexts were quantified. For each population of single-point mutations, a chi-squared test was used to determine whether the calculated $\Delta\Delta G$ distribution deviated from Gaussianity (alpha < 0.05 with Bonferroni correction) (see Figure 2-3). Based on this test, mutations at positions 2D and 4A most frequently demonstrated dependence on the peptide sequence context. These sites lie towards the ends of the peptide sequence. The same trend exists for analysis of all point mutations occurring in 25 or more unique contexts.
Figure 2-3: Boxplots of ΔΔG measurements for sets of single point mutations that were observed in more than 100, 50 or 25 unique contexts (top, middle, bottom panels). The set of single point mutations is colored red if the distribution significantly deviated from Gaussianity (alpha < 0.05 with Bonferroni correction). Single point mutations that deviate from Gaussianity are observed more frequently at positions 2D and 4A.
Table 2.5: Double mutant cycles identified as having large $|\Delta\Delta G|$ from SORTCERY

<table>
<thead>
<tr>
<th>UID</th>
<th>Background</th>
<th>XY</th>
<th>X'Y</th>
<th>XY'</th>
<th>X'Y'</th>
<th>Positions</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>E-X-WAREYGAQLDRFASD-Y-N</td>
<td>QF</td>
<td>RF</td>
<td>QY</td>
<td>RY</td>
<td>1F-4A</td>
</tr>
<tr>
<td>B</td>
<td>EQWAREIGAQ-X-RR-Y-ADDIN</td>
<td>AI</td>
<td>II</td>
<td>AY</td>
<td>IY</td>
<td>3A-3D</td>
</tr>
<tr>
<td>C</td>
<td>EQWARE-X-GAQLGR-Y-AGDFN</td>
<td>YF</td>
<td>VF</td>
<td>YV</td>
<td>VV</td>
<td>2D-3D</td>
</tr>
<tr>
<td>D</td>
<td>EQWAREYGAQL-X-RYA-Y-DFN</td>
<td>ED</td>
<td>HD</td>
<td>EQ</td>
<td>HQ</td>
<td>3B-3F</td>
</tr>
<tr>
<td>E</td>
<td>EQWAREYGAQL-X-RIA-Y-DFN</td>
<td>GN</td>
<td>GD</td>
<td>RN</td>
<td>RD</td>
<td>3B-3F</td>
</tr>
</tbody>
</table>

Detecting epistasis using double mutant cycles

Given the evidence of non-additivity from the single-point mutation analysis, I further searched for examples of residue-residue coupling by listing out all double mutant cycles identified in the pilot_xlr dataset [46].

Double mutant cycles are sets of 4 sequences composed of a parent sequence containing residues X and Y (XY); two single-point variants of the parent (X'Y, XY'); and the corresponding double mutant (X'Y'). The coupling energy between residues X and Y is computed as $\Delta\Delta G = (AG_{XY} - AG_{XY'}) - (AG_{X'Y} - AG'_{Y'})$. If the change in free energy with a double mutation differs from the sum of changes in free energy with the single mutations, then the residues at the two positions are coupled.

16,590 unique double mutant cycles were identified from the pilot_xlr dataset (see Figure 2-4). For 11,467 unique cycles, $\Delta\Delta G$ values were within 1 standard deviation of zero (see Methods). However, the remaining 5,123 double mutant cycles gave coupling energies of up to 2.5 kcal/mol.

I selected five double mutant cycles that showed high $\Delta\Delta G$ for verification via cell surface titration. Within each double mutant cycle set, the component sequences were required to have unimodal profiles across the SORTCERY gates. The SORTCERY profiles of the five double mutant cycles are shown in Figure 2-5. A unique ID, A-E, identifies each double mutant cycle, and the constituent sequences can be determined by substituting the indicated residues into the background sequence (see Table 2.5).

The peptides were individually tested for binding to Bcl-xL by yeast titration to
<table>
<thead>
<tr>
<th>Peptide Name</th>
<th>Attempts made to titrate peptide</th>
<th>Average Kₐ</th>
<th>std Kₐ</th>
<th># of Kₐ measurements</th>
</tr>
</thead>
<tbody>
<tr>
<td>EQWAREYGAQLDRFASDFN</td>
<td>Bcl-xL</td>
<td>3</td>
<td>25.9</td>
<td>13.4</td>
</tr>
<tr>
<td>EQWAREYGAQLDRFASDYN</td>
<td>Bcl-xL</td>
<td>4</td>
<td>55.6</td>
<td>31</td>
</tr>
<tr>
<td>ERWAREYGAQLDRFASDFN</td>
<td>Bcl-xL</td>
<td>3</td>
<td>14.8</td>
<td>3.8</td>
</tr>
<tr>
<td>ERWAREYGAQLDRFASDYN</td>
<td>Bcl-xL</td>
<td>3</td>
<td>26.2</td>
<td>12.2</td>
</tr>
<tr>
<td>EQWAREYGAQARRIADDIN</td>
<td>AI</td>
<td>3</td>
<td>28</td>
<td>4.5</td>
</tr>
<tr>
<td>EQWAREYGAQARRIADDIN</td>
<td>AY</td>
<td>3</td>
<td>45.2</td>
<td>25.7</td>
</tr>
<tr>
<td>EQWAREYGAQIRRIADDIN</td>
<td>II</td>
<td>4</td>
<td>1.7</td>
<td>1.3</td>
</tr>
<tr>
<td>EQWAREYGAQIRRIADDIN</td>
<td>IY</td>
<td>4</td>
<td>18.6</td>
<td>10.8</td>
</tr>
<tr>
<td>EQWAREYGAQLGRFAGDFN</td>
<td>VF</td>
<td>3</td>
<td>108.7</td>
<td>38.6</td>
</tr>
<tr>
<td>EQWAREYGAQLGRVAGDFN</td>
<td>VV</td>
<td>3</td>
<td>54.8</td>
<td>14.8</td>
</tr>
<tr>
<td>EQWAREYGAQLGRFAGDFN</td>
<td>VF</td>
<td>4</td>
<td>29.6</td>
<td>17</td>
</tr>
<tr>
<td>EQWAREYGAQLGRVAGDFN</td>
<td>YV</td>
<td>3</td>
<td>112</td>
<td>19</td>
</tr>
<tr>
<td>EQWAREYGAQLERYADDIN</td>
<td>ED</td>
<td>4</td>
<td>23.2</td>
<td>11.3</td>
</tr>
<tr>
<td>EQWAREYGAQLERYADDIN</td>
<td>EQ</td>
<td>4</td>
<td>24.1</td>
<td>9.5</td>
</tr>
<tr>
<td>EQWAREYGAQLERYADDIN</td>
<td>HD</td>
<td>3</td>
<td>2.6</td>
<td>1.1</td>
</tr>
<tr>
<td>EQWAREYGAQLERYADDIN</td>
<td>HQ</td>
<td>4</td>
<td>42.1</td>
<td>15.3</td>
</tr>
<tr>
<td>EQWAREYGAQLGRRIADDIN</td>
<td>GD</td>
<td>4</td>
<td>3.1</td>
<td>0.9</td>
</tr>
<tr>
<td>EQWAREYGAQLGRRIADDIN</td>
<td>GN</td>
<td>4</td>
<td>27.8</td>
<td>17.1</td>
</tr>
<tr>
<td>EQWAREYGAQLRRRIADDIN</td>
<td>RD</td>
<td>2</td>
<td>3.3</td>
<td>4.5</td>
</tr>
<tr>
<td>EQWAREYGAQLRRRIADDIN</td>
<td>RN</td>
<td>5</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 2.6: Kₐ values for peptides selected for verification in the double mutant cycle analysis.
Figure 2-4: Distribution of $|\Delta \Delta \Delta G|$ calculated from all double mutant cycles - $|\Delta \Delta \Delta G|$ was calculated for 16,590 double mutant cycles in pilot_xlr and plotted in ascending order. The blue error bars illustrate one standard deviation. 11,467 unique double mutant cycles were within 1 standard deviation of zero.

determine how well SORTCERY can identify non-additive double mutant cycles (see Appendix for the yeast titrations). Of these five cycles, double mutant cycle C had the four peptide affinities ranked identically as was observed using SORTCERY (see Table 2.6 and Figure 2-5).

Figure 2-6 shows double mutant cycle C. In this double mutant cycle, positions 2D and 3D were varied in the context EQWARE-2D-GAQLGR-3D-AGDFN. The four sequences analyzed were: EQWAREYGAQLGRVAGDFN, EQWAREVGGAQLGRFAGDFN, EQWAREYGAQLGRFAGDFN and EQWAREYGAQLGRVAGDFN. In order of affinity, EQWAREYGAQLGRFAGDFN is the tightest binder, followed by EQWAREYGAQLGRVAGDFN, EQWAREVGAQLGRFAGDFN, and then EQWAREYGAQLGRVAGDFN. Single-point mutations from V to F at 3D or V to Y at 2D are both destabilizing, while a double point mutation that introduces both changes is stabilizing. Thus, this is an example where the context makes a dramatic effect on the influence of a point mutation. The mutated sites are positioned one
Figure 2-5: Double mutant cycle peptide profiles across SORTCERY gates - For each set of double mutant cycles, the probability distribution across SORTCERY gates of the four component peptides is plotted. Peptides that have high affinity to the receptor are expected to fall in lower numbered gates. The labels are annotated with the measured $K_d$ in nM from cell surface titration. Double mutant cycles A, B, and E, did not fully agree with measured $K_d$. $K_d$ values measured for double mutant cycles C and D agree with the SORTCERY within measurement noise.
Figure 2-6: Double Mutant Cycle C - (A) The component peptide members of double mutant cycle C were individually titrated on the yeast cell surface to determine a $K_d$. (B) A pivoted rocking motion of the peptide in the groove is a plausible mechanism to explain the observed non-additivity. Dissociation constants for different peptides binding to Bcl-xL are given in nM. Binding of peptides with one large residue and one small residue is less favorable than binding by peptides with two small or two large residues. Data in panel A collected by L. Stretz.
heptad away from each other and both orient side chains towards the receptor. The observed energetic coupling is thus consistent with a structural shift that favors similarly sized residues at 2D and 3D. A shifted BH3 binding mode in the groove has been previously observed for Bfl-1 specific peptides [52].

For the double mutant cycles that did not agree with SORTCERY, we further investigated each component peptide to identify what could have caused this difference. Previously measured standards for SORTCERY reported a strong Pearson correlation of 0.88 between the binding free energy of the individually measured peptides and $\tilde{A}$. However, the 20 peptides identified for this analysis demonstrated a weak Pearson correlation of 0.54. This weak correlation suggested that there were errors mapping $\tilde{A}$ to affinity, or experimental noise.

One explanation for this discrepancy was identified by studying the FACS plots of the individual peptide titrations. For peptide EQWAREYGAQLRRIANDFN in double mutant cycle E, the saturated binding profile observed on the FACS plot was distinctly lower than the other saturated peptides measured in parallel. SORTCERY assumes that all displayed peptides when fully bound, will give rise to the same saturated ratio of binding:expression signal [81]. The lower saturation point explains why SORTCERY assigned it a weak $\tilde{A}$ while individual titrations report it as a tight 3 nM binder.

Another explanation for why SORTCERY can be inconsistent with the titrations is from contamination. For example, double mutant cycle A included two peptides that were outside of the designed input library. We cannot report on the reliability of these measurements because we do not know whether the peptides are real mutants that appeared before sorting or whether the peptides are DNA sequencing artifacts.
Figure 2-7: Schematic of the protocol used to train receptor-specific models. Binding energies determined by SORTCERY were used as input to train regression models that relate peptide sequence to binding affinity. By modeling with an amino-acid encoding and a linear kernel, we can extract out weights that can be visualized to learn about the determinants of binding. At each varied peptide position, every amino acid is assigned a weight that describes the residue’s energetic contribution to binding. Negative weights are stabilizing, and positive weights are destabilizing.
2.2.3 Modeling the relationship between sequence and binding free energy with SORTCERY data

Regression modeling

To generate a model that relates peptide sequence to binding free energy, I applied regression to learn functions of the form $f(\text{sequence}) = E_{\text{SORTCERY}}$ for each dataset. Sequences were encoded with a binary, amino-acid encoder to map each peptide to a vector (see Methods). Two alternative encoders were also evaluated but did not demonstrate any performance benefits (see nested cross-validation below). I applied support vector regression to train two models for each dataset. One model assumes independence between the variable sites, and the second model accounts for cooperativity between pairs of sites.

An amino-acid encoder applied with a linear kernel solves for the independent contribution of each residue to the binding free energy. An advantage of this simple model is that it is possible to visualize the weights of all residues at each position, to learn the residue preferences, as shown in Figure 2-7. For example, the models captured the hydrophobic vs. hydrophilic preferences of the amphipathic helix. Positions 2D, 3A, 3D, and 4A are buried and face the hydrophobic groove while 3B and 3F are more solvent-exposed. I can extract and visualize the weights from a regression model with a second-order polynomial kernel, but these higher-order weights are not as intuitive as the linear models (see Methods).
Figure 2-8: Nested cross-validation performance for models trained on each of the 10 datasets, with both linear and second order polynomial models. Light grey bars correspond to the average $R^2$ of models trained and evaluated on the same data for the amino acid encoding. Red and dark grey bars correspond the average $R^2$ of models evaluated data held out from training, for polynomial and linear models respectively. The green error bars correspond to the standard deviation of the 10 folds of nested cross-validation. Second order polynomial models consistently outperformed linear models trained on the same dataset.
Figure 2-9: Nested cross-validation performance for models trained with the five-factor encoding. Colors are as in Figure 2-8.
Figure 2-10: Nested cross-validation performance for models trained with the eleven-factor encoding. Colors are as in Figure 2-8.
Figure 2-11: Correlation of the mapped binding energies for each pair of SORTCERY replicates. The x and y axis represent the mapped binding free energies from each replicate.
Nested cross validation quantifies generalization performance and provides values for model hyperparameters

To select the best model hyperparameters and estimate each model's generalization performance, I applied nested cross-validation (NCV) (see Methods). In Figure 2-8, I report the average $R^2$ of the model fits for training and validation subsets. Second-order polynomial models evaluated on the validation data consistently outperformed linear models trained on the same data by an average of 9.1 percent. This observation suggests that linear models are underfitting where they fail to capture some replicable structure in the data, which the polynomial models capture through the pair terms. The linear models showed characteristics of underfitting such that they had high bias and low variance. High biased models make assumptions that limit model performance. Low variance models are insensitive to changes in the input training set. The linear models demonstrated low variance where the models plateaued in performance on both training and validation datasets at lower $R^2$ than the second-order polynomial models. (i.e., linear models trained on the pilot_x1r, m1, and m1r datasets). These results imply that the models were not sufficient complex to capture more signal from the dataset. The second-order models demonstrated higher variance, such that the gap between train and test performance was large, but these models still achieved higher validation performance than linear models.

I further observed that the validation performance between replicate datasets could be very different, which indicates that the datasets have variable quality. For example, models trained on the x100 dataset showed poor performance in nested cross-validation, compared to models trained on the x100r dataset. Despite both datasets having access to the same model complexity in training, and many of the same input sequences, modeling with the x100 dataset had irreducible error, which implies lower quality data. (Irreducible error is error that cannot be reduced by choosing a better model. It is due to the randomness in a system). The inconsistency in model performance also agrees with the poor replicate reproducibility between the x100 and x100r datasets (see Figure 2-11). These observations suggest that we can
detect poor experimental data using cross-validation.

I further applied nested cross-validation to compare the performance of models that use alternative residue encodings. I tested two reduced feature encodings that represent each residue as a vector of 5 or 11 features, as opposed to the 20 used in an amino-acid encoding. The 5-feature encoder represents each residue by a vector derived from principal coordinate analysis of 237 physical-chemical properties [67]. The 11-feature encoder represents each residue as a vector of numerical values that describe physicochemical properties of the amino acid residues [109]. The comparison of amino-acid, 5-, or 11-feature models demonstrated comparable performance (see Figure 2-9, 2-10). Therefore, I opted to continue using the amino-acid encoding to maintain interpretability.

The hyperparameters learned from nested-cross validation were applied to train a final linear and second-order polynomial model for each dataset, as described in the Methods.

**Benchmarking SORTCERY models**

To further evaluate the SORTCERY models, I compared the best models — as identified from NCV — to alternative protein interaction models. To determine whether higher order models improve performance over linear models, I tested the two model variants on data held out from training and on new sequence spaces. Finally, I evaluated the models on datasets previously reported in the literature.

**Comparisons of regression models with other methods of predicting interaction affinities** To compare the regression models to other methods of predicting interaction affinity, I benchmarked against naive models such as a Blosum-based predictor, empirical models such as PSSMspot and PSSMfreq, structure-based statistical models like STATIUM, and advanced parameterized structural models like FoldX. I applied these models to rank peptides in each SORTCERY dataset and computed the Spearman $\rho$ correlation of the predicted ranks to observed ranks (Figure 2-12). These scores were compared to the average 10-fold cross-validated Spearman $\rho$ correlation.
of the regression models (see Methods). I report a Spearman correlation because not all models are expected to predict scores with a linear relationship to binding free energies.

The Blosum scoring model is a naive model which applies the Blosum62 scoring matrix to calculate the similarity of an input sequence to a given reference. To score the SORTCERY datasets, I calculated the similarity of each sequence to the high-affinity native binder Bim. This simple model performed significantly better than random and captured the correct trend for most datasets. However, this model underperformed the other models that have more prior knowledge.

The PSSMspot model is an empirical scoring matrix built from SPOT array data for BH3 peptide mutants binding to Bcl-xL, Mcl-1 or Bfl-1. The SPOT experiments measured the binding affinity of a parent Bim BH3 peptide and 180 point mutants of the parent sequence. The PSSMspot model calculates the energy of substitution for a residue by taking the negative logarithm of the ratio of fluorescence intensity for the substitution to the intensity of the wild-type BH3 [25]. The PSSMspot models were applied to score their respective receptor specific dataset and achieved modest performance with an average Spearman $\rho$ of 0.41 across all datasets (Figure 2-12).

STATIUM is a structure-specific statistical scoring function that evaluates peptide binding to a pre-specified template structure based on the frequency that specific residue interactions are found in the PDB [25]. STATIUM achieved comparable performance with PSSMspot with an average Spearman $\rho$ of 0.38 across all datasets.

FoldX is a complex empirical force field developed for the rapid evaluation of the effect of mutations on protein stability [92]. Like STATIUM, FoldX does not use any system-specific binding data, other than an input crystal structure of the complex. FoldX achieved similar performance to STATIUM.

The PSSMfreq model was built off a multiple sequence alignment of the top 30% of the test dataset. These models derive weights empirically from the SORTCERY datasets by applying maximum likelihood (see Methods). The PSSMfreq method outperformed STATIUM and FoldX in all test cases with an average Spearman $\rho$ of 0.55.
Figure 2-12: Alternative modeling benchmarks on SORTCERY data - The 10 SORTCERY datasets in Table 2.4 were scored with a variety of protein-interaction models to benchmark model performance. The values and the heatmap colors report the Spearman correlation of predicted rank to observed rank for each model and dataset. All models outperformed the naive Blosum model on most tests. The second order polynomial model demonstrated the best average performance across all datasets.

The cross-validated regression models achieved the highest average Spearman $\rho$ performance (linear: 0.64, poly:0.67), demonstrating the performance benefits of specialized data-driven models.

**Comparisons of regression models on replicate exclusive peptides** In the nested-cross validation tests, polynomial models outperformed the linear models. To determine how generalizable this pattern is, I tested the models on data never observed in the training set. I generated test datasets that were non-overlapping with the training data using replicate experiments. For each SORTCERY experiment, the input population of cells tested was sampled and grown from a common library stock. Due to random sampling, the peptides observed in each experiment were different.
Table 2.7: Evaluation of model performance to score sequences from a different experiment - Polynomial and linear models trained on pilot_xlr and x1 datasets were evaluated on non-overlapping sequences from the other data set. Each row is a model and each column is a test dataset. Pearson correlation of the predicted energy and the observed binding energy is reported. The bolded values highlight the model performance on fair tests where the model did not observe any test data during training.

<table>
<thead>
<tr>
<th>Test</th>
<th>pilot_xlr (Pearson r)</th>
<th>x1 (Pearson r)</th>
</tr>
</thead>
<tbody>
<tr>
<td>pilot_xlr_linear</td>
<td>0.84</td>
<td>0.36</td>
</tr>
<tr>
<td>pilot_xlr_poly</td>
<td>0.92</td>
<td>0.45</td>
</tr>
<tr>
<td>x1_linear</td>
<td>0.58</td>
<td>0.8</td>
</tr>
<tr>
<td>x1_poly</td>
<td>0.62</td>
<td>0.92</td>
</tr>
</tbody>
</table>

Thus, for every pair of replicates, there exists a set of peptides exclusive to each of the two datasets. As a test of generalization, models were trained on a given dataset and evaluated on the peptides exclusive to its replicate dataset. These tests demonstrated that second-order polynomial models outperformed linear models when evaluating non-overlapping sequences drawn from the same sequence space. In eight out of ten tests, the polynomial models outperformed the linear models (see Figure 2-13). The Pearson correlation of predicted to observed in these tests have high variability for reasons I speculate about in the Discussion.

**Comparisons of regression models across library spaces** To evaluate whether polynomial regression models offered performance benefits for extrapolating across library spaces, I tested linear and second-order polynomial models for their ability to correctly predict the affinity of peptides outside of the input library space. Given that the pilot library and the specificity library have non-overlapping sequence spaces, I tested the performance of models trained on the pilot_xlr dataset to extrapolate to data from the x1 dataset and vice versa. I found that the polynomial models trained on the pilot_xlr dataset outperformed the linear models when evaluated on the x1 dataset. (Poly: 0.45 vs. Linear 0.36 Pearson r). Similarly, polynomial models trained on the x1 dataset outperformed linear models when evaluated on pilot_xlr dataset (Poly: 0.62 vs. Linear 0.58 Pearson r) (Table 2.7). I further speculate about the performance differences in the Discussion.
Figure 2-13: Replicate holdout tests - Linear and polynomial models were trained on each dataset in Table 2.4. Each model was evaluated on the set of sequences exclusive to its replicate dataset. (ie. a model trained on dataset x1 was tested on sequences exclusive to dataset x1r.) The Pearson correlation of the predicted binding energies to the observed binding energies is reported. This analysis shows that second-order polynomial models outperform linear models on 8 out of 10 tests. The table describes the number of peptides unique to each replicate dataset.
Figure 2-14: Model extrapolation performance on Bcl-2 protein-protein interaction datasets. Bcl-2 protein-protein interaction datasets were collected from the literature and evaluated with the best receptor specific models. For the Bcl-xL, Mcl-1, and Bfl-1 affinity tests, the performance is reported as a Spearman $\rho$ correlation coefficient. For the binary discrimination of Bfl-1 binders vs. non-binders and Bcl-xL vs. Mcl-1 specificity, the area under the ROC curve is reported. Datasets suffixed with "sub" are subsets of the data that included only the sampled residues in the specificity library.
<table>
<thead>
<tr>
<th>Dataset name</th>
<th>Format</th>
<th>Context</th>
<th>Relevant receptors</th>
<th>Full dataset count</th>
<th>Subset used testing count</th>
<th>Overlap with spec library</th>
<th>Overlap with pilot library</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>KD_joes_fpkd</td>
<td>Numerical</td>
<td>K_d</td>
<td>ALL</td>
<td>36</td>
<td>36</td>
<td>0</td>
<td>0</td>
<td>36 BH3 peptides computationally identified and validated to bind human Bcl-2 proteins. [26]</td>
</tr>
<tr>
<td>KD_s360_mccl</td>
<td>Numerical</td>
<td>SPOT</td>
<td>Bim</td>
<td>359</td>
<td>359</td>
<td>0</td>
<td>35</td>
<td>SPOT array testing 360 select peptides with mutations in the pilot library(Dutta) - [29]</td>
</tr>
<tr>
<td>KD_s360_bchxl</td>
<td>Numerical</td>
<td>SPOT</td>
<td>Bim</td>
<td>359</td>
<td>359</td>
<td>0</td>
<td>35</td>
<td>SPOT array testing 360 select peptides with mutations in the pilot library(Dutta) - [29]</td>
</tr>
<tr>
<td>RANK_bchxl_enrh</td>
<td>Rank</td>
<td>Deep Sequence</td>
<td>Bim</td>
<td>Bcl-xL</td>
<td>3645</td>
<td>400</td>
<td>0</td>
<td>47</td>
</tr>
<tr>
<td>RANK_bfil_comp</td>
<td>Rank</td>
<td>Deep Sequence</td>
<td>Puma</td>
<td>Bf-1</td>
<td>4780</td>
<td>400</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SPOT_bchxl_bim_1000</td>
<td>Numerical</td>
<td>SPOT</td>
<td>Bim</td>
<td>Bcl-xL</td>
<td>171</td>
<td>171</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>SPOT_bchxl_bim_1000_sub</td>
<td>Numerical</td>
<td>SPOT</td>
<td>Bim</td>
<td>Bcl-xL</td>
<td>68</td>
<td>68</td>
<td>0</td>
<td>22</td>
</tr>
<tr>
<td>SPOT_bchxl_bim_100</td>
<td>Numerical</td>
<td>SPOT</td>
<td>Bim</td>
<td>Bcl-xL</td>
<td>171</td>
<td>171</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>SPOT_mccl_bim_100</td>
<td>Numerical</td>
<td>SPOT</td>
<td>Mcl-1</td>
<td>171</td>
<td>171</td>
<td>0</td>
<td>25</td>
<td>SPOT array data from Bim BH3 substitution arrays at positions exploring all residues (except Cys and Met) at 2d, 2e, 2g, 3a, 3b, 3d, 3e, 3f, 3g, and 4a [29]</td>
</tr>
<tr>
<td>SPOT_mccl_bim_100_sub</td>
<td>Numerical</td>
<td>SPOT</td>
<td>Mcl-1</td>
<td>68</td>
<td>68</td>
<td>0</td>
<td>22</td>
<td>SPOT array data from Bim BH3 substitution arrays at positions exploring all residues (except Cys and Met) at 2d, 2e, 2g, 3a, 3b, 3d, 3e, 3f, 3g, and 4a [29]</td>
</tr>
<tr>
<td>SPOT_mccl_noxa_100</td>
<td>Numerical</td>
<td>SPOT</td>
<td>Noxa</td>
<td>172</td>
<td>172</td>
<td>0</td>
<td>0</td>
<td>SPOT array data from Noxa BH3 substitution arrays at positions exploring all residues (except Cys and Met) at 2d, 2e, 2g, 3a, 3b, 3d, 3e, 3f, 3g, and 4a [29]</td>
</tr>
<tr>
<td>SPOT_mccl_noxa_100_sub</td>
<td>Numerical</td>
<td>SPOT</td>
<td>Noxa</td>
<td>69</td>
<td>69</td>
<td>0</td>
<td>0</td>
<td>SPOT array data from Noxa BH3 substitution arrays at positions exploring all residues (except Cys and Met) at 2d, 2e, 2g, 3a, 3b, 3d, 3e, 3f, 3g, and 4a [29]</td>
</tr>
<tr>
<td>SPOT_bchxl_bad_100</td>
<td>Numerical</td>
<td>SPOT</td>
<td>Bad</td>
<td>Bcl-xL</td>
<td>172</td>
<td>172</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SPOT_bchxl_bad_100_sub</td>
<td>Numerical</td>
<td>SPOT</td>
<td>Bad</td>
<td>Bcl-xL</td>
<td>70</td>
<td>70</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SPOT_bfil_bim_100</td>
<td>Numerical</td>
<td>SPOT</td>
<td>Bf-1</td>
<td>171</td>
<td>171</td>
<td>0</td>
<td>25</td>
<td>SPOT array data from Bim BH3 substitution arrays at positions exploring all residues (except Cys and Met) at 2d, 2e, 2g, 3a, 3b, 3d, 3e, 3f, 3g, and 4a [29]</td>
</tr>
<tr>
<td>SPOT_bfil_bim_100_sub</td>
<td>Numerical</td>
<td>SPOT</td>
<td>Bf-1</td>
<td>68</td>
<td>68</td>
<td>0</td>
<td>22</td>
<td>SPOT array data from Bim BH3 substitution arrays at positions exploring all residues (except Cys and Met) at 2d, 2e, 2g, 3a, 3b, 3d, 3e, 3f, 3g, and 4a [29]</td>
</tr>
<tr>
<td>KD_gelm_bchxl</td>
<td>Numerical</td>
<td>K_d</td>
<td>Bim</td>
<td>Bcl-xL</td>
<td>52</td>
<td>52</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>KD_gelm_bchxl_sub</td>
<td>Numerical</td>
<td>K_d</td>
<td>Bim</td>
<td>Bcl-xL</td>
<td>15</td>
<td>15</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>KD_gelm_mccl</td>
<td>Numerical</td>
<td>K_d</td>
<td>Bim</td>
<td>Mcl-1</td>
<td>52</td>
<td>52</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>KD_gelm_mccl_sub</td>
<td>Numerical</td>
<td>K_d</td>
<td>Bim</td>
<td>Mcl-1</td>
<td>15</td>
<td>15</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2.8: Affinity extrapolation data
<table>
<thead>
<tr>
<th>Dataset name</th>
<th>Format</th>
<th>Data source</th>
<th>Context</th>
<th>Relevant receptors</th>
<th>Full dataset count</th>
<th>Subset used for testing count</th>
<th>Overlap with spec library</th>
<th>Overlap with pilot library</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bim</td>
<td>Binary</td>
<td>Deep Sequence</td>
<td>Bim</td>
<td>Bfl-1</td>
<td>39375</td>
<td>400</td>
<td>0</td>
<td>0</td>
<td>Sequences identified from deep sequencing of binding and non-binding populations of yeast displayed peptides with FACS [52]</td>
</tr>
<tr>
<td>Puma</td>
<td>Binary</td>
<td>Deep Sequence</td>
<td>Puma</td>
<td>Bfl-1</td>
<td>47167</td>
<td>400</td>
<td>4</td>
<td>0</td>
<td>Sequences identified from deep sequencing of binding and non-binding populations of yeast displayed peptides with FACS [52]</td>
</tr>
<tr>
<td>DeepSeqSpec</td>
<td>Binary</td>
<td>Deep Sequence</td>
<td>Bim</td>
<td>Bcl-xL/Mcl-1</td>
<td>4770</td>
<td>400</td>
<td>0</td>
<td>19</td>
<td>Sequences identified from deep sequencing of Bcl-xL and Mcl-1 competition sorts (Dutta) [29]</td>
</tr>
<tr>
<td>Spec</td>
<td>Binary</td>
<td>Deep Sequence</td>
<td>Bim</td>
<td>Bcl-xL/Mcl-1</td>
<td>73</td>
<td>73</td>
<td>0</td>
<td>34</td>
<td>Sequences identified as Bcl-xL or Mcl-1 specific from individual affinity measurements (Dutta) [29]</td>
</tr>
</tbody>
</table>

Table 2.9: Categorical extrapolation data
Comparisons of linear vs. polynomial regression models on non-SORTCERY datasets From past literature and extensive study of Bcl-2 family proteins, there exist several medium-sized datasets for which I can benchmark the SORTCERY models. I applied these datasets to evaluate model performance to rank binding affinity, separate peptides based on specificity, and discriminate binders from non-binders. I tested the best linear and second-order polynomial models trained on the pilot_xlr, x1, m1r, and f100 datasets. For the affinity ranking tests, I evaluated 8 Bcl-xL, 6 Mcl-1, and 3 Bfl-1 affinity datasets. For the specificity tasks, I evaluated two Bcl-xL vs. Mcl-1 datasets. For the binder discrimination task, I evaluated two Bfl-1 binder/non-binder datasets. Descriptions of all of the test datasets can be found in the Table 2.8 and 2.9. I reported Spearman rank correlation for the affinity tests and area under the curve of the ROC curve (TPR/FPR) for the binary tests.

In this benchmarking test, I had two objectives: first, to understand the limitations of the regression models as applied to non-SORTCERY data, and second, to investigate whether polynomial models offered any performance benefits in this new sequence space. In these extrapolation tests, I concluded that regression model performance is dependent on the test dataset. As expected, the empirical regression models performed poorly when evaluated on datasets that include residues never evaluated. For these datasets, re-evaluating models on subsets that included only the sampled residues demonstrated improved performance (see Figure 2-14). I identify and comment on a subset of the benchmarking observations below.

For Bcl-xL binding, the datasets describe numerical Kd values, SPOT array fluorescence data, and peptide rankings derived from multiple rounds of enrichment using yeast-surface display. I scored these datasets using the linear and polynomial models trained on pilot_xlr or x1 datasets. Of interest is the spot 360 test set, which included peptides in the same sequence space as the library used to train the pilot_xlr model: The pilot_xlr training set included 35 of the 359 peptides in the test set. Prediction using the polynomial model trained on pilot_xlr achieved a Spearman ρ of 0.64. Both models trained using the x1 dataset performed surprisingly well where the polynomial model trained on the x1 dataset achieved a Spearman ρ of 0.54, de-
spite not having any of the SPOT array peptide sequences in the training set. A second test set of interest is the SPOT dataset that tested mutations on another BH3 scaffold named BAD. Despite the scaffold differences, pilot_x1r and x1 models were capable of capturing the single-point mutation effects with Spearman $\rho$ as high as 0.53.

Data available for Mcl-1 binding include numerical $K_d$ values and SPOT array data. I scored these datasets with linear and polynomial models trained on the m1r dataset. Of interest in these tests is the poor performance of both the linear and polynomial models to score mutations measured in the context of the Noxa BH3 peptide. Unlike models for scoring Bcl-xL binding, which correctly ranked the mutations on the BAD scaffold well (0.53), the Mcl-1 models could not correctly rank the effect of these mutations on the Noxa scaffold (-0.18).

Data available for Bfl-1 binding include $K_d$ values, sequences ranked from Bfl-1 competition sorts, and SPOT data. I scored these datasets with the linear and polynomial models trained on the f100 dataset. Of interest here is the significant increase in spearman correlation when restricting the SPOT Bfl-1 data to the set of sequences that only include mutations in the specificity library (-0.25 to 0.26).

For the binary classification test to distinguish Bcl-xL vs. Mcl-1 specific binders, I scored the test sets using the best (as determined by NCV) linear and polynomial models trained on data from the pilot_x1r, x1, and m1r datasets. I evaluated two test sets that varied the same positions as in the pilot library. I found that the Mcl-1 models performed well, with AUC between 0.87 and 1 on both tests. This is interesting given that some of the positions varied in the test set were not observed in the m1r dataset. The models trained on x1 data performed poorly in comparison to those trained on the pilot_x1r data. This is not surprising, considering that the pilot_x1r library covered much of the same sequence space as the test dataset. One interesting finding was the increase in performance of the x1 models when comparing linear vs. polynomial models (0.23 units).

For the binary separation of binders vs. non-binders, I evaluated the Bfl-1 linear and polynomial models trained in the f100 dataset. I observed high performance with
AUC near 1, which was interesting given that the models observed 4 of the 400 data points in training.

Finally, to determine whether the polynomial models improve extrapolation performance relative to linear models, I tabulated how many times the polynomial models outperformed the linear ones. I found that in a majority of the tests, the polynomial models outperformed the linear models: 10/12 of the pilot_xlr models, 11/12 of the x1 models, 8/10 of the m1r models, 3/4 of the f100 models, 2/2 of the pilot_xlr models for the specificity tests and 1/2 of the tests for Bfl-1 binding.

**Comparisons of regression models on novel affinity space** To test whether a regression approach can extrapolate outside of an input affinity space, I tested multiple models trained with stratified data. The pilot_xlr dataset was partitioned into five sets, A-E, from which I trained linear and second-order polynomial models with nested cross-validation (see Figure 2-15). Partitions A-D were equally split subsets of the resolvable peptides, and partition E is the subset of the non-resolvable weak binders (see Methods). All of the models trained on partial data exhibited good performance when classifying extreme affinity differences. These models typically reported an AUC between 0.96 to 1 when distinguishing set A and set E. AUC scores dropped as the task shifted to classifying affinity targets that were closer in affinity. The performance of the first-order linear models and second-order polynomial models was comparable in this test.

2.3 Methods

2.3.1 Experimental data collection

Experimental sorts

The pilot library was constructed as described by Dutta et al. [31]. The specificity library was constructed as described by Jenson et al. [52]. Data from the pilot library was collected as described by Reich et al. [81]. Similar experimental steps
Figure 2-15: Evaluating generalizability across affinity ranges. The pilot x1r dataset was split into 5 parts (A-E) consisting of 4 equally split subsets of resolvable binders (A-D) and a set of weak binders with unresolved affinities. Sequences with unresolved affinities (E) have probability distribution profiles where the peak occurs at gate 11 or 12. Models were trained on each subset or a combination of subsets in nested cross-validation. The models were evaluated on their performance to correctly categorize the tight vs. weak sequences in a given test set. Each row describes the subsets combined for the test. Values are reported as area under the ROC curve (AUC). Bins are highlighted green to indicate that there was no intersection of sequences in training and testing sets.
were applied to collect data from the specificity library, except for two changes in downstream sequencing, as described below.

The first change was to linearize the circular DNA plasmid before the first PCR, which increases DNA yield for deep sequencing. The second change was to add multiplexing barcodes. For the initial investigation of the pilot library, each dataset was collected in a single deep sequencing experiment. Subsequent experiments tested more receptors and receptor concentrations; thus protocols were developed to pool the experiments to reduce cost. An additional set of barcodes was attached at the 3' end of the DNA to identify the source experiment uniquely (ATCACG, ACAGTG, CGATGT, CAGATC, GATCAG, GCCAAT, TTAGGC).

In total, 12 FACS sorts were performed and pooled to be deep sequenced in parallel. The deep sequencing runs were performed using an Illumina HiSeq or NextSeq sequencer. The first Illumina dataset included data for 6 sorts that reported on the binding affinity of peptides in the specificity library to Bcl-xL, Mcl-1, or Bfl-1 at 1 nM and 100 nM concentrations. The second Illumina dataset included data for a single sort of the specificity library against the Bfl-1 receptor at 100 nM. The third Illumina dataset included data for five sorts of the specificity library against Bcl-xL at 100 nM and 1 nM, Mcl-1 at 1 nM in replicate, and Bfl-1 at 100 nM.

**Selecting input library for SORTCERY**

To select a diverse pool of ~10,000 BH3 sequences for the multi-receptor SORTCERY experiments, we grew the six yeast display libraries described in Jenson et al. and pooled the libraries before incubating with 100 nM Bfl-1, Mcl-1, or Bcl-xL [52]. Cells were sorted into 12 gates set to separate binders of different affinities as described in Reich et al. [81]. Sorted cells were grown overnight in SD+CAA. An equal number of cells from the Bfl-1, Mcl-1, and Bcl-xL sorts were pooled together to make a final pool of ~10,000 cells. Of the ~3,333 cells from each sort, most were selected from the highest affinity gate (~540 cells), and the fewest were selected from the lowest affinity gate (~25 cells) with a linear sampling gradient in between. The mixed library was grown overnight and stored in glycerol stocks.
To experimentally determine affinities of yeast-displayed peptides for Bfl-1, Mcl-1, and Bcl-xL, we sorted the mixed library into 12 affinity gates and subsequently deep-sequenced DNA from cells collected in each gate following the SORTCERY protocol described in detail by Reich et al. [81].

2.3.2 Computational processing of SORTCERY data

Filtering sequences for high-fidelity reads

Deep sequencing datasets were filtered for high quality reads with at least 99% base call accuracy and for specific multiplex barcodes used to identify each experiment individually. Paired-end reads that did not overlap were discarded, and the overlapping paired-end reads were reassembled into a contig. Unique contigs that had at least 100 reads were processed for further analysis.

Generating clone profiles over gates and average affinity coordinates

Clonal cell counts per gate were estimated as a function of the deep sequencing read counts. Given that a different number of cells are collected per gate, the read counts do not map directly to cell counts. To calculate the cell count for sequence \( x \) in a given gate \( i \), \( n_i(x) \), we first calculated the clone’s relative frequency in that gate as the number of reads for sequence \( x \) in gate \( i \), \( r_i(x) \), divided by the sum of all reads for all sequences for gate \( i \). The clone’s relative frequency was then scaled by the observed number of cells recorded to hit gate \( i \) in a fixed amount of time, \( c_i \).

\[
n_i(x) = \left( \frac{r_i(x)}{\sum_{j=1}^{12} r_i(j)} \right) \times c_i \tag{2.1}
\]

The calculated cell counts were normalized to determine the probability distribution over gates for each sequence. The probability of finding clone \( x \) in gate \( i \) is given by:

\[
p_i(x) = \frac{n_i(x)}{\sum_{k=1}^{12} n_k(x)} \tag{2.2}
\]
To mitigate the effects of sequencing error on our analysis, DNA sequences were clustered by sequence similarity using USEARCH with a 3 percent identity cutoff [32]. Within each DNA cluster, the sequence with the most reads was assigned as the parent of the cluster, and all other sequences were assigned as daughters. Daughter probability distributions over gates were compared to parent profiles. A daughter sequence was split into its own cluster if its probability distribution over gates differed significantly from the parent probability distribution (chi-squared test with Bonferroni correction: alpha-value < 0.005/\# clones). Otherwise, daughter sequences were combined with parent sequences and each cluster was assigned a new probability distribution profile over gates (see Figure 2-16). The new probability distribution was calculated such that the number of reads per gate \( r_i(x) \) is equal to the sum of reads per gate for all sequences in the cluster. Sequences that only occurred in one gate were removed because profiles for individual clones measured independently always span multiple gates. We also removed sequences with non-unimodal profiles using a custom python script that looked for patterns of monotonic increase followed by monotonic decrease.

The probability distribution of cells over gates for each sequence approximates the distribution of measured clones along an axis of affinity [81]. Profiles were used to compute a mean affinity coordinate, \( \bar{A}(x) \), for each sequence using eq. 2.3, where \( i \) is the gate identity.

\[
\text{Mean affinity coordinate}(x) = \bar{A}(x) = \sum_{i=1}^{12} i \cdot p_i(x) \tag{2.3}
\]

DNA sequences were translated into protein sequences for all subsequent analyses, yielding a list of protein sequences and their associated \( \bar{A}(x) \) values. Redundant protein-to-energy mappings originating from synonymous mutation were removed.

**Filtering to remove experimental artifacts**

Sequences that may have originated from spurious mutation, PCR error, and cross-library contamination were removed from the dataset by filtering for only sequences
Figure 2-16: Removing DNA artifacts - A computational pipeline was implemented to remove DNA artifacts from single-point mutations. DNA reads were clustered by sequence identity. Within a single cluster, the clonal probability distributions across the gates were compared. The sequences were split off into separate clusters if the distributions across the gates differed. The sequences were clustered again by amino acid sequence.
that matched the designed library input using regex (regular expression) on the DNA level.

**Classifying resolvable vs. non-resolvable sequences**

Sequences were classified into three categories of affinity (unresolvable tight, resolvable, and unresolvable weak). These sequences were separated based on the shape of the probability distribution function of the peptide across FACS gates. If the mode of the distribution occurred in gate 1, the binder was classified as unresolvable tight. If the mode of the distribution occurred in gate 11 or 12, the binder was classified as unresolvable weak. The remaining sequences were classified as resolvable.

**Mapping mean affinity coordinate to ΔG**

The $\tilde{A}(x)$ assigned to each clone reflects the relative binding affinity, which does not permit comparison across experiments. To compare affinity values across experiments, we used standards to calibrate the affinity axis to report binding free energy in kcal/mol. Standards were selected to span the SORTCERY affinity range of each experiment. For each standard, a binding curve was measured via yeast titration and fit to give the standard free energy of binding. We applied linear regression to map $\tilde{A}(x)$ values to energies. Although a linear fit is an approximation of the true relationship between $\tilde{A}(x)$ and $\Delta G_{\text{binding}}$, fitting the theoretical curve gave minimal differences in values.

**Selecting the ten best datasets**

We collected 12 datasets that quantify receptor specific binding for peptides in the specificity library. After processing the data to determine the $\tilde{A}(x)$ for peptides in each dataset, we applied pairwise correlation analysis to each replicate dataset. From this approach, we identified that 8 of the 12 datasets demonstrated good replicate agreement (Pearson r $> 0.75$). Including the pilot library datasets, we have a total of ten highly reproducible receptor specific sequence-to-affinity datasets. See Figure 2-11 for the pairwise comparison of replicate experiments.
Calibrating the SORTCERY axis of affinity

39 peptides from the pilot library were measured for binding to calibrate the axis of affinity for the pilot sorts. 19 sequences were previously measured by Reich et al., and an additional 20 sequences were measured as part of this work, as part of the double mutant cycle study [81].

16 sequences representing peptides that were present in each of six SORTCERY experiments were chosen to serve as standards for calibrating the axis of affinity. The standards were chosen to span the range of affinities that were quantifiable under the experimental conditions used. To measure dissociation constants on the cell surface, clonal populations of yeast cells were titrated with increasing concentrations of Mcl-1, Bcl-xL or Bfl-1, and the median binding signal of expressing cells was measured using FACS. Curves were fit to a 1:1 hyperbolic binding model. For some clones, we observed a decrease in signal at high concentrations. Such anomalous data points, with signal decreasing below the maximum with increasing concentration, were removed from the analysis (see Appendix). I selected the standards to be measured and analyzed the data. Experiments were performed by Lindsey Stretz under the direction of Justin Jenson.

2.3.3 Epistasis analysis

Single point mutation analysis

Given a list of sequence-to-affinity mappings for thousands of peptide ligands, I identified all pairs of sequences that differ by a single amino-acid mutation and calculated their respective $\Delta \Delta G$ values. Pairs of sequences with the same sequence difference were grouped. For groups with greater than 100 occurrences, a chi-squared test was used to determine whether the population deviated from Gaussianity with significance p-value $<0.05$ with Bonferroni correction. Under the assumption of a fixed binding mode and a context-independent contribution of each residue to binding, $\Delta \Delta G$ for a given single point mutation is expected to be constant with Gaussian noise regardless of the context in which the mutation occurs. Significant deviation from Gaussianity
implies epistasis.

**Double point mutation analysis**

Double mutant cycles were identified by finding pairs of sequences that differed by exactly two amino acid mutations. A pair was retained only if both of the component single point mutants also existed in the dataset. The sets of 4 sequences were arranged alphabetically and unique sequence sets were retained. One standard deviation of propagated error was calculated from the sum of four Gaussians with 0 mean and 0.26 standard deviation using standard propagation of uncertainty theory. The value of 0.26 was calculated from the RMSE of the mapped SORTCERY mean affinity coordinate in kcal/mol and the measured standards.

**2.3.4 Regression modeling to relate sequence to $\Delta G_{binding}$**

**Support Vector Regression**

Support vector regression (SVR) models were trained to predict SORTCERY-measured affinity from protein sequence. SVR solves for a predictor $(\hat{y}_i)$ that has at most $\epsilon$ deviation from the observed value $(y_i)$: $(\hat{y}_i - y_i < \epsilon)$. The predictor is equal to the dot product of the weight vector $w$ and the encoded input $x_i$ plus an intercept $b$: $\hat{y}_i = w^T x_i + b$. As this constrained problem is not always feasible, slack variables $(\zeta_i)$ are introduced for each data point and minimized. This results in the primal form of the SVR regression problem, which balances model complexity and performance as described below.

$$
\begin{align*}
\min_{w,b,\zeta,\xi^*} & \frac{1}{2} w^T w + C \sum_{i=1}^{n} (\zeta_i + \xi_i^*) \\
\text{subject to} & \quad y_i - w^T \phi(x_i) - b \leq \epsilon + \zeta_i , \\
& \quad w^T \phi(x_i) + b - y_i \leq \epsilon + \xi_i^* , \\
& \quad \zeta_i, \xi_i^* \geq 0, i = 1, ..., n
\end{align*}
$$

(2.4)
The $\epsilon$ parameter defines a range of insensitivity to noise, and the $C$ parameter defines the cost of adding slack to the model. $C$ can be interpreted as a scalar that varies the complexity of the model. Smaller $C$ allows the model to be simpler by permitting more slack on the model optimization. The $\epsilon$ parameter and $C$ parameter for the final models were identified via nested cross-validation. SVR was used as implemented in SciKit-Learn.

Model encodings

**Amino acid encoding** The amino-acid encoding is a naive categorical encoder that represents every position as a vector of 20 binary indicators that capture which amino acid is present at that position. The entire peptide is encoded as a vector of length $N \times 20$ where $N$ is the length of the peptide (19 or 22 amino residues long in this work). For a peptide of length 22, as for Bcl-2-peptide interactions, the resulting encoding is a binary vector of length 440, with 22 non-zero values.

**11 factor encoding** The 11-Factor encoder used here is described by Liu et al. [67]. This encoder represents each position as a vector of 11 numerical values that describe physicochemical properties of the amino acid residues. This feature encoding has the potential to be more generalizable because it does not require the amino acid to be observed to score it. Rather, scores for unobserved residues can be determined based on the physiochemical similarities/differences with observed residues. For each residue, the vector assigns a value for steric parameter, hydrogen bond donors, hydrophobicity scale, hydrophilicity scale, average accessible surface area, van der Waals parameter R0, van der Waals parameter epsilon, free energy of solution in water, average side chain orientation angle, polarity, and isoelectric point. The resulting vector is of length $N \times 11$, where $N$ is the length of the peptide.

**5 factor encoding** A five-factor encoding was applied as described by Venkatarajan et al. Each residue is represented by a 5-dimensional vector derived from principal coordinate analysis of 237 physical-chemical properties [109].
Figure 2-17: Nested Cross-Validation schematic - (see Methods). The figure describes 3-fold nested cross-validation.

**Kernels**

We tested a linear kernel $x^T z$ and a second order polynomial kernel $(x^T z)^2$ in our modeling. In this context, $x$ and $z$ are vectors that encode the sequence. This primal optimization problem is solved in the dual form by the construction of the Lagrangian objective function [97].

**Nested cross-validation**

A generalizability score was determined for linear and second-order models trained on each dataset as shown in Figure 2-8 via 10-fold nested cross-validation. Each input dataset was split into 10, top-level subsets. Each top-level subset (1/10 of data) was used as a validation dataset ($V$) that reported the performance of models trained on the remaining 90% of data ($Tr$).

The models trained on the dataset $Tr$ retrieve their hyperparameters from ten-fold cross-validation within the dataset $Tr$. Each dataset $Tr$ was partitioned in 10 parts, where each part ($Te$) was used to test models trained on the remaining 90% of data ($S$). A grid of hyperparameters was evaluated ($\epsilon$ and $C$ ) on each $Te$.
and the best performing hyperparameters informed the models trained on Tr. The grid of parameters explored $2^{-14} - 2^6$ for C and epsilon values from 0 to 1, in 0.05 intervals. The hyperparameters that returned the best average performance across all Te datasets were used in training with Tr. The performance of models trained on Tr and evaluated on V report an estimate of generalizability. The best hyperparameters for the final models were selected by the parameters which performed the best on the Te datasets. (The V dataset is only used to report generalizability.) A schematic of nested cross-validation is shown Figure 2-17. The SVR parameters used for each model are shown in Table 2.10.

### Cross-validation

Ten-fold cross-validation was applied to calculate the average Spearman ρ score for each model. The full dataset was split into ten partitions for testing. For each test
set, a model was trained on the remaining 90% of the data. The $\rho$ correlation of the predicted value to the observed value was averaged across all 10 test sets. The parameters used for training the models were selected from nested cross-validation.

**Extrapolation across novel affinity space**

The pilot_xlr dataset was stratified into five sets, A-E, from tightest to weakest. Non-resolvable weak peptide sequences were assigned to set E. The resolvable peptide sequences were equally partitioned into 4 equal sets. Models were trained on each partition and tasked to discriminate the tight vs. weak peptide sequences given a combination of partitions.

### 2.3.5 Alternative models

**Position specific scoring matrices (PSSM)**

Position Specific Scoring Matrices (PSSM) are models that score sequences as a linear sum of residue/position weights. The weights for scoring peptides are stored in a 20 x N matrix ($M$) where the 20 rows represent the residue identity ($r$), and $N$ is the length of the peptide. In this work, weights for scoring matrix $M$ were taken from experimental SPOT array experiments or derived from frequencies in specific multiple-sequence alignments, as previously described (also, see below)[29]. Peptide sequences are scored as a linear sum of weights, or logarithms of weights, as shown below. $I$ is the identity function that identifies whether or not the residue/position pair is present in the peptide.

$$\text{Score(peptide)} = \sum_{n \in N} \sum_{r \in \{A...Y\}} -\log(M_{n,r})I(\text{peptide}_{n,r}) \quad (2.5)$$

**PSSM$_{freq}$** The weights for the frequency based PSSM used for benchmarking are derived from the tightest 30% of peptide sequences of the test set. Each matrix cell describes the probability that a given residue occurs at the specified position. Pseudocounts were added to be able to score all residues. $C$ is the function that
counts how many times residue r occurs at position n.

\[ M(n, r) = \frac{C(n, r) + 1}{\sum_{AA} (C(n, r) + 1)} \]  \hspace{1cm} (2.6)

**PSSM_{spot}**  Peptide SPOT arrays are assays used to study protein binding interactions. Peptides are synthesized on membranes where each spot represents thousands of copies of a specific peptide. Protein binding can be assayed by incubating the spots with the respective tagged target, washing, and measuring the fluorescence intensity of the tagged target with antibodies. Previous SPOT array experiments in the Keating lab tested the effect of single point mutations at 10 sites in the wild-type Bim BH3 sequence. We applied the SPOT arrays from DeBartolo et al. to score and rank each test set [25]. In the equation below, \( F \) is a function of the intensity of the observed spot fluorescence for a given position \( (n) \) and residue \( (r) \) on the spot array [30].

\[ M(n, r) = \frac{F(n, r)}{F(WT)_{avg}} \]  \hspace{1cm} (2.7)

**Blosum matrices**

A Blosum62 scoring matrix was used to evaluate how well distance from high-affinity native interaction partners correlates with affinity. I compared each sequence to its the Bim BH3 to score each sequence.

\[ \text{Score}(SEQ) = - \left( \sum_{i=1}^{n} Blosum_{62}(WT_i, SEQ_i) \right) \]  \hspace{1cm} (2.8)

**STATIUM**

STATIUM is a structure-specific statistical scoring function that scores peptide binding to a known template based on the frequency with which specific residue interactions are found in the PDB [25]. STATIUM models were constructed for the structures 3FDL (Bcl-x\(_L\) bound to Bim BH3), 3MQP (BfFl-1 bound to Noxa BH3), and 3PK1 (Mcl-1 bound to Bax BH3)
FoldX

FoldX is an empirical force field for scoring protein stability [92]. It generates mutant structures using a probability-based rotamer library, and scores complexes using empirically weighted physical terms. We applied FoldX to generate structures for all sequences observed in pilot_x1r. The sequences were modeled on 3FDL (Bcl-xL), 3MQP (Bfl-1), and 3PK1 (Mcl-1).

2.4 Discussion

In this chapter, I described how raw SORTCERY datasets were processed to produce quantitative measurements of affinities for thousands of cell-surface interactions. Through experimental data collection and curation efforts, we have provided the largest public Bcl-2 paralog protein-peptide affinity datasets available. These datasets can be used to improve our understanding of the determinants of Bcl-2 family protein binding, to model Bcl-2 family specificity, and to benchmark models.

The new wealth of data helps us generalize and refine our modeling assumptions. Most statistical models assume that multiple mutations make additive contributions to affinity. However, non-additivity can have significant effects on protein binding [113] [47] [44]. Identifying when multiple mutations sum independently or non-independently can have significant implication in modeling and design. By analyzing $\Delta \Delta G$ values calculated from high-throughput SORTCERY experiments, I found evidence of non-additivity for BH3 peptide binding. Subsequent experiments confirmed that at least one double mutant cycle demonstrated non-additivity when the component peptides were analyzed individually. I showed that by capturing second-order cooperative terms in second-order polynomial kernels, the higher-order models exhibited better generalizability than the models that made linear additive assumptions. The identification and capture of non-additivity demonstrated the potential of combining high-throughput data and computational modeling.

Below, I discuss the rationale behind the modeling choices made and discuss some caveats of our experimental approach.
To model binding affinity with the data collected from SORTCERY, regression models were applied to learn the sequence-to-affinity relationship empirically. One advantage of learning from sequence data is that these models do not require a crystal structure. Sequence-based models provide convenience and flexibility to rapidly score millions of sequences since models need not search through structural space before scoring. However, modeling in sequence space requires that an adequate encoding and kernel are selected to capture the complexity of the system without overfitting.

One aspect of the modeling that I investigated is the sequence encoding. One might expect models that capture the physicochemical nature of different amino acids to perform better since observations made for one residue can help inform scoring of others. In my analyses, reduced residue encodings did not offer any performance benefit. These encodings may obscure the real signal, by obfuscating slight nuances in residue differences. This could be true if the reduced features miss some crucial aspect of a particular residue, for instance, by insufficiently distinguishing the steric of differently shaped sidechains (e.g., leucine vs. isoleucine). Also, our datasets are large relative to what is required for a linear amino-acid encoding, which reduces the benefits observed when grouping residues.

An amino-acid encoding has several modeling advantages. First, this encoding is readily interpretable. Model weights can be easily visualized to understand the modeling decisions made because there exists a direct relationship to the input data. In addition, this encoding is useful because weights can easily be used in protein design protocols, including those that are based on integer linear programming as described in Chapter 3. Reduced encodings, such as the 5-factor and 11-factor encoding that I tested, have complex mappings to the 20 amino acids, which makes sequence optimization difficult.

One limitation of regression models that use amino-acid encoding is the lack of predictive power over unsampled residues. Unobserved residues at a given position are unweighted, which limits models to scoring sequences that are in the same sequence space. One solution is to vary more sites and residues in the combinatorial library. However, there is a tradeoff between increasing the diversity of the library and
obtaining adequate coverage of combinations of residues. Sparsely sampling an extensive combinatorial library makes it harder to observe local changes in the landscape. Given experimental limitations in the number of cells collected, sampling from more diverse sequence space could additionally decrease the chances of finding binders.

When modeling with a data-driven approach, one critical concern is overfitting. Overfitting occurs when an overly complicated model is used to describe the data — including noise in the data — instead of capturing the actual underlying relationship between sequence and binding affinity. A hallmark of overfitting is a higher performance on training datasets but decreased performance on unobserved datasets (because weights are too specific for a given training dataset). For a more representative performance score, it is necessary to tune model hyperparameters to adjust the model complexity. I addressed overfitting by applying nested cross-validation to select for parameters that balance model complexity and performance.

Testing on alternative datasets is another way of evaluating extrapolation, but the tests are not always easy to interpret. It is necessary to understand how the alternative datasets differ from the input training space, before accepting the performance metrics. For example, in benchmarking against peptides exclusive to each model’s replicate dataset, the scores were highly variable. Although the two datasets are experimental replicates, the performance differences may come from inconsistencies in the dynamic range of affinities captured. A shift in the FACS gates during collection could inflate the number of weak binders for one replicate, which changes the difficulty of the prediction task.

Often with knowledge-based models, there are inherent limitations. The knowledge-based potentials that I built in this work are not generalizable outside of the Bcl-2 family. Models in this study have been tuned to predict affinity for each receptor in the Bcl-2 family of proteins and are relevant only for predicting interactions with these proteins. Furthermore, the benchmarking results from testing Bcl-xL models between library spaces suggest that the models have limited extrapolation potential to new sequence spaces. These model biases are critical to consider when benchmarking the models and interpreting the model performance.
One limitation of SORTCERY is the limited dynamic range of affinities that can be measured. Specifically, at both the high and low extremes of a relatively narrow range, the mean-affinity coordinate values no longer correlate linearly with the free energy of binding. Furthermore, for some clones, the maximum signal is close to or beyond the dynamic range of the experiment, making the affinity unresolvable (yet providing an upper or lower bound). We observed that low-affinity peptides have less reproducible mean affinity coordinates from experiment to experiment, and the Spearman correlation of the replicate datasets typically improves after removing unresolved weak binders (see Figure 2-19). The datasets collected for Bcl-xL affinity using 1 nM and 100 nM receptor concentrations give us an opportunity to expand SORTCERY affinity beyond its limited dynamic range. However, this is still under investigation (see Figure 2-18). Poor resolution at the extremes of the measurable affinity range makes it difficult to combine datasets measured at multiple concentrations.

Another limitation of SORTCERY is the need to measure protein standards to map the average affinity coordinates to binding free energy. Currently, identifying the affinity of individual one-at-a-time peptides is a time-consuming process. A potential future modification is to perform SORTCERY at many receptor concentrations to generate binding curves as done in Tite-Seq [1]. The Conclusion discusses several additional ways to improve SORTCERY.

The development of accurate affinity models can help further understand and solve many protein-protein interaction engineering problems. Scoring functions can be applied to identify new protein-protein interactions in the proteome, design new, tight, and specific binders; predict evolutionary sequence relationships, and provide insight into the mechanisms of binding. For example, Hui et al. applied a PDZ classifier to scan the proteome to predict new protein-protein interactions [50]. Having accurate models is even more important when comparing evolutionarily related receptors. Due to similar binding preferences, it is more difficult to target specific members of a family of proteins. With more accurate specialized models to describe each receptor binding preferences, we can gain new insight as to how each receptor interacts with its target. The analysis in this chapter sets the preliminary work critical for application
in Chapter 3 to design for specificity.
Figure 2-18: Comparison of SORTCERY average affinity coordinates (\( \bar{A} \)) for peptides assayed at 100 nM and 1 nM.
Figure 2-19: Comparison of ranks - Unresolved vs. resolved subsets. For each of the 5 replicate datasets, ranks or raw values are plotted for each pair. In the left panel, the replicate ranks are compared. In the middle panel, the mapped energies are compared. In the right panel, the ranks of a subset of the data (called the resolvable set) are compared. Un-resolvable weak binders demonstrate lower replicate reproducibility as indicated by the spread near the high ranking peptides.
Chapter 3

Mapping sequence landscape to design specificity within protein families

Collaborator notes:
J. Jenson collected the SORTCERY data for the specificity library and performed the titration measurements for the designed sequences
L. Stretz performed the titration measurements for the peptide standards and prepared the DNA for deep sequencing
3.1 Introduction

Evolutionary gene duplication and mutation have created paralogous families of structurally conserved domains. Paralogous protein domains frequently bind to similar yet distinct motifs and exhibit individual specificity profiles. Knowing the binding partners of a protein is critical for determining its function and for predicting the consequences of disrupting its interactions. From a therapeutic perspective, it can be useful to disrupt the interactions of just one family member, e.g., to block a specific aberrant pathway without disrupting the function of other family members. Given the similarities of sequence and structure between paralogs, designing peptides that bind specifically to a single family member can be a difficult task.

Bcl-2 family proteins are a good model system in which to study protein-protein interaction specificity. Five anti-apoptotic Bcl-2 family proteins are often overexpressed in cancers. Bcl-2 family proteins sequester BH3 motifs in pro-apoptotic proteins which, when unbound, can trigger cell death. The binding behavior of anti-apoptotic Bcl-2 proteins constitutes an apoptotic blockade, which can render chemotherapeutics that work by triggering apoptosis ineffective. Molecules that bind anti-apoptotic Bcl-2 proteins competitively and block BH3 engagement can overcome such apoptotic blockades, and thereby promote cell death. Engineering tight and selective inhibitors to block specific anti-apoptotic proteins is an attractive approach to cancer therapy. Many pharmaceutical companies are pursuing the design of small molecules with this function. Paralog-selective inhibitors are desirable because they can avoid unnecessarily altering normal biological functions. For example, small molecule ABT-263 inhibits both Bcl-2 and Bcl-xL and is effective in inducing death of cancer cells. However, inhibition of Bcl-xL results in dose-limiting thrombocytopenia, an undesired side effect.

An alternative to the challenge of identifying small molecules to inhibit Bcl-2 protein-protein interactions is to design peptides that can mimic how nature performs this task. Peptides have some liabilities as therapeutics, but advances in cell delivery foreshadow significant future promise. The design of tight binding and selective pep-
tide inhibitors also faces obstacles. In particular, the number of potential peptidic inhibitors is vast, which necessitates developing strategies to guide the search of the sequence space. We need protein design methodologies to discover novel therapeutic options.

Protein design aims to identify rare functional sequences in a vast universe of possibilities. Systematic mutagenesis can provide insights to guide the search, but assays of protein interactions have traditionally been slow and expensive, limiting the amount of data available [25]. To design novel function, we need information about large sequence spaces, and we need to capture this in models that can then be used to guide design.

Simple position weight matrices (PWMs) derived from point-mutation surveys of sequence space are one example of a model that can be applied to design. Rational design with data from PDZ spot arrays achieved the successful design of novel "super binders" with increased affinity [115]. However, such modeling approaches make strong assumptions. PWMs assume a conserved binding mode in which residues contribute independently to binding. Structure-based, pseudo-physical models such as Rosetta can relieve this assumption and have advantages in that they provide a general approach. However, these methods require accurate modeling of structure, which is not always feasible. The construction of homology models for proteins for which structure is not available can be inaccurate. Rosetta computational models have been applied to design proteins that bind specifically to Bcl-2 family proteins. Berger et al. designed Bcl-2 family specific inhibitors by designing mutations in a three-helix bundle scaffold that contained a native BH3 hot spot residues scaffold to make receptor-specific contacts within and outside of the conserved BH3 binding cleft [7].

The need to find tight binders has led to the development of combinatorial library screening technologies capable of screening libraries of up $10^{10}$ variants [87]. These assays can identify receptor-specific peptides by applying rounds of screening of surface or phage-displayed libraries with increasing stringency to get the desired function. For example, Dutta et al., Foight et al., and Jenson et al. applied yeast
display libraries and rational mutagenesis to design Bcl-xL, Mcl-1, and Bfl-1 specific peptides, respectively [31] [37] [52].

Recently it has become possible to adapt library screens to provide information about the binding behavior of thousands of members [40]. Example screens include quantitative enrichment studies, SORTCERY, and Tite-seq [81] [1]. We can capitalize on increasing amounts of quantitative data to find new receptor specific binders by taking a more systematic approach to model the data, and to use the resulting models to extrapolate into new sequence spaces.

The work in this chapter lays the groundwork for an approach to capitalize on the increased availability of binding data. I describe how I surveyed the Bcl-2 family binding sequence space, built empirical models, and designed new sequences in unexplored sequence space. My work thus provides an example of a new approach, with a complete cycle of data collection, model building and design that led to the discovery of novel peptide sequences with desired binding profiles.
Figure 3-1: Three computationally designed libraries targeting Mcl-1 (red), Bfl-1 (green) and Bcl-xL (blue) were synthesized in the context of Bim and Puma BH3 scaffolds. In the left-most panel, "X" symbols indicate the positions that were varied in each library. The six libraries were expressed on the yeast cell surface, cell populations were pooled, and a subset of $10^4$ clones was evaluated for binding to the three receptors via a 12 gate FACS scheme. Deep sequencing and computational analysis were applied to reconstruct the individual peptide profiles across the 12 gates and determine mean affinity coordinates for each peptide. Standards were measured to calibrate the mean affinity coordinate ($A$) of each peptide to binding free energy in kcal/mol.
3.2 Results

The interaction analyses and peptide designs that I report here utilize data from a high-throughput experiment in which a set of BH3-like peptides was assayed for binding to the three paralogous proteins Bcl-xL, Mcl-1, and Bfl-1. The sequences of screened peptides came from studies by Jenson et al., who used structure-based modeling and previous experimental data to design three combinatorial libraries. Each of these libraries consisted of $\sim 10^7$ members and was intended to enrich for sequences selective for just one of Bcl-xL, Mcl-1 or Bfl-1 [52]. In that work, a computational optimization procedure was applied to choose degenerate codons that introduced diversity in a targeted way at 7-8 amino-acid positions of two different native BH3 sequences: Bim and Puma. As input into the experiments described here, we generated a pooled library of $\sim 10^4$ clones drawn from the three libraries. To restrict library size while retaining binders with a range of affinities for each receptor, we used FACS to partition each library into 12 sub-pools according to the affinity for the intended target. We then mixed clones from different pools in pre-determined ratios and took $\sim 10^4$ clones from this population as the input.

3.2.1 SORTCERY provides affinities for peptide library members binding to Mcl-1, Bcl-xL, and Bfl-1

The pooled library was FACS sorted twice against each of the three target receptors, Bcl-xL, Mcl-1, and Bfl-1 at 1 nM, 1 nM, and 100 nM, respectively, to produce six high-quality SORTCERY datasets as described in Chapter 2 (see Table 2.4). We collected data for Bfl-1 binding at a higher receptor concentration than for Bcl-xL or Mcl-1 because the binding signal was low at 1 nM. After computational filtering, we identified between 1292 and 3489 unique peptide probability distribution functions over gates (PDFs) in the six experiments. In total, the assay generated high-quality data for 5769 unique peptides. For each peptide, we calculated a mean-affinity coordinate ($\bar{A}$) from the PDF, equal to the sum of the weighted gate indices (Figure 3-1 and Chapter 2). Peptides observed in repeated experiments had highly reproducible
Standards can be used to map SORTCERY mean-affinity coordinates to free energies of binding

Theory predicts that SORTCERY ($\bar{A}$) will linearly correlate with binding free energies over a certain resolution range under certain conditions [81]. To test and calibrate this relationship, cells displaying peptide standards spanning the SORTCERY dynamic range were individually titrated with each of the three receptors to determine quantitative binding affinities (see Appendix for titration curves). Measured ($\bar{A}$) correlated with individually measured binding free energies with Pearson r between 0.81 and 0.92. I used this observation to map ($\bar{A}$) to binding free energies in kcal/mol (see Chapter 2). Linear fits gave RMSE between 0.34-0.56 kcal/mol over a range of dissociation constants from 0.07 nM up to 250 nM (Figure 3-2) The majority of peptides measured had binding free energies around -10 kcal/mol ($K_d = \sim 50$ nM).

3.2.2 Multiple-receptor SORTCERY experiments can map the specificity landscape of peptide ligands

Quantitative analysis of receptor specific binding affinities reflect receptor similarities

I compared how similar Mcl-1, Bfl-1, and Bcl-xL receptor binding preferences are by correlating the binding free energy of peptides in the intersection of each receptor-specific dataset (see Figure 3-3). Mcl-1 and Bfl-1 receptors have the highest binding similarity for the peptides tested (Pearson r = 0.47-0.55). This is consistent with the observation that Bfl-1 shares 38% binding-groove sequence identity with Mcl-1 but only 28% binding-groove identity with Bcl-xL [52] [35]. However, Mcl-1 and Bcl-xL receptors have a lower correlation (Pearson r = 0.15-0.34) despite having 39% binding-groove similarity.
Figure 3-2: Experimentally determined receptor binding energies for 2679, 1292, or 3480 peptides binding to Bcl-xL, Bfl-1, or Mcl-1 were identified, respectively. A histogram of the binding free energies shows that the peptides have cell-surface affinities between -14 and -9 kcal/mol. More Mcl-1 tight binders were identified than tight Bfl-1 or Bcl-xL binders. Experiments to generate the data underlying these distributions were performed by J. Jenson.

Figure 3-3: Pairwise correlations (given as Pearson r) of the 6 experimental datasets demonstrate high replicate reproducibility and binding preference similarities/differences between receptors. The receptors that share the greatest similarity in binding profiles are Mcl-1 and Bfl-1, which contrasts with the low similarity observed between the binding profiles of Mcl-1 and Bcl-xL.
Figure 3-4: Plotting the binding specificity landscape in 2D. The binding energies for 420 sequences with experimentally measured affinities for Mcl-1, Bfl-1, and Bcl-xL can be plotted in a trisected plane to visualize the specificity space. Each peptide is described by 3 vectors, each projecting the negative binding free energy for Mcl-1, Bfl-1, or Bcl-xL, from the origin (0,0) toward one of three corners of an equilateral triangle. The vectors are summed to produce a coordinate in 2D space that quantifies the peptide binding preference for the three receptors. This projection removes the absolute quantification of affinity and emphasizes paralog binding selectivity. The plot shows that our quantified binders included more Mcl-1-specific sequences than Bfl-1 or Bcl-xL-specific sequences.

Quantification of receptor specific energy measurements provides a view of the library specificity landscape

To examine peptide specificity preferences, I investigated the 420 peptides with measured affinities for Mcl-1, Bfl-1, and Bcl-xL. To illustrate the binding selectivity landscape, I plotted each peptide as a coordinate in a trisected plane. Peptides selective for Mcl-1 occurred with higher frequency over peptides selective for Bfl-1 and Bcl-xL. Multiple Mcl-1 selective peptides exist in this set of tested sequences with over 100-fold selectivity. However, few peptides exist with >10-fold selectivity for Bfl-1 or Bcl-xL (see Figure 3-4).
3.2.3 Modeling can integrate data to extrapolate to new sequence spaces

Mapping a peptide sequence on the specificity landscape map requires the peptide-receptor binding affinity for all three target receptors. However, given that each sorting experiment randomly sampled the pooled library, the number of peptides that have measured affinities for all three receptors is much smaller than the number of peptides in any individual dataset. To map more of the binding landscape, I used Mcl-1, Bfl-1, and Bcl-xL affinity models of the type described in Chapter 2 to predict binding free energy from protein sequence.

Regression modeling of sequence-to-binding free energy

As described in Chapter 2, to model the relationship between peptide sequence and binding free energy for each receptor, I applied support vector regression to train interpretable additive models [97]. I expressed the binding free energy as a sum of the contributions of each residue and each residue pair (see Figure 3-5). I encoded peptide sequences with an amino acid binary encoding that maps a given sequence into a vector of binary indicators. An amino-acid encoding encodes a 22mer peptide as a vector of length 440 where each binary indicator describes whether one of 20 amino acid residues is present at a specified position. The encoding can further describe all residue pair combinations by explicitly expanding the vector to describe all pair terms. For the same 22mer peptide, the expanded vector describes 440 single term indicators and 96,580 pair term indicators. A regression problem of this magnitude can be solved by applying support vector machines with a second order polynomial kernel.

Modeling with support vector regression is advantageous in two ways. First, polynomial kernels provide an efficient way to optimize models that include dependencies between parameters. Second, weights for single-residue and residue-pair terms can be used in subsequent constrained optimization problems. I took advantage of this to explore the Pareto frontier of specificity, examine the relationship of specificity and
For variants of each of the six receptor-specific datasets, I applied nested cross-validation to determine the best hyperparameters for modeling (see Methods). Hyperparameters $c$ and $\epsilon$ vary the model complexity to account for noise and overfitting. After grid searching the parameter space, the best dataset-specific models report an average $r^2$ between 0.55 and 0.73. I trained the final models for prediction with the complete datasets and the hyperparameters identified from nested cross-validation. The three best models are second-order polynomial models trained on the x1, f100, and m1r datasets discussed in Chapter 2 (see Methods).

Expanding the Pareto frontier of selectivity by integrating library mutations

Given receptor-specific affinity models, I broadened the coverage of the specificity landscape beyond the set of sequences with affinity measurements for all three receptors (420). I did this by applying models that predict the affinity for all unique sequences observed (5769), all input library sequences ($10^7$), and beyond, to what I call the integrated library, the combinatorial sequence space formed by integrating mutations from the six libraries ($10^{14}$). To illustrate the expansion of the specificity
landscape beyond input library space, I plotted the x,y coordinate extremes for all peptides in the input library. I further solved for sequences in the integrated library that maximized the extremes of selectivity and plotted the coordinates (see Methods). The plotted boundaries show that selectivity is predicted to be as much as tenfold higher in the integrated space than in the strict input space (see Figure 3-6). Modeling the space predicted modest opportunity to increase Mcl-1 specificity in comparison to Bfl-1 or Bcl-xL target specificity.

**Tradeoffs of affinity and selectivity - high-affinity Bfl-1 peptides are predicted to be cross-reactive with Mcl-1**

To further investigate how receptor selectivity varies with peptide affinity, I quantified fold selectivity as a function of binding affinity for each pair of receptors (see Methods). I observed receptor-dependent tradeoffs of affinity and specificity. For those peptides predicted to bind tightest to Mcl-1, I observed high selectivity against Bcl-xL and Bfl-1. Similarly, peptides predicted to bind most tightly to Bcl-xL had high selectivity against Bfl-1 and Mcl-1. Interestingly, peptides predicted to bind Bfl-1 tightly were predicted to be selective against Bcl-xL, but have extensive cross-selectivity with Mcl-1 (see Figure 3-7).

### 3.2.4 Design of monospecific and bispecific peptides by solving for optima in the sequence-energy landscape

**Designing sequences for targeted specificity with ILP**

To test our ability to traverse the specificity landscape with our models, we tasked ourselves to produce six sets of peptides with select specificity characteristics. We aimed to design three sets of monospecific peptides that only bind Mcl-1, Bfl-1, or Bcl-xL, and three sets of bispecific peptides that bind two out of three receptors, Mcl-1 and Bcl-xL, Mcl-1 and Bfl-1, or Bcl-xL and Bfl-1.
Figure 3-6: Extrapolating into the specificity landscape. Receptor-specific affinity models were used to explore specificity space. Black points plot the specificity coordinates of the 420 sequences with experimentally measured Mcl-1, Bfl-1, and Bcl-xL binding energies. Orange points plot the specificity coordinates of all peptides observed to bind at least one receptor, using the model to predict the coordinates. Models can also define extremes in the specificity space for all theoretical library members, including those that were not observed experimentally (Teal). A further extrapolation can map the predicted boundaries of specificity for an integrated library space that includes all substitutions that were sampled in any of the six original input libraries (Red). See Methods for details.
Figure 3-7: Specificity and affinity tradeoffs for BH3 peptides binding to Mcl-1, Bfl-1, and Bcl-xL. All 27,696,384 peptides in the designed library were computationally scored against Mcl-1, Bfl-1, and Bcl-xL. For a given target receptor, all peptides were binned by the predicted target affinity. The median fold selectivity against each off-target ($K_d$ off-target / $K_d$ target ) is plotted for each bin. The shaded fill indicates the 25th and 75th percentiles. This analysis predicts that high-affinity binders of Bfl-1 will also interact tightly with Mcl-1.
Figure 3-8: Designed peptides plotted in the specificity landscape. Selective peptides were designed to bind to Mcl-1 (Red), Bfl-1 (Green), or Bcl-xL (Blue), without cross-reacting with the other two receptors. Dual selective peptides were designed to interact with Mcl-1/Bfl-1 (Yellow), Bcl-xL/Bfl-1 (Aqua), or Mcl-1/Bcl-xL, without binding tightly to the off-target receptor. The sequences logos are composed of the sequences that were tested experimentally.
To design novel monospecific and bispecific peptides, I applied constrained optimization to identify candidate peptide sequences that satisfied select specificity objectives. Weight terms were extracted from the best models and applied to solve constrained integer linear programming (ILP) problems. ILP was iteratively applied to design up to 200 sequences for each specificity objective (see Methods).

For the three monospecific design objectives, I optimized the peptides to bind the target receptor maximally. The designed sequences maximized affinity to the target receptor, while constrained to bind tighter than 10 nM to the target and weaker than \( \sim 100 \) nM to the off-target receptors. From the 200 sequences designed for each set, I selected 12 peptides for each receptor to be experimentally validated via cell surface binding experiments (see Methods). See Table 3.1 for the peptide sequences.

For the bispecific designs, I optimized peptides to maximize target binding affinity, maximize specificity, or minimize off-target binding affinity. I optimized the bispecific Mcl-1/Bfl-1 peptides for maximal binding selectivity, i.e., the maximum difference in binding affinity between the targets and the off-target (see Methods). I generated 200 candidate peptide sequences and selected seven for experimental validation (see below, and Methods). See Table 3.2 for the peptide sequences.

The bispecific peptides for Bcl-x\(_L\)/Bfl-1 and Mcl-1/Bcl-x\(_L\) were designed in two rounds. In the first round, I optimized the peptides to minimize the off-target affinity while constrained to bind each target tighter than 50 nM and bind to the off-target weaker than 580 nM. I generated 200 candidate sequences for each set and selected seven for experimental validation. In the second round of design, I optimized the peptides to maximize the target affinity while constrained to bind both targets tighter than 5 nM and bind to the off-target weaker than \( \sim 46 \) nM. (see Methods). I generated 200 candidate sequences for each set, and selected 11 bispecific Bcl-x\(_L\)/Bfl-1 and 10 Mcl-1/Bcl-x\(_L\) designs for experimental validation. To distinguish these two rounds of bispecific designs, the peptide names in Table 3.2 are suffixed with 'on' or 'off' to indicate optimization for the highest on-target affinity or the minimum off-target affinity.

The peptides selected for experimental validation were those that demonstrated
<table>
<thead>
<tr>
<th>Label</th>
<th>Seq</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>GRRVRIAQGLRAGDQLDAYG</td>
</tr>
<tr>
<td>F2</td>
<td>GQRVRIAQGLRAGDQLDAYG</td>
</tr>
<tr>
<td>F3</td>
<td>GRRVHIAAGLRTGDQLEAQG</td>
</tr>
<tr>
<td>F4</td>
<td>GQRVHIAAGLRTGDQLEAYG</td>
</tr>
<tr>
<td>F5</td>
<td>GQRVVHIAAGLRTGDQUEAQG</td>
</tr>
<tr>
<td>F6</td>
<td>GQRVVQIAAAGLRTGDQLEKYG</td>
</tr>
<tr>
<td>F7</td>
<td>GQRVVQIAAAGLRTGDQLEKQG</td>
</tr>
<tr>
<td>F8</td>
<td>GRRVQIAAGLRTGDQLEAQG</td>
</tr>
<tr>
<td>F9</td>
<td>GRRVRIAQGLRAGDQLDXYG</td>
</tr>
<tr>
<td>F10</td>
<td>GRRVQIAAGLRTGDQLEKYG</td>
</tr>
<tr>
<td>F11</td>
<td>GQRVVQIAQGLRAGDQLEKYG</td>
</tr>
<tr>
<td>F12</td>
<td>GRRVQIAQGLRAGDQLEKQG</td>
</tr>
<tr>
<td>X1</td>
<td>GQTLIWYGSLRRADEFAKQR</td>
</tr>
<tr>
<td>X2</td>
<td>GQTLIWYGSLRRADEFAKQR</td>
</tr>
<tr>
<td>X3</td>
<td>GQPLIWFAGSLRRADEFAKQR</td>
</tr>
<tr>
<td>X4</td>
<td>GQTLIWYGSLRRADEFAKQR</td>
</tr>
<tr>
<td>X5</td>
<td>GQTAIWYGSLRRADEFAKQR</td>
</tr>
<tr>
<td>X6</td>
<td>GQSLIWFAGSLRRADEFAAQR</td>
</tr>
<tr>
<td>X7</td>
<td>GQPLIWFAGSLRRADEFAAQR</td>
</tr>
<tr>
<td>X8</td>
<td>GQSMIWFAGSLRRADEFAKQR</td>
</tr>
<tr>
<td>X9</td>
<td>GQTLIWYGSLRRADEFAAQR</td>
</tr>
<tr>
<td>X10</td>
<td>GQSLIWFAGSLRRADEFAAQR</td>
</tr>
<tr>
<td>X11</td>
<td>GQTLIWYGSLRRADEFAAQR</td>
</tr>
<tr>
<td>X12</td>
<td>GQSLIWFAGSLRRADEFAAQR</td>
</tr>
<tr>
<td>M1</td>
<td>GRSELEVQVELRIGDIVVAYF</td>
</tr>
<tr>
<td>M2</td>
<td>GRSELEVQVELRIGDIVVAYF</td>
</tr>
<tr>
<td>M3</td>
<td>GRSELEVQVELRIGDIVVAYF</td>
</tr>
<tr>
<td>M4</td>
<td>GRSELEVQVELRIGDIVVAYF</td>
</tr>
<tr>
<td>M5</td>
<td>GRSELEVQVELRIGDIVVAYF</td>
</tr>
<tr>
<td>M6</td>
<td>GRSELEVQVELRIGDIVVAYF</td>
</tr>
<tr>
<td>M7</td>
<td>GRSELEVQVELRIGDIVVAYF</td>
</tr>
<tr>
<td>M8</td>
<td>GRSELEVQVELRIGDIVVAYF</td>
</tr>
<tr>
<td>M9</td>
<td>GRSELEVQVELRIGDIVVAYF</td>
</tr>
<tr>
<td>M10</td>
<td>GRSELEVQVELRIGDIVVAYF</td>
</tr>
<tr>
<td>M11</td>
<td>GRSELEVQVELRIGDIVVAYF</td>
</tr>
<tr>
<td>M12</td>
<td>GRSELEVQVELRIGDIVVAYF</td>
</tr>
</tbody>
</table>

Table 3.1: Monospecific peptides designed
<table>
<thead>
<tr>
<th>Label</th>
<th>Seq</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF1</td>
<td>GRRIDEIAQILRRIGDHIEKYI</td>
</tr>
<tr>
<td>MF2</td>
<td>GRWIDQIAQFLRRIGDHIEKYI</td>
</tr>
<tr>
<td>MF3</td>
<td>GRRVDEIAQILRRIGDNIEEYI</td>
</tr>
<tr>
<td>MF4</td>
<td>GRRVDEIAQILRRIGDNINEYI</td>
</tr>
<tr>
<td>MF5</td>
<td>GRRIDEIAQILRRIGDHVEKYI</td>
</tr>
<tr>
<td>MF6</td>
<td>GRRVDEIAQILRRIGDNVTTYI</td>
</tr>
<tr>
<td>MF7</td>
<td>GRRVDEIAQILRRIGDQIEEYI</td>
</tr>
<tr>
<td>XF1.off</td>
<td>GRRVVGQQLKRLADEYHKA</td>
</tr>
<tr>
<td>XF2.off</td>
<td>GRREVWSQLKRIADQFQKYL</td>
</tr>
<tr>
<td>XF3.off</td>
<td>GREEWLSQYKLRIALDFQKYL</td>
</tr>
<tr>
<td>XF4.off</td>
<td>GREEWLSQYKLRIADMFQKYL</td>
</tr>
<tr>
<td>XF5.off</td>
<td>GREEWLSQYKLRIAHLDFQKYL</td>
</tr>
<tr>
<td>XF6.off</td>
<td>GQRVDDFGQGLKRVADEYHQA</td>
</tr>
<tr>
<td>XF7.off</td>
<td>GRRVWLSQYKLRIADQFQTYL</td>
</tr>
<tr>
<td>MX1.off</td>
<td>GRSQIYWQELVGGDNHAYR</td>
</tr>
<tr>
<td>MX2.off</td>
<td>GRSQIYWQFLVSGVNAAYR</td>
</tr>
<tr>
<td>MX3.off</td>
<td>GRSQIYDFQVLRSGDENAAYR</td>
</tr>
<tr>
<td>MX4.off</td>
<td>GRSQIYDFQVLRYYADYNAAYR</td>
</tr>
<tr>
<td>MX5.off</td>
<td>GRSQIYDFQVLRYGDVNAAYR</td>
</tr>
<tr>
<td>MX6.off</td>
<td>GRSQIYWQELVSGDNHAYR</td>
</tr>
<tr>
<td>MX7.off</td>
<td>GRSQIYWQELVSGDNHAYR</td>
</tr>
<tr>
<td>XF01.on</td>
<td>GQRLIWSAGLRRLADEFDKQA</td>
</tr>
<tr>
<td>XF02.on</td>
<td>GQRIIWIAEELRAADELDKI</td>
</tr>
<tr>
<td>XF03.on</td>
<td>GQRIIWIAEELRADAQLDAQI</td>
</tr>
<tr>
<td>XF04.on</td>
<td>GQRIIWIAEELRLADELDKV</td>
</tr>
<tr>
<td>XF05.on</td>
<td>GQRIIWIAEELRADAQDLKQY</td>
</tr>
<tr>
<td>XF06.on</td>
<td>GQRIIWIAEELRRAADQLKQY</td>
</tr>
<tr>
<td>XF07.on</td>
<td>GQALIWIGAEELRRADEFNKQL</td>
</tr>
<tr>
<td>XF08.on</td>
<td>GQRLIWAELRRLADEFDKQL</td>
</tr>
<tr>
<td>XF09.on</td>
<td>GQPLIWAELRRLADEFNKQV</td>
</tr>
<tr>
<td>XF10.on</td>
<td>GQRLIWAELRRLADDFDKQY</td>
</tr>
<tr>
<td>XF11.on</td>
<td>GQRLIWAELRRLADEFNKQA</td>
</tr>
<tr>
<td>MX01.on</td>
<td>GQWLHWAELRIADEFAQY</td>
</tr>
<tr>
<td>MX02.on</td>
<td>GQWLHYWAELVRIADDFLAAQR</td>
</tr>
<tr>
<td>MX03.on</td>
<td>GQSLIWFIALRIADEFAQY</td>
</tr>
<tr>
<td>MX04.on</td>
<td>GQWLWYIAELRIADEFAQY</td>
</tr>
<tr>
<td>MX05.on</td>
<td>GQWLWYIALRIADEFAQY</td>
</tr>
<tr>
<td>MX06.on</td>
<td>GQWLWYIALRRADEFAQY</td>
</tr>
<tr>
<td>MX07.on</td>
<td>GQWLWYWAELRRADEFAQY</td>
</tr>
<tr>
<td>MX08.on</td>
<td>GQWLWYIAELRADDFAQY</td>
</tr>
<tr>
<td>MX09.on</td>
<td>GQWLWYIAELRADDFAQY</td>
</tr>
<tr>
<td>MX10.on</td>
<td>GQSLIWFIALRRADEFAQY</td>
</tr>
</tbody>
</table>

Table 3.2: Bispecific peptides designed
the best scores according to an average of 6 different models (see Methods). Peptides in the final set of designs were 4-10 mutations away from any previously observed sequence (Figure 3-9). Sequence logos built from the designed sequences illustrate the sequence composition changes across the specificity landscape (see Figure 3-8). For example, Bcl-xL specific peptides prefer Ala at position 3E whereas Bfl-1 and Mcl-1 binders prefer a Gly at 3E. Also, the designs included residues that have previously been observed to confer specificity such as 4A_V for Mcl-1 specificity [29].

**FACS analysis of designed peptides validates designs**

To test whether the designed sequences have the desired specificity, we evaluated the peptides for binding using cell-surface display read out using FACS at four receptor concentrations (1000 nM, 100 nM, 10 nM and 1 nM). The median binding fluorescence signal of each clonal population was measured to estimate the receptor specific binding affinity (see Figure 3-10 and Figures 3-11—3-18 ). 36 of 36 monospecific peptides demonstrated the designed selectivity for the target receptor. The target binding signal for the designs are high, which suggest tight affinity and the off-target binding signal for the designs are low, which suggest weak affinity.

Of the bispecific designs, 15 of the 42 peptides expressed the desired specificity. The seven Mcl-1/Bfl-1 bispecific peptides demonstrated high binding signal to the target receptors, and weak binding signal to the off-target.

The Bcl-xL/ Bfl-1, and Bcl-xL/Mcl-1 bispecific peptides demonstrated variable success. For peptides that minimized off-target binding affinity, the designed peptides did not bind to both of the intended targets. The Bcl-xL/ Bfl-1 bispecific designs only bound Bcl-xL, and the Bcl-xL/Mcl-1 bispecific designs did not bind any of the three receptors.

The designed bispecific peptides that maximized target binding affinity yielded success. In this set of designs, 3 of the 11 bispecific Bcl-xL/Bfl-1 designs and 5 of the 10 Mcl-1/Bcl-xL bispecific designs demonstrated selectivity for the intended targets (see Figure 3-10).
Figure 3-11: Cell surface titration - Bcl-xL specific designs - Designed clones were verified for their specificity preferences via sparse titrations at 1000 nM, 100 nM, 10 nM, and 1 nM. The median binding signal of the expressing population is presented. Red, green, and blue bars correspond to the receptors Mcl-1, Bfl-1, and Bcl-xL respectively.
Mcl-1 specific replicate 1

Figure 3-12: Cell surface titration - Mcl-1 specific designs - Designed clones were verified for their specificity preferences via sparse titrations at 1000 nM, 100 nM, 10 nM, and 1 nM. The median binding signal of the expressing population is presented. Red, green, and blue bars correspond to the receptors Mcl-1, Bfl-1, and Bcl-xL respectively.
Bfl-1 specific replicate 1

Bfl-1 specific replicate 2

Figure 3-13: Cell surface titration - Bfl-1 specific designs - Designed clones were verified for their specificity preferences via sparse titrations at 1000 nM, 100 nM, 10 nM, and 1 nM. The median binding signal of the expressing population is presented. Red, green, and blue bars correspond to the receptors Mcl-1, Bfl-1, and Bcl-xL respectively.
Figure 3-14: Cell surface titration - Mcl-1 and Bfl-1 bispecific designs - Designed clones were verified for their specificity preferences via sparse titrations at 1000 nM, 100 nM, 10 nM, and 1 nM. The median binding signal of the expressing population is presented. Red, green, and blue bars correspond to the receptors Mcl-1, Bfl-1, and Bcl-x\(_L\) respectively.

Figure 3-15: Cell surface titration - Mcl-1 and Bcl-x\(_L\) bispecific designs (off) - Designed clones were verified for their specificity preferences via sparse titrations at 1000 nM, 100 nM, 10 nM, and 1 nM. The median binding signal of the expressing population is presented. Red, green, and blue bars correspond to the receptors Mcl-1, Bfl-1, and Bcl-x\(_L\) respectively.
Figure 3-16: Cell surface titration - Bcl-xL and Bfl-1 bispecific designs (off)- Designed clones were verified for their specificity preferences via sparse titrations at 1000 nM, 100 nM, 10 nM, and 1 nM. The median binding signal of the expressing population is presented. Red, green, and blue bars correspond to the receptors Mcl-1, Bfl-1, and Bcl-xL respectively.
Figure 3-17: Cell surface titration - Mcl-1 and Bcl-xL bispecific designs (on) - Designed clones were verified for their specificity preferences via sparse titrations at 1000 nM, 100 nM, 10 nM, and 1 nM. The median binding signal of the expressing population is presented. Red, green, and blue bars correspond to the receptors Mcl-1, Bfl-1, and Bcl-xL respectively.
Figure 3-18: Cell surface titration - Bcl-xL and Bfl-1 bispecific designs (on)- Designed clones were verified for their specificity preferences via sparse titrations at 1000 nM, 100 nM, 10 nM, and 1 nM. The median binding signal of the expressing population is presented. Red, green, and blue bars correspond to the receptors Mcl-1, Bfl-1, and Bcl-xL respectively.
3.3 Methods

3.3.1 Modeling sequence to binding energy

The linear and second-order polynomial models used in this chapter are trained similarly as described in Chapter 2. These models only differ in the input dataset such that the datasets used in this analysis came from an earlier iteration of data curation. The datasets are within 5% of the dataset sizes described in Chapter 2. See Chapter 2 for details on support vector regression.

Encoding protein sequences as vectors

For an additive regression model, a $n$ length sequence $x$ can be encoded as a binary vector, such that each element of the vector reports if the amino acid indexed by $j$ is present at position $i$. I refer to this simple binary encoding as $\phi(x)$. $([x:1, x:2, ..., x:i:j, ..., x:n:19, x:n:20])$ In a second order regression model, the vector must additionally describe all position_residue pair terms. I can extend the vector with additional binary indicators to report on the presence of the position_residue pairs. Each pair term can be referenced by subscripts $i, j, k,$ and $l$, such that $i$ and $k$ reference the positions and $j$ and $l$ reference the residues in the pair. The binary indicator for an independent position_residue term is notated $x_{i:j}$. The binary indicator for a pair of position_residue terms is notated $x_{i:j:k:l}$. The expanded vector, notated as $\Phi(x)$, is the concatenation of all independent and pair position_residue term indicators. ($[x:1,..., x:i:j,..., x:n:20, x:1:1:1,..., x:i:j:k:l,..., x:n:20:n:20]$)

Receptor-specific functions

The predicted binding energy of a sequence is equal to the dot product of the receptor-specific weights $(w)$ and the one-dimensional binary vector $\Phi(x)$, plus constant $b$. The dot product is equivalent to the sum of all binary weights and pairwise weights present in the sequence $x$. 

123
\[ y(x) = w^T \Phi(x) + b = \sum_{i=1}^{n} \sum_{j=1}^{20} \sum_{k=i+1}^{n} \sum_{l=1}^{20} x_{i:j:k:l} + \sum_{i=1}^{n} \sum_{j=1}^{20} \sum_{k=i+1}^{n} \sum_{l=1}^{20} x_{i:j:k:l}w_{i:j:k:l} + \sum_{i=1}^{n} \sum_{j=1}^{20} \sum_{k=i+1}^{n} \sum_{l=1}^{20} x_{i:j:k:l}w_{i:j:k:l} + b \]  

(3.1)

The best receptor-specific scoring models for Bcl-x\textsubscript{L}, Mcl-1, and Bfl-1 were the polynomial models trained from the x1, m1r, and f100 datasets. I refer to these receptor specific energy functions as \( X(x) \), \( M(x) \), and \( F(x) \) below, where \( x \) is a peptide sequence.

### 3.3.2 Extracting pairwise weights from second-order polynomial models

Each receptor-specific energy function takes on the same form, and differs only by the input weights. The weights are required to parameterize the scoring function, but the number of parameters in the polynomial model makes it difficult to solve for the weights in primal form. Support vector machines can solve for higher-order relationships by applying kernels to avoid the explicit mapping of the higher-order terms. I applied a second-order polynomial kernel with binary encoded sequences \( \phi(x) \) to avoid the explicit mapping of the sequence \( x \) to \( \Phi(x) \). Support vector regression models as implemented in Scikit-learn solve the objective function in dual form and return the dual coefficient as a solution to the fitting problem. To extract out the pairwise component weights that correspond to \( \Phi(x) \) from the dual coefficient, I applied the following function.

\[ w = \sum_{t=1}^{m} \alpha_t \Phi(x_t) \]  

(3.2)

This equation tells us that the weight vector \( w \) is equal to the weighted sum of each training data point. \( \Phi(x_t) \) is the expanded polynomial vector of a given sequence \( t \). The dual coefficient \( \alpha_t \) reports on the weighted contribution of the data point to the weight vector [97].
3.3.3 Design with integer linear programming

I applied ILP to solve for the vector \( \Phi(x) \) that maximizes or minimizes the receptor-specific scoring function shown above. I pursued six design objectives with integer linear programming (ILP). Monospecific peptides were designed to bind to Bcl-xL, Mcl-1, or Bfl-1. Bispecific peptides were designed to bind Bcl-xL and Mcl-1, Bcl-xL and Bfl-1, or Mcl-1 and Bfl-1. These peptides were designed to bind selectively to their intended receptors by maximizing the target binding affinity, maximizing the specificity gap, or by minimizing the off-target binding affinity while imposing constraints on target and off-target binding affinities. For each set of designs, I iteratively identified 200 sequence candidates with ILP, from which I selected 7-12 sequences for experimental validation.

**General constraints for designing peptides using ILP in polynomial space**

Without constraints, ILP will include any combination of single and pair terms that maximize the scoring function. To design a feasible peptide, I imposed constraints to capture the formal requirements of a peptide [56]. For example, a peptide must only have one residue at each position. Also, if a position-residue pair term is included, then the component position-residue terms must also be included. The equations below formalize the constraints applied to enforce design of a valid peptide.

\[
\begin{align*}
\text{min or max} \quad & \sum_{i=1}^{n} \sum_{j=1}^{20} x_{i;j} w_{i;j} + \\
& \sum_{i=1}^{n} \sum_{j=1}^{20} \sum_{k=i+1}^{n} \sum_{l=1}^{20} x_{i;j;k;l} w_{i;j;k;l} + \\
& \sum_{i=1}^{n} \sum_{j=1}^{20} \sum_{k=i}^{n} \sum_{l=j+1}^{20} x_{i;j;k;l} w_{i;j;k;l} + b \\
\text{subject to }& x_{i;j} \in (0, 1) \\
& x_{i;j;k;l} \in (0, 1) \\
& \forall i \in (1..n) \sum_{j=1}^{20} x_{i;j} = 1 \\
& \forall i \in (1..n) \forall k \in (i..n) \forall j \in (1..20) \quad (-1)x_{i;j} + \sum_{l=1}^{20} x_{i;j;k;l} = 0 \\
& \forall i \in (1..n) \forall k \in (1..n) \forall l \in (1..20) \quad (-1)x_{i;j} + \sum_{j=1}^{20} x_{i;j;k;l} = 0
\end{align*}
\]
Library constraints added to improve confidence

To improve the likelihood that the new designs would succeed, I added constraints on which residues were allowed in the designs. I restricted the designed sequences to only include residues in the input library space used for SORTCERY experiments (see Chapter 2) and excluded all Cys residues. At positions that were not varied in the libraries used for model development, design was allowed to select residues from either the Bim or Puma wild-type sequences, which expanded the potential sequence space to the order of $10^{14}$. I also added constraints to prevent inclusion of residues with low confidence weights: The sequence space was restricted to allow only position_residue terms that were observed at least 25 times in all three training datasets. The models are prone to overfit rare examples, thus, adding constraints to require that a given residue appear often helps to prevent designing in spaces with low confidence.

Designing monospecific peptides (maximizing target affinity)

To design monospecific peptides, I solved for sequences that maximized binding affinity to the target receptor (minimized binding energy). Constraints were imposed to require that the designs bind the target with $\Delta G \leq -10.9$ kcal/mol (10 nM) and bind the two off targets with $\Delta G \geq -9.5$ kcal/mol ($K_d$ greater than 107.9 nM). The off-target affinity boundary was selected to be near the mean binding energy of the non-resolvable weak binders. The equations below show the constraints applied for the design of a Bcl-xL receptor-specific peptide. The peptide feasibility constraints described above were also applied in all design calculations.
\[
\begin{align*}
\min & \quad X(x) \\
\text{subject to} & \quad F(x) > -9.5 \quad \text{Bfl-1 off-target constraint} \\
& \quad M(x) > -9.5 \quad \text{Mcl-1 off-target constraint} \\
& \quad X(x) < -10.9085 \quad \text{Bcl-xL target constraint} \\
& \quad \sum (x_{i,j} \notin S) = 0 \quad S \text{ is the set of position-residue terms w/ } \geq 25 \text{ counts - Frequency constraint} \\
& \quad \sum (x_{i,j} \notin L) = 0 \quad L \text{ is the set of position-residue terms in the library - Library constraint}
\end{align*}
\]

(3.4)

I applied the ILP optimization routine to iteratively select for the 200 best candidate peptide sequences for each receptor. At each iteration of optimization, the best solution was identified, \( x_s \), and added as constraint \( (x \neq x_s) \) such that the next iteration was forced to return new sequence. To select the peptide sequences to be experimentally tested, I applied a consensus-based approach by rescoring all 200 candidates using six receptor-specific SVR models. I trained six SVR models with linear or polynomial kernels on three different sets of experimental data. The three datasets describe the receptor-specific data from replicate 1, replicate 2, or a union of replicate 1 and replicate 2. I averaged the output of all six models first, and selected the twelve sequences with the highest average specificity for experimental validation. The average specificity was calculated as the mean of the predicted binding energy differences between the target and off-targets.

**Designing bispecific peptides (maximizing specificity)**

To design bispecific peptides that bind to Mcl-1 and Bfl-1, without binding tightly to Bcl-xL, I solved for the top 200 sequences that maximized the gap between Bfl-1 and Bcl-xL binding energies. (This is equivalent to finding a sequence that maximizes the difference of the Bcl-xL and Bfl-1 model weights). Constraints were added to require that the Mcl-1 and Bfl-1 binding energies differ by at most 0.2 kcal/mol. No
additional binding energy constraints were applied to the target or off-target binding energies.

\[
\begin{align*}
\text{max} & \quad X(x) - F(x) \\
\text{subject to} & \quad F(x) - M(x) \leq 0.2 \quad \text{Target similarity constraint} \\
& \quad F(x) - M(x) \geq -0.2 \quad \text{Target similarity constraint} \\
& \quad \sum (x_{ij} \notin S) = 0 \quad S \text{ is the set of position residue terms with } \geq 25 \text{ counts} - \text{Frequency constraint} \\
& \quad \sum (x_{ij} \notin L) = 0 \quad L \text{ is the set of position residue terms in the library} - \text{Library constraint} \\
\end{align*}
\]

(3.5)

To select the set of sequences to be experimentally tested, I re-scored all designs using the six SVR models for each of the three receptors as described above. I picked the seven bispecific sequences with the highest average specificity for experimental validation.

Designing bispecific peptides (minimizing off-target affinity)

For the bispecific designs targeting Bcl-xL and Bfl-1 (with specificity against Mcl-1) or the bispecific designs targeting Mcl-1 and Bcl-xL (with specificity against Bfl-1), I solved for the top 200 sequences that minimized the off-target affinity. Constraints were added to require tight binding to the two targets, \( \Delta G \leq -10 \text{ kcal/mol} \), and weak binding to the off-target, \( \Delta G \geq -8.5 \text{ kcal/mol} \).

The following equations show the constraints for designing Bcl-xL and Mcl-1 bispecific peptides.
\[
\begin{align*}
\text{max} & \quad F(x) \\
\text{subject to} & \quad F(x) > -8.5 & \text{Bfl-1 off-target constraint} \\
& \quad M(x) < -10 & \text{Mcl-1 target constraint} \\
& \quad X(x) < -10 & \text{Bcl-xL target constraint} \\
& \sum (x_{ij} \notin S) = 0 & S \text{ is the set of position\_residue terms w/ } \geq 25 \text{ counts -- Frequency constraint} \\
& \sum (x_{ij} \notin L) = 0 & L \text{ is the set of position\_residue terms in the library -- Library constraint}
\end{align*}
\] (3.6)

To select the set of sequences to be experimentally tested, I re-scored all designs using the six SVR models for each of the three receptors as described above. For each of the bispecific objectives, I picked the seven bispecific sequences with the highest average specificity for experimental validation.

**Designing bispecific peptides (maximizing target affinity)**

For the second round of bispecific peptide design, I added additional residue constraints, optimized target binding as an objective function (rather than maximizing the specificity gap or minimizing the stability of the off-target), and added requirements, as described below, to select for a more diverse set of sequences.

To prevent the potential inclusion of destabilizing residues in the designs, I limited the sequence space to include only those residues observed in tight binding peptides for the target (peptides with SORTCERY $\Delta G < -10.5 \text{ kcal/mol}$). I added constraints to require that the predicted binding energy for each of the two targets be less than -11.3 kcal/mol. This boundary was chosen because it was the minimal predicted binding affinity among the successful mono-receptor specific sequences. Furthermore, I chose to minimize the binding energy of the targets instead of maximizing the off-target binding energy. Given that ILP can only optimize one objective at a time, I solved two optimization problems for each bispecific objective, one for each target...
receptor. To enforce specificity, I imposed a constraint on the off-target binding energy: \( \Delta G \geq -10 \text{ kcal/mol} \). The constraints for this optimization, using Bcl-xL as the optimization target, are shown below.

\[
\begin{align*}
\text{min} & \quad X(x) \\
\text{subject to} & \quad F(x) > -10 \quad \text{Bfl-1 off-target constraint} \\
& \quad M(x) < -11.3 \quad \text{Mcl-1 target constraint} \\
& \quad X(x) < -11.3 \quad \text{Bcl-xL target constraint} \\
& \quad \sum (x_{ij} \notin S) = 0 \quad S \text{ is the set of position_residue terms with } \geq 25 \text{ counts - Frequency constraint} \\
& \quad \sum (x_{ij} \notin L) = 0 \quad L \text{ is the set of position_residue terms in the library - Library constraint} \\
& \quad \sum (x_{ij} \notin B) = 0 \quad B \text{ is the set of position_residue terms that can be found in tight binders - Residue constraint (3.7)}
\end{align*}
\]

To select the ten best Bcl-xL and Bfl-1 bispecific peptides and ten best Mcl-1 and Bcl-xL bispecific sequences, I applied the consensus-based method to compute average scores for each design binding to each receptor. I selected the top 5 sequences, based on affinity for one of the targets, that had at least two mutations from higher ranking designs.

### 3.3.4 Plotting the coordinates of peptide specificity

**Plotting the binding specificity landscape in 2D**

The binding energies for sequences with affinity values for Mcl-1, Bfl-1, and Bcl-xL were plotted in a trisected plane to visualize the specificity space. Each peptide is described by 3 vectors, each projecting the negative binding free energy for Mcl-1, Bfl-1, or Bcl-xL, from the origin \((0,0)\) toward one of three corners of an equilateral triangle. The Mcl-1 vector extends from \((0,0)\) to \((0,1)\). The Bfl-1 vector extends from \((0,0)\) to \((-0.5, \sin(120))\). The Bcl-xL vector extends from \((0,0)\) to \((-0.5, -\sin(120))\). The
vectors are summed to produce a coordinate in 2D space that quantifies the peptide binding preference for the three receptors.

**Input library boundary**

The input library is the set of all possible sequences that could exist in six libraries designed by Jenson et al. (see Table 2.2) [52]. I identified the boundary of specificity for the input library by estimating a solution to the concave hull problem for the full set of 27,696,384 library members. After mapping each sequence to an x,y coordinate, I estimated the shape by plotting the minimum and maximum y values for the set of sequences binned by the x values rounded to the nearest hundredth, and vice versa.

**Integrated library boundary**

The integrated library is the set of all possible sequences that could be created by mixing mutations from all 6 libraries in Jenson et al. [52]. I calculated the boundary of specificity for the integrated library space via ILP. For any given peptide, the binding affinity for the three receptors determines the angular and radial coordinate. At any given coordinate, there are two target receptors and an off-target receptor. The off-target receptor is the receptor with the lowest binding affinity. The energy gap between the tighter of the two target receptors and the off-target receptor determines the radial position. The affinity ratio between the two target binders determines the angular coordinate. For every pair of target receptors, Mcl-1/Bcl-x\textsubscript{L}, Mcl-1/Bfl-1, Bcl-x\textsubscript{L}/Bfl-1, I fixed the energy gap between the two receptors as a constant between -6.8 and 6.8 kcal/mol and solved for the sequence with the lowest off-target affinity via ILP. This optimization problem identifies a set of sequences that define a boundary of specificity. Points plotted on the boundary were additionally constrained by the requirement to have 25 observations per receptor.
3.3.5 Plotting specificity vs. affinity tradeoffs

All 27,696,384 peptides in the input library were computationally scored against Mcl-1, Bfl-1, and Bcl-xL using the same second-order polynomial models used for design. For a given target receptor, all peptides were binned by the predicted target affinity. The median affinity of the off-target is plotted for each bin.

3.3.6 Experimental evaluation of the computational designs

Binding of the designs to the three Bcl-2 proteins was tested using yeast-surface display at protein concentrations of 1000 nM, 100 nM, 10 nM, and 1 nM [52]. The median binding signal of the binding population was recorded for each clone and applied to estimate an approximate affinity. For calculations below, the binding signal is normalized by subtracting out the binding signal measured with no receptor.

Each peptide was assigned to one of 5 categories that reports an estimated affinity to the specified receptor. Categories A-E describe the estimated affinities in descending order. For each peptide-receptor pair, a half-max binding signal was calculated from the 4 titration measurements. The half-max signal is equal to the max of the normalized binding signals, divided by 2. The categorical affinity that was assigned to a peptide-receptor interaction reflects the lowest receptor concentration at which the normalized binding signal exceeds the half-max binding signal. For example, an interaction was assigned to Category A if the normalized binding signal at 1 nM, was greater than the half-max signal. A peptide-interaction was assigned to category E if the max binding signal was less than half of the binding signal of Bim at 1000 nM receptor concentration.

3.4 Discussion

Anyone tasked with designing a 22mer selective peptide for either Bcl-xL, Mcl-1, or Bfl-1 faces an astronomically large search space of $20^{22}$ (4.2x$10^{28}$). The challenge of
identifying a desired rare variant in this search space is immense, but in this chapter, we demonstrated a synergistic computational and combinatorial library approach to map the binding specificity landscape of the Bcl-2 family of proteins.

Prior efforts to design selective peptides for Bcl-x_L, Mcl-1, and Bfl-1 used combinatorial library screening, sometimes guided by a computational model, to identify a handful of selective peptides [31] [37] [52]. In this work, we designed and validated a dozen novel selective peptides for each receptor that were very different from any previously observed peptide. We anticipate that we could use a similar approach to design additional selective peptides in the future.

The models that guided our design were built using new high-resolution data generated by the SORTCERY assay. This high-resolution FACS assay was applied to quantitatively rank the binding affinity of a sample of ~10,000 cells from a pool of mixed libraries. From this pool of cells, we identified ~5000 unique peptides that we mapped to an affinity (in kcal/mol) for at least one of the three receptors. With this data in hand, we applied support vector regression to learn receptor specific models that relate peptide sequence to binding affinity.

One highlight of our modeling approach was the incorporation of cooperative pairwise features. Prior work suggested that this might be important for modeling BH3 peptide binding to Bcl-2 family proteins. For example, Jenson et al. demonstrated that there is a context dependence to the effect of substitutions made in Bim vs. Puma BH3 peptides. Specificity inducing mutations in the Puma BH3 context did not provide specificity in the Bim BH3 context [52]. DeBartolo et al. reported modest a correlation between the mutational effects of point residue changes in Bim vs. Noxa BH3 peptides context for Mcl-1 binding, and Bim vs. Bad context for Bcl-x_L binding [25]. Furthermore, our nested cross-validation tests showed us that second-order models have higher predictive performance than the first-order models.

The models that we trained are parameterized linear functions that can be analyzed and used for design in conjunction with integer linear programming. Most efforts to design selectivity apply small changes to a binding peptide that is already known to bind to the target of interest [37]. In contrast, our approach to design did
not start with a single known binder, but mixed mutations from different scaffolds and different libraries, guided by a computational scoring function. Our successes designing monospecific and bispecific sequences, represent significant progress in protein engineering using a novel approach.

Given the success of our single-receptor specific peptide designs, we challenged ourselves to generate bispecific peptides and were successful in designing several peptides that bind to Mcl-1 and Bfl-1 but not Bcl-xL. Such selective peptides have potential as novel therapeutics. For example, lymphoma cell lines exposed to the Bcl-2 antagonist ABT-737 upregulate both Mcl-1 and Bfl-1 [116]. It is also known that inhibiting Bcl-xL can lead to thrombocytopenia. Thus, hitting Mcl-1 and Bfl-1 without hitting Bcl-xL may be a viable approach to treating some cancers. In general, designing inhibitory peptides that are specific for one or two members of a protein family, but not all, may provide a route to treatments that avoid unnecessary side effects.

Although the design of Mcl-1/Bcl-xL and Bcl-xL/Bfl-1 bispecific peptides was unsuccessful at first, the second round of design yielded peptides that expressed the desired specificity attributes. In a second round of design, the peptides were optimized for specificity by maximizing on-target binding affinity rather than minimizing off-target affinity. This improvement suggests that to design for specificity, we should prioritize binding tightly to the target receptor before introducing mutations that destabilize off-target binding. By minimizing off-target affinity, we included more residues that achieve specificity by sacrificing on-target stability. Given that the unsuccessful peptides all failed to bind the targets, it is likely that the models underestimated the weights of the destabilizing residues. The FACS assay is limited in resolution by a lower limit where all weak binders are assigned the same binding value. Regression models trained on this data will underestimate the severity of including a destabilizing residue because of the capped range.

Although we achieved several notable successes in BH3 peptide engineering using this new approach, this effort to map out the Bcl-2 family specificity landscape required a substantial input of experimental work for a single family of interactions.
This approach for modeling protein-protein interactions lacks the generalizability of other structural and statistics based methods, but the resulting models achieve high accuracy in the limited sequence space. As the efficiency of high-throughput experimentation increases and cost decreases, this data-driven approach may become more readily applicable to other protein families.
Figure 3-9: Boxplot of the maximum sequence similarity of the designed peptides to any previously observed sequence. Designed peptides were all at least 4 mutations from any previously observed peptide measured by SORTCERY, and some differed in 10 positions from the closest characterized library member.
Figure 3-10: Experimental binding profiles of the computationally designed peptides interacting with Mcl-1 (M), Bfl-1 (F), or Bcl-xL (X), measured by yeast cell-surface sparse titrations. The peptide-receptor interaction affinity is categorized into affinity bins A-E (tight to weak). The designed monospecific peptides demonstrated binding to the intended targets. Bispecific Mcl-1/Bfl-1 (MF) binders selectively bound their intended dual targets. Bcl-xL/Bfl-1 (XF.off) and Mcl-1/Bcl-xL (MX.off) bispecific designs that were designed to minimize off-target binding only hit one of the intended targets, or no targets respectively. Redesigned XF.on and MX.on peptides, which were designed to optimize on-target binding affinity, yielded successful selective designs. Experiments performed by J. Jenson.
Chapter 4

Conclusions and future directions

Most of what we know about how protein sequence and structure control protein-protein interactions is obtained from low-throughput experiments and from models that approximate the physics of molecular associations. With recent technological developments in deep sequencing — which can combine with library display technologies to report on protein interactions and even protein-interaction affinities — we have new tools available that can be used to explore broader parts of the protein interaction universe. This technological advance offers opportunities to obtain a better understanding of the determinants of protein-protein binding and protein interaction specificity.

In this thesis, I presented computational elements of a pipeline for modeling and designing peptides that interact with members of the Bcl-2 family proteins. I contributed data curation methods for processing high-throughput experimental data, and I investigated the benefits of complementing combinatorial library screening with data-driven modeling. I showed that for the datasets that are now available, it is possible to go beyond simple models that treat all residue contributions independently. Also, I demonstrated examples of receptor-peptide interactions for which capturing cooperativity provides performance improvements.

Based on the combinatorial library screening data and models that I derived by using it, I successfully designed 36 Bcl-2 monospecific peptides for Bcl-xL, Mcl-1, and Bfl-1 (12 each). I also successfully designed seven bispecific peptides that bound
tightly to Mcl-1 and Bfl-1; three bispecific peptides that bound to Bcl-xL and Bfl-1; and eight bispecific peptides that bound to Mcl-1 and Bcl-xL without making strong interactions with the off-targets. These designed BH3 peptides may find application as therapeutics or as diagnostic tools [27].

In this chapter, I summarize the data that I will be contributing to the field, discuss ways to improve SORTCERY as an experimental assay and speculate on some of the future potentials of modeling with data from combinatorial library screens.

### 4.1 Application of collected data and designs

The data that I have curated consist of 10 datasets that demonstrate high replicate reproducibility and high correlation of affinity assignments with values determined by individual measurements made on peptide standards. These data describe the binding affinity of 11,061 unique peptides binding to Bcl-xL, Mcl-1, or Bfl-1 proteins (See Table 2.4).

Protein interaction datasets of this size and quality are rare but are valuable for developing and testing new modeling methods. The availability of these new datasets will benefit the computational structural biology community by serving as a benchmark that can be used to compare old and new approaches. Such data can be applied to test and further advance structure-based methods such as Rosetta [2]. In fact, we have already begun working with Frank DiMaio at the University of Washington to use the data from SORTCERY in improving parameterization of the Rosetta scoring function. In the future, I predict that there will be many opportunities to combine high-throughput experimental data with structural based modeling. Our Bcl-2 family data provides such an opportunity to explore what approaches are possible.

In addition to curated datasets, I designed novel tight and selective peptides that have potential therapeutic applications for treating chemoresistant cancers [74] [108]. Translation of Bcl-2-targeting peptides into the clinic will require delivery to the cytoplasm of cells, where Bcl-2 proteins function. This has been accomplished by introducing hydrocarbon crosslinks that stabilize alpha-helical structure and promote
cell uptake [60] [99] [83] [84]. Intracellular delivery can also be accomplished using delivery vehicles such as nanoparticles, which our lab investigates [14]. Bcl-2 paralog-selective peptides can also be applied in BH3 profiling to assess the dependency of primary cancer cells on specific Bcl-2 proteins for survival; an application that does not require that peptides transit the plasma membrane since the assay can be run using permeabilized cells [27]. With further work, the selective peptides in this study can also be modified using hydrocarbon stapling for improved pharmacological properties.

4.2 Improving the SORTCERY assay

Advances in surface-display, FACS, and deep sequencing technologies have led to the development of high-throughput assays like SORTCERY. However, there are still improvements that can be made to generate even more reliable data. One consideration when deep sequencing DNA from peptide-displaying cells is the fidelity of the DNA sequence. Protocols like SORTCERY depend on quantification of the DNA reads; thus sequence biases or experimental artifacts can affect the affinity measurements. In SORTCERY, sequence artifacts can arise at many steps and can come from experimental contamination, random mutation, PCR error or bias, or other problems.

Early in this project, I observed that many of the sequences identified in our earliest SORTCERY runs did not originate from the library sampled, but instead came from other libraries studied in the laboratory. We determined that this problem arose from PCR amplification of small amounts of contaminating DNA, and my labmates developed an experimental protocol to reduce this contamination such that it was not detectable in later runs. However, one way we protected against contamination corrupting the data in the earlier experiments was by having restricted libraries. This allowed us to filter for DNA sequences that came from our designed library as we intentionally encoded it. If the library were not constrained to a specific library composition, (i.e., designed with NNN/NNK codons), it would be difficult to distinguish whether or not a sequence originates from a sequencing error, growth mutation,
contamination, or from the degenerate codon library.

PCR bias is a concern for SORTCERY and assays like it because we assume that the quantitative count of the deep sequencing reads can be used to reconstruct the distribution of clones across the gates. If there is amplification bias or error, this mapping can be inaccurate. One solution to reduce DNA amplification bias is to introduce randomized barcodes in the first step of PCR [94] [58]. Randomized barcodes uniquely identify and tag input DNA sequences to track any downstream amplification bias. These unique identifiers could provide a better estimate of the actual number of cells displaying each sequence that pass through each gate, and thus give a more accurate report of affinity.

An artifact of the SORTCERY assay that came to light during this work is that clones occasionally show distinct saturating binding signals. SORTCERY assumes that all displayed peptides when fully bound, will give rise to the same saturated ratio of binding:expression signals. Reich et al. provided some experimental data to support this assumption [81]. However, as we examined a higher number of individual peptide standards as part of this work, we observed that some sequences do not reach the expected upper saturation limit. We have not yet determined why this is the case. It could be that peptides with fast dissociation kinetics give rise to lower signals because the bound receptor decreases upon washing. However, preliminary studies did not show a correlation between off-rates and low saturation signals. One way to more accurately measure affinities could be to build binding curves using SORTCERY by repeating the experiment at multiple receptor concentrations, as done by Adams et al. [1]. Using a broader range of receptor concentrations could also improve the dynamic range of affinities that can be measured using SORTCERY. Members of our laboratory are testing this approach, and the data-processing tools and scripts that I have generated will be useful for testing this extension of the method.

Although combinatorial library screening can provide a sizable amount of data compared to traditional low-throughput methods for measuring protein-protein interactions (e.g., \( \sim 10^4 \) data points compared to 10), the sequence space of potential binders is still well beyond what high-throughput experimental methods can fully
explore. Screening approaches must focus on sequence spaces that are enriched in binders to maximize the information gained from these techniques. One strategy we found useful was to screen multiple localized, distinct, sequence spaces. Strikingly — given that we have observed non-additive contributions between some combinations of residues — we were able to use data collected in multiple local sequence spaces to build a unified model. In our initial input libraries, the peptides originated from six individual libraries. Although each library covered a distinct localized space, we combined the observations into a single model and demonstrated success in navigating the mixed library space, as shown by our successful protein designs.

4.3 The future of modeling using large protein interaction datasets

In my thesis, I applied SVM regression to model the relationship between peptide sequence and binding affinity. One future improvement is to incorporate information about confidence levels of the input observation into the training of the model. Every peptide has a read count that reflects how many times the peptide occurs in the deep-sequencing dataset. Profiles that are determined by a higher number of reads are more likely to provide accurate estimates of affinity. Including this information in modeling could help models capture the signal over the noise.

More complex machine learning models can replace the SVM models used in this study. SVMs offer interpretability, but as datasets continue to grow, future modeling approaches may trade off interpretability for better performance. Models may expand in complexity by using higher order kernels, or by switching to alternative learning models such as neural networks. Although higher order models may offer performance benefits, they are more likely to overfit and are difficult to diagnose when they fail.

The problem that I tackled here could also potentially be addressed using a Bayesian approach. Bayesian models learn prior beliefs from external data and update the priors with every new observation. Including a prior can make a model
more generalizable, by providing information about residues that are unobserved in the training data. Input priors for modeling Bcl-2 family proteins could potentially be generated using mutational data collected using SPOT array experiments from Dutta et al. and DeBartaolo et al. [30] [25]. The challenge that arises when using Bayesian models is how to integrate all of the data into a cohesive model. Considering that data can have many formats and scales, e.g., coming from experiments that only resolve binders from non-binder data vs. those that provide quantitative affinity values, the challenge is how to normalize all of the datasets to combine the data into an integrated model.

Although data-driven models are tailored to make predictions for a given system, these approaches can combine with more general models for studying protein-protein interactions. Models like Rosetta and FoldX can synergize with empirical models to verify desired trends. For example, while designing the 36 receptor-specific peptides, we observed that FlexPepDock broadly agreed with our intended specificity. By applying a consensus-based approach in design, we can be more confident in our predictions before putting in the experimental resources required to validate the predictions.

4.4 Summary

The future of high-throughput binding affinity assays will bring "big data" into the field. In this thesis, I have demonstrated an example computational curation, modeling, and design pipeline for application to future datasets. However, the analysis performed in this work is not exhaustive, and there are potentially additional analyses on these datasets. (i.e., answer questions of whether there exist discrete binding motifs, or determine what attributes facilitate selectivity).

The methods and screening strategies investigated here can potentially be applied to study different domain-peptide interactions. In the Keating lab, there is ongoing work on applying combinatorial library screening to study TRAF domains, which bind to short linear motifs, with lower affinities than are observed for Bcl-2
family protein complexes. Using these methods to study protein families such as PDZ/SH2/SH3/WW could be an exciting new direction. In this thesis, I described methodology as it applies to predict binding, but similar assays and modeling techniques can also be used to study other functions, such as thermostability, cell permeabilization, immunogenicity, and more. It is with the combination of all of these methods that we can begin to chart a global roadmap for the future of sequence design.
Appendix A

Appendix
Figure A-1: Yeast titration of peptides selected for double mutant cycle analysis.
Figure A-2: Yeast titration of peptide standards used to calibrate Bfl-1 SORTCERY experiments.
Figure A-3: Yeast titration of peptide standards used to calibrate Bcl-xL SORCERY experiments.
Figure A-4: Yeast titration of peptide standards used to calibrate Mcl-1 SORCERY experiments.
Bibliography


157


[89] Deanne W. Sammond, Ziad M. Eletr, Carrie Purbeck, and Brian Kuhlman. Computational design of second-site suppressor mutations at protein-protein


