Computational Design of Protein Therapeutics with Reduced Immunogenicity through Structural Modeling of Protein Interactions

Chris King

A dissertation
submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Washington
2014

Reading Committee:
David Baker, Chair
Phil Bradley
Barry Stoddard

Program Authorized to Offer Degree:
Biochemistry

University of Washington

Abstract

Computational Design of Protein Therapeutics with Reduced Immunogenicity through Structural Modeling of Protein Interactions

Chris King

Chair of the Supervisory Committee:
Professor David Baker
Department of Biochemistry

Proteins possess huge potential as therapeutic agents for the control and modulation of human physiology. Protein interactions regulate most physiological processes, mediating the connection between atomic-scale self-assembly and macroscale health and disease. Natural proteins often display exquisite specificity and high affinity for molecular targets while avoiding detection and elimination by the host immune system. The design of synthetic, non-natural proteins to bind these molecular targets presents the opportunity to suppress, regulate, or enhance the cellular control processes underlying the physiology of both normal and disease states. Here, we demonstrate the development, application, and testing of computational protein design algorithms to predict protein-binding specificity, model the energetics of designed protein interactions, and reduce the immunogenicity of protein therapeutics.

Table of Contents

Chapter 1

Background: Biological Therapeutics and Computational Protein Design

Proteins as Drugs: Classification and Application.
Proteins represent the fastest-growing class of pharmaceuticals for a diverse range of clinical and biomedical applications. Since the early 1980s, proteins have played an increasing role in the development of many drugs, vaccines, and diagnostics [1]. These therapies typically fall into the category of antibodies, antibody Fc fusions, blood/clotting factors, growth factors, metabolic enzymes, interferons, hormones, and engineered protein scaffolds [2]. Protein drugs may simply replace or supplement a natural molecule that is missing or deficient, regulate a natural cellular process, induce a novel cellular process, or target molecular payloads to specific areas of the body, individual cells, or even sub-cellular compartments [3]. With over 130 protein drugs currently on the market [3] and many more in clinical development, these biologics are receiving increasingly greater focus in research and development to modify an increasingly broad range of biological processes.

Antibodies and Natural Proteins as Biological Therapeutics: Opportunities and Challenges
Antibody-based drugs represent by far the largest and, to-date, most successful class of protein therapeutics. Though nature has evolved a myriad of different molecular architectures uniquely suited to different chemical, cellular, and physiological contexts, antibodies represent over half of the biologics approved by the FDA in the last few years [5]. Though antibodies continue to show great clinical value, they are no panacea for all applications, as dysregulation of cellular pathways often involves a number of interconnected molecular entities, all of which may not all be readily targeted by antibodies due to the unique pharmacokinetic and pharmacodynamic profile displayed by this class of drug [6]. Natural ligands, receptors, and enzymes have also been developed as protein therapeutics, utilizing the preexisting binding or catalytic activity evolved by nature. Insulin and human growth hormone have shown particularly success in treating disease with little or no modifications to the wild-type protein [7,8]. However the diversity of protein architectures and functions offered by the human genome is limited in scope, such that modified human and nonhuman proteins have also been repurposed for use as molecular therapeutics. Though offering a wider range of structures and functions, nonhuman proteins present their own unique challenges, as they typically induce unfavorable immune responses upon systemic delivery, leading to host production of neutralizing anti-drug antibodies and even anaphylactic shock (Table 1). Nature does not always provide the tools necessary to effect new physiological states or restore ideal

cellular function in diseased biological systems, necessitating the development of new molecules to resolve pathologies untreatable with currently available drugs.

Immunogenic Protein	Therapeutic Application
L-asparaginase	enzyme replacement therapy
Streptokinase	thrombolysis in myocardial infarction
B-lactamase	targeted prodrug activation in antibody-conjugated enzymes
Carboxypeptidase G2	targeted prodrug activation in antibody-conjugated enzymes
Endotoxin A	targeted cell-death in recombinant immunotoxins
Ricin A Chain	targeted cell-death in recombinant immunotoxins
Erythropoietin	stimulation of erythrocyte production
CNTF	photoreceptor protection in macular degeneration
Flagellin FliC (CBLB502)	tumor suppression, radioprotective

Table 1. Protein therapeutics currently on the market or in active development that suffer from host immune reactions.

History of Protein Therapeutics, Current Challenges, and Future Directions

Insulin marked the first drug successfully created from recombinant DNA technology in 1982 [19]. Soon thereafter, other recombinant versions of human proteins were developed to replace their natural counterparts deficient in some patients or to augment natural pathways, such as human growth hormone, interferon-alpha, and erythropoietin [20]. Next, modified versions of these natural proteins became available. Injected insulin fails to mimic endogenous insulin due to slow action and rapid clearance from the blood. Mutant forms of insulin were developed to increase response time, prolong half-life in the body, and improve physiological profile [21]. For other therapeutics, glycosylation sites were added or removed to enhance therapeutic activity, as in the case of darbepoetin-α, a variant of epoetin-α with two additional glycosylation sites that extended serum half-life to three times that of its precursor [22]. Recombinant fusions of antibody Fc domains presented another technology for increasing serum half-life [23]. Proteins of less than 70 kDa in size are quickly eliminated from the bloodstream in minutes to hours, a fact that can render their therapeutic effects null. By fusing the Fc domain from antibodies onto protein drugs, half-life and stability can be extended greatly due to Fc-mediated interactions with the salvage receptor FcRn. Besides increasing half-life, Fc fusions may also increase solubility, aid expression and purification, and increase valency, a set of factors that led to the production of etancercept (Enbril®), a TNF receptor 2-Fc fusion that became the top-selling protein therapeutic in 2009 [23]. Antibodies represent the largest class of currently available biotherapeutics. Their rapid growth since the mid-1990s was made possible by chimerization and humanization methods [24] that ameliorated the immunogenicity problems of murine antibodies. Later, the development of transgenic mice to produce fully human antibodies [25] and production of antibody fragments in *E. coli* fermentation [26] further sped growth of this class. Though most Ab therapeutics take the form of

full-length native-like IgG molecules, FaB fragments and antibody-drug conjugates are gathering increasing attention as potent cancer therapies [23].

Though recombinant human proteins and monoclonal antibodies have experienced great success over the last 32 years, they are not without their shortcomings. First is their extremely high cost; protein production in specialized cell lines can be expensive and inefficient, especially when multi-gram treatment schedules of the drug in question are necessary. Druggable targets are also generally confined to extracellular molecules and cell surface receptors; intracellular delivery of large proteins, though possible in the lab, remains a largely open problem in the clinical domain. Delivery is not only an intracellular problem. Proteins are denatured and digested in the gut and usually require intravenous delivery, and even then, are prevented from entering the central nervous system by the blood-brain barrier. Additionally, the large size of many biologics, particularly antibodies, leads to inefficient tissue penetration for most tumors [27]. Lastly, protein immunogenicity remains a frequent problem even for exogenously delivered human proteins, leading to the production of neutralizing anti-drug antibodies (ADAs) and, even worse, the potential for the development of an adaptive immune response against similar endogenous proteins. This scenario was exemplified by administration of PEGylated TPO, where some patients developed a concomitant antibody response to endogenous TPO, leading to persistent anemia even after administration of the drug was ceased.

The great success of human biosimilars and monoclonal antibodies has led to increasing focus on a new class of drugs: engineered protein scaffolds. Like antibodies, these next-generation therapeutics may be engineered through rational design or high-throughput selection (e.g. phage display) to bind a wide range of molecular targets with high affinity and specificity, but are based on a diverse array of different template protein folds. This variability in the selection of source scaffold allows for the possibility of engineering small, monomeric protein drugs with high solubility and stability that lack the disulfide bonds, glycosylation sites, and other post-translational modifications that hinder cheap production of many biologics, allowing for straight-forward, large-scale fermentation in bacteria or yeast expression systems [28]. The small size and high stability of engineered scaffolds may also allow for new routes of administration, such as aerosolized pulmonary delivery, as well as facilitating greater systemic penetration into dense tissues. Additionally, the inherently modular nature of many designed scaffolds may lead to relatively easy combination of different binding domains into multivalent molecules with high avidity or multispecific molecules able to bind multiple targets simultaneously. Unfortunately, the novel qualities of this class come hand-in-hand with greater risk of immunogenicity, as non-human epitopes in the proteins are more likely to be recognized as non-self by the immune system, especially for therapies requiring repeated administration. Many of the novel scaffold-based drugs in current clinical development are based on human domains, such as DARPins from ankyrin repeat domains or Avimers® from LDLR-A modules, a fact that may or may not circumvent clinical immunogenicity issues before approval. Further clinical

testing will be required to actually gauge the promise of this class, as only one novel scaffold therapy, ecallantide (Kalbitor®) has so far been approved at the time of this writing.

The Adaptive Immune Response and Molecular Basis of Immunogenicity

All protein-based drugs have the potential to elicit an immune response from the patient, and non-human proteins present a particularly difficult challenge. Immune responses may affect drug efficacy, safety, potency, and pharmacokinetic/dynamic profile. In particular, the development of anti-drug antibodies (ADAs) by the patient may neutralize the drug, or even lead to dangerous cross-reaction with endogenous proteins [14]. Production of ADAs may proceed in both a T-cell dependent and -independent manner. T cell independent responses are usually initiated when B cell receptors bind to a repeated structural motif in the offending antigen such as repeat domains, structured aggregates, or post-translational glycans, leading to transient expression of typically lower-affinity IgM antibodies. In contrast, production of higher-affinity matured IgG ADAs requires activation of T cells by T cell epitopes in the protein drug.

Adaptive immune responses fall broadly into two categories: classical and breach-of-tolerance. The classical adaptive immune pathway responds to molecules recognized as “non-self”. This pathway plays out in the typical healthy responses to pathogens and vaccines. Therapeutic proteins that are originally deficient in the patient, human proteins modified from wild-type, or non-human proteins will all typically activate the adaptive immune system through the classical pathway, as the patient’s immune system does not recognize these molecules as “self”. The less well-understood pathway involves a breach of immune tolerance. This can occur spontaneously, leading to autoimmune disorders, or can occur when delivering human proteins exogenously even when the protein is originally present in the patient. Factors leading to breach-of-tolerance responses may include presence of epitope-like sequences, T cell cross-reactivity, innate immune activation, and protein aggregation.

The classical adaptive immune response is typically initiated when the protein drug is taken up by antigen-presenting cells (APCs) through pinocytosis, phagocytosis, or recepetor-mediated or –independent endocytosis. The antigen is unfolded and through a series of progressively more acidic endosomal compartments before subsequent fusion with the MHC class II compartment where a diverse set of proteolytic enzymes processes the antigen into variable-length peptide fragments. A specific subset of these peptides are selectively loaded onto human-leukocyte antigen (HLA) MHC class II receptors (MHC-II) before being shuttled back to the cell membrane for extracellular display of antigen fragments on MHC cell surface receptors. MHC molecules are highly polymorphic in the human population, and a given MHC allelic variant will bind only a unique repertoire of peptide sequences with high affinity. There are six main heterozygous MHC-II genes in humans (DPA1, DPB1, DQA1, DQB1, DRA, DRB1), all of which have many allelic variants, such that each person has a unique complement of up to 12 MHC-II receptors, each with a

unique peptide binding specificity profile. Once displayed on the cell surface, this MHC-peptide complex may be bound by T cell receptors (TCR) on the surface of T cells. If a TCR does bind the MHC-peptide complex with sufficient affinity, signal transduction through the T cell membrane may lead to activation of the T cell. Though extracellular proteins are typically processed through the class II pathway just described, cross-presentation can sometimes divert the antigen response into the class I pathway, leading to loading of antigen peptide fragments onto MHC-I molecules which display there own unique set of peptide binding specificities. Cross-presentation is a less likely and less well-understood phenomenon that is reviewed in sufficient detail elsewhere [13].

Experimental Characterization and Reduction of Protein Immunogenicity

In vitro T cell assay methods may be used to gauge the potential for a new drug candidate to elicit a harmful immune response before costly and time-consuming in clinical trials begin. Enzyme-linked immunosorbant spot-forming (ELISpot) assays can be used to measure cytokine produced by T cells stimulated ex vivo after introduction of the drug to the cell culture. These assays are similar to classical ELISA screens, but can give qualitative measurements of the number of cytokine-producing cells. T cell proliferation may be measured by incubating antigen-experienced T cells with MHC-peptide complexes in the presence of radioactive nucleotides, leading to radiolabeling of proliferating T cell pools but not resting inactive pools. T cells expressing TCRs that recognize a specific epitope may be selectively identified using MHC tetramers. Here, a tetrameric complex of MHC-II receptors conjugated with a linker to the epitope of interest are fluorescently labeled, allowing for staining of T cells reactive to that epitope sequence and rapid isolation with fluorescence-activated cell sorting (FACS).

After identification of T cell epitopes in the protein drug, experimental "de-immunization" efforts to reduce the immunogenicity of the therapeutic typically rely on multiple cycles of epitope mutation and re-characterization of both therapeutic activity and immunogenicity. Unfortunately, it is difficult to determine *a priori* which mutations to the protein will still allow for proper folding and function of the molecule, nor is it immediately obvious what mutations will eliminate the epitope without creating a new epitope. Mutations that still allow for functional activity of the protein may be identified through alanine-scanning mutagenisis or through random mutation and selection of active drug variants. Removal of T cell epitopes from staphylokinase constitutes the first published deimmunization effort in 2002 [15]. Here, alanine mutations were made at MHC anchoring residues, and active mutants were found that eliminated T cell response and in vivo immunogenicity. Since then epitope removal efforts have been applied to many different targets, and to a number of commercial protein therapeutics by Biovation [16], Epimmune [17], Genencor [18], and others, using a diverse range of different methods. Deimmunization of protein therapeutics is a rapidly changing field garnering greater and greater focus with the increasing number of immunogenic biologics both in active development and currently on the market.

Designer Proteins as Synthetic Biological Therapeutics: Affinity, Specificity, and Immunogenicity

Computational protein design may provide the tools necessary to create a novel class of therapeutic molecules with tunable biophysical properties such as increased specificity in molecular recognition, tighter binding to molecular targets, and reduced immunogenicity in systemic delivery. Computational protein design has already shown substantial success in design of stable non-natural protein folds [9], high-affinity protein binders for novel protein targets [10], ligand-binding proteins for synthetic organic molecules [11], and new enzymes to catalyze non-natural chemical reactions [12]. All these successes have hinged on our ability to correctly model the molecular energetics of protein interfaces and our ability to use these models to predict the structural and functional effects of design mutations. By leveraging the power of this approach to molecular design, we may open the door for the development of novel therapies targeting a wide range of currently untreatable diseases.

Sequence Optimization in Computational Protein Design

The protein design problem is typically considered as the inverse of the protein-folding problem: given a three-dimensional structural template, find a sequence that will adopt that structure spontaneously. The fixed-backbone approach to protein design samples different residue identities at fixed positions in the protein structure (Fig. 1). At each sequence position in the protein chain, different amino acids may adopt different sidechain conformations according to intra- and inter-molecular energetics. To sample this vast sequence and structure space, sidechain torsional conformations are discretized into discrete rotational isomers, “rotamers”, and bond lengths and angles are held fixed. In protein crystal structures, most amino acids adopt one of a discrete set of sidechain conformations as a function of backbone angle and local sterics. This allows for the clustering of the conformations into a small number of discrete rotameric states. During the design process, different rotamers of different identities are swapped in and out of different sequence positions in order to minimize the predicted folding free energy of the protein. Brute force sampling of all 20^N possible sequences and associated conformational states is unpractical and computationally intractable. Combinatorial optimization of the protein sequence can be carried out in a biased stochastic manner using multiple rounds of Metropolis-Monte Carlo and simulated annealing, allowing for fast convergence to near-optimal sequences.

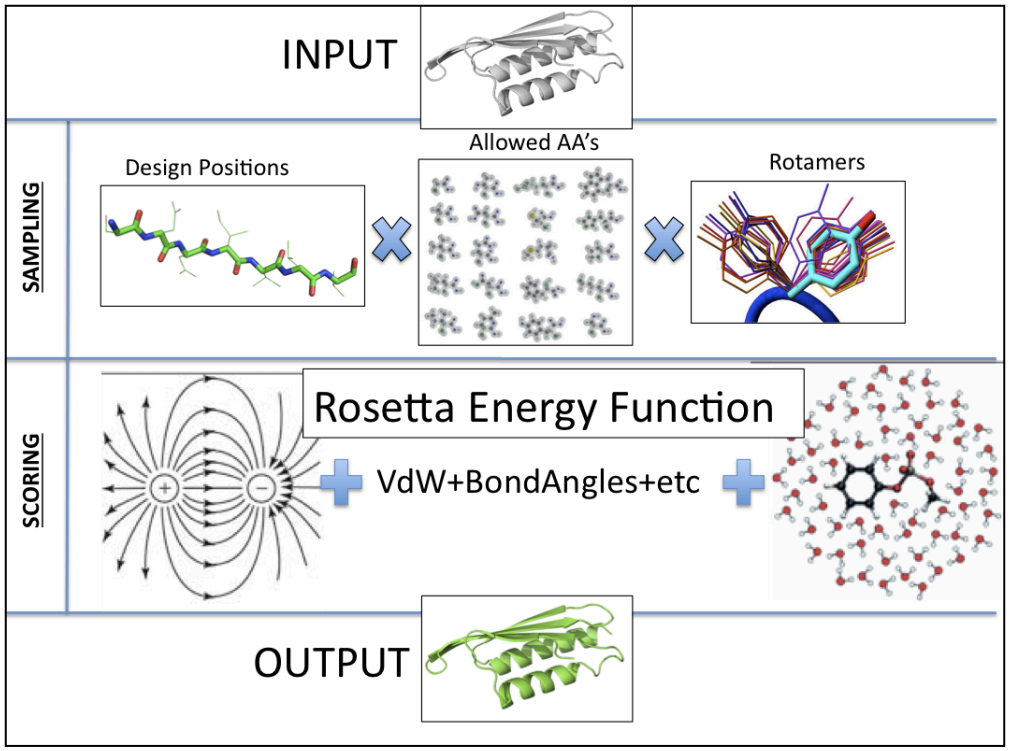


Figure 1. Fixed-backbone protein design proceeds by rapid iteration of two interrelated steps: sampling and scoring. A predefined set of design positions, allowed amino acids at those positions, and a discrete number of sidechain rotamers for each amino acid are sampled stochastically to optimize the computed energy. This energy is evaluated for each state in the scoring phase, where intra- and inter-molecular energies are calculated as function of the three-dimensional coordinates and chemical connectivity of all the atoms in the simulation.

Modeling Energetics in Computational Protein Design

A quantitative model of the intra- and intermolecular forces governing protein folding and binding is critical to our ability to design new proteins. These forces can be put into three categories: chemical bonds, protein-protein forces, and protein-water-forces. The conformational preferences of bonded atoms obey quantum mechanical rules that are difficult to calculate and are outside the scope of this work. However, we may glean the emergent effects of these forces by observing high-resolution crystal structures and collecting statistics on bond lengths, angles, and dihedrals as a function of amino acid type and local chemical environment. These statistics can be used to construct heuristically motivated potentials that recapture these phenomena and bias bonded atoms into native-like conformations. The increasingly vast number of crystal structures available for training these statistical models allows for further conditioning on priors such as local sequence, local structure, solvent exposure, and other local environmental factors

Protein-protein forces may be captured by a sum of distance-dependent inter-atomic interaction potentials. Van der Waals forces are typically accounted for with a standard Lennard-Jones curve; an easily computed spline of the emergent effects

of the Pauli exclusion principle (non-polar repulsion) and electron polarization (non-polar attraction). Electrostatic interactions are often simply modeled as Coulomb charges in a dielectric medium. Because polar water molecules will orient partial charges toward solvent-exposed charged groups, electrostatic interactions can be screened and dampened when they occur in or near the polar solvent environment; when charges are buried in the core of the protein, the dielectric environment more closely approximates a vacuum with a relative increase the strength of charge-charge interactions. Accounting for the dielectric environment in a computationally efficient manner is the subject of active development, from simple and fast distance-dependent dielectric models to complex and slow Poisson-Boltzmann calculations. Hydrogen bonds are unique in that they exhibit both ionic-like and covalent-like characteristics. Similar to the modeling of bond torsional angles, the complex quantum effects underlying hydrogen bond energies and geometries can be captured by collecting statistics from high-resolution crystal structures, facilitating rapid evaluation of approximate hydrogen bond forces to bias simulations toward native-like geometries and design of native-like binding motifs.

Protein-water forces are critical to the evaluation of protein design solutions, as the "hydrophobic effect" remains the primary driving force behind protein folding and self-assembly. These effects are very difficult to model accurately unless bulk water molecules are accounted for explicitly, as in all-atom molecular dynamics simulations. Unfortunately, the massive sampling of sequence space required by protein design renders concomitant sampling of all solvent degrees of freedom intractable at best. To work around this problem, implicit solvation models attempt to account for water molecules in a continuous, mean-field manner as a function of the protein atomic configurations. Solvent-exclusion models account for water effects by calculating the energetic costs of desolvation for each atom. Every chemical group is assigned a reference free energy of solvation, representing the energy released or absorbed when that chemical group is fully solvated with water molecules. Other protein atoms in the vicinity of a given atom "exclude" some fraction of the total water molecules surrounding the atom in the fully-solvated state. The energetic penalty of excluding solvent from interaction with the atom or chemical group is thus proportional to both the volume of space excluded and the reference free energy of solvation. Variants of this model may account for water hydrogen bonding and the anisotropic distribution of solvent around polar chemical groups with an exclusion model dependent on the precise orientations of protein-protein inter-atomic geometries.

Chapter 2

Structure-based prediction of protein-peptide specificity in Rosetta

Disclosure: This chapter has been published as

King, Christopher A., and Philip Bradley. "Structure-based prediction of protein–peptide specificity in Rosetta." *Proteins: Structure, Function, and Bioinformatics*78.16 (2010): 3437-3449.

My contributions consisted of devising the algorithms and protocols described, implementing those protocols in the Rosetta molecular modeling package, performing calculations and analyzing data, and writing the full text of the paper.

Abstract:

Protein-peptide interactions mediate many of the connections in intracellular signaling networks. A generalized computational framework for atomically precise modeling of protein-peptide specificity may allow for predicting molecular interactions, anticipating the effects of drugs and genetic mutations, and redesigning molecules for new interactions. We have developed an extensible, general algorithm for structure-based prediction of protein-peptide specificity as part of the Rosetta molecular modeling package. The algorithm is not restricted to any one peptide-binding domain family and, at minimum, does not require an experimentally characterized structure of the target protein nor any information about sequence specificity, although known structural data can be incorporated when available to improve performance. We demonstrate substantial success in specificity prediction across a diverse set of peptide-binding proteins, and show how performance is affected when incorporating varying degrees of input structural data. We also illustrate how structure-based approaches can provide atomic-level insight into mechanisms of peptide recognition and can predict the effects of point mutations on peptide specificity. Shortcomings and artifacts of our benchmark predictions are explained, and limits on the generality of the method are explored. This work provides a promising foundation upon which further development of completely generalized, *de novo* prediction of peptide specificity may progress.

Keywords: peptide specificity, peptide recognition domains, linear binding motifs, peptide design, molecular modeling, Rosetta PepSpec

Abbreviations:

PRD, peptide recognition domain, PDB, Protein Data Bank, PWM, position-weight matrix

Introduction:

The small number of genes in the human genome is able to produce organisms of such astounding complexity due in large part to the signaling networks created by protein-protein interactions. Such interactions are often mediated by peptide recognition domains (PRDs), modular protein domains that bind specifically to one or more short, linear amino acid sequences. The topology of protein signaling networks depends in part on these specific intermolecular binding events. The ability to predict these interactions based solely on the sequence of the PRD circumvents the need for time- and cost-intensive experiments and is a significant step toward prediction of entire signaling networks using genomic data.

Past approaches to peptide binding specificity prediction fall on a spectrum that runs from data-intensive statistical and machine-learning approaches at one end to biophysical and computational chemistry methods at the other. The probabilistic methods at one end of this spectrum rely on the previously observed behavior of a molecular system, leveraging data from actual peptide binding experiments to train a classifier to discriminate between binding and non-binding peptides or predict binding affinities and peptide binding profiles. Such methods have been used extensively in machine learning-based prediction of MHC-peptide specificity [1-3] and PDZ domain specificity [4]. These methods have the advantage of being fast and sometimes extremely accurate; however, they typically require large amounts of experimental training data, and thus may fail for systems that have not been well characterized experimentally. By contrast, physical/structural methods rely on basic principles of chemistry and physics in order to predict the relative binding affinities of different peptide ligands from the precise three-dimensional structure of the protein-peptide complexes. Prediction of binding affinity is often based on *ab initio* free energy calculations as per classical molecular mechanics or semi-empirical force fields [5-8]. Though these *ab initio* methods can be accurate even in the absence of experimental data or when non-canonical binding modes are important, they require large computational resources, making it challenging to explore a large number of peptide sequences. Additionally, they typically rely on an experimentally characterized protein structure as input. Many methods for peptide specificity prediction fall somewhere between these two extremes. For example, the Predikin webserver uses a database of "substrate-determining residues" manually curated from X-ray protein structures, along with public databases of known kinase targets, to predict the specificity of Ser/Thr kinases [9]. Hou et al. used a combination of homology modeling, molecular dynamics, and machine learning to predict the substrates of SH3 domains [10]. In another study, Sánchez et. al used the FoldX force field and *in silico* site-directed mutagenesis of known crystal structures to predict peptide binding affinities and interaction maps for SH2 domains [11]. All of these methods rely on a large number of experimentally determined structures from the PDB, and are primarily tailored to one specific class of PRD.

Here, we describe *pepspec*, a flexible structure-based algorithm for prediction of protein-peptide specificity developed within the Rosetta molecular modeling package [12]. The *pepspec* algorithm is essentially an anchored, flexible-backbone peptide docking and design algorithm in which the sequence and structure of the peptide are simultaneously optimized. Rather than performing global peptide docking searches, *pepspec* requires as input an approximate location for a key "anchor" residue of the peptide; the remainder of the peptide is assembled from fragments as in *de novo* structure prediction and refined with simultaneous sequence optimization. Backbone flexibility of the protein is incorporated implicitly by docking into a structural ensemble for the protein partner. The algorithm builds upon established strengths of the Rosetta modeling package in conformational sampling and protein sequence design. We show promising benchmark results for a diverse set of peptide-binding proteins, and demonstrate that use of atomically precise structural modeling not only permits accurate predictions without the use of any experimental specificity data, but may elucidate structural mechanisms of specificity as well. This work extends the work of Sood et. al [13] in which Rosetta was used to design peptide extensions onto PRD-bound peptides. However, because *pepspec* prediction of a target protein's specificity profile does not require the structure of the protein bound to a peptide, we are able to make predictions for both unbound structures and homology models. We present results from a number of different benchmark tests of increasing difficulty, and evaluate the major hurdles and strategies for increasing the accuracy of the algorithm in cases where high-resolution structural modeling becomes difficult.

Methods:

The *pepspec* application requires two inputs: a set of backbone coordinates for the target protein obtained via experiment or computational modeling, and an approximate docking location for a single anchor residue in the peptide. The use of an anchor residue is motivated by the existence in many PRD families of a key peptide residue whose docking location is highly constrained: in the case of kinases, this is the residue to be phosphorylated; for phospho-binding families such as the SH2 domain this is the phosphorylated residue; for families such as the PDZ domain that recognize terminal peptides, the anchor is the terminal residue itself. For all the modeling simulations described in this work, the peptide anchor residue is specified by the user, and the approximate binding location is defined by the position of the anchor residue in one or more structures of homologous protein-peptide complexes. The protocol then proceeds in three stages (Figure 1). Briefly, a structural ensemble of the target protein is generated and the data on approximate anchor location are used to constrain the docking of a new, user-defined anchor residue to the surface of the target protein, thus generating an ensemble of target protein structures docked to a single amino acid. Next, these structures serve as the starting point for a flexible-backbone peptide design protocol followed by fixed-backbone sequence sampling. After generating a large number of candidate peptides, sequences are

sorted by predicted binding energies and compiled into a specificity prediction for the target PRD.

This basic protocol can be modified by the user to incorporate additional structural data. If a peptide-bound structure of the target protein itself is available, the fixed-backbone sequence sampling phase can be applied independently, allowing for local exploration of sequence specificity within the vicinity of the input structure. If several homologous peptide-bound structures are available, these can be processed to generate loose constraints on the backbone conformation of the peptide away from the anchor residue.

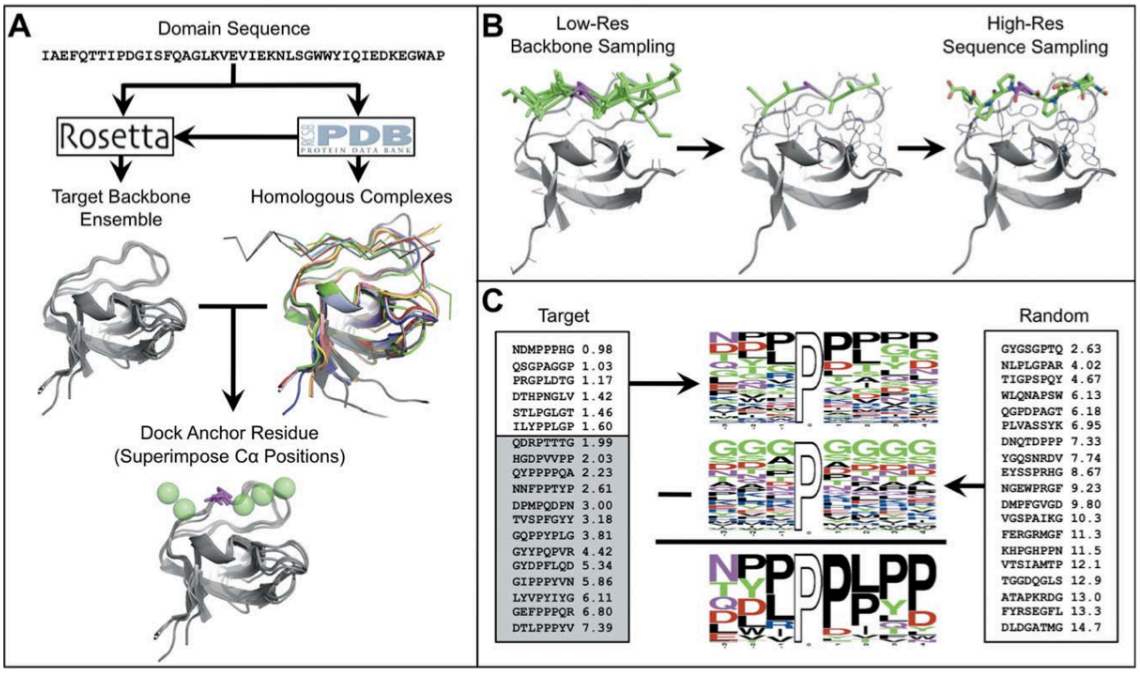


Figure 1. Overview of algorithm for peptide specificity prediction.
The peptide specificity algorithm consists of three stages: (A) a prep stage in which publicly available structural data are used to obtain a PRD model, dock an anchor residue, and optionally derive Ca constraints, (B) a peptide design stage in which high affinity peptides are built from the anchor residue, and (C) a postprocessing stage in which a specificity PWM is generated from the designed peptides and normalized by background noise.

Input structure preparation and diversification

Target protein backbone coordinates were obtained from either the Protein Data Bank (PDB) [14] or via homology modeling. All PDB-derived coordinates were crystal structures with resolution less than 3 Å. To ensure homology modeling tests of non-trivial difficulty, we excluded template proteins with more than 50% sequence identity to the target protein. Sequence alignments were performed with the ClustalW web server [15]. Initial backbone coordinates based on these alignments

were constructed using Modeller [16], after which Rosetta was used for local backbone and side-chain optimization and diversification. The Rosetta *relax* application [12] was used for homology models based on unbound template structures, whereas the Rosetta *backrub* application [17] was used for diversification of native unbound structures and homology models based on bound templates. This distinction was made because homology models based on bound templates, though more similar to the native peptide-bound conformation, may be disfavored when in the unbound state. Use of the less aggressive *backrub* algorithm prevented these models from moving too far away from the hypothetical bound state.

Anchor docking with constraints

Structural alignments for the target PRDs and their homologues were performed using the *align* command in PyMol [18] before loading the structures into Rosetta. The preparatory stage of the *pepspec* application begins with docking of anchor residues to members of the input target structure ensemble. The program constrains docking to a small area of the protein's surface by calculating averages and standard deviations for all rigid body degrees of freedom between the protein and the anchor residues of the aligned homolog-peptide complexes. At each iteration of docking, a new user-defined anchor residue is created in this averaged orientation and perturbed by up to one standard deviation, after which docking to the protein surface is attempted with a combination of small Gaussian perturbations of the rigid-body degrees of freedom and rotamer-based sidechain optimization and minimization of both the anchor residue and neighboring protein residues. If clashes cannot be resolved, the simulation is terminated and begins again. This generates an ensemble of protein-anchor complexes that can be later be filtered by score to isolate energetically favorable conformations.

Optionally, homologous structures can be used to generate a set of coordinate constraints on the peptide residue Cα atoms for use in the subsequent peptide design phase. For each peptide Cα position, a flat-bottom harmonic potential is calculated from the set of aligned homolog peptides according to Equation 1. The potential is centered on the mean Cα position and has a minimum zero-potential well width equal to the distance from the mean Cα to the farthest homolog Cα coordinate. To avoid over-constraining the search space with insufficient data, the well width increases as the relative number of data points for that position decreases. During the subsequent peptide design phase, the total score penalty can be calculated as:

$$E^{cst} = \sum_{i=1}^{N} \begin{cases} \left(\dfrac{\left\| x_i - \mu_i \right\| - t_i}{\sigma_i} \right)^2, & \left\| x_i - \mu \right\|_i > t_i \\ 0, \text{ else} \end{cases} \qquad \textbf{(1)}$$

$$t_i = t_i^{\min} * \left(2 - \left(\frac{n_i}{N} \right)^{1/3} \right) \quad \textbf{(2)}$$

$$t_i^{\min} = \max \left\{ \left\| x_{i,1}^h - \mu_i \right\|, \ldots \left\| x_{i,N}^h - \mu_i \right\| \right\} \quad \textbf{(3)}$$

where, for each peptide position i, x_i is the ith Cα coordinate of the peptide, $x^h_{i,j}$ is the ith Cα coordinate of homolog j, μ_i and σ_i are the average and standard deviation for position i Cα coordinates over all N homologs, n_i is the number of homolog structures that contain an aligned peptide residue at that position, t_i is the zero-potential tolerance radius for that position, and E^{cst} is the total energy penalty from the constraint.

Flexible-backbone peptide design

After the peptide anchor residue has been docked, a peptide backbone is generated using a reduced-atom model of the protein-peptide complex. That is, all protein residues except proline and glycine are mutated to alanine, and all peptide residues except the anchor residue are represented as glycine. First, a user-defined number of peptide residues is added N- and C-terminal to the anchor residue using random Φ/Ψ angles. Peptide backbone conformations are then sampled in a 1000-step Metropolis Monte Carlo search using insertion of random Φ/Ψ/Ω angle triplets (1-mer fragments [19]) chosen from the PDB. Backbone fragments are selected without regard to amino acid identity, such that the glycine representation of the peptide does not directly influence backbone dihedral angle preferences. Use of alanine residues for the reduced atom peptide model was not found to increase performance in benchmark testing, and actually prevented sufficiently broad backbone sampling in a number of cases (data not shown). Reduced-atom models are scored using a sum of the Rosetta full-atom repulsive term, a function penalizing residues that move more than 6 Å away from the nearest protein residues, and, optionally, the harmonic constraint scores mentioned above. The best-scoring structure from each search is then recovered and checked for intermolecular atomic clashes by replacing the protein's native sidechains. If any clashes occur, the program attempts to resolve them by performing a fixed-backbone rotamer optimization of the protein residues in the context of the new poly-glycine peptide using the full-atom Rosetta potential function. If all clashes are resolved, the peptide is used as a template for subsequent sequence optimization.

Sequence optimization and sampling of the peptide backbone proceed as follows. The peptide sequence and the neighboring protein sidechains are simultaneously optimized with a rotamer-based simulated annealing Monte Carlo search based on the Dunbrack rotamer library [20]. The sequence specificity of the peptide is then explored using a Monte Carlo/Minimization (MCM) sequence sampling protocol. Each step of this search involves random mutation of one peptide residue, rotameric optimization of the mutated residue and all neighboring residues, and gradient-

based minimization of all sidechain degrees of freedom, followed by scoring with the standard Rosetta full-atom potential function. This final MCM sequence sampling phase can be applied independently to known PRD-peptide structures when the backbone coordinates of the bound peptide are known.

Sequence and score data are written out after each iteration of sequence sampling for later processing. A binding score is calculated as the total Rosetta score of the complex minus the sum of the scores of the unbound peptide and protein. Unbound scores are calculated simply by deleting one or the other binding partner, repacking all sidechains, and rescoring. Because the backbone is frozen during rescoring, different members of the PRD target ensemble are calculated to have different unbound scores. Though unphysical, this was seen to improve specificity predictions, as the noise generated in the target backbone diversification often overshadowed the more subtle energetic effects of favorable peptide-binding conformations.

The computational cost of the algorithm varies with the length of peptide and the number of residues at the interface. Full-length simulations ran between 100 and 300 cpu-hours on 64-bit 2.33 GHz Intel Xeon processors.

Post-processing

After all peptides have been generated, sequences are sorted by binding score and the best 1% scoring sequences are compiled into a position-weight matrix (PWM), a simple but versatile representation of sequence specificity [21]. To increase the strength of the PWM signal and eliminate artifacts of the sampling protocol and scoring scheme, this PWM is then normalized by a background PWM. The background PWM was generated by combining *pepspec* sequence-score output for all structures in *peptiDB* [22], a non-redundant dataset of 103 high-resolution protein-peptide complex crystal structures. Instead of seeding the peptide design protocol with a structural ensemble of one target PRD, the application was seeded with these 103 diverse proteins. For each peptide generated, a random structure from the *peptiDB* was chosen and all peptide residues were removed except for a randomly chosen anchor position, followed by the unconstrained peptide design protocol. All sequences thus generated were sorted by binding score, after which the low 10% scoring sequences were compiled into a PWM. This 10% cutoff threshold was used so as to insure a diverse set of PRD systems were included in the background PWM, as a 1% cutoff was seen to include a relatively small number of proteins from the peptiDB. To normalize a target PWM with this background signal, the background PWM is directly subtracted from the target PWM. Elements of the resulting matrix that fall below zero are simply set to zero, after which columns are renormalized to unity. All PWM graphics were created using WebLogo [23].

Benchmark reference dataset

A benchmark target set was chosen to cover four of the most well characterized families of PRDs: kinase, SH2, SH3, and PDZ. One representative protein was chosen for testing in each PRD family. We attempted to choose the most prototypical example of each family; details of our target selection rationale can be found in Supplementary Materials. To benchmark our prediction algorithm, it was also necessary to compile experimental reference specificity data for each PRD target and compare to our predictions. We decided to compare PWMs based on experimental data to PWMs generated by *pepspec*. Though discussion as to the validity of the positional independence assumption of the PWM is ongoing [24], we chose to use this representation based on its relative simplicity and widespread use. Experimental PWMs are usually compiled simply from an alignment of some number of known protein-binding peptides. PWMs derived from phage display data can vary significantly from experiment to experiment due to different codon bias in different phage strains, artifacts due to phage propagation optimization, and different solid supports and affinity partners [25]. In addition, the information content of an *in vitro* experiment-based PWM will vary with the number of sequences published, a variable affected by the authors' definition of a "high-affinity" sequence. When only a small number of sequences are published, phage display PWMs converge narrowly on one or more optimal sequences, resulting in very high information content PWMs that may miss biologically relevant secondary and tertiary preferences (Figure 2). PWMs based on aligned *in vivo* substrates, however, can carry with them specificity biases independent of peptide affinity due to biological factors such as intracellular co-localization, scaffold proteins, and codon biases. These sources often result in PWMs with very low information content. For the sake of consistency, and a balance between these two extremes, we chose to use the Scansite webserver [26] where possible to gather data for reference specificity PWMs. Scansite PWMs are based on amalgamated experimental binding data from oriented peptide array screening and phage display experiments. Because the internal Scansite scoring matrices could not be obtained directly, we used the webserver to extract the top-scoring 2000 hits from the mammalian SWISS-PROT database for each PRD and compiled these sequences into PWMs. This number of sequences is comparable to the number of sequences (2000-4000) used to build the predicted PWMs. Because PSD-95 PDZ data was not available in the Scansite database, a reference PWM was compiled from 93 high affinity peptide sequences isolated from a phage display experiment by Tonikian et al. [27].

PWM scoring and evaluation of performance

We chose to use the Frobenius norm as a measure of the 'distance' between the reference and predicted PWMs (simply the square root of the summed squared deviations in amino acid frequencies, [21]), as this seemed the most straightforward and common scoring metric available, and has an easily visualized geometric interpretation. To evaluate the significance of a given reference-prediction PWM distance, we transformed the value of the Frobenius norm into a P-value by scoring

1000 random PWMs against the reference. Random PWMs were generated so as to display a distribution of information content similar to the reference dataset. Details can be found in Supplementary Materials.

Comparison of performance with other methods was complicated by a lack of standardization in the output of structure-based specificity methods. Many structure-based methods involve some classifier for discriminating binding peptides, such as the SH3 peptide prediction algorithm of Hou et al [10]. Avoiding a binary distinction between "binding" and "non-binding" peptides precludes simple comparison in this case. Fair comparison with methods such as Predikin [9] that are based on extrapolation of known peptide specificity data is also not straightforward, since we have chosen for our benchmark set proteins with well-characterized specificities that were likely used in training these methods. On the other hand, comparison with flexible peptide docking methods is made difficult by the ensemble output of *pepspec*; the output consists of thousands of peptide sequences and structures, which are compiled into a PWM. More closely related structure-based approaches like that of Sanchez et al. [11] compared predicted vs. experimental binding affinities for SH2-peptide complexes. Our method transforms sequence-structure output into a PWM that is then processed further, limiting the ability for direct comparison. For these reasons, we have calculated PWM-based scores and P-values in order to provide some objective measure of the relative performance of our method.

Results:

Peptide-bound structures can be used to predict the distribution of information content

Before attempting *de novo* specificity prediction, we first tested the ability of Rosetta *pepspec* to recapitulate peptide sequence specificity when given the crystal structure of a native protein-peptide complex. This makes use only of the final sequence sampling phase of the full *pepspec* protocol, relying on experimental data for protein and peptide backbone coordinates. As in the full protocol, we generated a structural ensemble by applying the Rosetta relax application to the coordinates of the input structure. This was immediately followed by the final phase of the peptide design protocol: conservative fixed-backbone MCM sequence sampling of the peptide ligand. All peptide sequences were then compiled into a PWM. As seen in the *bound +fixed backbone* row of Figure 2, the optimal sequences do not deviate from the sequences of the original crystal structures (Table S1 and Table S2), likely due to the lack of substantial backbone sampling in this more conservative protocol. This provides a limited test of the sequence optimization protocol: strong biases in either the scoring function or the sequence sampling could be expected to produce substantial divergence from the crystal structure sequences. More interesting, however, is the correlation of the positional information content between the predicted and reference logos. Information content is a quantifiable measure of the

importance of a given peptide residue to the overall specificity of the binding event [28]. As seen in Figure 3, the predicted values do indeed trend somewhat with the reference data, with a Pearson correlation coefficient of 0.43. The accuracy of these predictions is limited by the quality of the match between the consensus sequence and the peptide sequence in the input structure, as peptide residues mismatched with the preferred residue at information-rich positions are not likely to be good predictors of information content. For instance, in the kinase test, a strong preference for Gly is predicted at the P-4 position, whereas the reference PWMs display very little specificity at P-4. This is due to the fact that the input peptide-bound crystal structure was solved with a peptide that contains Gly at P-4 and backbone torsion angles that mostly disallow the other 19 amino acids. A similar situation also occurs in the PDZ domain test, where Pro is predicted frequently at the P-1 position due to the input peptide backbone torsion angles at P-1 giving a higher score to proline residues.

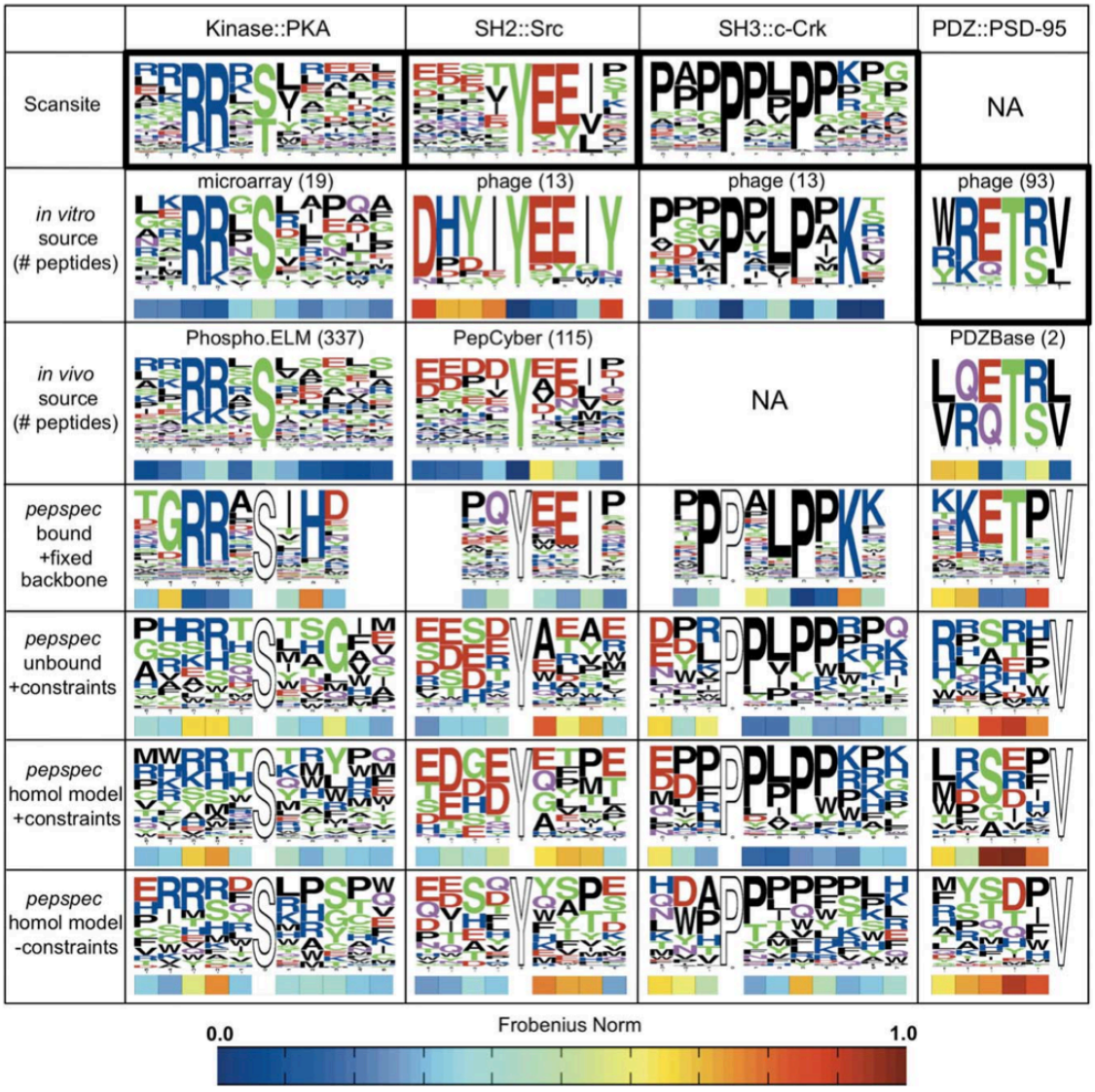


Figure 2. Benchmark results for peptide specificity prediction.
Benchmark results for recapitulation of binding specificity for four different PRDs. Rows 1-3 show experimental PWMs from various sources (see text). Bold boxes denote reference PWMs. *In vivo* sequences were collected from publicly available databases [30,38,39]. *In vitro* sequences were collected from various published peptide specificity experiments [27,40-42]. Color bars below PWMs indicate single position Frobenius norms relative to the reference PWM, from dark blue (0.0) to dark red (1.0). Anchor residues are colored white and were not predicted or scored. Rows 4-7 show results for four benchmark tests with increasing degrees of difficulty.

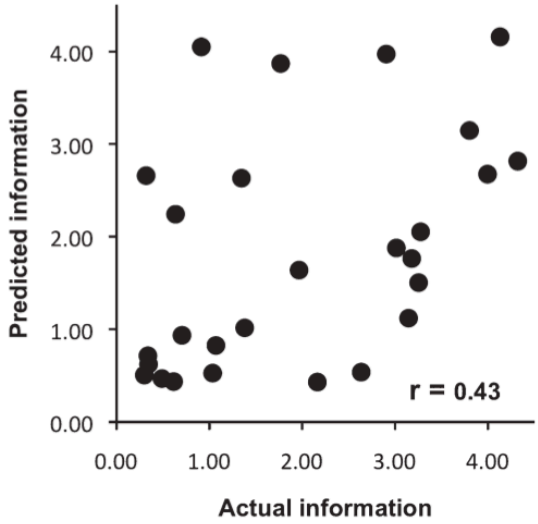


Figure 3. Prediction of information content using native bound structures.
Predicted vs. actual information content by position for PWMs generated from native peptide-bound structures. The data give a Pearson correlation of 0.43.

The *pepspec* application can make predictions for unbound structures and homology models

Because the majority of PRD structures in the PDB have not been solved in complex with their peptide ligands, one of our goals is to predict peptide specificity using structures of unbound proteins. Additionally, we wish to predict peptide specificity for proteins that are not present in the PDB. To accomplish this, we included in our benchmark test unbound crystal structures for all four PRD targets, and we used homology modeling to generate structural models for all four targets. To exclude trivial homology modeling cases, we only used template proteins with less than or equal to 50% sequence identity to the target protein. These tests rely only on the input structure and the structures of homologous peptide-bound PRDs (Table S1). The *pepspec* predictions starting from unbound structures and homology models are shown in Figure 2, rows labeled *unbound + constraints* and *homol model + constraints*, respectively. In what follows, we discuss the results for each benchmark

target individually, focusing on those positions with the highest information content in the reference PWMs.

For the kinase PKA, Rosetta did recapitulate the most salient features of the reference PWM; the preference for Arg/Lys at P-2 and P-3 was recognized, and the preference for beta-branched hydrophobic residues at P+1 was captured to a small extent. However, the relative importance of these positions was not captured by the prediction. The homology model-based prediction is comparable to the unbound structure-based prediction, even though the model had a higher Cα RMSD to the native structure than the unbound high-resolution crystal structure (data not shown). This is due to the fact that, in this case, the homology template was taken from a peptide-bound kinase structure, necessary due to the high intrinsic plasticity and the large conformational transitions that often occur upon peptide binding in the Ser/Thr kinase family [29].

Rosetta's performance for the Src SH2 domain tests was only marginally successful in both the unbound and homology tests. The acidic residue specificity for N-terminal, P+1, and P+2 positions was partly recovered in both tests. However, Rosetta failed to recapitulate the preference for Ile at P+3. Interestingly, the predicted PWMs are much more similar to a PWM compiled from known *in vivo* ligands in the PepCyber database [30]. For the *in vivo* PWM in Figure 4A, the P+3 position shows only a slight preference for Ile over other small hydrophobic residues. Also, more acidic residues are preferred in the N-terminal positions, and the P+1 position shows some specificity for Ala, a signal recovered in the *unbound +constraints* test. The fact that hydrophobic residues were recovered at the P+3 position shows that Rosetta did recognize the chemical environment of this position's binding pocket, but the precise shape of the pocket in both the unbound crystal structure and the homology model was too dissimilar to the bound crystal structure's conformation to accurately match the preferred peptide residue (Figures 4B and 4C).

The c-Crk SH3 domain predictions were the most successful of the benchmark set. Both the unbound and homology tests fairly accurately recovered the strong signal for the PLPP sequence at positions P+1 to P+4. The preference for Lys at P+5 was recovered in the *homol model +constraints* test, and both tests correctly predicted Lys/Pro/Arg as the top three residues at P+5 and Pro as the top residue at P+6. The PSD-95 PDZ domain tests resulted in the poorest benchmark predictions, as only the P-4 position was predicted accurately in both tests. The most specific position, the Thr at P-2, was recovered as the second most preferable residue for the *unbound +constraints* test, but was completely missed in the *homol model +constraints* test. This is due in part to Rosetta's implementation of the environment dependence of hydrogen bonding, which down-weights hydrogen bonds between residues that are considered exposed to solvent. Elimination of this environment dependence allowed for improved recovery of P-2 Thr (data not shown), but resulted in an overall decrease in performance for the entire benchmark set.

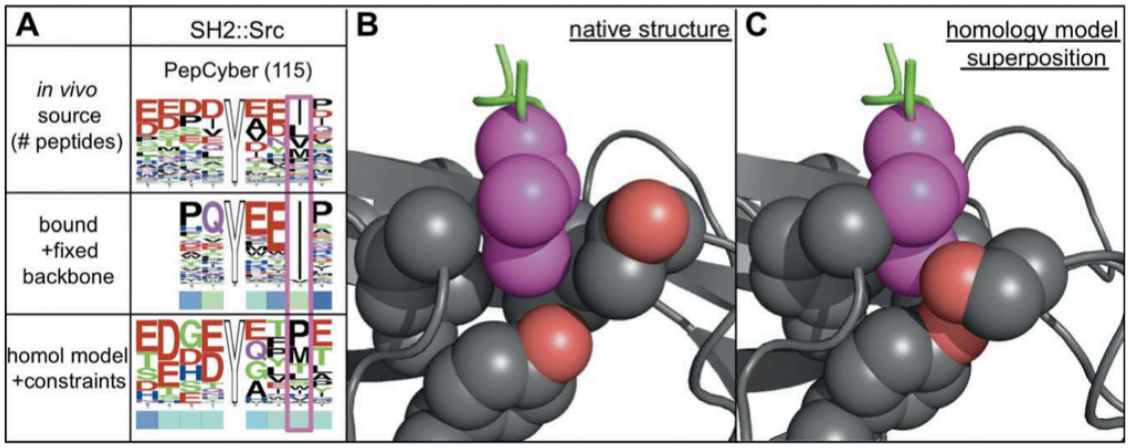


Figure 4. Protein backbone flexibility must be modeled precisely for Src SH2 peptide binding.
(A) SH2 benchmark tests compared to a reference PWM compiled from aligned *in vivo* targets. Note that this reference PWM differs from that of Figure 2. Predictions for the P+3 position (magenta box) match many of the less preferred residues in native peptide ligands. Comparison of the P+3 binding pocket in the native bound structure (B) and the native peptide superimposed on the homology model (C) reveals differences in the P+3 binding pocket that result in clashes between the conserved isoleucine (magenta) and the protein (gray).

Pepspec is extensible to use with limited structural data

Though structural methods for specificity prediction are computationally intensive, they have the advantage of being extensible to problems with very limited experimental data. The *pepspec* algorithm requires at minimum only one structure of a homologous protein bound to a peptide ligand. This is needed only to place the anchor residue, after which peptide design proceeds unconstrained. To show the degree to which the algorithm performs with this minimal amount of data, we predicted PWMs for our homology modeling benchmark set using only one homologous complex per target (Table S1). For the kinase and the SH2 domain, we used the complex structure from which the original homology model was derived. Because the SH3 and PDZ homology models were not based on structures of peptide-bound proteins, we used instead the complex structure of the protein most similar in primary sequence to the homology template as measured by BLAST E-value [31]. Because standard deviations are undefined for a single structure, the anchor docking translation and rotation deviations were manually set to 0.5 Å and 5°, respectively.

Results of this prediction can be seen in the *homol model -constraints* row of Figure 2. Surprisingly, elimination of Cα constraints did not consistently decrease performance (Table S2). For the kinase PKA, total PWM distance was lower than in either of the tests utilizing constraints. Due to the large number of acidic residues on the surface of the kinase, Rosetta could design arginine residues to take advantage of potential salt bridges not seen in crystal structures of the complex. As seen in Figure 5A, accurate prediction of P-4 specificity is a result of the unconstrained, flexible backbone methodology. A representative member of the predicted low-energy peptide ensemble provides a possible mechanism for P-4 Arg specificity not apparent from the peptide-bound crystal structure (Figures 5B and 5C). A number of low-energy designed peptides with Arg at P-4 were seen to form salt bridges with Asp241 and/or Glu203. These peptide configurations were disallowed by implementation of homologous constraints, but were sampled by the algorithm in the unconstrained simulation.

In contrast to the removal of kinase peptide constraints, the unconstrained SH3 domain prediction shows a substantial decrease in performance relative to the simulations employing prior knowledge of SH3 domain binding modes. Though the canonical PXXP motif was recovered, the signal is very weak, and deteriorates further at more distal peptide positions. In an unconstrained search, the volume of space to be explored increases very quickly as the peptide grows from the anchor residue to further and further positions. The homologous constraints allow for a more localized search within a smaller volume where the biologically relevant peptide ligands are likely to be found.

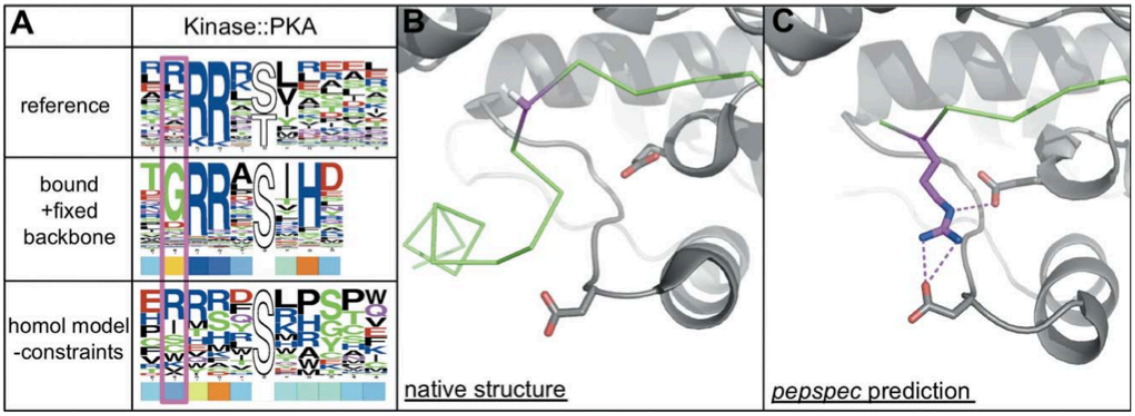


Figure 5. Structural predictions offer possible mechanism of P-4 kinase specificity.
(A) Kinase benchmark tests compared to reference PWM. P-4 specificity prediction (magenta box) fails for the near-native backbone test but succeeds for the *de novo* test. The bound crystal structure (B) does not explain the preference for Arg at P-4 (magenta). A member of the low-energy peptide ensemble from the *pepspec* simulation (C) achieves two salt bridge interactions with the kinase at P-4, providing a possible mechanism for Arg specificity.

Performance is a function of homology modeling difficulty

The success of homology model-based predictions is necessarily limited by the quality of the initial structural models, and the quality of these models is typically dictated by the degree of homology between the target protein and the template protein. To quantify this trend, we generated a total of four different homology models for the c-Crk SH3 domain starting from templates with a range of sequence similarities to the target protein. Template proteins were found with a BLAST search of the target sequence, and selected to represent E-values spanning 6 orders of magnitude. As seen in Figure 6, the accuracy of specificity predictions does trend as expected with the sequence similarity of the template protein.

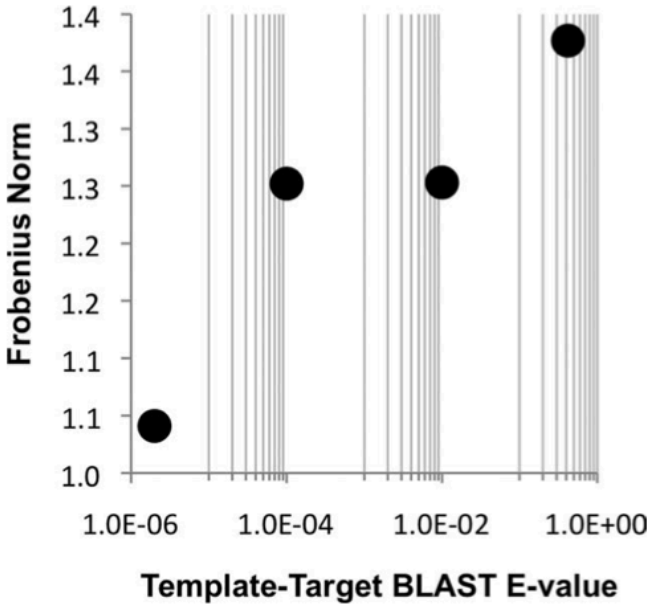


Figure 6. Homology modeling difficulty affects accuracy of c-Crk SH3 predictions.
Frobenius norm values for c-Crk SH3-N domain target predictions. Each point represents a prediction based on a different homology model, each model derived from a different template protein. The degree of similarity between the target sequence and the template sequence is captured by the E-value associated with the two sequences' BLAST alignment.

***Pepspec* can be applied to members of any domain family**

One of the primary advantages of structural methods is the minimal experimental information necessary to make predictions. To illustrate this, we chose to make peptide specificity predictions for a pSer/pThr-binding FHA domain. The details of binding specificity for this domain family are not as well characterized as for the other four benchmark targets, and we could not find any published examples of peptide specificity prediction methods applied to this domain family. Based on the

availability of experimental specificity data and a structure in the PDB, we chose the FHA1 domain of Rad53 as our target, generating a backbone ensemble from the refined NMR structure of the unbound protein. To predict the location of the conserved pThr anchor residue, we used the coordinates of only one homologous protein complex, the FHA domain of RNF8, which contains only 35% identical PRD residues and a dissimilar peptide ligand. A previous peptide library study has found Rad53 FHA1 to display strong preference for Asp at the P+3 position, as can be seen in Figure 7A, republished with permission from the study of Yongkiettrakul *et al.* [32]. In that study, the authors identified a single point mutation, G135D, that dramatically reduced the protein's specificity for this position. We attempted to replicate this experiment *in silico*. As seen in Figure 7B, the strong preference for Asp at P+3 in the wild-type protein was indeed recovered, as was the effect of the mutation in reducing the preference for Asp and reducing the overall information content of the P+3 position. The experimental residue frequencies for the mutant protein display very low overall specificity, with some preference for many hydrophobic residues, whereas the predicted specificity profile shows a slightly greater overall specificity, with very little preference for hydrophobic residues. This is due to the fact that our PWM normalization tends to lead to overestimates of information content (Figure 3) and most of the peptide backbones sampled in this case were very solvent-exposed. The crystal structure of wild-type Rad53 FHA1 in complex with peptide ligand [33] shows that P+3 Asp specificity is mediated by an ionic interaction with Rad53's Arg83 (Figure 7C). This interaction is indeed recovered by our peptide design algorithm as seen in a member of the low-energy peptide ensemble in Figure 7D. Furthermore, the effect of the G135D mutation is rationalized in this structural context. Gly135 is adjacent to Arg83, such that a mutation to Asp135 prevents local binding of acidic peptide residues due to charge repulsion and/or competition for binding to Arg83. This example illustrates the potential of structural modeling methods to generate predictions for and yield insights into systems with limited prior experimental data.

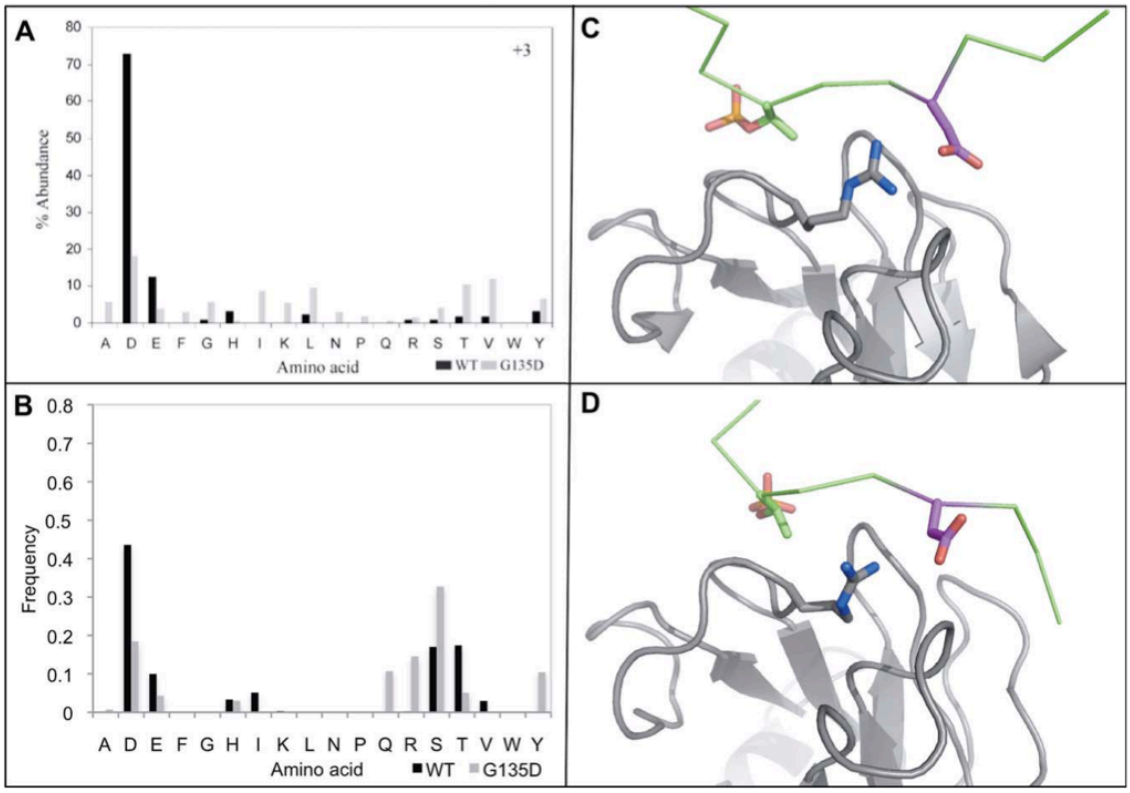


Figure 7. Prediction of Rad53 FHA1 specificity and effects of G135D mutation. (A) Experimentally derived amino acid preferences for the P+3 position of the wild-type Rad53 FHA1 domain (black) and the G135D mutant (gray) (reprinted with permission from Figure 5 of [32]). (B) P+3 preferences predicted by *pepspec* compare favorably with experimental data. (C) Crystal structure of Rad53-1 bound to high affinity peptide ligand. (D) Lowest RMSD Rosetta designed peptide after low energy filtering correctly predicts Arg83 mediated P+3 specificity.

Discussion

We have developed a new flexible-backbone peptide design algorithm, *pepspec*, and evaluated its ability to make predictions of peptide binding specificity across a range of benchmark tasks. The algorithm combines extensive sampling of the peptide backbone through anchored docking and fragment assembly together with large-scale exploration of peptide sequence space using Rosetta's sequence redesign methodology. Backbone flexibility of the protein partner is included implicitly by docking to a diverse structural ensemble. By emphasizing peptide and protein backbone sampling, we have tailored the algorithm for prediction scenarios in which the starting protein model is inaccurate, and knowledge of peptide conformation is limited. The algorithm is able to make non-trivial predictions for targets from a diverse range of PRD families using only unbound crystal structures

or homology models. In cases where additional structural information is available, this information can be incorporated to focus the conformational search. The generality of the algorithm is exemplified by accurate prediction of wild-type and mutant Rad53 FHA1 domain specificity using a minimum of experimental structural data. By taking a structural approach, we were able to make a testable hypothesis about the mechanism of PKA P-4 specificity that is not explained by the crystal structure, and were able to accurately recapitulate the known mechanism of FHA P+3 specificity.

At present, the *pepspec* algorithm requires as input a structural model of the target protein together with an approximate binding location for a single anchor residue in the peptide. In the tests described here, the binding location for this anchor residue is defined by one or more experimentally determined peptide-bound structures of related proteins. The need for a user-defined anchor residue and an approximate binding pocket is a clear limitation of the method, however three recent studies suggest avenues for relaxing this requirement. Petsalaki *et al.* [34] derived spatial position specific scoring matrices for 23 amino acid types by analyzing known protein-peptide complexes in the PDB; these matrices were used successfully to predict the locations of peptide residues on the surfaces of unbound protein structures. Raveh *et al.* [35] introduced a novel flexible peptide docking algorithm that met with considerable success in structure prediction of known protein-peptide complexes. Brenke *et al.* [36] developed the FTMAP algorithm for predicting the consensus binding sites for a number of small organic probe molecules. Combining *pepspec* with sources of approximate structural binding data such as these might allow for true *de novo* prediction of peptide binding specificity.

Challenges for structure-based specificity prediction

Our results clearly demonstrate that structure-based prediction of binding specificity through atomistic modeling is accompanied by significant challenges. Simultaneous design of both structure and sequence results in a massive solution space that must be explored. At present, it is not known how to precisely model the complicated energetics of protein-peptide binding at speeds necessary to search through this space in a reasonable amount of time. In particular, computation of accurate binding energies for peptides is hampered by the potential diversity of their unbound conformations; approximations to binding energy [37] that are reasonably accurate for protein-protein and protein-DNA interactions break down for these highly flexible ligands. To circumvent this limitation, other methods have tailored scoring functions or classifiers specifically to capture the features common to one class of PRD [10] or have constrained their search to the neighborhood of known peptide structures [11]. Most notably, Smith and Kortemme have recently met with considerable success applying Rosetta to the prediction of Erbin PDZ domain peptide specificity (*T. Kortemme, personal communication*). By generating a conformational ensemble in the neighborhood of an experimentally determined peptide-bound PDZ domain structure, they have successfully optimized peptide sequences with a genetic algorithm and modified scoring function for accurate

prediction of both wild-type and mutant Erbin PDZ specificities. This demonstrates that success in modeling can be achieved by focusing sampling to the neighborhood of probable solutions and focusing scoring to the most relevant interactions. It was our goal to develop a domain-independent method theoretically extensible to any protein family or structure, such that we did not bias scoring or constrain the structure to any single binding mode or interaction type. To prevent over-training, we did not re-optimize the standard Rosetta all-atom scoring function. The modest accuracy of the specificity predictions in the more challenging test cases reflects the challenges associated with atomically detailed structure-based peptide-binding specificity prediction.

Protein backbone diversification is necessary for accurate predictions

Instead of designing peptides for binding to single protein structures, we chose to diversify the input backbone coordinates into an ensemble of structures and combine specificity predictions for all of them. This was essential when making use of unbound protein structures and homology models, as slight changes in the protein backbone can have dramatic effects on predicted peptide binding ability. Our protein backbone sampling was not always aggressive enough, however. Our inability to predict the optimal conformation of the two Src SH2 loops comprising the P+3 binding pocket prevented accurate determination of P+3 specificity for both the unbound native structure and the homology model (Figure 4). The plasticity inherent in flexible structures such as these can result in significant differences between bound and unbound conformations. This highlights a key challenge for approaches that either ignore conformational flexibility of the binding protein, or sample it implicitly as we do through the use of structural ensembles. Accurate modeling of the protein's bound conformation may be essential for prediction of the optimal peptide ligand; at the same time, the protein conformation of the bound state may be energetically unfavorable in the absence of peptide, so that accurate structure prediction necessitates inclusion of the peptide ligand. Future work may involve concerted sampling of both protein and peptide backbone degrees of freedom for identification of the mutually optimal conformations.

Background PWM normalization boosts signal over sampling and scoring noise

The large degree of sampling involved in the *pepspec* protocol introduces a bias towards smaller residues in PWMs built directly from the low-energy sequences. This is due to the fact that, after the low-resolution backbone sampling step, smaller residues simply have a much greater chance of avoiding steric clashes with protein atoms during the high-resolution sidechain sampling phase. We also found our raw PWMs to exhibit rather low average information content of only 0.38 bits compared to the reference PWM average of 1.97 bits. To measure these biases, we ran a *pepspec* simulation drawing not from an ensemble of one target protein, but from a collection of 103 non-redundant protein-peptide complex structures. This generated a background PWM (Figure S1), not associated with any specific PRD,

that captures the inherent biases of the computational methodology and scoring scheme. To correct for these biases, we found it advantageous to use this background PWM in a normalization of any predicted PWMs, thereby boosting the signal-to-noise ratio and increasing the average information content to 1.72 bits. The normalization procedure was formulated to insure an amplification of signals that exceed levels expected by background noise, while preventing amplification of rare artifacts not captured by the background PWM. We initially considered normalization by calculating percentage enrichment over background frequencies. However, this was found to overweight signals from low frequency residues that nevertheless were relatively enriched to a great degree in the predicted PWM. To address this, we simply subtracted matrices directly and eliminated frequencies that fell below zero, as this offered a more conservative and ultimately more effective approach to eliminating background and boosting signal.

Conclusion:

We have described a flexible-backbone peptide design algorithm and its application to the prediction of peptide binding specificity. The algorithm can also be used for the design of high-affinity binding peptides, for example as competitive inhibitors of naturally occurring protein-peptide interactions. At present, the algorithm requires as input a structural model of the target protein together with an approximate binding location for a single anchor residue in the peptide. Our benchmark tests indicate that the algorithm is capable of making non-trivial predictions even when the starting models for the target protein are inaccurate. Although there is substantial room for improvement, we expect that by highlighting a few of the challenges inherent in structure-based specificity prediction, and evaluating potential solutions, this work may provide a foundation for future progress toward fast and reliable prediction of protein-protein interactions.

Supplementary Materials:

Benchmark Target Selection

Targets had to meet a number of requirements. Each had to have been structurally characterized both in complex with a peptide ligand and in unbound from, and each needed publicly available specificity data. Beyond this, representatives were chosen primarily by what seemed to be the most prototypical and/or well-studied member of the family. For the kinase test, we chose the cAMP-dependent Protein Kinase A (PKA), as this was the first kinase to be structurally characterized, and has the most entries in the PDB. For the SH2 domain, we chose the Src SH2 domain, as this is the protein from whence the acronym derives (Src Homology 2). For the SH3 domain, we chose the c-Crk SH3-N domain for its interesting specificity profile and availability of experimental data. We decided not to use the Src SH3 domain because its specificity profile is a simple polyproline peptide, and we wanted to attempt recapitulation of more complex sequence features. Only Type II SH3 ligand orientations were considered for this study. For the PDZ domain, we chose the PSD-

95 PDZ3 domain, as this is the first protein represented by the "PDZ" acronym (PSD-95, Dlg, ZO1).

PWM P-values

The Frobenius norm is calculated simply by concatenating all columns of a matrix into a single vector, then computing the Euclidean distance between this vector and another, similarly formed vector. To evaluate the significance of a given reference-prediction PWM distance, we transformed the value of the Frobenius norm into a P-value by scoring 1000 random PWMs against the reference PWM, thus defining a Gaussian distribution from which P-values could be obtained. Random PWMs were generated as follows. For each PWM position p, 20 random residue weights $\{w_{p,1}...w_{p,20}\}$ were chosen from a uniform distribution on (0,1), and another random value S_p was chosen from (0,S_{max}) for information content scaling. Each residue weight in position p was then raised to the power S_p before the weight vector sum was normalized to unity. This allowed for both a random distribution of weights for a given position, and a random distribution of information content throughout the PWM. For our reference set, the average of information content per position (excluding anchor positions) was 1.960 bits. This compares favorably with an average 1.965 bits for the random PWMs generated with S_{max} = 21, as opposed to only 0.288 bits for the naïve case of S_{max} = S = 1.

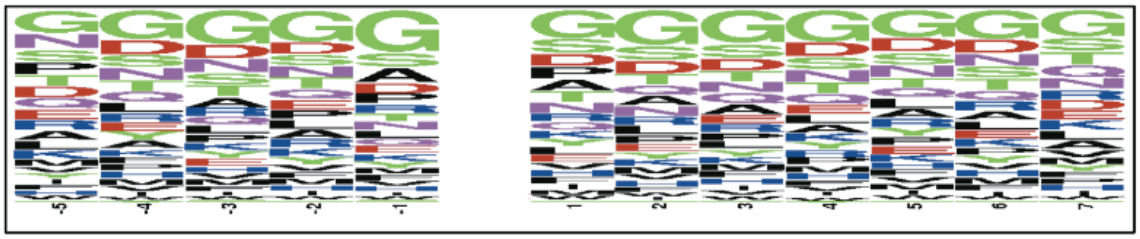


Figure S1. Background PWM logo used for normalization of raw predicted PWMs.

By combining peptide design solutions for a non-redundant set of 103 peptide-binding protein structures, we were able to measure the background sequence bias of our specificity prediction methodology.

Family	Kinase	SH2	SH3	PDZ	FHA
Protein	cAMP-dependent protein kinase A	Src SH2	c-Crk SH3-N	PSD-95 PDZ3	Rad53-1
Bound PDB ID	1L3R	1SPS	1CKA	1TP3	1J4P
Bound Peptide Sequence	TGRRApSIHD	PQpYEEIP	PPPALPPKK	KKETPV	KKMTFQpTPTDPLE
Unbound PDB ID	1BKX	1SPR	1M30	1TQ3	1J4O
Homology Template PDB ID	1O6L	1LCJ	1CSK, 2DRK, 1YWO, 1RUW	2JIN	-
Homology Template Sequence Identity	43%	50%	42%, 38%, 33%, 37%	46%	-
Docking and Constraint Homolog PDB ID's	1O6L, 2BIL, 2PHK, 2Q0N	1AYA, 1AYB, 1AYC, 1H9O, 1I3Z, 1LCJ, 1LKK, 1TZE, 2CI9, 2CIA, 2HDX, 2IUH, 2IUI, 2VIF	1N5Z, 1OEB, 1SEM, 1SSH, 1UJ0, 1UTI, 1W70, 1YWO, 1ZUK, 2AK5, 2DF6, 2DRK, 2O9V, 2V1R, 2VWF, 2W0Z, 2W10	1MFL, 1MFG, 1BE9, 1L6O, 1RZX, 1V1T, 1W9O, 2I04, 2I0I, 2I0L, 2QT5	-
Docking-only Homolog PDB ID's	1O6L	1LCJ	1SSH	2G2L	2PIE

Table S1. Summary of Protein Data Bank structures used in this study.

Frobenius Norms							
Protein	**cAMP-dependent protein kinase A**	**Src SH2**	**c-Crk SH3-N**	**PSD-95 PDZ3**	**c-Crk SH3-N (model 2)**	**c-Crk SH3-N (model 3)**	**c-Crk SH3-N (model 4)**
bound +fixbb	1.261	0.901	1.254	1.313	-	-	-
unbound +csts	1.443	1.445	1.277	1.68	-	-	-
homol +csts	1.401	1.511	1.041	1.847	1.377	1.253	1.252
homol -csts	1.362	1.457	1.519	1.753	-	-	-
P-values							
Protein	**cAMP-dependent protein kinase A**	**Src SH2**	**c-Crk SH3-N**	**PSD-95 PDZ3**	**c-Crk SH3-N (model 2)**	**c-Crk SH3-N (model 3)**	**c-Crk SH3-N (model 4)**
bound +fixbb	2.57E-07	1.56E-09	8.94E-09	2.08E-06	-	-	-
unbound +csts	5.39E-05	1.05E-02	2.18E-08	2.07E-02	-	-	-
homol +csts	1.76E-05	3.09E-02	7.67E-13	1.92E-01	8.27E-09	8.60E-09	7.95E-07
homol -csts	5.84E-06	1.30E-02	6.16E-05	6.32E-02	-	-	-

Table S2. Benchmark PWM prediction Frobenius norms and P-values.

Chapter 3

Improvements in Energetic Modeling of Biomolecular Systems at Protein Interfaces

Abstract

Structure-based design of biomolecular interfaces relies on the accurate prediction of intermolecular energetics and precise recapitulation of experimentally verified molecular geometries. Improvements in the energy function used for design can only be accomplished through quantitative characterization of geometric artifacts and rigorous testing of hypothesis-driven modifications to the energy function. Here, we describe standardized scientific benchmark tests on representative sets of high resolution protein-protein and protein-ligand interface crystal structures for the evaluation and energetic analysis of current computational protocols, then leverage these insights to introduce modifications to our energetic calculations that improve prediction of both structure and sequence in protein design calculations.

Introduction

Biological macromolecules adopt stable structures via minimization of the Gibbs free energy. To predict the conformations of these polymers, it is essential that one employ some scheme to estimate the relative stability of that molecule in different conformations; likewise, when designing biopolymers, one must accurately predict the relative stabilities of different sequences in some structural or functional context. This typically takes the form of an energy function or, more generally, scoring function, that ascribes some value to a given molecular system that correlates to the Gibbs free energy to some degree. There exists a polarity of thought in how best to accurately predict the lowest-energy structures/sequences for a given system. On one end of this polarity, probabilistic and machine-learning techniques leverage large amounts of experimental data to train some function or classifier to discriminate correct from incorrect structures based on some number of features of the input data. These methods are often extremely fast, but require re-parameterization every time a new feature type is included, and may suffer from wild inaccuracies when modeling types of systems for which little experimental data exists. On the other end, physical or theoretical methods attempt to use quantum mechanics and/or statistical mechanics to formally calculate the thermodynamics of the system from first principles given some approximations; the total energy of the system is expressed as a sum of terms with clear physical origins. These methods require little parameterization and can be extremely accurate, but are often computationally intractable for molecular systems of reasonable size.

The Rosetta scoring function falls somewhere between these two extremes by combining physical and statistical terms; the total energy of the system is expressed as a sum of terms with identifiable physical origins, but the functions that

calculate these terms are either refined with experimental data or based entirely on distributions of features in known structures. This approach, while effective, results in significant double-counting of many different physical interactions [1], and makes the direct effects of modification or improvement to any one score term very hard to predict. Only through rigorous testing benchmarks based on high quality experimental data and careful analysis of the interplay between different score terms may one hope to significantly improve such an energetic model.

Here, we describe the creation, testing, and analysis of automated computational protein design benchmarks utilizing high quality crystal structures of protein-protein and protein-ligand interfaces. The components of the Rosetta energy function driving predictions away from the accurate reference data are then elucidated and the effects of different modifications to the energy function are explored. Score function perturbations often have mixed effects on predictive performance, such that the net effect of a substantial modifications may appear to garner little or no benefit; however, computational tools for statistical analysis and visualization of these complex, interdependent effects were developed to ascertain not only the net performance of a given score function, but also the shifting interplay and balance between different score terms and the physicochemical contexts in which erroneous predictions are made.

Methods:

Score Function. In Rosetta, each molecule is represented as a polymer of one or more subunits, or residues. The total score of a biomolecular system is calculated as the sum of all residue-residue pairwise scores. Each pairwise score is calculated as a weighted sum of Rosetta score terms. The Rosetta all-atom scoring function employs a variety of physical and statistical terms to account for the physical forces known to drive biopolymer folding and geometric features found in experimentally determined biomolecular structures. Van der Waals forces are accounted for a with a Lennard-Jones model, with atomic radii and well depths adapted from the CHARMM19 parameter set [2]. Electrostatic interactions in protein-protein systems are scored with an atom-atom distance potential based on atom-type pair distances observed in protein crystal structures, or alternatively a Coulomb potential with a distance-dependent dielectric. In addition, hydrogen bonding is accounted for explicitly with an orientation-dependent potential based on donor-acceptor distances and angles in crystal structures. Solvent effects are represented implicitly using the model of Lazaridis and Karplus [3].An amino acid-dependent backbone dihedral potential has been derived from the Ramachandran backbone dihedral distributions found in crystal structures, and sidechain dihedral propensities are captured in a statistical potential derived from backbone-sidechain dihedral probability correlations observed in crystal structures [4]. Lastly, each residue type is associated with a constant, context-independent score, calibrated to reflect the average energy of that residue type in the reference (unfolded) state. The standard Rosetta score function with corrections, including the modified Dunbrack02

rotamer library, implemented by Yifan Song [1], is denoted as the standard Rosetta score function, "score12".

Energy Function Terms: Classification and Definitions

Solvation	fa_sol – implicit solvation energy
Electrostatics	fa_pair – statistical non-bonded atom pair distance potential (captures electrostatics) hack_elec – Coulombic electrostatics
Hydrogen Bonding	hbond_sr_bb – short-range backbone hydrogen bonds hbond_lr_bb – long-range backbone hydrogen bonds hbond_bb_sc – backbone-sidechain hydrogen bonds hbond_sc - sidechain-sidechain hydrogen bonds
Sidechain Torsional Strain	fa_dun – rotamer probability score (captures intra-residue strain) fa_intra_rep – intra-residue Van der Waals repulsive
Backbone Torsional Strain	p_aa_pp – backbone torsion score (from probability of amino acid given phi/psi) rama - another backbone torsion score (from prob. of phi/psi given amino acid) pro_close – penalty for proline rings not closed properly omega – penalty for non-planar backbone omega geometry
Disulfide Bonds	dslf_ss_dst – disulphide S-S distance restraint dslf_cs_ang – disulphide C-S angle restraint dslf_ss_dih – disulphide S-S dihedral restraint dslf_ca_dih – disulphide Cα dihedral restraint
Reference State	ref – reference (unfolded) state energy for design calculations. Should reflect average energy of residue type in unfolded state. Actually a residue type-specific slack variable/calibration factor.

Energy Function Term Scaling Coefficients

$$V_{design} = V_{interaction} + V_{hbond} + V_{dslf} + V_{struct}$$

$$V_{interaction} = 0.8 \cdot V_{fa\_atr} + 0.44 \cdot V_{fa\_rep} + 0.65 \cdot V_{fa\_sol} + 0.49 \cdot V_{fa\_pair} +$$

$$V_{hbond} = 0.585 \cdot V_{hbond\_sr\_bb} + 1.17 \cdot V_{hbond\_lr\_bb} + 1.17 \cdot V_{hbond\_bb\_sc} + 1.1 \cdot V_{hbond\_sc}$$

$$V_{dslf} = 0.5 \cdot V_{dslf\_ss\_dst} + 2 \cdot V_{dslf\_cs\_ang} + 5 \cdot V_{dslf\_ss\_dih} + 5 \cdot V_{dslf\_cs\_dih}$$

$$V_{struct} = V_{pro\_close} + 0.2 \cdot V_{rama} + 0.5 \cdot V_{omega} + 0.56 \cdot V_{fa\_dun} + 0.32 \cdot V_{p\_aa\_pp} -$$

Natural Protein-Protein Interface Benchmark Set. The protein-protein interface benchmark crystal structure set was obtained from a subset of the ZDOCK 4 benchmark set [5]. It excludes obligate complexes, complexes formed due to crystal contacts, and excludes complexes that show redundancy at the family level in SCOP [6]. Crystal structures of the unbound forms of every protein in the set are available, as are electron density maps. In total, 152 structures with a resolution of 3.0 Å or less were included in the test set. The monomeric protein benchmark set includes 69 crystal structures with a resolution less than 1.9 Å for which electron density maps could be obtained. These were selected to include all-alpha, all-beta, and alpha-beta proteins [7]. For each protein structure in the benchmark set, the residues of interest are first identified. For structures that contain intermolecular interfaces, only the interfacial residue's sidechains are optimized, whereas all residues' sidechains are optimized in monomeric proteins. Interfacial residues are defined as follows: any residue with any sidechain heavy atoms within 3.5 Å of another heavy atom in another molecule. Any residue containing a sidechain atom with less than 0.5 occupancy is ignored. Electron density maps (2mFo-DFc) were obtained for each protein in the benchmark sets via the Uppsala Electron Density Server (EDS) [8]. Each residue in the test set is first optimized (see below) using the electron density as a strong scoring constraint [9]. If the sidechain optimized using the electron density fits the electron density better (has a lower electron density constraint score), this optimized sidechain is treated as the "native" rotamer.

Designed Protein-Ligand Interface Benchmark Set. We collected 18 crystal structures from a diverse array of designed ligand-binding proteins and compared

each to the corresponding computational model that originally drove its experimental production and testing. The set includes the following: two designs of *de novo* esterase (Richter, F., unpublished); three designs to perform a Kemp elimination (Rothlisberger, unpublished); one design of a *de novo* Morita Baylis Hillman catalyst (Bjelic S., and Nivón L.G., unpublished); one design of a phosphorylated-ester binding protein (Nivón, L.G., unpublished); three designs of a binding protein for digoxigenin (Tinberg, C., unpublished); four designs to form a catalytic triad for hydrolysis (Sridhar unpublished); three designs to catalyze a Retro-Aldol reaction (Althoff, unpublished); and one designed to cleave organophosphate compounds (Sagar and Yakov, unpublished).

Sidechain Recapitulation Testing. Next, for each residue undergoing testing, the lowest-scoring sidechain conformation is identified. For rotamer recovery, the chemical identity of the residue is held fixed; for sequence recovery, it is allowed to sample all 20 canonical amino acids. Sidechains are optimized within the context of the experimental structure; that is, only one sidechain is optimized at a time, while all others remain fixed. First, all sidechain degrees of freedom undergo gradient-based minimization until convergence, and score data is saved. Rotamers are then "repacked", optimized via iterative sampling of the sidechain conformations in the Dunbrack Rotamer Library [4]. The chi angles corresponding to each rotamer bin center, plus 2 half standard deviations on each side of that bin, are sampled iteratively. Each rotamer conformation is subjected to gradient-based minimization before scoring again. The lowest-scoring rotamer is saved as the Rosetta prediction.

Weight Fitting. The weights for linear combination of Rosetta score terms, along with residue-specific reference energies, were optimized using optE (Leaver-Fay, unpublished). This application employs an iterative particle-swarm scheme for optimizing score term weights and reference energies to maximize sequence recovery for predefined residues across some number of input protein structures.

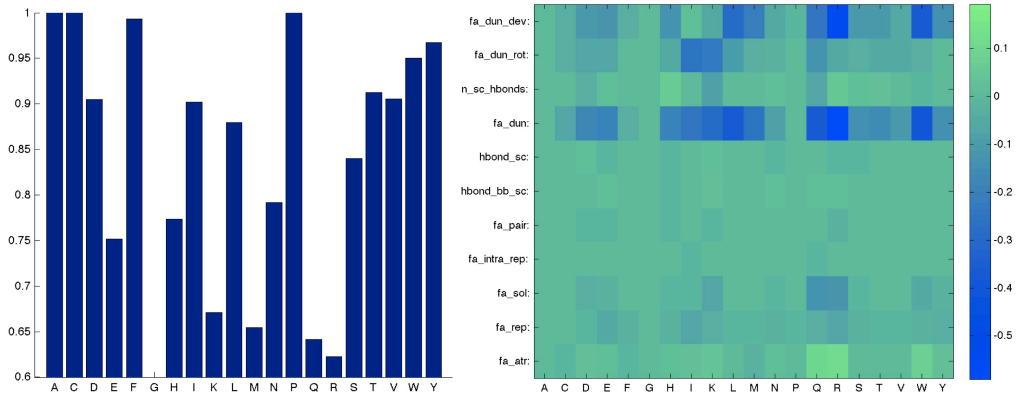
Sidechain Recovery Statistical Calculations. Before analyzing the results of sidechain recovery simulations, residue positions are filtered to remove any possibly erroneous experimental data. All crystal structure B- factors are normalized according to [10], and a normalized sidechain B-factor is calculated by averaging over normalized heavy atom B-factors. Any residue with a sidechain B-factor greater than 1.0 is ignored. Also, any sidechain detected to clash with another residue (LJ repulsive score greater than 5.0) in the native structure or in the gradient-minimized structure is also ignored.

After filtering, sidechain recovery is measured. Recovery rates were always normalized by residue type counts in the dataset; that is, relative abundance of one residue type over another in the dataset is corrected for and has no direct effect on recovery rate. For fixed-sequence simulations, rotamer recovery is gauged by whether or not the lowest-scoring rotamer has a sidechain heavy atom RMSD of less than 0.8 Å to the crystal structure sidechain. For design testing, success is gauged only by whether or not the correct residue type is recovered. Score covariances were analyzed as follows. For each residue, the change in each score term between

the native and predicted sidechain was saved. For each residue type, the covariance between the sidechain heavy atom RMSD and the change in each score term was measured. For sequence recovery testing, the RMSD was either set to 0 for a match or 1 for a mutation. Correlation coefficients were calculated from the RMSD-Δscore covariances by normalizing each covariance by the geometric mean of the variables' variances.

Results:

Sidechain Structure Prediction in Evolved Protein-Protein Interfaces and Energetic Analysis of Artifacts. For the protein-protein interface benchmark, rotamer recovery stands at 80.8%. As seen in Fig 1A, recovery at protein interfaces fares particularly poorly for K, M, Q, and R residues. This is not surprising to some extent, as R and K have four sidechain degrees of freedom, while M and Q have three, allowing for a greater number of chances for scoring errors during sidechain sampling. Which components of the score function are primarily responsible for deviations from the conformations that best fit the electron density? To answer this, we may compute the covariance between the change in each component of the score function and the sidechain RMSD of the rotamer repacked into the electron density vs. the rotamer repacked using only the score function. That is, for each residue type, how does the degree of "wrongness" of the prediction vary with the change in each score term? This can be visualized as the covariance matrix heatmap in Fig. 1B. Here, darker blue indicates a stronger negative correlation between the change in score and the change in structure; that is, darker blue indicates the score terms that are becoming more negative, and hence driving the change in structure. The row labeled "fa_dun", corresponding to the Dunbrack rotamer score, dominates the signal in the heatmap, pointing to this term as one of the primary culprits in failure of rotamer recovery.



1A 1B

Figure 1. Protein-protein interface rotamer recovery rates (1A) and covariance matrix (1B) for the standard Rosetta score function.

The rotamer score is composed of the sum of two terms: a rotamer probability score, and a rotamer deviation score. The first term reflects the log probability that a residue with given backbone angles exists in a given rotamer bin, and the second calculates a harmonic potential penalty for chi angle deviations from the center of that rotamer bin (Eqn. 1).

Eqn.1

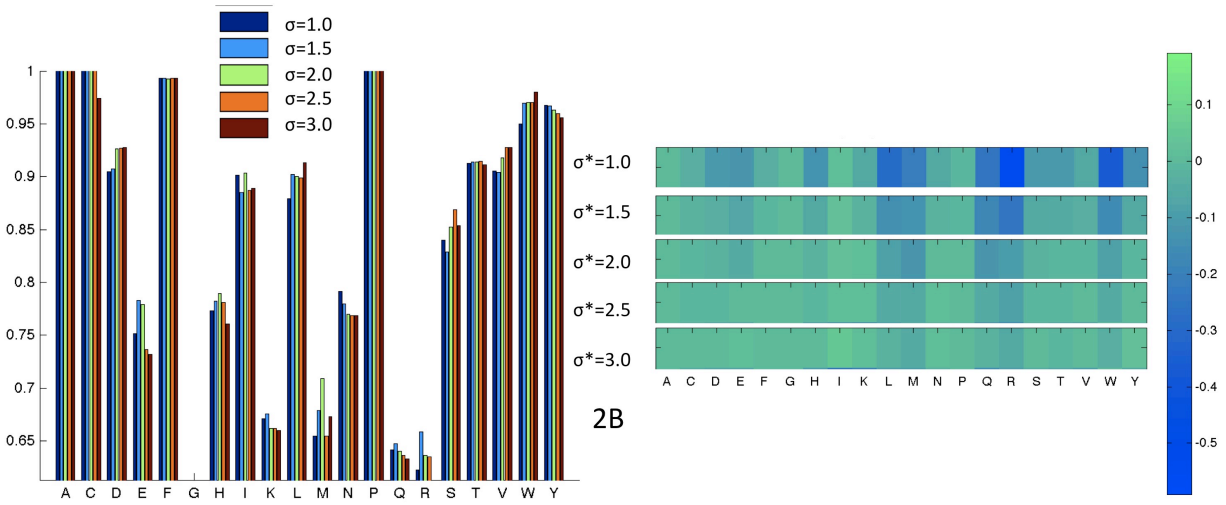
$$E_{dun} = E_{dun\_rot} + E_{dun\_dev}$$

$$E_{dun\_rot} = -\log(P_{rot}(\bar{\chi} \mid \Phi, \Psi))$$

$$E_{dun\_dev} = \sum_{chis} \frac{\chi - \mu}{\sigma}$$

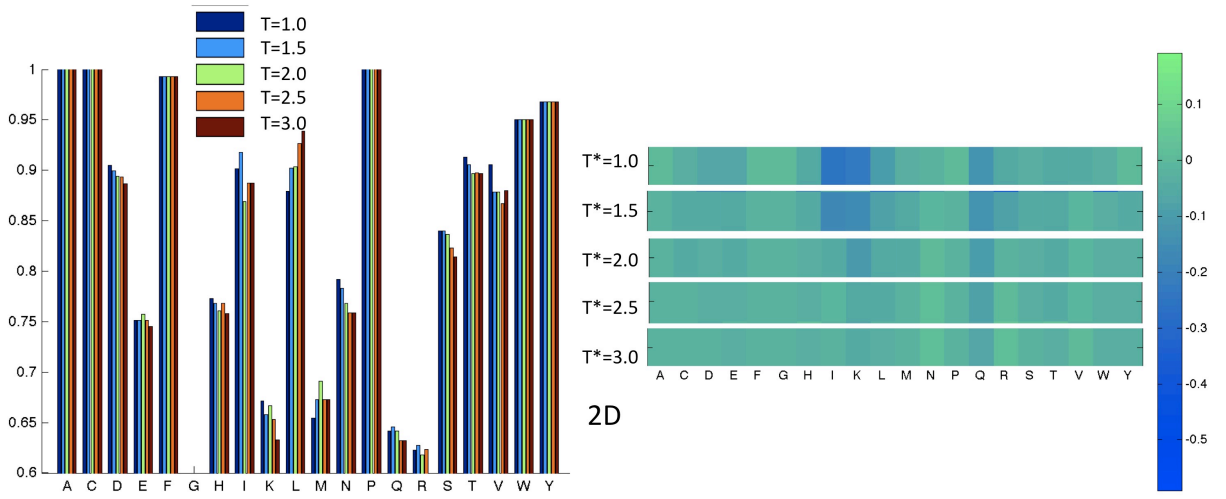
These terms are labeled “fa_dun_rot” and “fa_dun_dev”, respectively, at the top of the covariance heatmap in Fig. 1B. For L, M, Q, R, and W, we see a strong RMSD covariance with the drop in rotamer deviation score. To correct for these errors, we may decrease the chi deviation penalties by rescaling the rotameric chi standard deviations, effectively reducing the spring constant on the harmonic deviation potential. The results of this can be found in Fig. 2A. For each residue type, rotamer recovery peaks at some standard deviation scaling factor between 1.0 and 3.0. In Fig. 2B can be seen the RMSD - chi deviation penalty covariances for the different scaling factors and, as expected, we see a reduction in the signal corresponding to the chi deviation-driven errors. By combining these optimized rescaling factors for their respective residue types, average rotamer recovery can be increased from 80.8% to 82.3%.

In Fig. 1B, it can be seen that the rotamer probability score also has a substantial effect on rotamer recovery, especially for I and K residues. Thus, the effect of the rotamer probability score was adjusted in a similar fashion by smoothing the probability distributions for each residue type. That is, each rotamer probability was transformed into an energy via Eqn. 1, those energies were divided by some constant “temperature” factor, then transformed back into probabilities in the inverse fashion. This has the same effect as increasing the temperature of the Boltzmann probability distribution. The results of this can be seen in Fig. 2C and 2D. The expected effect is observed, but the changes in rotamer recovery are quite modest.



2A

2B



2C

2D

Figure 2. Protein-protein interface rotamer recovery rates for rotamer chi deviation penalty smoothing (2A) and probability density penalty smoothing (2C). The covariance matrix rows corresponding to the chi deviation penalty (2B) and the probability density penalty (2D) for different smoothing factors are shown. The covariances between these scores and RMSD decreases as the potentials are smoothed.

Sequence Prediction in Evolved Protein-Protein Interfaces and Energetic Determinants of Artifacts. If we use a protocol identical to the rotamer recovery protocol, but allow all residues at each position, we get sequence recovery data like that shown in Fig. 3A. Here, success is judged not by recovery of the native rotamer, but simply by whether or not the original residue type is recovered. In Fig. 3B, we

see that the rotamer score (fa_dun) remains a problem across multiple residue types. One way to address this is to apply the chi standard deviation rescaling parameters that were found to be optimal for each residue type in rotamer recovery; we can rescale the standard deviations by a different factor for each residue type. The results of this are also shown in Fig. 3A. Overall sequence recovery remains at 49%, and the effects here are much less clear, as some residues' recovery rates increase, while others decrease. Smoothing of the overall rotamer probability distributions has a similarly small effect on sequence recovery (data not shown).

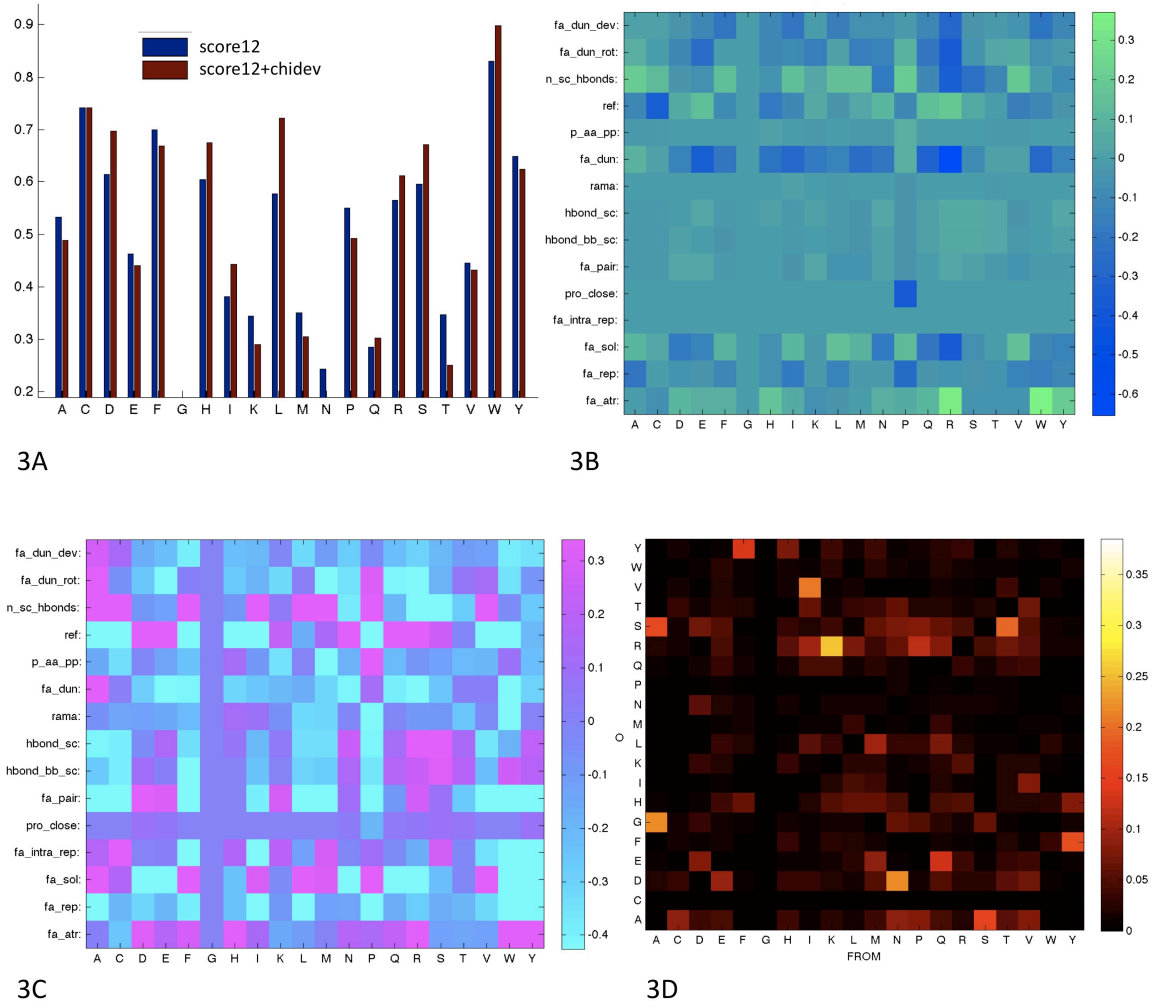


Figure 3. (3A) Protein-protein interface sequence recovery rates for single-residue design testing with standard score12 weights and with optimized chi deviation penalty smoothing. Covariances (3B) and correlation coefficients (3C) for score term changes with design failures. The heatmap rows in 3C are normalized by the variances of each associated score term, allowing for better visualization of more subtle score changes. The transition matrix in 3D shows only mutations, and has been normalized by sequence counts in the original benchmark set (by column).

What other effects are driving sequence recovery failure? To visualize this, we may take our covariance matrix, and normalize each covariance by the variances of the two observables. This way, each element of the correlation matrix is normalized within the change-in-energy range of that score term for that residue type. This allows us to better visualize which score terms are varying with RMSD/sequence recovery. This heatmap can be seen in Fig. 3C. Now we can see that reductions in solvation, the atom-pair term, and hydrogen bonding are also making contributions to sequence recovery failure. In particular, the solvation term shows a noticeable covariance and strong correlation with residues D, E, N, Q, and R. To get a better understanding why, it is useful to know which residue types are mutating into other residue types. This is visualized in Fig. 3D, where each element the number of native residue types (X-axis) that mutated into another residue type (Y-axis). A number of consistent mutations appear, but in particular there exist strong preferences for N->D and Q->E. As seen in Fig. 3A, sequence recovery fared particularly poorly for Asn and Gln residues. This was due in large part to the predicted solvation free energy penalty of water molecule occlusion by other surrounding residues (solvation score). In the Lazaridis-Karplus solvation model, each atom type is associated with a parameter controlling the energetic effect of complete burial (desolvation) of that atom type, ΔG_{free}^{solv}. In Rosetta, the ΔG_{free}^{solv} magnitudes for terminal carboxamide N and O atoms were increased over those published in the original model. This was done to compensate for erroneous design of Asn and Gln residues into the core of proteins (Brian Kuhlman, personal communication). Reversion to the original values leads to drastic improvement in recovery of Asn and Gln residues, as seen in Fig. 4A, "score12+lk". However, this also has the effect of a smaller decrease in performance for some hydrophobic residues and a larger problem with recovery of acidic Asp and Glu. Because of the nature of Rosetta's heuristically motivated score function, it is possible to address these new concerns, not by compromising on the solvation ΔG_{free}^{solv} parameters, but by modifying other score terms.

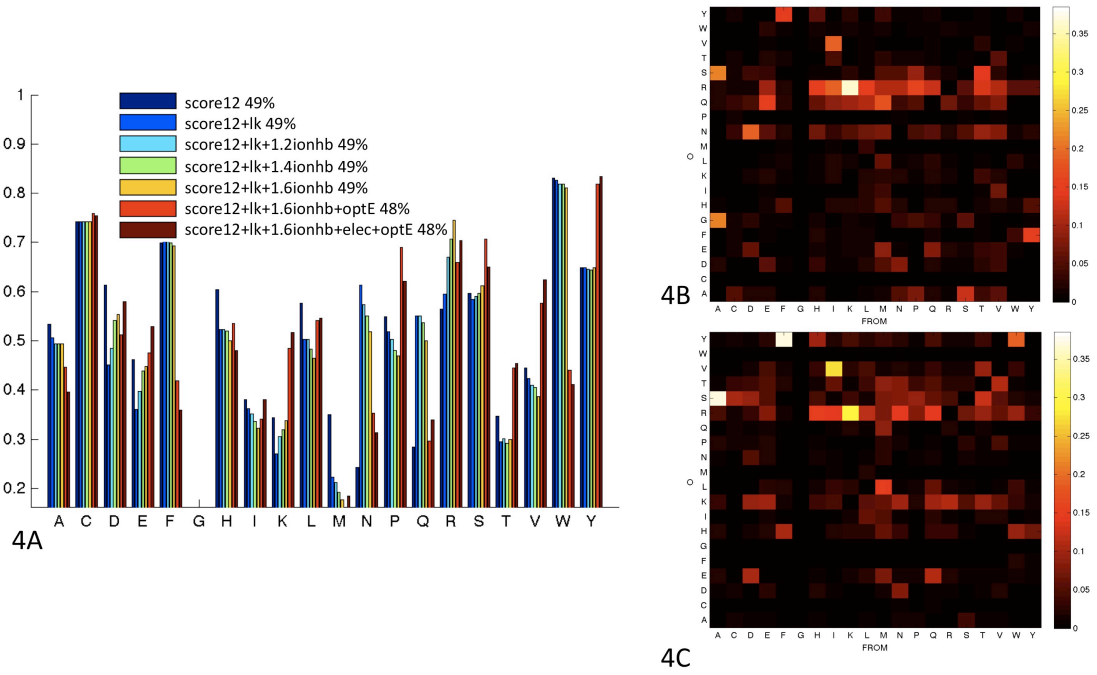


Figure 4. Protein-protein interface sequence recovery rates for score function modifications. (4A) Sequence recovery rates for six different score function modifications. Residue-residue transition probabilities comparing native sequences to design sequences for the original score12 function (4B) show substantial erroneous mutations of hydrogen bonding residues, whereas the modified score12+lk+1.6ionhb+optE function (4C) shows a significant reduction in these errors.

Rosetta's standard hydrogen bonding potential rewards polar and charged interactions equally. By modifying this potential to preferentially reward polar-charged and charged-charged hydrogen bonds, it is possible to recover accurate prediction of Asp and Glu residues (Fig. 4A, "score12+lk+1.6ionhb"). Unfortunately, upweighting of charged hydrogen bonds further decreased performance in predicting hydrophobic residues. This is to be expected because of the way in which hydrogen bonding and solvation must be carefully balanced against one another. At buried residue positions, such as the core of a protein, the reward for creating a hydrogen bond must be balanced against the penalty for burying it away from solvent. By raising only the strength of hydrogen bonds, one might expect erroneous prediction of buried hydrogen bonds in hydrophobic areas. This is exactly what is seen in Fig. 4B, with substantial mutations from hydrophobic residue into R, Q, and N. To correct for this, we may raise the weight of the solvation term to further reward buried hydrophobic residues. The solvation weight and the reference energies were optimized simultaneously using optE by training on the protein-

protein interface dataset. As expected, the solvation weight rose from 0.65 to 0.76. Unfortunately, reweighting the solvation term and reference energies had almost no effect on overall sequence recovery, which remained at 48%. As seen in Fig. 4A, "score12+lk+1.6ionhb+optE" and in Fig. 4C, increasing the strength of the solvation term reverts almost all the gains that were recovered in changing the N and Q solvation ΔG_{free}^{solv} parameters. Perhaps the electrostatics may need readjustment? To explore this, optE was re-run, replacing the statistical pair term with a more physical Coulombic electrostatic term, and allowing solvation, hydrogen bonding, and electrostatics to all vary their weights in a concerted manner. This, however, affected sequence recovery very little, as evidenced in Fig. 4A, "score12+lk+1.6ionhb+elec+optE". Interestingly, during weight optimization for this final weight set, the hydrogen bond weight decreased from 1.1 to 0.46, whereas the electrostatics term increased to 0.76, indicating that ionically strengthened hydrogen bonds may better be taken into account simply with an electrostatics bonus.

Structural Deviations in Designed Ligand-Binding Proteins and Energetic Analysis of Crystal Structures. By computing the geometrical deviations between computational models of inactive designs and their actual crystal structure conformations, we may evaluate the Rosetta energy function in the context of non-natural interfaces and quantify the relationship between sidechain structure prediction and mutation-driven deviations in the backbone of the original design template protein. Computational models used for the design of 18 non-natural proteins encompassing varying ligand-binding interfaces were compared to their associated crystal structures to explore the performance of current design algorithms. All designs were generated with a combination of previously-described computational protocols and manual design modifications by human designers based on visual inspection and physicochemical intuition. Rotamer recovery was calculated by residue type over the benchmark set. Hydrophobic residues such as V and F reached the highest recovery at about 90%, whereas long basic residues K and R reached levels of only 30% (Fig. 5A). Longer amino acids have more torsion angles and thus more degrees of freedom and exponentially greater possible conformations to adopt. Recovery of indivual chi angles decreases with successive torsions away from the main chain, as do the polar/charged amino acids on account of their less sterically constrained, more solvent accessible local environments (Fig. 5B). Rotamer recovery was significantly higher for D than N, despite their identical number of torsional degrees of freedom; similarly, F is recovered at greater rates than Y.

Failure to predict sidechain conformations may be driven by structural deviations in the protein backbone, a factor not accounted for in our fixed-backbone design approach. To measure this effect, we compared rotamer recovery summary statistics with whole-structure backbone RMSD to the crystal structure. As expected, we observe a negative correlation; rotamer recovery becomes more difficult as the actual structure of the main chain deviates further from the computational design model (Fig. 5C) with extremely low recovery rates for deviations greater than ~1Å

RMSD. Methods to accurately account for the main chain effects of design mutations in concert with fixed-backbone design simulations will be necessary to more accurately predict the energetic consequences of mutations to the sequence of native scaffolds used for design of protein with non-natural functions.

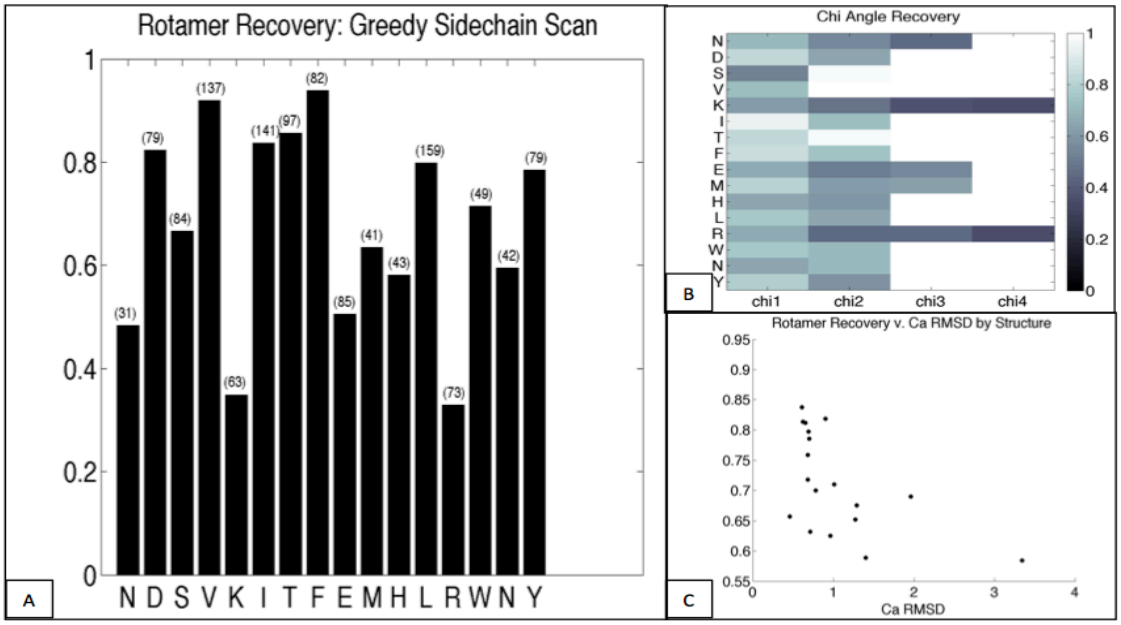


Figure 5**.** Rotamer recovery over a set of computational design/crystal structure pairs. (A) Rotamer recovery over the set of 18 structure pairs by amino acid. (B) χ angle recovery over the set, with higher recovery in lighter gray and lower in darker gray, as indicated in the legend. (C) Rotamer recovery over each pair vs. Cα RMSD shows lower recovery over 1 Å RMSD.

Discussion

Benchmark sets for design must include specificity data or functional context. Development of the Rosetta score function has been focused primarily on protein folding and sequence recovery in monomeric proteins. Protein folding is primarily driven by packing of hydrophobic residues, as these residues are under much more steric constraint, whereas hydrophilic residues on the surface primarily point out into solvent, and surface residues are generally conserved more for functional reasons than structural reasons[11]. Accordingly, whole-protein design typically fares much better for residues in the core for most sequence design methodologies[12],[13]. However, many of the practical applications of protein design (ligand binding, protein-DNA interactions, protein-protein interactions, enzyme design) require accurate modeling of electrostatic interactions of residues on the surface, a factor that has been overshadowed in the evolution of the Rosetta sore function. Though it

is known that protein interfaces are often not optimized for affinity, but for specificity as well[14], sequence conservation at protein interfaces in general is expected to be greater than that of typical non-interface protein surfaces. By testing design in the context of the interface partner, it may be possible to develop a greater understanding and ability to predict the energetics of electrostatically driven interactions. This is the reason that development thus far has focused on protein interfaces. Ideally, one would hope to gauge success of sequence design using actual specificity data: for which residues does successful recovery matter, and for which might failure be irrelevant? It will be difficult to substantially improve the score function for design treating all residue positions as equally important when some are in fact quite variable. Further work will incorporate development of acquisition of such benchmark data.

Reference state energies may need to be precomputed explicitly. For design calculations, it is essential that one estimate the energy of the reference state. For monomeric protein design, the reference state corresponds to the unfolded state ensemble, whereas, for protein-protein interface design, this corresponds to the unbound form of the protein. This is the reason why I have thus far focused primarily on protein-protein interfaces; for every single complex in the protein interface design benchmark, crystal structures exist the unbound forms of both proteins. Future work will allow more explicit modeling of the reference state for multistate design calculations, thus avoiding the problem of reference energy calibration for modification of other score terms, making evaluation of score function improvements much easier to accomplish.

How might the reference state problem be addressed more generally? Currently, the reference energies are simply calibrated on design benchmarks to optimize sequence recovery. This is not actually a reference energy: it is a slack variable, and the belief that it reflects only the reference state energy must be predicated on the belief that the reference state energy is the only energetic phenomena that has not been accounted for. Additionally, reference state energy calibrations can vary wildly with choice of dataset, and are very costly. It would be more ideal to precompute the reference state energies directly. In the protein design package DESIGNER[12], Jaramillo et al. precompute the reference energy for each residue type by taking a Boltzmann average over all sidechain rotamer degrees of freedom of a single amino acid with ideal propensity phi/psi angles. This allows them to simply recomputed these energies when score terms are added or modified, rather than fitting 20 free parameters on an arbitrary dataset every time. Ron Jarnak has attempted calculation of an explicit unfolded state energy in Rosetta by averaging energies of each residue type over a large set of protein fragments from the PDB, and has met with substantial success, but not quite to the level of the current reference energy fitting methodology (personal communication). This is because he failed to re-fit the reference energies with the inclusion of the unfolded state energies; introduction of 20 free parameters will always improve a model, but, if the magnitude of these corrections became smaller, one might be led to believe that the unfolded state energy was, by some estimation, being properly accounted for.

It may be necessary to estimate conformational entropy. The "folded state" exists in the realm of thermodynamic systems. There are some number of actual physical microstates that correspond to the "folded" macrostate. How we choose to define the "folded state" has a direct impact on the folding free energy of a protein. Thermodynamic system are not physical systems. A crystal structure exists in the realm of physical systems; it corresponds to one and only one microstate, that is, the 3-D coordinates of each atom. The Rosetta score function is calibrated to estimate free energies. How is this distinct from estimating enthalpies? The free energy should include an entropic correction to the enthalpy, because lower-enthalpy states may actually be unlikely to happen because of a greater entropic penalty. What is the change in the number of states in going from the unfolded to folded state? For an atomistic model, the entropy of the protein is always zero; it corresponds to exactly one microstate, that is, the coordinates of every atom. Thus, the folded state in Rosetta corresponds to exactly one protein microstate plus all the solvent microstates compatible with this structure. Conceptually, we are assuming a protein near 0° K but water at 298° K when we attempt to represent the folding free energy using only one structure. Why then, does the Rosetta score function appear to work fairly well for many protein folding applications? This is because success of the score function is gauged primarily upon recapitulation of features and structures seen in high-resolution crystal structures. It is a central dogma of structural biology that a crystal structure well represents the conformation of a protein at its global free energy minimum. The entropy of a perfect crystal lattice is 0; conceptually, the "effective temperature" of a crystal structure approaches 0° K. Thus, it is no surprise that the low-temperature approximation in Rosetta recapitulates structures and distributions of features in crystal structures. However, problems may occur when functional concerns become our criteria for success, specifically in design problems where success is gauged by whether the design performs some function at room temperature, where the effects of conformational entropy may be more pronounced.

How might one estimate conformational entropy? Xiang et al. have developed a method for estimating sidechain conformational entropy using the colony free energy, and were able to substantially increase rotamer recovery in monomeric proteins[15]. The same group also found improvements in loop prediction by adding to the force field a term that rewards structures with a low RMSD to other sampled structures[16], thus biasing predictions to structures with wider energy basins. Alternatively, one might imagine a MC/MD short simulation around the structure/sequence of interest and Boltzmann averaging all the structures deemed to fall within a certain radius of this structure. One could also attempt to estimate the degree of "flexibility" of a given conformation or sequence. For instance, if a given loop design was found to contact a ligand in a greater variety of distinct conformations in a short simulation, it could be given an entropic bonus to its total score. Kellogg and others have already developed tools in Rosetta for unbiased Monte Carlo sampling (unpublished) which can be used for generation of ensembles in the neighborhood of a target structure.

Chapter 4

Removing T cell Epitopes with Computational Protein Design

Disclosure: This chapter has been submitted for publication and is under review as

Chris King, Esteban N Garza, Ronit Mazor, Jonathan L Linehan, Ira Pastan, Marion Pepper, David Baker. Removing T cell Epitopes with Computational Protein Design. *PNAS.* (Submitted)

My contributions consisted of devising the algorithms and protocols described, implementing those protocols in the Rosetta molecular modeling package, performing calculations and analyzing data, and writing the full text of the paper.

Abstract:

Immune responses can make protein therapeutics ineffective or even dangerous. We describe a general computational protein design method for reducing immunogenicity by eliminating known and predicted T cell epitopes and maximizing the content of human peptide sequences without disrupting protein structure and function. We show that the method recapitulates previous experimental results on immunogenicity reduction, and use it to disrupt T cell epitopes in GFP and Pseudomonas exotoxin A without disrupting function.

Introduction

Immunogenicity is a major problem in the development of protein therapeutics. Repeated administration of a protein therapeutic can lead to B cell activation and production of antibodies rendering the therapeutic clinically ineffective or cross-reacting with host proteins (3). Affinity-maturation of antibody-producing memory B cells is initiated by T cell recognition of peptide epitopes displayed on MHC class II (MHCII) proteins on the surface of mature antigen-presenting cells. Immunogenicity may be reduced by eliminating known T cell epitopes from the protein sequence and/or increasing the prevalence of sequences already found in the host genome to which T cells would already be tolerant, an approach that has met with substantial clinical success in the humanization of recombinant antibodies (4). However, unlike antibodies, which have been extensively characterized, the mutational tolerance of most proteins is generally not known, and hence the extension of this approach to proteins of arbitrary structure and function remains a major challenge. Deimmunization efforts have relied for the most part on experimental characterization of a large number of point mutants followed by combination of individual mutations (2, 5).

To reduce or eliminate immunogenicity, it would be desirable to have a method that eliminates MHCII-binding epitopes and increases host sequence content without disrupting interactions essential for proper folding and function. The peptide-binding repertoire of many MHCII alleles has been extensively characterized (6), and a number of methods have been developed for predicting the affinity of novel peptides for a given MHCII (7). Coupling of epitope prediction methods with methods for predicting the structural and functional consequences of mutations offers the possibility of reducing the immunogenicity of a target protein without disrupting structure and function. Epitope prediction methods, homolog substitution matrices, and structural stability calculations have been combined to predict optimal epitope-eliminating mutations (8, 9). Epitope prediction methods have been integrated with structure-based protein design (10) by combining the 9mer epitope PROPRED matrices with protein design of all residues in a flexible backbone method that allows substantial redesign of protein cores. The combined method was able to eliminate epitope-like sequences while maintaining native-like values for a number of predicted protein stability metrics, but folding, function, and immunogenicity were not evaluated experimentally.

Here, we describe the integration of the Rosetta computational protein design method with experimental immunogenic epitope data, MHC epitope prediction tools, and host genomic data to enable the design of proteins with reduced immunogenicity while retaining function and stability. Our approach goes beyond PROPRED by implementing a more accurate machine-learning based epitope prediction method for 28 different H-2, HLA-DR, and HLA-DQ alleles, restricts design to select 15mer epitope regions, and utilizes a greedy stepwise protein design algorithm (11) to eliminate the most immunogenic epitopes with as few mutations as possible, avoiding disruptive core mutations likely to destabilize the protein. We compare the performance of our epitope predictor to PROPRED and another leading epitope prediction method for thirteen different human and mouse MHC alleles, demonstrate the effectiveness and generality of the method with *in silico* tests on previously characterized deimmunized protein targets, and show experimentally, for GFP in mice and Pseudomonas exotoxin A in humans, that the method eliminates T cell epitopes without disrupting function.

Results

Overview of computational method. For a given target protein and set of host MHC alleles, potential T cell epitopes are first identified using a support vector machine (SVM). These regions are then optimized to eliminate the T cell epitopes while retaining structure and function using an extension of the Rosetta all-atom protein design methodology with modifications to both the energy function used in the design calculations and the optimization procedure.

The energy function used in the sequence optimization is supplemented with two terms that incorporate immunologically relevant data. The first term calculates predicted epitope content using support vector machines (SVMs) trained with experimentally determined peptide-MHC binding data. Scores from SVMs for each MHC allele in each 15mer sequence frame are averaged, then summed over each

frame. The second term utilizes known host genome 9mer data and known epitope data, rewarding each host 9mer in proportion to its frequency of occurrence in the host genome, and penalizing known epitopes. Both deimmunization scores favor negatively charged residues on the surface of the protein, hence we also introduce a net charge constraint into the total objective function, penalizing deviations from the input protein formal charge. By weighting this term appropriately against the deimmunization terms, negatively charged residues may be placed at critical epitope-disrupting surface positions while compensatory positively charged residues are introduced at other positions.

Sequence optimization is carried out using a protocol that focuses on solutions that reduce the energy with a relatively small number of mutations and allows use of computationally expensive objective functions in design calculations that otherwise might be unacceptably slow. The energies of all point mutants at each design position are first computed and then sorted. At each position, all point mutants within a certain threshold of the lowest energy mutant are saved for subsequent combination. These lowest energy point mutations are then combined using a greedy stepwise, steepest descent heuristic (see Methods), allowing for structural relaxation at each step, until mutations have been attempted at all design positions. Multiple, diverse near-optimal designs can be generated in parallel by stochastically accepting the placement of near-optimal mutations during the combination process.

T cell Epitope Prediction. We initially trained our SVMs on experimentally measured MHC binding affinities for 26 allelic variants (12). For each MHC allele type represented in the training set, we collected all epitope sequences from the Immune Epitope Database (IEDB) known to elicit T cell activation (6). We restricted further analysis to alleles with at least 25 known T cell epitopes in the IEDB. In the absence of sufficient data on non-binding peptides for each allele, we assumed that most 15mers in the host genome would not constitute strong MHC binders, and generated negative data sets by randomly choosing 1,000 15mers from the human genome (13). We compared our SVM-based epitope predictions to those of previous methods using a subset of T cell epitopes withheld from initial training. Sensitivity and specificity were evaluated over all alleles for our method, NetMHCII-v2.2 (14), and PROPRED (15), and predictive performance was evaluated by calculating the area under the ROC curve (AUC) for each allele type both independently and over the entire set. Since PROPRED contains only matrices for DRB alleles, we combined testing data for the eight DRB alleles covered by all three methods and generated a standard ROC plot (Fig. 1A). The highest AUC was achieved by NetMHCII (AUC: 0.759). Rosetta performed comparably but slightly worse (AUC: 0.752), while PROPRED achieved significantly lower prediction accuracy (AUC: 0.710). Because more T cell epitopes have been characterized for some alleles in the test set, combining all testing data weights performance analysis toward alleles with more data points. If we average AUCs with equal weight over the shared DRB allele set, again, NetMHCII performs best (AUC: 0.792), with Rosetta slightly lower (AUC: 0.785) and PROPRED lower still (AUC: 0.771) (Fig. 1B).

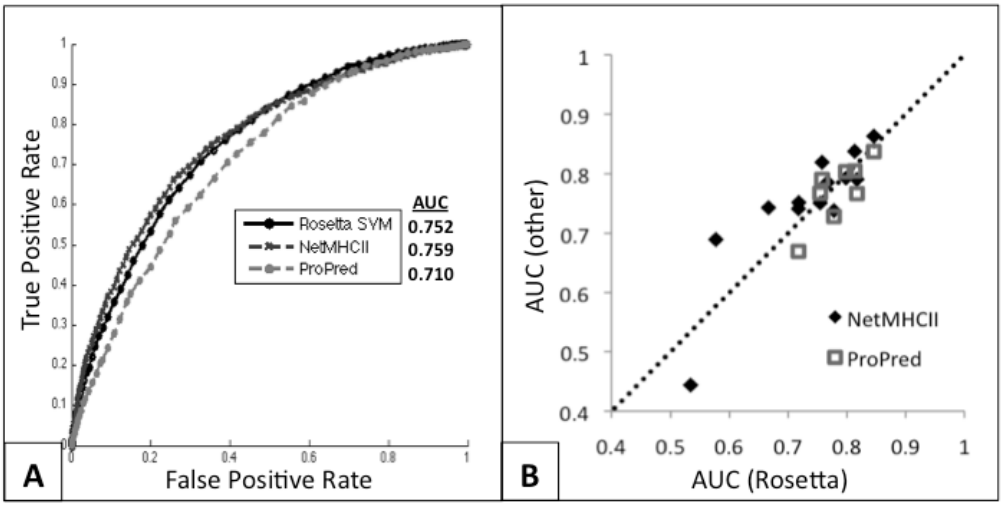


Figure 1. Performance of Rosetta SVM T cell epitope prediction. A) ROC curve true positive rate vs. false positive rate) for all testing data and comparison with current methods. Total Area Under Curve (AUC) is listed for each method. B) Predictive performance over each allele test set. X-axis: Rosetta AUC, Y-axis: other method AUC. Points below the 1:1 dotted line indicate where Rosetta performs better than other methods

Large-scale benchmarking and calibration of design method. We first tested the ability of the method to eliminate putative human T cell epitopes from eight proteins from pathogenic organisms that contain known MHC-binding epitopes. The computational design protocol was used to simultaneously eliminate all predicted epitopes for eight representative human DRB1 alleles (Text S1), collectively covering almost 95% of the human population (16). To evaluate the tradeoff between epitope removal and protein stability, multiple design simulations were carried out with an increasing weight on the SVM-based epitope scoring term. Increasing the weight on this term decreases the number of MHCII predicted epitopes and increases the Rosetta energy (Fig. 2A).

Because amino acid substitutions predicted to disrupt peptide-MHC binding might destabilize the overall protein, a balance between the stability of the protein and disruption of possible MHC binding must be sought. A weight on the epitope scoring term of 2.0 eliminates 79% of the predicted epitopes and 84% of the known epitopes without increasing Rosetta energy above that of the native protein (Table S1). Because sequence changes are permitted only at critical predicted epitope positions, the number of mutations is minimized, thus allowing for substantial reduction in predicted immunogenicity while maintaining average sequence identity at 66%. Similar calculations were carried out varying the weight of the term favoring 9mer sequences found in the human genome (Fig. 2B). As the weight increases, the average number of human genome 9mers over the designed regions increases and the Rosetta energy becomes more unfavorable. At a weight of 3.5, human 9mer sequences increase from 0% to 4.3% of redesigned epitopes while the average Rosetta energy increases only 10% over baseline. These weights were used for the remainder of this work unless otherwise noted in the Methods.

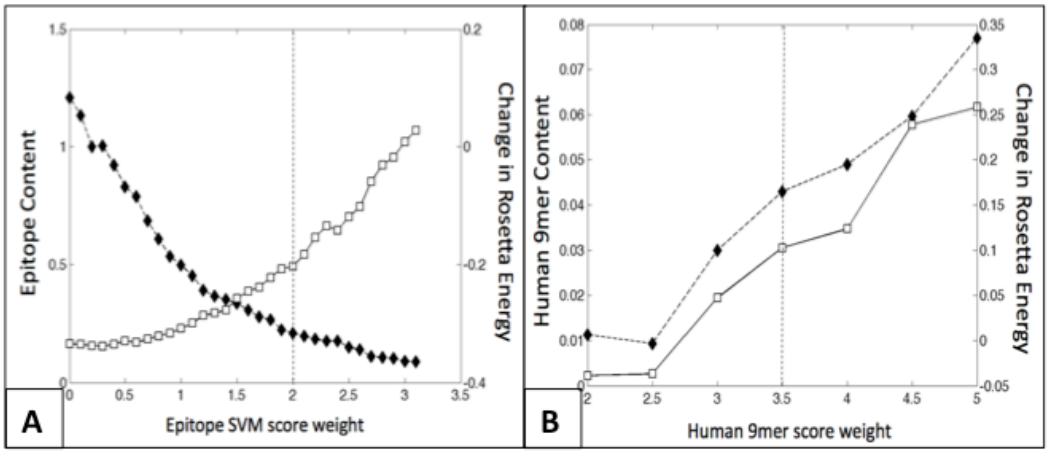


Figure 2. Tradeoffs between Rosetta energy and extent of deimmunization. Rosetta energy (open squares) of redesigned proteins increases while epitope content decreases (A) and human 9mer count increases (B) (filled diamonds) as the weights on the associated score terms are increased.

Recapitulation of previous immunogenicity reduction data. We next attempted to recapitulate the results of previous experimentally validated protein deimmunization efforts where immunogenicity was reduced without disrupting biological function. Cantor et al. (1) sought to remove T cell epitopes from *E. coli* L-asparaginase II (EcAII), an enzyme approved for treatment of acute lymphoblastic leukemia, while maintaining native-like enzymatic activity and protein stability. They first employed a neutral drift selection scheme to identify allowable mutations in each of three predicted HLA-DRB1*04:01 epitopes, then combined mutations in each epitope to produce a fully functional enzyme with reduced immunogenicity *in-vivo*. We applied our protocol to the biologically active tetrameric state of EcAII, constraining design to the residue positions chosen for randomization in the previous study, and compared our predictions against the mutations identified in the selection screen and the predicted immunogenicity of the resultant design sequence. Following Rosetta redesign for all 24 epitope residues, 2 out of 5 mutations assumed residue identities found in the sequences of experimentally isolated activity-preserving variants, with one mutation occurring at a position not tested in the above study. In addition, 2 of 3 redesigned epitopes have predicted HLA-DRB1*04:01 binding affinity lower than the peptides tested in the previous study (Table 1).

Tangri et al. (2) attempted the deimmunization of erythropoietin (Epo), a growth hormone used in treatment of anemia and myelodysplasia. They screened a set of Epo-derived peptides for binding to a set of 15 different HLA alleles, targeted two potential epitopes for deimmunization by screening a small set of point mutants in each epitope for reduction of T cell activation, and combined point mutants and screened for both immunogenicity and biological activity. We applied our deimmunization design protocol to this protein, utilizing the structure of Epo bound to its native receptor and targeting the same epitope positions experimentally explored by Tangri et al. Because the HLA allele set of the human donors in the study is unknown, we designed simultaneously against all eight alleles in the HLA-DRB1 allele set. In the first epitope region, Rosetta introduces only one substitution,

the same mutation found to be optimal in (2). Rosetta similarly recovers the point mutation position in the second epitope (Table 1), but substitutes an alanine to preserve the protein's net charge and makes three additional mutations to further disrupt MHC binding. In both cases, the Rosetta design minimizes the number of predicted MHC binding peptides while minimizing the energy of the protein in the epitope regions.

	Sequence	Predicted IC50 (nM)	Rosetta Energy
EcAII: Epitope 1 (115-123) [rank 6]			
Native	MRPSTSMSA	194.3	-12.0
Cantor et al.	**V**RP**P**T**R**MS**P**	339.9	77.4
Rosetta	MRP**Q**T**F**MSA	87.2	-9.4
EcAII: Epitope 2 (216-224) [rank 11]			
Native	IVYNYANAS	217.3	-15.4
Cantor et al.	**V**VY**G**YANAS	195.8	-13.8
Rosetta	IVYNY**S**NA**M**	197.1	-11.2
EcAII: Epitope 3 (304-312) [rank 3]			
Native	VLLQLALTQ	135.3	-17.4
Cantor et al.	VLL**T**LALT**N**	122.4	-11.3
Rosetta	VLLQLAL**W**Q	193.1	-13.4
Epo: Epitope 1 (101-115) [rank 7]			
Native	GLRSLTTLLRALGAQ	7.7	-20.0
Tangri et al.	GLRSLT**D**LLRALGAQ	12.1	-20.3
Rosetta	GLRSLT**D**LLRALGAQ	12.1	-20.3
Epo: Epitope 2 (136-150) [rank 1]			
Native	DTFRKLFRVYSNFLR	5.0	-23.5
Tangri et al.	DTFRKLFRVY**D**NFLR	24.0	-19.9
Rosetta	DTFRK**E**F**FD**Y**A**NFLR	70.2	-13.7

Table 1. Recapitulation of previous L-asparaginase II (EcAII) deimmunization efforts (1) against one HLA allele and erythropoietin (Epo) (2) against eight HLA alleles. MHC IC50's are predicted by Rosetta SVM, Rosetta energies are sum of total residue energies over the epitope region. IC50s for Epo are listed as the lowest predicted across the allele set. Epitope ranks as a function of predicted immunogenicity are listed next to the residue ranges. Mutations are highlighted in bold, known activity-preserving mutations are underlined.

Epitope removal in superfolder GFP. As an experimental proof of concept, we chose the fluorescent reporter protein superfolder GFP (17) (sfGFP). GFP is used to identify and track genetically modified stem cells *in vivo* for numerous applications (18), but concern about the immunogenicity of cells expressing GFP remains (19). We sought to redesign sfGFP, eliminating T cell epitopes without disrupting

fluorescence. To do so, we targeted the top four predicted H-2-IAb epitopes in the sfGFP sequence (Fig. 3A). None of these epitopes were present in our known epitope database, though epitope 84 had been previously identified as immunodominant in wild-type GFP. The design algorithm is nearly deterministic; to generate multiple candidates for testing, we stochastically sampled alternative sequences by random inclusion of locally near-optimal mutations at each design position. Eight designs were chosen for testing based on sequence diversity and predicted stability (Table 2, Text S2). sfGFP and all eight designs expressed in *E. coli* and purified as soluble protein. To determine whether fluorescence was affected by the design mutations, emission and absorbance spectra were obtained for sfGFP and all eight deimmunized proteins. All eight designs showed fluorescence absorbance and emission spectra comparable to sfGFP with fluorescence excitation peaks at 485 nm (Fig. 4A, Fig. S1).

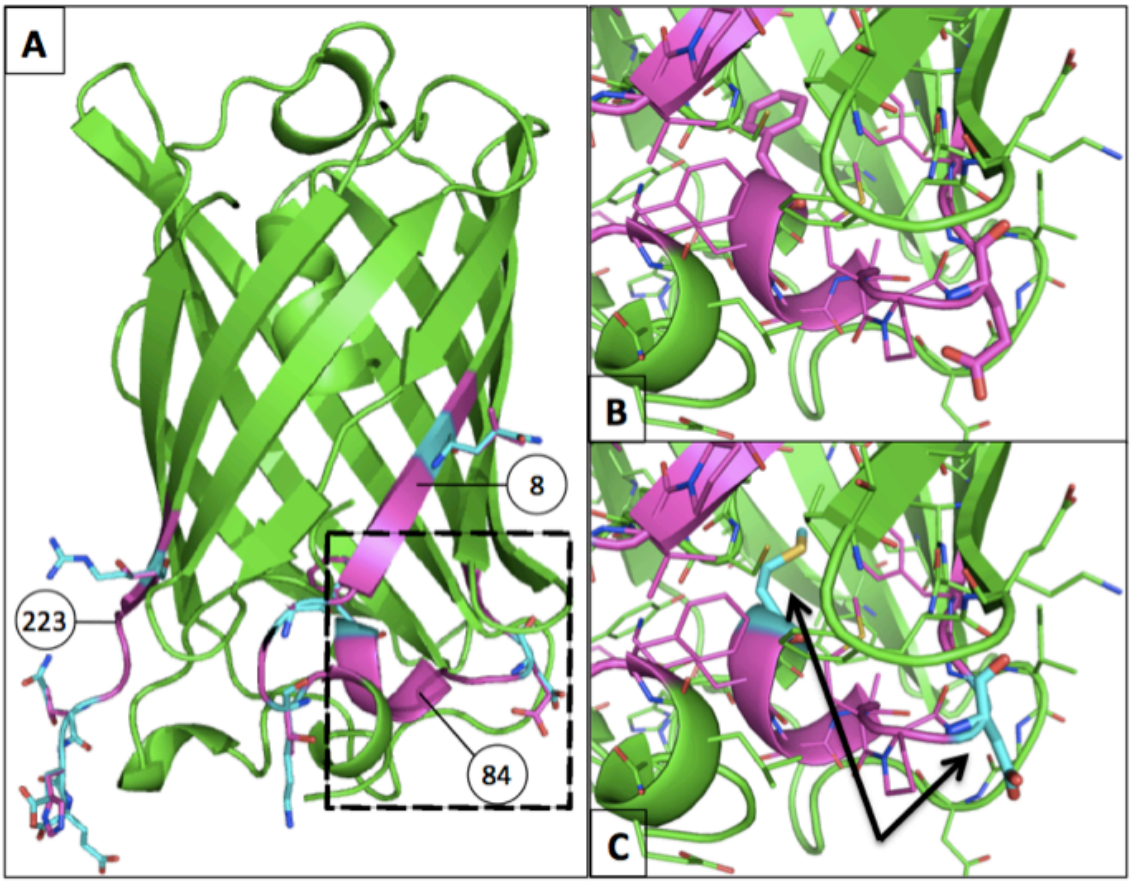


Figure 3. Rosetta design model for sfGFP deimmunization. Green: sfGFP, Magenta: predicted epitopes, Cyan: design mutations. A) published coordinates of superfolder GFP crystal structure. Both known and predicted epitopes were targeted for design. Epitope indices from Table 2 are labeled in circles. B) closeup of immunodominant epitope. C) Rosetta deimmunization design of B).

	sfGFP			sfGFP.di.v3.2		
Index	Native Seq	Predicted IC50 (nM)	Rosetta Energy	Design Seq	Predicted IC50 (nM)	Rosetta Energy
8	FTGVVPILV	548	-12.9	F**K**G**R**VPI**Q**V	998	-10.6
84	FKSAMPEGY	784	-8.4	**M**KSAMP**D**GY	4465	-8.9
223	FVTAAGITH	542	-8.4	FV**R**AAGI**QE**	3312	-7.9
224	VTAAGITHG	954	-7.15	V**R**AAGI**QEE**	2354	-6.5

Table 3. sfGFP epitopes targeted for redesign. Mutations are highlighted in bold.

We then investigated whether existing epitopes were correctly identified and removed and new epitopes not introduced by the design mutations. sfGFP and the deimmunized variant (sfGFP.di.v3.2) were chosen for immunological testing. This variant was selected as the most aggressive design because it had the highest Rosetta energy but still maintained function. GFP:I-Ab tetramer reagents were generated using both native and design peptide sequences for three of the predicted epitope regions. For both constructs, five mice were injected with the protein in complete Freund's adjuvant (CFA). After six days, spleens from all ten mice were stained with a multicombinatorial panel of GFP:I-Ab tetramers corresponding to the both the native and design sequences of all three predicted epitope regions (Table 2) and tetramer-positive cells were magnetically bead enriched. For mice challenged with wildtype sfGFP, flow cytometry confirmed epitope 84 as the immunodominant epitope, with epitope 223/224 recognized by a smaller number of T cells (Fig. 4B). For mice challenged with the deimmunized protein, all three tetramers corresponding to the three redesigned epitope regions failed to isolate T cells above background levels except for one mouse that responded weakly to the mutant epitope 223/224 (Fig. 4C). This confirmed that the design mutations effectively eliminated or greatly reduced the antigenicity of wildtype T cell epitopes without creating new immunodominant epitopes or disrupting fluorescence.

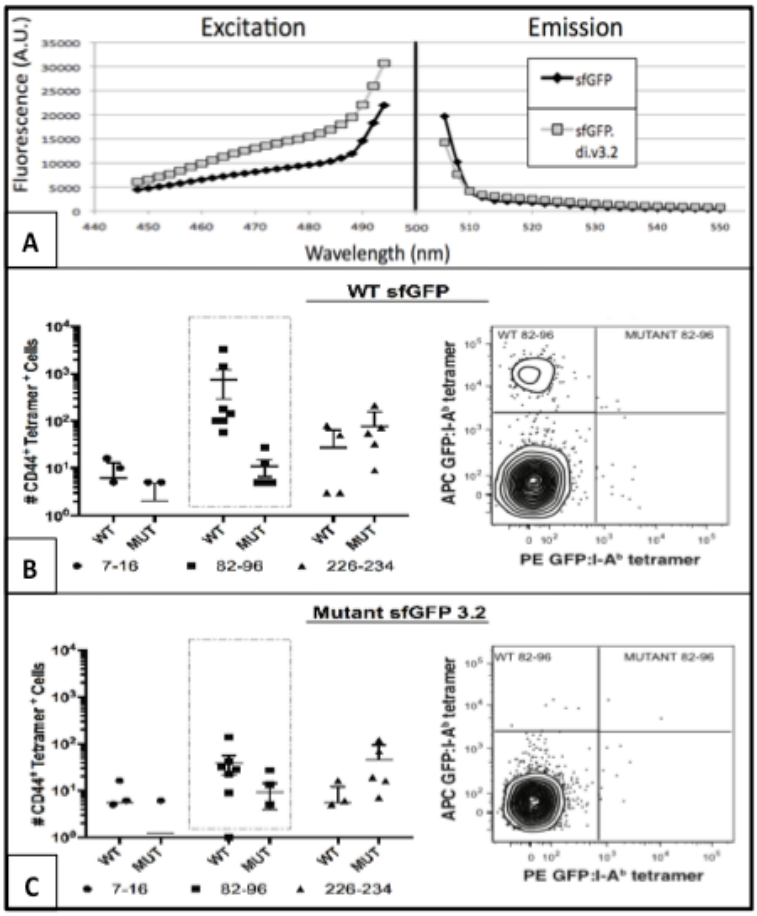


Figure 4. Redesign of sfGFP reduces T cell reactivity without disrupting fluorescence. A) Deimmunized sfGFP excitation and emission spectra. Excitation spectrum measured at 510nm emission. Emission spectra measured at 488 nm excitation. B) Flow cytometry analysis of tetramer-enriched populations of CD3+ CD4+ CD44+ GFP:I-Ab+ lymphocytes. Total CD44+ CD4+ tetramer positive cells for each of the six GFP:I-Ab tetramers in mice immunized with WT sfGFP. Immunization with the native sfGFP leads to the expansion and activation of CD4+ T cells responding to epitope 82-96. C) Total CD44+ CD4+ tetramer positive cells for each of the six GFP:I-Ab tetramers in mice immunized with the designed sfGFP 3.2. Mice immunized with the designed sfGFP 3.2 no longer respond to the native sfGFP 82-96 epitope or the novel designed epitope 82-96 in sfGFP 3.2.

Epitope removal in domain III of Pseudomonas exotoxin A. We next sought to remove T cell epitopes from the toxin domain of the cancer therapeutic HA22, a recombinant immunotoxin containing a 38 kDa fragment of Pseudomonas exotoxin A (PE38) (20), while maintaining cytotoxic activity. HA22 has been used successfully to treat refractory hairy cell leukemia in a recent phase 1 clinical trial (21), and has produced complete remission in several patients with acute lymphoblastic leukemia (22). However, HA22 has shown limited effectiveness in treating patients who are not immune compromised, as the presence of the bacterial PE38 moiety leads to host immune response and the production of neutralizing anti-drug antibodies (23). To address this issue, we targeted three epitope regions in PE38 previously identified in humans (24) and designed five mutations predicted to eliminate binding to a diverse set of 14 human HLA alleles while maintaining

favorable interactions with the toxin substrate, eukaryotic elongation factor 2 (Fig. S4). The five mutants were expressed and purified for subsequent testing of cytotoxicity in two Burkitt's lymphoma cell lines. Two mutants in the region 466-480, A476D and A476D+D474Y, had cytotoxic activity reduced by approximately 80%, but three mutants in region 547-564 displayed equal or greater cytotoxicity than the wild-type toxin (Fig. 5A). The two most active mutants (L552N and L552E) were chosen for further characterization of immunogenicity. Peripheral blood mononuclear cells (PBMCs) derived from two patients and one naïve donor were stimulated with mutant antigen, and IL-2 response was measured after restimulation with the wild-type and mutant epitope peptides. Both mutants caused a significant reduction in T cell response for both epitope peptides in all three samples ($p > 0.01$ in student T test) (Fig. 5B).

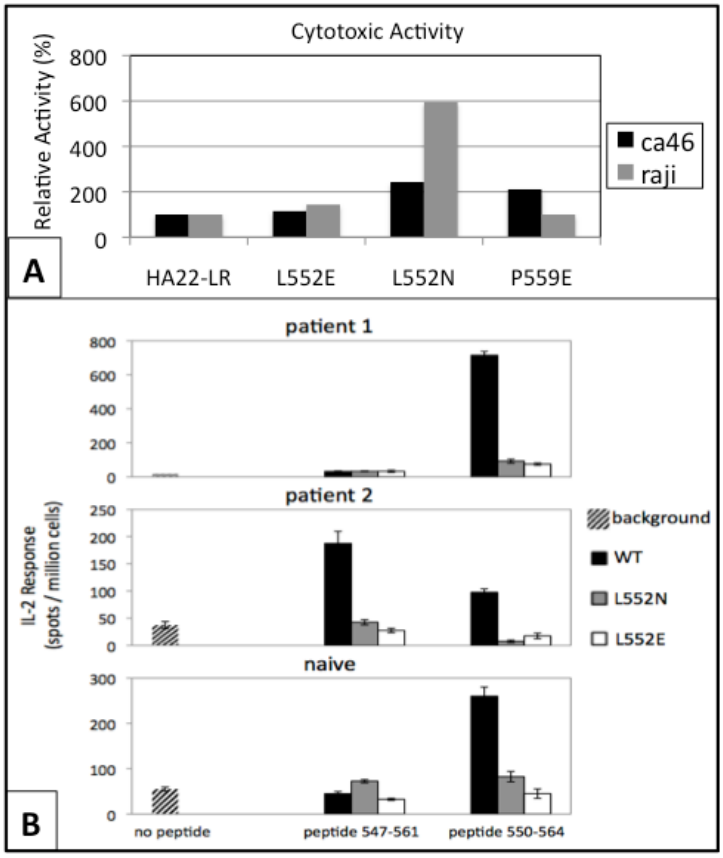


Figure 5. Redesign of exotoxin A reduces T cell reactivity without loss of function. A) Relative cytotoxicity for the original HA22-LR toxin and three computationally designed variants in two cell types. B) ELISpot IL-2 response was measured for PBMCs derived from two patients and one naïve donor after restimulation with two WT peptides and four mutant peptides.

Discussion

We have developed a computational protein design method that incorporates host genome information and MHC-binding prediction tools to reduce the immunogenicity of arbitrary protein targets. The method removes MHC epitopes and increases human sequence content while maintaining protein stability and

interactions with binding partners. Mutations predicted by the method partially recapitulate the mutations of previous successful deimmunization efforts. The effectiveness of the method was verified experimentally by successfully predicting and eliminating an immunodominant T cell epitope from sfGFP and eliminating a known T cell epitope from Pseudomonas exotoxin A. We redesigned these proteins to mitigate T cell immune responses in both mice and humans, respectively, demonstrating that the method presented is generally applicable to humans or other species for which sufficient MHC data exist.

Due to the possibility of disrupting folding and function of the target protein, most deimmunization efforts rely on experimental testing of a limited number of point mutants, followed by conservative attempts to combine a small subset of these mutations into a functional product. Though varied in their approaches, proprietary methods used by companies such as EpiVax (25) and Antitope (26) typically combine matrix-based epitope scoring, experimental characterization of MHC binding or T cell epitope mapping, step-wise mutation of antigenic amino acids, and introduction of tolerizing epitopes where possible. Using structure-based design simulations, we introduced 9 mutations into sfGFP simultaneously without disrupting function. One redesigned epitope lies largely buried in the core of the protein (Fig. 3B,C), requiring the selection of mutations that simultaneously eliminate MHC-binding propensity without disrupting the folding free energy of the protein. Eliminating such epitopes would be difficult using the above proprietary approaches, as multiple substitutions may be required to compensate for packing defects caused by mutations in the epitope region, and nonconservative mutations near the protein binding interface of exotoxin A would not have been predicted without a structure-based design calculation.

This work has focused only on eliminating the most immunoreactive epitopes for a given set of MHC alleles. Immunological testing of a larger number of human patients would be required to fully cover the breadth of HLA allotype diversity, and there may exist highly conserved T cell reactive amino acids in protein sequences for which no immune silencing mutations are possible, necessitating the prediction and removal of discontinuous B cell epitopes from the protein surface. Methods exist for prediction of B cell epitopes, but suffer from lower predictive power due to the difficulty of obtaining sufficient structural data for the entire repertoire of discontinuous three-dimensional epitopes (27). Nevertheless, B cell epitope removal methods have proven successful for a number of clinical targets (28, 29). Such methods could be incorporated into a comprehensive deimmunization pipeline, and further testing of systemic antibody response would be required to demonstrate complete immune evasion.

Here we have focused on application to protein therapeutics delivered extracellularly, where MHCII mediates the primary immunological pathway. The method is readily extensible to both MHCI and MHCII-based deimmunization, and is thus applicable to therapeutics expressed intracellularly, such as gene therapy products. Because epitope scores are averaged over all MHC allele SVMs, degenerate binding epitopes are penalized more strongly, and thus are the first to be targeted in our greedy design algorithm. This is critical, as mounting evidence points toward the correlation between the number of host MHC alleles a given epitope binds and

its propensity to initiate an immune reaction *in vivo* (2, 30). Our epitope prediction method was trained on large-scale peptide binding affinity data, though predictions can be made using other sources. When experimental binding constants are not available, MHC-peptide structure simulations have shown promise in calculating accurate sequence specificities based on MHC-peptide energetics (31). As improvements in energy functions lead to improvement in the prediction of the effects of mutations on stability and function, and high-throughput experimental MHC-peptide binding data become increasingly available, computational protein design will play an increasingly prominent role in development of next-generation protein therapeutics.

Methods

Penalizing Predicted Epitopes

Epitope SVM Construction: A detailed description of SVM construction is provided in the S.I. Briefly**,** one SVM model was trained for each MHC using publicly available peptide-MHC binding constants for 29 human and mouse MHC alleles. 15mer peptide sequences were first aligned, then encoded into numerical feature vectors as described in the S.I. Once encoded, SVM regression models were trained using libSVM to recapitulate peptide-MHC binding IC50 values by minimizing mean-squared error in five-fold cross-validation tests.

Epitope SVM Scoring. The total epitope prediction score of a protein during Rosetta design is calculated by sliding a 15 residue window across the protein sequence, calculating the average SVM binding score over all allele SVM models in the user-defined allele list, and summing the contributions from each overlapping sequence frame.

Rewarding 9mer sequences that occur in the host genome and penalizing known epitopes.

Host 9mer Database Construction. Protein translations of Ensembl gene predictions for *Homo sapiens* and *Mus Musculus* genomes were downloaded from ftp://ftp.ensembl.org/pub/current_fasta/ on 28 Feb. 2012. The number of occurrences of every unique contiguous 9mer peptide found in all hypothetical translation products was first calculated. Each 9mer was assigned a score that rewards common sequences. 9mers that occur between one and ten times were assigned a score of –log(n)-1, where n is the total count of the 9mer. 9mers that occur more than ten times were given a constant score of -2.0 to prevent domination of scoring by widespread repeat sequences. Thus, each unique sequence was given a score in the range [-2,-1].

Known Epitope Database Construction. All epitope sequences known to elicit T cell activation though the MHCII pathway in either humans or mice were downloaded from the IEDB on 28 Feb. 2012. 9mer core sequences were predicted as the epitope subsequence with the highest predicted MHC binding affinity as scored by Rosetta SVMs. All 9mer epitope sequences were assigned a constant score of 10.0.

9mer Database Scoring. 9mer subsequences with associated scores (genomic 9mers and known epitope sequences) are loaded from a user-supplied table at runtime and stored in a hash table for quick lookup during design. The total score is the sum of the scores for each 9mer subsequence.

Rosetta Design Calculations

Structure Preprocessing and Simulation Parameters. Before design calculations, all protein structures were subject to multiple cycles of backbone minimization and rotamer optimization with position restraints on sidechain heavy atoms, allowing for small structural changes to bring the structure to the local energy function minimum. For all design calculations, “talaris2013” Rosetta score weights were used; native residues were given a constant energy bonus of -0.2 REU, non-native residues a penalty of 0.2 REU, and non-native cysteine and histidine residues were disallowed in design.

Rosetta Greedy Optimization Design. All design simulations were implemented using Rosetta Scripts (32). Greedy sequence design and rotamer optimization was carried out using the Rosetta

greedy descent optimization algorithm as previously described (11). Details are provided in the S.I. Briefly, every amino acid point mutant is sampled independently and the total energy is stored. Optimal mutations are sorted by energy before combinatorial optimization is attempted in a rank-ordered, steepest descent fashion. Multiple diverse solutions can be generated by stochastically attempting combination of near-optimal substitutions at each position.

Deimmunization Design Simulations. Details of each design calculation are provided in the S.I. Briefly, all simulations followed a similar workflow as follows: Crystal structures were downloaded from the RCSB Protein Data Bank and pre-minimized with Rosetta as described above. Designable residues were selected on the basis of average predicted MHC binding affinity; all residues in any epitope frame that score above a certain threshold are selected for design. For comparison with previous experimental works, residues were selected so as to provide a meaningful comparison to the published data. Deimmunization design was then carried out as described above. For design targets undergoing experimental characterization, multiple design sequences were generated by randomly sampling near-optimal mutations at each design position and combining these mutations in a stochastic manner.

sfGFP Activity and Immunogenicity Assays

Identification of eGFP$_{82\text{-}96}$:I-A^b epitope. To begin to narrow down a region in the eGFP protein containing a CD4$^+$ T cell epitope in C57BL/6 mice, the eGFP nucleotide sequence (Clontech) was parsed into 5 equally sized fragments and each was inserted into the TOPO cloning site of pTrcHis2-TOPO vector, containing an IPTG-inducible promoter and a 6X His c-terminal epitope tag (Invitrogen). eGFP protein fragments were expressed in Rosetta 2 competent *E. coli* (EMD Millipore). Bacteria were lysed with BugBuster protein extraction reagent (Novagen), sonicated, and eGFP fragments were purified with His-Bind resin columns (Novagen). Next, individual mice were immunized subcutaneously with 25ug of purified whole eGFP protein (Biovision, Inc.) in Complete Feund's Adjuvant (CFA) [Sigma]. After 10 days, draining lymph node cells were negatively selected for CD4$^+$ T cells (Miltenyi Biotech) and an anti-interferon gamma ELISPOT assay was done by interrogating with the above purified eGFP protein fragments (antibodies from eBioscience, 96 well Multiscreen filter plates from Millipore). Upon identifying the eGFP protein fragment that gave the best interferon gamma-producing CD4$^+$ T cell response, a 15-mer overlapping peptide library (each offset by two amino acids) was constructed (Mimotopes). This library was then used for interrogation in another ELISPOT assay, as described above, to narrow down the 15-mer epitope to amino acids 82-96.

Immunizations. C57BL/6 mice were injected subcutaneously at the base of the tail with either 50ug sfGFP or 50ug sfGFP.di.v3.2 emulsified in 50ul complete Freund's adjuvant (Sigma-Aldrich). C57BL/6 mice were injected i.p. with 100 ug sfGFP or 100 ug sfGFP.di.v3.2 in aluminum hydroxide adjuvant (Brenntag).

Tetramer Production. Biotin-labeled soluble I-A^b molecules containing eGFP peptide (FKSAMPEGY) covalently attached to the I-A^b beta chain were produced in Drosophila melanogaster S2 cells, then purified and made into tetramers with streptavidin-phycoerythrin or streptavidin-allophycocyanin (Prozyme) as described (Moon et al., 2007).

Cell Enrichment and Flow Cytometry. All antibodies were from eBioscience unless noted. Single cell suspensions of spleens and lymph nodes were stained for 1 hour at room temperature with eGFP:I-A^b allophycocyanin tetramer. Samples were then enriched for bead-bound cells on magnetized columns and a portion was removed for counting as described (Moon et al., 2007). For identification of surface phenotype, the rest of the sample underwent surface staining on ice with a mixture of antibodies specific for B220 (RA3-6B2), CD11b (MI-70), CD11c (N418), CD44 (IM7; BD), CD4 (RM4-5; BD), CD3 (145-2C11), and CD8 (5H10; BioLegend) each conjugated with a different fluorochrome. Cells were then analyzed on a Canto (BD) flow cytometer. Data were analyzed with FlowJo software (TreeStar).

Fluorescence Spectra. Superfolder GFP samples were diluted to a uniform concentration of 9.8 uM in PBS and fluorescence spectra were measured on a SpectraMax plate reader. Excitation was measured from 448 nm to 500nm (2nm intervals) at 510 nm emission. Emission was measured from 498 nm to 550 nm (2nm intervals) at 488 nm excitation. Fluorescence spectra were normalized by subtracting the signal obtained from pure PBS buffer.

Exotoxin A Activity and Immunogenicity Assays

Construction, Expression, Purification and cytotoxic activity of recombinant immunotoxin. The mutations L552E, L552N and P559E were introduced into a plasmid expressing HA22 VH-PE38 using PCR overlap extension. The mutant RIT were purified as previously described (33). Cytotoxicity assays were performed on CD22+ human Burkitt lymphoma cell lines (CA46 and Raji). The assay was performed as previously described (24).

In Vitro Expansion and ELISpot. PBMC from two patients that were previously treated with PE38 RIT and one naïve donor were obtained after informed consent. The PBMC were stimulated with parent RIT, HA22-L552E or HA22-L552N and cultured in 37°c with 5% CO_2 for 14 days. IL2 (10U/ml) (Millipore,MA) was added every three days. On day 14, cells were harvested restimulated with either WT peptides 93 and 94 (GPEEEGGRLETILGW and EEGGRLETILGWPLA) or mutant peptides (GPEEEGGR**E**ETILGW, EEGGR**E**ETILGWPLA, GPEEEGGR**N**ETILGW and EEGGR**N**ETILGWPLA). IL2 ELISpot was used to detect T cell activation according to manufacturer instructions.

Rosetta Command-line Demo

Command line examples, input files, and instructions for running all protein design simulations are included in the archived demo `removing_tcell_epitopes_protein_design_demo.tar.gz` . The demo can be downloaded from https://zenodo.org/record/8436 .

Acknowledgement

This research was supported by the Defense Threat Reduction Agency and the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

Supporting Information

Text S1. Supporting Information Methods

Epitope SVM Construction Details. One SVM model was trained for each MHC using publicly available peptide-MHC binding constants for 29 human and mouse MHC alleles. 15mer peptide sequences were first aligned, then encoded into numerical feature vectors. Amino acid sequence encoding relied on two sources: BLOSUM62 amino acid transition probabilities and PSSMs values derived from the initial sequence alignment. Once encoded, nu-SVR regression models were trained using libSVM to recapitulate log-transformed IC50 values by minimizing mean-squared error in five-fold cross-validation tests.

Alignment of SVM Training Data Sequences. MHC peptide binding training data for 29 MHC alleles were downloaded from the IEDB, non-15mers were discarded, and sequences were split into binder and non-binder groups for each allele with a 1000 nM cutoff. Peptides in each group were aligned using NetMHCII v2.2 to find the highest scoring peptide core frame in each 15mer. Aligned 15mers were then extracted by including the 9mer core and 3 residues upstream and downstream, substituting "X" for gapped termini positions.

Calculation of MHC allele specific PSSMs. For each allele, a position-frequency matrix was calculated by counting residue frequencies from all aligned 9mer cores from peptides with IC50 < 1000 nM. Values less than 0.001 in the matrix were given a pseudo-value of 0.001 and columns renormalized. PSSM values were calculated as the log-odds ratio of each amino acid position frequency to the baseline frequency of that amino acid in the host organism.

Peptide Sequence SVM Encoding. Each 15mer sequence was encoded as a 240-element feature vector for SVM training as follows: Each amino acid of the 9mer core sequence is represented by a 21-element vector from the corresponding row of the probability-transformed BLOSUM62 matrix. Additionally, the upstream and downstream 3mers, the peptide-flanking residues (PFRs), were similarly encoded as a weighted average over the 3 positions with N-terminal weights of (1/6, 2/6, 3/6) and C-terminal weights of (3/6, 2/6, 1/6). Another 9 features are added by including the score for each 9mer core position from the PSSM described above.

SVM Training. SVM models were created using libSVM. A regression model was trained (nu-SVR) to recapitulate log-transformed IC50 values from the feature vectors described above. IC50 values were

transformed into scores according the same manner as Nielsen et al. (14), where each score S = 1-log(IC50)/log(50,000), such that the strongest binder receives a score of 1.0 and the weakest a score of 0.0. libSVM models were generated using the RBF kernel with the shrinking heuristic, and c, f parameters were chosen for each MHC allele model to minimize the average mean-squared error in five-fold cross-validation tests.

Deimmunization Design Simulation Details

Rosetta Greedy Optimization Design. All design simulations were implemented using Rosetta Scripts (32). Greedy sequence design and rotamer optimization was carried out using the Rosetta greedy descent optimization algorithm as previously described (11). First, every amino acid point mutant and rotamer state at every position is sampled independently, and, after rotamer optimization and gradient minimization of all neighbor sidechains within an 8 Å sphere, the total energy is stored. After every position's point mutants have been evaluated, substitutions at each position are sorted by energy, and positions are rank ordered by the value of the optimal substitution at each position. Substitutions are combined by first attempting placement of the optimal substitution at the optimal position, evaluating the total energy, and accepting if the total score improves. The substitution at the second ranked position is then attempted, and so on, until substitutions have been attempted. This approach converges reliably to identical solutions, though multiple diverse solutions can be generated by optionally attempting combination of near-optimal substitutions at each position, only considering substitutions whose scores remain within a certain threshold from the position's optimal substitution.

Large-scale Design Benchmarking. The crystal structures of eight proteins isolated from human pathogens, all containing known T cell epitopes, were downloaded from the RCSB Protein Data Bank. Target protein sequences were scanned for MHC-binding sequences using Rosetta as described above. Each sequence was scanned for eight HLA-DR alleles (HLA-DRB1*01:01, HLA-DRB1*03:01, HLA-DRB1*04:01, HLA-DRB1*07:01, HLA-DRB1*08:02, HLA-DRB1*11:01, HLA-DRB1*13:02, HLA-DRB1*15:01). Epitopes were identified as those with a log-averaged predicted IC50 less than or equal to 1500 nM. The predicted 15mer cores of all these epitopes were targeted for design. For the SVM score term, design simulations were carried out with varying score weights using the greedy optimization scheme described above. For host genome 9mer score term, three design simulations were carried out using Monte Carlo with 150 steps per designable position. Human 9mer content as a fraction of total epitopes subject to design and Rosetta Energy values were averaged from these simulations. Final predicted epitope count was calculated as those with a log-averaged predicted IC50 less than or equal to 500 nM.

L-asparaginase II Design Simulations. The crystal structure of L-asparaginase II (PDBID: 1NNS) was downloaded from the RCSB Protein Data Bank as the homo-tetrameric biological assembly. Design positions were restricted to those 9mer epitope regions identified by Cantor et al. The design simulation was carried out with greedy sequence optimization as mentioned above, with a subsequence SVM score weight of 1.0. Epitopes were designed using the SVM for HLA-DRB1*04:01. (L-asparaginase simulations only used one allele predictor to better match the experiment of Cantor et al. This single allele has lower average binding affinity and thus higher average scores than the eight-allele set used in the Epo simulations, so its weight was decreased to compensate.) Predicted IC50's are reported as the strongest binding epitope that overlaps the design target 9mer. Rank is reported as the highest-ranking epitope frame that encompasses all design positions. Rosetta energies for epitope regions are calculated by summing intra- and inter-residue energies over the target segment.

Erythropoietin Design Simulations. The crystal structure of erythropoietin complexed to the binding domain of the erythropoietin receptor (PDBID: 1EER) was downloaded from the RCSB Protein Data Bank. Design positions were restricted to those chosen by Tangri et al. for mutation (residues 102, 103, 104, 107, 141, 143, 144, 146, 147), using a subsequence SVM score weight of 3.5. Epitopes were redesigned using SVMs corresponding to eight HLA-DR alleles (HLA-DRB1*01:01, HLA-DRB1*03:01, HLA-DRB1*04:01, HLA-DRB1*07:01, HLA-DRB1*08:02, HLA-DRB1*11:01, HLA-DRB1*13:02, HLA-DRB1*15:01). Predicted IC50's are reported as the strongest binding epitope that overlaps the design target 9mer. Rank is reported as the highest-ranking epitope frame that encompasses all design positions. Predicted allele binders were calculated using an IC50 cutoff of 500 nM.

sfGFP Design Simulations. sfGFP design simulations utilized the available sfGFP crystal structure (PDBID: 2B3P). Eight designs were calculated, utilizing Rosetta both for predicting epitopes and predicting the mutations' effects on MHC binding. To generate multiple design sequences, candidate mutations at each design position included all amino acids within 1.5 REU of the lowest-energy mutation. Candidate mutations at each position were chosen randomly during the Greedy Optimization Design stage as described above.

Exotoxin A Design Simulations. Exotoxin A design simulations utilized the available exotoxin-eEF2 co-crystal structure (PDBID: 1ZM4). Epitopes were redesigned using SVMs corresponding to fourteen HLA-DR alleles (HLA-DRB1*01:01, HLA-DRB1*03:01, HLA-DRB1*04:01, HLA-DRB1*07:01, HLA-DRB1*08:02, HLA-DRB1*09:01, HLA-DRB1*11:01, HLA-DRB1*13:02, HLA-DRB1*15:01, HLA-DRB3*01:01, HLA-DRB4*01:01, HLA-DRB5*01:01, HLA-DQA1*05:01-DQB1*03:01, HLA-DQA1*03:01-DQB1*03:02). Three designs were calculated, utilizing Rosetta for predicting the mutations' effects on MHC binding of the known residue 466-480 and 547-564 epitope region. Candidate mutations at each position were chosen randomly during the Greedy Optimization Design stage as described above.

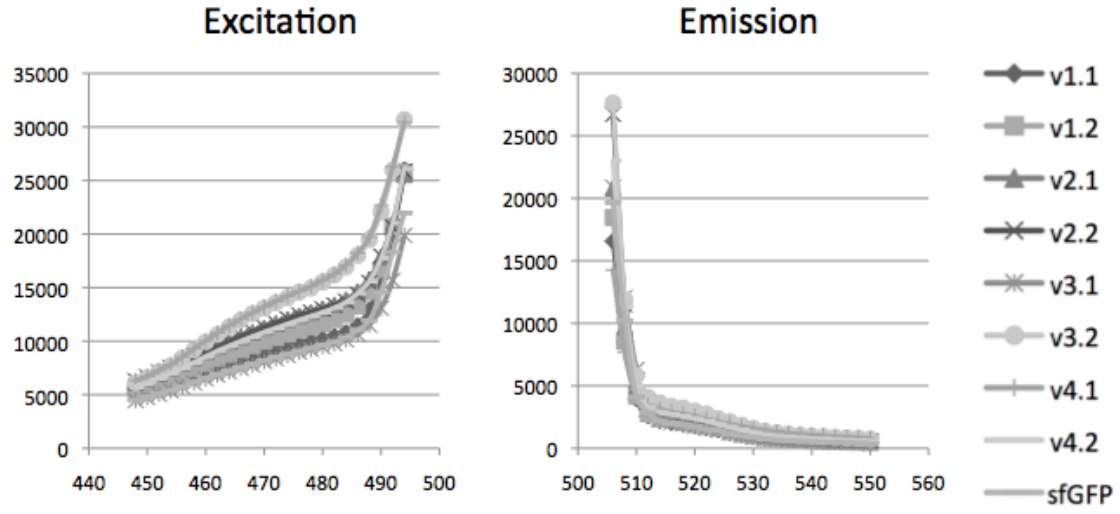


Figure S1. Native and deimmunized sfGFP excitation and emission spectra for all eight sfGFP designs. A) Excitation spectrum measured at 510nm emission. B) Emission spectra measured at 488 nm excitation.

Source	PDBID	Known Epitopes (Design/Native)	Predicted Epitopes (Design/Native)	Human 9mers	Delta R.E.	Sequence Identity
Tuberculosis esxB	3fav	2/13	0/4	0/0	-7.962	0.811
Arenavirus L-protein	3jsb	2/21	9/16	0/0	-23.113	0.827
Influenza Matrix M1 protein	1aa7	0/13	5/19	0/0	-25.347	0.857
HSV envelope protein D	1a8z	0/7	5/29	0/0	-29.163	0.853
Tuberculosis MPT63	1lmi	2/17	1/10	0/0	-32.871	0.893
SARS Nucleocapsid	2cjr	0/3	0/10	0/0	-8.404	0.861
SARS ORF9-B	2cme	0/3	5/14	0/0	-22.94	0.692
Malaria AMA1	2q8a	1/1	10/22	0/0	-32.925	0.918
Tuberculosis						

Table S1. Rosetta deimmunization of crystal structures with known MHC epitopes. The number of known and predicted epitopes (from the 8 HLA-DR allele set described in the main text) after design (Design) and in the original native sequence (Native) is shown for each target protein, along with change in Rosetta energy (Delta R.E.) and the fraction of residues not changed during design (Sequence Identity). Known epitope data was excluded from the simulation so as to provide an unbiased measure of the prediction and design

Text S2. sfGFP deimmunization design sequences alignment.

```
sfgfp             SKGEELFTGVVPILVELDGDVNGHKFSVRGEGEGDATNGKLTLKFICTTGKLPVPWPTLV
sfgfp.di.v2.2     SKGEELFKGRVPIQVELDGDVNGHKFSVRGEGEGDATNGKLTLKFICTTGKLPVPWPTLV
sfgfp.di.v1.2     SKGEELFLGRVPIQVELDGDVNGHKFSVRGEGEGDATNGKLTLKFICTTGKLPVPWPTLV
sfgfp.di.v4.2     SKGEELFLGRVPIQVELDGDVNGHKFSVRGEGEGDATNGKLTLKFICTTGKLPVPWPTLV
sfgfp.di.v4.1     SKGEELFLGRVPIQVELDGDVNGHKFSVRGEGEGDATNGKLTLKFICTTGKLPVPWPTLV
sfgfp.di.v1.1     SKGEELFLGRVPIQVELDGDVNGHKFSVRGEGEGDATNGKLTLKFICTTGKLPVPWPTLV
sfgfp.di.v3.1     SKGEELFLGRVPIQVELDGDVNGHKFSVRGEGEGDATNGKLTLKFICTTGKLPVPWPTLV
sfgfp.di.v3.2     SKGEELFKGRVPIQVELDGDVNGHKFSVRGEGEGDATNGKLTLKFICTTGKLPVPWPTLV
sfgfp.di.v2.1     SKGEELFTGVVQILVELDGDVNGHKFSVRGEGEGDATNGKLTLKFICTTGKLPVPWPTLV
                  ******* * * * ******************************************

sfgfp             TTLGYGVQCFSRYPDHMKRHDFFKSAMPEGYVQERTISFKDDGTYKTRAEVKFEGDTLVN
sfgfp.di.v2.2     TTLGYGVQCFSRYPDHMKRHDFFKSSMPDGYVQERTISFKDDGTYKTRAEVKFEGDTLVN
sfgfp.di.v1.2     TTLGYGVQCFSRYPDHMKRHDFFKSAQPDGYVQERTISFKDDGTYKTRAEVKFEGDTLVN
sfgfp.di.v4.2     TTLGYGVQCFSRYPDHMKRHDFFKSAMSDGYVQERTISFKDDGTYKTRAEVKFEGDTLVN
sfgfp.di.v4.1     TTLGYGVQCFSRYPDHMKRHDFFKSAMSDGYVQERTISFKDDGTYKTRAEVKFEGDTLVN
sfgfp.di.v1.1     TTLGYGVQCFSRYPDHMKRHDFFKSAQPDGYVQERTISFKDDGTYKTRAEVKFEGDTLVN
sfgfp.di.v3.1     TTLGYGVQCFSRYPDHMKRHDFMKSAMPDGYVQERTISFKDDGTYKTRAEVKFEGDTLVN
sfgfp.di.v3.2     TTLGYGVQCFSRYPDHMKRHDFMKSAMPDGYVQERTISFKDDGTYKTRAEVKFEGDTLVN
sfgfp.di.v2.1     TTLGYGVQCFSRYPDHMKRHDFFKSSMPDGYVQERTISFKDDGTYKTRAEVKFEGDTLVN
                  **********************:**:  :******************************

sfgfp             RIELKGIDFKEDGNILGHKLEYNFNSHNVYITADKQKNGIKANFKIRHNVEDGSVQLADH
sfgfp.di.v2.2     RIELKGIDFKEDGNILGHKLEYNFNSHNVYITADKQKNGIKANFKIRHNVEDGSVQLADH
sfgfp.di.v1.2     RIELKGIDFKEDGNILGHKLEYNFNSHNVYITADKQKNGIKANFKIRHNVEDGSVQLADH
sfgfp.di.v4.2     RIELKGIDFKEDGNILGHKLEYNFNSHNVYITADKQKNGIKANFKIRHNVEDGSVQLADH
sfgfp.di.v4.1     RIELKGIDFKEDGNILGHKLEYNFNSHNVYITADKQKNGIKANFKIRHNVEDGSVQLADH
sfgfp.di.v1.1     RIELKGIDFKEDGNILGHKLEYNFNSHNVYITADKQKNGIKANFKIRHNVEDGSVQLADH
sfgfp.di.v3.1     RIELKGIDFKEDGNILGHKLEYNFNSHNVYITADKQKNGIKANFKIRHNVEDGSVQLADH
sfgfp.di.v3.2     RIELKGIDFKEDGNILGHKLEYNFNSHNVYITADKQKNGIKANFKIRHNVEDGSVQLADH
sfgfp.di.v2.1     RIELKGIDFKEDGNILGHKLEYNFNSHNVYITADKQKNGIKANFKIRHNVEDGSVQLADH
                  ************************************************************

sfgfp             YQQNTPIGDGPVLLPDNHYLSTQSVLSKDPNEKRDHMVLLEFVTAAGITHG
sfgfp.di.v2.2     YQQNTPIGDGPVLLPDNHYLSTQSVLSKDPNEKRDHMVLLEFVTAAGIDDG
sfgfp.di.v1.2     YQQNTPIGDGPVLLPDNHYLSTQSVLSKDPNEKRDHMVLLEFVRAAGITDQ
sfgfp.di.v4.2     YQQNTPIGDGPVLLPDNHYLSTQSVLSKDPNEKRDHMVLLEFVRAAGITDG
sfgfp.di.v4.1     YQQNTPIGDGPVLLPDNHYLSTQSVLSKDPNEKRDHMVLLEFVEAAGIQEQ
sfgfp.di.v1.1     YQQNTPIGDGPVLLPDNHYLSTQSVLSKDPNEKRDHMVLLEFVTAGGIQEE
sfgfp.di.v3.1     YQQNTPIGDGPVLLPDNHYLSTQSVLSKDPNEKRDHMVLLEFVRAAGIQEE
sfgfp.di.v3.2     YQQNTPIGDGPVLLPDNHYLSTQSVLSKDPNEKRDHMVLLEFVRAAGIQEE
sfgfp.di.v2.1     YQQNTPIGDGPVLLPDNHYLSTQSVLSKDPNEKRDHMVLLEFVRAAGIDEG
```

********** *.** .

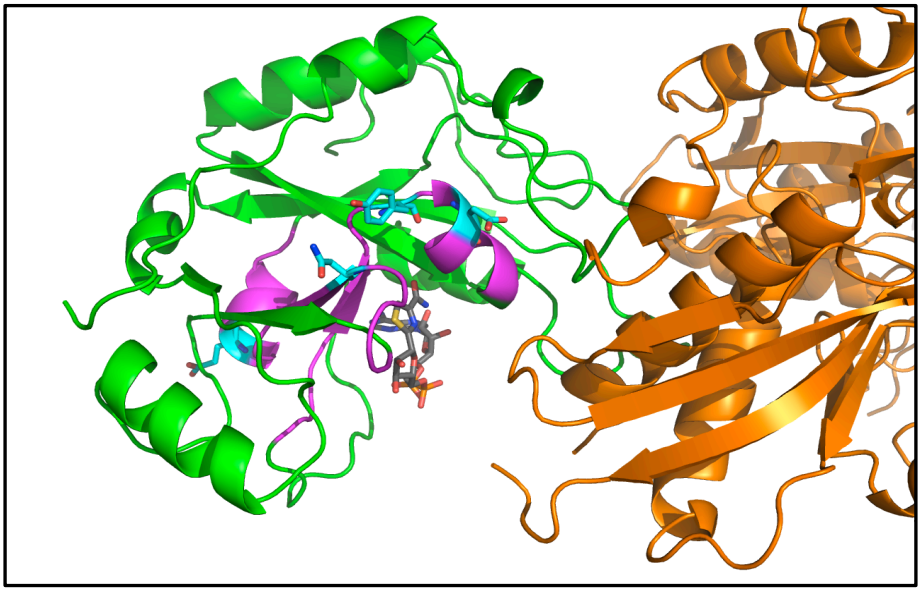


Figure S2. Rosetta design model for immunotoxin deimmunization. Green: Endotoxin A, Magenta: T cell epitopes, Cyan: design mutations, Orange: eEF-2

Bibliography

Chapter 1:

1. G. Walsh. Biopharmaceutical benchmarks 2010. Nat. Biotechnol., 28 (2010), pp. 917–924

2. N.C. Nicolaides, P.M. Sass, L. Grasso. Advances in targeted therapeutic agents. Expert Opin. Drug Discov., 5 (2010), pp. 1123–1140
3. B. Leader, Q.J. Baca, D.E. Golan. Protein therapeutics: a summary and pharmacological classification. Nat. Rev. Drug Discov., 7 (2008), pp. 21–39
4. J.M. Reichert. Metrics for antibody therapeutics development. mAbs, 2 (2010), pp. 695–700
5. J. Kling. Fresh from the biologic pipeline — 2010. Nat Biotechnol, 29 (2011), pp. 197–200
6. B. Leader, Q.J. Baca, D.E. Golan. Protein therapeutics: a summary and pharmacological classification. Nat Rev Drug Discov, 7 (2008), pp. 21–39
7. R.N. Brogden, R.C. Heel. Human insulin. A review of its biological activity, pharmacokinetics and therapeutic use. Drugs, 34 (1987), pp. 350–371
8. M.J. Henwood, A. Grimberg, T. Moshang Jr. Expanded spectrum of recombinant human growth hormone therapy. Curr Opin Pediatr, 14 (2002), pp. 437–442
9. Koga, Nobuyasu, et al. "Principles for designing ideal protein structures."*Nature* 491.7423 (2012): 222-227.
10. Fleishman, Sarel J., et al. "Computational design of proteins targeting the conserved stem region of influenza hemagglutinin." *Science* 332.6031 (2011): 816-821.
11. Tinberg, Christine E., et al. "Computational design of ligand-binding proteins with high affinity and selectivity." *Nature* 501.7466 (2013): 212-216.
12. Siegel, Justin B., et al. "Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction." *Science* 329.5989 (2010): 309-313.
13. Neefjes, Jacques, et al. "Towards a systems understanding of MHC class I and MHC class II antigen presentation." *Nature Reviews Immunology* 11.12 (2011): 823-836.
14. Casadevall, Nicole, et al. "Pure red-cell aplasia and antierythropoietin antibodies in patients treated with recombinant erythropoietin." *New England Journal of Medicine* 346.7 (2002): 469-475.
15. Warmerdam, Petra AM, et al. "Elimination of a human T-cell region in staphylokinase by T-cell screening and computer modeling." *THROMBOSIS AND HAEMOSTASIS-STUTTGART*- 87.4 (2002): 666-673.
16. Hellendoorn, K., et al. "Limiting the risk of immunogenicity by identification and removal of T-cell epitopes (DeImmunisation™)." *Cancer Cell International*4.Suppl 1 (2004): S20.
17. Tangri, Shabnam, et al. "Rationally engineered therapeutic proteins with reduced immunogenicity." *The Journal of Immunology* 174.6 (2005): 3187-3196.

18. Yeung, V. Peter, et al. "Elimination of an immunodominant CD4+ T cell epitope in human IFN-β does not result in an in vivo response directed at the subdominant epitope." *The Journal of Immunology* 172.11 (2004): 6658-6665.
19. Goeddel, David V., et al. "Expression in Escherichia coli of chemically synthesized genes for human insulin." *Proceedings of the National Academy of Sciences* 76.1 (1979): 106-110.
20. Leader, Benjamin, Quentin J. Baca, and David E. Golan. "Protein therapeutics: a summary and pharmacological classification." *Nature Reviews Drug Discovery* 7.1 (2008): 21-39.
21. Vigneri, R., S. Squatrito, and L. Sciacca. "Insulin and its analogs: actions via insulin and IGF receptors." *Acta diabetologica* 47.4 (2010): 271-278.
22. Kiss, Zoltán, et al. "Discovery and basic pharmacology of erythropoiesis-stimulating agents (ESAs), including the hyperglycosylated ESA, darbepoetin alfa: an update of the rationale and clinical impact." *European journal of clinical pharmacology* 66.4 (2010): 331-340.
23. Walsh, Gary. "Biopharmaceutical benchmarks 2010." *Nature biotechnology*28.9 (2010): 917.
24. Almagro, Juan C., and Johan Fransson. "Humanization of antibodies." *Front Biosci* 13 (2008): 1619-1633.
25. Lonberg, Nils. "Fully human antibodies from transgenic mouse and phage display platforms." *Current opinion in immunology* 20.4 (2008): 450-459.
26. Better, Marc, et al. "Escherichia coli secretion of an active chimeric antibody fragment." *Science* 240.4855 (1988): 1041-1043.
27. Seemann, G., et al. *Antibodies as carriers of cytotoxicity*. Basel, Switzerland: Karger, 1992.
28. Gebauer, Michaela, and Arne Skerra. "Engineered protein scaffolds as next-generation antibody therapeutics." *Current opinion in chemical biology* 13.3 (2009): 245-255.

Chapter 2:

1. Donnes P, Elofsson A. Prediction of MHC class I binding peptides, using SVMHC. Bmc Bioinformatics 2002;3:-.
2. Liu W, Meng XS, Xu QQ, Flower DR, Li TB. Quantitative prediction of mouse class I MHC peptide binding affinity using support vector machine regression (SVR) models. Bmc Bioinformatics 2006;7:-.
3. Salomon J, Flower DR. Predicting class II MHC-peptide binding: A kernel based approach using similarity scores. Bmc Bioinformatics 2006;7:-.
4. Chen JR, Chang BH, Allen JE, Stiffler MA, MacBeath G. Predicting PDZ domain-peptide interactions from primary sequences. Nat Biotechnol 2008;26(9):1041-1045.
5. Gan WX, Roux B. Binding specificity of SH2 domains: Insight from free energy simulations. Proteins-Structure Function and Bioinformatics 2009;74(4):996-1007.

6. Rognan D, Scapozza L, Folkers G, Daser A. Molecular-Dynamics Simulation of Mhc-Peptide Complexes as a Tool for Predicting Potential T-Cell Epitopes. Biochemistry 1994;33(38):11476-11485.
7. Suenaga A, Hatakeyama M, Ichikawa M, Yu X, Futatsugi N, Narumi T, Fukui K, Terada T, Taiji M, Shirouzu M, Yokoyama S, Konagaya A. Molecular dynamics, free energy, and SPR analyses of the interactions between the SH2 domain of Grb2 and ErbB phosphotyrosyl peptides. Biochemistry 2003;42(18):5195-5200.
8. Zhang J, King CA, Dalby K, Ren P. Conformational preference of ChaK1 binding peptides: a molecular dynamics study. PMC Biophys 2010;3(1):2.
9. Brinkworth RI, Breinl RA, Kobe B. Structural basis and prediction of substrate specificity in protein serine/threonine kinases. Proceedings of the National Academy of Sciences of the United States of America 2003;100(1):74-79.
10. Hou T, Xu Z, Zhang W, McLaughlin WA, Case DA, Xu Y, Wang W. Characterization of domain-peptide interaction interface: a generic structure-based model to decipher the binding specificity of SH3 domains. Mol Cell Proteomics 2009;8(4):639-649.
11. Sanchez IE, Beltrao P, Stricher F, Schymkowitz J, Ferkinghoff-Borg J, Rousseau F, Serrano L. Genome-wide prediction of SH2 domain targets using structural information and the FoldX algorithm. PLoS Comput Biol 2008;4(4):e1000052.
12. Raman S, Vernon R, Thompson J, Tyka M, Sadreyev R, Pei J, Kim D, Kellogg E, DiMaio F, Lange O, Kinch L, Sheffler W, Kim BH, Das R, Grishin NV, Baker D. Structure prediction for CASP8 with all-atom refinement using Rosetta. Proteins-Structure Function and Bioinformatics 2009;77 Suppl 9:89-99.
13. Sood VD, Baker D. Recapitulation and design of protein binding peptide structures and sequences. J Mol Biol 2006;357(3):917-927.
14. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28(1):235-242.
15. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. Clustal W and Clustal X version 2.0. Bioinformatics 2007;23(21):2947-2948.
16. Fiser A, Sali A. Modeller: generation and refinement of homology-based protein structure models. Methods Enzymol 2003;374:461-491.
17. Smith CA, Kortemme T. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. J Mol Biol 2008;380(4):742-756.
18. DeLano WL. The PyMOL Molecular Graphics System, http://www.pymol.org. 2002.
19. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. Journal of Molecular Biology 1997;268(1):209-225.

20. Dunbrack RL, Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Science 1997;6(8):1661-1681.
21. Habib N, Kaplan T, Margalit H, Friedman N. A novel Bayesian DNA motif comparison method for clustering and retrieval. PLoS Comput Biol 2008;4(2):e1000010.
22. London N, Movshovitz-Attias D, Schueler-Furman O. The structural basis of peptide-protein binding strategies. Structure 2010;18(2):188-199.
23. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res 2004;14(6):1188-1190.
24. Benos PV, Bulyk ML, Stormo GD. Additivity in protein-DNA interactions: how good an approximation is it? Nucleic Acids Research 2002;30(20):4442-4451.
25. Yang L, Nolan JP. High-throughput screening and characterization of clones selected from phage display libraries. Cytometry Part A 2007;71A(8):625-631.
26. Obenauer JC, Cantley LC, Yaffe MB. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Res 2003;31(13):3635-3641.
27. Tonikian R, Zhang Y, Sazinsky SL, Currell B, Yeh JH, Reva B, Held HA, Appleton BA, Evangelista M, Wu Y, Xin X, Chan AC, Seshagiri S, Lasky LA, Sander C, Boone C, Bader GD, Sidhu SS. A specificity map for the PDZ domain family. PLoS Biol 2008;6(9):e239.
28. Stormo GD, Fields DS. Specificity, free energy and information content in protein-DNA interactions. Trends Biochem Sci 1998;23(3):109-113.
29. Huse M, Kuriyan J. The conformational plasticity of protein kinases. Cell 2002;109(3):275-282.
30. Gong W, Zhou D, Ren Y, Wang Y, Zuo Z, Shen Y, Xiao F, Zhu Q, Hong A, Zhou X, Gao X, Li T. PepCyber:P∼PEP: a database of human protein protein interactions mediated by phosphoprotein-binding domains. Nucleic Acids Res 2008;36(Database issue):D679-683.
31. Tatusova TA, Madden TL. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol Lett 1999;174(2):247-250.
32. Yongkiettrakul S, Byeon IJ, Tsai MD. The ligand specificity of yeast Rad53 FHA domains at the +3 position is determined by nonconserved residues. Biochemistry 2004;43(13):3862-3869.
33. Yuan C, Yongkiettrakul S, Byeon IJ, Zhou S, Tsai MD. Solution structures of two FHA1-phosphothreonine peptide complexes provide insight into the structural basis of the ligand specificity of FHA1 from yeast Rad53. J Mol Biol 2001;314(3):563-575.
34. Petsalaki E, Stark A, Garcia-Urdiales E, Russell RB. Accurate prediction of peptide binding sites on protein surfaces. PLoS Comput Biol 2009;5(3):e1000335.
35. Raveh B, London N, Schueler-Furman O. Sub-angstrom modeling of complexes between flexible peptides and globular proteins. Proteins: Structure, Function, and Bioinformatics 2010;9999(999A):NA.

36. Brenke R, Kozakov D, Chuang GY, Beglov D, Hall D, Landon MR, Mattos C, Vajda S. Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. Bioinformatics 2009;25(5):621-627.
37. Humphris EL, Kortemme T. Design of multi-specificity in protein interfaces. PLoS Comput Biol 2007;3(8):e164.
38. Beuming T, Skrabanek L, Niv MY, Mukherjee P, Weinstein H. PDZBase: a protein-protein interaction database for PDZ-domains. Bioinformatics 2005;21(6):827-828.
39. Diella F, Gould CM, Chica C, Via A, Gibson TJ. Phospho.ELM: a database of phosphorylation sites--update 2008. Nucleic Acids Res 2008;36(Database issue):D240-244.
40. Khati M, Pillay TS. Phosphotyrosine phosphoepitopes can be rapidly analyzed by coexpression of a tyrosine kinase in bacteria with a T7 bacteriophage display library. Anal Biochem 2004;325(1):164-167.
41. Schutkowski M, Reimer U, Panse S, Dong L, Lizcano JM, Alessi DR, Schneider-Mergener J. High-content peptide microarrays for deciphering kinase specificity and biology. Angew Chem Int Ed Engl 2004;43(20):2671-2674.
42. Sparks AB, Rider JE, Hoffman NG, Fowlkes DM, Quillam LA, Kay BK. Distinct ligand preferences of Src homology 3 domains from Src, Yes, Abl, Cortactin, p53bp2, PLCgamma, Crk, and Grb2. Proc Natl Acad Sci U S A 1996;93(4):1540-1544.

Chapter 3:

1. Yifan Song MT, Andrew Leaver-Fay, James Thompson, and David Baker. Structure guided forcefield optimization unpublished 2011.
2. Brooks BR, Brooks CL, 3rd, Mackerell AD, Jr., Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M. CHARMM: the biomolecular simulation program. J Comput Chem 2009;30(10):1545-1614.
3. Lazaridis T, Karplus M. Effective energy function for proteins in solution. Proteins 1999;35(2):133-152.
4. Dunbrack RL, Jr., Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci 1997;6(8):1661-1681.
5. Hwang H, Vreven T, Janin J, Weng Z. Protein-protein docking benchmark version 4.0. Proteins 2010;78(15):3111-3114.
6. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 2004;32(Database issue):D226-229.
7. Tyka MD, Keedy DA, Andre I, Dimaio F, Song Y, Richardson DC, Richardson JS, Baker D. Alternate states of proteins revealed by detailed energy landscape mapping. J Mol Biol 2011;405(2):607-618.

8. Kleywegt GJ, Harris MR, Zou JY, Taylor TC, Wahlby A, Jones TA. The Uppsala Electron-Density Server. Acta Crystallogr D Biol Crystallogr 2004;60(Pt 12 Pt 1):2240-2249.
9. Frank DiMaio DB. Increasing the Radius of Convergence of Molecular Replacement by Density and Energy Guided Protein Structure Optimization. Nature 2011.
10. Smith DK, Radivojac P, Obradovic Z, Dunker AK, Zhu G. Improved amino acid flexibility parameters. Protein Sci 2003;12(5):1060-1072.
11. Elcock AH. Prediction of functionally important residues based solely on the computed energetics of protein structure. J Mol Biol 2001;312(4):885-896.
12. Jaramillo A, Wernisch L, Hery S, Wodak SJ. Folding free energy function selects native-like protein sequences in the core but not on the surface. Proc Natl Acad Sci U S A 2002;99(21):13554-13559.
13. Jaramillo A, Wodak SJ. Computational protein design is a challenge for implicit solvation models. Biophys J 2005;88(1):156-171.
14. Humphris EL, Kortemme T. Design of multi-specificity in protein interfaces. PLoS Comput Biol 2007;3(8):e164.
15. Xiang Z, Steinbach PJ, Jacobson MP, Friesner RA, Honig B. Prediction of side-chain conformations on protein surfaces. Proteins 2007;66(4):814-823.
16. Xiang Z, Soto CS, Honig B. Evaluating conformational free energies: the colony energy and its application to the problem of loop prediction. Proc Natl Acad Sci U S A 2002;99(11):7432-7437.

Chapter 4:

References

1. Cantor JR*, et al.* (2011) Therapeutic enzyme deimmunization by combinatorial T-cell epitope removal using neutral drift. *Proceedings of the National Academy of Sciences* 108(4):1272-1277.
2. Tangri S*, et al.* (2005) Rationally engineered therapeutic proteins with reduced immunogenicity. *The Journal of Immunology* 174(6):3187-3196.
3. Zubler RH (2001) Naive and memory B cells in T-cell-dependent and T-independent responses. *Springer seminars in immunopathology*, (Springer), pp 405-419.
4. Goldenberg MM (1999) Trastuzumab, a recombinant DNA-derived humanized monoclonal antibody, a novel agent for the treatment of metastatic breast cancer. *Clinical therapeutics* 21(2):309.
5. Harding FA*, et al.* (2005) A B-lactamase with reduced immunogenicity for the targeted delivery of chemotherapeutics using antibody-directed enzyme prodrug therapy. *Molecular cancer therapeutics* 4(11):1791-1800.
6. Vita R*, et al.* (2010) The immune epitope database 2.0. *Nucleic acids research* 38(suppl 1):D854-D862.
7. Nielsen M, Lund O, Buus S, & Lundegaard C (2010) MHC Class II epitope predictive algorithms. *Immunology* 130(3):319-328.

8. Andrew SP, Griswold KE, & Bailey-Kellogg C (2011) Optimization of therapeutic proteins to delete T-cell epitopes while maintaining beneficial residue interactions. *Journal of bioinformatics and computational biology* 9(02):207-229.
9. Parker AS, Zheng W, Griswold KE, & Bailey-Kellogg C (2010) Optimization algorithms for functional deimmunization of therapeutic proteins. *BMC bioinformatics* 11(1):180.
10. Choi Y, Griswold KE, & Bailey-Kellogg C (2013) Structure-based redesign of proteins for minimal T-cell epitope content. (Translated from eng) *J Comput Chem* 34(10):879-891 (in eng).
11. Nivon LG, Bjelic S, King C, & Baker D (2013) Automating Human Intuition for Protein Design. *Proteins*.
12. Anonymous (Immune Epitope Database: MHC-II Binding Dataset.
13. Flicek P*, et al.* (2012) Ensembl 2012. (Translated from eng) *Nucleic Acids Res* 40(Database issue):D84-90 (in eng).
14. Nielsen M, Justesen S, Lund O, Lundegaard C, & Buus S (2010) NetMHCIIpan-2.0-Improved pan-specific HLA-DR predictions using a novel concurrent alignment and weight optimization training procedure. *Immunome research* 6(1):9.
15. Singh H & Raghava GPS (2001) ProPred: prediction of HLA-DR binding sites. (Translated from English) *Bioinformatics* 17(12):1236-1237 (in English).
16. Southwood S*, et al.* (1998) Several common HLA-DR types share largely overlapping peptide binding repertoires. *The Journal of Immunology* 160(7):3363-3373.
17. Pedelacq JD, Cabantous S, Tran T, Terwilliger TC, & Waldo GS (2006) Engineering and characterization of a superfolder green fluorescent protein (vol 24, pg 79, 2005). (Translated from English) *Nature Biotechnology* 24(9):1170-1170 (in English).
18. Brazelton TR & Blau HM (2005) Optimizing Techniques for Tracking Transplanted Stem Cells In Vivo. *STEM CELLS* 23(9):1251-1265.
19. Persons DA*, et al.* (1998) Use of the green fluorescent protein as a marker to identify and track genetically modified hematopoietic cells. *Nat Med* 4(10):1201-1205.
20. FitzGerald DJ, Wayne AS, Kreitman RJ, & Pastan I (2011) Treatment of Hematologic Malignancies with Immunotoxins and Antibody-Drug Conjugates. (Translated from English) *Cancer Res* 71(20):6300-6309 (in English).
21. Kreitman RJ*, et al.* (2012) Phase I trial of anti-CD22 recombinant immunotoxin moxetumomab pasudotox (CAT-8015 or HA22) in patients with hairy cell leukemia. *Journal of Clinical Oncology* 30(15):1822-1828.
22. Wayne AS*, et al.* (2011) A novel anti-CD22 immunotoxin, moxetumomab pasudotox: phase I study in pediatric acute lymphoblastic leukemia (ALL). *Blood*, (AMER SOC HEMATOLOGY 1900 M STREET. NW SUITE 200, WASHINGTON, DC 20036 USA), pp 113-113.
23. Hassan R*, et al.* (2007) Phase I study of SS1P, a recombinant anti-mesothelin immunotoxin given as a bolus IV infusion to patients with mesothelin-

expressing mesothelioma, ovarian, and pancreatic cancers. *Clinical Cancer Research* 13(17):5144-5149.

24. Mazor R*, et al.* (2012) Identification and elimination of an immunodominant T-cell epitope in recombinant immunotoxins based on Pseudomonas exotoxin A. (Translated from English) *P Natl Acad Sci USA* 109(51):E3597-E3603 (in English).
25. De Groot AS, Terry F, Cousens L, & Martin W (2013) Beyond humanization and de-immunization: tolerization as a method for reducing the immunogenicity of biologics. *Expert review of clinical pharmacology* 6(6):651-662.
26. Baker M & Carr F (2010) Pre-clinical considerations in the assessment of immunogenicity for protein therapeutics. (Translated from eng) *Curr Drug Saf* 5(4):308-313 (in eng).
27. Greenbaum JA*, et al.* (2007) Towards a consensus on datasets and evaluation metrics for developing B,Äêcell epitope prediction tools. *Journal of Molecular Recognition* 20(2):75-82.
28. Nagata S & Pastan I (2009) Removal of B cell epitopes as a practical approach for reducing the immunogenicity of foreign protein-based therapeutics. *Advanced drug delivery reviews* 61(11):977-985.
29. Onda M*, et al.* (2008) An immunotoxin with greatly reduced immunogenicity by identification and removal of B cell epitopes. *Proceedings of the National Academy of Sciences* 105(32):11311-11316.
30. Hammer J*, et al.* (1993) Promiscuous and allele-specific anchors in HLA-DR-binding peptides. *Cell* 74(1):197.
31. Yanover C & Bradley P (2011) Large-scale characterization of peptide-MHC binding landscapes with structural simulations. *Proceedings of the National Academy of Sciences* 108(17):6981-6986.
32. Fleishman SJ*, et al.* (2011) RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite. *PloS one* 6(6):e20161.
33. Pastan I, Beers R, & Bera TK (2004) Recombinant immunotoxins in the treatment of cancer. *Antibody Engineering*, (Springer), pp 503-518.

Vita

The author was physically instantiated as a human male in 1983 after crash-landing his transdimensional lightship somewhere near Austin in the hill country of central Texas. In order to interject some style into the tactless aesthetic of the extant singularity, he earned a Bachelor of Science degree in Biomedical Engineering at the University of Texas in 2007 before moving to Seattle to complete a Doctor of Philosophy in Biochemistry at the University of Washington in 2014.