THE BIOPHYSICAL BASIS OF A PROTEASE – SUBSTRATE INTERACTION LANDSCAPE

by

MANASI PETHE

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Chemistry and Chemical Biology

Written under the direction of

Professor Sagar D. Khare

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October 2017

ABSTRACT OF THE DISSERTATION

The biophysical basis of a protease – substrate interaction landscape

By MANASI PETHE

Dissertation Director:

Professor Sagar D. Khare

Characterizing the specificity of proteases is important to illuminate their role as signaling moieties in a range of diverse biological processes. Proteases often display multispecificity, which is the ability of a single receptor protein molecule to interact with multiple substrates. The ability to accurately recapitulate protease specificity profiles would aid in the design of custom proteases designed to cleave targets in biotechnology or therapeutic scenarios. Current specificity prediction methods use machine - learning techniques that are not generalizable and relatively slow, and thus limited in use for prediction and especially design of multispecificity.

We tackle these challenges using a two - pronged approach - by increasing the accuracy of scoring for biophysical protease substrate models as well as by hastening the process of sampling. We develop a general approach for prediction of protease specificity through the construction of high - resolution atomic models, using protein structure

modeling and biophysical energetic evaluation of enzyme substrate complexes. Specifically, we develop a discriminatory scoring function using enzyme design modules from Rosetta and Amber-MMPBSA. Analysis of structural models provides physical insight into the structural basis for the observed specificities. We further test the predictive capability of the model by designing and experimentally characterizing the cleavage of four novel substrate motifs for the Hepatitis C virus NS3/4A protease using an in vivo assay. The presented structure-based approach is generalizable to other protease enzymes with known or modeled structures, and complements existing experimental methods for specificity determination. To improve our sampling approach, we develop a rapid, flexible-backbone self-consistent mean field theory-based technique, MFPred, for multispecificity modeling at protein-peptide interfaces. We benchmark our method by predicting experimentally determined peptide specificity profiles for a range of receptors. Our approach should enable the design of a wide range of altered receptor proteins with programmed multispecificities.

Viral systems encoding proteases are exemplars of multispecificity. Multispecific proteases mediate the precise cleavage of the polyprotein during replication and viral assembly. The HCV NS3/4A protease is a multispecific protease, which is likely a result of both positive selection pressure to maintain cleavability of its four native substrates, i.e. known sites on the polyprotein, and negative selection pressure to avoid cleavage of other sites in the polyprotein. We map the specificity landscape of the HCV NS3/4A protease to obtain a comprehensive understanding of the protease–substrate interaction network. Using an in vivo yeast surface display assay, Fluorescence Assisted Cell

Sorting, Next Generation Sequencing technology and computational modeling using Rosetta and Amber packages, we were able to reconstruct the entire (3.2 million sequences) HCV NS3/4A substrate landscape learning from the sequences identified in our experiment, using an SVM based approach.

The work discussed in this thesis gives us insight into the biophysical basis of protease specificity. This work can further be used in rational design of custom proteases and in understanding the mechanisms underlying co-evolution of protease substrate interactions in viral proteases, as well as robustness of the interaction.

ACKNOWLEDGEMENTS

At the outset I want to thank my advisor, Dr. Sagar Khare for his constant guidance, support, suggestions, critique, and for always encouraging me to stand for the ideas I believed in. I would also like to whole - heartedly thank Dr. Case, Dr. Marcotrigiano, Dr. Burley and Dr. Khiabanian for advice, support and encouragement when I needed it the most. To Dr. Celia Schiffer, Dr. Amy Keating and Dr. Theresa Choi, I express my appreciation for their insights and time. I also want to express my warmest gratitude to all the members of the Khare lab for the engaging discussions - scientific and otherwise, food, support and friendship. First and foremost Dr. Srinivas Annavarappu and Aliza Rubenstein for close collaborations and helpful discussions. I would like to specially mention Nancy Hernandez, Kristin Blacklock, Lu Yang, William Hansen, Dr. Brahm Yachnin, Elliott Dolan, and Dmitri Zorine for being the most wonderful labmates and friends.

I would be remiss not to mention the wonderful people I met over the course of the PhD - Sushmita Patwardhan, Malathi Kalyanikar, Sneha Raghunandan, Rohit Gupte, Dr. Prasad Subramanium and Bharat Murali, who were my first friends when I started at Rutgers and have continually supported me since, both professionally and personally. Additionally, Maulik, Myth, Mig, Abhi, Ava, Tanvi, Kaizad, Vineet, Sani and Sattu for making life in Jersey feel like home! My warmest thanks to Jamie Palmer (Woody's cafe) for handing me my lunch with a never waning smile. Last but not least, heartfelt thanks to my parents

for always giving me two polar opinions – invariably making me think about all possible outcomes, believing in me constantly and unconditionally supporting my choices.

Parts of the thesis have been previously published as follows:

Chapter 2 of the thesis has been published as:

Pethe MA, Rubenstein A, Khare SD, Large scale structure based prediction and identification of novel protease substrates by computational protein design, Journal of Molecular Biology (2017); 429(2): 220-236. doi: 10.1016/j.jmb.2016.11.031

Chapter 3 has been published as:

Rubenstein, A.B., Pethe, M.A., and Khare, S.D. (2017). MFPred: Rapid and accurate prediction of protein-peptide recognition multispecificity using self-consistent mean field theory. PLoS Comput Biol 13, e1005614.

Chapter 4 has been published as a preprint at: Manasi A. Pethe, Aliza B. Rubenstein, Dmitri Zorine, Sagar D. Khare, Biophysical determinants of mutational robustness in a viral molecular fitness landscape, bioRxiv 172197; doi: https://doi.org/10.1101/172197

DEDICATION

To my parents,
Aayo and Babi
For the constant love and support.

To Guppa
For always making me believe that I can.

Table of Contents

List of Tables

List of Illustrations

Chapter 1: Introduction

Proteases (also known as peptidases, proteinases) are enzymes that cleave the peptide bond. Proteases were perceived to be enzymes with a principally catabolic function, especially digestive enzymes such as trypsin and chymotrypsin,(Bender & Kaiser 1962; Kasserra & Laidler 1969; Celis-Guerrero et al. n.d.) that demonstrate wide specificities of cleavage thus mediating the truncation of proteins into smaller peptide fragments e.g. Digestive enzyme trypsin is secreted as trypsinogen(Abita et al. 1969) - an inactive precursor, further activated by enteropeptidase. Once activated, trypsin itself continues to activate its inactive zymogen. Digestive proteases often display broad specificity and are thus secreted as zymogens(Khan & James 1998) to regulate indiscriminate proteolytic activity. Proteases that regulate biological pathways have a narrower specificity profile, e.g. caspases(Riedl & Shi 2004; Pop & Salvesen 2009) are involved in apoptotic pathway regulation, cathepsin B(Alapati et al. 2014) is involved in tumor metastasis, MMP2(Jezierska & Motyl 2009; Bauvois 2012) modulates cancer cell migration and growth. Several such examples underline the fact that proteases are ubiquitous in biological regulation and contribute to delicate modulation of pathway regulation to carry out diverse functions. The promiscuity/ specificity of a protease cleavage profile is observed to dictate its functional role in a biological context.

1.1. Diversity in mechanism, occurrence, architecture and specificity

Proteases have evolved multiple times to perform the same reaction via completely different mechanisms, using a variety of different active site architectures and are found to exist in all five kingdoms: Animalia, Plantae, Fungi, Bacteria, Archaea,

Viruses(Nemova & Lysenko 2013). They display a variety of folds and thus a diverse range of interactions at the substrate – protease interface. Several commonly encountered interface interactions between the protease and substrate are hydrogen bonding (HCV protease - substrate(Lin 2006), TEV protease substrate(Phan et al. 2002)), shape complementarity(Prabu-Jeyabalan et al. 2002; Romano et al. 2010; Shen et al. n.d.)(HCV substrate, HIVPR - 1 substrate), electrostatic binding (Harris et al. 1998; Casciola-Rosen et al. 2006; Matthews et al. 1994; Rockwell et al. 2002; Walker et al. 1994)(Granzyme B, Furin). Proteases bind their substrates in a variety of different ways. TEV protease adopts a two-domain antiparallel beta barrel fold wherein the catalytic domain is located at the interface between the two domains creating a specific groove for the substrate binding. This kind of an interface creates specific pockets that define grooves for the side chains of the substrate to fit in, with favorable contacts. Specificity is endowed by large contact surface between the substrate and enzyme (Phan et al. 2002). Most proteases that adopt this fold create a closed binding groove for the substrate have narrow specificity profiles e.g. TEV protease (ENLYFQ -- G), TVMV (ETVRFQ -- G/S), and the 3C protease family of enzymes. In contrast, enzymes such as the HCV protease that adopt the chymotrypsin - fold of proteases, as well as the HIV protease 1 – which is a dimer, have exposed active sites – with defined substrate pockets at only a few positions. Substrate binding in such cases is primarily governed by favorable hydrogen bonds rather than the architecture of the active site groove. HCV protease has a network of hydrogen bonds between the backbone of a bound substrate and a beta strand on the protease. This strand runs parallel to the substrate - binding groove and mediates the binding interaction between the substrates for proteases that display a chymotrypsin like fold. The HIV PR1

does not have a preferred consensus substrate sequence, instead the selection for substrates that are cleaved occurs by choosing substrates that "fit" best in the substrate - binding pocket. Proteases such as HCV and HIV demonstrate relaxed, multi specificities in their substrate preference both in high throughput assays as well as in a biological context. Another commonly seen mode of substrate binding is favorable electrostatic interactions. The substrates for such proteases are highly enriched in charged residues that are complementary to the charge on the protease active site e.g. Furin, Granzyme. Proteases such as HTRA-1(Clausen et al. 2011), have a catalytic domain and an additional binding domain. All of the specificity is dictated by the binding domain where as the catalytic part is indiscriminate.

1.2. Biophysical consequence of substrate binding, chemical transformation on protease specificity

Across diverse protease families it is observed that the active site architecture has a dual role - substrate binding as well as efficient catalysis. Schechter and Berger(Schechter & Berger 1967) devised a series of experiments to determine proteolysis rates for poly-alanine peptides of various lengths. Based on the proteolytic constant for substrates of different polyalanine lengths, Schechter and Berger determined the sites on the substrate, which play an important role in efficient binding and catalysis events. They described seven sites on the substrate on either side of the scissile bond – the N terminal sites were termed the S sites labeled as S1, S2, … SN outward from the scissile bond where as the C terminal sites were labeled as S' sites labeled S1', S2', S3', … SN'. The sites on the protease side were correspondingly labeled as P1 and P1', etc. This nomenclature is

extensively used and is highly effective as a descriptor for protease – substrate binding specificities. Apart from the binding site architecture, the mechanistic steps involved in the proteolysis reaction also contribute to the substrate selectivity(Hedstrom 2002). For instance, serine proteases act via a multistep mechanism consisting of the following steps (a) formation of enzyme substrate complex (b) formation of the acyl enzyme intermediate (c) hydrolysis of acyl enzyme intermediate. Notably that the hydrolysis step (product dissociation) is assumed to be faster than the chemical transformation step. If substrate selection could be considered as governed by the rate determining step, it is assumed to be primarily influenced by the formation of the acyl enzyme intermediate and much less by binding affinity.

In summary, the architecture of the binding site influences substrate selectivity by allowing/ disallowing certain residue types to "fit" into the active site substrate pockets, whereas the mechanistic steps contribute to substrate selectivity by enhancing the rate of chemical transformation for certain substrates over others. Substrates are considered to be better "cleaved" than others depending on their comparative rates of chemical transformation, rather than a comparison between substrate binding affinity. Thus, when constructing a static biophysical model of proteolysis, the most accurate structural representation of substrate selection is the acyl enzyme intermediate step. It is well understood that the active site geometry and substrate binding groove as well as the overall fold of the protease are universal determinants of specificity and function(Tyndall et al. 2005; Hedstrom 2002). In understanding the biophysical rules that underlie protease specificity, it is important to contemplate the role of substrate binding (energetic

determinants in binding) as well as catalytic turnover geometry (formation of the intermediate, geometry of scissile bond and active site residues).

1.3. Statement of the problem

In consideration of the biological discussion of understanding protease specificity this dissertation attempts to answer the following questions: Is there a common underlying biophysical basis of protease specificity? If so, can we use these biophysical rules to recapitulate known specificities across diverse sets of proteases? Can we understand protease specificity well enough for a diverse set of enzymes to not only recapitulate known/ existing specificities, but also to modulate it and engineer new specificities that are unexplored in nature? Specifically we explored the interaction landscape for a multi-specific viral protease – Hepatitis C NS3 protease – by asking, if the protease is capable of cleaving a more diverse substrate profile than is sampled by it in its native biological context.

1.4. Protease specificity prediction and design

Proteases are useful synthetic biology tools. For mass spectrometry experiments, a common preparation step involves proteolysis of complex protein mixtures. For protein purification experiments, the constructs often are fused with an N or C terminal His tag, which bind to the Nickel resin during purification. These tags need to be cleaved off in order to regain enzyme folding, function, sometimes so that the enzymes can crystallize well. Proteases, especially TEV and thrombin(Waugh 2011), act as cleavage agents facilitate the removal of the His tag.

Proteases have widespread use in industrial, biotechnological and therapeutic scenarios. USFDA approved uses for proteases include use as digestive aids (Zenpep), u-PA (urokinase) and t – PA (reteplase and tenecteplase) indicated for thrombotic disease, Factor IX indicated in hemophilia, as well as Botulinum Toxin A (Botox)(Craik et al. 2011). Several of these proteases although extremely active, are very unstable for administration directly to the blood stream. Protease engineering efforts are directed in two ways – (a) increasing the half-life/ stability of these enzymes(Craik et al. 2011; Taguchi et al. 1998; Choudhury et al. 2010) and (b) efforts geared toward increasing activity(Varadarajan et al. 2005; Yi et al. 2013; Flowers & Ann 2013; Li et al. n.d.; Guerrero et al. 2016; Wang et al. 2016; Chang et al. 1994; Khouri et al. 1991; Hill et al. 2016). Competing antibody based therapies work through a stoichiometric effect, hence an advantage of enzyme - based therapies is that lower doses can be administered due to catalytic turnover of the therapeutic protease(Craik et al. 2011). There are several instances of increase in localized protease activity in tumorigenic cells. Recently designed cancer therapies use pro-drug approaches that are activated in the presence of up-regulated cellular proteases(Choi et al. 2012). These therapies provide a means to reduce cytotoxicity, which is a commonly known side effect.

Several successful specificity-switching studies(Hill et al. 2016; Hashimoto et al. 2011) have been performed on enzymes such as TEV protease, caspases and metalloproteases. The studies focus on proteases that have similar active site architecture. It has also been observed that the broad/narrow nature of specificity changes with the architecture of the

active site, which is predetermined by the overall protease fold. In order to successfully predict and modulate protease specificity it is evident that we need to understand better how the rules of specificity are governed by the active site architecture and identity of residues in the fold. Elucidating the specificity profile for proteases and designing experiments outside of their biological context enables further understanding of specificity rules. Several library-based techniques are used to investigate the substrate site preferences of protease enzymes. While positional scanning libraries(Backes et al. 2000; Schneider & Craik 2009) are highly recommended for investigating the average preferences at a given position on the substrate site, microarrarys(Salisbury et al. 2002; Gosalia et al. 2005) offer the advantage of quicker testing times for individual substrates and thus to discover substrate cooperativity effects that are invariably averaged out by other techniques. Gosalia and Diamond(Gosalia & Diamond 2003) developed a nano-droplet microarray using glycerol and DMSO to form a suspension of the peptides. Using an aerosol to activate the system, the study was able to not only detect positional preferences but also detect covariance, which would have been averaged out by methods like positional scanning. Biological display systems (phage display(Ratnikov et al. 2009; Matthews et al. 1994; Smith 1985; and & Petrenko 1997), mRNA display(Liu et al. 2000; Amstutz et al. 2001) and ribosome display(He & Taussig 2002; Zahnd et al. 2007), bacterial display(Kenrick & Daugherty 2010; Daugherty 2007; Getz et al. 2012), yeast display(Gai & Wittrup 2007; Yi et al. 2015; Park et al. 2006; Cherf & Cochran 2015), mammalian display(Ho & Pastan 2009; Zhou et al. 2010; Bowers et al. 2014)) have been used traditionally to test a high number of sequences (10^7 -10^9). While prokaryotic display techniques have been used previously, eukaryotic display techniques such as

yeast surface display and mammalian display are better suited to realize better expression and folding of the proteins to be tested. Cell based techniques also lend the added advantage of being amenable to rounds of directed evolution that are useful in protein engineering(Bloom & Arnold 2009; Traxlmayr & Shusta 2017; Boder & Wittrup 2000).

Experimental exploration of sequence space is often rendered incomplete due to the demands of time, resource and other limiting factors. The results are also prone to biases depending on the assay system. Computational biology tools serve to support such experimental searches by guiding experimental techniques to narrow down searchable sequence space(Punta et al. 2008). The proteolysis reaction is tied to the spatial conformation of the two proteins, their relative configuration related to one another and conformational changes that may occur during a binding step. It has been previously shown that cleavage can be predicted with accuracy from knowledge of the amino acid sequence alone(von Heijne 1986). These computational techniques rely on machine learning(Tarca et al. 2007), pattern recognition and structure - energetics based macromolecular modeling software(Maximova et al. 2016; Alford et al. 2017). Several pattern recognition-based tools such as MEROPS(Rawlings et al. 2010), PoPs(Boyd et al. 2004), Prosper(Song et al. 2012) have been developed in order to predict protease recognition profiles. While these tools are useful in predicting profiles of extremely specific proteases they do not perform accurately for proteases that demonstrate wider, relaxed specificities. Once trained on a protease, the algorithm cannot be used in a transferrable manner across all protease classes. The algorithms are trained based on knowledge from experiment, which might introduce sampling bias into the training set

from which rule extraction occurs. Thus the training occurs on an incomplete dataset and the recognition software does not reflect the underlying biological model of the protease but rather is indicative of the sampling limits of the experiment.

Because of diversity in protease families the development of prediction software often tied to a single family. Research favored artificial neural networks and then shifted towards SVMs. For some linearly separable datasets, linear SVMs or decision trees could also be useful(duVerle & Mamitsuka 2012). One of the strengths of supervised learning models is to be able to work across various kinds of biological models. The quality of the supervised learning method is based upon feature selection – amino acid sequences are an obvious choice however some empirical data indicates that secondary or tertiary structural preference has shown to improve learning(Sakai et al. 1987). For feature inputting, the amino acid sequences need to be translated as a vector - encoding scheme for the supervised learning algorithm to work. The most commonly used scheme is the canonical binary encoding where each residue is assumed to be a unique 20- long binary vector(Qian & Sejnowski 1988). The problem with this scheme is that the amino acids are considered to be equidistant from one another and this scheme has no basis in chemistry or biology. However, this method has been noted to generally work better than other encoding schemes(Yang & Chou 2004; Barkan et al. 2010). A common issue is the imbalance between the positive (cleaved) and negative (uncleaved) classes for a proteolytic learning study. To encounter this, one can increase the penalty factor associated with misclassification of positive class sequences or increase the sampling of a subset of negative sequences to increase the parity between the two data classes(Akbani

et al. 2004). Supervised learning methods are frequently evaluated using an AUROC (Area Under the Receiver Operator Characteristic)(Hand & Till 2001). Efforts to increase the AUROC value sometimes leads to over-fitting of the data, and the best way to generalize this is to use cross validation and a considerably large experimental set for training to ensure that feature set size is not large compared to the learning data.

One of the commonly used methods to study cleavability of a substrate is to calculate the position specific probability of occurrence for each amino acid. This is represented via a sequence logo, which can represent the likelihood of occurrence corresponding to the size of the represented amino acid at a certain substrate position. Work by Poorman et al(Poorman et al. 1991) is equivalent to the work on position specific scoring matrix (PSSM)(Henikoff & Henikoff 1994). PSSMs have been applied to protease systems like caspases with a certain amount of success(Backes et al. 2005; Wilkins et al. 1999; Garay-Malpartida et al. 2005). However, the biologically inaccurate assumption that substrate positions are independently governed in substrate recognition events makes it harder to apply this to other complex protease systems. Several other techniques such as artificial neural networks, decision trees, rule extraction methods, hidden markov models and kernels each with its own set of advantages and limitations(duVerle & Mamitsuka 2012).

Calculation of interaction energies at the protease substrate interface can be performed using macromolecular modeling software such as Rosetta(Leaver-Fay et al. 2011; Fleishman et al. 2011; Alford et al. 2017), AMBER(Case et al. 2005),etc. Recapitulation of protease specificity using energy based techniques promises to capture the underlying

biophysical rules governing specificity. Thus, learning these biophysical rules would enable us to recapitulate protease specificity for narrow/broadly specific proteases. This method would also be transferable across various protease datasets irrespective of sampling and training biases. However, due to the intricate nature of interactions that need to be computed to calculate all possible interaction energies, several of theses interactions have to be simplified in the energy models. A result of this simplification is that certain interactions are disregarded. Sequence based learning can help to provide orthologous information that could be missing in our repertoire of energy based learning. Thus with current technologies considered our best strategy to decode protease specificity is to use both sequence information as well as energetic interactions in prediction and design.

1.5. Viral proteases: a model multi-specific system

Many viruses use a replication strategy that involves the translation of a large polyprotein, which is cleaved into its functional units by cellular or viral proteases. Viral proteases face heavy selection pressure since they need to evolve to cleave only specific parts of the polyprotein (positive selection) and not the rest (negative selection). Viral proteases are thus natural model systems that have evolved for multi-specific cleavage(Yost & Marcotrigiano 2013). The cleavage sites that are N terminal to the protease are cleaved by signal peptidases and the NS2 auto-protease(Carrère-Kremer et al. 2004). The four canonical cleavage sites are cleaved specifically by the NS3 protease. This cleavage event is highly specific since it allows for further maturation and viral assembly. The protease has evolved to cleave its canonical sequences but not other sites

on the viral polyprotein. The NS5B (RNA polymerase) present in this system is highly error prone(Ribeiro et al. 2012). It has an error rate two fold higher than the human polymerase enzymes. This creates quasi-species of the virus with each replication cycle. In spite of the high mutational rate, this protease has evolved to maintain its specific yet broad cleavage profile.

There are several viral systems that contain multi-specific proteases of similar nature. Nature has successfully modeled multi -specificity into viral protease systems, very precisely cleaving only the sequence of interest. Modeling our design strategy using viral proteases as a template would help to unravel the biophysical rules of multispecificity.

1.6. Protease – substrate interaction landscape

The concept of an evolutionary landscape is used to describe the process of genetic drift on a gene, protein, population or species. In the context of a protein, we can visualize the protein to be sampling several areas of sequence space. This searchable sequence space lies on the x - y plane of the evolutionary landscape. The z plane describes the function, stability, foldability, in other words a “fitness” parameter expressed as height where fittest variants lie on “peaks” and unfit variants are sampled in “valley” regions. The idea of representing genotype – phenotype correlation in a 3D space was first described by Sewall Wright (Sewall Wright 1932). If movements along the landscape lead to changes in fitness that are small then this leads to landscapes that are smooth, whereas if movements along landscapes lead to large changes the landscape is described to be

rugged. The term “evolutionary landscape” is used interchangeably with “adaptive landscape” and “fitness landscape”.

Maynard Smith (1970) was the first to look at molecular evolution as a landscape phenomenon(Svensson & Calsbeek 2012). This means to observe the biophysical landscape of a protein as a network of mutants as being steps away from each other. Thus, there needed to be paths connecting two functional forms of a protein. Maynard Smith was the first to point out that these steps (proteins along the connecting paths) may not be as/ at all functional. This means that moves in different directions of this landscape may not all be movements in an upward direction. This points to the neutral drift theory of evolution. While the fitness landscapes have been investigated since as early as 1932, it has only recently become possible to experimentally explore fitness landscapes via advances in library design and next generation sequencing technologies(Head et al. 2014). In our work, we study the biophysical interactions between a viral protease – substrate as an interaction landscape, keeping the protease sequence constant and varying substrate sequence at all positions of the specificity determining N-terminal pentapeptide.

We envision the problem of protease specificity as a landscape that we need to traverse. Substrates for a given protease can be ranked according to their cleavability. Thus a “good” substrate would be situated on a high, wide peak and would thus be cleavable whereas a “bad” substrate would be situated in a valley and be functionally uncleaved. The ability to predict the consequence of each mutation on specificity would supplement the process of rational design. In traversing the landscape we come across a biological

phenomenon termed epistasis(Breen et al. 2012; Sailer et al. 2017; Weinreich et al. 2005). Epistasis is said to occur when the effect of single mutants is not additive to the effect of the double. For instance, an example where two single mutants A and B are cleaved however the double mutant AB is uncleaved would be a presentation of epistasis. Efforts have driven further understanding and even prediction of epistasis – most notably investigations carried out by Harms(Harms & Thornton 2013; Sailer & Harms 2017; Sailer et al. 2017), Tokuriki(Miton & Tokuriki 2016), etc. Studies by Tokuriki et al have shown that the order of mutations matters in consideration of which functional peak is reached and that it is not necessary that the most functional intermediates will lead to the fittest point on the interaction landscape.

1.7. Outline of the dissertation

Using the aforementioned design tools our overarching goal was to take forward steps toward engineering a library of proteases – Restriction endopeptidases. Similar to restriction enzymes that exist to cleave very specific DNA sequences, we were motivated to develop a variety of proteases that could selectively and specifically cleave protein sequences of choice. This enables several applications for protease use in therapy, biotechnology and synthetic biology. We tackle this problem of specificity by using available cleaved and uncleaved substrate sets (from high - throughput assay literature) to develop a generalizable, biophysical structure based energy function that increases the ability of Rosetta's energy function to distinguish between cleaved and uncleaved substrates for a myriad of proteases (Chapter 2). We also investigate faster techniques to recapitulate multispecificity in proteases by developing a faster sampling technique,

MFPred – discussed in Chapter 3. Further, (Chapter 4) we explore the substrate protease interaction of a multispecific viral protease – Hepatitis C NS3 protease – known to cleave four canonical substrates as part of its biological role. We explore the biophysical basis of this interaction by generating a library of substrates and testing it using a high - throughput yeast based assay. Results from this exploratory study as well as further computation using an SVM based approach, lead to insights in evolutionary trends and future directions in the protease design field.

1.8. References

Abita, J.P., Delaage, M. & Lazdunski, M., 1969. The Mechanism of Activation of Trypsinogen The Role of the Four N-Terminal Aspartyl Residues. European J. Biochem, 8, pp.314–324.

Akbani, R., Kwek, S. & Japkowicz, N., 2004. Applying Support Vector Machines to Imbalanced Datasets. In Springer, Berlin, Heidelberg, pp. 39–50.

Alapati, K. et al., 2014. uPAR and cathepsin B-mediated compartmentalization of JNK regulates the migration of glioma-initiating cells. Stem cell research, 12(3), pp.716–29.

Alford, R.F. et al., 2017. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. Journal of Chemical Theory and Computation, 13(6), pp.3031–3048.

Amstutz, P. et al., 2001. In vitro display technologies: novel developments and applications. Current opinion in biotechnology, 12(4), pp.400–5.

G.P.S. & Petrenko, V.A., 1997. Phage Display. Chem Rev. 97(2), pp.391-410.

Backes, B.J. et al., 2000. Synthesis of positional-scanning libraries of fluorogenic peptide substrates to define the extended substrate specificity of plasmin and thrombin. Nature biotechnology, 18(2), pp.187–93.

Backes, C. et al., 2005. GraBCas: a bioinformatics tool for score-based prediction of Caspase- and Granzyme B-cleavage sites in protein sequences. Nucleic acids research, 33(Web Server issue), pp.W208-13.

Barkan, D.T. et al., 2010. Prediction of protease substrates using sequence and structure

features. Bioinformatics (Oxford, England), 26(14), pp.1714–22.

Bauvois, B., 2012. New facets of matrix metalloproteinases MMP-2 and MMP-9 as cell surface transducers: Outside-in signaling and relationship to tumor progression. Biochimica et Biophysica Acta (BBA) - Reviews on Cancer, 1825(1), pp.29–36.

Bender, M.L. & Kaiser, E.T., 1962. The Mechanism of Trypsin-catalyzed Hydrolyses. The Cinnamoyl-trypsin Intermediate. Journal of the American Chemical Society, 84(13), pp.2556–2561.

Bloom, J.D. & Arnold, F.H., 2009. In the light of directed evolution: pathways of adaptive protein evolution. Proceedings of the National Academy of Sciences of the United States of America, 106 Suppl 1(Supplement 1), pp.9995–10000.

Boder, E.T. & Wittrup, K.D., 2000. Yeast surface display for directed evolution of protein expression, affinity, and stability. Methods in enzymology, 328, pp.430–44.

Bowers, P.M. et al., 2014. Mammalian cell display for the discovery and optimization of antibody therapeutics. Methods, 65(1), pp.44–56.

Boyd, S.E. et al., 2004. PoPS: a computational tool for modeling and predicting protease specificity. Proceedings / IEEE Computational Systems Bioinformatics Conference, CSB. IEEE Computational Systems Bioinformatics Conference, pp.372–81.

Breen, M.S. et al., 2012. Epistasis as the primary factor in molecular evolution. Nature, 490(7421), pp.535–538.

Carrère-Kremer, S. et al., 2004. Regulation of hepatitis C virus polyprotein processing by signal peptidase involves structural determinants at the p7 sequence junctions. The Journal of biological chemistry, 279(40), pp.41384–92.

Casciola-Rosen, L. et al., 2006. Mouse and Human Granzyme B Have Distinct Tetrapeptide Specificities and Abilities to Recruit the Bid Pathway *.

Case, D.A. et al., 2005. The Amber biomolecular simulation programs. Journal of Computational Chemistry, 26(16), pp.1668–1688.

Celis-Guerrero, L.E. et al., Characterization of Proteases in the Digestive System of Spiny Lobster (Panulirus interruptus).

Chang, T.K. et al., 1994. Subtiligase: A tool for semisynthesis of proteins. Biochemistry, 91, pp.12544–12548.

Cherf, G.M. & Cochran, J.R., 2015. Applications of Yeast Surface Display for Protein Engineering. Methods in molecular biology (Clifton, N.J.), 1319, pp.155–75.

Choi, K.Y. et al., 2012. Protease-activated drug development. Theranostics, 2(2), pp.156–78.

Choudhury, D. et al., 2010. Improving thermostability of papain through structure-based protein engineering. Protein Engineering, Design and Selection, 23(6), pp.457–467.

Clausen, T. et al., 2011. HTRA proteases: regulated proteolysis in protein quality control. Nature Reviews Molecular Cell Biology, 12(3), pp.152–162.

Craik, C.S., Page, M.J. & Madison, E.L., 2011. Proteases as therapeutics. The Biochemical journal, 435(1), pp.1–16.

Daugherty, P.S., 2007. Protein engineering with bacterial display. Current Opinion in Structural Biology, 17(4), pp.474–480.

duVerle, D.A. & Mamitsuka, H., 2012. A review of statistical methods for prediction of proteolytic cleavage. Briefings in Bioinformatics, 13(3), pp.337–349.

Fleishman, S.J. et al., 2011. RosettaScripts: A Scripting Language Interface to the Rosetta Macromolecular Modeling Suite V. N. Uversky, ed. PLoS ONE, 6(6), p.e20161.

Flowers, C.A. & Ann, C., 2013. Engineering and analysis of protease fine specificity via site-directed mutagenesis. University of Texas

Gai, S.A. & Wittrup, K.D., 2007. Yeast surface display for protein engineering and characterization. Current opinion in structural biology, 17(4), pp.467–73.

Garay-Malpartida, H.M. et al., 2005. CaSPredictor: a new computer-based tool for caspase substrate prediction. Bioinformatics, 21(Suppl 1), pp.i169–i176.

Getz, J.A., Schoep, T.D. & Daugherty, P.S., 2012. Peptide Discovery Using Bacterial Display and Flow Cytometry. In pp. 75–97.

Gosalia, D.N. et al., 2005. Profiling serine protease substrate specificity with solution phase fluorogenic peptide microarrays. PROTEOMICS, 5(5), pp.1292–1298.

Gosalia, D.N. & Diamond, S.L., 2003. Printing chemical libraries on microarrays for fluid phase nanoliter reactions. Proceedings of the National Academy of Sciences, 100(15), pp.8721–8726.

Guerrero, J.L., O'Malley, M.A. & Daugherty, P.S., 2016. Intracellular FRET-based Screen for Redesigning the Specificity of Secreted Proteases. ACS Chemical Biology, 11(4), pp.961–970.

Hand, D.J. & Till, R.J., 2001. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. Machine Learning, 45(2), pp.171–186.

Harms, M.J. & Thornton, J.W., 2013. Evolutionary biochemistry: revealing the historical and physical causes of protein properties. Nature Reviews Genetics, 14(8), pp.559–571.

Harris, J.L. et al., 1998. Definition and redesign of the extended substrate specificity of granzyme B. The Journal of biological chemistry, 273(42), pp.27364–73.

Hashimoto, H. et al., 2011. Structural basis for matrix metalloproteinase-2 (MMP-2)-selective inhibitory action of β-amyloid precursor protein-derived inhibitor. The Journal of biological chemistry, 286(38), pp.33236–43.

He, M. & Taussig, M.J., 2002. Ribosome display: cell-free protein display technology. Briefings in functional genomics & proteomics, 1(2), pp.204–12.

Head, S.R. et al., 2014. Library construction for next-generation sequencing: overviews and challenges. BioTechniques, 56(2), p.61–4, 66, 68, passim.

Hedstrom, L., 2002. Serine Protease Mechanism and Specificity. Chemical Reviews, 102(12), pp.4501–4524.

von Heijne, G., 1986. A new method for predicting signal sequence cleavage sites. Nucleic acids research, 14(11), pp.4683–90.

Henikoff, S. & Henikoff, J.G., 1994. Position-based sequence weights. Journal of molecular biology, 243(4), pp.574–8.

Hill, M.E. et al., 2016. Reprogramming Caspase-7 Specificity by Regio-Specific Mutations and Selection Provides Alternate Solutions for Substrate Recognition. ACS Chemical Biology, 11(6), pp.1603–1612.

Ho, M. & Pastan, I., 2009. Mammalian Cell Display for Antibody Engineering. In Methods in molecular biology (Clifton, N.J.). pp. 337–352.

Jezierska, A. & Motyl, T., 2009. Matrix metalloproteinase-2 involvement in breast cancer progression: a mini-review. Medical science monitor : international medical journal of experimental and clinical research, 15(2), p.RA32-40.

Kasserra, H.P. & Laidler, K.J., 1969. Mechanisms of action of trypsin and chymotrypsin. Canadian Journal of Chemistry, 47.

Kenrick, S.A. & Daugherty, P.S., 2010. Bacterial display enables efficient and quantitative peptide affinity maturation. Protein Engineering Design and Selection, 23(1), pp.9–17.

Khan, A.R. & James, M.N., 1998. Molecular mechanisms for the conversion of zymogens to active proteolytic enzymes. Protein science : a publication of the Protein Society, 7(4), pp.815–36.

Khouri, H.E. et al., 1991. Engineering of papain: selective alteration of substrate specificity by site-directed mutagenesis. Biochemistry, 30(37), pp.8929–8936.

Leaver-Fay, A. et al., 2011. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods in enzymology, 487, pp.545–74.

Li, Q. et al., 2017. Profiling Protease Specificity: Combining Yeast ER Sequestration Screening (YESS) with Next Generation Sequencing. ACS Chem Biol., 12(2), pp. 510-518

Lin, C., 2006. HCV NS3-4A Serine Protease, Tan SL. Horizon Bioscience. Hepatitis C Viruses: Genome and Molecular Biology

Liu, R. et al., 2000. Optimized synthesis of RNA-protein fusions for in vitro protein selection. Methods Enzymol., 318. pp. 268–293.

Matthews, D.J. et al., 1994. A survey of furin substrate specificity using substrate phage display. Protein science : a publication of the Protein Society, 3(8), pp.1197–205.

Maximova, T. et al., 2016. Principles and Overview of Sampling Methods for Modeling Macromolecular Structure and Dynamics B. L. de Groot, ed. PLOS Computational Biology, 12(4), p.e1004619.

Miton, C.M. & Tokuriki, N., 2016. How mutational epistasis impairs predictability in protein evolution and design. Protein science : a publication of the Protein Society, 25(7), pp.1260–72.

Nemova, N.N. & Lysenko, L.A., 2013. Biological significance of protease diversity. Paleontological Journal, 47(9), pp.1085–1088.

Park, S. et al., 2006. Limitations of yeast surface display in engineering proteins of high thermostability. Protein Engineering Design and Selection, 19(5), pp.211–217.

Phan, J. et al., 2002. Structural basis for the substrate specificity of tobacco etch virus protease. The Journal of biological chemistry, 277(52), pp.50564–72.

Poorman, R.A. et al., 1991. A cumulative specificity model for proteases from human immunodeficiency virus types 1 and 2, inferred from statistical analysis of an extended substrate data base. The Journal of biological chemistry, 266(22), pp.14554–61.

Pop, C. & Salvesen, G.S., 2009. Human Caspases: Activation, Specificity, and Regulation *. J Biol Chem. 284(33), pp. 21777-81.

Prabu-Jeyabalan, M., Nalivaika, E. & Schiffer, C.A., 2002. Substrate shape determines specificity of recognition for HIV-1 protease: analysis of crystal structures of six substrate complexes. Structure (London, England : 1993), 10(3), pp.369–81.

Punta, M. et al., 2008. The Rough Guide to In Silico Function Prediction, or How To Use Sequence and Structure Information To Predict Protein Function F. Lewitter, ed. PLoS Computational Biology, 4(10), p.e1000160.

Qian, N. & Sejnowski, T.J., 1988. Predicting the secondary structure of globular proteins using neural network models. Journal of molecular biology, 202(4), pp.865–84.

Ratnikov, B., Cieplak, P. & Smith, J.W., 2009. High throughput substrate phage display for protease profiling. Methods in molecular biology (Clifton, N.J.), 539, pp.93–114.

Rawlings, N.D., Barrett, A.J. & Bateman, A., 2010. MEROPS: the peptidase database. Nucleic acids research, 38(Database issue), pp.D227-33.

Ribeiro, R.M. et al., 2012. Quantifying the Diversification of Hepatitis C Virus (HCV) during Primary Infection: Estimates of the In Vivo Mutation Rate C. O. Wilke, ed. PLoS Pathogens, 8(8), p.e1002881.

Riedl, S.J. & Shi, Y., 2004. Molecular mechanisms of caspase regulation during apoptosis. Nature Reviews Molecular Cell Biology, 5(11), pp.897–907.

Rockwell, N.C. et al., 2002. Precursor processing by kex2/furin proteases. Chemical reviews, 102(12), pp.4525–48.

Romano, K.P. et al., 2010. Drug resistance against HCV NS3/4A inhibitors is defined by the balance of substrate recognition versus inhibitor binding. Proceedings of the National Academy of Sciences of the United States of America, 107(49), pp.20986–91.

Sailer, Z.R. et al., 2017. High-order epistasis shapes evolutionary trajectories J. Krug, ed. PLOS Computational Biology, 13(5), p.e1005541.

Sailer, Z.R. & Harms, M.J., 2017. Detecting High-Order Epistasis in Nonlinear Genotype-Phenotype Maps. Genetics. Available from: https://doi.org/10.1534/genetics.116.195214

Sakai, K. et al., 1987. A unique specificity of a calcium activated neutral protease indicated in histone hydrolysis. Journal of biochemistry, 101(4), pp.911–8.

Salisbury, C.M., Maly, D.J. & Ellman, J.A., 2002. Peptide Microarrays for the Determination of Protease Substrate Specificity. Journal of the American Chemical Society, 124(50), pp.14868–14870.

Schechter, I. & Berger, A., 1967. On the size of the active site in proteases. I. Papain. Biochemical and Biophysical Research Communications, 27(2), pp.157–162.

Schneider, E.L. & Craik, C.S., 2009. Positional Scanning Synthetic Combinatorial Libraries for Substrate Profiling. In Methods in molecular biology (Clifton, N.J.). pp. 59–

78.

Sewall Wright, 1932. The roles of mutation, inbreeding, crossbreeding, and selection in evolution. Proceedings of the Sixth International Congress on Genetics, pp.355–366.

Shen, Y. et al., Testing the Substrate-Envelope Hypothesis with Designed Pairs of Compounds. ACS Chem Biol, 8(11), pp. 2433-2441

Smith, G., 1985. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. Science, 228(4705).

Song, J. et al., 2012. PROSPER: an integrated feature-based tool for predicting protease substrate cleavage sites. PloS one, 7(11), p.e50300.

Svensson, E.I. & Calsbeek, R., 2012. The adaptive landscape in evolutionary biology, Oxford University Press.

Taguchi, S., Ozaki, A. & Momose, H., 1998. Engineering of a Cold-Adapted Protease by Sequential Random Mutagenesis and a Screening System. Applied and Environmental Microbiology, 64(2), pp.492–495.

Tarca, A.L. et al., 2007. Machine Learning and Its Applications to Biology. PLoS Computational Biology, 3(6), p.e116.

Traxlmayr, M.W. & Shusta, E. V., 2017. Directed Evolution of Protein Thermal Stability Using Yeast Surface Display. In Methods in molecular biology (Clifton, N.J.). pp. 45–65.

Tyndall, J.D.A., Nall, T. & Fairlie, D.P., 2005. Proteases universally recognize beta strands in their active sites. Chemical reviews, 105(3), pp.973–99.

Varadarajan, N. et al., 2005. Engineering of protease variants exhibiting high catalytic activity and exquisite substrate selectivity. Proceedings of the National Academy of Sciences of the United States of America, 102(19), pp.6855–60.

Walker, J.A. et al., 1994. Sequence specificity of furin, a proprotein-processing endoprotease, for the hemagglutinin of a virulent avian influenza virus. Journal of virology, 68(2), pp.1213–8.

Wang, H. et al., 2016. Engineering of a Bacillus amyloliquefaciens Strain with High Neutral Protease Producing Capacity and Optimization of Its Fermentation Conditions S. Yang, ed. PLOS ONE, 11(1), p.e0146373.

Waugh, D.S., 2011. An overview of enzymatic reagents for the removal of affinity tags. Protein expression and purification, 80(2), pp.283–93.

Weinreich, D.M., Watson, R.A. & Chao, L., 2005. Perspective: Sign Epistasis and

henetic constraint on evolutionary trajectories. Evolution, 59(6), p.1165.

Wilkins, M.R. et al., 1999. Protein identification and analysis tools in the ExPASy server. Methods in molecular biology (Clifton, N.J.), 112, pp.531–52.

Yang, Z.R. & Chou, K.-C., 2004. Bio-support vector machines for computational proteomics. Bioinformatics, 20(5), pp.735–741.

Yi, L. et al., 2013. Engineering of TEV protease variants by yeast ER sequestration screening (YESS) of combinatorial libraries. Proceedings of the National Academy of Sciences of the United States of America, 110(18), pp.7229–34.

Yi, L. et al., 2015. Yeast Endoplasmic Reticulum Sequestration Screening for the Engineering of Proteases from Libraries Expressed in Yeast. Methods in molecular biology (Clifton, N.J.), 1319, pp.81–93.

Yost, S.A. & Marcotrigiano, J., 2013. Viral precursor polyproteins: keys of regulation from replication to maturation. Current opinion in virology, 3(2), pp.137–42.

Zahnd, C., Amstutz, P. & Plückthun, A., 2007. Ribosome display: selecting and evolving proteins in vitro that specifically bind to a target. Nature Methods, 4(3), pp.269–279.

Zhou, C. et al., 2010. Development of a novel mammalian cell surface antibody display platform. mAbs, 2(5), pp.508–18.

Chapter 2: Large-scale structure-based prediction and identification of novel protease substrates using computational protein design

2.1. Abstract

Characterizing the substrate specificity of protease enzymes is critical for illuminating the molecular basis of their diverse and complex roles in a wide array of biological processes. Rapid and accurate prediction of their extended substrate specificity would also aid in the design of custom proteases capable of selectively and controllably cleaving biotechnologically or therapeutically relevant targets. However, current in silico approaches for protease specificity prediction, rely on, and are therefore limited by, machine learning of sequence patterns in known experimental data. Here, we describe a general approach for predicting peptidase substrates de novo using protein structure modeling and biophysical evaluation of enzyme-substrate complexes. We construct atomic resolution models of thousands of candidate substrate-enzyme complexes for each of five model proteases belonging to the four major protease mechanistic classes – serine-, cysteine-, aspartyl- and metallo-proteases, and develop a discriminatory scoring function using enzyme design modules from Rosetta and Amber-MMPBSA. We rank putative substrates based on calculated interaction energy with a modeled near-attack conformation of the enzyme active site. We show that the energetic patterns obtained from these simulations can be used to robustly rank and classify known cleaved and uncleaved peptides and that these structural-energetic patterns have greater discriminatory power compared to purely sequence-based statistical inference. Combining sequence and energetic patterns using machine-learning algorithms further

improves classification performance, and analysis of structural models provides physical insight into the structural basis for the observed specificities. We further tested the predictive capability of the model by designing and experimentally characterizing the cleavage of four novel substrate motifs for the Hepatitis C virus NS3/4 protease using an in vivo assay. The presented structure-based approach is generalizable to other protease enzymes with known or modeled structures, and complements existing experimental methods for specificity determination.

2.2. Introduction

Proteolytic cleavage is a ubiquitous post-translational modification that controls the transmission of biological information(López-Otín & Bond 2008; Hedstrom 2002a; Hedstrom 2002b). Proteases encompass a structurally and mechanistically diverse class of enzymes that display a range of cleavage specificities reflecting their complex and diverse biological roles (Hedstrom 2002b; Tyndall et al. 2005; Powers et al. 1993; Rawlings & Salvesen 2013). For example, proteases involved in digestion and extracellular matrix degradation, e.g. trypsins and matrix metalloproteases, respectively, show relatively relaxed specificity profiles(Rawlings et al. 2010), whereas those involved in apoptotic and thrombolytic cascades, e.g. caspases(Julien et al. 2016) and thrombin(Di Cera & Cantwell 2001), respectively, are more selective in their cleavage motifs. In many viruses, protease-mediated cleavage of the viral polyprotein at specific sites is crucial for viral maturation(Scheel & Rice 2013); as a result, these enzymes are highly selective in cleaving only a small set of polypeptide sequences, while not acting on other sequences in the polyprotein. Accordingly, these enzymes have been successful drug targets for

developing anti-viral therapies (Drag & Salvesen 2010; Eder et al. 2007). Thus, proteases are exemplars of enzymatic multispecificity, which have likely evolved to act upon and cleave a range of substrates – their specificity profile – while simultaneously avoiding the cleavage of other substrates (Tawfik 2014). Modeling of protease substrate specificity would illuminate the structural and physiochemical basis of these observed positive and negative selectivities, and aid protease biology by identifying novel substrates and biological roles of proteolysis.

Experimental methods to characterize protease specificity (Poreba & Drag 2010) range from low-throughput methods in which individual peptides or mixtures of peptides are assayed for cleavage(Turk et al. 2001; Backes et al. 2000; Fretwell et al. 2008) to high-throughput methods that allow identification of substrates on a proteome-wide scale (van den Berg & Tholey 2012; Ratnikov et al. 2009; Agard et al. 2012; Julien et al. 2016; Vizovišek et al. 2016). However, substrate sequence space is large and different proteome-wide datasets often have little overlap, suggesting that a large number of substrate sequences remain to be identified. Moreover, each experiment is limited to a single enzyme variant (typically the wild type). Computational approaches could, in principle, enable more rapid construction of specificity profiles, especially for naturally occurring or drug-resistant protease variants, and/or assist in library design for experimental specificity determination in a specific region of sequence space. Pattern recognition-based approaches have been used to predict substrate sequence preferences for various proteases based on machine learning from available experimental data (Barkan et al. 2010; Boyd et al. 2004; Song et al. 2011; Song et al. 2012; Verspurten et

al. 2009; Li et al. 2012). However, these sequence-only approaches are constrained by the quality of the input data, and cannot be generalized to other proteases, or to variants of the same protease enzyme.

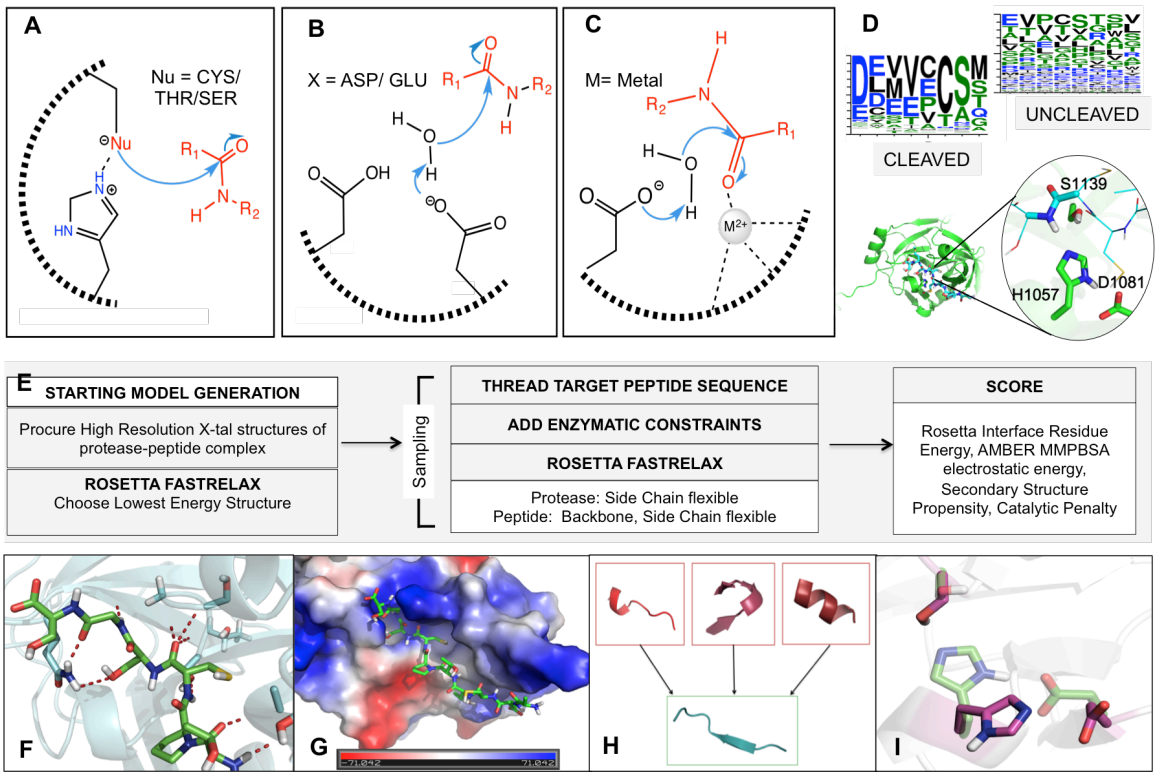


Figure 1.1. Overview of a general, energy-based discriminator
An illustration of the mechanism of steps leading to the formation of a common tetrahedral intermediate (TI) for serine-, cysteine-, threonine (A), aspartic, glutamic (B), and metallo-proteases (C). Protease active site cleft is depicted as a dashed arc. (D) Generation of atomic resolution models of the near attack conformation using high - resolution crystallographic structures and known cleaved and uncleaved sequence datasets. (E) The resulting complexes were allowed to relax into a minimum energy conformation using the described protocol (FastRelax) and scored using a linear combination of (F) the sum of the interface residues' Rosetta energy, (G) the sum of the interface residues' AMBER MMPBSA electrostatic scores, (H) a score that describes the propensity of the peptide to adopt an extended conformation (reorganization penalty), and (I) the deviation of the active cleft residues from the idealized active conformations (a pseudo score-term). The linear combination of weighted scores were recombined according to this equation: Total_score = w1*Rosetta_Interface_Energy(Protease energy) + w2*Rosetta_Interface_Energy (Peptide energy) + w3* Catalytic constraint penalty + w4 *Reorganization Penalty + w5* Electrostatic Binding Energy; where w1 =1, w2 =1, w3 = 3.5, w4 = 0.01 , w5 = 0.5

Proteolysis is a multi-step reaction involving the binding of the substrate and subsequent nucleophilic attack on the carbonyl group carbon of the scissile peptide bond to yield a tetrahedral intermediate (TI; Figure 1.1A-C) (Hedstrom 2002a). Steps after TI formation are mechanism-dependent: in cysteine, serine (and threonine) proteases, the intermediate disproportionates to yield one product and the reaction proceeds via the formation of an enzyme-bound intermediate that is deacylated to yield the second product (Figure 1.1A). In aspartic (and glutamic), and metallo-proteases, which use a hydroxide nucleophile generated from a bound water molecule, the tetrahedral intermediate directly disproportionates into both products (Figure 1.1B,C). In principle, different steps could determine substrate specificity depending on the substrate and the mechanism under consideration. However, for all proteases, regardless of the mechanistic class they belong to, the first step, i.e., enzyme nucleophilic attack is required for turnover (Hedstrom 2002a). This observation led us to hypothesize that a model of the enzyme with the bound substrate and catalytic machinery modeled in a near-nucleophilic attack conformation would enable us to capture the energetics involved in substrate recognition and specificity.

Here, we develop a predictive biophysical model aimed at uncovering the underlying rules that govern protease-peptide molecular recognition and test its ability to classify known protease substrates from uncleaved ones. We construct a discriminative scoring function that includes descriptors of the energetics (including long-range electrostatic interactions) at the interface of the protease–peptide complex, the geometric compatibility of the substrate with the catalytically active state of the protease, and the

reorganization penalty of a given substrate to adopt a favorable conformation in the protease active site (Tyndall et al. 2005). We demonstrate the predictive capacity of this discriminator by the recapitulation of known cleavage specificities of five experimentally characterized proteases representing all the major mechanistic protease classes (Powers et al. 1993) (serine, cysteine, aspartic, and metallo- proteases). We demonstrate an application of our biophysical discriminator by exploring previously uncharacterized, novel sequence motifs cleaved by the HCV NS3/4A protease via a yeast surface display-based assay (Yi et al. 2013) to identify novel cleaved sequences. Our biophysical structure-based model should allow the prediction of substrate specificities of experimentally uncharacterized proteases as well as protease variants (e.g. drug-resistant variants) and enable the structure-based design of proteases targeted to novel substrates.

2.3. Results

2.3.1. Rationale for the curation of Benchmark Datasets:

To develop and test a general structure- and energy-based prediction approach for protease specificity, we curated benchmark sequence sets for five diverse proteases. Each of these exhibit diverse mechanisms of action, varied folds and biological functions – TEV Protease (cysteine proteases), HCV NS3/4A protease (serine proteases), Granzyme B (serine protease), HIV Protease-1 (aspartyl protease) and Matrix Metalloprotease -2 (Metalloprotease). The sequence sets were composed of cleaved and uncleaved sequences identified in experiments or generated by examining naturally occurring targets (and non-targets) of each protease (see Materials and Methods). We preferentially chose datasets in which cleaved and uncleaved sequences were identified in the same

experiment. For HCV NS3/4A protease, HIV Protease 1 and Granzyme B, we were able to identify experiment-derived datasets (Shiryaev et al. 2012; Rögnvaldsson et al. 2009a; Barkan et al. 2010). For TEV protease and MMP2 protease, we were able to obtain experimentally cleaved datasets(Kostallas et al. 2011; Boulware et al. 2010; Ratnikov et al. 2014) but uncleaved sequences were not available. Therefore, we generated a synthetic dataset of uncleaved sequences using a two-residue protein walk approach, utilized in previous computational and experimental work (Shiryaev et al. 2012; Barkan et al. 2010). It is possible that these synthetically generated uncleaved sequences may include a small number of cleaved sequences. However, experimental results from Shiryaev et al (Shiryaev et al. 2012) suggest that misclassification of uncleaved sequences obtained using this approach is low. Therefore, in the absence of a directly experimentally determined uncleaved dataset for TEV protease and MMP2, we utilized this previously validated approach for uncleaved dataset creation.

2.3.2. Developing an energetic discriminatory scoring function based on structural simulations:

We hypothesized that determinants of substrate cleavage include (a) protease-peptide interfacial interactions, (b) the adoption of a catalytically competent conformation of the protease active site machinery in the bound state (near-attack conformation), and (c) a reorganization penalty that captures the propensity of a given substrate to adopt the extended conformation required for positioning the scissile bond in a cleavage-prone location in the protease active site. We created atomic resolution models for each peptide-protease complex and computed each of these terms as described below.

To model the conformation of each substrate peptide complexed with the active conformation of the protease, we created atomic resolution models within the context of the Rosetta macromolecular modeling software. Each known peptide substrate was threaded on the respective modeled near-attack conformation generated from the protease crystal structures (Figure 1.1D), and the resulting complex was allowed to computationally relax into a local energy minimum using Rosetta FastRelax(Tyka et al. 2011), followed by scoring this modeled conformation using Rosetta and Amber's MMPBSA modules (Figure 1.1E).

In addition to the interaction energy evaluated using Rosetta (Figure 1.1F), which includes a model of electrostatics, (called fa_elec), we also evaluated binding electrostatics by using Amber's MMPBSA module (Figure 1.1G). We reasoned that the Rosetta energy function has been weight optimized for all of its component terms including fa_elec. Thus, we decided to include fa_elec even upon inclusion of the AMBER electrostatics score. We included two other terms in our discriminator scoring function: First, we included a term ("reorganization penalty") that captures the propensity of a given substrate to adopt the extended conformation observed in crystal structures of all proteases (Figure 1.1H). Second, the deviation of the active site from ideal catalytic geometry (a pseudo-energy term) upon energy minimization (Figure 1.1I), which captures the fit of a given substrate to the catalytically competent conformation of the protease, was included. These scores – energetic descriptors of the peptide-protease complex in a near-attack conformation – were combined using a linear weighting

approach to obtain a discriminatory score function such that lower scores are predicted to energetically fit better in the active site (Figure 1.1F-I).

2.3.3 Recapitulation of known protease specificity profiles:

Each predicted substrate-binding set for each protease consists of a large set of evaluated peptide sequences, atomic-resolution bound structures, and predicted binding energies of individual peptides to the near-attack state of the enzyme. We compared our predictions with experimentally determined specificity data from peptide library screening. Briefly, in these experiments, peptide (or peptide-cDNA fusion) libraries are generated and treated with protease of interest, cleaved and uncleaved populations of peptides are captured and identified using (deep) sequencing or mass spectrometry, and cleavage probability is assigned using Enrichment of a given peptide sequence in the cleaved population versus the uncleaved.

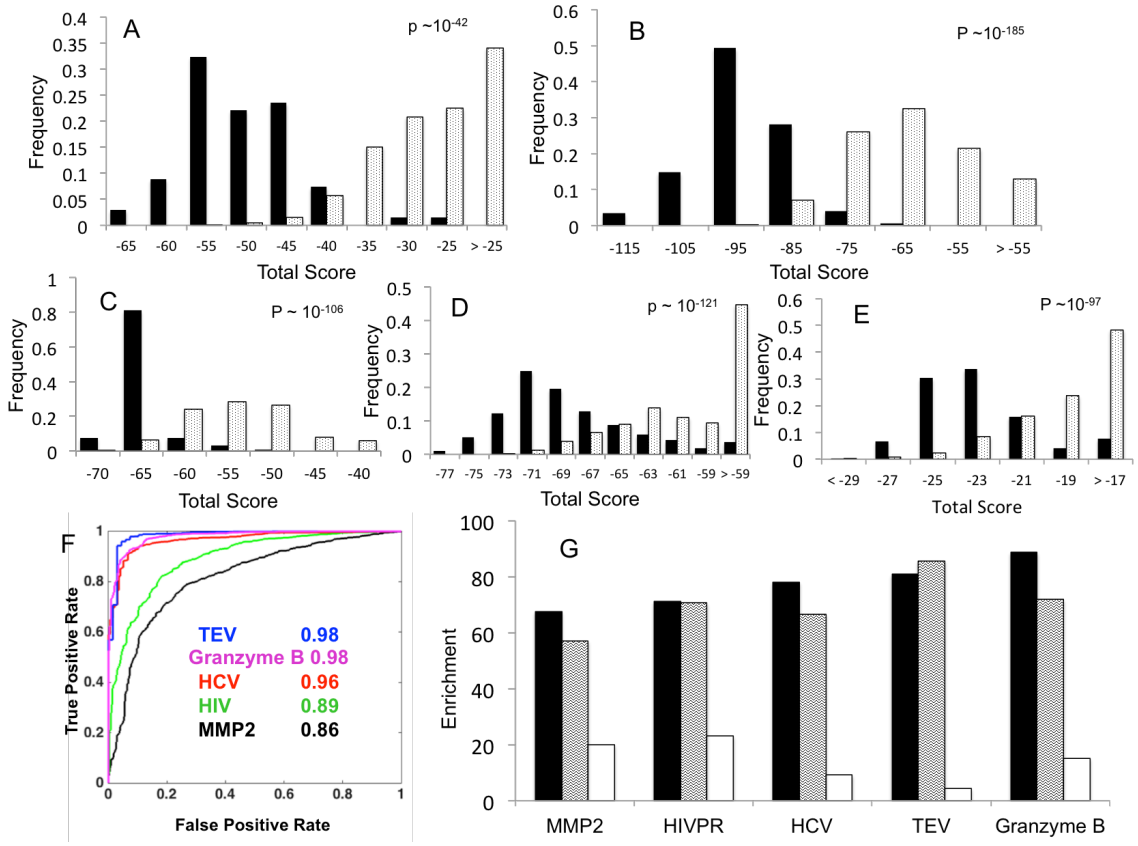


Figure 1.2. Distribution of Discriminator Scores.
Score distributions for cleaved sequences (depicted in black) and uncleaved (depicted in dotted bars) for (A) TEV protease (B) Granzyme B (C) HCV (D) HIV (E) MMP2. The p–values were calculated using a Wilcoxon rank test. A threshold based binary classification of sequences into cleaved and uncleaved sequences using these scores was performed and the auROC (F) for the five proteases are indicated. (G) Enrichment of true cleaved sequences in the top-ranked pools. Enrichment ratio (black bars) = #true cleaved/ # of cleaved sequences in datatset. Background Enrichment (white bars), which represents fraction of cleaved sequences in the dataset, and Enrichment obtained from SitePrediction model (wavy bars) with 20% of the known cleaved sequences. In each case, the structure-based discriminator performs comparably to or better than SitePrediction.

We found that for each of the five proteases, the distribution of discriminator scores was bimodal and cleaved and uncleaved sequences were separated in a statistically significant manner (p-values calculated using the Wilcoxon rank test; Figure 1.2A-E). To quantify the performance of the discriminator in the task of separating cleaved from uncleaved

substrates, we performed a score threshold-based binary classification of the sequences into cleaved and uncleaved sets and calculated the area under the resulting receiver-operator curve (auROC; perfect discrimination would yield an auROC of 1.0; the expected auROC for a random ordering of the peptides is 0.5). The auROC values for the five proteases ranged between (0.86 for MMP-2 to 0.98 for TEV-PR), demonstrating robust discrimination using energetics (Figure 1.2G). The critical point of the auROC plot represents the optimal tradeoff between false positive and false negative rates. We found that false positive rates at critical points ranged from 0.04 (TEV-PR) to 0.24 (MMP-2), suggesting robust discrimination of the substrates into cleaved and uncleaved sets with a small but significant false positive rate (Table 1.1). We note that weights used for combining the five score terms were initially optimized to maximize discrimination for HCV NS3/4A protease (five weight terms over approximately 2100 data points), yet TEV-PR displays the best performance in terms of both auROC and critical point values using this weight set. These results demonstrate the generality and robustness of the energy-based scoring function.

Protease	TPR	FPR
HCV	0.92	0.08
TEV	0.96	0.04
HIV	0.82	0.18
Granzyme B	0.93	0.07
MMP2	0.76	0.24

Table 1.1: True Positive and False Positive Rates observed for critical point of auROC

To evaluate the ability of the discriminator to identify cleaved sequences from the entire pool of sequences – a task that would aid in novel substrate identification – we calculated

the fraction of truly cleaved sequences in the top-scoring $N_{cleaved}$ sequences, where $N_{cleaved}$ is the number of cleaved sequences in the dataset. This Enrichment value is compared to background Enrichment, i.e. fraction of cleaved sequences in the dataset (reflecting a scenario when the ranking is performed by randomly shuffling the list of sequences). We find that in all cases a significantly higher fraction of sequences was enriched compared to the background with Enrichment ratios ranging from 3-fold (HIV-PR) to 19-fold (TEV-PR) (Figure 1.2F). We compared the Enrichment obtained using our discriminator with that obtained using SitePrediction (Verspurten et al. 2009)– a sequence-based machine learning method that relies on training with experimental data. For each protease, we trained a SitePrediction model with randomly chosen 20% of the known cleaved sequences and used the remaining dataset for testing. For all proteases, we find that our unbiased, biophysics-based approach yielded similar or higher Enrichment values as SitePrediction models trained separately on each individual protease. The lack of training on known experimental data makes the structure-based discriminator more widely applicable.

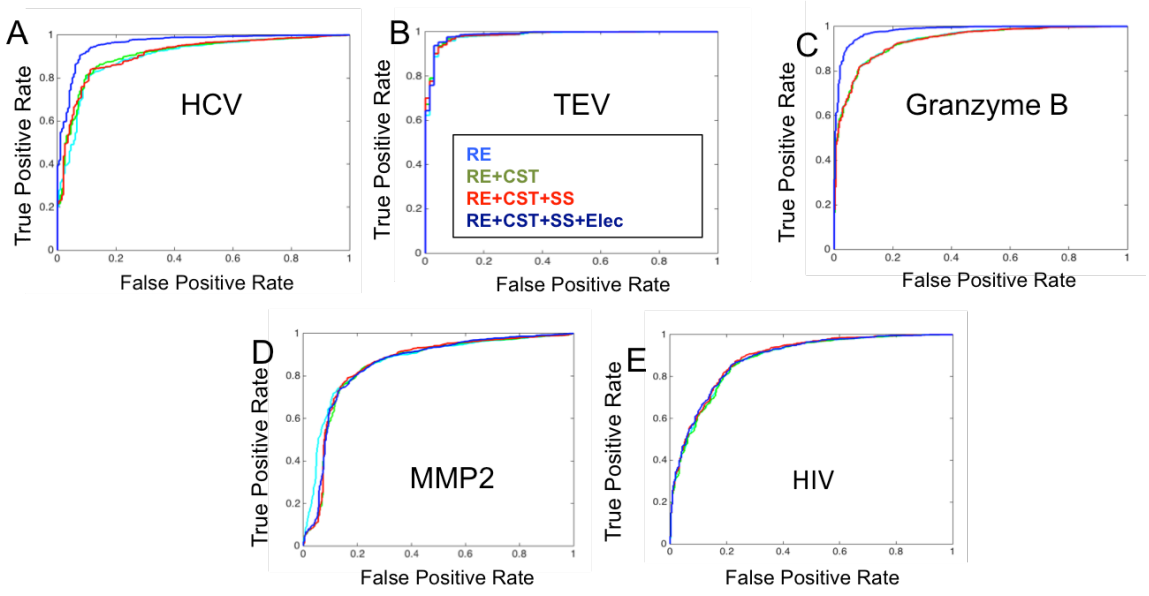


Figure 1.3. The additive effect of each energy term to the auROC.
Each plot shows the representative ROC curve for Rosetta Energy (sum total of peptide and protease interface energy; depicted in light blue), Rosetta Energy + constraint score (green), Rosetta Energy + constraint score + secondary structure propensity (red), Rosetta Energy + constraint score + secondary structure propensity + Electrostatic binding energy (dark blue). All score terms are seen to contribute to the discriminative efficiency of the score function.

2.3.4. Optimization of scoring and sampling strategies:

To investigate the contribution of each score term and its weight in the discriminator scoring function, we evaluated the discrimination performance of various score term combinations. We found that while the majority of the discriminatory power could be attributed to Rosetta interface residue energies, all five terms do contribute to the observed prediction metrics when they are serially included along with the Rosetta energy. While the increases in auROC compared to Rosetta energies-only scoring functions were modest, Enrichment values benefited significantly by the inclusion of the additional terms e.g., for Granzyme B inclusion of the AMBER electrostatics score and secondary structure propensity increases Enrichment from 0.70 to 0.87 (Figure 1.3, Table 1.2). As auROC measures the overall difference in the two distributions (cleaved and uncleaved) and Enrichment measures the rank ordering of sequences, we conclude that inclusion of additional terms serves to subtly alter the calculated energy landscape and "rescue" some false negatives (cleaved sequences that score comparably to low-energy uncleaved ones).

Protease		RE+CST	RE+CST+Elec	RE+CST+Elec+SS
Granzyme B	Enrichment	0.70	0.68	0.87
	Fold increase	4.6	4.5	5.7
	AUC	0.93	0.93	0.98
HCV	Enrichment	0.64	0.76	0.80
	Fold increase	6.2	7.3	7.6
	AUC	0.93	0.97	0.97
TEV	Enrichment	0.72	0.72	0.80
	Fold increase	16.68	16.68	18.35
	AUC	0.98	0.98	0.98
HIV	Enrichment	0.69	0.68	0.69
	Fold increase	3.2	3.2	3.2
	AUC	0.90	0.90	0.90

Table 1.2: Results of a calculation to investigate the additive effect of each score term in the discriminatory score function

We next investigated whether optimization of weights of the energetic scoring terms could improve performance. We used a grid-based optimization scheme in weight space to maximize Enrichment. While keeping Rosetta protease energy fixed, we optimized four free parameters by enumerating all combinations of peptide residue energy (0.3-1.3 in increments of 0.1, constraints (2.5-3.5 in increments of 0.1), secondary structure (0.005-0.02 in increments of 0.005), and electrostatics (0.1-0.3 in increments of 0.05). The ranges were chosen after a coarse-grained parameter sweep to find good starting parameters, and by considering the orders of magnitudes of raw scores of the score terms. For example, the raw score for the Secondary Structure Propensity term ranges between 0-200 (number of fragments from the top 200 that have an RMSD greater than 3.0 Å compared to the crystallographic conformation of the peptide). As the Rosetta

residue energy weight was 1, we explored weight ranges of 0.005-0.02 for the secondary structure term. The results of this optimization are listed in Table 1.3.

	Enrichment	protease	peptide	cst	ss	Elec
TEV	0.8088	1	0.3	2.5	0.005	0.25
HCV	0.7806	1	1	3.5	0.001	0.5
HIV	0.7112	1	0.8	3.4	0.013	0.1
GrB	0.8867	1	0.5	2.5	0.005	0.3
MMP2	0.6747	1	0.7	2.6	0.007	0.1

Table 1.3: Results of a grid-based optimization scheme to maximize enrichment

We next examined the impact of sampling flexibility of the backbone and side chain degrees of freedom (DOF) at the protease-peptide interface (Figure 1.4) and found that limiting the backbone degrees of freedom of the protease, while sampling the full backbone DOFs of the peptide, yielded the highest Enrichment values. Previous studies with farnesyltransferase enzyme similarly observed that greater sampling of the peptide degrees of freedom increased performance (London et al. 2011). When the protease backbone was allowed to move in an unconstrained manner, several uncleaved sequences adopted energetically favorable conformations. While some of these false positives can be attributed to limitations of the simulation force fields and sampling strategies, these results indicate that side chain flexibility in the protease pockets coupled to peptide backbone flexibility are key contributors to the molecular recognition observed at these interfaces.

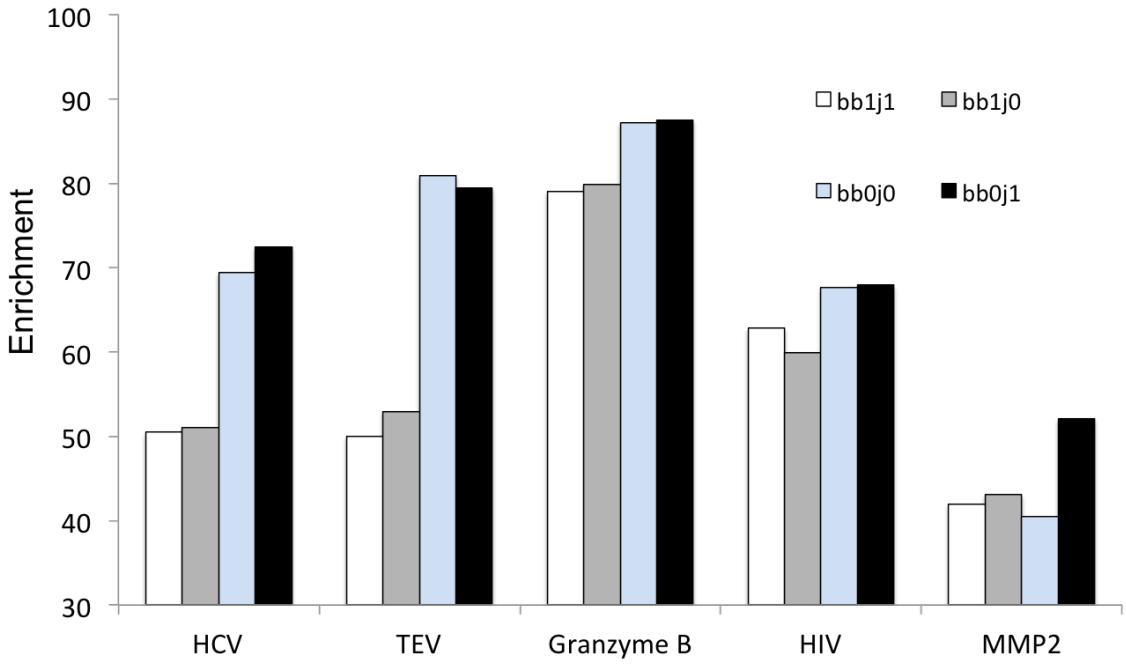


Figure 1.4. Impact of sampling flexibility of the protease backbone and sidechain degrees of freedom.
The peptide backbone and sidechains were flexible in all of the simulations depicted in the figure. "bb" refers to the backbone of the protease such that bb=0 indicates that the backbone was not allowed to relax, bb=1 that backbone was allowed to relax. "j" refers to the rigid body freedom of the peptide with respect to the protease. j=0 means that rigid body freedom was constrained during the simulation; j=1 rigid body flexibility allowed during simulation. The highest efficiency of discrimination was observed when the protease backbone was not allowed to relax, and the protease sidechains were flexible during the simulation.

Finally, we explored the contribution of maintaining, during each simulation, the scissile peptide bond in a near-attack conformation with respect to the protease catalytic machinery using geometric constraints, by performing simulations without these geometric constraints, and/or removing the constraint scores from the discriminator scoring function. In each case, a decrease in Enrichment was observed (Figure 1.5), providing further support for our rationale that specificity in protease-peptide molecular recognition is not simply a ground state binding phenomenon, but is contingent upon the

relative energetics of the near attack substrate conformation during the nucleophilic attack step.

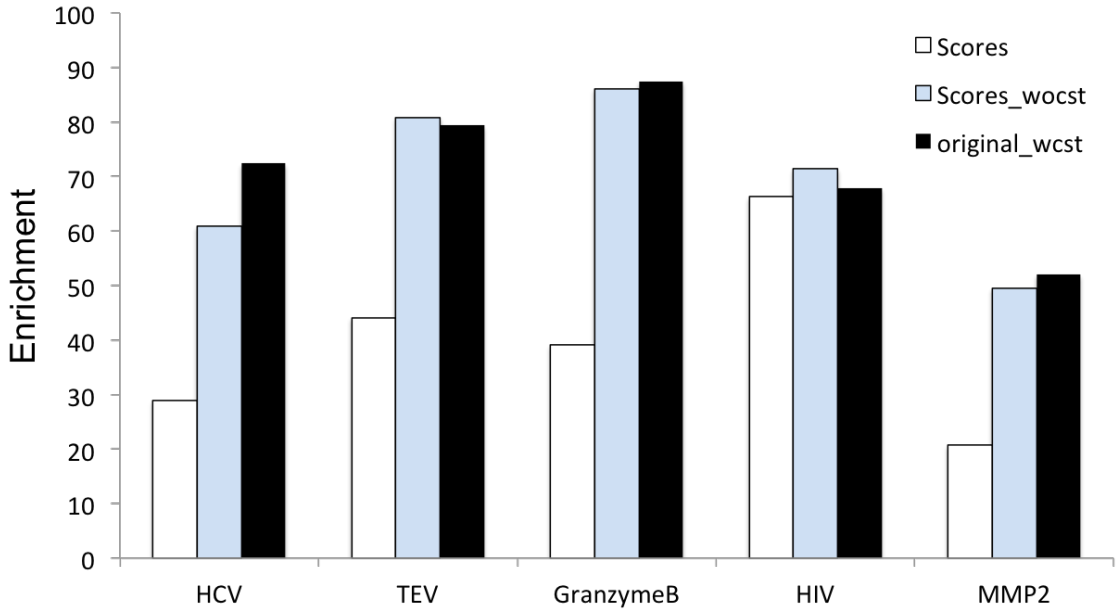


Figure 1.5. Contribution of maintaining near attack conformation with respect to protease catalytic machinery.
Three FastRelax protocols were performed to compare the effect of the presence of catalytic constraints during the FastRelax and scoring stage. Scores (white bars) depict enrichment values obtained when enzymatic constraints were excluded in the FastRelax step but were included in the scoring step. Scores_wocst (blue) depict experimental results where constraints were excluded from the FastRelax step as well as from the scoring calculation. Original_wcst (black) depict experimental results where FastRelax was performed with constraints and the constraint score was included in calculation of Enrichment. Highest enrichment is observed when catalytic constraints are included in both the FastRelax as well as scoring steps.

2.3.5. Combining sequence and energetic signatures using machine learning leads to higher discriminatory power

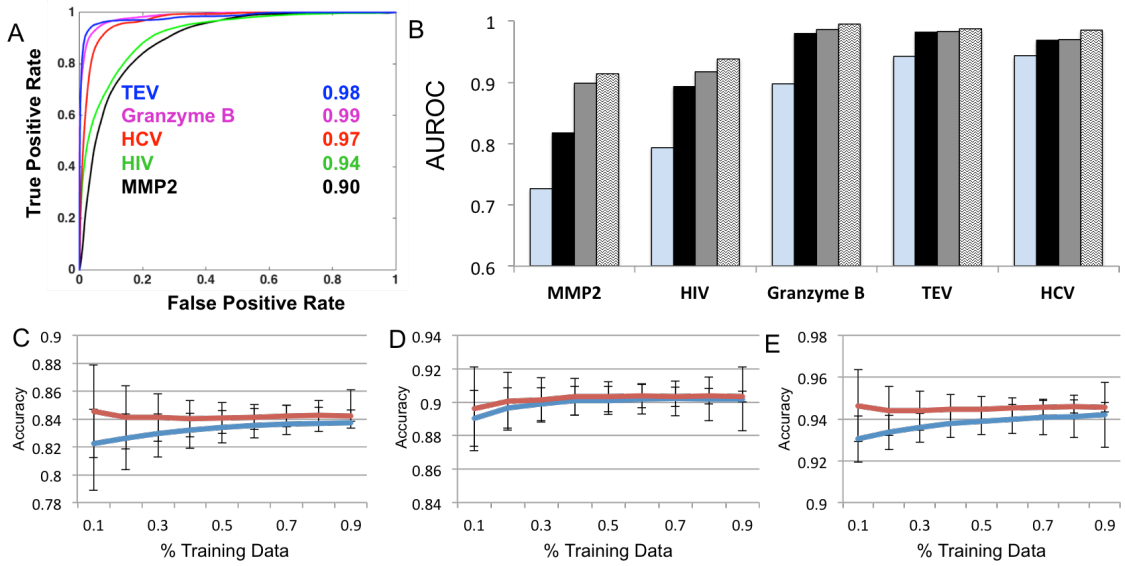


Figure 1.6. Combining sequence and energy signatures leads to higher discriminatory power
(A) The energetic features were used to train an SVM using a radial based function, which yielded higher auROC values for all proteases as compared to a linear combination of optimized weights. (B) auROCs obtained from support vector machines (SVMs) trained with sequence only(blue), energetic only(gray) and both sequence and energetic features(wavy) in a 5-fold cross-validation test. Black bars indicate auROC for the linear combination of weighted score terms. The combination of sequence and energy features consistently results in higher auROC values. (C) Accuracy as a function of training set size used for training for the (C) sequence, (D) energetic features, and (E) both sequence and energetic features for the HCV protease. The accuracy values are not altered appreciably when a significantly smaller training dataset is used. In-set classification and generalization curves converge as a progressively higher fraction of the dataset is used for training. The classification curve is shown in red whereas the generalization curve is depicted in blue.

Current approaches for protease specificity prediction, including the SitePrediction tool discussed above, PCSS server (Barkan et al. 2010) and PROSPER(Song et al. 2012), use machine learning of sequence patterns in known experimental data. To more extensively compare our structure-based specificity prediction with current sequence-based approaches we trained support vector machines (SVMs) with sequence-only, energetic-only and both sequence and energetic features (Methods). For the energy-based SVM, the (unweighted) energy terms described above were treated as features ("interface protease

residue energy", "interface peptide residue energy", "constraints energy", "reorganization penalty" and "MMPBSA electrostatic binding energy"), whereas sequence-based features were generated using a protocol described by Barkan et al.(Barkan et al. 2010). We found robust discrimination of the substrate sequences using energy-based SVMs trained individually on each protease in 5-fold cross-validation test (Figure 1.6A). The values of auROC obtained using these SVMs are higher than those obtained with scoring using a linear weighting scheme (Figure 1.6B, black and gray bars), due likely to the use of a non-linear kernel function and training on individual datasets. When compared to a purely sequence-based SVM, the energy-based SVM consistently leads to higher auROC values for all datasets, and an SVM constructed based on sequence and energy features displays a high AUC value when compared to solely sequence-based and energy-based based SVMs (Figure 1.6B). These results indicate that structural/energetic features contribute information that is orthogonal to that obtained from sequence-only features.

To ensure that the increased discriminatory ability observed upon combining sequence- and energy-based features is not a result of data over-fitting, we performed a cross-validation procedure where in-set training (classification) and out-of-set testing (generalization) was performed by randomly splitting the datasets into training and test subsets (Baugh et al. 2016) . We find that the performance of the method as indicated by the accuracy of prediction, does not appreciably alter when a significantly smaller training dataset is used for the energy-based SVMs, and the classification and generalization performance converge as the training set size increases (Figure 1.6C-E, Figure 1.7). The convergence between classification and generalization occurs at higher

training set fraction for the sequence-based SVMs than energy-based ones, demonstrating that the key energetic signatures underlying discrimination can be captured with a smaller dataset compared to the corresponding sequence signatures (Figure 1.7). Thus, energetic feature-based SVMs can outperform sequence-based ones, and the two sets of features can be combined to obtain more accurate classification than either set of features independently.

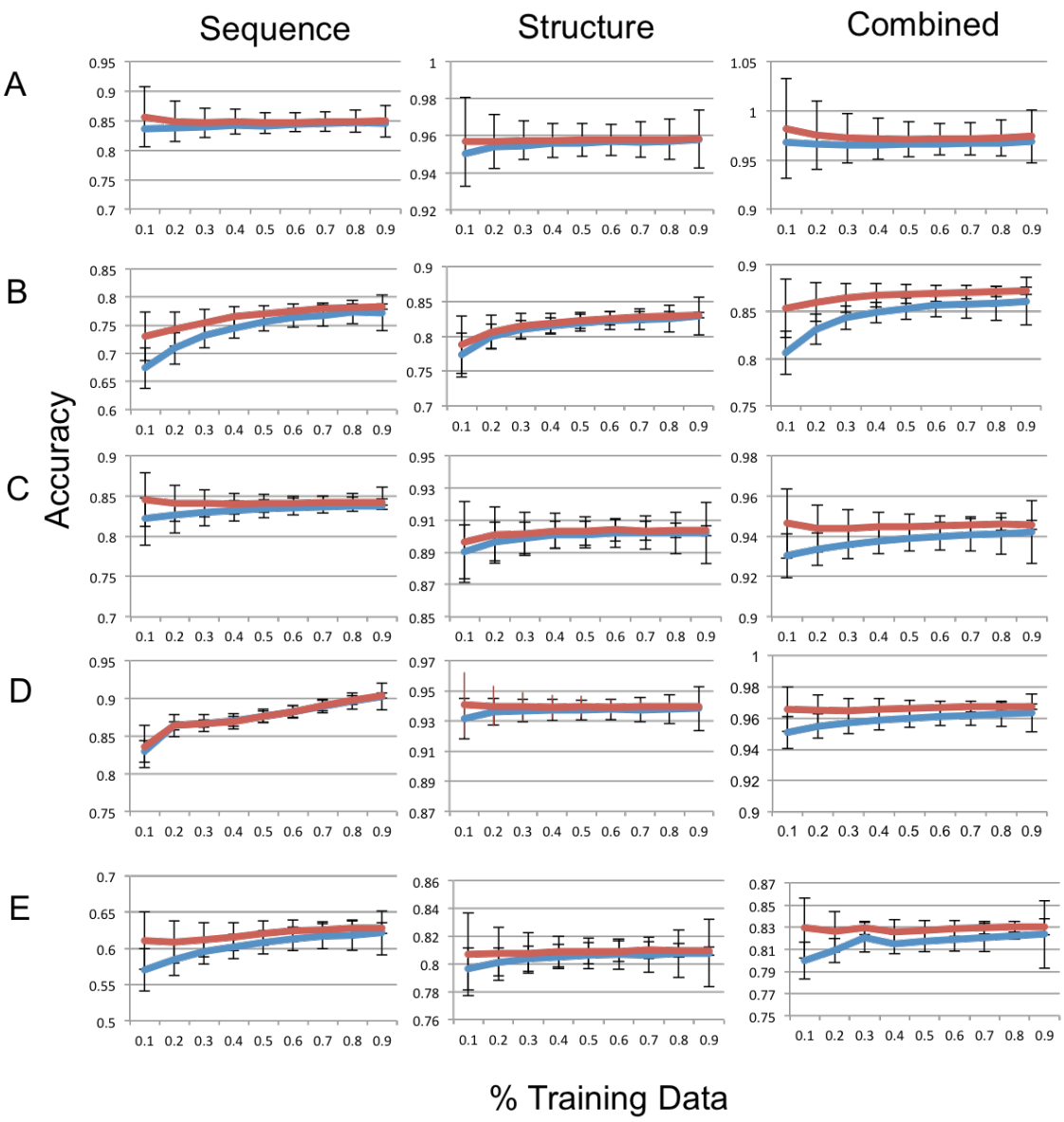


Figure 1.7. Accuracy versus Training Data size plots for Sequence, Structure and Combination SVMs.
To avoid over-fitting we performed a jack-knifing procedure where classification and generalization was performed by randomly splitting the datasets into training and test. (A) TEV (B) HIV (C) HCV (D) Granzyme B (E) MMP2

2.3.6. Multi-body interaction networks at the interface underlie improved discrimination

To investigate the underlying reasons for the observed increase in prediction efficiency when structural features are used, we identified several peptide sequences that are consistently misclassified by the sequence-based approach but are correctly classified by the structure-based approach. In several cases, we find that the increased classification ability could be attributed to interaction networks composed of multiple substrate and protease residues. A sequence-only approach would require a significantly larger training data than a relatively unbiased energy-based approach to directly "learn" multi-body correlations (interactions).

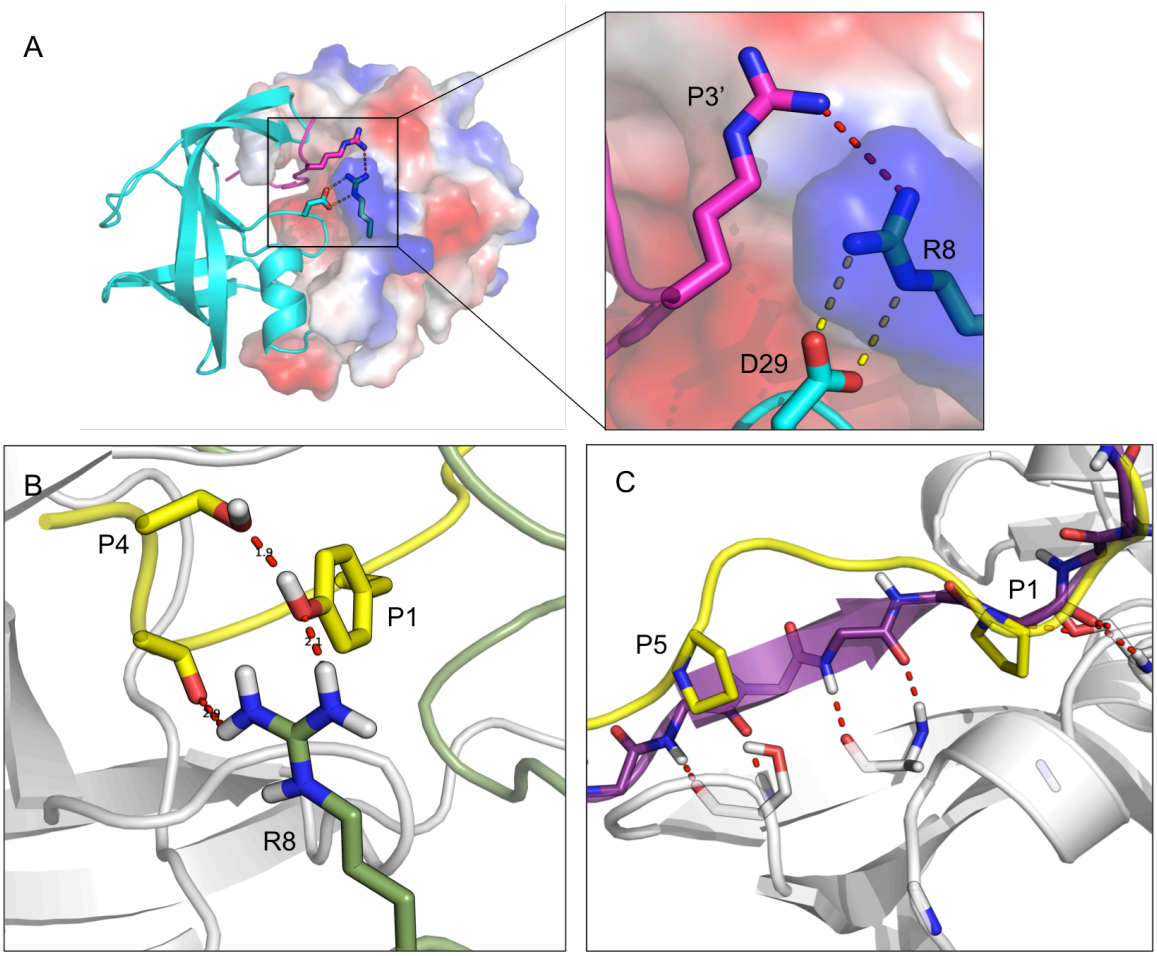


Figure 1.8. Multi-body interaction networks at the interface underlie improved discrimination.
Several sequences are misclassified by the Sequence-Based Discriminator whereas they are correctly classified when the Structure based Discriminator is used. (A) The sequence

‘GPGTARSP’ is misclassified by the sequence based SVM as ‘cleaved’ for the HIV protease sequence set. Residue P3’ of the peptide is packed in the vicinity of ARG 8; which is involved in a key interaction with ASP 29 necessary in maintaining the dimer interface. The P3’ – ARG 8 repulsion leads to a destruction of one of the key interactions involved in dimer interface stabilization. One half of dimer surface is shown as a cartoon representation and the other as a charged surface in order to highlight the dimer interface of the HIV protease. This electrostatic repulsion is captured by the energy-based approach but not the sequence based approach, leading to a misclassification by the latter (B) ‘SQAYPIVQ’ is misclassified as an uncleaved sequence present in the HIV protease sequence set. The P1 tyrosine residue (yellow) along with the serine at P4 forms a favorable hydrogen bond network with ARG 8 (green) allowing for substrate cleavage. This favorable hydrogen-bonding network is likely not directly recognized by the sequence-based approach. (C) ‘KPAIIPDR’ belongs to the HCV Protease sequence set which is misclassified as cleaved by the sequence-based approach. The presence of proline at positions P5 and P1 (yellow) bends the substrate chain in an orientation that is unfavorable for cleavage. The extended conformation of a peptide, which allows hydrogen bond formation, leading to binding of the peptide and eventually cleavage, is highlighted (purple). The Rosetta energies correctly detect this disruption of the hydrogen bond network caused by the presence of proline residues between peptide (yellow) and protease.

Three examples of these interaction networks are described below:

1. The structure-based discriminator can identify context-dependence of the substrate residue interactions more readily than a sequence-based approach, especially in cases where sequence preference at a given substrate site is not pronounced. For example, for the HIV protease, cleavage occurs between small non-polar amino acids and sequence preference at any other site is not particularly pronounced. As a result, GPGTASRP (Figure 1.8A) is misclassified as “cleaved” by the sequence-based SVM for HIV protease-1. There are no pronounced sequence preferences at position P3’ (Figure 1.10). The structural model of this sequence, however, shows that the guanidinium group of the arginine sidechain (P3’) is packed in the vicinity of R8, a key residue, whose interaction with D29 is critical for HIV protease structural (dimer) stability (Appadurai & Senapati

2016). Thus, the presence of an arginine at this P3' position would lead to lack of cleavage of the substrate, unless a secondary interaction relieves the electrostatic repulsion between the substrate arginine sidechain and the guanidinium group of R8. The subtle balance of these protease-substrate interactions can be captured by the electrostatics calculations in our approach.

2. The energy-based discriminator is able to detect hydrogen bond networks between substrate residues, including those mediated by the protease structure. For example, for the sequence SQAYPIVQ (Figure 1.8B), the sidechain of the tyrosine residue at position P1 forms a hydrogen bond network with the P4 position on the substrate and the R8 of the protease chain. This likely allows the protease to recognize and cleave this substrate.

3. Another set of interactions that our structural approach correctly characterizes are those mediated by proline and glycine residues, as these have specific backbone preferences that can affect the peptide backbone conformation. Figure 1.8C is an example of a sequence, KPAIIPDR, which is experimentally shown to be uncleaved by the HCV NS3/4A protease. The sequence-only approach misclassifies this sequence as cleaved, likely on account of the non-polar isoleucine residues at the P1, P1' residues. However, the proline residues present at P5 and P1 substrate positions bend the substrate backbone into a conformation that results in the disruption of the stabilizing backbone hydrogen bond network, which drives the extended substrate conformation optimal for cleavage. The Rosetta energy function detects the disruption of this backbone hydrogen bond

network, and thus the energy-based approach correctly classifies this sequence as 'uncleaved'.

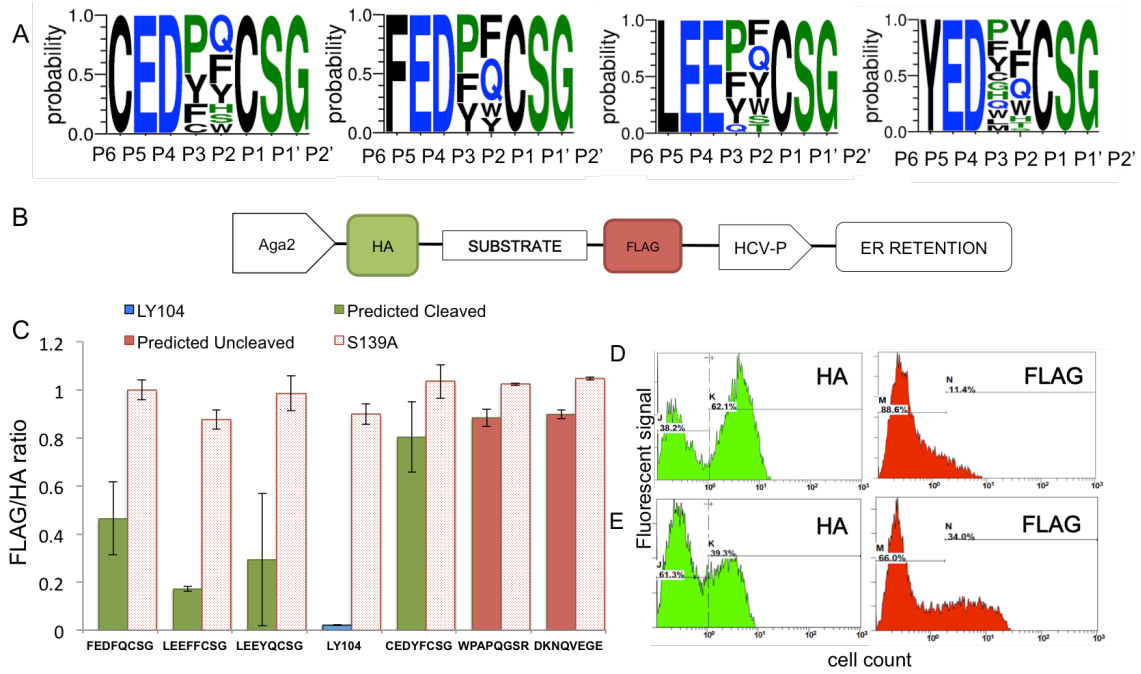


Figure 1.9. Discovering novel sequence specificities HCV/NS3 4A Protease
(A) Sequence Logo plots of the identified four novel sequence motifs whose scores overlapped with the cleaved sequences in the benchmark dataset (B) Schematic of the vector (LY104) used for the YESS assay. The vector contains Aga2 cell surface signaling moiety followed by the substrate flanked between HA tag and FLAG tag which can be detected on the cell surface by fluorescently tagged antibodies. The protease and substrate are co-expressed in the ER of the yeast cell. If cleavage occurs the FLAG:HA ratio is 0, if substrate is uncleaved ratio is 1. (C) Results of the YESS assay test of the predicted cleaved sequences. Three out of the four tested sequences (predicted cleaved; green bar) showed a FLAG:HA ratio <0.5. The positive control (wild type shown in blue) showed an expected low FLAG/HA ratio whereas the negative control (known and predicted uncleaved sequences, red bars) showed high FLAG:HA ratios >0.85. The protease activity knockout mutant S139A (dotted red bars) showed FLAG:HA ratio >0.85 for all sequences, confirming that the sequences were cleaved because of the co-expressed HCV NS3/4A protease from the assay vector and not an endogenous yeast ER enzyme. (D) Cell cytometry histograms of LEEFFCSG, predicted cleaved sequence showing a 62.1% cell population signal for HA tag, 11.4% cell population signal for FLAG, thus showing a FLAG:HA ratio of 0.18 (E) Cell cytometry histograms for the negative control sequence DKNQVEGE, showing a 38.3% cell population signal for HA tag, and 34.0% for FLAG tag, thus exhibiting a FLAG:HA ratio of 0.88.

2.3.7. Discovering novel sequence specificities HCV NS3/4A Protease

To further investigate the predictive ability of the energetic-discriminator in a blind test, we used our simulations to identify novel cleaved substrates for the HCV NS3/4 protease. The residue identities on the substrate peptide at positions P6 through P2 were sampled and scored as described in Methods using the structure-based discriminator. A total of 26,400 candidate sequences were evaluated (out of the possible 20^5 = 3.2 million) in a two-step procedure of sequence sampling as described in Methods, low-scoring sequences were clustered and were further pruned to identify sequence motifs that were novel (i.e., absent from the dataset used for developing the discriminator). We identified four such sequence motifs (Figure 1.9A), whose scores overlapped with the distribution of scores obtained from known cleaved sets. At least one peptide sequence was selected from three of the four identified motifs, and these were tested experimentally using a Yeast Endoplasmic Reticulum Sequestration Screen (YESS system) based assay (Yi et al. 2013; Yi et al. 2015) (Figure 1.9B).

In this assay, the protease and substrate are co-expressed in active forms in the ER of yeast, and the substrate is targeted to the cell surface by fusion to the cell surface protein Aga2p. Proteolysis is detected using fluorescent antibodies against the HA and FLAG tags that flank the substrate. We confirmed that the cleavage of the wild type substrate sequence (DEMEECA- canonical HCV NS3/4A cleavage sequence present between NS4A/4B on the polyprotein) results in the detachment of the FLAG tag from the AGA2

surface-signaling moiety, thus resulting in a FLAG:HA ratio of zero for complete cleavage and a ratio of one for no cleavage when an inactive variant of the protease (S139A) is used (Figure 1.9C). Several previous studies (Shiryaev et al. 2012; Grakoui, McCourt, et al. 1993; Grakoui, Wychowski, et al. 1993)have shown that the HCV protease cleaves between C/S or C/A residues (P1/P1') – however, the specificity at other positions can be broad and has not been explored fully. In all our predicted substrates (that we tested experimentally) the P1/P1' positions are still maintained as the known canonical sequence C/S, and our goal was prediction of different P6-P2 patterns. We, therefore, reasoned that the cleavage position of our substrates would not be altered as they retain the canonical P1/P1' cleavage pattern. The FLAG and HA signals were detected using flow cytometry. The observed FLAG/HA ratios (Figure 1.9 C, D) demonstrate that three out of four predicted sequences showed cleavage with ratios <0.5, whereas control assays with the S139A inactive protease variant showed significantly higher (>0.85) ratio, demonstrating that the observed cleavage is not due to a non-specific endogenous yeast enzyme.

Out of the four sequences that are predicted as cleaved, one sequence – CEDYFCSG – shows a high FLAG/HA ratio, and represents a prediction failure. These results are consistent with the ~75% True Positive and ~25% False Positive rates (Figure 1.2F) observed in the performance of the discriminator on known cleaved and uncleaved datasets, i.e., approximately one out of four sequences identified is expected to be a false positive sequence. We also identified two predicted uncleaved substrates, and these show lack of cleavage when co-expressed with either wild type protease or the inactive

protease variant, as expected. The FLAG:HA ratios for the novel identified substrates are higher than positive control LY104, indicating that the substrates identified are suboptimal. However, our test for novel substrates is particularly stringent as we chose sequence motifs that have previously not been identified in multiple studies of HCV NS3/4 protease. Thus, the developed discriminative score function and validating assay provide a method to screen for potential novel biological targets of this viral protease that is also a drug target.

2.4. Discussion

Proteolytic cleavage is a key component of diverse and ubiquitous biological processes such as apoptosis, blood clotting, viral maturation, and cancer(Puente et al. 2003). Developing a generalizable, predictive model for protease specificity would enable identification of potential novel substrates for furthering our understanding of protease biology and enhancing our ability to design inhibitor small molecules to chosen proteases. We developed a structure-based approach for specificity prediction using Rosetta and Amber force fields that provides atomic resolution insights into the molecular recognition at protease-substrate interfaces. We found that structural models robustly recapitulate known protease specificities for each of the four major protease classes (serine, cysteine, aspartic, and metallo-proteases) with little training on experimental data, and in several cross-validation tests. When combined with a machine learning algorithm our energy-based approach outperforms current bioinformatics-based approaches (Song et al. 2011) on benchmark sets, and a further increase in discrimination

is achieved when both structure-based and sequence-based approaches are combined. To further test the utility of our approach in a blind manner, we used it to predict four novel substrate sequences for HCV NS3/4A protease, tested these predictions experimentally, and found that three of the four novel predicted cleaved sequences were cleaved by the protease; a success rate similar to the benchmark set was achieved in the blind experimental test.

The value of using energetic information in the discriminator is evident in the protease structure-dependent interaction networks that are captured in the energetic signatures. These interaction networks are equivalent to pairwise and multi-body correlations in the sequence data. Given 20 amino acid types at every substrate peptide position, a relatively large number of training sequences are required to "learn" pairwise and higher-order correlations between positions, whereas only ~2000 sequences (among them ~200 cleaved) are available in the experimental benchmark datasets. The structure-guided, energy-based discriminator has the advantage of being generalizable, relatively unbiased and is able to recapitulate key interactions that stabilize the peptidase – peptide interface as well as predict novel interactions not present in the training data. Success in using structure-based energetic signatures and molecular docking for binding partner identification has been achieved for several peptide recognition modules such as SH3 and PDZ domains(Hou et al. 2008; Teyra et al. 2012; Li et al. 2011; Smith & Kortemme 2010; Crivelli et al. 2013), major histocompatibility complex (Yanover & Bradley 2011) and for the enzymes methyltransferase(Lanouette et al. 2015), farnesyltransferase (London et al. 2011), and HIV protease (Chaudhury & Gray 2009; Jensen et al. 2014).

We show here that a structure-based approach, guided by the knowledge of mechanism, can be successfully integrated with machine learning to predict substrates for a mechanistically diverse enzyme family such as proteases with high accuracy.

Proteolytic sites in full-length proteins are more often found in exposed regions of the structure, and more frequently in flexible loops and beta conformations compared to buried regions and alpha helices (Agard et al. 2012). A substrate sequence generally adopts an extended conformation in the protease active site (Tyndall et al. 2005), and surface-exposed loops and beta-strand regions are likely to pay a smaller reorganization penalty to adopt this extended conformation. Therefore, we incorporated the local structure preferences of the substrates in our datasets by computing local sequence-structure compatibility – an implicit assumption in our approach is that every candidate peptide sequence is equally accessible to the protease active site. This assumption is valid when analyzing the extended substrate specificity of the protease, but for the task of predicting cleavage sites in a given whole protein sequence, additional solvent accessibility and structural information are expected to modulate cleavability. Barkan et al. have shown that incorporation of such features improved prediction of cleavage sites in whole protein sequences. Furthermore, Julien et al. (Julien et al. 2016) found that cleavage efficiencies of protein substrates identified using a high throughput mass spectrometry-based approach and their synthetic peptide counterparts were correlated. Taken together, it appears likely that local primary sequence specificity (modeled here) largely determines the identity of cleavage sites, although the context of the cleavage site modulates the kinetics of cleavage.

Comparing the performance of the discriminator for the different protease systems included in the benchmark set highlights the strengths and limitations of our approach. Highest Enrichment of cleaved sequences in the top-ranked population is observed for TEV and Granzyme B proteases (Figure 1.2), where the active site is relatively rigid and steric effects and hydrogen bond interactions are the major contributors to specificity, highlighting the strength of the Rosetta force field in modeling these effects. However, performance is more modest for the metalloenzyme MMP-2, which features a zinc ion in the active site, and for the HIV protease, in which loop residues mediate molecular recognition. For these systems, inaccuracies in the modeling of flexibility of the active site conformation, and lack of explicit consideration of entropy changes can lead to increased misclassification. More exhaustive sampling of the backbone degrees of freedom of the loop structural elements is likely to improve performance as observed in other studies of peptide-protein molecular recognition (Smith & Kortemme 2011; London et al. 2011). Finally, while modeling catalytic residue conformations using geometric constraints appears to be a reasonable approximation for most systems considered here as evidenced by success in discrimination, electronic effects may be involved in the vicinity of the active site, especially for the metalloenzyme MMP-2. We also investigated alternative protonation states of key catalytic residues (nucleophiles serine, cysteine, hydroxyl and bases histidine, aspartic acid) in the MM-PBSA pipeline, but these charge changes did not lead to any appreciable increase in the performance (data not shown). It is likely that quantum mechanical (QM) calculations may be required to model these effects more accurately. However, the high computational cost of detailed QM

simulations precludes the use of such calculations for the thousands of substrate-enzyme pairs considered in our study. Advances in QM simulation methodology(Liu et al. 2015) and computational infrastructure are likely to bridge this gap in the future.

In contrast with sequence-based specificity prediction approaches, the unbiased nature of the biophysical substrate specificity predictor developed here should allow the modeling of specificity of protease variants for which experimental data are not available, such as newly emerged drug-resistant variants(Romano et al. 2012) of viral proteases as well as newly-discovered and/or uncharacterized proteases, whose sequences are homologous to proteases of known structure. Energy-based specificity prediction will also aid in the design of protease variants targeted to specific substrates. Current approaches for protease design rely on library-based screening/selection(Varadarajan et al. 2008; Yi et al. 2013; Boulware et al. 2010) in vivo. These directed evolutionary trajectories often proceed via incremental "generalist"(Khersonsky & Tawfik 2010) intermediates that display relaxed specificity, and are, therefore, toxic to cells (or the proteases undergo self-cleavage) and are never identified in the selection. A structure-guided computational design approach based on the evaluation of interaction energies of substrates with protease variants should allow for multiple simultaneous substitutions ("jumps" in the sequence landscape) to allow specificity switching without generating generalist toxic intermediates. Combining structural computation using the discriminator described here with directed evolution should enable more efficient protease specificity design.

2.5. Materials and Methods

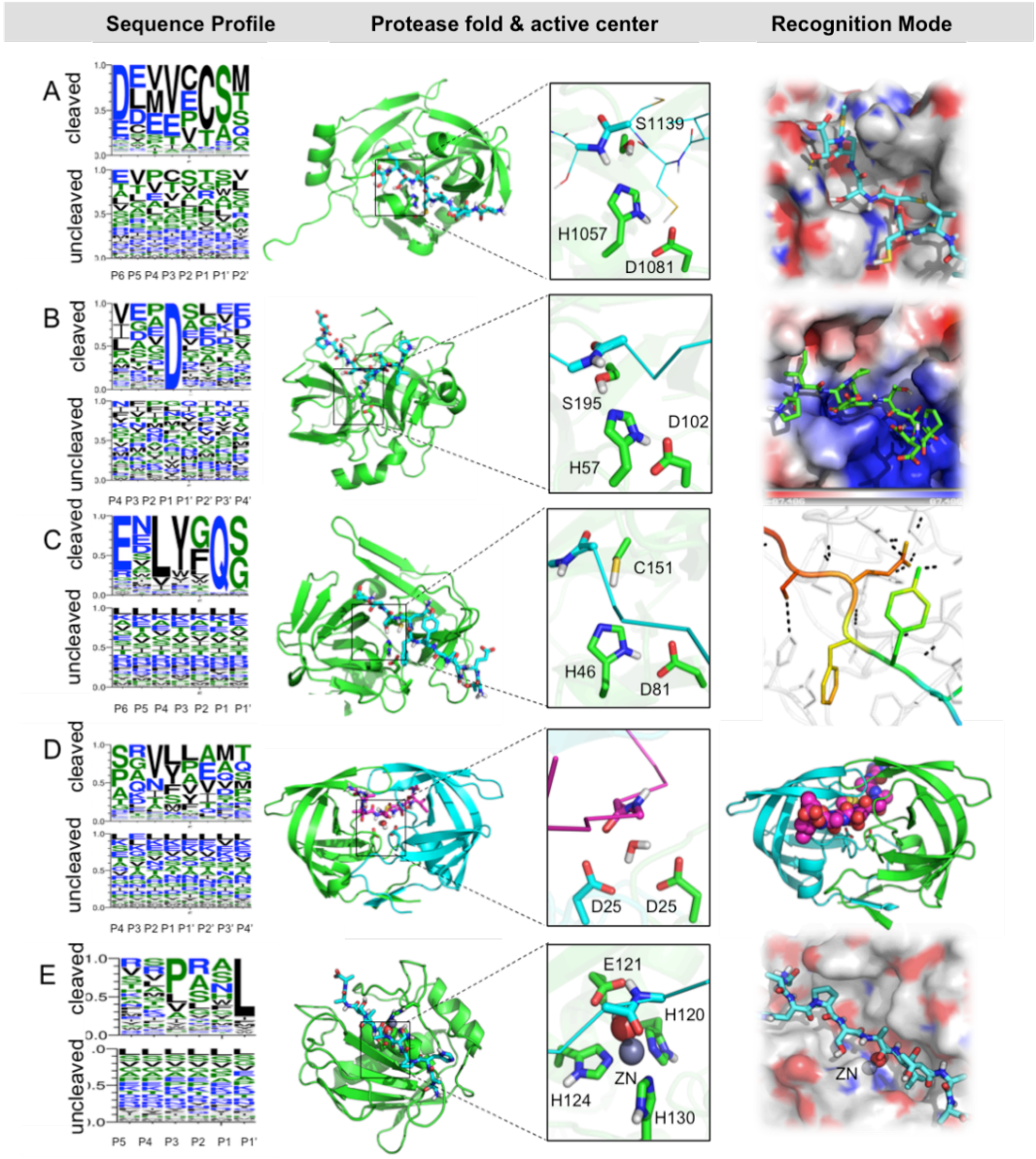


Figure 1.10. The cleaved and uncleaved dataset distributions, model generation and active site geometry of the starting crystal structure and mode of recognition of proteases used in the study

(A) HCV Protease (PDB ID: 3M5N), a serine protease shows recognition via interfacial hydrogen bonding. (B) Granzyme B (PDB ID: 1FI8) a serine protease shows an electrostatic mode of substrate recognition (C) TEV Protease, (PDB ID: 1LVB), a cysteine protease displaying extensive hydrogen bonding at the protease-substrate interface (E) HIV Protease I (PDB ID: 1MT9), a symmetric aspartyl protease, working

via proposed recognition mechanism - substrate-envelope hypothesis. (F) MMP2 (PDB ID: 3AYU) includes a zinc catalytic center

2.5.1. Curation of Benchmark Datasets

Each protease used in the study exhibits diverse mechanisms of action, interface recognition modes, varied folds and biological functions (Figure 1.10) – e.g. TEV Protease (cysteine proteases), HCV NS3/4A protease (serine proteases), Granzyme B (serine protease), HIV Protease-1 (aspartyl protease) and Matrix Metalloprotease -2 (Metalloprotease). The sequences of cleaved and uncleaved substrate peptides for each protease were obtained as detailed below:

HCV protease: We obtained the cleaved and uncleaved sequence sets from a deep sequencing study by Shiryaev et al (Shiryaev et al. 2012). Only sequences with signals above a threshold (Z-score value> 3) at all three time points in their study were considered in order to avoid noise from deep sequencing analyses. We also incorporated sequences from a study by Rögnvaldsson et al(Rögnvaldsson et al. 2009b). Merging both individual sets generated a set with 196 cleaved and 1943 uncleaved sequences.

HIV-PR: 374 cleaved and 1251 uncleaved sequences were obtained from Rögnvaldsson et al(Rögnvaldsson et al. 2009b) .

TEV protease: The cleaved set of 68 sequences was curated from results obtained by Kostallas et al. (Kostallas et al. 2011) and Boulware et al. (Boulware et al. 2010). Due to

the absence of a large uncleaved sequence dataset for the TEV protease, we synthetically generated the uncleaved dataset using a two-residue walk on the TEV polyprotein sequence. The TEV protease is expected to cleave only at one specific site in the polyprotein. Half of the sequences were randomly discarded to generate a dataset of 1520 uncleaved sequences. We ensured that the sequence distribution was not biased toward any specific amino acid type at any peptide position (Figure 1.10).

Granzyme B: The cleaved sequence set was obtained and uncleaved sequence set was adapted from Barkan et al. (Barkan et al. 2010). A subset of the uncleaved sequences was randomly chosen and the amino acid identity at P1 was randomly mutated to all amino acid identities except aspartate and glutamate. A total of 353 cleaved and 1973 uncleaved sequences were chosen.

Matrix Metalloprotease*:* The cleaved sequence set of 455 sequences was obtained from Ratnikov et al (Ratnikov et al. 2014). To curate the uncleaved sequence set, we scanned the CutDB(Igarashi et al. 2007) database for MMP-2 protein substrates. Excluding the known cut sites in these proteins, the rest of the protein sequence was treated as uncleaved using a two-residue walk to generate an uncleaved sequence set of 1818 sequences for MMP-2.

2.5.2. Starting model generation for simulations:

Protease	PDB ID	Resolution	Model Generation
HCV NS3/4A Protease	3M5L, 3M5N	1.9 Å	The P' residues of the bound peptide were built by overlaying PDB ID: 3M5N and PDB ID: 3M5L (inhibitor bound crystal structure) thus allowing us to build a complete substrate bound complex
TEV Protease	1LVB, 1LVM	2.2 Å	Starting model generated from PDB by reverting C151A to WT
MMP2	3AYU, 1BQQ	2.0 Å	Starting model was generated by superimposing PDB ID: 1BQQ with PDB ID: 3AYU(MMP2). The N terminal (P side) residues of the substrate were extended outward to build the complete substrate and were then relaxed to find an optimal substrate conformation
Granzyme B	1FI8	2.2 Å	The interface of the ecotin chain in the crystal structure, spanning eight residue substrate chain was used as the starting point for further calculations
HIV Protease 1	1MT9	2.0 Å	Starting model generated by inverting D25N and V82N from crystal structure to native residue identities

Table 1.4: Details of starting Model Generation for five proteases

We constructed models of peptide-protease bound complexes using high-resolution crystal structures culled from the Protein Data Bank (PDB) (Table 1.4)(Romano et al.

2010; Prabu-Jeyabalan et al. 2003; Waugh et al. 2000; Phan et al. 2002; Hashimoto et al. 2011). Crystal structures were filtered based on the following criteria: a resolution lower than 2.6 Å and a peptide or peptidomimetic inhibitor bound in the crystal structure. We remodeled the crystallographic conformation of the bound peptide to mimic the near-attack conformation for nucleophilic addition step of the proteolysis reaction by enforcing catalytic geometries obtained from mechanistic quantum mechanics simulations and/or crystal structures of proteases bound to inhibitors during Rosetta FastRelax simulations. The selected crystal structures were optimized using a Rosetta FastRelax protocol to find a low energy, stable structure, which was used as a starting point in further calculations. Constraints were applied during FastRelax in order to maintain active site geometry and keep the protease in a catalytically active conformation. Co-ordinate constraints were also applied to the protease backbone to ensure that the structure does not drift away from the crystallographic conformation, while still minimizing energy, as previously described (Nivón et al. 2013).

2.5.3. Calculating Rosetta and Amber energies:

Starting from the relaxed crystal structure described above, we threaded the candidate peptide sequences to generate models of the protease-peptide complex corresponding to each sequence. The energy of the resulting conformation was minimized with constraints using Rosetta FastRelax and ten models were generated for each sequence. During this protocol, the protease backbone was constrained, protease side chains were allowed complete conformational flexibility, whereas peptide side chains and backbone were allowed to sample all degrees of freedom including backbone, sidechain and rigid body

orientation with respect to the protease. The side chains of the catalytically active residues were constrained with respect to the scissile peptide bond of the substrate using enzyme design-style Rosetta constraints. This model represents a pre-transition state near-attack conformation for each of the peptide substrates for the protease. The resulting models were scored with Rosetta's Talaris2013 energy function.

Total residue energies for protease interface residues were extracted for all ten structures representing a single sequence, averaged and stored as "protease energy". Interface residues were defined as those whose C-alpha atom was within 8 Å of any peptide residue's C-alpha atom. We experimented with 8, 10, and 12 Å as the cutoff distance for defining the protease shell, but we found that the discriminator performance was robust to this cutoff value. The sum of total residue energies over all peptide residues was averaged and stored as "peptide energy". Total interface energy was defined as the sum of protease and peptide energies. These models were also scored for "constraint energy" based on the deviation of active site residues geometries from idealized ones. Each energy term was used as a feature during machine learning (see below).

Sampling of the peptide backbone and protease and peptide side chains degrees of freedom was performed before calculating scores for a given complex structure. We optimized the structure sampling protocol by investigating several combinations of sidechain and backbone flexibility for the peptide and the protease, and their relative rigid-body transform. Allowing peptide backbone and sidechain flexibility, and protease sidechain flexibility afforded the highest discriminatory capability (Figure 1.4). All

calculations were performed with the interface RosettaScripts(Fleishman et al. 2011; Richter et al. 2011). Sample xml files used can be found in Supplementary Methods. The AMBER Tools 12 MMPBSA(Miller et al. 2012) application was used to calculate the electrostatic contribution to the bound state energy over the unbound energy for the protease–peptide complex. Run scripts are provided in Supplementary Methods.

2.5.4. Local sequence-structure compatibility

Rosetta's FragmentPicker (Gront et al. 2011) Tool was used to analyze the propensity of a peptide sequence to adopt an extended conformation that is found in protease active sites. We picked 200 fragments for a given peptide sequence, and calculated the RMSD of each fragment with the bound conformation of the peptide. The number of fragments with RMSD > 2.0 in the set of 200 top fragments compared to the bound conformation was used as the score.

2.5.5. Support Vector Machines

An SVM constructs a hyper plane between two sets of data points in multi-dimensional "feature" space, based on a predefined kernel function in order to maximally separate the two datasets. We used the built-in SVM function (MATLAB 2015) with a radial-based kernel function following Barkan et al. (Barkan et al. 2010). In the RBF kernel, parameters C and γ need to be adjusted: C, also called cost factor, is a regularization parameter that controls the trade-off between maximizing the margin and minimizing the prediction error, while γ is a kernel-type parameter that dominates the generalization ability of SVM by regulating the amplitude of the kernel function. We optimized the

training parameters of SVM based on 5-cross-validation tests. C- and γ-values of 10 and 10, respectively, were used.

Sequence features: Each position within the sequence was considered to be one feature. The one letter amino acid codes were transformed into an index, which was calculated from the rank of the amino acid residue in an alphabetical ordering of all amino acids as well as on its position in the sequence from N to C terminus on the substrate chain as in Barkan et al.(Barkan et al. 2010). All 20 amino acids at each position in the peptide were assigned a number using the formula n*20+i, where n represents the position of the residue in the peptide sequence and i represents the position of the residue in an alphabetical ordering of amino acids by their one letter code.

Structure features: Each contributing discriminator energy score was imported into the SVM as an independent feature. The structure-based Rosetta energies ("Interface residue peptide energy", "Interface residue protease energy", "Reorganization penalty", "constraint energy") and Amber energy ("electrostatic energy") were used as features. The SVMs were cross-validated using an 80-20 bootstrap over 1000 iterations.

2.5.6. Generation of a computational library for HCV NS3/4A substrate from P6 through P2 positions:

The mutational scanning was executed in two parts. We generated models of the protease–peptide complex for substrate positions P6 through P4, energy minimized and scored them using the computational protocol descried above. Ten structures were

generated for each sequence. The models were evaluated using the weighted optimized energies as used in the discriminator. The top scoring 66 sequences were identified, and 26,400 models were generated by sampling P3 and P2 substrate positions for each sequence. These 26,400 models were subjected to energy minimization and score calculations as previously described. To calculate their final score, Rosetta interface energy, constraint energy, and AMBER MMPBSA electrostatic energy were used at the optimized Enrichment values. To reduce computational costs, the reorganization penalty score was not included in the final score calculation since it did not measurably change the auROC value in the benchmark set (Figure 1.3). The sequences that lay in the score distribution of the native cleaved sequences were further analyzed. These were filtered to be most different from the initial HCV cleaved sequence distribution and clustered using Hamming distance into 4 main sequence pools- CED*, LEE*, FED*, YED*. Representative sequences from the first three sequence clusters were tested experimentally.

2.5.7. Flow Cytometry:

We used the Yeast ER Sequestration and Screening Assay (YESS) for in vivo testing of predicted substrates of the HCV protease. The LY104 construct for the assay was a gift from Y. Li, B. Iverson, and G. Georgiou (University of Texas at Austin). The sequences to be tested were cloned into LY104 using a Restriction Free Cloning method(Bond & Naus 2012). Table 1.5 lists all the primers associated with the cloning protocol.

Sequence	Primers
LEEFFC SG	FOR: CGGTAGCGGAGGCGGAGGGTCGTTGGAAGAATTCTTCTGTTCAGG

	C
	REV: CTGCCTTTATCATCATCATCTTTATAATCACTGCCGCCTGAACAGAAGAATTCTTCC
LEEYQCSG	FOR: CGGTAGCGGAGGCGGAGGGTCGTTGGAAGAATATCAATGTTCAGGCG
	REV: CTGCCTTTATCATCATCATCTTTATAATCACTGCCGCCTGAACATTGATATTCTTCCAA
CEDYFCSG	FOR: CGGTAGCGGAGGCGGAGGGTCGTGTGAAGATYMTTTCTGTTCAGGCG
	REV: CTGCCTTTATCATCATCATCTTTATAATCACTGCCGCCTGAACAGAAAKRATCTTCACA
FEDFQCSG	FOR: CGGTAGCGGAGGCGGAGGGTCGTTCGAAGATTTCCAATGTTCAGGC
	REV: CTGCCTTTATCATCATCATCTTTATAATCACTGCCGCCTGAACATTGGAAATCTTCG

Table 1.5: Primers used for molecular cloning the sequences to be tested in the YESS assay into the assay (LY104) vector using RF cloning

The positive control and test plasmids were then transformed into the EBY100 competent yeast strain. They were plated on selective complete (SC) media (20 g/L glucose) with a selective amino acid mix (-Trp, - Ura). After two days of growth, a single colony was transferred to a 2 mL SC media culture tube supplemented with 2 µL of 1000x antibiotics (carbenicillin, kanamycin). The growth cultures were incubated for ~24h (OD_{600} 2.0 – 3.0) in a 30 ^{o}C shaking incubator. 1.5 x 10^{7}cells(OD_{600} ~0.5) were pelleted and resuspended in 2 mL induction media (20 g/L galactose, 2 g/L glucose) supplemented with 2 µL each of 1000x antibiotics (carbenicillin, kanamycin). The induction cultures were grown overnight at 30 ^{o}C to an OD_{600} of 1-1.5. All spins in the protocol were done at 3000 r.c.f for 5 min. The induced cultures were pelleted and washed with 500 µL PBS

followed by 500 µL PBS+ 0.5% BSA. 1 µL of each antibody stain(anti-FLAG, anti-HA) was incubated with 10^7 cells for 30 min at 4 ^oC. The samples were resuspended by vortexing and incubated at RT for an additional 30 min. The cells were washed with 100µL PBS with 0.5% BSA, pelleted and then resuspended in 500 µL PBS. Samples were diluted to achieve a final concentration of 10^6 cells/mL and then FITC (anti-HA) and PE(anti-FLAG) intensities were detected using a Flow Cytometer (Beckman Coulter Gallios).

2.6. References

Agard, N.J. et al., 2012. Global kinetic analysis of proteolysis via quantitative targeted proteomics. Proceedings of the National Academy of Sciences of the United States of America, 109(6), pp.1913–8.

Appadurai, R. & Senapati, S., 2016. Dynamical Network of HIV-1 Protease Mutants Reveals the Mechanism of Drug Resistance and Unhindered Activity. Biochemistry, 55(10), pp.1529–40.

Backes, B.J. et al., 2000. Synthesis of positional-scanning libraries of fluorogenic peptide substrates to define the extended substrate specificity of plasmin and thrombin. Nature biotechnology, 18(2), pp.187–93.

Barkan, D.T. et al., 2010. Prediction of protease substrates using sequence and structure features. Bioinformatics (Oxford, England), 26(14), pp.1714–22.

Baugh, E.H. et al., 2016. Robust classification of protein variation using structural modelling and large-scale data integration. Nucleic acids research, 44(6), pp.2501–13.

van den Berg, B.H.J. & Tholey, A., 2012. Mass spectrometry-based proteomics strategies for protease cleavage site identification. Proteomics, 12(4–5), pp.516–29.

Bond, S.R. & Naus, C.C., 2012. RF-Cloning.org: an online tool for the design of restriction-free cloning projects. Nucleic acids research, 40(Web Server issue), pp.W209-13.

Boulware, K.T., Jabaiah, A. & Daugherty, P.S., 2010. Evolutionary optimization of peptide substrates for proteases that exhibit rapid hydrolysis kinetics. Biotechnology and bioengineering, 106(3), pp.339–46.

Boyd, S.E. et al., 2004. PoPS: a computational tool for modeling and predicting protease specificity. Proceedings / IEEE Computational Systems Bioinformatics Conference, CSB. IEEE Computational Systems Bioinformatics Conference, pp.372–81.

Di Cera, E. & Cantwell, A.M., 2001. Determinants of thrombin specificity. Annals of the New York Academy of Sciences, 936, pp.133–46.

Chaudhury, S. & Gray, J.J., 2009. Identification of structural mechanisms of HIV-1 protease specificity using computational peptide docking: implications for drug resistance. Structure (London, England : 1993), 17(12), pp.1636–48.

Crivelli, J.J. et al., 2013. Simultaneous prediction of binding free energy and specificity for PDZ domain-peptide interactions. Journal of computer-aided molecular design, 27(12), pp.1051–65.

Drag, M. & Salvesen, G.S., 2010. Emerging principles in protease-based drug discovery. Nature reviews. Drug discovery, 9(9), pp.690–701.

Eder, J. et al., 2007. Aspartic proteases in drug discovery. Current pharmaceutical design, 13(3), pp.271–85.

Fleishman, S.J. et al., 2011. RosettaScripts: A Scripting Language Interface to the Rosetta Macromolecular Modeling Suite V. N. Uversky, ed. PLoS ONE, 6(6), p.e20161.

Fretwell, J.F. et al., 2008. Characterization of a randomized FRET library for protease specificity determination. Molecular bioSystems, 4(8), pp.862–70.

Grakoui, A., McCourt, D.W., et al., 1993. Characterization of the hepatitis C virus-encoded serine proteinase: determination of proteinase-dependent polyprotein cleavage sites. Journal of virology, 67(5), pp.2832–43.

Grakoui, A., Wychowski, C., et al., 1993. Expression and identification of hepatitis C virus polyprotein cleavage products. Journal of virology, 67(3), pp.1385–95.

Gront, D. et al., 2011. Generalized fragment picking in Rosetta: design, protocols and applications. PloS one, 6(8), p.e23294.

Hashimoto, H. et al., 2011. Structural basis for matrix metalloproteinase-2 (MMP-2)-selective inhibitory action of β-amyloid precursor protein-derived inhibitor. The Journal of biological chemistry, 286(38), pp.33236–43.

Hedstrom, L., 2002a. Introduction: Proteases. Chem Rev., 102(12), pp. 4429-4430.

Hedstrom, L., 2002b. Serine Protease Mechanism and Specificity. Chemical Reviews, 102(12), pp.4501–4524.

Hou, T. et al., 2008. Characterization of domain-peptide interaction interface: a case study on the amphiphysin-1 SH3 domain. Journal of molecular biology, 376(4), pp.1201–14.

Igarashi, Y. et al., 2007. CutDB: a proteolytic event database. Nucleic acids research, 35(Database issue), pp.D546-9.

Jensen, J.H. et al., 2014. In silico prediction of mutant HIV-1 proteases cleaving a target sequence. PloS one, 9(5), p.e95833.

Julien, O. et al., 2016. Quantitative MS-based enzymology of caspases reveals distinct protein substrate specificities, hierarchies, and cellular roles. Proceedings of the National Academy of Sciences of the United States of America, 113(14), pp.E2001-10.

Khersonsky, O. & Tawfik, D.S., 2010. Enzyme promiscuity: a mechanistic and evolutionary perspective. Annual review of biochemistry, 79, pp.471–505.

Kostallas, G., Löfdahl, P.-Å. & Samuelson, P., 2011. Substrate profiling of tobacco etch virus protease using a novel fluorescence-assisted whole-cell assay. PloS one, 6(1), p.e16136.

Lanouette, S. et al., 2015. Discovery of substrates for a SET domain lysine methyltransferase predicted by multistate computational protein design. Structure (London, England : 1993), 23(1), pp.206–15.

Li, B.-Q. et al., 2012. Prediction of protein cleavage site with feature selection by random forest. PloS one, 7(9), p.e45854.

Li, N. et al., 2011. Characterization of PDZ domain-peptide interaction interface based on energetic patterns. Proteins, 79(11), pp.3208–20.

Liu, F. et al., 2015. Quantum Chemistry for Solvated Molecules on Graphical Processing Units Using Polarizable Continuum Models. Journal of chemical theory and computation, 11(7), pp.3131–44.

London, N. et al., 2011. Identification of a novel class of farnesylation targets by structure-based modeling of binding specificity. PLoS computational biology, 7(10), p.e1002170.

López-Otín, C. & Bond, J.S., 2008. Proteases: multifunctional enzymes in life and disease. The Journal of biological chemistry, 283(45), pp.30433–7.

Miller, B.R. et al., 2012. MMPBSA.py : An Efficient Program for End-State Free Energy Calculations. Journal of Chemical Theory and Computation, 8(9), pp.3314–3321.

Nivón, L.G. et al., 2013. A Pareto-Optimal Refinement Method for Protein Design

Scaffolds Y. Zhang, ed. PLoS ONE, 8(4), p.e59004.

Phan, J. et al., 2002. Structural basis for the substrate specificity of tobacco etch virus protease. The Journal of biological chemistry, 277(52), pp.50564–72.

Poreba, M. & Drag, M., 2010. Current strategies for probing substrate specificity of proteases. Current medicinal chemistry, 17(33), pp.3968–95.

Powers, J.C. et al., 1993. Proteases--structures, mechanism and inhibitors. Agents and actions. Supplements, 42, pp.3–18.

Prabu-Jeyabalan, M. et al., 2003. Viability of a drug-resistant human immunodeficiency virus type 1 protease variant: structural insights for better antiviral therapy. Journal of virology, 77(2), pp.1306–15.

Puente, X.S. et al., 2003. Human and mouse proteases: a comparative genomic approach. Nature reviews. Genetics, 4(7), pp.544–58.

Ratnikov, B., Cieplak, P. & Smith, J.W., 2009. High throughput substrate phage display for protease profiling. Methods in molecular biology (Clifton, N.J.), 539, pp.93–114.

Ratnikov, B.I. et al., 2014. Basis for substrate recognition and distinction by matrix metalloproteinases. Proceedings of the National Academy of Sciences, 111(40), pp.E4148–E4155.

Ratnikov, B.I. et al., 2014. Basis for substrate recognition and distinction by matrix metalloproteinases. Proceedings of the National Academy of Sciences of the United States of America, 111(40), pp.E4148-55.

Rawlings, N.D., Barrett, A.J. & Bateman, A., 2010. MEROPS: the peptidase database. Nucleic acids research, 38(Database issue), pp.D227-33.

Rawlings, N.D. & Salvesen, G., 2013. Handbook of Proteolytic Enzymes. In Handbook of Proteolytic Enzymes.

Richter, F. et al., 2011. De novo enzyme design using Rosetta3. PloS one, 6(5), p.e19230.

Rögnvaldsson, T. et al., 2009a. How to find simple and accurate rules for viral protease cleavage specificities. BMC bioinformatics, 10, p.149.

Rögnvaldsson, T. et al., 2009b. How to find simple and accurate rules for viral protease cleavage specificities. BMC bioinformatics, 10(1), p.149.

Romano, K.P. et al., 2010. Drug resistance against HCV NS3/4A inhibitors is defined by the balance of substrate recognition versus inhibitor binding. Proceedings of the National Academy of Sciences of the United States of America, 107(49), pp.20986–91.

Romano, K.P. et al., 2012. The Molecular Basis of Drug Resistance against Hepatitis C Virus NS3/4A Protease Inhibitors A. Gamarnik, ed. PLoS Pathogens, 8(7), p.e1002832.

Scheel, T.K.H. & Rice, C.M., 2013. Understanding the hepatitis C virus life cycle paves the way for highly effective therapies. Nature Medicine, 19(7), pp.837–849.

Shiryaev, S.A. et al., 2012. New details of HCV NS3/4A proteinase functionality revealed by a high-throughput cleavage assay. PloS one, 7(4), p.e35759.

Smith, C.A. & Kortemme, T., 2011. Predicting the tolerated sequences for proteins and protein interfaces using RosettaBackrub flexible backbone design. PloS one, 6(7), p.e20451. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21789164

Smith, C.A. & Kortemme, T., 2010. Structure-based prediction of the peptide sequence space recognized by natural and synthetic PDZ domains. Journal of molecular biology, 402(2), pp.460–74.

Song, J. et al., 2011. Bioinformatic approaches for predicting substrates of proteases. Journal of bioinformatics and computational biology, 9(1), pp.149–78.

Song, J. et al., 2012. PROSPER: an integrated feature-based tool for predicting protease substrate cleavage sites. PloS one, 7(11), p.e50300.

Tawfik, D.S., 2014. Accuracy-rate tradeoffs: how do enzymes meet demands of selectivity and catalytic efficiency? Current opinion in chemical biology, 21, pp.73–80.

Teyra, J., Sidhu, S.S. & Kim, P.M., 2012. Elucidation of the binding preferences of peptide recognition modules: SH3 and PDZ domains. FEBS letters, 586(17), pp.2631–7.

Turk, B.E. et al., 2001. Determination of protease cleavage site motifs using mixture-based oriented peptide libraries. Nature biotechnology, 19(7), pp.661–7.

Tyka, M.D. et al., 2011. Alternate states of proteins revealed by detailed energy landscape mapping. Journal of molecular biology, 405(2), pp.607–18.

Tyndall, J.D.A., Nall, T. & Fairlie, D.P., 2005. Proteases universally recognize beta strands in their active sites. Chemical reviews, 105(3), pp.973–99.

Varadarajan, N. et al., 2008. Highly active and selective endopeptidases with programmed substrate specificities. Nature chemical biology, 4(5), pp.290–4.

Verspurten, J. et al., 2009. SitePredicting the cleavage of proteinase substrates. Trends in biochemical sciences, 34(7), pp.319–23.

Vizovišek, M. et al., 2016. Current trends and challenges in proteomic identification of protease substrates. Biochimie, 122, pp.77–87.

Waugh, S.M. et al., 2000. The structure of the pro-apoptotic protease granzyme B reveals the molecular determinants of its specificity. Nature structural biology, 7(9), pp.762–5.

Yanover, C. & Bradley, P., 2011. Large-scale characterization of peptide-MHC binding landscapes with structural simulations. Proceedings of the National Academy of Sciences of the United States of America, 108(17), pp.6981–6.

Yi, L. et al., 2013. Engineering of TEV protease variants by yeast ER sequestration screening (YESS) of combinatorial libraries. Proceedings of the National Academy of Sciences of the United States of America, 110(18), pp.7229–34.

Yi, L. et al., 2015. Yeast Endoplasmic Reticulum Sequestration Screening for the Engineering of Proteases from Libraries Expressed in Yeast. Methods in molecular biology (Clifton, N.J.), 1319, pp.81–93.

2.7. Supplementary Methods:

The MMPBSA calculation includes the following steps:

1. Preparation of AMBER input .pdb files
2. Preparation of input parameter and topology files
3. MMPBSA Calculation

Description of each of the steps below:

In order to transform a pdb file into an AMBER readable format the hydrogens and virtual atoms are stripped. The subsequent file is loaded into AMBER using the following script using a tleap interface.

```
source leaprc.gaff
source leaprc.ff12SB
loadamberparams frcmod.ionsjc_tip3p
d$i = loadpdb "toload_$i.pdb"
addions d$i Cl- 0
charge d$i
saveamberparm d$i d$i.prmtop d$i.inpcrd
quit
```

The files saved as d$i.prmtop and d$i.inpcrd are inputs to the ante-MMPBSA.py program which generates the receptor-ligand, receptor only and ligand only topology files. An AMBER topology file is used to specify atom types, charges, etc. The inpcrd / input coordinate file is used to build the connections which forms the overall structure of the pdb.

ante-MMPBSA.py -p d$i.prmtop -c d_c$i.prmtop -s @Cl-

ante-MMPBSA.py -p d_c$i.prmtop -r d_r$i.prmtop -l d_l$i.prmtop -n : “residue range”

Residue range: specify the pose numbering of the peptide

The final step involves using the inpcrd and prmtop files to calculate the MMPBSA contribution of the complex. This is done by calculating the electrostatic energy of the peptide and protease separately as well as in a bound state

The following commandline is used for MMPBSA calculation

MMPBSA.py -O -i mmpbsa.in -o FINAL_RESULTS_MMPBSA.dat -sp d$i.prmtop -cp d_c$i.prmtop -rp d_r$i.prmtop -lp d_l$i.prmtop -y *.inpcrd

For MMP2: The pdbs in these cases needed to be analyzed differently because of the presence of heteroatoms such as Zinc and Water that are involved in the active sites respectively.

The water is modeled using the TP5.lib and the following command is added to the prep script

Sample Scripts:

Sample xml for initial Relax:

```
<dock_design>
    <SCOREFXNS>
      <myscore weights=enzdes.wts/>
    </SCOREFXNS>
    <TASKOPERATIONS>
      <ProteinInterfaceDesign name=pido design_chain2=0 modify_after_jump=1/>
      <InitializeFromCommandline name=init/>
      <ReadResfile name=rrf filename="PATH TO RESFILE"/>
    </TASKOPERATIONS>

    <FILTERS>
    </FILTERS>
```

<MOVERS>

<AddOrRemoveMatchCsts name=cstadd cst_instruction=add_new/>

<FastRelax name=fastrelax scorefxn=myscore repeats=8 task_operations=pido,init>

<MoveMap name=mm>

<Chain number=2 chi=1 bb=1/>

<Chain number=1 chi=1 bb=1/>

<Jump number =1 setting=1/>

</MoveMap>

</FastRelax>

<TaskAwareMinMover name =min_pro task_operations=rrf scorefxn=myscore chi=1 bb=0 jump=0/>

<PackRotamersMover name=repack task_operations=rrf/>

<ConstraintSetMover name=protease_cst cst_file="PATH_TO_PROTEASE_BACKBONE_HEAVY_ATOM_CONSTRAINT_FILE"/>

</MOVERS>

<APPLY_TO_POSE>

</APPLY_TO_POSE>

<PROTOCOLS>

<Add mover_name=protease_cst/>

```
        <Add mover_name=repack/>
        <Add mover_name=min_pro/>
        <Add mover_name=cstadd/>
        <Add mover_name=fastrelax/>
</PROTOCOLS>
</dock_design>
```

Command line:

```
~<PATH_TO_ROSETTA_BIN> rosetta_scripts.static.linuxgccrelease  -jd2:ntrials 1 -
nstruct 20  -parser:protocol <PATH_TO_RELAX_XML> -database
<PATH_TO_DATABASE> -out::prefix Job_${i}_ -s <PATH_TO_STARTING_PDB> -
run:preserve_header -enzdes::cstfile <PATH_TO_CONSTRAINT_FILE>  -
out:file:output_virtual @<PATH_TO_FLAGS_FILE>
```

Sample Script For Mutate, FastRelax, Scoring

```
#MUTATERUN
<PATH_TO_EXECUTABLE>/rosetta_scripts.static.linuxgccrelease -nstruct 10  -
jd2:ntrials 1 -parser:protocol <PATH_TO_XML> -database <PATH_TO_DATABASE>
-out::prefix $1_mut_ -s <PATH_TO_STARTING_PDB> -enzdes:cstfile
<PATH_TO_CSTFILE> -run:preserve_header @<PATH_TO_FLAGSFILE> >
design.log
```

find `pwd` -name "$1_mut_*00*pdb" > tlist

cp ~/Rosetta/main/database/scoring/weights/talaris2013 ./

#SCORINGRUN

~/Rosetta/main/source/bin/rosetta_scripts.static.linuxgccrelease -jd2:ntrials 1 -parser:protocol <PATH_TO_SCORING_XML> -database <PATH_TO_DATABASE> -out::prefix Scores_ -l tlist -in:file:native <PATH_TO_STARTINGPDB> -run:preserve_header @<PATH_TO_FLAGSFILE> -score:weights talaris2013 > scoring.log

ls Scores_*.pdb > slist

#CSTRUN

~/Rosetta/main/source/bin/rosetta_scripts.static.linuxgccrelease -jd2:ntrials 1 -parser:protocol <PATH_TO_XML> -database ~/Rosetta/main/database/ -out::prefix $1_cst_ -l tlist -enzdes:cstfile <PATH_TO_CSTFILE> -run:preserve_header @<PATH_TO_FLAGSFILE> -jd2:enzdes_out > cst.log

Protease Mutate:

<dock_design>

<SCOREFXNS>

<myscore weights=enzdes.wts/>

```
    </SCOREFXNS>
    <TASKOPERATIONS>
      <ProteinInterfaceDesign name=pido design_chain2=0 modify_after_jump=0/>
      <InitializeFromCommandline name=init/>
      <ReadResfile name=rrf filename="PATH_TO_RESFILE"/>
    </TASKOPERATIONS>

    <FILTERS>
    </FILTERS>

    <MOVERS>
      <MutateResidue name=mut1 target=Res#1 new_res=DM1/>
      <MutateResidue name=mut2 target= Res#2 new_res=DM2/>
      <MutateResidue name=mut3 target= Res#3new_res=DM3/>
      <MutateResidue name=mut4 target= Res#4 new_res=DM4/>
      <MutateResidue name=mut5 target= Res#5 new_res=DM5/>
      <MutateResidue name=mut6 target= Res#6 new_res=DM6/>
<AddOrRemoveMatchCsts name=cstadd cst_instruction=add_new/>
      <FastRelax name=fastrelax scorefxn=myscore repeats=8 task_operations=pido,init>
      <MoveMap name=mm>
                <Chain number=2 chi=1 bb=1/>
                <Chain number=1 chi=1 bb=0/>
                <Jump number =1 setting=1/>
```

```
        </MoveMap>
        </FastRelax>
        <TaskAwareMinMover name =min_pro task_operations=rrf chi=1 bb=0 jump=0/>
        <PackRotamersMover name=repack task_operations=rrf/>

</MOVERS>
<APPLY_TO_POSE>
</APPLY_TO_POSE>

     <PROTOCOLS>
            <Add mover_name=mut1/>
            <Add mover_name=mut2/>
            <Add mover_name=mut3/>
            <Add mover_name=mut4/>
            <Add mover_name=mut5/>
            <Add mover_name=mut6/>
          <Add mover_name=repack/>
            <Add mover_name=cstadd/>
            <Add mover_name=fastrelax/>
</PROTOCOLS>
</dock_design>
```

SCORING XML

CST XML

```
<dock_design>
    <SCOREFXNS>
        <myscore weights=enzdes.wts/>
    </SCOREFXNS>
    <TASKOPERATIONS>
     <InitializeFromCommandline name=init/>
    </TASKOPERATIONS>

    <FILTERS>
     <EnzScore name="cstenergy" scorefxn=myscore whole_pose=1 score_type=cstE
energy_cutoff=99999.0/>
    </FILTERS>

    <MOVERS>
     <AddOrRemoveMatchCsts name=cstadd cst_instruction=add_new/>
    </MOVERS>

    <APPLY_TO_POSE>
    </APPLY_TO_POSE>

    <PROTOCOLS>
```

```
        <Add mover_name=cstadd/>
        <Add filter_name=cstenergy/>
      </PROTOCOLS>
</dock_design>
```

AMBER MMPBSA

```
cat >tleap.in <<EOF
source leaprc.gaff
source leaprc.ff12SB_manasi
loadamberparams frcmod.ionsjc_tip3p
loadamberparams frcmod.ionslrcm_hfe_tip3p
d$i = loadpdb "toload_$i.pdb"
charge d$i
saveamberparm d$i d$i.prmtop d$i.inpcrd
quit
EOF
tleap -f tleap.in

ante-MMPBSA.py -p d$i.prmtop -c d_c$i.prmtop -s @Cl-
ante-MMPBSA.py -p d_c$i.prmtop -r d_r$i.prmtop -l d_l$i.prmtop -n :199-208
```

```
MMPBSA.py -O -i mmpbsa.in -o FINAL_RESULTS_MMPBSA.dat -cp d_c$i.prmtop -
rp d_r$i.prmtop -lp d_l$i.prmtop -y d$i.inpcrd
```

MATLAB

```
function [test, testlab, ttcleaved, to, ts, train, trainlab,  a, f, X, Y, T, AUC, AUCav, Std,
Performanceav,Stdp] = coduh(A, LABELS, cleaved, uncleaved, boxconstraint, rbfsigma)

clearvars -except A LABELS cleaved uncleaved boxconstraint rbfsigma TABLE

X = [];
Y = [];
T = [];
AUC = [];


 [numberofelements len] = size(A);
tic

for s = 1:1000

    zcleaved = ceil(0.2%*cleaved);
    zuncleaved = ceil(0.2%*uncleaved);
```

```
    ttcleaved = randperm(cleaved,zcleaved);
%generatingRandomFromNumLength
    ttuncleaved = randperm((numberofelements - cleaved), zuncleaved) + cleaved;
    t = vertcat(ttcleaved',ttuncleaved');
    to(:,s) = vertcat(ttcleaved',ttuncleaved');
    ts(s) = length(t);
    z = zcleaved + zuncleaved;

test(:,:,s) = A(t,:);
testlab(:,:,s) = LABELS(t,:);

    x = numberofelements - z;
    train(:,:,s) = zeros(x, len);
    trainlab(:,:,s)= cell(x,1);

    clear n1;
    n1 = 1;

    for i = 1:numberofelements

       if i ~= t(:)
          train(n1,:,s) = A(i,:);
          trainlab(n1,s) = LABELS(i);
```

```
        n1 = n1 + 1;
      end

  end
```

```
svmrbf =[];
svmrbf=svmtrain(train(:,:,s), trainlab(:,s), 'kernel_function', 'rbf', 'boxconstraint',
boxconstraint, 'rbf_sigma', rbfsigma);


%%TEST%%
V = svmclassify(svmrbf,test(:,:,s));
result = transpose(V);
a(:,s)=transpose(result);

shift = svmrbf.ScaleData.shift;
scale = svmrbf.ScaleData.scaleFactor;
Xnew = bsxfun(@plus,test(:,:,s),shift);
Xnew = bsxfun(@times,Xnew,scale);
sv = svmrbf.SupportVectors;
alphaHat = svmrbf.Alpha;
bias = svmrbf.Bias;
```

```
kfun = svmrbf.KernelFunction;

kfunargs = svmrbf.KernelFunctionArgs;

f(:,s) = kfun(sv,Xnew,kfunargs{:})'*alphaHat(:) + bias;


[X(:,s),Y(:,s),T(:,s),AUC(s)] = perfcurve(testlab(:,:,s), f(:,s) ,'CLEAVED', 'Xcrit','reca',

'YCrit', 'prec' );


AUCav = mean(AUC);

Std = std(AUC);


%ACCURACY

tf(:,s) = strcmp (a(:,s), testlab(:,s));

Performance(s) = sum(tf(:,s)) / numel(a(:,s));

Performanceav = mean(Performance);

Stdp = std(Performance);


% %TRAIN

Vtrain = svmclassify(svmrbf,train(:,:,s));

resulttrain = transpose(Vtrain);


%clear train end

atrain(:,s)=transpose(resulttrain);
```

```
shift = svmrbf.ScaleData.shift;
scale = svmrbf.ScaleData.scaleFactor;

Xnew1 = bsxfun(@plus,train(:,:,s),shift);
Xnew1 = bsxfun(@times,Xnew1,scale);
sv = svmrbf.SupportVectors;
alphaHat = svmrbf.Alpha;
bias = svmrbf.Bias;
kfun = svmrbf.KernelFunction;
kfunargs = svmrbf.KernelFunctionArgs;
ftrain(:,s) = kfun(sv,Xnew1,kfunargs{:})'*alphaHat(:) + bias;
display(f(:,s));

[Xtraintemp,Ytraintemp,Ttraintemp,AUCtrain(s)]=
perfcurve(trainlab(:,:,s),ftrain(:,s),'CLEAVED');
[r] = length(Xtraintemp);
Xtrain(1:r, s) = Xtrain(1:r, s) + Xtraintemp;
Ytrain(1:r, s) = Ytrain(1:r, s) + Ytraintemp;
Ttrain(1:r, s) = Ttrain(1:r, s) + Ttraintemp;

clear Xtraintemp Ytraintemp Ttraintemp
```

```
[Xtrain(:,s),Ytrain(:,s),Ttrain(:,s),AUCtrain(s)]=

perfcurve(trainlab(:,:,s),ftrain(:,s),'CLEAVED');


AUCtrainav = mean(AUCtrain);

Stdtrain = std(AUCtrain);


tftrain(:,s) = strcmp (atrain(:,s), trainlab(:,s));

Performancetrain (s)= sum(tftrain(:,s)) / numel(atrain(:,s));

Performancetrainav = mean(Performancetrain);

Stdptrain = std(Performancetrain);


s

end

toc
```

Chapter 3: MFPred - Rapid and Accurate Prediction of Protein-peptide Recognition Multispecificity Using Self-Consistent Mean Field Theory

3.1. Abstract

Multispecificity – the ability of a single receptor protein molecule to interact with multiple substrates – is a hallmark of molecular recognition at protein-protein and protein-peptide interfaces, including enzyme-substrate complexes. The ability to perform structure-based prediction of multispecificity would aid in the identification of novel enzyme substrates, protein interaction partners, and enable design of novel enzymes targeted towards alternative substrates. The relatively slow speed of current biophysical, structure-based methods limits their use for prediction and, especially, design of multispecificity. Here, we develop a rapid, flexible-backbone self-consistent mean field theory-based technique, MFPred, for multispecificity modeling at protein-peptide interfaces. We benchmark our method by predicting experimentally determined peptide specificity profiles for a range of receptors: protease and kinase enzymes, and protein recognition modules including SH2, SH3, MHC Class I and PDZ domains. We observe robust recapitulation of known specificities for all receptor-peptide complexes, and comparison with other methods shows that MFPred results in equivalent or better prediction accuracy with a ~10-1000-fold decrease in computational expense. We find that modeling bound peptide backbone flexibility is key to the observed accuracy of the method. We used MFPred for predicting with high accuracy the impact of receptor-side mutations on experimentally determined multispecificity of a protease enzyme. Our

approach should enable the design of a wide range of altered receptor proteins with programmed multispecificities.

3.2. Introduction

Many natural proteins, including signal transduction hubs and enzymes that process biological information, have evolved to be multispecific – they participate in specific interactions with several interaction partners (Kim et al. 2006; Erijman et al. 2011). Evolution of multispecificity includes selection for both positive and negative specificity, involving recognition and non-recognition, respectively, of sets of interaction partners (Tawfik 2014). Most multispecific interactions arise when the active site of a single receptor protein interacts with multiple binding partners of differing sequence (Schreiber & Keating 2011). Nature uses structurally conserved protein-recognition domains (PRDs), e.g., SH2, SH3 and PDZ domains, to mediate many multispecific interactions (Schutkowski et al. 2004; Khati & Pillay 2004; Tonikian et al. 2008; Vouilleme et al. 2010; Stiffler et al. 2007; Sparks et al. 1996). Thus, it is crucial that methods that model and modulate PRD specificity are able to accurately recapitulate their multispecific nature.

Similar to cascades composed of multispecific PRDs like SH3, SH2 and PDZ domains that mediate signal transduction, proteolytic cascades are ubiquitous in the post-translational transduction of biological information (Li et al. 2013). Protease activity and selectivity is involved in a diverse range of biological processes including digestion, blood clotting, apoptosis and cancer (Chapman et al. 1997; Hirsch et al. 1998; Monahan

& Di Paola 2010; Pampalakis & Sotiropoulou 2007). Proteases are inherently multispecific such that they recognize and proteolyze (or cleave) a range of substrates (positive specificity) while not recognizing others (negative specificity) (Tawfik 2014). For example, viral proteases such as HCV protease that are involved in viral maturation cleave only specific sites in the viral polyprotein but do not cleave others (Scheel & Rice 2014). These proteases may also have evolved the ability to cleave specific host proteins (Kerekatte et al. 1999). Prediction of protease multispecificity is, therefore, key for identifying their substrates under healthy and disease conditions. Additionally, designed proteases with programmed multispecificity have the potential to be used as therapeutics and protein-level knockout reagents in cell culture (Craik et al. 2011). The ability to manipulate protease specificity computationally would enable the creation of such designer proteases with dialed-in recognition specificity, thereby providing tools to interrogate and intervene in biological processes.

Rational modulation of protein-protein or protein-peptide interaction multispecificity has met with limited success, except in a few notable cases, such as coiled-coil interfaces (Newman & Keating 2003; Havranek & Harbury 2002). In principle, computational structure-based modeling methods should be able to recapitulate and modulate multispecificity. In fact, several methods relying on, among others, Monte-Carlo (MC) simulations in sequence and conformation space, and genetic algorithms (GA) have been developed to predict PRD multispecificity (King & Bradley 2010; Smith & Kortemme 2010; Wollacott & Desjarlais 2001; Lanouette et al. 2015; Grigoryan et al. 2009). However, these methods are limited by the time required to enumerate a sufficiently large

number of sequences to sample the substrate/peptide sequence space. As multispecific design entails additional sampling of (thousands) of receptor variants and modeling the multispecificity of each variant separately, using current methods to design receptors for and against specificity profiles is not computationally feasible.

We have developed a structure-based method that eliminates the expense of explicit sequence enumeration in multispecificity modeling. The method uses a self-consistent **M**ean-**F**ield theory-based **Pred**iction (MFPred) approach that expresses specificity as a sitewise probability distribution function that can be calculated relatively rapidly. We have benchmarked MFPred on four diverse proteases and compared the results to MC- and GA-based methods. MFPred has comparable accuracy to MC-based and GA-based methods and provides a tens- to thousands-fold speedup. We demonstrate the generality of MFPred by obtaining significant multispecificity predictions for five diverse classes of protein-recognition domains (PRDs). Finally, as a proof-of-concept for design, we demonstrate that MFPred can recapitulate experimentally determined changes in specificity profiles due to receptor-side mutations.

3.3. Results

3.3.1. Self-Consistent Mean Field Theory-Based Specificity Profile Prediction Algorithm

To predict the specificity profile, we consider an ensemble of peptide backbone conformations bound to a receptor. For each peptide backbone conformation, we simultaneously sample all rotameric conformations of all amino acids at all peptide

residue positions while keeping the receptor backbone and sidechains in their crystallographic conformations. The sidechain conformations at a given peptide position are sampled in the "mean field" of all other sidechain conformations at all other positions and (fixed) receptor residues, as described in Methods. Next, the contribution of each peptide backbone conformation at each peptide position is accounted for by Boltzmann averaging the mean-field specificity profile solution obtained in the previous step. The final specificity profile is constructed by combining these individual predictions. While the sequence specificity prediction described here can be performed using any (pairwise decomposable) energy function, we implemented our prediction method in the context of the Rosetta modeling suite, thus combining its sophisticated energy function with the speed of mean-field sampling (Figure 2.1).

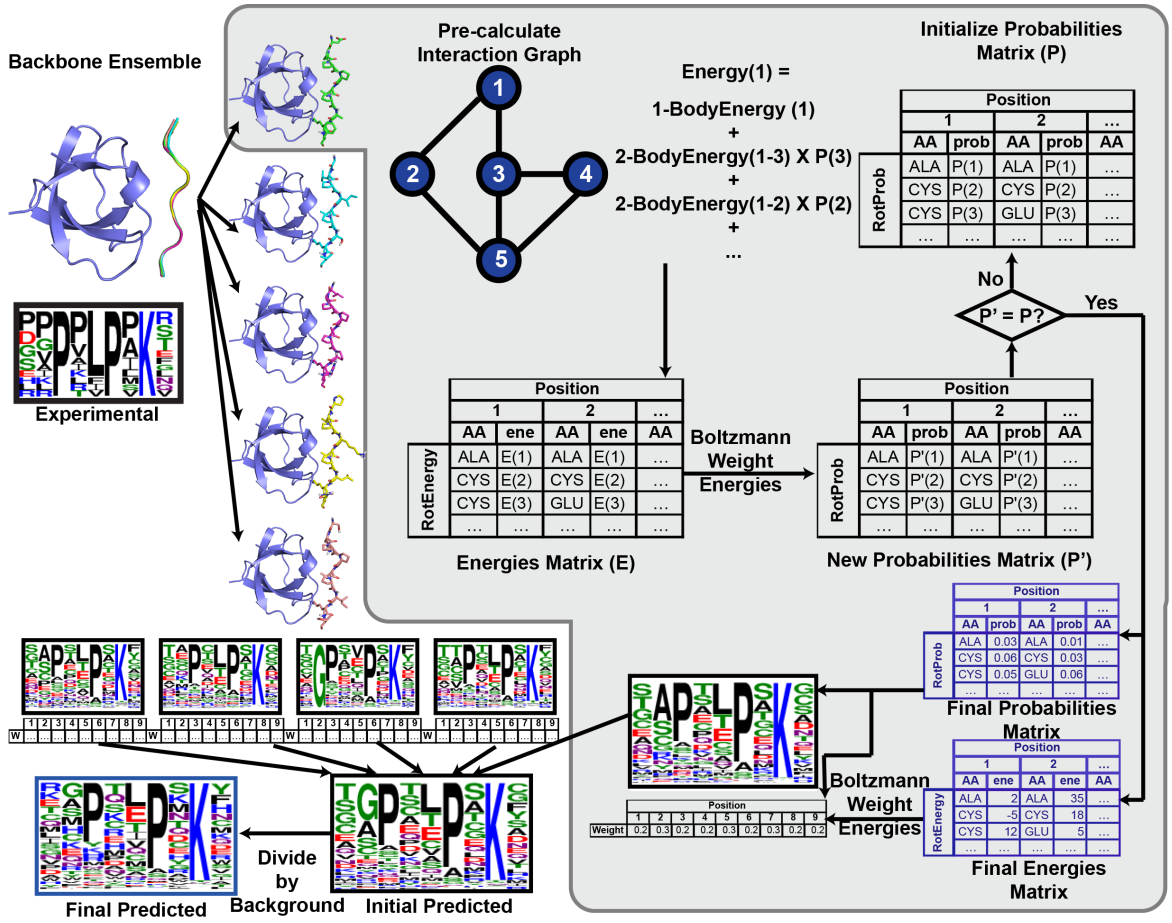


Figure 2.1. MFPred workflow.
MFPred input is a backbone ensemble of a protein/peptide complex, which is generated from a protein structure from the PDB (1CKA here) as described in Methods. For each backbone, Rosetta pre-calculates the interaction graph, which stores intrinsic rotamer one-body energies on the vertices (blue circles) and matrices of rotamer-rotamer two-body energies on the edges (black lines). A probabilities matrix (P) is initialized. Mean-field energies (E) are calculated using the interaction graph and P, and a new matrix, P' is generated from E. If P' is equal to P, convergence has been reached. If not, the process is repeated by updating P with a combination of P and P'. Once convergence is reached, the final energies matrix and probabilities matrix is used to generate the Boltzmann weights of each backbone position, which is then used to average all the backbone specificity profiles together. This specificity profile is divided by the background specificity profile to reach the final predicted specificity profile.

3.3.2. Rationale for Choice of Benchmark Datasets

To test our MFPred method, we sought to first recapitulate experimentally determined specificity profiles of a variety of PRDs. We chose PRDs where both structural as well as

specificity information has been experimentally determined. We focused primarily on protease enzymes for methodology development, and tested the generality of our approach with previously developed benchmarks for multispecificity prediction on PRDs such as a kinase enzyme, and SH3, SH2, MHC, and PDZ domains.

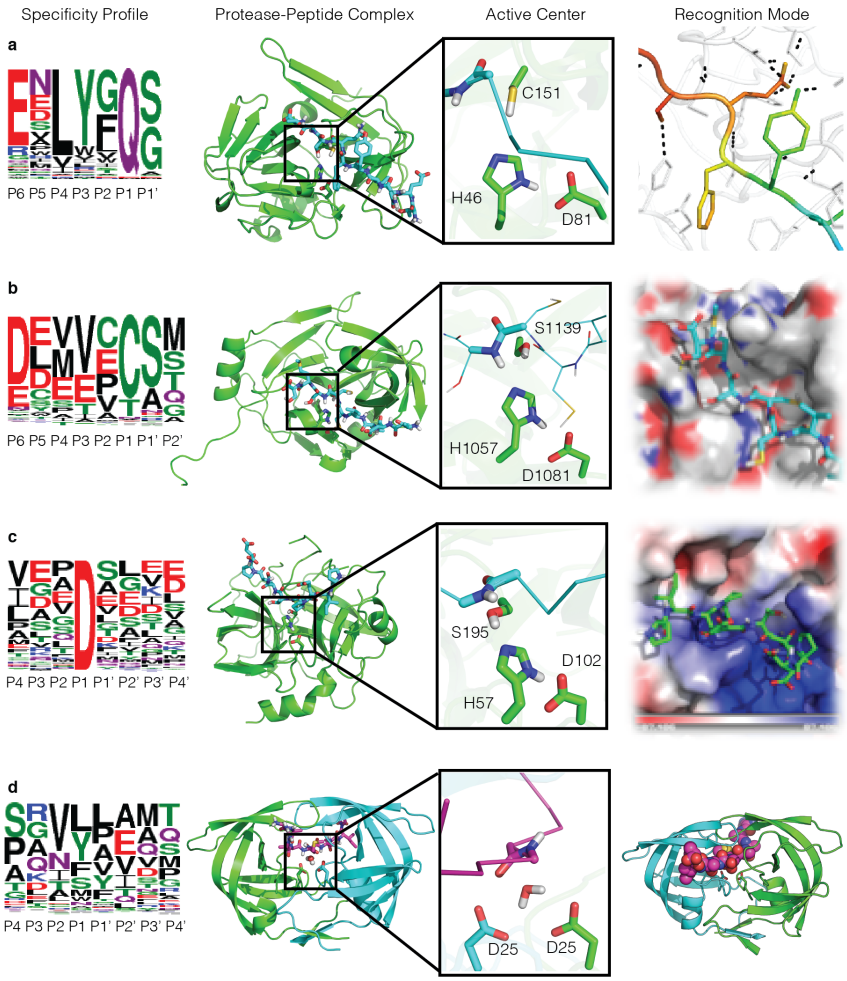


Figure 2.2. Protease benchmark specificity profiles, models, active centers, and recognition modes.
(a) Tobacco etch virus (TEV) protease is a cysteine protease displaying extensive hydrogen bonding recognizes substrates via interfacial hydrogen bonding at the protease substrate interface. (b) Hepatitis C virus (HCV) NS3 protease, a serine protease recognizes substrates through electrostatic interactions (c) Granzyme B, a serine protease recognizes substrates through electrostatic interactions (d) Human immunodeficiency virus (HIV) protease I, a symmetric aspartyl protease, has been proposed to recognize substrates via the substrate – envelope hypothesis.

Protease set. We benchmarked our method on four protease enzymes that had both high-resolution crystal structures with a bound peptide in the Protein Data Bank (PDB) and experimental cleavage data (see Methods for details). The chosen proteases represent the vast diversity seen in structural fold, biological function, and mechanism of action amongst the protease enzyme family (Figure 2.2). Additionally, there is a mix of highly conserved and less specific positions among their specificity profiles, thus enabling us to determine how well MFPred performs with regard to varying degrees of flatness in the experimental specificity profile.

Testing on protein-recognition domains. To test the generality of the MFPred method, we curated a dataset consisting of a variety of non-protease PRDs that had high-resolution crystal structures as protein-peptide complexes in the PDB and experimental binding specificity data available. We tested fourteen PRDs that comprise five classes of PRDs: kinases, SH2 domains, SH3 domains, PDZ domains, and MHC-I proteins. Including these diverse domains allows us to test the method on a range of underlying recognition modes, binding affinities and specificities; while proteases bind with relatively high dissociation constants to their substrates (K_M ~10 uM), SH2 domains have been known to bind with dissociation constants as low as 0.3 nM (Felder et al. 1993).

The binding specificities and mechanisms for each of these domains are distinct, thereby adding to the diversity of the test set. PDZ domains bind up to 7 C-terminal residues in a highly specific manner (Tonikian et al. 2008). SH3 domains bind proline-rich regions that

often form PPII helices (Sparks et al. 1996). SH2 domains show a preference for pTyr-containing peptides (Waksman et al. 1993), while the context surrounding the pTyr residue determines the specificity of the peptide towards a distinct SH2 domain (Domchek et al. 1992). Kinases are one of the largest families in the eukaryotic genome and share a common fold that allows for the binding of ATP and a Ser, Thr, or Tyr residue-containing substrate (Ubersax & Ferrell 2007). Finally, MHC-I domains bind short pathogenic peptides to be presented to cytotoxic T lymphocytes (CTLs). MHC-I domains are promiscuous and may bind many peptides; generally, one or two substrate positions are conserved, while others are tolerant to mutations (Lundegaard et al. 2010).

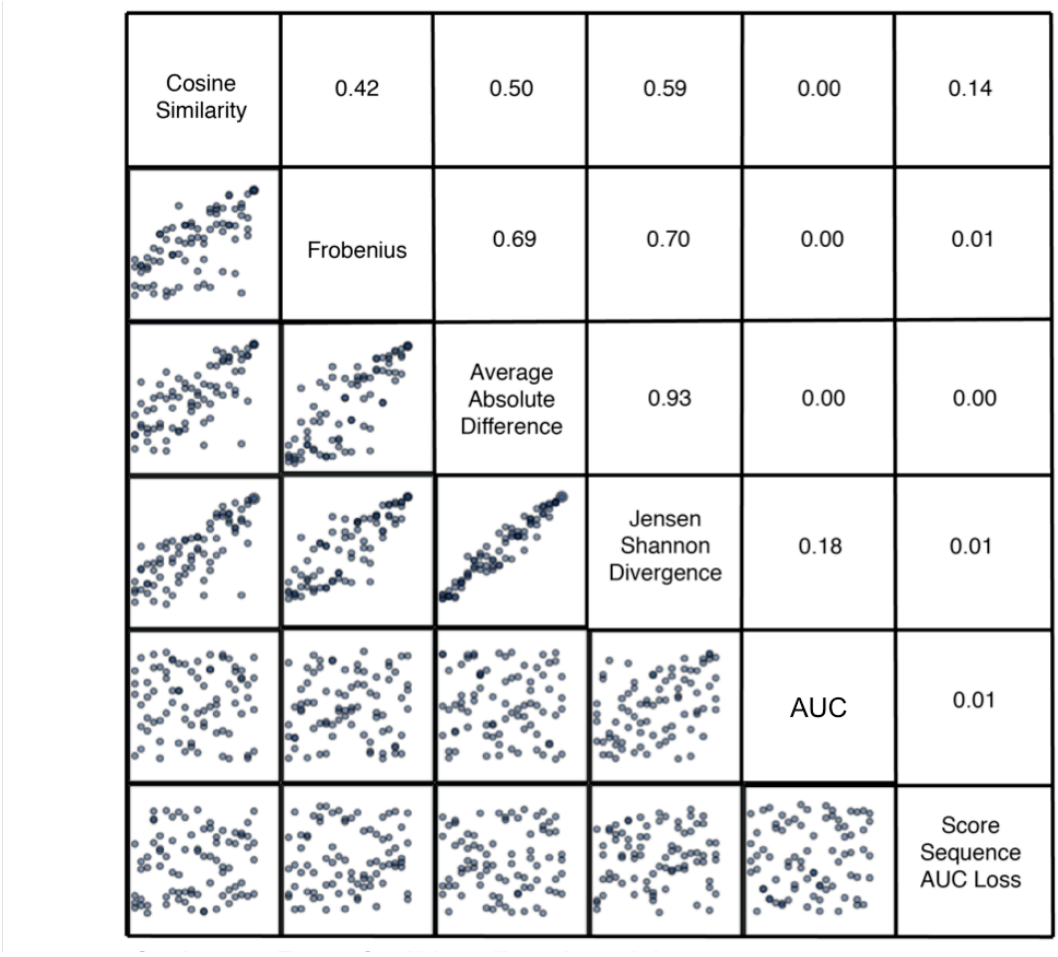


Figure 2.3. Specificity profile metric correlation

Correlation coefficients between pairs of metrics are shown in the upper diagonal while scatterplots are shown in the lower diagonal. Cosine similarities and AUC values are shown as 1 – cosine and 1- AUC, respectively, so that a lower value represents a better prediction. Scatterplot points are colored by the number of bits in the predicted profile, with a darker blue representing fewer bits, or more peaked profiles

3.3.3. Choosing Metrics for Evaluation of Prediction Accuracy

We evaluated the performance of MFPred by quantifying the differences between predicted and experimentally determined specificity profiles using several metrics (see S1 Note for detailed descriptions of these metrics). Four of these metrics, the cosine similarity, Frobenius norm, average absolute distance (AAD) and Jensen-Shannon divergence (JSD) are correlated, as shown in Figure 2.3. The Frobenius norm and AAD are distance-based metrics that have been used previously to compare profiles (Smith & Kortemme 2010; King & Bradley 2010). The Frobenius norm is more sensitive to flatness in the specificity profile than the AAD (Figure 2.4). Additionally, we evaluated the profiles by their cosine similarity, which is another distance-based metric that is less sensitive to flatness than either AAD or Frobenius norm. The Jensen-Shannon divergence (JSD) has also been used in the past to evaluate profiles (King & Bradley 2010). We used cosine distance as the general score of a profile, as it is easy to visualize and interpret. It falls between 0 and 1, where 0 denotes a random prediction and 1 denotes a perfect prediction. For each position, we evaluated the significance of its JSD score by scoring 100,000 random profiles against the experimental profile and thus determining the p-value of the JSD score (see S1 Note for details).

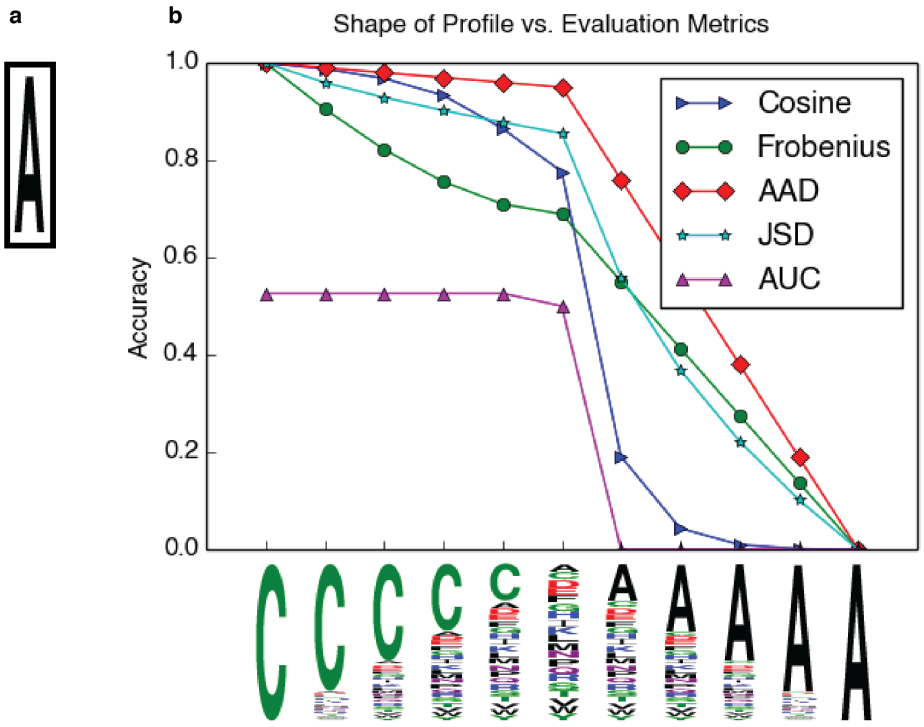


Figure 2.4. Profile shape affects evaluation metrics differently
(a) "Experimental" profile to compare to. (b) Each metric is affected differently by the shape of the profile (x- axis). Accuracy is normalized for all metrics so that the worst metric corresponds to one. Both AUC and cosine are subtracted from 1, as well. Cosine similarity varies slightly with regard to flatness of the profile, whether or not the most frequent amino acid is correct. Frobenius distance varies more than the cosine similarity; it decreases somewhat consistently with the shape of the profile. While AAD does not vary much with regard to flatness when the most frequent amino acid is incorrect, it decreases very quickly when the most frequent amino acid is correct. JSD also varies more frequent amino acid is incorrect, t is ~0.5 (or random), and if the most frequent amino acid is correct, it is zero.

We also used a second metric as a general score for each profile: area under the ROC (receiver operating characteristic) curve (AUC) is a non-distance-based metric that evaluates predictions based on their ranking more tolerated amino acids correctly (Smith & Kortemme 2010). It is relatively unaffected by flatness (Figure 2.4) but will not evaluate well if either the experimental or predicted profile is close to uniform. It is not correlated with the above metrics. Additionally, we developed a new metric, Score

Sequence AUC Loss (SSAL), which encapsulates the efficacy of the predicted specificity profile in differentiating between substrates which are recognized and cleaved by a given protease (cleaved sequences) and substrates which are not cleaved by that protease (uncleaved sequences). A perfect prediction scores an SSAL of zero. It does not correlate well with any other metric (Figure 2.3).

3.3.4. Recapitulation of protease specificity profiles

Proteolysis is a multi-step reaction, which involves substrate peptide binding, the formation of a tetrahedral intermediate (acylation) and hydrolytic cleavage of the tetrahedral intermediate (deacylation). We have previously found that modeling a near-attack conformation for the acylation step was successful in discriminating between known cleaved and uncleaved peptides (Pethe et al. 2017). Therefore, starting from structures of protease-substrate complexes in a near-attack conformation, we performed MFPred-based specificity prediction. We found that MFPred robustly recapitulates protease specificity profiles (Figure 2.5b) in our benchmark set. The cosine similarities of the entire profiles range from 0.66 to 0.89, AUC ranges from 0.73 to 0.86, and SSAL ranges from 0.21 to 0.002. Out of 31 substrate positions across the protease dataset, 20 were predicted with a significant JSD p-value. The best prediction is obtained for the common biotechnologically used protease TEV-PR. The predicted profile has a high cosine similarity of 0.89 (1 would be a perfectly accurate prediction). The primarily steric and hydrogen-bonding-based nature of molecular recognition at TEV-PR-substrate interfaces is well suited to the strengths of the Rosetta energy function underlying MFPred. Similarly, the profiles of HCV protease and granzyme B (GrB) protease are

also generally recapitulated with a high degree of accuracy, except for positions with no marked preference for specific amino acids (flat positions) – positions P5 and P2 in HCV protease and positions P4, P1', and P2' in granzyme B protease. We attribute the lack of correlation at these flat positions to small errors in energy evaluations being equivalent to the size of the energy gaps being modeled, thus leading to erroneous ranking. Challenges in measuring prediction accuracy at flat positions have indeed been noted before (Smith & Kortemme 2010).

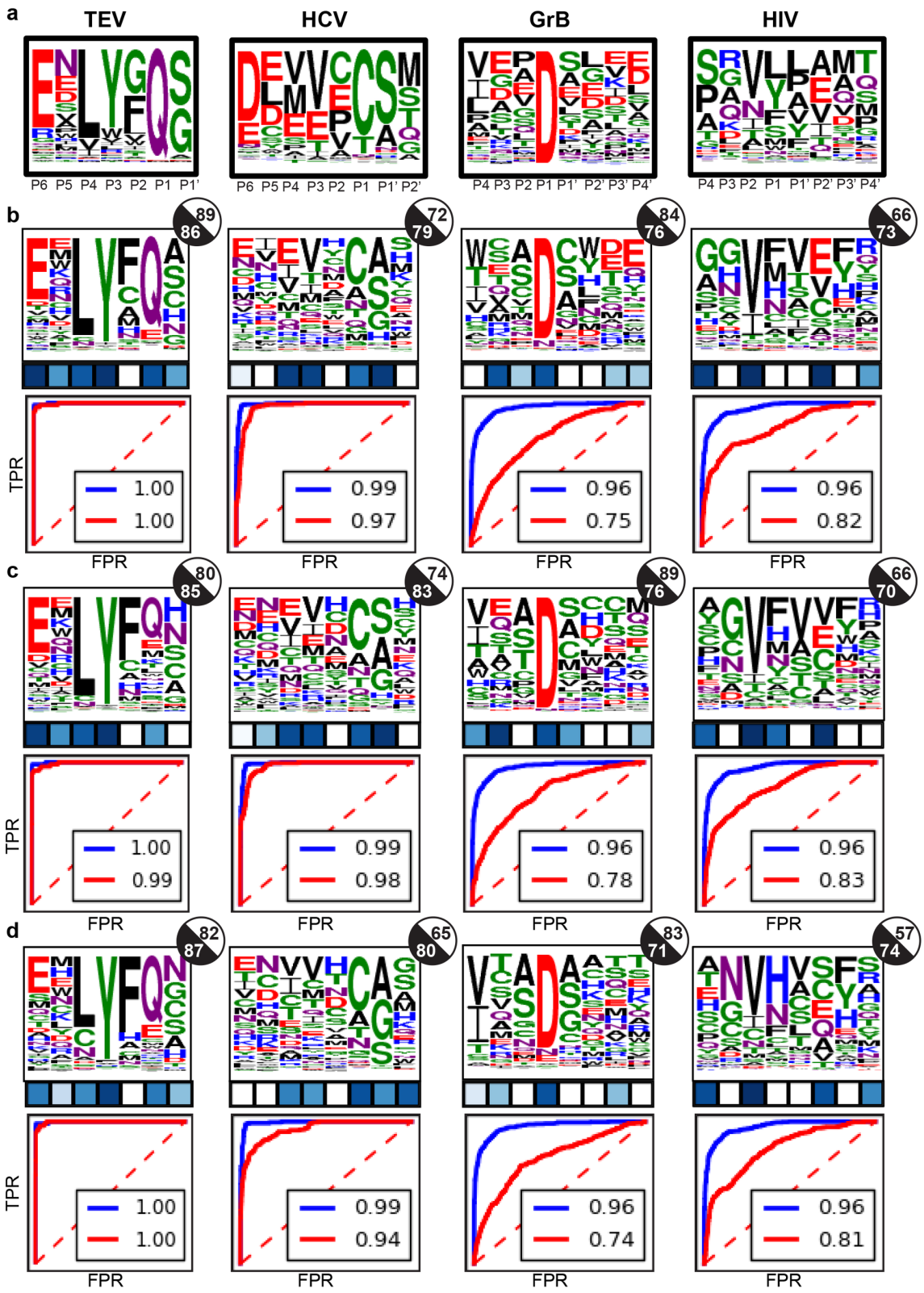


Figure 2.5. Comparison of backbone ensemble generation methods.

(a) Experimental specificity profiles. **(b)** MFPred on FastRelax backbone ensemble. The *p*-value of the JSD for a given position is represented by the color of the square under that position; white denotes a p-value > 0.5 and dark blue denotes a p-value of 0. A given circle to the right of a profile represents the cosine similarity (white) and AUC (black) of that profile. The ROC plots beneath each profile depict the SSAL calculation via the experimental ROC (blue) and predicted ROC (red) with their respective AUC values. **(c)** MFPred on FlexPepDock backbone ensemble. **(d)** MFPred on Backrub backbone ensemble.

The worst performance among the proteases in the benchmark set is observed for the prediction of HIV protease-1 (HIVPR1) specificity. This protease is known to have a relaxed specificity profile, with preference for small hydrophobic residues at P1 and P1' positions. The cavity of HIV protease-1 is large and peptides may adopt large variations in backbone conformation depending on their sidechains. Additionally, substrate binding involves flexibility on the protease side, with two loops ("flaps") that are mobile and close over the binding pocket. Incorporation of greater backbone flexibility on both the receptor and peptide parts of the HIVPR1-peptide interface may help improve predictions, as previously observed by us and others (London et al. 2011; Smith & Kortemme 2011; Pethe et al. 2017).

3.3.5. Modeling Backbone Flexibility is Key for Prediction Accuracy

To determine the contribution of modeling backbone flexibility to the accuracy of prediction and to investigate if backbone sampling could be optimized for specificity prediction, we generated MFPred profiles with different levels of backbone flexibility.

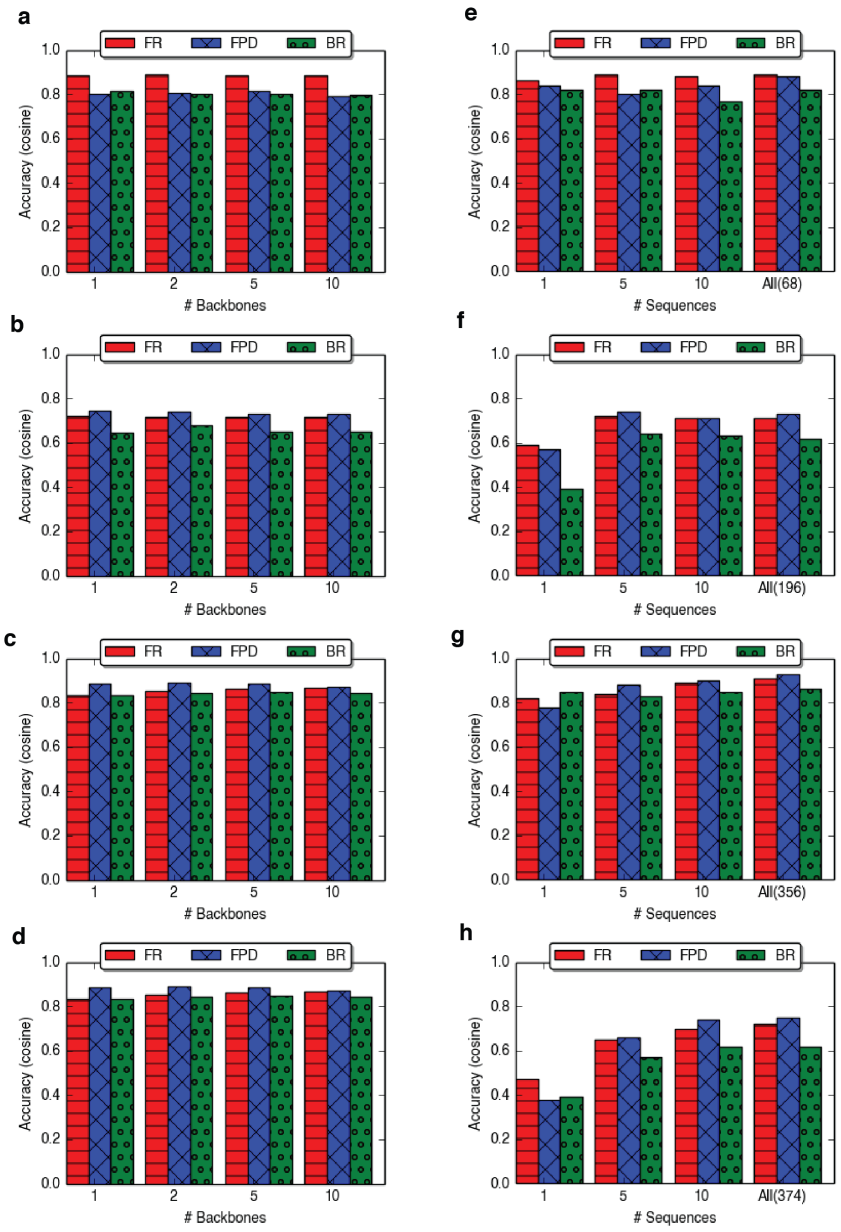


Figure 2.6. Number of sequence vs. accuracy and number of backbones vs. accuracy for methods of backbone ensemble generation

(a)- (d) Number of backbones per sequence vs. accuracy for TEV, HCV, Granzyme B and HIV, respectively. Each protocol begins with five sequences, which are then relaxed using FR, FPD or BR 1,2,5 or 10 times each. (e)-(h) Number of sequences vs. accuracy for TEV, HCV, Granzyme B and HIV, respectively. Number of sequences is varied over 1,5,10, all experimentally derived sequences, which is different for each protease.

Table 2.1. Results of all methods of backbone generation - FastRelax (FR), FlexPepDock (FPD), and backrub (BR) - on variously-sized backbone ensembles.

Protease	Method	#Seq	Cosine	Frob	AAD	JSD	AUC	SSAL	Bits
TEV	**FR**	1	0.86	1.06	0.04	0.22	0.87	0.00	0.43
		5	0.89	0.85	0.04	0.21	0.86	0.00	-0.34
		10	0.88	0.86	0.04	0.20	0.91	0.00	-0.55
		All (68)	0.89	0.84	0.03	0.20	0.91	0.00	-0.69
	FPD	1	0.84	1.08	0.04	0.23	0.86	0.00	0.23
		5	0.80	1.10	0.04	0.27	0.85	0.01	-0.64
		10	0.84	0.99	0.04	0.24	0.91	0.00	-0.64
		All (68)	0.88	0.87	0.04	0.20	0.91	0.00	-0.72
	BR	1	0.82	1.11	0.04	0.25	0.84	0.00	-0.06
		5	0.82	1.06	0.05	0.26	0.87	0.00	-0.70
		10	0.77	1.17	0.05	0.29	0.89	0.00	-0.91
		All (68)	0.82	1.06	0.05	0.27	0.89	0.00	-0.87
HCV	**FR**	1	0.59	1.37	0.06	0.35	0.77	0.08	-0.51
		5	0.72	1.13	0.05	0.31	0.79	0.02	-1.28
		10	0.71	1.15	0.05	0.30	0.82	0.02	-1.28
		All (196)	0.71	1.14	0.05	0.29	0.84	0.02	-1.29
	FPD	1	0.57	1.45	0.06	0.35	0.76	0.09	-0.39
		5	0.74	1.10	0.05	0.30	0.83	0.02	-1.29
		10	0.71	1.14	0.05	0.30	0.80	0.01	-1.29
		All (196)	0.73	1.12	0.05	0.28	0.87	0.01	-1.35
	BR	1	0.39	1.67	0.06	0.44	0.69	0.17	-0.83
		5	0.64	1.23	0.05	0.32	0.80	0.05	-1.20
		10	0.63	1.25	0.06	0.32	0.81	0.04	-1.22
		All (196)	0.62	1.26	0.05	0.32	0.81	0.05	-1.31
GrB	**FR**	1	0.82	0.85	0.04	0.23	0.71	0.20	0.60
		5	0.84	0.73	0.04	0.20	0.76	0.21	0.07
		10	0.89	0.60	0.03	0.17	0.80	0.17	0.06
		All (356)	0.91	0.53	0.03	0.13	0.87	0.15	-0.08
	FPD	1	0.78	1.04	0.04	0.25	0.72	0.19	0.83
		5	0.88	0.62	0.03	0.17	0.76	0.18	0.10
		10	0.90	0.59	0.03	0.15	0.80	0.17	0.02
		All (356)	0.93	0.49	0.03	0.11	0.83	0.13	-0.08
	BR	1	0.85	0.74	0.04	0.22	0.71	0.19	0.38
		5	0.83	0.74	0.04	0.20	0.71	0.22	0.14
		10	0.85	0.70	0.04	0.19	0.72	0.22	0.09
		All (356)	0.86	0.68	0.04	0.18	0.72	0.21	0.08
HIV	**FR**	1	0.47	1.55	0.06	0.42	0.66	0.17	0.96
		5	0.65	0.96	0.05	0.27	0.73	0.14	-0.01
		10	0.70	0.88	0.04	0.23	0.78	0.08	-0.04
		All (374)	0.72	0.82	0.04	0.21	0.81	0.05	-0.21
	FPD	1	0.38	1.78	0.07	0.47	0.69	0.22	1.22
		5	0.66	0.96	0.05	0.28	0.70	0.13	-0.04
		10	0.74	0.81	0.04	0.22	0.78	0.07	-0.18
		All (374)	0.75	0.77	0.04	0.19	0.83	0.05	-0.32
	BR	1	0.39	1.48	0.06	0.41	0.67	0.23	0.47
		5	0.57	1.06	0.05	0.30	0.74	0.15	-0.04
		10	0.62	0.98	0.05	0.27	0.73	0.14	-0.11
		All (374)	0.62	0.96	0.05	0.27	0.73	0.11	-0.16

Most Similar	1.00	0.00	0.00	0.00	1.00	0.00	0.00
Most Different	0.00	$\sqrt{(2n)}$[1]	0.06	1.00	0.00	1.00	4.32

First, we found that predictions generated by starting from a single crystallographically-determined backbone structure for the peptide led to poor accuracy for HCV and HIV proteases (panels f,h in Figure 2.6), indicating that incorporating peptide backbone diversity is a key requirement for the observed accuracy of prediction. Second, we generated peptide backbone ensembles by threading on a varying number of known substrate (cleaved) peptides using three different Rosetta-based backbone sampling protocols (FastRelax (Tyka et al. 2011), FlexPepDock (Raveh et al. 2010), and Backrub (Smith & Kortemme 2008)) separately to further diversify the peptide backbone ensemble. In each case, geometric constraints (Pethe et al. 2017) were used to limit the scissile peptide bond to a near-attack conformation and the catalytic residues to an active conformation. The MFPred simulations were then performed on all backbone ensembles and their results were compared to each other (Figure 2.5, Table 2.1).

While the algorithm is relatively robust to the method of backbone generation as long as scissile bond geometry is maintained, the FastRelax (FR) protocol has a small improvement in overall performance over the FlexPepDock (FPD) protocol, with 20 significant p-values (out of 31) for FR vs. 19 for FPD, and FPD has a minor increase in overall performance over Backrub (BR), with 19 significant p-values for FPD vs. 18 for BR. The profile for TEV-PR is predicted best by FR, due to better prediction of Q at P1 and S at P1'. In the case of HIV protease-1, FR recapitulates the profile better than FPD and BR do. However, the performance of FPD is marginally better than that of FR and

significantly more accurate than that of BR in the cases of HCV protease and granzyme B protease.

To determine how MFPred accuracy depends on the number and sequences of known cleaved substrates used to generate the backbone ensemble, we generated a peptide backbone conformational ensemble that was independent of peptide sequence. For all positions on the peptide backbone, we enumerated every combination of phi/psi dihedral angles that were x-15, x, and x+15, where x is the dihedral angle of the relaxed crystal structure peptide backbone. The resulting structures were filtered to remove those with clashes and to preserve hydrogen-bond interactions. The remaining structures were further clustered by all-heavy-atom RMSD of the peptide residues (see S2 Note for details) and MFPred was performed on the cluster centers. The resulting predictions are significantly less accurate than those of FR, FPD, or BR (Figure 2.7), indicating that successful prediction requires a backbone ensemble that is optimally positioned in the binding site for cleavage.

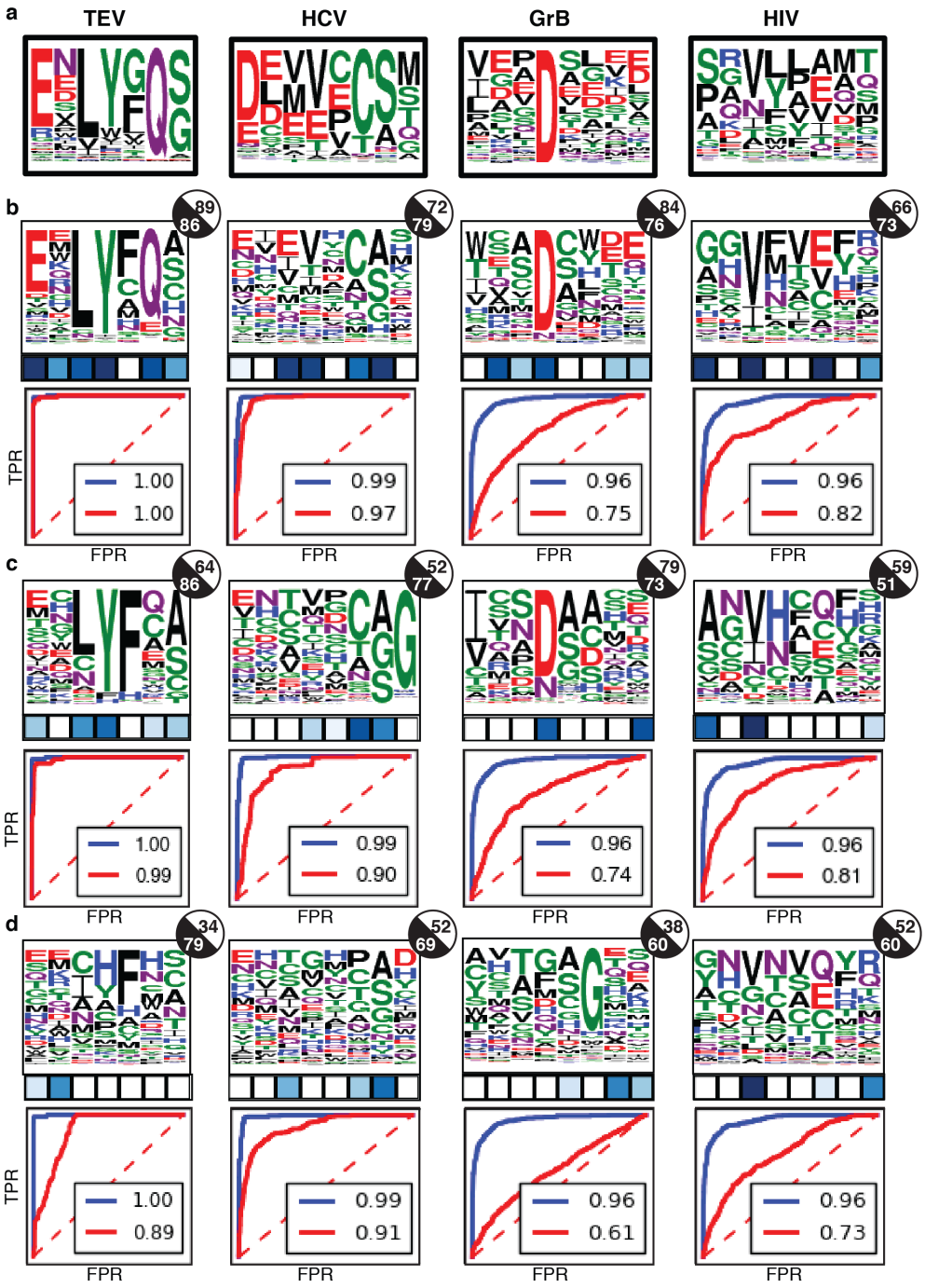


Figure 2.7. Incorporating cleaved sequences into backbone ensemble generation improves MFPred's accuracy.
(a) Experimental specificity profiles (b) Results of running MFPred on backbone ensemble of five cleaved sequences FastRelaxed (c) Results of running MFPred on backbone ensemble generated by enumerating combinations of phi/psi angles. (d) Results of running MFPred on backbone ensemble of five uncleaved sequences FastRelaxed.

As a second test of the dependence of MFPred on the cleaved sequence information, we threaded five known uncleaved (i.e., not bound by the protease in a productive conformation) sequences on the peptide backbone and then performed FastRelax on the resulting structures. The prediction accuracy of MFPred decreased on these structures (Figure 2.7), to the extent that the specificity profiles are almost uniform. Therefore, diversifying the peptide structure in suboptimal sequence space led to worse predictions than those obtained while diversifying it without any sequence information.

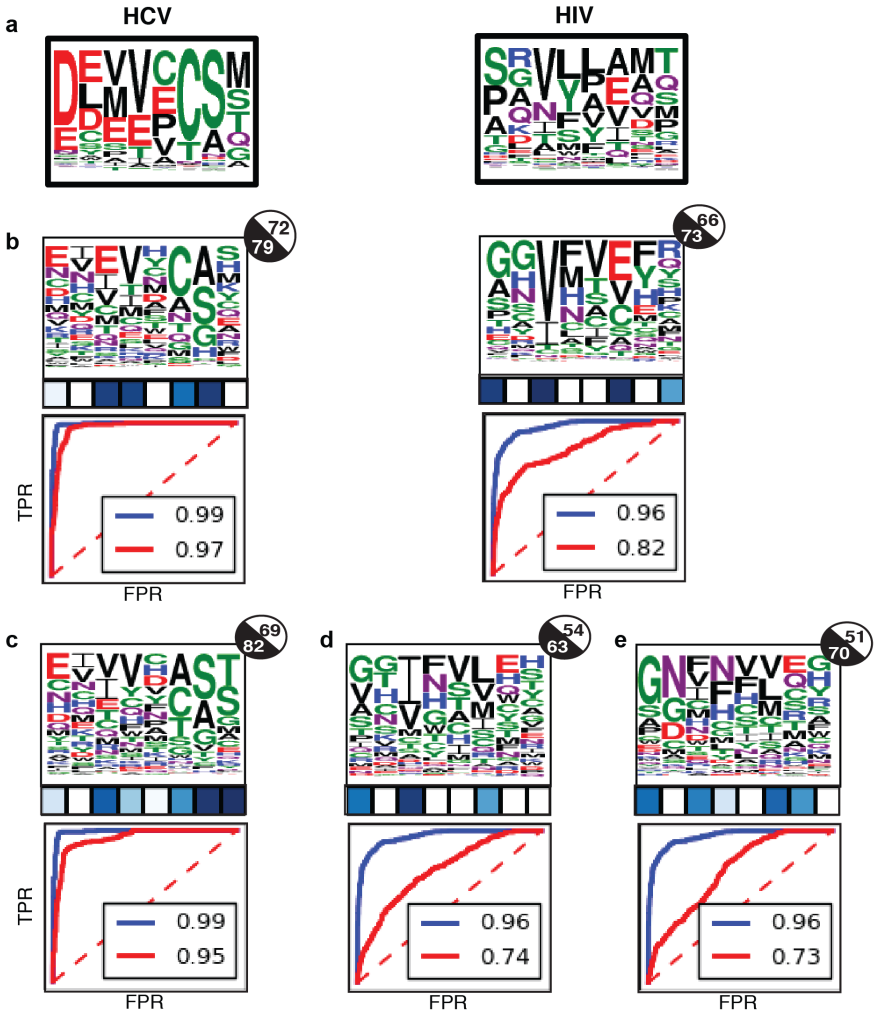


Figure 2.8. Using structures of receptor peptide complexes vs. apo structures improves the accuracy of MFPred.

(a) Experimental specificity profiles. (b) MFPred prediction on receptor – peptide complexes. (c) MFPred prediction on HCV NS3 Protease apo structure(PDB 3KF2) (d) MFPred prediction on HIV protease 1 closed form apo structure (PDB: 2HB4). (e) MFPred prediction on HIV protease 1 open form apo structure (PDB: 1PCO)

Next, to determine the impact of starting from bound complexes to generate MFPred predictions, we performed MFPred simulations on apo structures of two proteases: HCV NS3/4A protease and HIV protease-1 (Figure 2.8). As HIV protease-1 has two flaps that can assume either a closed or open form (Heaslet et al. 2007), we used both a 'closed apo' structure and an 'open apo' structure for our simulations. In each case the protease all-atom RMSD between bound and open states, as determined by PyMol (Anon n.d.), were 1.04 Å, 1.85 Å, and 2.00 Å. In all three cases, MFPred accuracy was higher when starting from the bound complex compared to the apo state. While the number of significant p-values remains similar, the overall cosine similarities, AUC, and SSAL decreased for the apo structure-based simulations. Additionally, the information content decreased significantly for the apo structures of HIV (0.72-0.74 bits) as opposed to the bound complex (1.18 bits). Overall, the prediction accuracies between apo and bound states were more similar for the HCV protease where small backbone changes in the protease are incurred upon binding, compared to HIV protease where larger differences in prediction accuracy were apparent. These results suggest that especially in cases where there is significant backbone conformational change in the receptor upon peptide binding, such as the HIV protease, the incorporation of receptor flexibility may be needed for maintaining MFPred accuracy.

Finally, to investigate the dependence of performance accuracy on the number of known cleaved (recognized) sequences, we executed MFPred simulations on backbone

ensembles generated from differing numbers of starting peptide sequences threaded on to the crystallographic backbone conformation. We varied the number of sequences used to generate the backbone ensemble from one sequence to five sequences to ten sequences to all known sequences in the benchmark set. We found that MFPred is highly dependent on N, the number of cleaved sequences used, when N is small (panels e-h in Figure 2.6). However, as N increases, this effect is decreased. For TEV-PR and HCV protease, which have relatively few sequences (68 and 198 respectively), the prediction accuracy plateaus after ten sequences, although in some cases it may fluctuate slightly from five to ten to all sequences. However, for granzyme B and HIV proteases (356 and 374 cleaved sequences respectively), the accuracy of MFPred has a minor increase from ten to all sequences. Thus, there is a near-maximum of accuracy for each system; once that point of diminishing returns has been reached, incorporating more cleaved sequences does not lead to significant increases in the accuracy.

Table 2.2: Effect of various Rosetta settings on MFPred predictions on five sequence backbones.

Protease	Method	Cosine	Frob	AAD	JSD	AUC	SSAL	Bits
TEV	**Current**	0.89	0.85	0.04	0.21	0.86	0.00	-0.34
	Dun02	0.86	0.97	0.04	0.24	0.86	0.00	-0.14
	Ex1aro,ex2aro	0.89	0.85	0.04	0.21	0.86	0.00	-0.34
	Ex3,ex4	0.88	0.87	0.04	0.22	0.86	0.00	-0.38
	No input sc	0.88	0.88	0.04	0.22	0.86	0.00	-0.46
	Pack prot 4	0.81	1.07	0.04	0.25	0.90	0.00	-0.56
	Pack prot 6	0.81	1.07	0.04	0.25	0.91	0.00	-0.59
	Pack prot 8	0.81	1.07	0.04	0.25	0.91	0.00	-0.60
HCV	**Current**	0.72	1.13	0.05	0.31	0.79	0.02	-1.28
	Dun02	0.64	1.24	0.06	0.35	0.78	0.02	-1.19
	Ex1aro,ex2aro	0.72	1.13	0.05	0.31	0.79	0.02	-1.28
	Ex3,ex4	0.71	1.14	0.05	0.31	0.77	0.03	-1.27
	No input sc	0.71	1.15	0.05	0.31	0.78	0.02	-1.29
	Pack prot 4	0.67	1.20	0.06	0.33	0.73	0.04	-1.21
	Pack prot 6	0.67	1.20	0.06	0.33	0.74	0.04	-1.21
	Pack prot 8	0.67	1.20	0.06	0.33	0.74	0.04	-1.20
GrB	**Current**	0.84	0.73	0.04	0.20	0.76	0.21	0.07
	Dun02	0.82	0.78	0.04	0.23	0.79	0.22	0.21
	Ex1aro,ex2aro	0.84	0.73	0.04	0.20	0.76	0.21	0.07
	Ex3,ex4	0.84	0.73	0.04	0.20	0.76	0.21	0.08
	No input sc	0.84	0.73	0.04	0.20	0.75	0.22	0.06
	Pack prot 4	0.81	0.80	0.04	0.23	0.77	0.25	0.22
	Pack prot 6	0.80	0.82	0.04	0.23	0.75	0.26	0.18
	Pack prot 8	0.81	0.80	0.04	0.23	0.76	0.25	0.21
HIV	**Current**	0.65	0.96	0.05	0.27	0.73	0.14	-0.01
	Dun02	0.59	1.08	0.05	0.32	0.68	0.14	0.10
	Ex1aro,ex2aro	0.65	0.96	0.05	0.27	0.73	0.14	-0.01
	Ex3,ex4	0.65	0.97	0.05	0.27	0.71	0.14	-0.01
	No input sc	0.63	0.98	0.05	0.28	0.70	0.15	-0.06
	Pack prot 4	0.63	1.01	0.05	0.30	0.71	0.14	0.08
	Pack prot 6	0.61	1.04	0.05	0.32	0.71	0.15	0.11
	Pack prot 8	0.60	1.05	0.05	0.31	0.69	0.15	0.05
Most Similar		1.00	0.00	0.00	0.00	1.00	0.00	0.00
Most Different		0.00	$\sqrt{(2n)}$[1]	0.06	1.00	0.00	1.00	4.32

Besides determining that the level of backbone sampling was optimal for prediction, we also optimized sidechain sampling (Table 2.2). Using an older version of the rotamer library (2002) (Dunbrack 2002) decreased scores for all systems. Increasing the fineness

of rotamer chi-angle sampling or removing the starting sidechain conformation from the rotamer sampling had little impact on the results. Packing protease sidechains around the peptide (between distances of 4-8 Angstroms) decreased the accuracy of the results. This may be explained by the finding that hot spot residues at protein-protein interfaces often adopt strained rotamer configurations (Watkins et al. 2016); packing protease interface sidechains while designing peptide residues within MFPred may force protease sidechains to adopt conformations that are unfavorable for productive substrate binding.

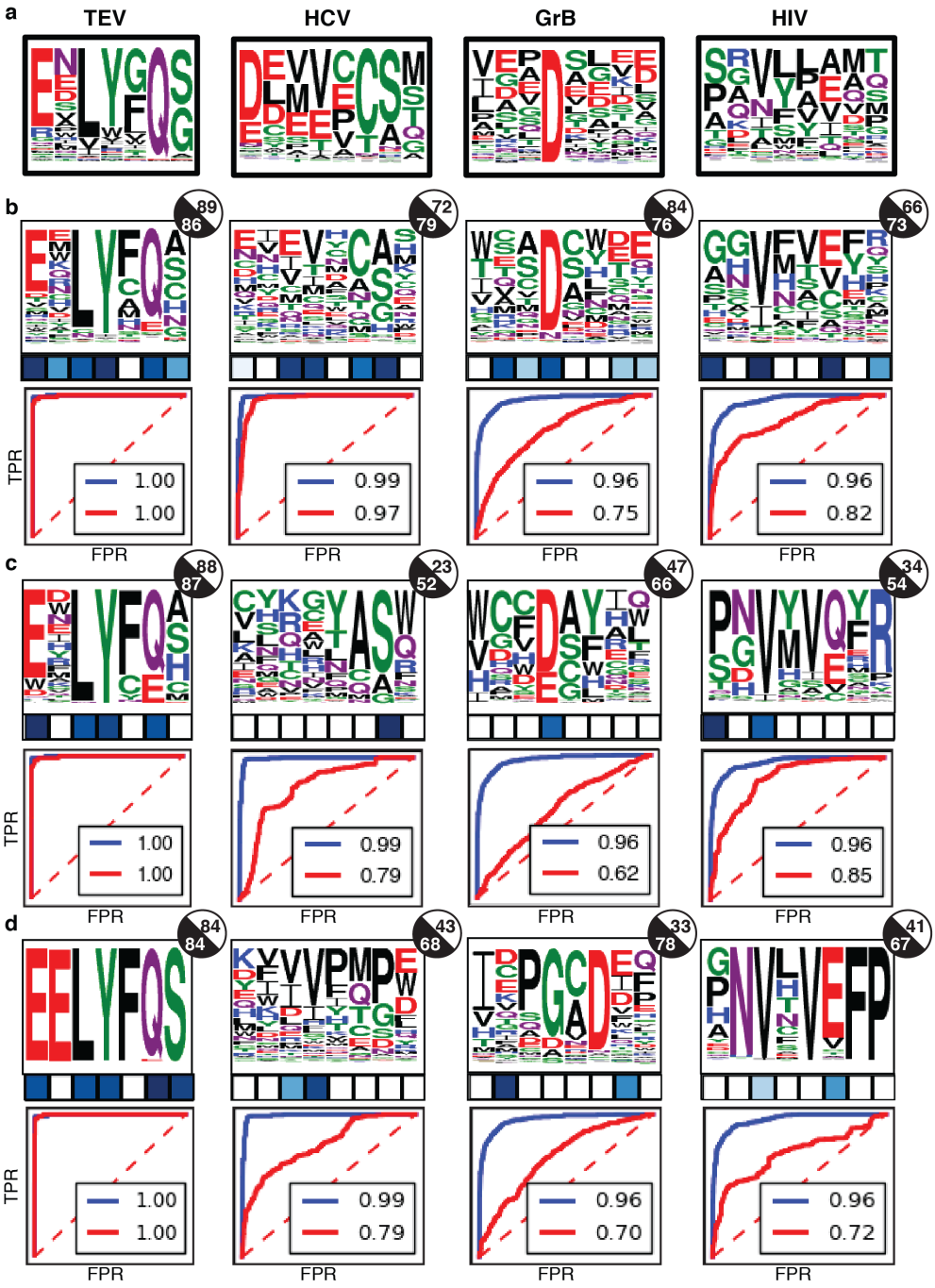


Figure 2.9. MFPred vs. other Rosetta prediction techniques on ensemble of five sequences.
(a) Experimental specificity profiles. (b) MFPred. (c) pepspec. (d) sequence_tolerance.

3.3.6. Comparison of MFPred with Other Structure-Based Approaches

Table 2.3. Results of all methods - MFPred (MF), sequence_tolerance (ST), and pepspec (PS) - on variously-sized backbone ensembles.

Prote	Metho	#Seq	Time(m)	Cosi	Fro	AA	JS	AU	SSA	Bit
TEV	**MF**	1	0.18	0.86	1.06	0.0	0.2	0.8	0.00	0.4
		5	0.80	0.89	0.85	0.0	0.2	0.8	0.00	-
		10	2.08	0.88	0.86	0.0	0.2	0.9	0.00	-
		All (68)	11.97	0.89	0.84	0.0	0.2	0.9	0.00	-
	ST	1	195.65	0.84	1.49	0.0	0.2	0.8	0.00	1.8
		5	923.91	0.84	1.49	0.0	0.2	0.8	0.00	1.7
		10	1827.32	0.84	1.49	0.0	0.2	0.8	0.00	1.8
		All (68)	12333.94	0.84	1.44	0.0	0.2	0.8	0.00	1.6
	PS	1	17.46	0.72	1.50	0.0	0.3	0.8	0.01	0.8
		5	96.01	0.85	1.06	0.0	0.2	0.9	0.00	0.4
		10	189.43	0.82	1.17	0.0	0.2	0.8	0.00	0.3
		All (68)	1290.41	0.86	1.04	0.0	0.2	0.8	0.00	0.2
HCV	**MF**	1	0.68	0.59	1.37	0.0	0.3	0.7	0.08	-
		5	3.61	0.72	1.13	0.0	0.3	0.7	0.02	-
		10	7.14	0.71	1.15	0.0	0.3	0.8	0.02	-
		All (196)	132.15	0.71	1.14	0.0	0.2	0.8	0.02	-
	ST	1	115.04	0.30	1.77	0.0	0.5	0.6	0.30	-
		5	574.01	0.43	1.54	0.0	0.4	0.6	0.21	-
		10	1101.15	0.44	1.49	0.0	0.4	0.7	0.17	-
		All (196)	22239.05	0.43	1.51	0.0	0.4	0.6	0.17	-
	PS	1	17.78	0.24	2.19	0.0	0.6	0.6	0.34	0.6
		5	91.68	0.37	1.69	0.0	0.5	0.5	0.20	-
		10	171.30	0.61	1.30	0.0	0.3	0.7	0.05	-
		All (196)	3462.64	0.63	1.26	0.0	0.3	0.7	0.05	-
GrB	**MF**	1	0.34	0.82	0.85	0.0	0.2	0.7	0.20	0.6
		5	2.39	0.84	0.73	0.0	0.2	0.7	0.21	0.0
		10	5.24	0.89	0.60	0.0	0.1	0.8	0.17	0.0
		All (356)	145.63	0.91	0.53	0.0	0.1	0.8	0.15	-
	ST	1	114.80	0.28	2.02	0.0	0.4	0.7	0.26	1.2
		5	544.28	0.33	1.71	0.0	0.3	0.7	0.26	0.6
		10	1109.45	0.35	1.62	0.0	0.3	0.8	0.17	0.5
		All (356)	39036.17	0.34	1.67	0.0	0.3	0.8	0.21	0.5
	PS	1	19.58	0.62	1.45	0.0	0.5	0.6	0.38	1.5
		5	101.24	0.63	1.15	0.0	0.3	0.7	0.34	0.6
		10	203.69	0.76	0.99	0.0	0.2	0.7	0.27	0.6
		All (356)	6814.15	0.88	0.64	0.0	0.1	0.8	0.18	0.1
HIV	**MF**	1	0.23	0.47	1.55	0.0	0.4	0.6	0.17	0.9
		5	1.29	0.65	0.96	0.0	0.2	0.7	0.14	-
		10	3.15	0.70	0.88	0.0	0.2	0.7	0.08	-
		All (374)	110.65	0.72	0.82	0.0	0.2	0.8	0.05	-
	ST	1	92.37	0.40	2.48	0.0	0.6	0.6	0.19	2.7
		5	453.18	0.41	2.20	0.0	0.5	0.6	0.24	2.1
		10	907.90	0.45	2.05	0.0	0.5	0.7	0.16	1.9
		All (374)	34090.45	0.48	1.81	0.0	0.4	0.7	0.14	1.3
	PS	1	23.05	0.37	2.13	0.0	0.6	0.5	0.22	2.0
		5	109.77	0.55	1.54	0.0	0.4	0.6	0.11	1.2
		10	218.41	0.53	1.51	0.0	0.3	0.7	0.16	1.0
		All (374)	8134.56	0.57	1.23	0.0	0.2	0.7	0.10	0.3

Most Similar	1.00	0.00	0.0	0.0	1.0	0.00	0.0
Most Different	0.00	√(2n	0.0	1.0	0.0	1.00	4.3

[1]n refers to the number of positions in the profile

We compared our results to the two previously developed methods for specificity prediction that have been implemented in the Rosetta software. MFPred performed with comparable or greater accuracy than the sequence_tolerance (Smith & Kortemme 2010) and pepspec (King & Bradley 2010) methods (Table 2.3). Additionally, MFPred was between 23-fold to 120-fold faster than the pepspec method and between 154-fold to 1154-fold faster than the sequence_tolerance method, depending on the number of peptide backbone conformations and rotamers (Table 2.3). Furthermore, MFPred is more accurate on single backbones and smaller backbone ensembles than the other two methods; when performed on a backbone ensemble generated from five substrate sequences, MFPred predicts 19 out of 31 positions with a significant *p*-value, whereas only 11 of the positions predicted by sequence_tolerance and 8 of the positions predicted by pepspec yield significant p-values (Figure 2.9). When executed on a single backbone conformation, MFPred predicts 12 positions with a significant p-value, while both sequence_tolerance and pepspec predict only 8 positions with a significant p-value. Both sequence_tolerance and pepspec are designed to be used with larger peptide ensembles – their success is dependent on a diverse backbone ensemble – and, as expected, their prediction accuracy increases as the number of backbones in the ensemble rises (Figure 2.10a-d), with sequence_tolerance predicting 15 significant positions and pepspec predicting 16 significant positions on the backbone ensemble generated from all cleaved sequences (Figure 2.11). When performed on this expanded backbone ensemble, MFPred prediction accuracy was also higher, with 25 significant predictions. Thus,

compared to two state-of-the-art existing methods, MFPred-based predictions are of comparable or higher accuracy, and can be obtained with 10-1000-fold higher computational efficiency.

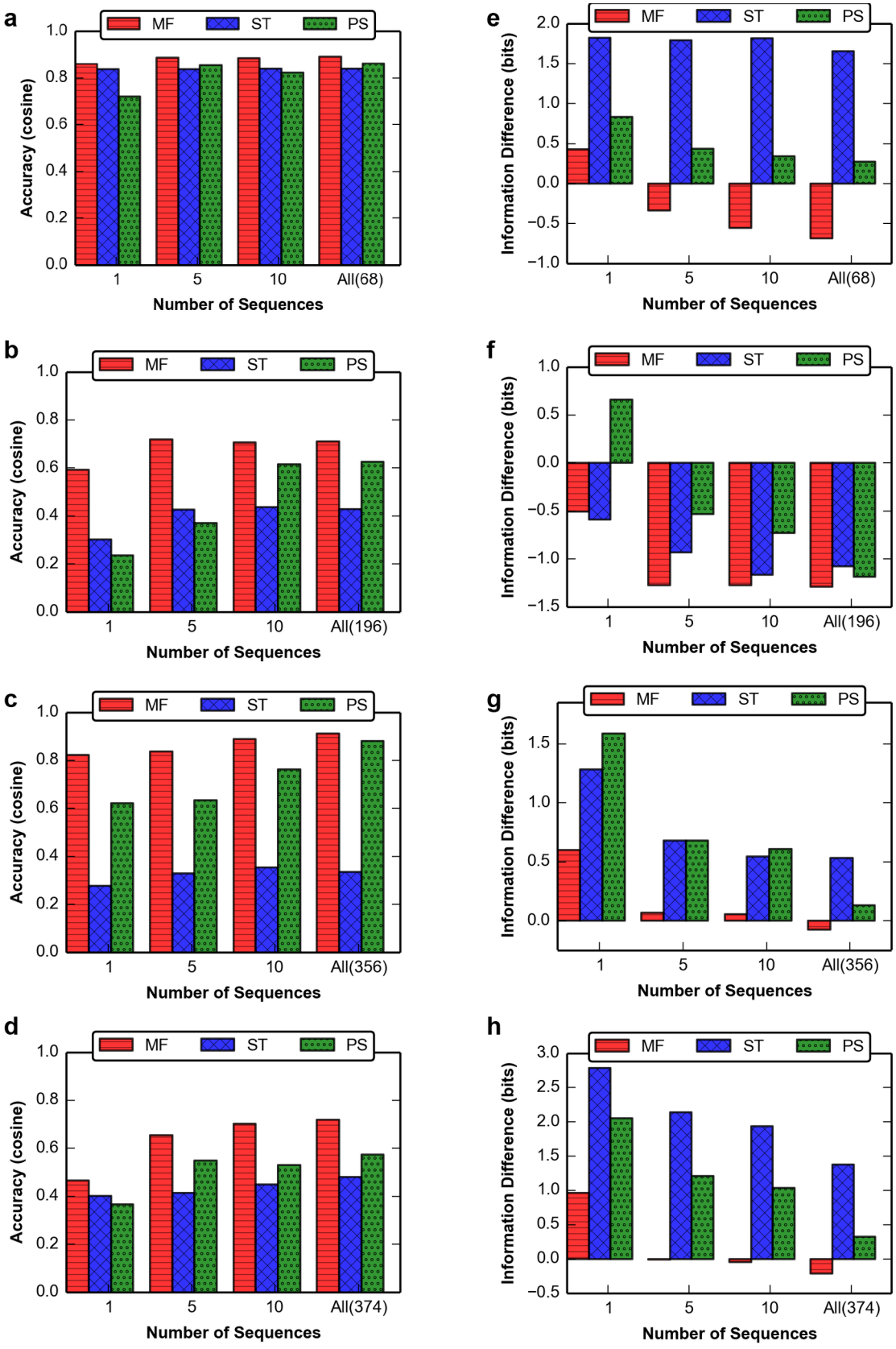

a
MF
ST
PS
Accuracy (cosine)
1.0
0.8
0.6
0.4
0.2
0.0
1
5
10
All(68)
Number of Sequences
e
MF
ST
PS
Information Difference (bits)
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
1
5
10
All(68)
Number of Sequences
b
MF
ST
PS
Accuracy (cosine)
1.0
0.8
0.6
0.4
0.2
0.0
1
5
10
All(196)
Number of Sequences
f
MF
ST
PS
Information Difference (bits)
1.0
0.5
0.0
−0.5
−1.0
−1.5
1
5
10
All(196)
Number of Sequences
c
MF
ST
PS
Accuracy (cosine)
1.0
0.8
0.6
0.4
0.2
0.0
1
5
10
All(356)
Number of Sequences
g
MF
ST
PS
Information Difference (bits)
1.5
1.0
0.5
0.0
1
5
10
All(356)
Number of Sequences
d
MF
ST
PS
Accuracy (cosine)
1.0
0.8
0.6
0.4
0.2
0.0
1
5
10
All(374)
Number of Sequences
h
MF
ST
PS
Information Difference (bits)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
1
5
10
All(374)
Number of Sequences

Figure 2.10. Number of sequences vs. accuracy and information for methods of profile prediction

(a)-(d) Number of sequences vs. accuracy for TEV, HCV, GrB, and HIV, respectively. Number of sequences is varied over 1-5-10-All experimentally derived sequences, which is different for each protease. **(e)-(h)** Number of sequences vs. information content (i.e. shape of profile) difference for TEV, HCV, GrB, and HIV, respectively. Information difference is equal to the predicted bits minus the experimental bits. An information difference that is close to zero approximates the experimental information content well; a highly positive information difference indicates a more peaked predicted than experimental profile while a highly negative information difference denotes a flatter predicted than experimental profile.

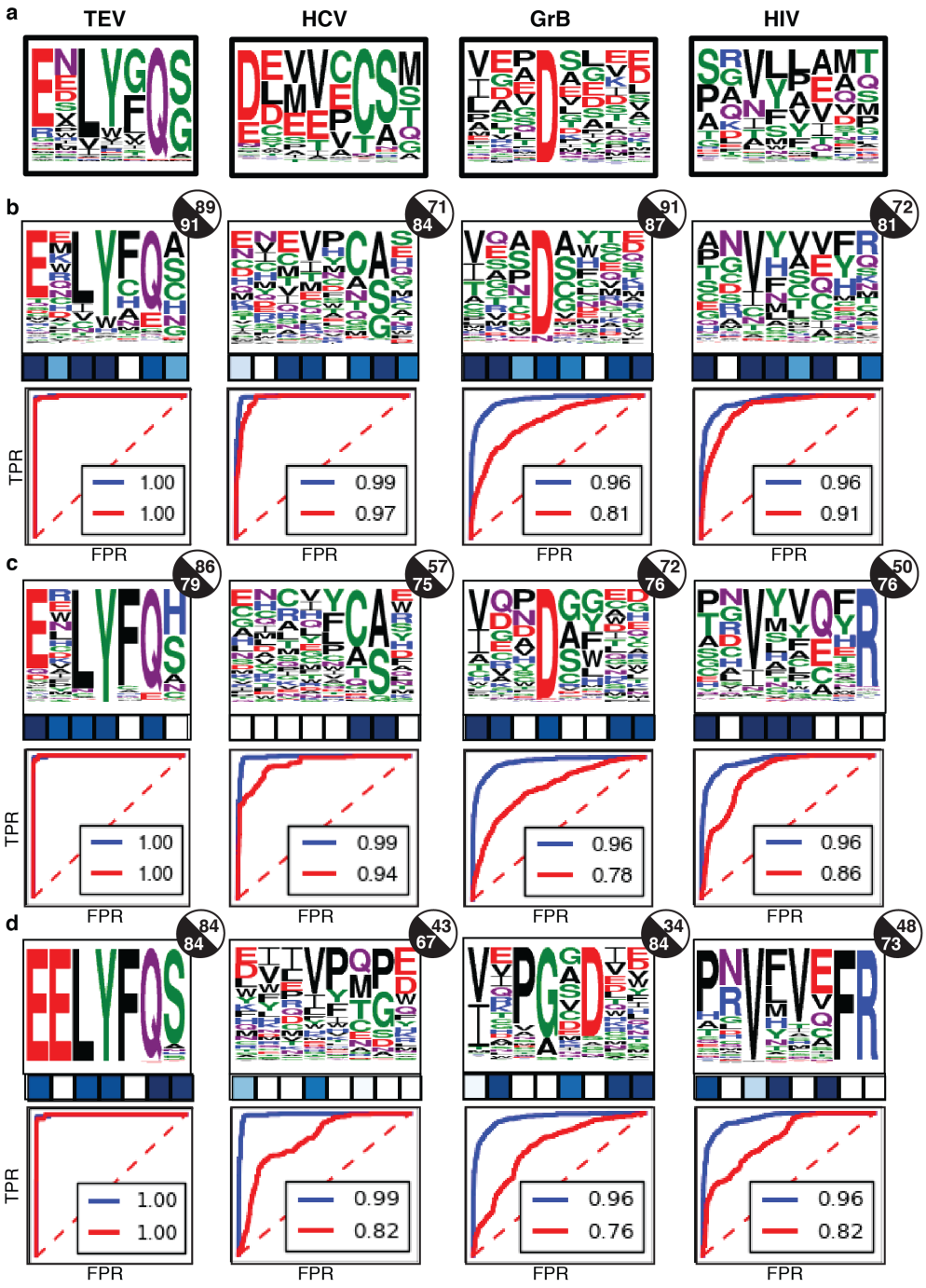


Figure 2.11. MFPred vs. other Rosetta prediction techniques on ensemble of all sequences.
(a) Experimental specificity profiles. (b) MFPred. (c) pepspec. (d) sequence_tolerance.

Besides informing us about the accuracy and speed of MFPred relative to existing methods, the comparison of MFPred to pepspec and sequence_tolerance allows us to categorize inaccuracies in MFPred predictions into those obtained from incorrect sequence sampling and those due to the Rosetta energy function or incomplete backbone conformational diversity. For example, MFPred on all cleaved backbones does not recover the experimentally determined high frequency for G at P2 of TEV-PR. Since both pepspec and sequence_tolerance also do not recover G at P2 with the same peptide backbone conformational ensemble, we attribute this inaccuracy to imperfections in the underlying Rosetta energy function and/or an incomplete peptide backbone ensemble used for prediction.

Generally, MFPred predicts lower information content (i.e. flatter shape) for the profiles than both sequence_tolerance and pepspec (Table 2.3, Figure 2.10e-h). In the cases of granzyme B protease and HIVPR1, the predicted lower information content is reflective of the experimentally determined profiles; however, in the case of TEV-PR MFPred underestimates the information content relative to pepspec and sequence_tolerance. All protocols underestimate the information content of the profile of HCV protease. This underestimation may be due to an incomplete experimental dataset or sampling/scoring inaccuracies as discussed above. Overall, the difference between the predicted information content and the experimental information content was smaller for MFPred than for sequence_tolerance and pepspec, especially when performed with smaller backbone ensembles.

3.3.7. Generalizing MFPred to other Protein-Recognition Domains

To investigate the generality of our method for specificity prediction, we utilized the MFPred method to predict the specificity profiles for a variety of peptide-recognition domains: kinase, SH2, SH3, PDZ, and MHC domains. We achieved 17 significant p-values out of 31 positions and high cosine similarities (0.77-0.85) for three out of five PRD classes: PKA (kinase), Src (SH2), and c-Crk (SH3) domains (Figure 2.12). However, these three systems had lower AUCs (0.60-0.65). This may be due to the inadequacy of AUC as a metric for scoring positions that have low information content in the experimentally-derived profile; if few of the experimental amino acid frequencies are greater than 10%, the AUC reveals little about the prediction accuracy.

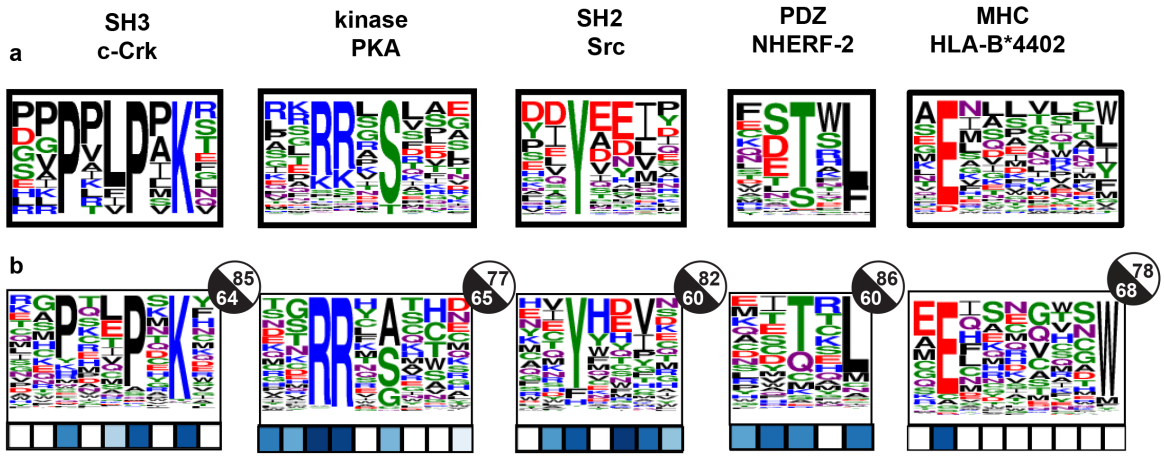


Figure 2.12: Generalize MFPred to PRD benchmark.
(a) Experimental specificity profiles. **(b)** MFPred prediction. The *p*-value of the JSD for a given position is represented by the color of the square under that position; white denotes a p-value > 0.5 and dark blue denotes a p-value of 0. A given circle to the right of a profile represents the cosine similarity (white) and AUC (black) of that profile. For the PDZ domain, prediction was performed at a kT of 0.6, which was found to be optimal for PDZ domains.

We predicted the specificity profiles of seven different PDZ domains: NHERF-2 PDZ2, PSD-95, AF-6 PDZ, Erbin PDZ, MPDZ-13, ZO-1 PDZ1, and DLG1-2 PDZ (Figure 2.12,

Figure 2.13). The specificity of NHERF-2 PDZ-2 was already predicted computationally by Zheng et al. (Zheng et al. 2015), who were able to achieve good prediction via the use of CLASSY and FlexPepDock. King and Bradley previously predicted the specificity profile for PSD-95 computationally using pepspec (King & Bradley 2010), while the five other PDZ domain specificities were previously predicted by Smith and Kortemme via sequence_tolerance (Smith & Kortemme 2010). Six out of seven PDZ domains were predicted with medium to high accuracies, with cosine similarities of 0.63-0.86, AUCs of 0.60 to 0.88, and 25 out of 38 significant p-values. However, the prediction accuracy of the final PDZ domain, AF-6 PDZ was much lower, with a cosine similarity of 0.43, AUC of 0.59, and no significant p-values. This low accuracy may be due to the flexibility of the AF-6 PDZ domain, which has been known to bind in multiple binding modes and can be characterized as belonging to multiple classes of PDZ domain specificity (Chen et al. 2007; Fujiwara et al. 2015). Similar to the HIVPR1 case above, addition of receptor flexibility to MFPred may assist in AF-6 specificity profile recapitulation.

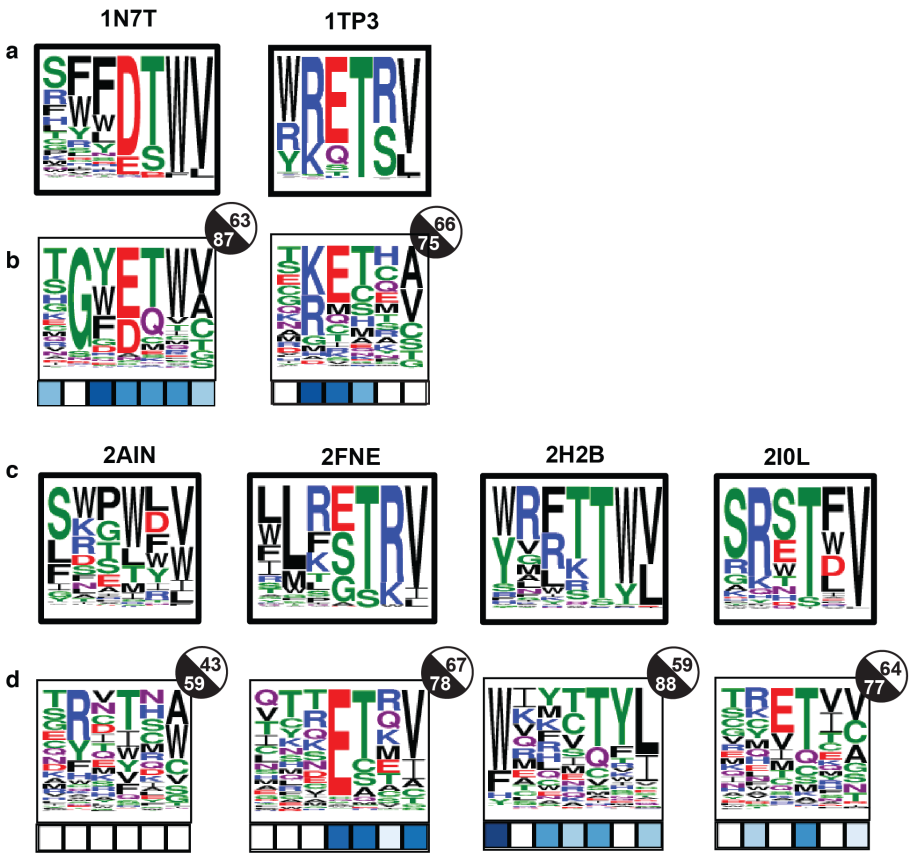


Figure 2.13. MFPred prediction for six PDZ domains.
(a,c) Experimental specificity profiles. (b,d) MFPred prediction. Prediction was performed at a kT of 0.6, which was found to be optimal for PDZ domains.

Finally, we tested the performance of MFPred on predicting MHC-I peptide recognition specificities. We selected four MHC-I domains with crystallographic structure availability and a large pool of known peptide binders (Vita et al. 2015). The experimentally derived specificity profiles for the MHCs were highly conserved at one or two positions but relatively flat at others (Figure 2.12, Figure 2.14). The MFPred predictions reflected this pattern: while 30 out of 36 positions had p-values that were not significant, due to the high tolerance of a diversity of amino acid at those positions, the cosine similarity of the predictions was high (0.63-0.78), reflecting good overall profile

recapitulation (Figure 2.12, Figure 2.14). These results indicate that robust and accurate predictions of the specificity profiles of a variety of peptide-recognition domains can be obtained using the MFPred approach, pointing to its wide applicability, especially for cases where receptor backbone flexibility is minimal. Improved modeling of backbone conformational diversity, an area where methodological improvements are needed (Khare & Fleishman 2013), is likely to improve prediction accuracy further.

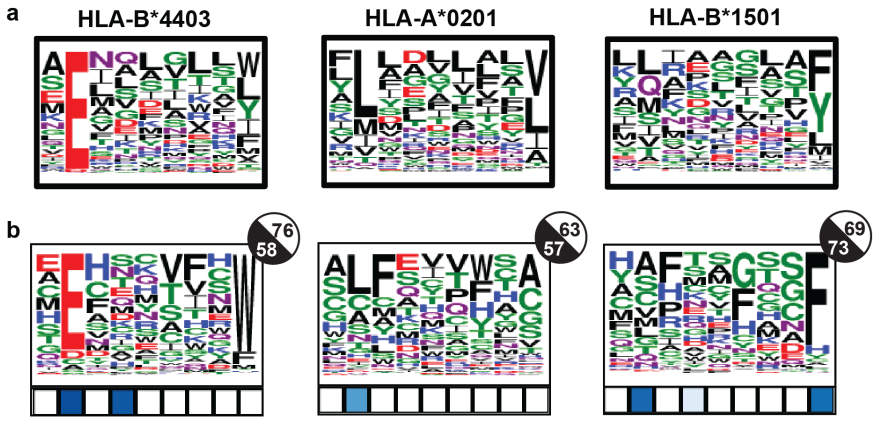


Figure 2.14. MFPred prediction for three MHC-I domains.
(a) Experimental specificity profiles. (b) MFPred prediction.

Prediction of changes in multispecificity upon receptor mutation

When used to design receptors for and against specificity profiles, MFPred should be able to accurately recapitulate changes in specificity profiles due to protease mutations, when simulations are performed on a constant set of backbones. As a proof of concept, we predicted the changes in the specificity profiles of two variants of granzyme B protease for which altered multispecificity has been experimentally determined (Figure 2.15). R192E granzyme B protease and R192E/N218A granzyme B protease have been shown to have decreased specificity for glutamic acid and increased specificity for lysine and arginine at P3 (Harris et al. 1998; Ruggles et al. 2004). To investigate whether MFPred

can recapitulate mutant specificity profiles without changing the peptide backbone, we modeled the variants of granzyme B protease by performing the necessary mutations in Rosetta on the five FastRelaxed granzyme B protease backbones.

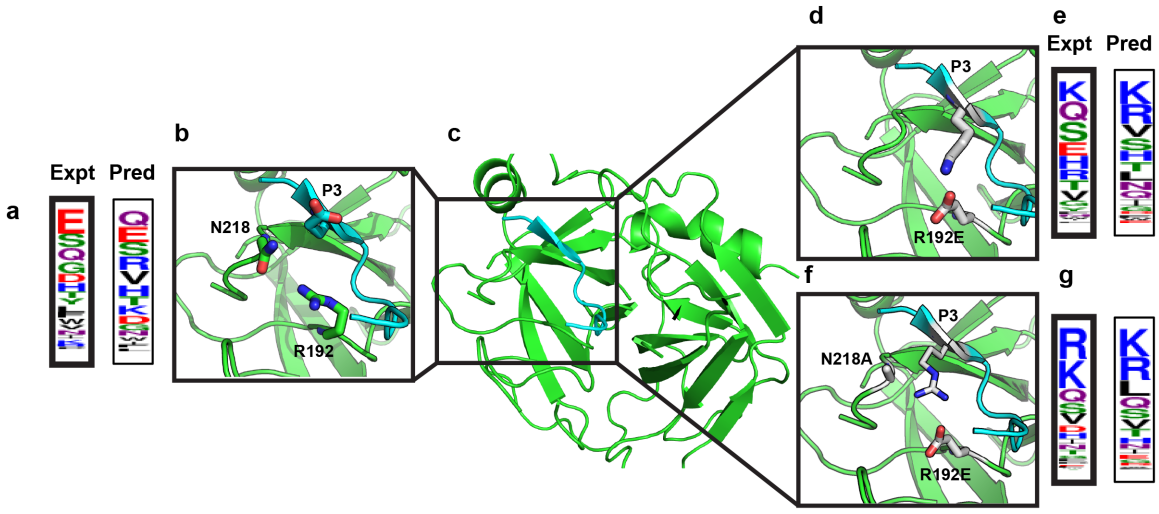


Figure 2.15. Proof-of-concept for design. Changes in specificity profile upon granzyme B protease mutation are recapitulated by MFPred.
(a) Experimental (bold) specificity (average of Harris et al. (Harris et al. 1998) and Ruggles et al. (Ruggles et al. 2004)) and predicted P3 specificity for WT granzyme B protease. **(b)-(c)**, WT granzyme B protease structure. **(d)** R192E granzyme B protease active site. **(e)** Experimental specificity (bold) (Harris et al. 1998) and predicted P3 specificity for R192E granzyme B protease. **(f)** R192E/N218A granzyme B protease active site. **(g)** Experimental specificity (bold) (Ruggles et al. 2004) and predicted P3 specificity for R192E/N218A granzyme B protease.

The MFPred-predicted specificity profile for the mutated structures accurately recapitulated the experimentally predicted specificity profile for the mutants. In the case of R192E, the change from a positively-charged arginine to a negatively-charged glutamic acid yields an increased frequency of positive amino acids such as lysine and arginine and a decreased frequency of negative amino acid glutamic acid. MFPred predicts the shift toward lysine and arginine and away from glutamic acid correctly, although it upweights the frequency of arginine and downweights the frequency of

glutamic acid relative to the experimental profile. In the case of R192E/N218A, the shift towards arginine and lysine is even more pronounced in the experimentally-derived profile. Sterically, the mutation of N to A may allow for the longer sidechains of R and K (relative to E) to fit at P3. MFPred correctly predicts this shift as well. The sensitivity of MFPred to altered multispecificity at a given position due to a given receptor mutation should enable its use in designing for or against a given specificity profile.

3.4. Discussion

Protein-peptide interactions underlie much of biology, and the ability to computationally manipulate these interactions would enable intervention in many biological processes. The rational design of receptor proteins, including enzymes that act upon peptide substrates, for and against peptide recognition specificity profiles is an open challenge. Such design would benefit from a specificity profile prediction technique that is both (i) rapid enough to be used in each step of the design process, and (ii) able to predict changed specificity for receptor variants with a constant peptide backbone conformational ensemble. The MFPred method developed here represents a step forward in achieving in both of these goals. MFPred is able to predict profiles for both proteases and a diverse set of PRDs, and it can recapitulate changes in the profile of variant granzyme B. This result sets the stage for application of the MFPred algorithm to enable the design of proteins for and against specificity profiles, by combining the MFPred algorithm with multi-state design (Leaver-Fay et al. 2011).

The MFPred method, implemented in the context of the Rosetta software, performs specificity profile prediction with equivalent or better accuracy when compared to two previously developed methods (pepspec, sequence_tolerance) in the Rosetta framework, but with a significant decrease in run time (~10- to 1000-fold). Practically, this means that given a receptor variant and a peptide backbone ensemble, a specificity profile can be obtained, on a standard single processor, on a time-scale of seconds vs. hours required for other approaches. While pepspec and sequence_tolerance are less accurate on a smaller peptide backbone ensemble, MFPred is relatively robust to the size of the backbone ensemble. Additionally, MFPred can predict information content (determined from the amino acid frequency distribution at a given peptide position) better than other methods (Figure 2.10e-h). The ability to recapitulate information content should enable design for a narrow or wide range of amino acid types at a given peptide position, thereby allowing greater control over binding selectivity. The speed, prediction accuracy on a small backbone ensemble, and robust recapitulation of information content of MFPred are due to the mean-field approach of MFPred: rather than attempt to enumerate many sequences on varying backbones, MFPred predicts a specificity profile by treating amino acid energies as a Boltzmann probability distribution. However, optimal sampling of the peptide backbone conformational space by MFPred does require some prior knowledge in the form of several (~5) recognized substrates, which is not required for pepspec or sequence_tolerance.

While MFPred can rapidly and consistently generate recognition profiles with high accuracy compared to experimental data, it was not possible to achieve a perfect

prediction using MFPred. Several reasons may underlie these limitations of MFPred. First, our experimental dataset may be incomplete: it comprises various in vitro and in vivo sources in the literature, each of which may have their biases. In vitro experimental profiles vary with the definition of a cleaved sequence; when few sequences are included in this definition, the profile will converge on a few optimal sequences. In vivo experimental profiles are subject to biases due to biological factors (King & Bradley 2010). Second, any specificity prediction challenge is composed of several, smaller problems – sampling the vast sequence space, sampling the significantly larger conformational space, and scoring the structures – each of contributes multiplicatively to the error-rate. In our study, the sequence sampling problem is solved by MFPred itself. As it is an approximation, MFPred may not sample the sequence space effectively; the free parameters, which are optimized for overall success, are sub-optimal for each system. This is especially true in the case of the temperature parameter, which we found to be the most system-dependent. Thus, application of MFPred to domain families that are not included in our benchmark set may require further system-specific optimization of model parameters to achieve comparable accuracy. In terms of structure sampling, our method of utilizing a small number of known recognized peptides to generate a backbone ensemble is an attempt to more efficiently sample the large backbone conformational space (which also determines sidechain sampling due to the use of a backbone-dependent rotamer library (Shapovalov & Dunbrack 2011)); however, this space is so large, especially in the case of a flexible binding pocket such as the HIV protease-1, that sampling efficiency is still limited. The sampling of receptor backbone flexibility is also required in such cases, as evidenced by decreased prediction accuracy when the apo-

structure of the complex is used (Figure 2.8). Finally, we score the structures using an empirical energy function (from Rosetta); subtle errors in the energy function may also contribute to the observed inaccuracies. As both conformational and sequence sampling in the MFPred approach rely on, and are limited by, the underlying rotamer library and energy function as implemented in Rosetta, improvements in these features (Park et al. 2016; Shapovalov & Dunbrack 2011) should yield higher accuracy predictions.

3.5. Methods

3.5.1. Inputs

Table 2.4: Details of model generation for four proteases and fourteen PRDs

Protein	PDB ID	Resolution	Notes
HCV NS3/4A **Protease**	3M5L, 3M5N	1.9 Å	The P' residues of the bound peptide were built by overlaying PDB ID: 3M5N and PDB ID:3M5L (inhibitor-bound crystal structure) thus allowing us to build a complete substrate bound complex
HCV NS3 **Protease** (apo)	3KF2	2.5 Å	PDB ID: 3KF2, the apo structure of HCV NS3 protease, was superimposed with the complex built from 3M5L and 3M5N (above) and the peptide from that model was added to the apo structure to generate the starting model.
TEV **Protease**	1LVB, 1LVM	2.2 Å	Starting model generated from PDB by reverting C151A to WT

Granzyme B (**Protease**)	1FI8	2.2 Å	The interface of the ecotin chain in the crystal structure, spanning eight residue substrate chain was used as the starting point for further calculations
HIV **Protease** 1	1MT9	2.0 Å	Starting model generated by inverting D25N and V82N from crystal structure to native residue identities
HIV **Protease** 1 (apo)	2HB4	2.15 Å	PDB ID: 2HB4, the closed-form apo structure of HIV protease-1, was superimposed with the complex built from 1MT9 (above) and the peptide from that model was added to the apo structure to generate the starting model.
HIV **Protease** 1 (apo)	2PC0	1.4 Å	PDB ID: 2PC0, the open-form apo structure of HIV protease-1, was superimposed with the complex built from 1MT9 (above) and the peptide from that model was added to the apo structure to generate the starting model.
c-Crk **SH3**-N	1CKA	1.5 Å	
cAMP-dependent PKA (**kinase**)	1L3R	2.0 Å	

Src **SH2**	1SPS	2.7 Å	
PSD-95 **PDZ3**	1TP3	1.99 Å	
NHERF-2 **PDZ2**	2HE4	1.45 Å	
AF-6 **PDZ**	2AIN	(NMR)	First model in NMR ensemble was taken.
Erbin **PDZ**	1N7T	(NMR)	First model in NMR ensemble was taken.
MPDZ-13 (**PDZ**)	2FNE	1.83 Å	
ZO-1 **PDZ1**	2H2B	1.6 Å	
DLG1-2 (**PDZ**)	2I0L	2.31 Å	
HLA-A*0201 (**MHC)**	1QSF	2.8 Å	

HLA-B*1501 (**MHC)**	1XR9	1.79 Å	
HLA-B*4402 (**MHC)**	1M6O	1.6 Å	
HLA-B*4403 (**MHC)**	1N2R	1.7 Å	

Structure Preparation. Crystal structures of the four protease-peptide complexes, fourteen protein-recognition domains, and three protease apo structures were procured from the Protein Data Bank (PDB) (Table 2.4) (Phan et al. 2002; Prabu-Jeyabalan et al. 2003; Waugh et al. 2000; Romano et al. 2010; Saro et al. n.d.; Madhusudan et al. 2002; Wu et al. 1995; Elkins et al. 2007; Skelton 2003; Appleton et al. 2006; Zhang et al. 2007; Ding et al. 1999; Røder et al. 2006; Macdonald et al. 2003; Cummings et al. 2010; Heaslet et al. 2007; Waksman et al. 1993; Chen et al. 2007). Structures were filtered for a resolution equal to or lower than 2.8 Å and a bound peptide or peptidomimetic inhibitor. Active site mutations were reverted to the wild-type residues.

The selected crystal structures were optimized using Rosetta FastRelax to find a low energy structure, which was used as a starting point in further calculations. In the case of the protease enzymes, constraints were applied to catalytic residues during FastRelax to

maintain active site geometry and keep the protease in a pre-transition-state near-attack conformation, and coordinate constraints were applied to the backbone to ensure that the enzyme did not unfold; we did not apply constraints in the general PRD benchmark, as constraints were found to decrease prediction accuracy in those cases. Peptide side chains and backbone were allowed to sample all degrees of freedom including rotation, translation, and rigid body orientation with respect to the protease. The models were scored with Rosetta's talaris2013 energy function.

The apo crystal structures were aligned with the relaxed models of the protease-peptide complexes using PyMol (Anon n.d.), and the peptides from the protease-peptide complexes were placed within the apo models. The crystal structures were further optimized using Rosetta FastRelax as described above.

Experimental Sequence Profiles and Cleaved/Uncleaved Sequences. The sequences of cleaved and uncleaved substrate peptides for each protease and bound peptides for each PRD were obtained as described in Table 2.5. For further details on the curation of the protease datasets, please see our recent study (Pethe et al. 2017). To generate a specificity profile for each protease, we first removed duplicates from the set of cleaved peptides and then calculated the frequency of each amino acid at each position. We followed the same procedure for the PRDs; however, we did not remove duplicates from those sets. The sequence sets are provided in S1 Dataset.

Table 2.5. Substrates for proteases and PRDs.

Protease	**# Cleaved**	**# Uncleaved**	**References**
TEV-PR	68	1520	• Kostallas et al. (Kostallas et al. 2011) • Boulware et al. (Boulware et al. 2010)
HCV protease	196	1943	• Shiryaev et al. (Shiryaev et al. 2012) • Rögnvaldsson et al. (Rögnvaldsson et al. 2009)
Granzyme B protease	353	1973	• Barkan et al. (Barkan et al. 2010)
HIV-PR	374	1251	• Rögnvaldsson et al. (Rögnvaldsson et al. 2009)
PRD	**#Bound in vitro**	**#Bound in vivo**	**References**
c-Crk SH3-N	13	N/A	• Sparks et al. (Sparks et al. 1996)
cAMP-dependent PKA	346	19	• PhosphoELM (Dinkel et al. 2011) • Schutkowski et al. (Schutkowski et al. 2004)
Src SH2	13	117	• PepCyber (Gong et al. 2008) • Khati et al. (Khati & Pillay 2004)
PSD-95 PDZ3	93	2	• PDZBase (Beuming et al. 2005) • Tonikian et al. (Tonikian et al. 2008)
NHERF-2 PDZ2	132	N/A	• Vouilleme et al. (Vouilleme et al. 2010) • Stiffler et al. (Stiffler et al. 2007) • Tonikian et al. (Tonikian et al. 2008)
AF-6 PDZ	176	N/A	• Tonikian et al. (Tonikian

			et al. 2008)
Erbin PDZ	86	N/A	• Tonikian et al. (Tonikian et al. 2008)
MPDZ-13 (PDZ)	91	N/A	• Tonikian et al. (Tonikian et al. 2008)
ZO-1 PDZ1	71	N/A	• Tonikian et al. (Tonikian et al. 2008)
DLG1-2 (PDZ)	58	N/A	• Tonikian et al. (Tonikian et al. 2008)
HLA-A*0201 (MHC)	3273	N/A	• Vita et al. (Vita et al. 2015)
HLA-B*1501 (MHC)	1187	N/A	• Vita et al. (Vita et al. 2015)
HLA-B*4402 (MHC)	236	N/A	• Vita et al. (Vita et al. 2015)
HLA-B*4403 (MHC)	207	N/A	• Vita et al. (Vita et al. 2015)

3.5.2. Backbone Ensemble Generation

We generated a flexible backbone ensemble by constructing models of the proteins bound to several cleaved sequences, and then diversifying those models via FastRelax (Tyka et al. 2011), FlexPepDock (Raveh et al. 2010), or Backrub (Smith & Kortemme 2008) backbone sampling protocols, as described in detail below. For each protein, N cleaved sequences were chosen from the dataset by sorting the sequences in alphabetical order and then choosing evenly spaced sequences from the sorted dataset. Two alternative methods of picking cleaved sequences - randomly, or at even intervals from a set sorted by hamming distance from an arbitrarily chosen cleaved sequence - did not impact the results.

Then those N cleaved sequences were threaded onto the original FastRelaxed protein-peptide complex to create N structure-sequence models. Each model was subjected to 10

trajectories of FastRelax simulations, 10 trajectories of FlexPepdock refine simulations, or 10 trajectories of Backrub simulations, and the resulting 10 models were considered to be the backbone conformational ensemble. As we found that the FastRelax protocol was more accurate than FlexPepDock and Backrub, we used FastRelax alone in the final version of the protocol. The model was constrained to active catalytic geometry for the proteases; we did not apply constraints to the PRD systems. Finally, the x lowest-scoring models for each sequence (with x dependent on the protocol in question, and generally set as 1) were chosen as the final backbone ensemble.

3.5.3. Mean-Field Algorithm

Various self-consistent mean-field theory-based methods have been developed for use in protein sidechain packing and design (Koehl & Delarue 1994; Delarue & Koehl 1997; Lee 1994; Voigt et al. 2001; Saven & Wolynes 1997; Xiao et al. 2014; Mendes et al. 1999; Kono 1996). In the canonical self-consistent mean field theory-based method for protein sidechain packing as proposed by Koehl and Delarue (Koehl & Delarue 1994), the energy landscape is investigated by using an effective energy potential to approximate the effects of all possible rotamers at all positions to be modeled. Thus, the mean-field energy of rotamer *r* occurring at position *i* is determined by Eq. 1:

$$E(i,r) = e(i_r) + \sum_{j=1, j\neq i}^{N} \sum_{s=1}^{K_j} e(i_r, j_s) P(j,s) \quad (1)$$

$e(i_r)$ represents the one-body energy of the rotamer, or the energy between a residue and the fixed components of the protein. $e(i_r, j_s)$ represents the two-body energy between a

rotamer *r* at position *i* and a rotamer *s* at position *j*. Energies are truncated at a threshold that we optimized as a free parameter. *P(j, s)* represents the probability of rotamer *s* occurring at position *j* and is initially given as $1/K_j$, where K_j is the total number of available rotamers at position *j* (obtained from a rotamer library).

A probability matrix (**P**) of size $N \times K_{max}$, where N is the number of positions to be analyzed and K_{max} is the maximum number of rotamers at any position, is used to model the probabilities of each rotamer occurring. Once the effective energy of each rotamer is determined using (1), the probability of each rotamer is:

$$P(j,s) = \frac{e^{-\beta E(j,s)}}{\sum_{x=1}^{K_j} e^{-\beta E(j,x)}} \quad (2)$$

β (= $1/k$T) is also optimized as a free parameter. The algorithm iterates between the two equations until convergence is reached. We use a pre-calculated interaction graph in Rosetta (Leaver-Fay et al. 2005) to store the one-body and two-body energies, which do not change between iterations, so the iteration is rapid. Convergence is improved with the use of a memory in the updating of **P**, so that the probability matrix after iteration *x* is given by $P_x = \lambda P_{x-1} + (1-\lambda)P_x$, where λ is a free parameter between 0 and 1. Once convergence is reached, the probability matrix **P** can be used to obtain the probability for every rotamer.

We extended the algorithm for use with a flexible backbone and with any given amino acid alphabet. Given an ensemble of backbone conformations, the probability matrix **P** is calculated for each backbone using the canonical self-consistent mean field method,

while allowing each position to take on any amino acid, so that the vector for that position contains all the rotamers for all amino acids at that position. $P_{aa}(bb, i)$, the probability of amino acid *aa* occurring at position *i* in backbone *bb*, is determined for all amino acids at all positions in all backbones:

$$P_{aa}(bb,i) = \frac{\sum_{r=1}^{K_{aa}} P_{bb}(i,r) \Big/ K_{aa}^{\gamma}}{\sum_{x=1}^{20} \sum_{r=1}^{K_x} P_{bb}(i,r) \Big/ K_x^{\gamma}} \tag{3}$$

where K_{aa} is the number of rotamers available to amino acid aa at position *i*, and γ is a free parameter optimized to 0.8 in our implementation. Dividing the sum of probabilities over all amino acids by K_{aa}^{γ} thus corrects for cases where numerous rotamers of an amino acid artificially inflate the probability of a specific amino acid occurring (Figure 2.16). The probability matrices for all backbones are then averaged together using a Boltzmann-weighting scheme in a two-step process. First, $E_{bb}(i,aa)$, the weighted sum of the energies for rotamers of amino acid *aa* at position *i* in backbone *bb*, divided by K_{aa}^{γ}, is calculated (Eq. 4). Then $E_{bb}(i,aa)$ is used to find $W(i)$, the probability of backbone bb occurring at position i (Eq. 5). *M* is the number of (peptide) backbones in the ensemble.

$$E_{bb}(i,aa) = \frac{\sum_{r=1}^{K_{aa}} E_{bb}(i,r) P_{bb}(i,r)}{K_{aa}^{\gamma}} \tag{4}$$

$$W(i) = \frac{e^{-\beta \sum_{aa=1}^{20} E_{bb}(i,aa)}}{\sum_{s=1}^{M} e^{-\beta \sum_{aa=1}^{20} E_s(i,aa)}} \tag{5}$$

Finally, a weighted average P is determined and taken to be the predicted specificity profile for that protease:

$$P(i, aa) = \sum_{bb=1}^{M} P_{aa}(bb, i) W(i) \tag{6}$$

Thus, MFPred can be used for prediction of multispecificity for both one backbone and multiple backbone conformations.

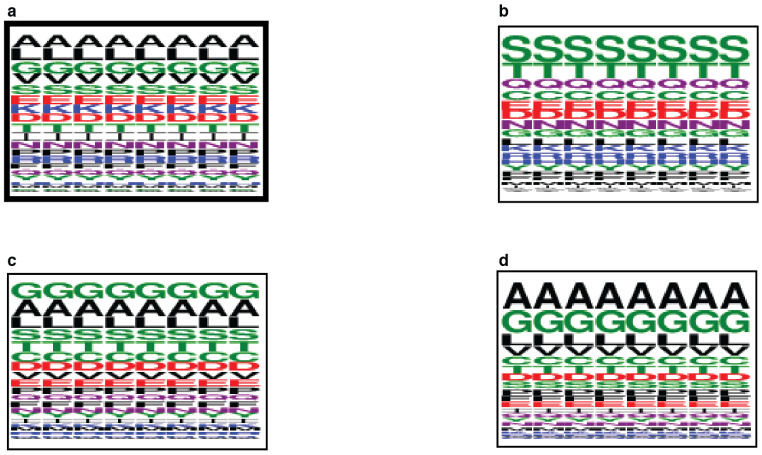


Figure 2.16. The need for γ in the mean-field algorithm when averaging rotamers of an amino acid to find the probability of that amino acid.
(a) Background amino acid composition as defined in Rosetta database (P_AA). This is the gold-standard which we attempted to match in our background profile generation (see Methods.3). (b) MFPred's background prediction with γ=0, i.e. the rotamer probabilities are simply summed to find the amino acid probability. Serine and threonine are overrepresented as the Rosetta Dunbrack library contains many more rotamers for S and T, and glycine and alanine are underrepresented due to having only one rotamer each. (c) MFPred's background prediction with γ=0.8 (current settings). This is closest to the P_AA distribution (Frobenius distance of 0.24). (d) MFPred's background prediction with γ=1.0, i.e. the amino acid probability is simply the average of the rotamer probabilities. While this is better than γ=0, alanine and glycine are now overrepresented and serine and threonine are underrepresented. Frobenius distance is 0.39.

3.5.4. Parameter Optimization of MFPred

To optimize four free parameters for MFPred (lambda, γ, threshold, and *k*T), we enumerated all combinations of lambda (0.25, 0.5, 0.75), γ (0, 0.2, 0.4, 0.6, 0.8, 1.0), threshold (5, 10, 50, 100, 250, 500), and *k*T (0.2, 0.4, 0.6, 0.8, 1.0). We selected 68 structures from the peptiDB (a peptide-protein complex database) (London et al. 2010) that met our criteria of having at least eight peptide residues. The structures were input into MFPred as a backbone ensemble and all combinations of the above parameters were tested. The resulting background specificity profiles were compared to the background residue distribution in the Rosetta database (Figure 2.16, Figure 2.17) and the combination of parameters with the lowest cosine distance from the known background distribution was chosen as our final set of parameters. While varying lambda had little impact on the results, all other parameters had a significant, system-dependent impact on the results.

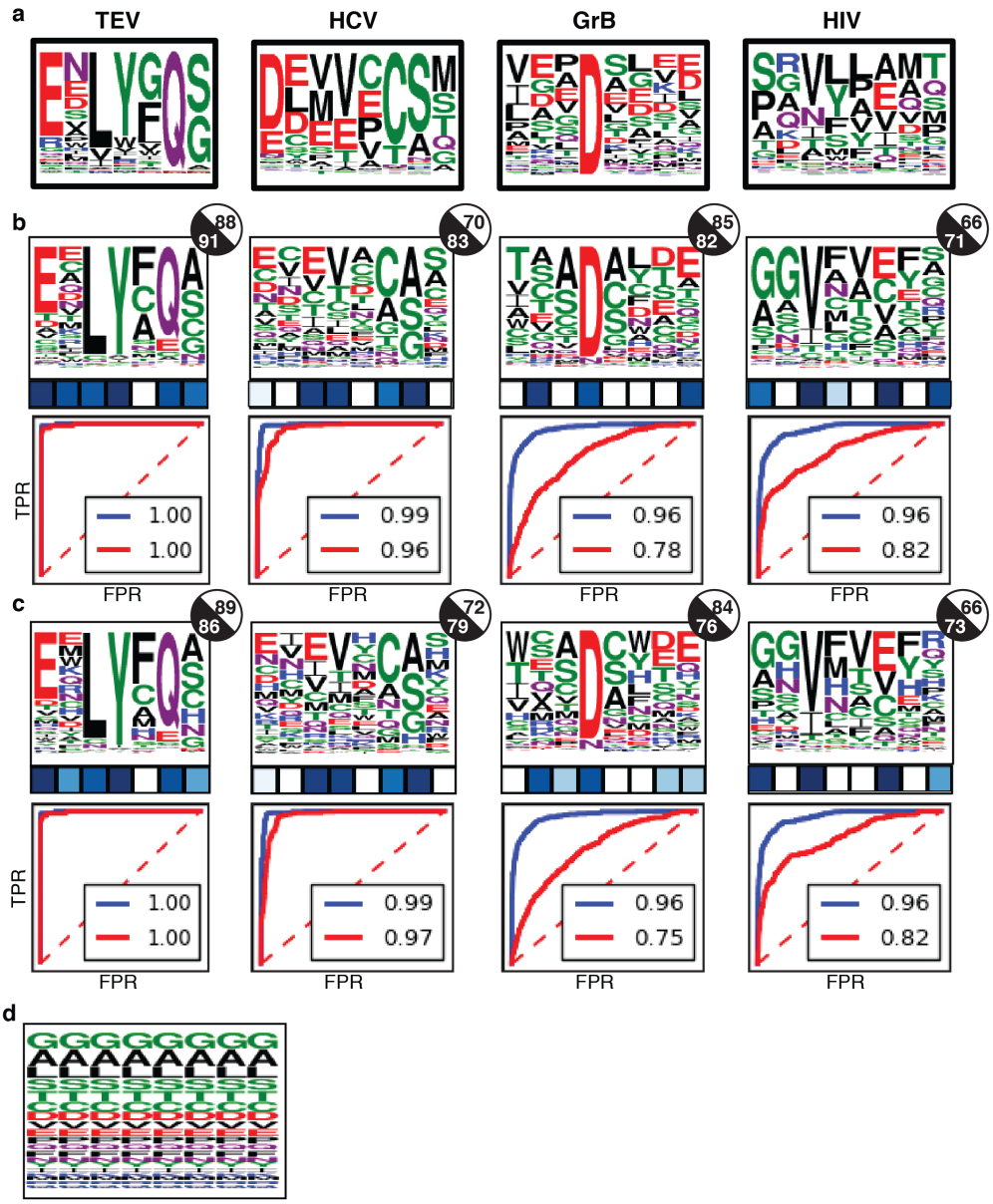


Figure 2.17. Enriching specificity profiles over background specificity profile improves accuracy.
(a) Experimental specificity profiles. (b) Initial MFPred-predicted specificity profiles. (c) Specificity profiles divided by background specificity profile. (d) Background specificity profile.

3.5.5. Enrichment over Background

Since the MFPred predictions did include some noise due to the background distribution, we divided its predictions by the background profile to find the final prediction. The

background profile was determined by averaging the frequencies of each position in the peptiDB profile. We divided each amino acid frequency in the initial predicted profile by the frequency of that amino acid in the background profile to find the final profile (Figure 2.17).

3.5.6. Software Availability

MFPred is available as a RosettaScripts Mover within the master branch of Rosetta. Sample cases for how to use MFPred can be found in S2 Note and in online Rosetta documentation.

3.6. References

Anon, The PyMol Molecular Graphics System. , p.Version 1.8.0.3, Schrodinger, LLC.

Appleton, B.A. et al., 2006. Comparative Structural Analysis of the Erbin PDZ Domain and the First PDZ Domain of ZO-1. Journal of Biological Chemistry, 281(31), pp.22312–22320.

Barkan, D.T. et al., 2010. Prediction of protease substrates using sequence and structure features. Bioinformatics (Oxford, England), 26(14), pp.1714–22. Available at: http://bioinformatics.oxfordjournals.org/content/26/14/1714.abstract

Beuming, T. et al., 2005. PDZBase: A protein-protein interaction database for PDZ-domains. Bioinformatics, 21(6), pp.827–828.

Boulware, K.T., Jabaiah, A. & Daugherty, P.S., 2010. Evolutionary optimization of peptide substrates for proteases that exhibit rapid hydrolysis kinetics. Biotechnology and bioengineering, 106(3), pp.339–46.

Chapman, H.A., Riese, R.J. & Shi, G.P., 1997. Emerging roles for cysteine proteases in human biology. Annual Reviews in Physiology, 59, pp.63–88.

Chen, Q. et al., 2007. Solution structure and backbone dynamics of the AF-6 PDZ domain/Bcr peptide complex. Protein Science, 16(6), pp.1053–1062.

Craik, C.S., Page, M.J. & Madison, E.L., 2011. Proteases as therapeutics. Biochemical

Journal, 435, pp.1–16.

Cummings, M.D. et al., 2010. Induced-fit binding of the macrocyclic noncovalent inhibitor TMC435 to its HCV NS3/NS4A protease target. Angewandte Chemie - International Edition, 49(9), pp.1652–1655.

Delarue, M. & Koehl, P., 1997. The inverse protein folding problem: self consistent mean field optimisation of a structure specific mutation matrix. Pac.Symp.Biocomput., p.109.

Ding, Y.H. et al., 1999. Four A6-TCR/peptide/HLA-A2 structures that generate very different T cell signals are nearly identical. Immunity, 11(1), pp.45–56.

Dinkel, H. et al., 2011. Phospho.ELM: A database of phosphorylation sites-update 2011. Nucleic Acids Research, 39(SUPPL. 1), pp.D261-7.

Domchek, S.M. et al., 1992. Inhibition of SH2 domain/phosphoprotein association by a nonhydrolyzable phosphonopeptide. Biochemistry, 31, pp.9865–9870.

Dunbrack, R., 2002. Rotamer Libraries in the 21st Century. Current Opinion in Structural Biology, 12(4), pp.431–440.

Elkins, J.M. et al., 2007. Structure of PICK1 and other PDZ domains obtained with the help of self-binding C-terminal extensions. Protein Science, 16, pp.683–694.

Erijman, A., Aizner, Y. & Shifman, J.M., 2011. Multispecific recognition: mechanism, evolution, and design. Biochemistry, 50, pp.602–611.

Felder, S. et al., 1993. SH2 domains exhibit high-affinity binding to tyrosine-phosphorylated peptides yet also exhibit rapid dissociation and exchange. Molecular and Cellular Biology, 13(3), pp.1449–1455.

Fujiwara, Y. et al., 2015. Crystal structure of afadin PDZ domain-nectin-3 complex shows the structural plasticity of the ligand-binding site. Protein Science, 24(3), pp.376–385.

Gong, W. et al., 2008. PepCyber:P~PEP: a database of human protein protein interactions mediated by phosphoprotein-binding domains. Nucleic Acids Res, 36(Database issue), pp.D679-83.

Grigoryan, G., Reinke, A.W. & Keating, A.E., 2009. Design of protein-interaction specificity gives selective bZIP-binding peptides. Nature, 458(7240), pp.859–864.

Harris, J.L. et al., 1998. Definition and redesign of the extended substrate specificity of granzyme B. Journal of Biological Chemistry, 273(42), pp.27364–27373.

Havranek, J.J. & Harbury, P.B., 2002. Automated design of specificity in molecular

recognition. Nature Structural Biology, 10, pp.45–52.

Heaslet, H. et al., 2007. Conformational flexibility in the flap domains of ligand-free HIV protease. Acta Crystallographica Section D: Biological Crystallography, 63(8), pp.866–875.

Hirsch, T. et al., 1998. Caspases : Enemies Within. Science, 281(August), pp.1312–1316.

Kerekatte, V. et al., 1999. Cleavage of Poly(A)-binding protein by coxsackievirus 2A protease in vitro and in vivo: another mechanism for host protein synthesis shutoff? Journal of virology, 73, pp.709–717.

Khare, S.D. & Fleishman, S.J., 2013. Emerging themes in the computational design of novel enzymes and protein – protein interfaces. FEBS Letters, 587(8), pp.1147–1154.

Khati, M. & Pillay, T.S., 2004. Phosphotyrosine phosphoepitopes can be rapidly analyzed by coexpression of a tyrosine kinase in bacteria with a T7 bacteriophage display library. Analytical Biochemistry, 325(1), pp.164–167.

Kim, P. et al., 2006. Relating Three-Dimensional Structure to Protein Network Provides Evolutionary Insights. Science, 314(December), pp.1938–1941.

King, C.A. & Bradley, P., 2010. Structure-based prediction of protein– peptide specificity in Rosetta. Cancer Research, pp.3437–3449.

Koehl, P. & Delarue, M., 1994. Application of a Self-consistent Mean Field Theory to Predict Protein Side-chains Conformation and Estimate Their Conformational Entropy. Journal of Molecular Biology, 239(2), pp.249–275.

Kono, H., 1996. A new method for side-chain conformation prediction using a Hopfield network and reproduced rotamers. Journal of computational chemistry, 17(14), pp.1667–1683.

Kostallas, G., Löfdahl, P.-Å. & Samuelson, P., 2011. Substrate profiling of tobacco etch virus protease using a novel fluorescence-assisted whole-cell assay. PloS one, 6(1), p.e16136.

Lanouette, S. et al., 2015. Discovery of substrates for a SET domain lysine methyltransferase predicted by multistate computational protein design. Structure (London, England : 1993), 23(1), pp.206–15.

Leaver-Fay, A. et al., 2011. A generic program for multistate protein design. PLoS ONE, 6(7).

Leaver-Fay, A., Kuhlman, B. & Snoeyink, J., 2005. An adaptive dynamic programming algorithm for the side chain placement problem. Pacific Symposium on Biocomputing,

pp.16–27.

Lee, C., 1994. Predicting protein mutant energetics by self-consistent ensemble optimization. Journal of Molecular Biology, 236(3), pp.918–939.

Li, Q. et al., 2013. Commercial proteases: present and future. FEBS Letters, 587, pp.1155–1163.

London, N. et al., 2011. Identification of a novel class of farnesylation targets by structure-based modeling of binding specificity. PLoS computational biology, 7(10), p.e1002170.

London, N., Movshovitz-Attias, D. & Schueler-Furman, O., 2010. The Structural Basis of Peptide-Protein Binding Strategies. Structure, 18(2), pp.188–199.

Lundegaard, C. et al., 2010. Major histocompatibility complex class I binding predictions as a tool in epitope discovery. Immunology, 130(3), pp.309–318.

Macdonald, W.A. et al., 2003. A naturally selected dimorphism within the HLA-B44 supertype alters class I structure, peptide repertoire, and T cell recognition. The Journal of experimental medicine, 198(5), pp.679–691.

Madhusudan et al., 2002. Crystal structure of a transition state mimic of the catalytic subunit of cAMP-dependent protein kinase. Nature Structural & Molecular Biology, 9(4), pp.273–277.

Mendes, J., Soares, C.M. & Carrondo, M.A., 1999. Improvement of side-chain modeling in proteins with the self-consistent mean field theory method based on an analysis of the factors influencing prediction. Biopolymers, 50(2), pp.111–131.

Monahan, P. & Di Paola, J., 2010. Recombinant Factor IX for Clinical and Research Use. Seminars in Thrombosis and Hemostasis, 36(5), pp.498–509.

Newman, J.R.S. & Keating, A.E., 2003. Comprehensive identification of human bZIP interactions with coiled-coil arrays. Science, 300(5628), pp.2097–2101.

Pampalakis, G. & Sotiropoulou, G., 2007. Tissue kallikrein proteolytic cascade pathways in normal physiology and cancer. Biochimica et Biophysica Acta - Reviews on Cancer, 1776(1), pp.22–31.

Park, H. et al., 2016. Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. Journal of Chemical Theory and Computation, 12(12), pp.6201–6212.

Pethe, M.A., Rubenstein, A.B. & Khare, S.D., 2017. Large-Scale Structure-Based Prediction and Identification of Novel Protease Substrates Using Computational Protein

Design. Journal of Molecular Biology, 429(2), pp.220–236.

Phan, J. et al., 2002. Structural basis for the substrate specificity of tobacco etch virus protease. The Journal of biological chemistry, 277(52), pp.50564–72.

Prabu-Jeyabalan, M. et al., 2003. Viability of a drug-resistant human immunodeficiency virus type 1 protease variant: structural insights for better antiviral therapy. Journal of virology, 77(2), pp.1306–15.

Raveh, B., London, N. & Schueler-Furman, O., 2010. Sub-angstrom modeling of complexes between flexible peptides and globular proteins. Proteins: Structure, Function and Bioinformatics, 78(9), pp.2029–2040.

Røder, G. et al., 2006. Crystal structures of two peptide-HLA-B*1501 complexes; structural characterization of the HLA-B62 supertype. Acta Crystallographica Section D: Biological Crystallography, 62(11), pp.1300–1310.

Rögnvaldsson, T. et al., 2009. How to find simple and accurate rules for viral protease cleavage specificities. BMC bioinformatics, 10(1), p.149.

Romano, K.P. et al., 2010. Drug resistance against HCV NS3/4A inhibitors is defined by the balance of substrate recognition versus inhibitor binding. Proceedings of the National Academy of Sciences of the United States of America, 107(49), pp.20986–91.

Ruggles, S.W., Fletterick, R.J. & Craik, C.S., 2004. Characterization of structural determinants of granzyme B reveals potent mediators of extended substrate specificity. Journal of Biological Chemistry, 279(29), pp.30751–30759.

Saro, D. et al., Structure of the third PDZ domain of PSD-95 protein complexed with KKETPV peptide ligand. To be Published.

Saven, J.G. & Wolynes, P.G., 1997. Statistical mechanics of the combinatorial synthesis and analysis of folding macromolecules. Journal of Physical Chemistry B, 101(41), pp.8375–8389.

Scheel, T. & Rice, C., 2014. Understanding the HCV life cycle paves the way for highly effective therapies. Nat.Med, 19(7), pp.837–849.

Schreiber, G. & Keating, A.E., 2011. Protein binding specificity versus promiscuity. Current Opinion in Structural Biology, 21(1), pp.50–61.

Schutkowski, M. et al., 2004. High-content peptide microarrays for deciphering kinase specificity and biology. Angewandte Chemie - International Edition, 43(20), pp.2671–2674.

Shapovalov, M. V & Dunbrack, R.L., 2011. A smoothed backbone-dependent rotamer

library for proteins derived from adaptive kernel density estimates and regressions. Structure, 19(6), pp.844–858.

Shiryaev, S.A. et al., 2012. New details of HCV NS3/4A proteinase functionality revealed by a high-throughput cleavage assay. PloS one, 7(4), p.e35759.

Skelton, N.J., 2003. Origins of PDZ Domain Ligand Specificity. STRUCTURE DETERMINATION AND MUTAGENESIS OF THE ERBIN PDZ DOMAIN. Journal of Biological Chemistry, 278(9), pp.7645–7654.

Smith, C.A. & Kortemme, T., 2008. Backrub-Like Backbone Simulation Recapitulates Natural Protein Conformational Variability and Improves Mutant Side-Chain Prediction. Journal of Molecular Biology, 380(4), pp.742–756.

Smith, C.A. & Kortemme, T., 2010. Structure-based prediction of the peptide sequence space recognized by natural and synthetic PDZ domains. Journal of molecular biology, 402(2), pp.460–74.

Smith, C. & Kortemme, T., 2011. Predicting the tolerated sequences for proteins and protein interfaces using RosettaBackrub flexible backbone design. PLoS ONE, 6(7), p.e20451.

Sparks, A.B. et al., 1996. Distinct ligand preferences of Src homology 3 domains from Src, Yes, Abl, Cortactin, p53bp2, PLCgamma, Crk, and Grb2. Proceedings of the National Academy of Sciences of the United States of America, 93(4), pp.1540–1544.

Stiffler, M.A. et al., 2007. PDZ domain binding selectivity is optimized across the mouse proteome. Science, 317(5836), pp.364–369.

Tawfik, D.S., 2014. Accuracy-rate tradeoffs: how do enzymes meet demands of selectivity and catalytic efficiency? Current opinion in chemical biology, 21, pp.73–80.

Tonikian, R. et al., 2008. A specificity map for the PDZ domain family. PLoS Biology, 6(9), pp.2043–2059.

Tyka, M.D. et al., 2011. Alternate states of proteins revealed by detailed energy landscape mapping. Journal of molecular biology, 405(2), pp.607–18.

Ubersax, J.A. & Ferrell, J.E., 2007. Mechanisms of specificity in protein phosphorylation. Nature Reviews Molecular Cell Biology, 8, pp.530–541.

Vita, R. et al., 2015. The immune epitope database (IEDB) 3.0. Nucleic Acids Research, 43(D1), pp.D405–D412.

Voigt, C.A. et al., 2001. Computational method to reduce the search space for directed protein evolution. Proceedings of the National Academy of Sciences, 98(7), pp.3778–

3783.

Vouilleme, L. et al., 2010. Engineering peptide inhibitors to overcome PDZ binding promiscuity. Angewandte Chemie - International Edition, 49(51), pp.9912–9916.

Waksman, G. et al., 1993. Binding of a high affinity phosphotyrosyl peptide to the Src SH2 domain: crystal structures of the complexed and peptide-free forms. Cell, 72, pp.779–790.

Watkins, A.M., Bonneau, R. & Arora, P.S., 2016. Side-chain conformational preferences govern protein−protein interactions. Journal of the American Chemical Society, 138, p.10386−10389.

Waugh, S.M. et al., 2000. The structure of the pro-apoptotic protease granzyme B reveals the molecular determinants of its specificity. Nature structural biology, 7(9), pp.762–5.

Wollacott, A.M. & Desjarlais, J.R., 2001. Virtual interaction profiles of proteins. Journal of molecular biology, 313(2), pp.317–342.

Wu, X. et al., 1995. Structural basis for the specific interaction of lysine-containing proline-rich peptides with the N-terminal SH3 domain of c-Crk. Structure, 3(2), pp.215–226.

Xiao, X., Hall, C.K. & Agris, P.F., 2014. The design of a peptide sequence to inhibit HIV replication: a search algorithm combining Monte Carlo and self-consistent mean field techniques. Journal of biomolecular structure & dynamics, 32(10), pp.1523–1536.

Zhang, Y. et al., 2007. Structures of a human papillomavirus (HPV) E6 polypeptide bound to MAGUK proteins: mechanisms of targeting tumor suppressors by a high-risk HPV oncoprotein. Journal of virology, 81(7), pp.3618–3626.

Zheng, F. et al., 2015. Computational design of selective peptides to discriminate between similar PDZ domains in an oncogenic pathway. Journal of Molecular Biology, 427(2), pp.491–510.

3.7. S1 Note. Explanation of metrics.

We used several metrics and distances to evaluate specificity profile predictions. The Frobenius distance is defined as:

$$Frobenius(E,P) = \sqrt{\sum_{i=1}^{N}(E_i - P_i)^2}$$

where E is a vector of experimentally determined amino acid frequencies and P is a vector of predicted frequencies. To calculate the Frobenius distance of the entire profile, we simply flattened the experimental and predicted profiles into one vector each. Two identical probability distributions have a Frobenius distance of 0, while two most divergent distributions have a Frobenius distance of $(2n)^{1/2}$, where n is equal to the number of positions in the profile.

The Average Absolute Distance (AAD) is defined as:

$$AAD(E,P) = \frac{1}{N}\sum_{i=1}^{N}|E_i - P_i|$$

Again, to calculate the AAD of the entire profile, we flattened each profile to a single vector. AAD ranges between 0 to 1, with 0 as the best score and 1 as the worst score. According to Smith and Kortemme, an AAD of less than 6% (or 0.06) is considered to be a good prediction.

The cosine similarity is defined as:

$$Cosine(E,P) = \frac{\sum_{i=1}^{N} E_i P_i}{\sqrt{\sum_{i=1}^{N} E_i^2}\sqrt{\sum_{i=1}^{N} P_i^2}}$$

We flattened each profile to a single vector. Two identical specificity profiles have a cosine distance of 1 whereas two most divergent profiles have a similarity of 0.

Jensen-Shannon Divergence (JSD) is defined as:

$$JSD(E,P) = H\left(\sum_{i=1}^{N} 0.5E_i + 0.5P_i\right) - 0.5\sum_{i=1}^{N} H(E_i) - 0.5\sum_{i=1}^{N} H(P_i)$$

where H is Shannon entropy, defined as:

$$H(E) = -\sum_{i=1}^{N} E_i \log_2 E_i$$

We calculated the JSD of the entire profile by averaging the JSD of each vector (or position) in the profile. A JSD of zero denotes two identical profiles, whereas a JSD of 1 denotes two entirely divergent profiles. While JSD is not considered a proper metric, it does provide information regarding how divergent two profiles are.

Area under the ROC curve, or AUC, as developed by Smith and Kortemme (Smith & Kortemme, 2010) , is another measure that we used to evaluate the profiles. We plotted an ROC curve for each predicted profile based on how well the most frequent experimental amino acids (defined as > 10%) are recapitulated in the predicted profile. We then calculated the area under the curve, which denotes the probability that the predicted profile ranks a positive amino acid as higher than a negative amino acid. An AUC of 1 represents a perfect prediction, while an AUC of 0.5 is equivalent to a random prediction.

Last, we developed a new distance, referred to as the Score-Sequence AUC Loss (SSAL). This distance also takes advantage of an ROC curve, although this one is slightly different. We use the experimental profile to generate a score for each cleaved and uncleaved sequence by taking the sum of the probabilities of each amino acid in the sequence occurring at its position:

$$Score(S) = \sum_{i=1}^{len(S)} E_i(S_i)$$

We then plot an ROC curve that demonstrates how well the scores rank the cleaved vs. uncleaved sequences and calculate its AUC. We repeat the entire process with the predicted profile, and then subtract the predicted ROC-AUC from the experimental ROC-AUC. The result is the SSAL, which denotes how well the predicted profile differentiates between cleaved/uncleaved sequences vs. the experimental profile.

In order to transform the values of the distances to p-values, we generated 100,000 random profiles by randomly sampling columns of our protease and PRD experimental profile library and randomly shuffling the amino acid identity of their frequencies so as to generate profiles with similar information content. We then calculated their per-column and overall distance from each experimental profile for each of the six measures. The ranking of a given predicted profile distance value in its given distance list was then used to find the p-value.

References

Smith CA, Kortemme T. 2010. Structure-Based Prediction of the Peptide Sequence Space Recognized by Natural and Synthetic PDZ Domains. J Mol Biol.,402(2). pp:460–74.

3.8. S2 Note. Supplementary Software.

Running Entire MFPred pipeline:

Inputs

- Crystallographic pdb of protein-peptide complex
- List of five substrate sequences to thread on

Process

1. Initial Relax
    a. Run on initial crystallographic pdb to get rid of internal clashes
2. Thread Peptide-FastRelax
    a. Run this step for each substrate sequence
3. MFPred
    a. Choose the lowest-scoring pdb from 2a for each substrate sequence and use a list of paths to these pdbs as the input for MFPred
4. Distances.py (optional)

Outputs

- Transfac file for each pdb and averaged transfac file
- Distance file (distances per-column and overall)

Initial Relax

Inputs

1. <PATH_TO_XTAL_PDB> Crystallographic pdb of protein-peptide complex

Retrieve from pdb

2. **<PATH_TO_ENZDES_CSTFILE> (for proteases only)**

   Generate yourself based on protease catalytic geometry

3. **<PATH_TO_COO_CSTFILE> (for proteases only)**

   Use a modified version of sidechain_cst_3.py (at /source/src/apps/public/relax_w_allatom_cst/sidechain_cst_3.py in the Rosetta source code) to generate constraints with settings of 0.1 and 0.5 on the protease atoms.

4. **<RESFILE>**

   NATRO all, NATAA peptide residues

5. **<XML_FILE>**

   Sample xml:

```
<ROSETTASCRIPTS>
    <SCOREFXNS>
        <ScoreFunction name="myscore" weights="<SCORE_FUNCTION>".wts/>
    </SCOREFXNS>
    <TASKOPERATIONS>
        <ProteinInterfaceDesign design_chain2="0" modify_after_jump="1"
name="pido"/>
        <InitializeFromCommandline name="init"/>
        <ReadResfile name="rrf"/>
    </TASKOPERATIONS>
    <FILTERS/>
    <MOVERS>
```

```
        <AddOrRemoveMatchCsts cst_instruction="add_new" name="cstadd"/>
        <FastRelax name="fastrelax" repeats="8" scorefxn="myscore"
task_operations="pido,init">
            <MoveMap name="mm">
                <Chain bb="1" chi="1" number="2"/>
                <Chain bb="1" chi="1" number="1"/>
                <Jump number="1" setting="1"/>
            </MoveMap>
        </FastRelax>
        <TaskAwareMinMover bb="0" chi="1" jump="0" name="min_pro"
scorefxn="myscore" task_operations="rrf"/>
        <PackRotamersMover name="repack" task_operations="rrf"/>
        <ConstraintSetMover name="protease_cst"/>
    </MOVERS>
    <APPLY_TO_POSE/>
    <PROTOCOLS>
        <Add mover_name="protease_cst"/>
        <Add mover_name="repack"/>
        <Add mover_name="min_pro"/>
        <Add mover_name="cstadd"/>
        <Add mover_name="fastrelax"/>
    </PROTOCOLS>
</ROSETTASCRIPTS>
```

6. <PATH_TO_FLAGS>

-mute core.io.database

-packing::use_input_sc

-packing::extrachi_cutoff 1

-packing::ex1

-packing::ex2

-linmem_ig 10

-out:file::output_virtual

Process

Run on initial crystallographic pdb to get rid of internal clashes.

Command Line:

<ROSETTA_BIN>rosetta_scripts.static.linuxgccrelease -jd2:ntrials 1 -nstruct 1000 -parser:protocol <XML_FILE> -database <ROSETTA_DB> -s <PATH_TO_XTAL_PDB> -run:preserve_header -enzdes::cstfile <PATH_TO_ENZDES_CSTFILE> -constraints:cst_file <PATH_TO_COO_CSTFILE> -resfile <PATH_TO_RESFILE> @<PATH_TO_FLAGS>

Outputs

1000 “relaxed” pdb files. Use lowest scoring pdb file as input for the next step.

Remarks

Differences between protease and PRD:

Protease:

command line includes: -enzdes::cstfile <PATH_TO_ENZDES_CSTFILE> -

constraints:cst_file <PATH_TO_COO_CSTFILE>

<SCORE_FUNCTION>: talaris2013_cst

PRD:

command line does not include constraint parameters

<SCORE_FUNCTION>: talaris2013

Thread Peptide-FastRelax

Inputs

1. **<STARTING_RELAXED_MODEL> Lowest scoring pdb from Initial Relax step.**
2. **<PATH_TO_ENZDES_CSTFILE> (for proteases only)**

   Generate yourself based on protease catalytic geometry
3. **<RESFILE>**

   NATRO all, NATAA peptide residues
4. **<XML_FILE>**

   Sample xml:

   <ROSETTASCRIPTS>

   <SCOREFXNS>

   </SCOREFXNS>

   <TASKOPERATIONS>

```
    <ProteinInterfaceDesign name="pido" design_chain2="0"
modify_after_jump="1" />
    <InitializeFromCommandline name="init"/>
    <ReadResfile name="rrf" filename=<RESFILE> />
  </TASKOPERATIONS>

  <FILTERS>
  </FILTERS>

  <MOVERS>
   <MutateResidue name="mut1" target="<PEPT_RES1>" new_res="DM1"/>
   <MutateResidue name="mut2" target="<PEPT_RES2>" new_res="DM2"/>
   <MutateResidue name="mut3" target="<PEPT_RES3>" new_res="DM3"/>
   <MutateResidue name="mut4" target="<PEPT_RES4>" new_res="DM4"/>
   <MutateResidue name="mut5" target="<PEPT_RES5>" new_res="DM5"/>
   <MutateResidue name="mut6" target="<PEPT_RES6>" new_res="DM6"/>
   <MutateResidue name="mut7" target="<PEPT_RES7>" new_res="DM7"/>
   <AddOrRemoveMatchCsts name="cstadd" cst_instruction="add_new" />
   <FastRelax name="fastrelax" repeats="8" task_operations="pido,init">
   <MoveMap name="mm">
            <Chain number="2" chi="1" bb="1"/>
            <Chain number="1" chi="1" bb="0"/>
            <Jump number="1" setting="1"/>
```

```
        </MoveMap>
        </FastRelax>
        <PackRotamersMover name="repack" task_operations="rrf"/>

    </MOVERS>
    <APPLY_TO_POSE>
    </APPLY_TO_POSE>

        <PROTOCOLS>
            <Add mover_name="mut1"/>
            <Add mover_name="mut2"/>
            <Add mover_name="mut3"/>
            <Add mover_name="mut4"/>
            <Add mover_name="mut5"/>
            <Add mover_name="mut6"/>
            <Add mover_name="mut7"/>
            <Add mover_name="cstadd"/>
            <Add mover_name="repack"/>
            <Add mover_name="fastrelax"/>
    </PROTOCOLS>
    </ROSETTASCRIPTS>
```

5. **<PATH_TO_FLAGS>**

    -mute core.io.database

-packing::use_input_sc

-packing::extrachi_cutoff 1

-packing::ex1

-packing::ex2

-linmem_ig 10

-out:file::output_virtual

Process

Run on lowest scoring relaxed pdb from Initial Relax one time per substrate sequence. Substitute your peptide sequence for <PEPT_RES1>, etc. in xml script. Add more <MutateResidue> movers as needed. Generates 10 relaxed protease-peptide complexes with that substrate sequence threaded on. Select lowest-scoring complex from these 10 complexes for MFPred step.

Command Line:

<ROSETTA_BIN>rosetta_scripts.static.linuxgccrelease -nstruct 10 -jd2:ntrials 1 -parser:protocol <XML_FILE> -database /home/arubenstein/Rosetta/main/database/ <CONST_ARG> -s <STARTING_RELAXED_MODEL> -run:preserve_header -overwrite @<PATH_TO_FLAGS> -score:weights <SCORE_FUNCTION>

Outputs

10 “relaxed” pdb files. Use lowest scoring pdb file as input for the next step.

Remarks

Differences between protease and PRD:

Protease:

command line includes <CONST_ARG>: -enzdes::cstfile

<PATH_TO_ENZDES_CSTFILE>

<SCORE_FUNCTION>: talaris2013_cst

PRD:

command line does not include constraint parameters

<SCORE_FUNCTION>: talaris2013

MFPred

Inputs

1. **<PATH_TO_INPUT_PDB> Lowest scoring pdb from Initial Relax step.**
2. **<LIST_PDB_COMPLEXES>**

   List of paths to lowest-scoring pdbs for each of the Thread Peptide runs in the previous step.
3. **<RESFILE>**

   NATRO all, NATAA peptide residues that should not be designed (flanking residues), ALLAA peptide residues for which a specificity profile should be predicted.
4. **<XML_FILE>**

   Sample xml:

   <ROSETTASCRIPTS>

<TASKOPERATIONS>

<InitializeFromCommandline name="init" />

<ReadResfile name="rrf" />

</TASKOPERATIONS>

<SCOREFXNS>

</SCOREFXNS>

<FILTERS>

</FILTERS>

<MOVERS>

<GenMeanFieldMover name="boltz" threshold="5" lambda_memory="0.5" tolerance="0.0001" temperature="0.8" task_operations="rrf,init"/>

</MOVERS>

<APPLY_TO_POSE>

</APPLY_TO_POSE>

<PROTOCOLS>

<Add mover_name="boltz"/>

</PROTOCOLS>

</ROSETTASCRIPTS>

5. **<PATH_TO_FLAGS>**

-mute core.io.database

-packing::use_input_sc

-packing::extrachi_cutoff 1

-packing::ex1

-packing::ex2

-out:file::output_virtual

6. **<EXPT_SPEC_PROFILE> (optional)**

   Path to known (experimentally-derived) specificity profile. MFPred protocol will output certain distances from this profile in the log if this parameter is given.

7. **<ROT_NORM_PARAM> (optional)**

   This is the γ parameter described in the paper. The default is 0.8.

8. **<BB_AVERAGE_PARAM>**

   This is the γ parameter described in the paper. The default is 0.8.

Process

Run on backbone ensemble as generated in Thread Peptide step. Runs MFPred algorithm on residues that are designated as packed/designed in the TaskOperations.

Command Line:

<ROSETTA_BIN>rosetta_scripts.static.linuxgccrelease -database <ROSETTA_DB> -parser:protocol <XML_FILE> -s <PATH_TO_INPUT_PDB> -rot_norm_weight <ROT_NORM_PARAM> -bb_average_weight <BB_AVERAGE_PARAM> -spec_profile <EXPT_SPEC_PROFILE> -bb_list <LIST_PDB_COMPLEXES> -dump_transfac <PATH_TO_OUTPUT_TRANSFAC> -resfile <RESFILE> -nooutput true –score:weights talaris2013 @<PATH_TO_FLAGS>

Outputs

Log contains probabilities per rotamer, probabilities per amino acid, and distances from experimental specificity profile (if provided). If <PATH_TO_OUTPUT_TRANSFAC> is provided, dumps one transfac file per backbone, file with backbone Boltzmann probabilities, and one averaged transfac file for the ensemble as a whole.

Distances.py

Inputs

1. **Transfac file as output by MFPred**
2. **Experimental specificity profile**

Process

```
import os
import sys
import numpy as np
import math
from sklearn import metrics
import matplotlib.pyplot as plt
from pylab import *


def binarizeList ( firstList ):
    binary_freq = []


    choose_val = 0.10
```

```
    max_val = max(firstList)
    if max_val < 0.10:
        if max_val > 0.09:
          choose_val = 0.09
        elif max_val > 0.08:
            choose_val = 0.08
        elif max_val > 0.07:
            choose_val = 0.07

    for val in firstList:
        if val > choose_val:
            binary_freq.append( 1 )
        else:
            binary_freq.append( 0 )
    return binary_freq

def areaUnderCurve ( firstList, secondList ):
    binary_freq = binarizeList( firstList )
    fpr, tpr, _ = metrics.roc_curve(binary_freq, secondList)
    auc = metrics.auc(fpr,tpr)
    return auc

def shannonEntropy( firstList ):
```

```
    sE = -1.0 * np.sum( [ p * math.log(p,2) for p in firstList if p != 0.0 ] )
    return sE

def JSDivergence( firstList, secondList ):

    firstSE = shannonEntropy( firstList )

    secondSE = shannonEntropy( secondList )

    combList = [ 0.5 * fL + 0.5 * sL for fL,sL in zip(firstList, secondList) ]

    combSE = shannonEntropy( combList )

    return combSE - 0.5 * firstSE - 0.5 * secondSE

def cosineDist( firstList, secondList):

    dotP = np.dot(firstList, secondList)

    sqrt_1 = math.sqrt( np.sum( np.power( firstList,2 ) ) )
    sqrt_2 = math.sqrt( np.sum( np.power( secondList,2 ) ) )

    return dotP/(sqrt_1 * sqrt_2)
```

```
def frobDist( firstList, secondList):

    diff_lists = np.subtract(firstList,secondList)
    terms = np.power( diff_lists,2)
    return math.sqrt( np.sum( terms ) )

def aveAbsDist( firstList, secondList ):

    diff_lists = np.fabs( np.subtract( firstList, secondList) )
    return sum( diff_lists ) / len( diff_lists )

def readSpecProfileList( filename ):
    with open(filename) as transfac_file:
        transfac = transfac_file.readlines()

    motifWidth = len(transfac)-2

    aaAlpha = transfac[1].split()[1:]

    freq = [{k: 0.0 for k in aaAlpha} for i in range(motifWidth)]

    t_read = transfac[2:]
```

```
    for pos,line in enumerate( t_read,0 ):
        for aa_ind,f in enumerate( line.split()[1:], 0):
            freq[pos][aaAlpha[aa_ind]] = float(f)

    freqList = [ [ val for key,val in sorted(pos.iteritems()) ] for pos in freq ]

    return freqList

def main(args):
    infile = args[1]
    infile_expt = args[2]


    expt = os.path.basename(infile_expt).rstrip()
    expt = expt.rsplit('.',1)[0]

    tokens=infile.rsplit('.',1)
    file=tokens[0]

    outfile= '%s_dist.txt' % (file)
    outfile_heat= '%s_heat.png' % (file)
```

```
freq_in = readSpecProfileList( infile )

freq_expt = readSpecProfileList( infile_expt )

nda_freq_in = np.array( [ freq_in] )

nda_freq_expt = np.array( [ freq_expt] )

flat_freq_in = np.ndarray.flatten( nda_freq_in )

flat_freq_expt = np.ndarray.flatten( nda_freq_expt )


c = [ cosineDist( i, g ) for i,g in zip( freq_in, freq_expt ) ]

f = [ frobDist( i, g ) for i,g in zip( freq_in, freq_expt ) ]

a = [ aveAbsDist( i, g ) for i,g in zip( freq_in, freq_expt ) ]

jsd1 = [ JSDivergence ( i, g ) for i,g in zip( freq_in, freq_expt )]

auc = [ areaUnderCurve ( i, g ) for i, g in zip( freq_expt, freq_in )]

avg_c = cosineDist( flat_freq_in, flat_freq_expt )

avg_f = frobDist( flat_freq_in, flat_freq_expt )

avg_a = aveAbsDist( flat_freq_in, flat_freq_expt )

avg_jsd = np.sum(jsd1) / len(jsd1)

avg_auc = np.sum(auc) / len(auc)


c.append(avg_c)

f.append(avg_f)

a.append(avg_a)

jsd1.append(avg_jsd)

auc.append(avg_auc)
```

```
    dist_out = open(outfile,"w")


    dist_out.write("Metric\t")
    dist_out.write("\t".join([ "Col{0}".format(i) for i in xrange(1,len(c)) ]))
    dist_out.write("\tAvg\nCosine\t")


    dist_out.write("\t".join(map(str,c)))
    dist_out.write("\nFrobenius\t")
    dist_out.write("\t".join(map(str,f)))
    dist_out.write("\nAAD\t")
    dist_out.write("\t".join(map(str,a)))
    dist_out.write("\nJSD\t")
    dist_out.write("\t".join(map(str,jsd1)))
    dist_out.write("\nAUC\t")
    dist_out.write("\t".join(map(str,auc)))
    dist_out.write("\n")


    dist_out.close()


if __name__ == "__main__":
  main(sys.argv)
```

Outputs

Distances file: the name of this file is <INPUT_FILE>_dist.txt. Contains one line per metric. Each line contains one value per column and the last value is the average of the columns.

Non-MFPred pipeline software – used for controls and/or optimization of protocol:

Backbone Ensemble Generation

Thread Peptide Alone (pre-flexpepdock or pre-backrub)

Command Line:

<ROSETTA_BIN>rosetta_scripts.static.linuxgccrelease -nstruct 1 -jd2:ntrials 1 -parser:protocol <XML_FILE> -database <ROSETTA_DB> <CONST_ARG> -s <STARTING_RELAXED_MODEL> -run:preserve_header -overwrite @<PATH_TO_FLAGS>

Sample xml:

```
<ROSETTASCRIPTS>
    <SCOREFXNS/>
    <TASKOPERATIONS>
        <InitializeFromCommandline name="init"/>
        <ReadResfile filename=”<RESFILE>” name="rrf"/>
    </TASKOPERATIONS>
    <FILTERS/>
```

```
<MOVERS>
        <MutateResidue name="mut1" target="<PEPT_RES1>" new_res="DM1"/>
        <MutateResidue name="mut2" target="<PEPT_RES2>" new_res="DM2"/>
        <MutateResidue name="mut3" target="<PEPT_RES3>" new_res="DM3"/>
        <MutateResidue name="mut4" target="<PEPT_RES4>" new_res="DM4"/>
        <MutateResidue name="mut5" target="<PEPT_RES5>" new_res="DM5"/>
        <MutateResidue name="mut6" target="<PEPT_RES6>" new_res="DM6"/>
        <MutateResidue name="mut7" target="<PEPT_RES7>" new_res="DM7"/>
      <AddOrRemoveMatchCsts cst_instruction="add_new" name="cstadd"/>
      <PackRotamersMover name="repack" task_operations="rrf,init"/>
</MOVERS>
<APPLY_TO_POSE/>
<PROTOCOLS>
        <Add mover_name="mut1"/>
        <Add mover_name="mut2"/>
        <Add mover_name="mut3"/>
        <Add mover_name="mut4"/>
        <Add mover_name="mut5"/>
        <Add mover_name="mut6"/>
        <Add mover_name="mut7"/>
        <Add mover_name="cstadd"/>
        <Add mover_name="repack"/>
</PROTOCOLS>
```

</ROSETTASCRIPTS>

Resfile:

NATRO all, NATAA peptide residues

Flags:

-mute core.io.database

-packing::use_input_sc

-packing::extrachi_cutoff 1

-packing::ex1

-packing::ex2

-linmem_ig 10

-out:file::output_virtual

FlexPepDock

Command line:

<ROSETTA_BIN>_scripts.static.linuxgccrelease -parser:protocol

~/mean_field/xml/flexpepdock.xml -database <ROSETTA_DB> -s

<STARTING_THREADED_MODEL> -ex1 -ex2 -ex1aro -ex2aro -extrachi_cutoff 0 -

nstruct 10 -enzdes:cstfile <PATH_TO_ENZDES_CSTFILE> -score:weights

talaris2013_cst -run:preserve_header -packing:use_input_sc

Sample xml:

<ROSETTASCRIPTS>

```
    <TASKOPERATIONS>
    </TASKOPERATIONS>
    <SCOREFXNS>
    </SCOREFXNS>
    <FILTERS>
    </FILTERS>
    <MOVERS>
        <AddOrRemoveMatchCsts name="cstadd" cst_instruction="add_new" />
        <FlexPepDock name="fpd" pep_refine="1" />
    </MOVERS>
    <APPLY_TO_POSE>
    </APPLY_TO_POSE>
    <PROTOCOLS>
        <Add mover_name="cstadd"/>
        <Add mover_name="fpd"/>
    </PROTOCOLS>
</ROSETTASCRIPTS>
```

Backrub

Command line:

<ROSETTA_BIN>backrub_cst.linuxgccrelease -run:preserve_header -score:weights talaris2013_cst -database <ROSETTA_DB> -s <STARTING_THREADED_MODEL> -ex1 -ex2 -ex1aro -ex2aro -extrachi_cutoff 0 -backrub:minimize_movemap

<MOVEMAP_FILE> -backrub:ntrials 10000 -backrub:pivot_residues 215 216 217 218 219 220 221 222 223 224 -overwrite -enzdes:cstfile <PATH_TO_ENZDES_CSTFILE> -packing:use_input_sc

Movemap:

RESIDUE * CHI

JUMP * YES

CHAIN 2 BBCHI

Backrub_cst app:

This app is a version of the general backrub app that includes Enzdes style constraint as a mover. Currently, the general backrub app has been moved to a new Mover called BackrubProtocol mover – had this been available at the time of benchmarking, this would have been used instead.

Enumerate_dihedral

Command line:

<ROSETTA_BIN>enumerate_dihedral.linuxgccrelease -database <ROSETTA_DB> -s <STARTING_RELAXED_MODEL> -anchor_res <FIXED_RES_P1> -run:preserve_header

Enumerate dihedral app:

This app is in my pilot apps folder within the Rosetta source code (Rosetta/main/source/src/apps/pilot/arubenstein/enumerate_dihedral.cc).

Clustering via AmberTools cpptraj:

Run tleap to convert pdb to topology and coordinate files:

tleap.in:

source leaprc.ff14SB

source leaprc.phosaa10

loadAmberParams frcmod.ionsjc_tip3p

pdb = loadpdb <PDB_NAME>

addions pdb Cl- 0

addions pdb Na+ 0

#solvatebox pdb TIP3PBOX 10.0

saveamberparm pdb <PDB_NAME>.top <PDB_NAME>.crd

Run:

tleap -f tleap.in

Run cpptraj to cluster:

cpptraj.in file:

parm <TOPO_FILE_1>

trajin <COORD_FILE_1>

parm <TOPO_FILE_2>

trajin <COORD_FILE_2>

.

.

.

cluster hieragglo clusters <N_CLUSTERS> rms :<PEPT_BEG_RES>-<PEPT_END_RES> repout <N_CLUSTERS> repfmt pdb

Run:

cpptraj -i 'cpptraj.in'

Multispecificity Prediction Controls for MFPred

Monte-Carlo (pepspec)

Command-line:

<ROSETTA_BIN>mc_no_sa.linuxgccrelease -database <ROSETTA_DB> -pepspec:pdb_list <BACKBONE_ENSEMBLE_LIST> -save_low_pdbs false -pepspec:n_peptides 1 -pepspec:use_input_bb true -ex1 -ex2 -extrachi_cutoff 0 -pepspec:diversify_lvl 50 -pepspec:run_sequential -use_input_sc

Mc_no_sa app:

This app is a version of the general pepspec app that includes profiling (necessary to extract running times and determine speedup).

Genetic Algorithm (sequence_tolerance)

Command-line:

<ROSETTA_BIN>sequence_tolerance_control.linuxgccrelease -database <ROSETTA_DB> -s <MODEL_FROM_BACKBONE_ENSEMBLE> -ex1 -ex2 -ex1aro -ex2aro -extrachi_cutoff 0 -ms:generations 5 -ms:pop_size 2000 -ms:pop_from_ss 1 -ms:checkpoint:prefix <PREFIX> -ms:checkpoint:interval 200 -ms:checkpoint:gz -seq_tol:fitness_master_weights 1 1 1 2 -resfile <RESFILE>

Resfile:

NATAA residues according to seqtol_resfile.py, ALLAA peptide residues

Sequence_tolerance_control app:

This app is a version of the general sequence_tolerance app that includes profiling (necessary to extract running times and determine speedup).

Chapter 4: Biophysical determinants of mutational robustness in a viral molecular fitness landscape

4.1. Abstract

Biophysical interactions between proteins and peptides are key determinants of genotype-fitness landscapes, but an understanding of how molecular structure and residue-level energetics at protein-peptide interfaces shape functional landscapes remains elusive. Combining information from yeast-based library screening, next-generation sequencing and structure-based modeling, we report comprehensive sequence-energetics-function mapping of the specificity landscape of the Hepatitis C Virus (HCV) NS3/4A protease, whose function – site-specific cleavages of the viral polyprotein – is a key determinant of viral fitness. We elucidate the cleavability of 3.2 million substrate variants by the HCV protease and find extensive clustering of cleavable and uncleavable motifs in sequence space indicating mutational robustness, thereby providing a plausible molecular mechanism to buffer the effects of low replicative fidelity of this RNA virus. Specificity landscapes of known drug-resistant variants are similarly clustered. Our results highlight the key and constraining role of molecular-level energetics in shaping plateau-like fitness landscapes from quasi-species theory.

4.2. Introduction

RNA viruses, e.g., influenza, Hepatitis C Virus (HCV) and Human Immunodeficiency Virus (HIV), are under a heavy mutational load due to the extremely high error-rates of their RNA polymerases(Domingo and Holland, 1997; Holland et al., 1982; Lauring et al.,

2013). As a result of this low replication fidelity, these viruses exist as a population of variants called quasispecies (Andino and Domingo, 2015; Eigen, 1993), even within a single host individual (Cristina et al., 2007). While this genetic diversity and a large population size is believed to increase viral adaptive potential against antiviral therapies(Elde et al., 2012; Goldberg et al., 2012; Wilke et al., 2001), low replication fidelity may also lead to too many mutations, causing an "error catastrophe" and extinction(Eigen, 2002; Lauring and Andino, 2010). The underlying biomolecular structures and interactions in the virus must, therefore, be robust to genetic variability such that they provide a buffer against the deleterious impacts of a high mutational load (Elena et al., 2006; Masel and Siegal, 2009). Tawfik and co-workers have hypothesized that viral proteins possess "gradient robustness" in which individual mutations have small and largely additive effects on stability leading to a slower loss of function compared to "threshold robustness" exhibited by proteins in general (Tokuriki et al., 2009). It has been argued that mutational robustness may itself promote adaptiveness if the number of phenotypes accessible to a variant through mutation is smaller than the total number of phenotypes possible(Draghi et al., 2010; Wilke and Adami, 2003). How is mutational robustness encoded at the molecular level in RNA viruses such as HCV? How is structural integrity and interaction fidelity maintained in the face of a large mutational load, and what, if any, are the limits imposed by the underlying molecular interactions on mutational robustness and adaptive potential? The degeneracy of the genetic code, the thermodynamic and kinetic stabilities of RNA and proteins, and the presence of molecular chaperones, may all contribute to the robustness of the structures of individual viral biomolecules (Lauring et al., 2013). However, how viral protein-based interactions,

especially those that are critical for viral propagation, encode “fuzziness” (Tokuriki et al., 2009) leading to mutational robustness at the molecular level is not well understood.

At the molecular level, the balance between mutational robustness and functional plasticity is encapsulated in the notion of molecular fitness landscapes(Smith, 1970), which are high-dimensional maps that relate the function of individual biomolecular variants to their functional and/or evolutionary fitness (de Visser and Krug, 2014; Wright, 1931). Analysis of mutational trajectories on these landscapes provides insight into the constraints placed on evolution by the physiochemical properties of biomolecules, allowing, in principle, reconstruction as well as forward prediction of molecular evolution (Bridgham et al., 2006; Harms and Thornton, 2013; Kondrashov and Kondrashov, 2015; Romero and Arnold, 2009; Weinreich et al., 2006). The molecular fitness landscape has long been theoretically postulated (Smith, 1970) and recent empirically determined sequence-function mappings of proteins (Bandaru et al., 2017; Firnberg et al., 2014; Fowler et al., 2010; Hietpas et al., 2011; Kim et al., 2013; McLaughlin et al., 2012; Podgornaia and Laub, 2015; Sarkisyan et al., 2016; Wrenbeck et al., 2017) have enabled the partial construction of fitness landscapes. These reconstructed landscapes permit testing of possible evolutionary scenarios and provide insight into properties such as mutational robustness and non-additivity (epistasis) of mutational effects (Blanquart and Bataillon, 2016; Breen et al., 2012; Harms and Thornton, 2013; Hartl, 2014; Sailer and Harms, 2017a; Thyagarajan and Bloom, 2014; Weinreich et al., 2013; Wu et al., 2016). Empirical sequence-function relationships also enable

biomolecular engineering for new or improved functions (Jenson et al., 2017; Klesmith et al., 2015; McLaughlin et al., 2012; Tinberg et al., 2013; Whitehead et al., 2012).

Typically, sequence-function mapping of proteins and protein-protein interactions described above involves partial enumeration of the possible sequence diversity (for example, all single mutations and a subset of double mutations at a large number of protein residue positions) and high-throughput functional evaluation coupled with deep sequencing(Fowler and Fields, 2014; Klesmith et al., 2017; Reich et al., 2015). Statistical and/or biophysical models can be used to make inferences about the regions of sequence space not sampled(Jenson et al., 2017; Klesmith et al., 2017). However, comprehensive construction of the fitness landscape requires enumeration and evaluation of the complete sequence diversity (all higher-order mutations at all residue positions). Laub and co-workers have pioneered studies in which the entire combinatorial diversity is experimentally sampled, albeit at a smaller number of positions(Aakre et al., 2015; Podgornaia and Laub, 2015). The astronomical size of sequence space, however, makes the comprehensive experimental evaluation of sequence-function landscapes with any one experimental approach difficult. Computational biophysical methods may, in principle, assist in creation and analysis of functional and fitness landscapes (Rodrigues et al., 2016). Indeed, evolutionary landscapes of simple protein models, such as lattice models, have been extensively investigated using biophysical evolutionary theory and computational simulations(Bloom et al., 2004; Bornberg-Bauer and Chan, 1999; DePristo et al., 2005; Ding and Dokholyan, 2006; Drummond and Wilke, 2008; Echave and Wilke, 2017; Manhart and Morozov, 2015; Sailer and Harms, 2017b; Sikosek and Chan, 2014;

van Nimwegen et al., 1999; Yang et al., 2012), and deep connections with population genetics theories have been discovered (Bershtein et al., 2017; Echave and Wilke, 2017; Serohijos and Shakhnovich, 2014). While pioneering and crucial insights have been obtained in these studies, chemically realistic atomic-resolution structure-based elucidation of functional landscapes has not been performed so far, due both to high computational cost as well as inaccuracies in simulation force fields which preclude accurate biophysical evaluation of mutational effects on protein-protein interactions.

Here, we use a combination of experimental (biochemical) and computational techniques to elucidate the specificity landscape of the interaction between HCV NS3/4A protease enzymes and its substrates. This enzyme-substrate interaction is key for viral maturation as it cleaves exclusively at four specific sites in the viral polyprotein (Figure 3.1A) to release individual non-structural proteins(Scheel and Rice, 2013), and also mediates inactivation of key human immunity proteins(Meylan et al., 2005). The cleavage specificity of the protease is thus a key determinant of viral fitness, and its proper functioning includes negative specificity – the lack of cleavage of non-canonical sites on the viral protein and of most host cell proteins (Figure 3.1A). The molecular interactions underlying both positive and negative specificities must be robust to mutations as the HCV virus RNA polymerase has a high error-rate (Powdrill et al., 2011), but how and whether this robustness is encoded in the protease-substrate interactions is not known. Using yeast surface display, next-generation sequencing and a machine-learning approach which combines features from experimental data and atomistic computational simulations (utilizing the Rosetta and Amber force fields) that we recently developed

(Pethe et al., 2017; Rubenstein et al., 2017), we construct the specificity landscape (with cleavability assignments made for 3.2 million substrate pentapeptide sequences) of the HCV NS3/4A protease and three of its known drug-resistant variants(Romano et al., 2012). We demonstrate that energetic features of protease-substrate interactions inherently encode mutational robustness, and that the connectivity patterns in the specificity landscape may act as a "biophysical capacitor" for maintaining protease function in the face of high mutational load.

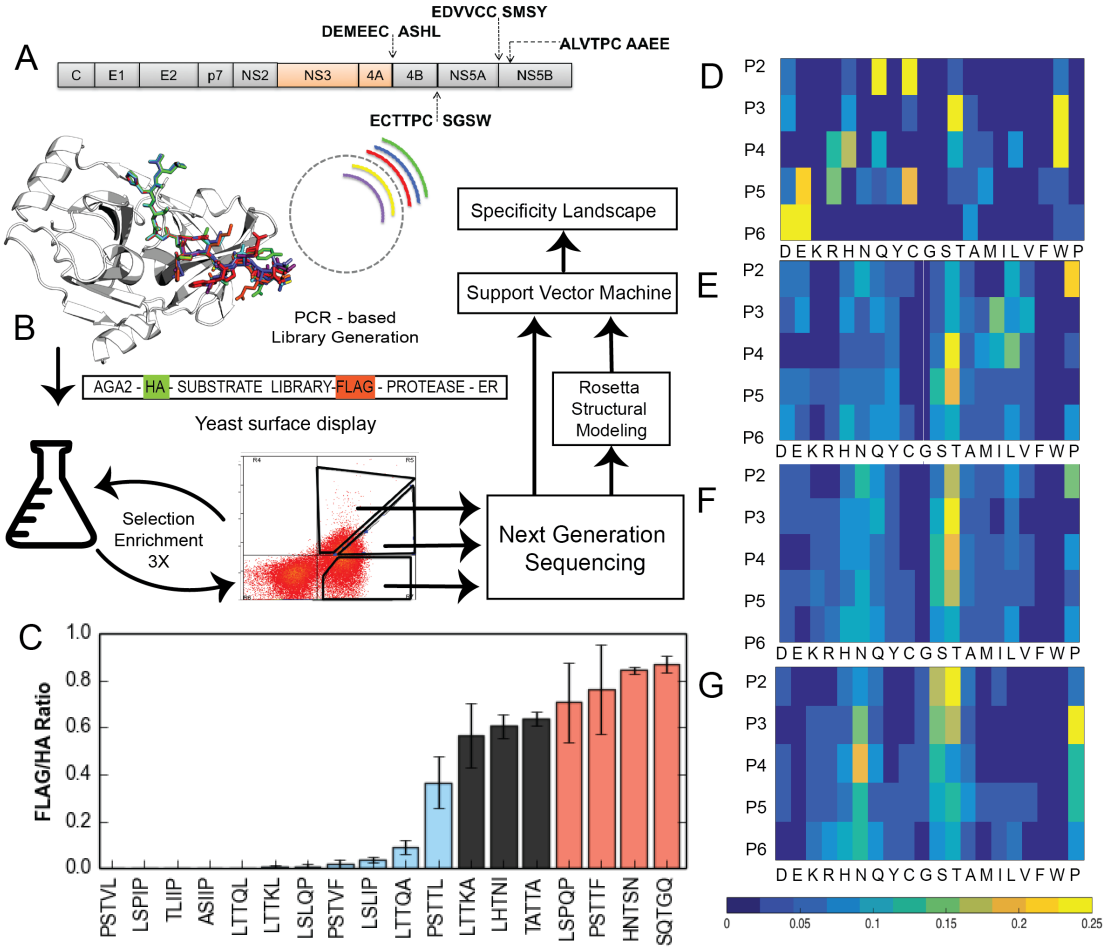


Figure 3.1. Overview of experimental workflow, validation of results
(A) The HCV viral polyprotein depicting marked biological cleavage sites for the HCV NS3/4A protease (B) overview of the experimental and computational workflow. (C) Validation of FACS gates for cleaved, partially cleaved and uncleaved sequences using

yeast surface display assay (D) Sequences taken from in vivo samples of HCV patients (8726) as compared to (E) sequences determined by our assay as cleaved(7472), (F) as partially cleaved (8737) and (G) as uncleaved (14702)

4.3. Results

HCV NS3/4A protease is known to cleave four canonical cleavage sites on the hepatitis C viral polyprotein (Fig. 3.1A), causing a cascade of viral assembly and maturation events. These cleavages (and a lack of cleavage of other parts of the polyprotein) are thus, critical for viral fitness. The high mutational load on the HCV polyprotein can lead to sequence variation in both the protease and substrate regions (Geller et al., 2016a). At the protein level, the distribution of mutational effects in a folded protein (protease) are modulated by both the thermodynamic stability and function (binding and cleavage), while the peptide substrate regions, which are found in flexible linker regions of the HCV polyprotein and connect component proteins, do not have a native tertiary structure. Therefore, we reasoned that a more direct sequence-cleavability mapping can be made for diversity in the substrate region without the need to additionally deconvolute the contribution from stability effects on tertiary structure. Secondly, it is more feasible to enumerate and evaluate by sequencing the substrate combinatorial diversity due to its shorter length (~7 residues) compared to the protease (>200 residues). Therefore, we mapped the viral protease-substrate interaction landscape for the HCV NS3/4A protease by considering all possible pentapeptide sequence combinations in its sequence recognition site at positions P6 through P2 following the Schechter and Berger nomenclature(Schechter and Berger, 1967). Positions P1 and P1', between which the scissile bond is present, were maintained as C and A, respectively, in this study. In the rest of this chapter, we refer to individual pentapeptide patterns (e.g. the canonical

cleavage sites DEMEE, EDVVC, ECTTP, ALVTP) and omit the identity of the P1,P1' residues.

4.3.1. Exploration of the (P6-P2) specificity landscape of the HCV NS3/4A protease reveals a diverse specificity profile

To mimic the viral intrachain arrangement of substrate libraries and the protease, we utilized a modified version of the assay described by Iverson, Georgiou and co-workers(Yi et al., 2013) as depicted in Figure 3.1B. A mutagenic library was created incorporating degenerate codons at P6-P2 specificity defining substrate positions (Benatuil et al., 2010; Kowalsky et al., 2015). In our assay, substrates are transported to the surface of yeast cells in a cleavage-dependent manner: the degree of cleavage is estimated by measuring the relative levels of substrate-flanking FLAG and HA tags using fluorescent, labeled antibodies. We have previously used this assay to test known and novel substrates of the HCV protease (Pethe et al., 2017). A first round of yeast surface display assay and Fluorescence Assisted Cell Sorting (FACS) was performed with an inactive protease variant (S139A) to select for high expression of library variants, for removing sequences containing stop codons in the substrate region, and to deplete substrate sequences that are cleaved by yeast ER proteases (Li et al., 2017).

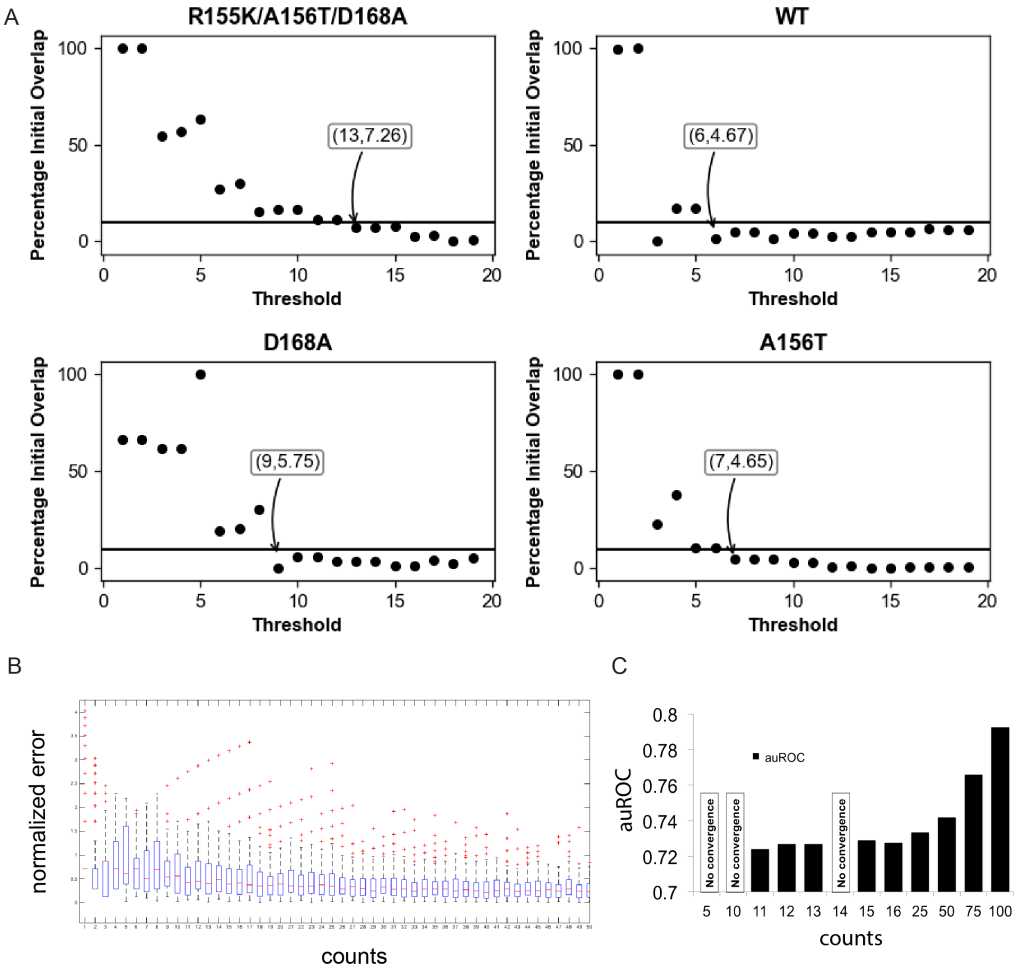


Figure 3.2. Threshold determination
(A) Threshold vs. percentage of initial overlap between cleaved and uncleaved sequences for all variants. The final threshold beyond which all other thresholds have a percentage overlap that is <= 10% is marked with an arrow (B) Duplicate population analysis. Normalized error is calculated for technical duplicates of cleaved samples by the formula: | (counts_S2- counts_S)|/ counts_S2, where sample S and S2 are technical duplicates (C) the Area Under the Curve for the ROC plot, when the SVM is used to classify cleaved versus uncleaved sequence pools at various count thresholds.

The resulting substrate variants from the pre-selection were subjected to rounds of yeast surface display assay and FACS with an active protease containing construct to select cleaved, partially cleaved and uncleaved variants using three sorting gates (Figure 3.1B), based on the relative levels of anti-HA and anti-FLAG fluorescence values (FLAG/HA

ratio, ranging between 0, for completely cleaved, and 1, for completely uncleaved). Sorting gates were defined based on the distribution of populations observed for known cleaved and uncleaved sequences (Pethe et al., 2017). This procedure was coupled with rounds of growth and selection to improve signal:noise for variants in each pool. Sequence profiles of the unselected population and isolated functional variants were determined using next-generation sequencing technology (Illumina NextSeq). Analysis of unique sequences in all sequenced pools showed that we identified a total of ~1.3 million sequences corresponding to ~30% of the possible amino acid diversity (3.2 million; Supplementary Methods). Analysis of sequencing and technical duplicates as well as overlap between the sequence pools was used to determine a count threshold (raw count 11) to remove noise from the sequencing data (Methods, Figure 3.2). Based on these criteria, we identified 7472, 8737 and 14702 unique pentapeptide sequences in the cleaved, partially cleaved and uncleaved pools. In parallel, we performed Rosetta simulations on all 3.2 million sequences in the P6-P2 region, and used a Support Vector Machine to predict the complete protease-substrate interaction landscape using sequence information procured from the aforementioned library and Rosetta-generated energetic features (Figure 3.1B).

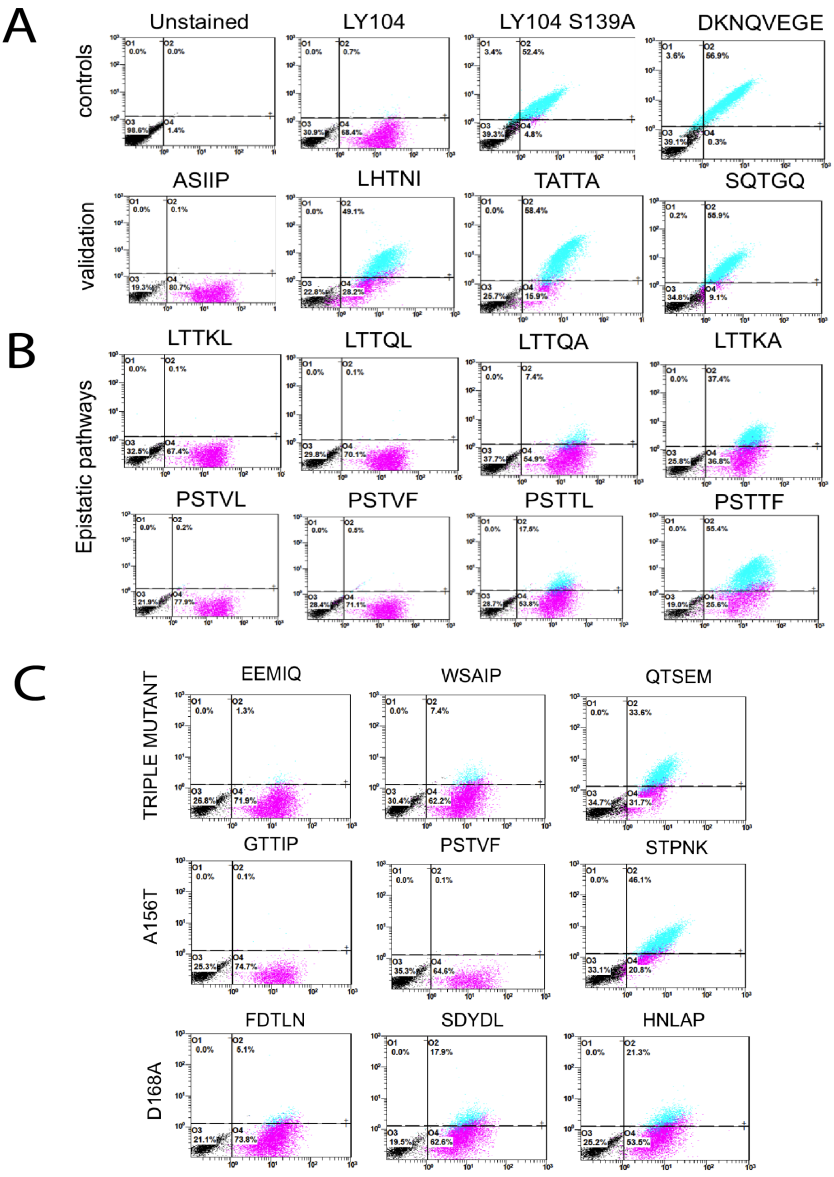


Figure 3.3. 2D plots of anti HA and anti FLAG antibody signals seen in the flow cytometry assay
(A) Display controls (B) Epistatic pathway validation (C) Drug resistant mutant validation plots

Several novel substrates identified from the three variant populations were tested as clonal populations in the yeast surface display assay system (Figure 3.1C, Figure 3.3) to validate that individual sequences fall into the gates used for selection from the library (Figure 3.4A-C). A subset of these sequences was also tested in vitro to ensure that the

cleavage properties observed in the yeast system were reproduced with purified protease and substrates (Figure 3.4D).

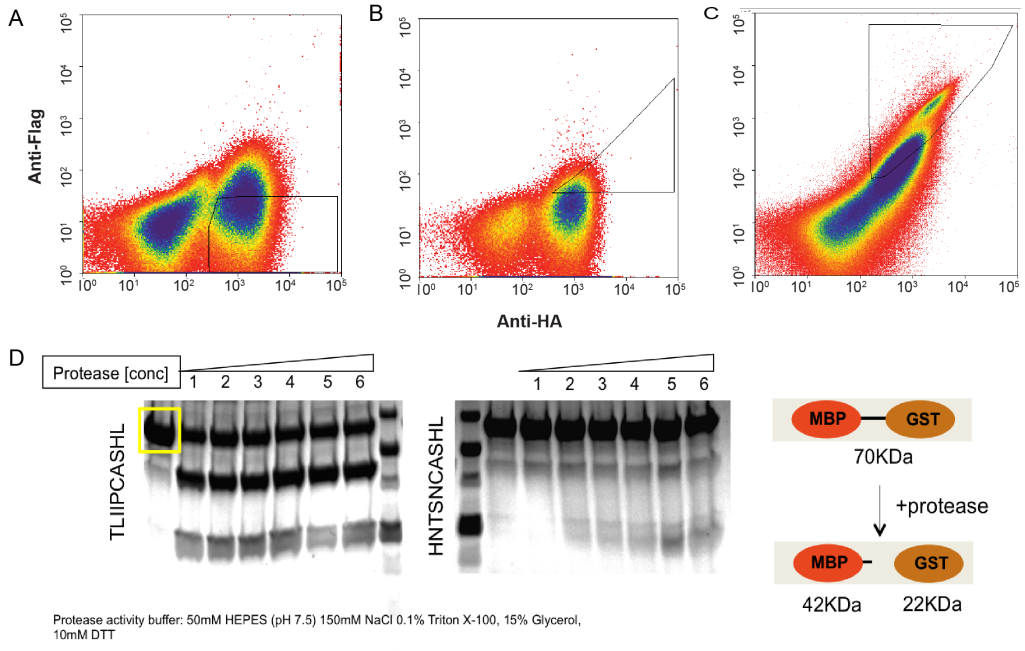


Figure 3.4. Flow cytometry 2D plots showing anti HA and anti FLAG stains for cell populations collected after enrichment round three
(A) Plot showing gate and cell population for cleaved (B) partially cleaved and (C) uncleaved populations (D) in vitro gel based assay using an MBP- GST fusion protein (70KDa). Upon overnight incubation with increasing concentrations of the protease – 500 nM, 700 nM, 1 µM, 2µM, 3µM, 4µM (well#1 through #6) results in cleavage for substrate TLIIPCASHL whereas HNTSNCASHL displays no cleavage

We next analyzed the profiles of sequences in each pool. For the cleaved sequence pool, the obtained substrate sequence ensemble has greater diversity compared to substrates identified from viral genomes sequenced from patient populations (Supplementary Methods, Figure 3.1D). For example, we observe that a more diverse subset of amino acids is tolerated at substrate positions P6 and P5 in our cleaved and partially cleaved pools (Figure 3.1E, F) whereas the patient isolated genomes display a high enrichment of Asp and Glu specifically at these positions. Another notable difference observed was the enrichment of small hydrophilic residues (Figure 3E, F), Ser (at P5) and Thr (at P4) in the

cleaved and partially cleaved populations, in contrast to enrichment at P3 and P2 in the uncleaved population (Figure 3.1G). Strikingly, we found prolines enriched at position P2 in the cleaved and partially cleaved populations and at P3 in the uncleaved populations, which corresponds well with the fact that 2 out of 4 canonical cleaved sequences have proline at P2 (ECTTP, ALVTP). While some of the above trends are also reflected in the sequences we tested in the course of our method validation (Figure 3.1C), it is evident that individual positional enrichments cannot be directly used to predict the pool assignments of individual sequences. For example, His is enriched at P6 in the cleaved sequence pool, however the sequence HNTSN is experimentally determined to be in the uncleaved pool (Figure 3.1C, Figure 3.4). While individual positional preferences of amino acids are useful, these results clearly indicated that molecular recognition between the protease and substrate pools is highly (sequence) context-dependent. We concluded that interactions between amino acids at various substrate positions (mediated possibly through interaction networks in the protease) influence the cleavability, thereby motivating the need for an analysis of the determined specificity landscape using properties of whole pentapeptide sequences.

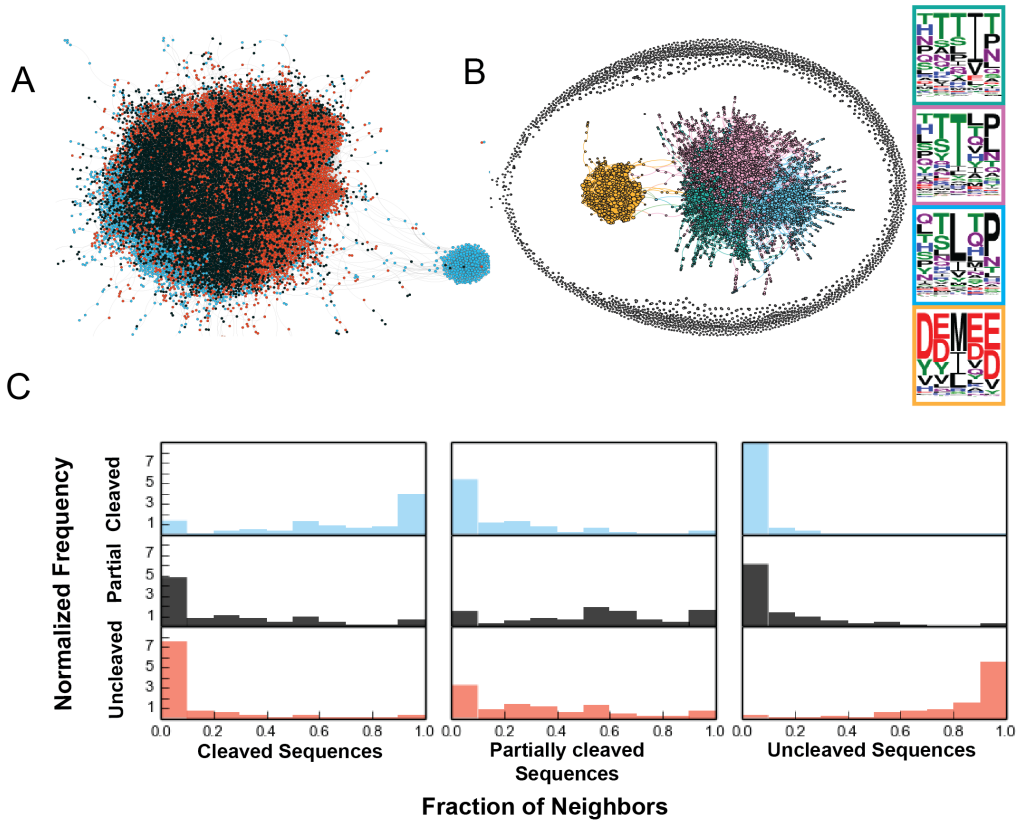


Figure 3.5. Force directed graph representation of experimental landscape; Neighbor analysis
(A) Force - directed graph of amino acid sequence space. Blue nodes are cleaved, red are uncleaved, and black is partially cleaved. Edges connect nodes that are within one hamming distance of each other (B) Force- directed graph of cleaved sequence. Colors denote clusters which are shown as specificity profiles outlined in the same color as the corresponding cluster (C) Frequency of neighbors for cleaved, partially cleaved, and uncleaved sequences denoting cleaved neighbors shown in blue bars, uncleaved neighbors depicted in red and partially cleaved neighbors depicted as black.

4.3.2. Clustering among cleaved, partially cleaved and uncleaved substrates

To visualize the functionally labeled sequence space of the experimentally derived substrates, we generated a force-directed graph (Figure 3.5A) (Amat, 2016; Jacomy et al., 2014) in which each node represents a sequence and is colored according to the functional pool to which it belongs. Nodes are connected by an edge if they differ by one

amino acid (Hamming distance = 1). Cleaved substrates exhibit significant clustering in the resulting graph (Figure 3.5A). To examine the landscape in greater detail around the cleaved sequences, we generated a sub-graph of the cleaved sequences (Figure 3.5B). We identified four clusters in this graph using the Gephi (Amat, 2016) modularity algorithm and determined corresponding profiles for each cluster. One identified cluster is clearly related to a canonical substrate, DEMEE. The other three clusters appear to have similarities with the other three canonical substrates (ALVTP, ECTTP, and EDVVC) but are less distinct from each other compared to the DEMEE cluster. These results indicate that the four canonical cleaved sites in the viral polyprotein are all members of mutationally robust clusters. Single amino acid changes within the cluster lead to other cleaved sequences, thereby buffering the impact of the heavy mutational load on the virus.

To determine if this clustering behavior observed in the cleaved sequence pool is also found in the partially cleaved and uncleaved pools, we calculated the fraction of neighbors for sequences with neighbors that belong to the same functional pool (Figure 3.5C). We find that similar to cleaved sequences, uncleaved sequences are also most frequently surrounded by uncleaved neighbors indicating clustering behavior for this functional pool as well. On average, cleaved sequence neighbors are 66.4% cleaved, and uncleaved sequence neighbors are 83.3% uncleaved. Partially cleaved sequences are the least clustered among the three pools, having on average 53% neighbors belonging to the same pool. These distributions indicate that in the specificity landscape, clusters of partially cleaved sequences surround clusters of cleaved and uncleaved ones.

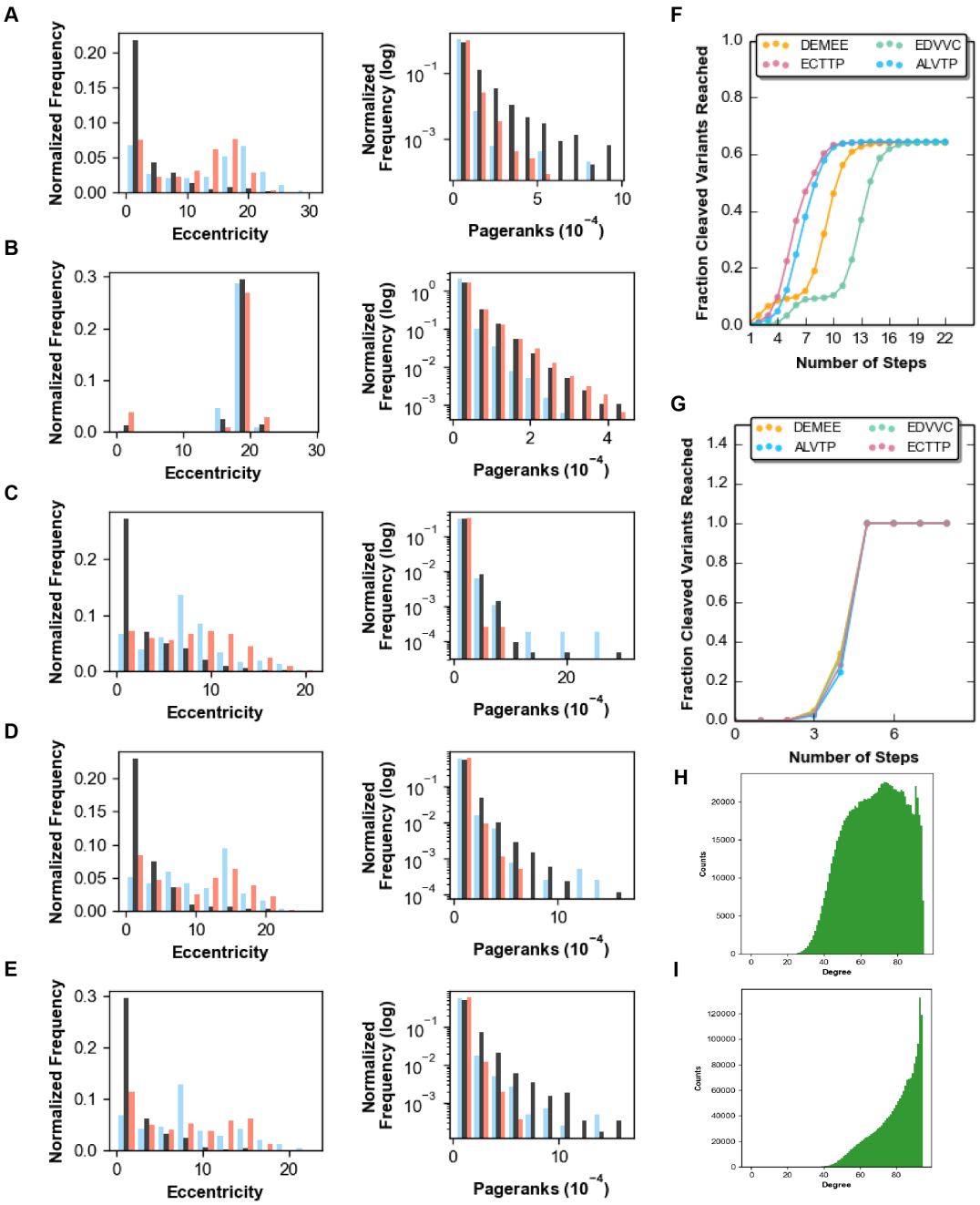


Figure 3.6. Graph metrics for WT and mutant protease
Cleaved (blue), uncleaved (red) and partially cleaved (black) graph metrics for (A) wild type HCV (B) randomly generated graph (C) R155K/A156T/D168A triple mutant (D) A156T and (E) D168A. Partially cleaved sequences generally have higher pageranks and lower eccentricity. Number of mutations vs. fraction cleaved variants reached for (F) experimental and (G) SVM- generated graphs. (H) Degree distribution for cleaved sequences in SVM derived graph (I) degree distribution for uncleaved sequences

To delineate how the three functional populations, which appear to be individually clustered in sequence space, are connected to each other, we used the PageRank metric (Brin and Page, 1998). This metric predicts the likelihood of reaching a node given a random walk on the substrate specificity landscape starting from a chosen sequence. Strikingly, partially cleaved substrates have higher pageranks (Figure 3.6A) than either cleaved or uncleaved substrates, indicating that they are most likely to be reached on long unbiased evolutionary trajectories starting from the canonical cleaved sequence DEMEE, the sequence that was used as the template for library generation. These connectivity patterns imply that partially cleaved node clusters may act as an evolutionary buffer on the substrate landscape, thereby enhancing robustness.

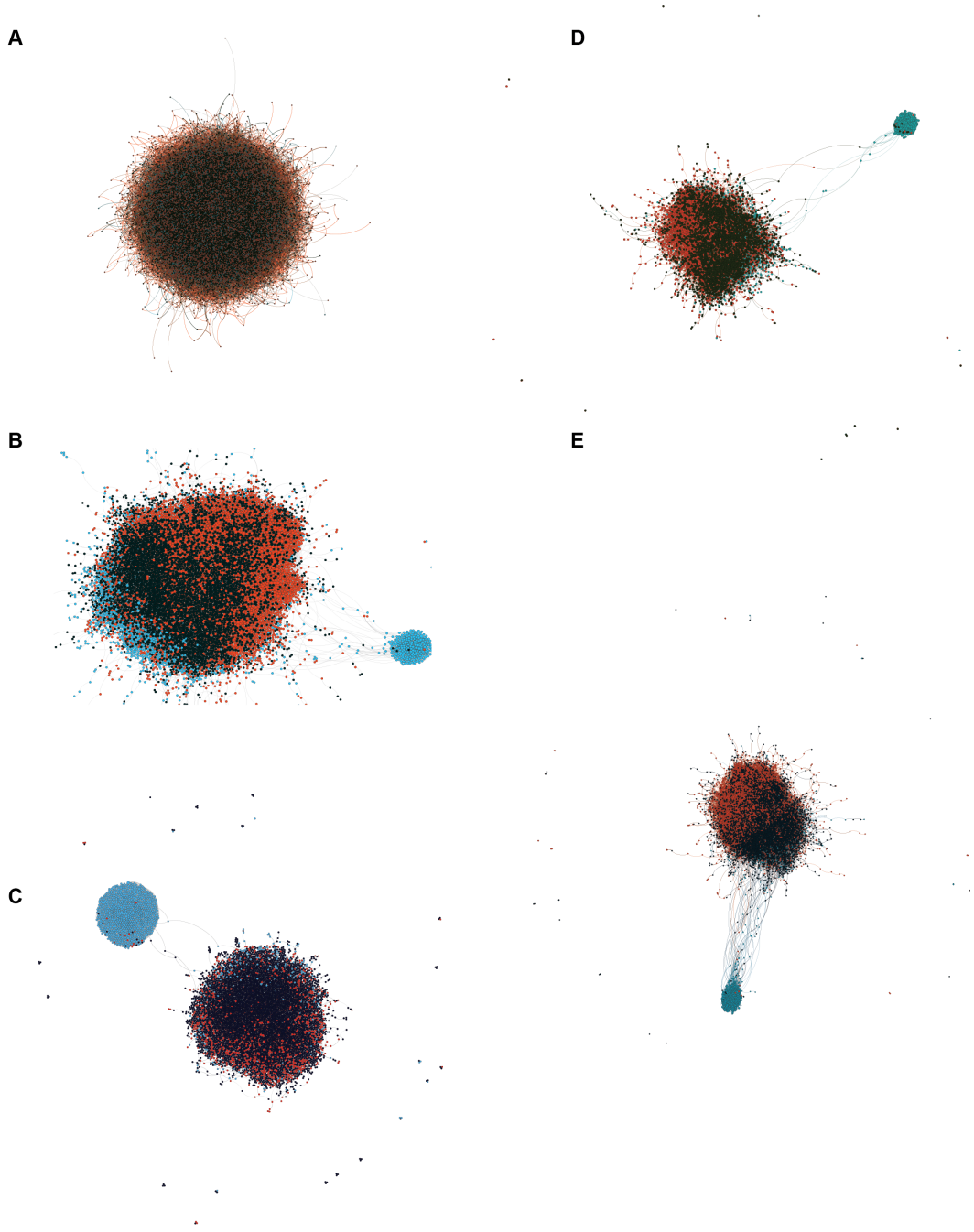


Figure 3.7. Force – directed graphs for WT and mutant proteases
(A) Randomly generated graph (B) wild type HCV protease (C) R155K/A156T/D168A triple mutant (D) D168A variant (E) A156T mutant

The graph generated by the experimentally derived sequences is incomplete (~30,000 nodes out of the 3.2 million possible). To test if the observed clustering and PageRank distributions are an artifact of the limited sampling in the experiment, we generated a

control random graph (Figure 3.7A) with the same number of nodes and edges, but having a randomly rewired connectivity. Both partially cleaved and uncleaved sequences are found to have higher pageranks than cleaved sequences in this random graph, indicating that the higher pageranks of partially cleaved sequences than cleaved and uncleaved sequences in the original experimental graph is significant.

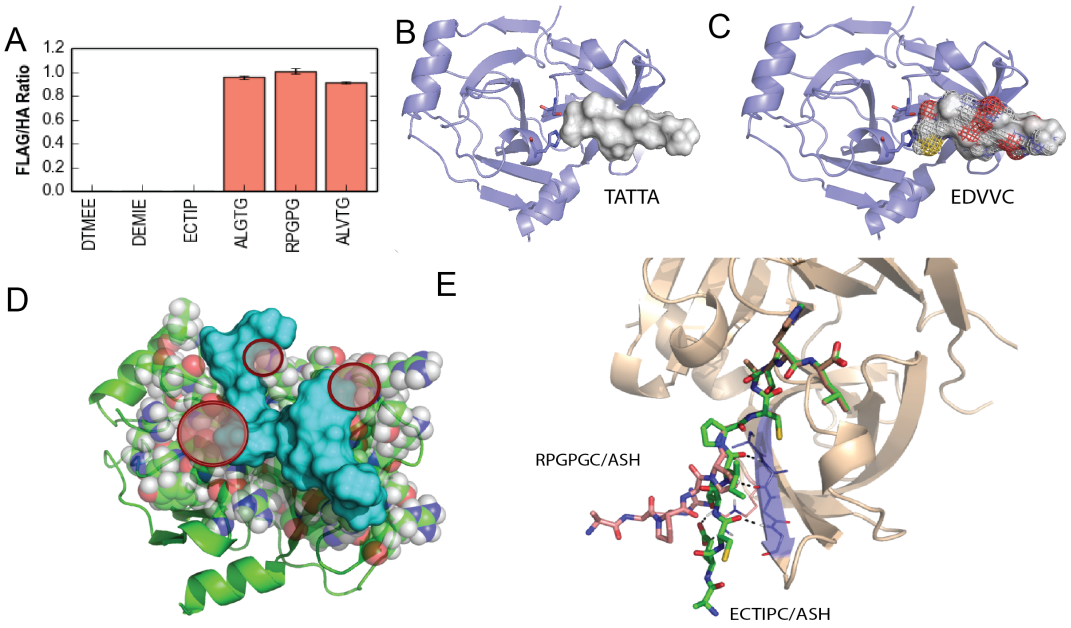


Figure 3.9. Structural basis for SVM prediction & validation
(A) Validation assay performed for three predicted cleaved and uncleaved sequences using a yeast surface display based technique (B) and (C) depict the volume occupied by TATTA and EDVVC, EDVVC occupies an optimal volume, making good contacts with the protease residue side chains. TATTA fits in the available space but does not make optimal contacts, thus resulting in suboptimal interaction energetics making TATTA a suboptimal substrate (D) Peptide (surface shown in blue) "FWPPM" sterically clashing against the protease chain (E) Structure of two models, ECTIP (cleaved) and RPGPG(uncleaved)

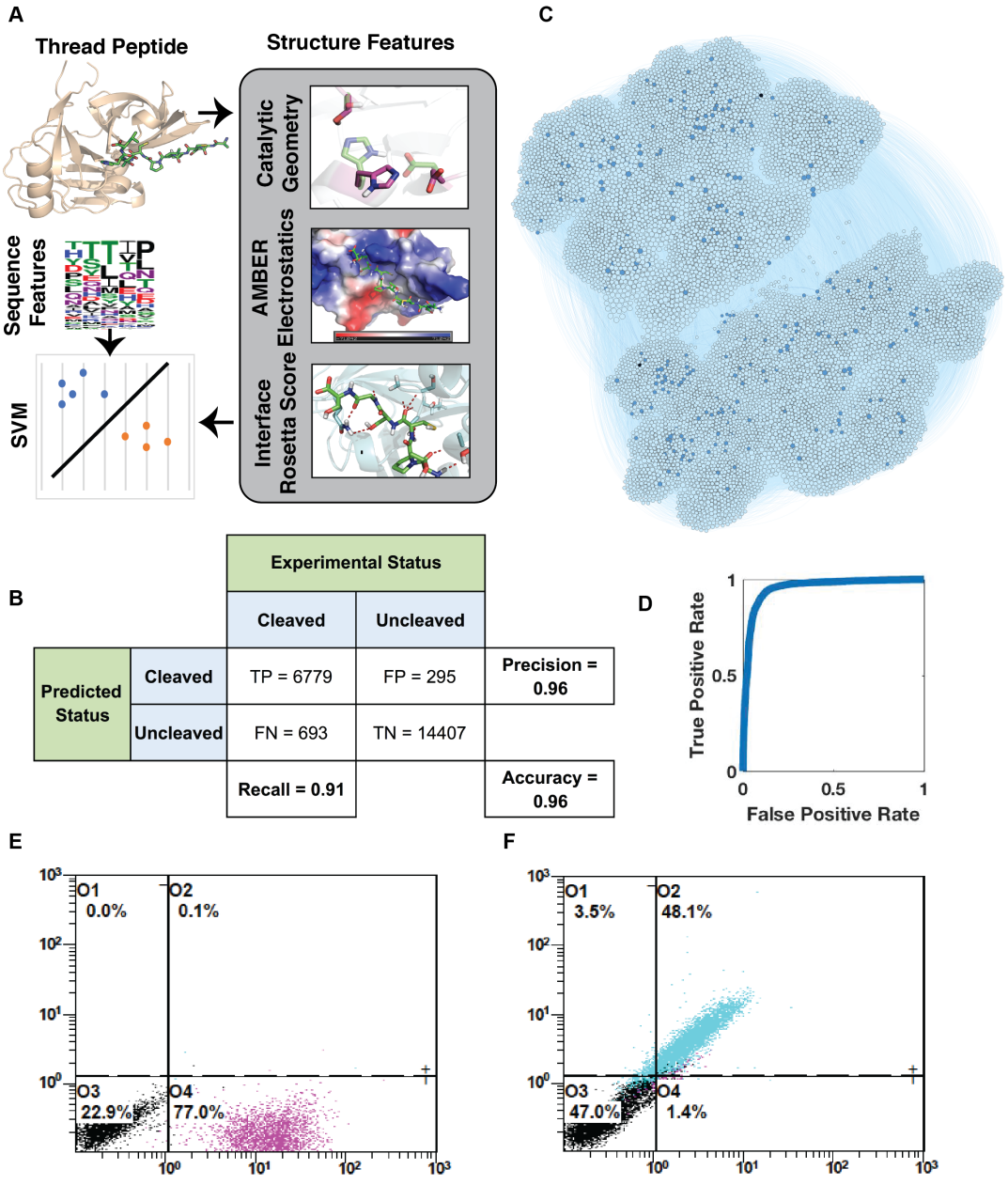


Figure 3.8. SVM generation workflow, contingency table and validation results
(A) Schematic workflow for SVM generation (B) Sub-graph of SVM predicted cleaved sequences with a distance > 2 from the SVM hyperplane. Experimental cleaved sequences are dark blue and experimental partially cleaved sequences are depicted as black. (C) Contingency table for SVM prediction (D) ROC plot of cross-validation on training set for SVM (E) Flow cytometry plot for ECTIP (SVM- predicted cleaved) (F) Flow cytometry plot for RPGPG (SVM – predicted uncleaved)

4.3.3. Energetic features derived from Rosetta modeling enable reconstruction of the complete protease-pentapeptide substrate landscape

While the experimentally-derived populations of the cleaved, partially cleaved and uncleaved sequences revealed striking clustering patterns in sequence space, it is not clear if these connectivity patterns would be preserved in a complete graph containing the complete diversity at five positions (3.2 million sequences). Therefore, to predict cleavability of all possible 3.2 million sequences in the interaction landscape, we used a Support Vector Machine (SVM)-based method that we developed previously (Pethe et al., 2017). Briefly, each sequence was threaded onto a bound complex based on a modeled near-attack conformation a crystal structure of the protease, and the complex was then relaxed to maintain favorable catalytic geometry. Energy evaluation of each of the 3.2 million complexes was performed using Rosetta and Amber simulation packages. A binary classification (cleaved/uncleaved) SVM was trained on a subset of experimentally identified sequences that passed a more stringent threshold of enrichment compared to the unselected pool in our assay (1817 cleaved and 3605 uncleaved sequences) as well as sequences identified by Shiryaev et al (Shiryaev et al., 2012) for a total of 7342 unique sequences. Training features consisted of structure-based features (energies of interaction) and sequence-based features (see Supplementary Methods, Figure 3.8A). We initially cross-validated the SVM on the training set using an 80:20 split with 100 iterations, which yielded an average AUROC of 0.96 (Figure 3.8B) indicating high recapitulation of training data (a perfect performance would lead to an AUROC of 1). We then used the SVM to predict cleaved and uncleaved labels for the remaining 3,192,658 sequences. These predictions have a precision of 0.96 at a recall

level of 0.91 for an overall accuracy of 0.96 (Figure 3.8B) for the experimentally-derived assignments that were left out of the training set (~10000 sequences). We experimentally validated cleavage predictions for six substrates as clonal populations using the yeast assay and find good agreement with the SVM-based predictions (Fig. 3.9A). We visualized a sub-graph of predicted cleaved sequences, present at a distance > 2 from the hyper-plane constructed by the SVM (Figure 3.8C). The experimentally identified cleaved sequences are recapitulated well, and distributed evenly across the predicted cleaved population.

4.3.4. Structural and energetic bases for observed specificity patterns

Having obtained and validated predictions of cleavability by combining experimental and computational data, we turned to structural models of protease-substrate complexes to obtain insight into the underlying structural basis of observed specificity patterns. For example, a comparative analysis of the partially cleaved substrate 'TATTA' and canonical substrate 'EDVVC' reveals that the former, composed of small residues does not completely occupy the substrate cavity volume (Figure 3.9B, C) whereas 'EDVVC' occupies the entire cavity. The lack of voids at the interface and several hydrogen bonds formed by the canonical lead to better binding (Binding interaction energy = -80.2 Rosetta energy units (Reu), as opposed to -77.5 Reu for TATTA), resulting in better cleavage for this substrate. Similarly, models of the uncleaved sequence FWPPM (Figure 3.9D) reveals that the side chains are found to have steric clashes with the protease side chains. Apart from sidechain-based interaction patterns, models also capture backbone conformational changes that affect the orientation of the substrate in the active site. For example, in the model corresponding to the sequence RPGPG (uncleaved), the proline

present at P3 in RPGPG (Fig. 3.9E) bends the peptide chain away from the protease, resulting in breaking of the crucial backbone hydrogen bond patterns that are characteristic of protease-substrate interactions (Tyndall et al., 2005).

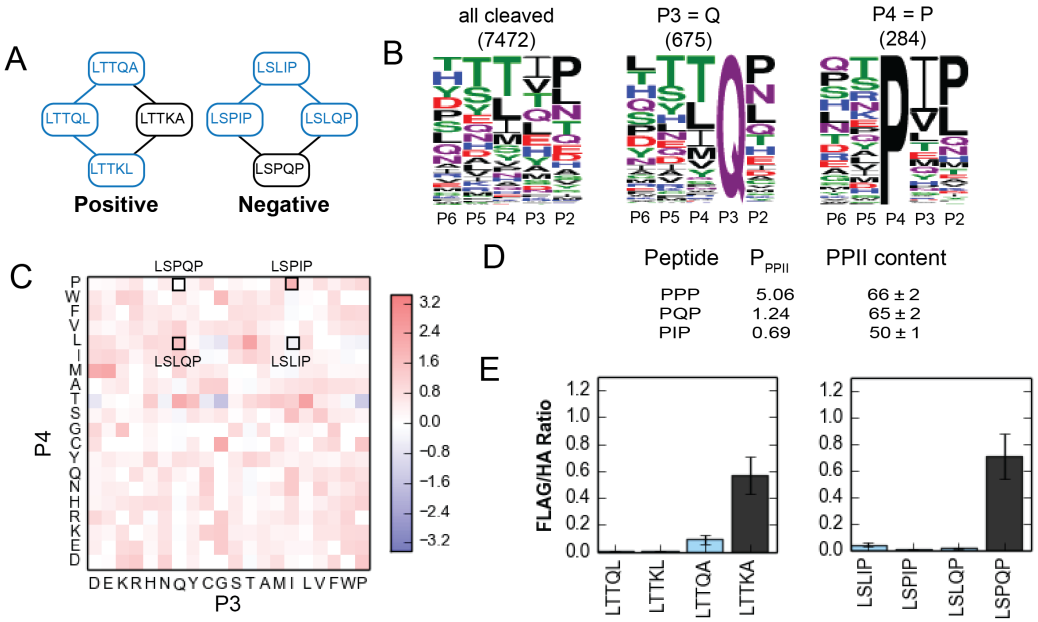


Figure 3.10. Structural basis underlying epistasis found on the interaction landscape.
(A) Examples of positive and negative epistasis. Cleaved sequences are highlighted in blue, partially cleaved in red. (B) Specificity profiles for entire cleaved set (left), sequences with glutamine at P3 (middle), and proline at P4 (right). (C) Heatmap of correlations between positions 3 and 4, as measured by mutual information. (D) Polyproline II structure propensity of peptides (see text). (E) Experimental validation of the sequences in both positive and negative epistatic pathways, performed using yeast surface display. Blue bars indicate sequences that are expected to be cleaved and black bars indicate sequences that are expected to be partially cleaved.

Structural analysis also allows rationalization of non-additive (epistatic) patterns between amino acid substitutions. We detected the presence of both positive and negative epistasis in our experimental data, and further investigated two cases (Figure 3.10A). We examined a predicted negative epistasis pathway (Figure 3.10B), where single-mutant P at position P4 and single-mutant Q at position P3 both result in a cleaved substrate but the

double-mutant PQ at position P3-P4 is uncleaved. We measured the mutual information (Figure 3.10C; Methods) between positions P3 and P4 in the experimentally derived cleaved sequence pool and found that both L at P3 and Q at P4 (corresponding to sequence LSLQP) and P at P3 and I at P4 (corresponding to sequence LSPIP) are correlated, indicating that these two amino acid preferences are found in the experimentally-derived cleaved population at a higher incidence than expected by their individual incidence. However, the correlation for P at P4 and Q at P3 (corresponding to sequence LSPQP) is low, suggesting that the PQ pattern is depleted in the cleaved sequence population. Structurally, the sequence LSPQP (Figure 3.10D) may have an increased PPII (polyproline-II) helix propensity (Kelly et al., 2001; Vila et al., 2004), causing the substrate to twist out of a catalysis-competent binding conformation in our models. The PPII helix propensity for the sequence LSPIP is lower thus resulting in retention of the extended substrate binding conformation that is favorable for catalysis (Tyndall et al., 2005). Thus, analysis of models of individual substrates provides atom-resolution insights into how the underlying biophysics of molecular recognition by the protease shapes the observed specificity landscapes, including non-additive effects.

Having validated (Figure 3.10E) these examples of double-mutant epistatic networks, we enumerated the double mutant epistatic networks present in the experimental data, and found that the majority of these epistatic networks (60.7%) involved cleaved and partially cleaved sequences only. The preponderance of epistatic networks at the cleaved/partially cleaved boundary indicates that the boundary between cleaved and partially cleaved sequences is more rugged than the boundary between cleaved and uncleaved sequences,

further highlighting the role of partially cleaved sequences as a biophysical buffer in sequence space, leading to "gradient robustness" proposed by Tawfik and co-workers.

4.3.5. Mutational robustness and possible evolutionary trajectories in the experimentally-determined and computationally reconstructed landscape

Having computed the entire P6-P2 specificity landscape, we next examined the connectivity patterns between cleaved and uncleaved sequences in this reconstructed landscape. As with the experimentally determined landscape, the reconstructed landscape also shows clear evidence of clustering between cleaved and uncleaved nodes (Figure 3.6E-I), indicating that mutational robustness extends to regions of sequence space not covered in our library, and is an essential feature of this protease-substrate interface. As our SVM-based approach is a binary classification scheme, partially cleaved sequences are classified in either cleaved or uncleaved pools. Attempts to build a 3-way classifier failed due both to the noise from the experiments as well as difficulty in estimating small energy differences in Rosetta simulations. Further improvements in each methodology may allow the prediction of partially cleaved sequences.

As the Hepatitis C virus is subject to a considerable amount of evolutionary drift, we investigated the impact of the pathways of drifting on the landscape on maintaining function. For the experimentally determined landscape, we calculated the number of mutations from each canonical sequence to the functional boundary and plotted the fraction of cleaved substrates that can be reached at each step (Figure 3.6E). The curves for both DEMEE and EDVVC reach a small initial plateau and then rise sharply,

indicating that both are surrounded by a cluster of cleaved sequences and then must bridge a largely non-functional region of the graph to reach the rest of the cleaved sequences, whereas the curves for both ALVTP and ECTTP rise steadily, indicating that the topology surrounding these sequences is less rugged.

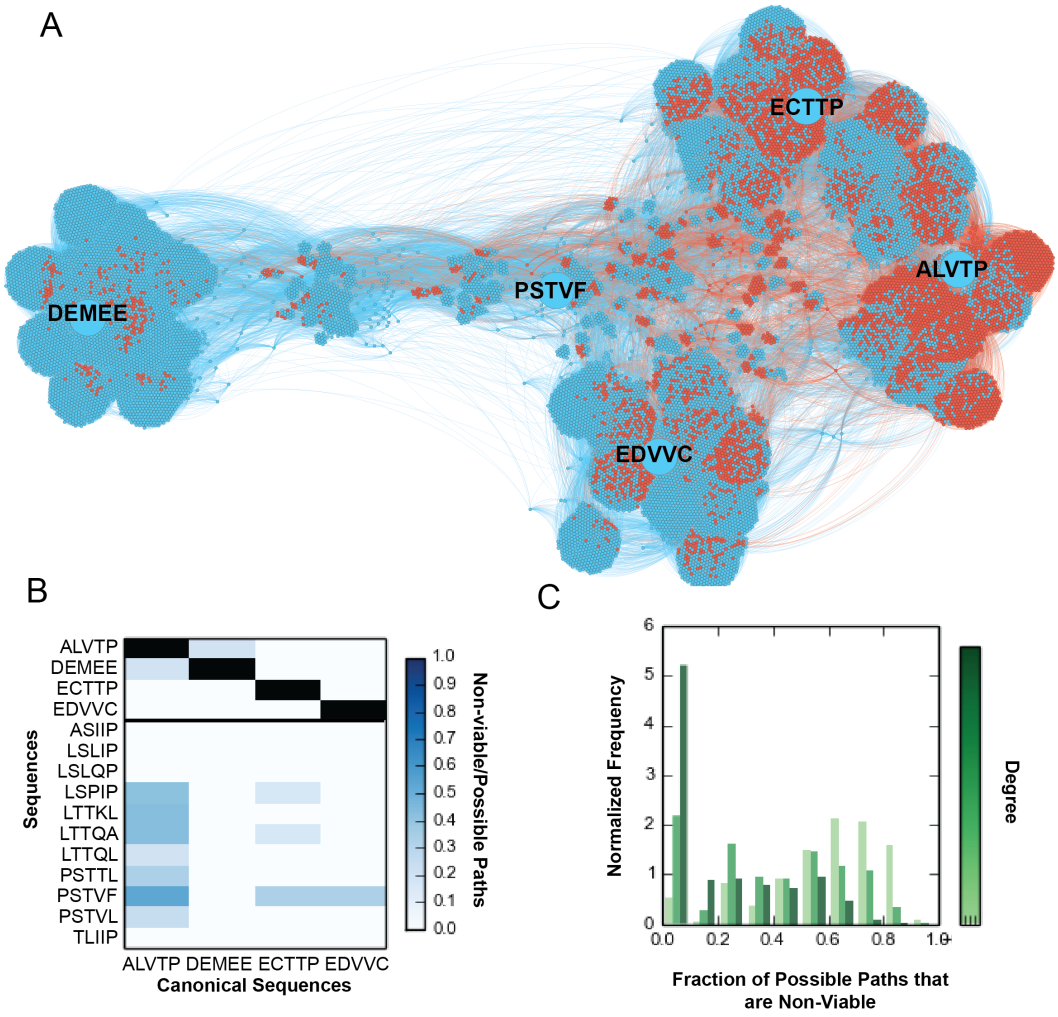


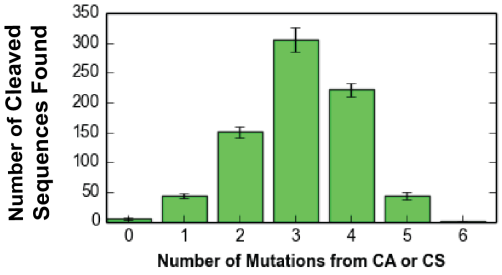
Figure 3.11. Force directed graph representation between five canonical and novel sequences and graph metrics for validation
(A) A Force-directed interaction graph between the five canonical sequences – DEMEE, ECTTP, EDVVC, ALVTP and the novel cleaved sequence PSTVF (depicted by large blue nodes). The graph depicts neighbors of all intermediate sequences between PSTVF and all canonical sequences. The cleaved sequences in the interaction pathways are denoted by blue nodes and the uncleaved are denoted by red (B) The fraction of uncleaved nodes present in the shortest paths from both canonical sequences and novel

sequences to all canonical sequences (C) Degree vs. fraction of the shortest paths uncleaved between all novel sequences and all canonical sequences.

Both the reconstructed and experimentally-derived landscapes feature several "novel" cleaved sequence patterns (defined as >3 substitutions away from a canonical recognition motif). To investigate if these novel sequences can be reached, as an example, we generated a sub graph of the sequence space connecting the canonical cleaved sequences (DEMEE, EDVVC, ECTTP, ALVTP) with each other as well as the novel cleaved sequences, e.g., PSTVF (Figure 3.11A). Analysis of all inter-node shortest paths on these networks shows that there exist many paths between canonical and novel sequences that do not include uncleaved nodes (viable paths) while some paths involve traversal of at least one predicted uncleaved node (unviable paths; Figure 3.11B). All canonical sequences are more highly connected to each other than to any of the novel sequence motifs, suggesting that the latter may be "kinetically" less accessible during evolutionary drifts. We calculated the fraction of non-viable paths between canonical sequences and compared it to the fraction of non-viable paths between canonical sequences and novel sequences. The latter shows a higher, albeit still small, fraction of non-viable paths (Figure 3.11B). We also find that those novel cleaved sequences that have a higher fraction of cleaved neighbors (higher degree) are more likely to have a higher fraction of viable trajectories to canonical nodes (Figure 3.11C). Thus, it appears that the higher single mutational robustness of a given novel sequence is correlated with its ability to be reachable from/to canonical sequences that are at least three amino acid substitutions away in sequence space. Further contributions from codon usage in the host context may modulate the reachability of different substrates by making some amino acid changes

even less likely. Our analysis above leaves out these contributions to selectively delineate the impact of amino acid-level effects.

A



B

Genotype	Average Cleaved	Average Uncleaved
1a	748	2246
1b	780	2221
2a	803	2200
3a	791	2193
4	792	2179
5a	780	2197
6a	779	2216

Figure 3.12. Evidence for negative selection of canonical substrate areas
(A) Bar plot depicting the number of DNA mutations required to mutate from current protein sequence to 'CS' which is the scissile bond sequence for the HCV NS3/4A protease (B) Table depicting the classification of all genotype derived 5-mers as classified by our SVM based predictor

4.3.6. Protease specificity landscape may contribute to purifying selection

Sequences of patient-derived genomes indicate that the HCV virus is under strong negative selection (Campo et al., 2008; Cuypers et al., 2015). Although the underlying mechanisms are not well understood, several factors have been invoked to explain the observation of a low dN/dS ratio (number of non-synonymous to synonymous substitutions in the genome) in the patient-derived populations including intrahost competition between quasispecies, and immune evasion (Skums et al., 2015). Given the centrality of the protease in viral maturation, we asked if maintenance of cleavability (and uncleavability) in different parts of the polyprotein also contributes to negative/purifying selection, and what, if any, are the limits imposed by the recognizability of different polyprotein regions by the protease on their variability.

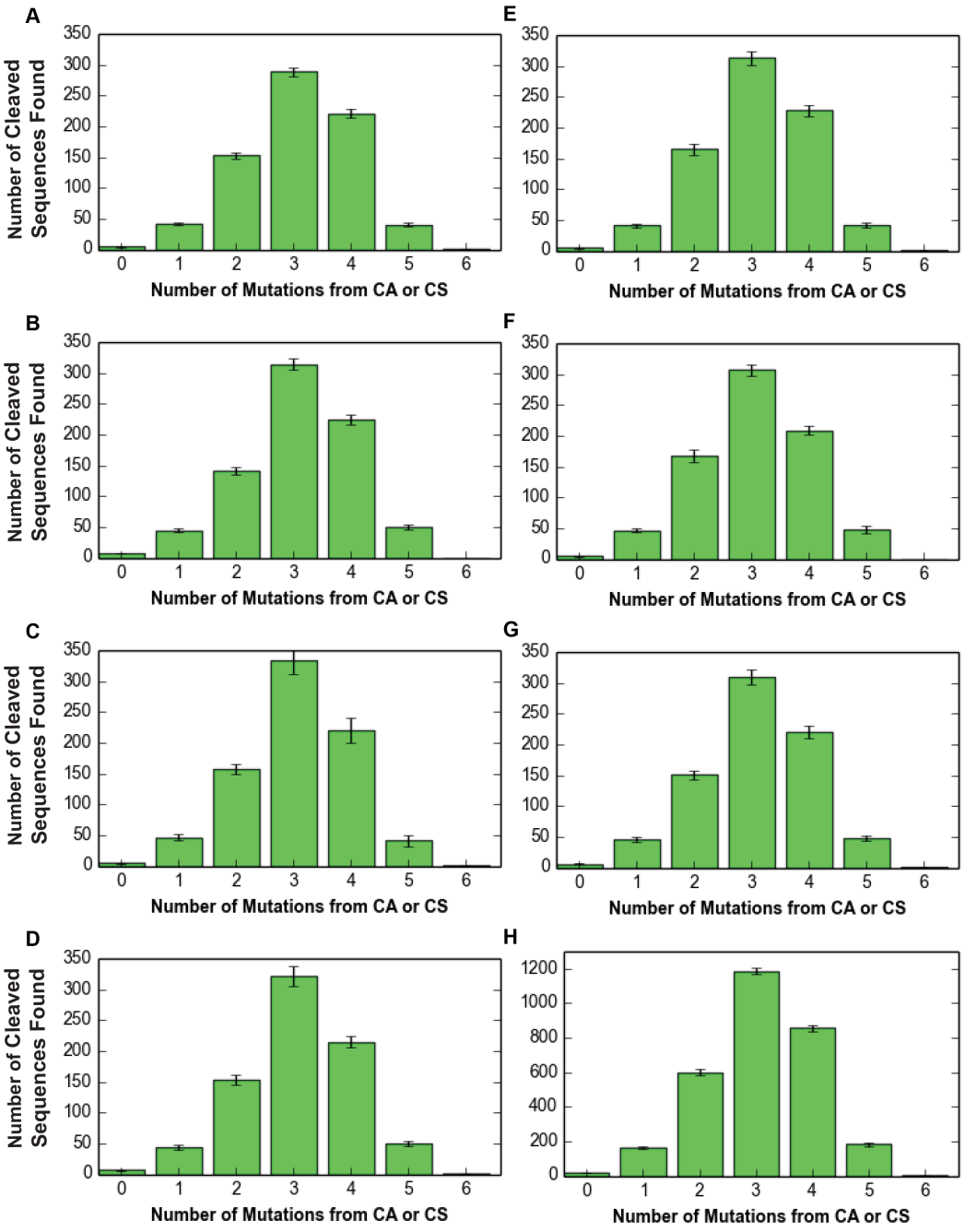


Figure 3.13. Plot depicting the number of DNA mutation required to mutate from current protein sequence to 'CS', which is the scissile bond sequence for the HCV NS3/4A protease for all genotypes

(A) strain 1a (B) strain 1b (C) strain 2 (D) strain 3 (E) strain 4 (F) strain 5 (G) strain 6 (H) control. Control is the distance from CA/CS for all 2-mers in all genotypes

As our reconstructed landscape provides information on all pentapeptide sequence combinations (followed by Cys-Ala), we asked if overlapping pentapeptides in the other parts of the polyprotein (apart from the known cleavage sites) are likely to be cleaved, especially if they acquired a Cys-Ala pattern in the two immediately downstream residues (thereby acquiring the necessary heptapeptide pattern that would be cleaved). If several regions of the polyprotein are poised to be cleaved upon acquisition of the Cys-Ala motif, an error catastrophe may ensue upon increasing the mutational load. We performed a genome-wide comparison of patient derived sequences with sequences predicted as cleaved by our SVM classifier. Each viral genome (Cuypers et al., 2015) was split into overlapping 5-mer peptide sequence fragments using a one-residue sliding window method. These 5-mers were compared to the pentapeptide sequences predicted by our approach as cleaved. If the patient-derived pentapeptide sequence was found in the cleaved pool, we calculated the minimum nucleotide mutational distance of the successive two residues from the DNA sequences that code for 'CA' and 'CS' which are known to be the canonical P1-P1' sites favoring cleavage by the HCV NS3/4A protease (Figure 3.1A). The results (Figure 3.12A,B, Figure 3.13A-H) indicate that the majority (~70%) of patient-derived translated pentapeptides are found the uncleaved pool. Of the remaining (~30%) 5-mer sequences that are identified as potentially cleavable (if they acquire a CA or CS as the following two amino acids), 74.1% pentapeptides from all genotypes of the virus require more than three nucleotide changes to acquire a 'CA' or 'CS' at the P1-P1' sites (Figure 3.13A-H). The avoidance of acquisition of a cleavable sequence in other regions of the protein, made feasible by codon usage, may thus, contribute to the previously described negative selection pressure on the HCV

genome(Campo et al., 2008), and may be reflected in the measured low dN/dS rates in the non-structural regions of the protein(Cuypers et al., 2015). Additional avoidance of non-productive cleavage may also result at the structural level from altered dynamics (Fuchs et al., 2014) and/or the post-translational structural context of the potentially cleavable regions – these may be buried (inaccessible to the protease) or adopt secondary structures that are incompatible with the extended conformation required to fit in the protease active site(Barkan et al., 2010; Julien et al., 2016).

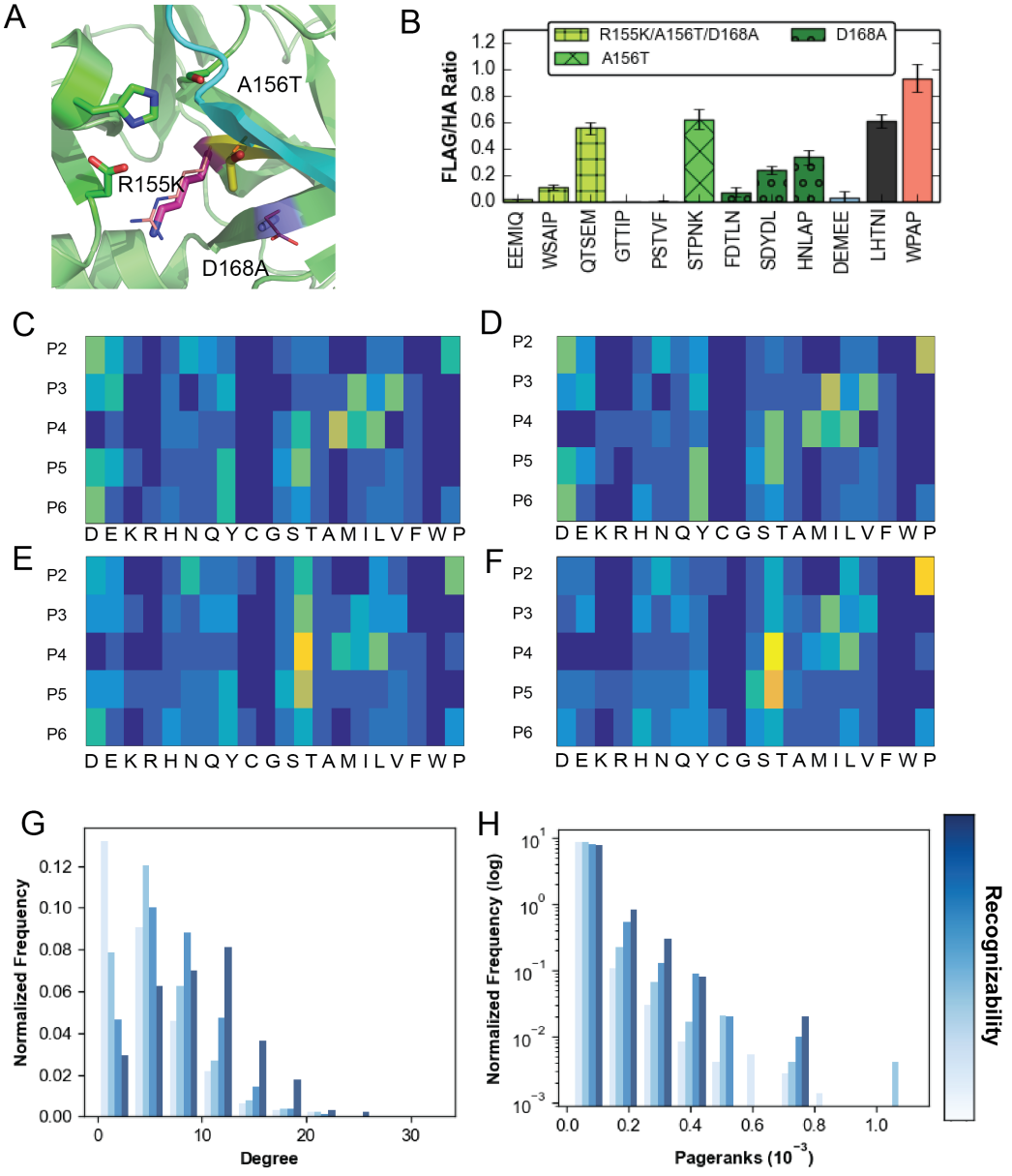


Figure 3.14. Validation, graph metrics and specificity profile for Drug resistant mutant proteases

(A) Drug-resistant variant structures. Mutations are outlined in sticks and WT residues in lines. Active site residues are represented as green sticks (B) Validation assay performed using yeast surface display for each of the mutants (C-F) Mutant specificity logos for the triple mutant, D168A, A156T and wild type showing that the mutants have very similar specificity profiles with slight variation as compared to the WT (G-H) Substrate sequences that are recognized by a greater number of protease variants have higher degrees (G) and pageranks (H)

4.3.7. Specificity landscapes of Drug Resistant Protease variants

As the NS3/4A protease plays a key role in the viral assembly and maturation process, it is a target for therapeutics that aim to neutralize viral activity. However, due to prevalence of quasispecies that are lurking at low levels in the population (Farci et al., 2000), several viral variants get exposed to the drug. Some of these develop resistance, and propagate to form Resistance Associated Variants (RAVs). To investigate how drug-resistant variants of the protease affect the mutational robustness, we explored the specificity landscape for three RAVs – A156T, D168A, R155K/A156T/D168A (Figure 3.14A). If the connectivity patterns of the sequences recognized by the RAVs are dramatically different (e.g., less clustered), it would indicate that their evolutionary fitness might be more limited under the heavy mutational load, as drifts on the substrate side would abolish the molecular interaction required for viral maturation. In this scenario, treatment with mutagens may be a desirable therapeutic strategy to induce error catastrophes. On the other hand, if similar mutationally robust connectivity is detected, the RAVs are likely to have a similar evolutionary potential as the wild type, and have an additional selective advantage in the population in the presence of the drug.

To obtain the landscapes of the protease variants (Figure 3.5B-E), we generated the library using a PCR amplification based strategy; isolated functional variants using FACS, deep sequenced the isolated populations and validated mutants (Figure 3.14B, Figure 3.5) identified from these populations using the yeast surface display assay. We find that the RAVs demonstrate a similar sequence profile to each other and to the wild type protease (Figure 3.14C-F). Upon comparing the graphical properties of the

specificity landscapes of the various protease variants, we observe that substrates that are experimentally detected in the cleaved pools of a greater number of protease variants are more reachable (higher pageranks) and more connected (higher degree) in each graph (Figure 3.14G,H). As our goal was to compare gross features of the specificity landscapes for the wild type and variant proteases, we did not perform detailed structure-based calculations for RAVs. Nonetheless, these data indicate that more recognizable substrates appear to be more robust to changes in the protease, and indeed, mutational robustness is a key feature of this specificity landscape.

4.4. Discussion

For RNA viruses, such as HCV, which have a high mutation rate, it has been hypothesized that viral evolution occurs via "survival of the flattest": the most conserved viral form is not necessarily the most fit, but instead is the one most robust to mutation – thus mutational robustness may provide an evolutionary advantage (Lauring and Andino, 2010; Lauring et al., 2013; Wilke et al., 2001). Our data, based on combining, using a machine learning framework, information gleaned from library screening in yeast, deep sequencing, and structure-based modeling, provide atomic-resolution insight into how mutational robustness may be encoded in the molecular recognition landscapes involved in viral maturation, and indicate that cleavage specificity of the HCV NS3/4A protease is robust to patient-derived mutations in both the substrate regions as well as the protease. However, molecular interaction between the protease and substrate, which key for viral survival, is but one of the many evolutionary forces at play, especially in the "wild"(Boucher et al., 2016). Other factors such as the intrahost population size, stability

and structure of the viral RNA genome, and interactions between the host and viral machineries and other environment dependent factors are also important to consider while considering evolutionary demands and trajectories.

We used a yeast surface display-based assay that relies on the cleavage of the substrate region in the ER of yeast followed by cell sorting into gates and deep sequencing. We note that our assay is qualitative, and does not permit association of the detected signal from deep sequencing with quantitative cleavability of substrates. Indeed, while we have validated that assignments to the three different pools is accurate with at least ~20 individual sequences, the identified cleaved and partially cleaved substrates may represent a wide range of catalytic efficiencies. A limitation of our technique is that it flattens this diversity into two pools. On the other hand, the assay construct with the protease and substrate on the same chain is a good representation of the situation in the virus, where the substrates of the protease are part of the same polyprotein (although both cis and trans cleavages occur) leading to high effective concentrations of substrates ([S] >> K_M) in vivo. Under these saturating conditions in the virus and in our assay, we argue that selectivity and catalytic efficiency are both determined to a great extent by the goodness of fit of various substrates in the protease active site (i.e. by the relative binding between the different substrates). Similarly, our machine learning approach to combine experimental and computational data also is not without errors, showing a false-positive rate of ~5-10% on the experimental data. While we have validated several predictions on individual sequences (Figures 3.1, 3.9, 3.14), it is possible that some individual sequences may be mispredicted. However, the overall trends regarding the connectivity patterns

observed for the entire landscape should be robust to the misprediction noise. Further ongoing development of the computational and experimental methods that we utilized is expected to help increase the accuracy of the approach.

HCV infects ~3% of the world population and the limited number of available viral genome sequences show low sequence heterogeneity in the substrate regions for the HCV protease. Nevertheless, resistance mutations upon protease inhibitor drug treatment arise in a facile manner in the patient population, suggesting that genetic heterogeneity (quasispecies) indeed exists, possibly at levels too low for being captured in patient-derived sequencing. Spontaneous emergence of diverse HCV protease mutations (including drug-resistant mutations) was demonstrated recently by Liu and colleagues in continuous evolution studies of the protease (Dickinson et al., 2014), as well as by Sanjuan and colleagues in viral replicon assays coupled to ultradeep sequencing (Geller et al., 2016b). Our results show how genetic heterogeneity is entirely consistent with the robustness of a key protease-peptide interaction in the virus, and therefore, provide a biophysical baseline for understanding evolvability of HCV, and for evaluating inhibitor drug resistance risks. For example, our analysis suggests that viral evolution occurring at the substrate sites on the polyprotein could also contribute to drug resistance. Due to the flatness of the specificity landscape and high inter-connectedness of partially cleaved and fully cleaved clusters, novel sequences that are better substrates of drug-resistant variants may easily arise. Thus, considering both substrate and protease variation in evaluating and designing anti-viral therapies may be necessary. This mode of substrate coevolution-based drug resistance has been observed in HIV-1(Dam et al., 2009). At the same time,

our analysis of the dominant HCV sequences obtained from patients suggests that the protease substrate interactions may also contribute to negative selection and help limit the acquisition of heterogeneity – the sequences of sites in the protease that are potentially cleavable upon acquisition of CA/CS at the P1-P1' junction (Figure 3.12) appear to be mutationally distant from doing so. Thus, the protease-substrate interaction landscape reveals that the balance between mutational robustness, negative selection and adaptive potential to environmental changes may be necessary to consider for understanding and therapeutic interventions.

In summary, our exploration of a viral molecular specificity landscape uncovers novel specificities for the HCV NS3/4A protease and data provides a biophysical basis for the mutational robustness observed for a key interaction required in HCV propagation. Given the widespread prevalence of HCV, insights obtained here may help in better understanding, and tackling the evolutionary trajectories of this ever-changing virus. The developed specificity landscape enumeration approach is general, and combining experimental deep sequencing and Rosetta-based structural modeling at a matching high throughput, followed by statistical machine learning, may be useful for elucidating a significantly larger space of sequence-function relationships for a variety of other systems.

4.5. References

Aakre, C.D., Herrou, J., Phung, T.N., Perchuk, B.S., Crosson, S., and Laub, M.T. (2015). Evolving new protein-protein interaction specificity through promiscuous intermediates. Cell 163, 594-606.

Amat, C.B. (2016). Gephi Cookbook. Rev Esp Doc Cient 39.

Andino, R., and Domingo, E. (2015). Viral quasispecies. Virology 479-480, 46-51.

Bandaru, P., Shah, N.H., Bhattacharyya, M., Barton, J.P., Kondo, Y., Cofsky, J.C., Gee, C.L., Chakraborty, A.K., Kortemme, T., Ranganathan, R., et al. (2017). Deconstruction of the Ras switching cycle through saturation mutagenesis. eLife 6.

Barkan, D.T., Hostetter, D.R., Mahrus, S., Pieper, U., Wells, J.A., Craik, C.S., and Sali, A. (2010). Prediction of protease substrates using sequence and structure features. Bioinformatics 26, 1714-1722.

Benatuil, L., Perez, J.M., Belk, J., and Hsieh, C.M. (2010). An improved yeast transformation method for the generation of very large human antibody libraries. Protein Eng Des Sel 23, 155-159.

Bershtein, S., Serohijos, A.W., and Shakhnovich, E.I. (2017). Bridging the physical scales in evolutionary biology: from protein sequence space to fitness of organisms and populations. Curr Opin Struct Biol 42, 31-40.

Blanquart, F., and Bataillon, T. (2016). Epistasis and the Structure of Fitness Landscapes: Are Experimental Fitness Landscapes Compatible with Fisher's Geometric Model? Genetics 203, 847-862.

Bloom, J.D., Wilke, C.O., Arnold, F.H., and Adami, C. (2004). Stability and the evolvability of function in a model protein. Biophys J 86, 2758-2764.

Bornberg-Bauer, E., and Chan, H.S. (1999). Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. Proc Natl Acad Sci U S A 96, 10689-10694.

Boucher, J.I., Bolon, D.N., and Tawfik, D.S. (2016). Quantifying and understanding the fitness effects of protein mutations: Laboratory versus nature. Protein Sci 25, 1219-1226.

Breen, M.S., Kemena, C., Vlasov, P.K., Notredame, C., and Kondrashov, F.A. (2012). Epistasis as the primary factor in molecular evolution. Nature 490, 535-538.

Bridgham, J.T., Carroll, S.M., and Thornton, J.W. (2006). Evolution of hormone-receptor complexity by molecular exploitation. Science 312, 97-101.

Brin, S., and Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. Computer Networks and Isdn Systems 30, 107-117.

Campo, D.S., Dimitrova, Z., Mitchell, R.J., Lara, J., and Khudyakov, Y. (2008). Coordinated evolution of the hepatitis C virus. Proc Natl Acad Sci U S A 105, 9685-9690.

Cristina, J., del Pilar Moreno, M., and Moratorio, G. (2007). Hepatitis C virus genetic variability in patients undergoing antiviral therapy. Virus Res 127, 185-194.

Cuypers, L., Li, G., Libin, P., Piampongsant, S., Vandamme, A.M., and Theys, K. (2015). Genetic Diversity and Selective Pressure in Hepatitis C Virus Genotypes 1-6: Significance for Direct-Acting Antiviral Treatment and Drug Resistance. Viruses 7, 5018-5039.

Dam, E., Quercia, R., Glass, B., Descamps, D., Launay, O., Duval, X., Krausslich, H.G., Hance, A.J., Clavel, F., and Group, A.S. (2009). Gag mutations strongly contribute to HIV-1 resistance to protease inhibitors in highly drug-experienced patients besides compensating for fitness loss. PLoS Pathog 5, e1000345.

de Visser, J.A., and Krug, J. (2014). Empirical fitness landscapes and the predictability of evolution. Nat Rev Genet 15, 480-490.

DePristo, M.A., Weinreich, D.M., and Hartl, D.L. (2005). Missense meanderings in sequence space: a biophysical view of protein evolution. Nat Rev Genet 6, 678-687.

Dickinson, B.C., Packer, M.S., Badran, A.H., and Liu, D.R. (2014). A system for the continuous directed evolution of proteases rapidly reveals drug-resistance mutations. Nat Commun 5, 5352.

Ding, F., and Dokholyan, N.V. (2006). Emergence of protein fold families through rational design. PLoS Comput Biol 2, e85.

Domingo, E., and Holland, J.J. (1997). RNA virus mutations and fitness for survival. Annu Rev Microbiol 51, 151-178.

Draghi, J.A., Parsons, T.L., Wagner, G.P., and Plotkin, J.B. (2010). Mutational robustness can facilitate adaptation. Nature 463, 353-355.

Drummond, D.A., and Wilke, C.O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134, 341-352.

Echave, J., and Wilke, C.O. (2017). Biophysical Models of Protein Evolution: Understanding the Patterns of Evolutionary Sequence Divergence. Annu Rev Biophys 46, 85-103.

Eigen, M. (1993). Viral quasispecies. Sci Am 269, 42-49.

Eigen, M. (2002). Error catastrophe and antiviral strategy. Proc Natl Acad Sci U S A 99, 13374-13376.

Elde, N.C., Child, S.J., Eickbush, M.T., Kitzman, J.O., Rogers, K.S., Shendure, J., Geballe, A.P., and Malik, H.S. (2012). Poxviruses deploy genomic accordions to adapt rapidly against host antiviral defenses. Cell 150, 831-841.

Elena, S.F., Carrasco, P., Daros, J.A., and Sanjuan, R. (2006). Mechanisms of genetic robustness in RNA viruses. EMBO Rep 7, 168-173.

Farci, P., Shimoda, A., Coiana, A., Diaz, G., Peddis, G., Melpolder, J.C., Strazzera, A., Chien, D.Y., Munoz, S.J., Balestrieri, A., et al. (2000). The outcome of acute hepatitis C predicted by the evolution of the viral quasispecies. Science 288, 339-344.

Firnberg, E., Labonte, J.W., Gray, J.J., and Ostermeier, M. (2014). A comprehensive, high-resolution map of a gene's fitness landscape. Mol Biol Evol 31, 1581-1592.

Fowler, D.M., Araya, C.L., Fleishman, S.J., Kellogg, E.H., Stephany, J.J., Baker, D., and Fields, S. (2010). High-resolution mapping of protein sequence-function relationships. Nat Methods 7, 741-746.

Fowler, D.M., and Fields, S. (2014). Deep mutational scanning: a new style of protein science. Nat Methods 11, 801-807.

Fuchs, J.E., von Grafenstein, S., Huber, R.G., Wallnoefer, H.G., and Liedl, K.R. (2014). Specificity of a protein-protein interface: local dynamics direct substrate recognition of effector caspases. Proteins 82, 546-555.

Geller, R., Estada, U., Peris, J.B., Andreu, I., Bou, J.V., Garijo, R., Cuevas, J.M., Sabariegos, R., Mas, A., and Sanjuan, R. (2016a). Highly heterogeneous mutation rates in the hepatitis C virus genome. Nat Microbiol 1.

Geller, R., Estada, U., Peris, J.B., Andreu, I., Bou, J.V., Garijo, R., Cuevas, J.M., Sabariegos, R., Mas, A., and Sanjuan, R. (2016b). Highly heterogeneous mutation rates in the hepatitis C virus genome. Nat Microbiol 1, 16045.

Goldberg, D.E., Siliciano, R.F., and Jacobs, W.R., Jr. (2012). Outwitting evolution: fighting drug-resistant TB, malaria, and HIV. Cell 148, 1271-1283.

Harms, M.J., and Thornton, J.W. (2013). Evolutionary biochemistry: revealing the historical and physical causes of protein properties. Nat Rev Genet 14, 559-571.

Hartl, D.L. (2014). What can we learn from fitness landscapes? Curr Opin Microbiol 21, 51-57.

Hietpas, R.T., Jensen, J.D., and Bolon, D.N. (2011). Experimental illumination of a fitness landscape. Proc Natl Acad Sci U S A 108, 7896-7901.

Holland, J., Spindler, K., Horodyski, F., Grabau, E., Nichol, S., and VandePol, S. (1982). Rapid evolution of RNA genomes. Science 215, 1577-1585.

Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. (2014). ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. Plos One 9.

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

Julien, O., Zhuang, M., Wiita, A.P., O'Donoghue, A.J., Knudsen, G.M., Craik, C.S., and Wells, J.A. (2016). Quantitative MS-based enzymology of caspases reveals distinct protein substrate specificities, hierarchies, and cellular roles. Proc Natl Acad Sci U S A 113, E2001-2010.

Kelly, M.A., Chellgren, B.W., Rucker, A.L., Troutman, J.M., Fried, M.G., Miller, A.F., and Creamer, T.P. (2001). Host-guest study of left-handed polyproline II helix formation. Biochemistry 40, 14376-14383.

Kim, I., Miller, C.R., Young, D.L., and Fields, S. (2013). High-throughput analysis of in vivo protein stability. Mol Cell Proteomics 12, 3370-3378.

Klesmith, J.R., Bacik, J.P., Michalczyk, R., and Whitehead, T.A. (2015). Comprehensive Sequence-Flux Mapping of a Levoglucosan Utilization Pathway in E. coli. ACS Synth Biol 4, 1235-1243.

Klesmith, J.R., Bacik, J.P., Wrenbeck, E.E., Michalczyk, R., and Whitehead, T.A. (2017). Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning. Proc Natl Acad Sci U S A 114, 2265-2270.

Kondrashov, D.A., and Kondrashov, F.A. (2015). Topological features of rugged fitness landscapes in sequence space. Trends Genet 31, 24-33.

Kowalsky, C.A., Klesmith, J.R., Stapleton, J.A., Kelly, V., Reichkitzer, N., and Whitehead, T.A. (2015). High-resolution sequence-function mapping of full-length proteins. PLoS One 10, e0118193.

Lauring, A.S., and Andino, R. (2010). Quasispecies theory and the behavior of RNA viruses. PLoS Pathog 6, e1001005.

Lauring, A.S., Frydman, J., and Andino, R. (2013). The role of mutational robustness in RNA virus evolution. Nat Rev Microbiol 11, 327-336.

Li, Q., Yi, L., Hoi, K.H., Marek, P., Georgiou, G., and Iverson, B.L. (2017). Profiling Protease Specificity: Combining Yeast ER Sequestration Screening (YESS) with Next Generation Sequencing. ACS Chem Biol 12, 510-518.

Manhart, M., and Morozov, A.V. (2015). Protein folding and binding can emerge as evolutionary spandrels through structural coupling. Proc Natl Acad Sci U S A 112, 1797-1802.

Masel, J., and Siegal, M.L. (2009). Robustness: mechanisms and consequences. Trends Genet 25, 395-403.

McLaughlin, R.N., Jr., Poelwijk, F.J., Raman, A., Gosal, W.S., and Ranganathan, R. (2012). The spatial architecture of protein function and adaptation. Nature 491, 138-142.

Meylan, E., Curran, J., Hofmann, K., Moradpour, D., Binder, M., Bartenschlager, R., and Tschopp, J. (2005). Cardif is an adaptor protein in the RIG-I antiviral pathway and is targeted by hepatitis C virus. Nature 437, 1167-1172.

Pethe, M.A., Rubenstein, A.B., and Khare, S.D. (2017). Large-Scale Structure-Based Prediction and Identification of Novel Protease Substrates Using Computational Protein Design. J Mol Biol 429, 220-236.

Podgornaia, A.I., and Laub, M.T. (2015). Protein evolution. Pervasive degeneracy and epistasis in a protein-protein interface. Science 347, 673-677.

Powdrill, M.H., Tchesnokov, E.P., Kozak, R.A., Russell, R.S., Martin, R., Svarovskaia, E.S., Mo, H., Kouyos, R.D., and Gotte, M. (2011). Contribution of a mutational bias in hepatitis C virus replication to the genetic barrier in the development of drug resistance. Proc Natl Acad Sci U S A 108, 20509-20513.

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

Rodrigues, J.V., Bershtein, S., Li, A., Lozovsky, E.R., Hartl, D.L., and Shakhnovich, E.I. (2016). Biophysical principles predict fitness landscapes of drug resistance. Proc Natl Acad Sci U S A 113, E1470-1478.

Romano, K.P., Ali, A., Aydin, C., Soumana, D., Ozen, A., Deveau, L.M., Silver, C., Cao, H., Newton, A., Petropoulos, C.J., et al. (2012). The molecular basis of drug resistance against hepatitis C virus NS3/4A protease inhibitors. PLoS Pathog 8, e1002832.

Romero, P.A., and Arnold, F.H. (2009). Exploring protein fitness landscapes by directed evolution. Nat Rev Mol Cell Biol 10, 866-876.

Rubenstein, A.B., Pethe, M.A., and Khare, S.D. (2017). MFPred: Rapid and accurate prediction of protein-peptide recognition multispecificity using self-consistent mean field theory. PLoS Comput Biol 13, e1005614.

Sailer, Z.R., and Harms, M.J. (2017a). Detecting High-Order Epistasis in Nonlinear Genotype-Phenotype Maps. Genetics 205, 1079-1088.

Sailer, Z.R., and Harms, M.J. (2017b). High-order epistasis shapes evolutionary trajectories. PLoS Comput Biol 13, e1005541.

Sarkisyan, K.S., Bolotin, D.A., Meer, M.V., Usmanova, D.R., Mishin, A.S., Sharonov, G.V., Ivankov, D.N., Bozhanova, N.G., Baranov, M.S., Soylemez, O., et al. (2016). Local fitness landscape of the green fluorescent protein. Nature 533, 397-401.

Schechter, I., and Berger, A. (1967). On the size of the active site in proteases. I. Papain. Biochem Biophys Res Commun 27, 157-162.

Scheel, T.K., and Rice, C.M. (2013). Understanding the hepatitis C virus life cycle paves the way for highly effective therapies. Nat Med 19, 837-849.

Serohijos, A.W., and Shakhnovich, E.I. (2014). Merging molecular mechanism and evolution: theory and computation at the interface of biophysics and evolutionary population genetics. Curr Opin Struct Biol 26, 84-91.

Shiryaev, S.A., Thomsen, E.R., Cieplak, P., Chudin, E., Cheltsov, A.V., Chee, M.S., Kozlov, I.A., and Strongin, A.Y. (2012). New details of HCV NS3/4A proteinase functionality revealed by a high-throughput cleavage assay. PLoS One 7, e35759.

Sikosek, T., and Chan, H.S. (2014). Biophysics of protein evolution and evolutionary protein biophysics. J R Soc Interface 11, 20140419.

Skums, P., Bunimovich, L., and Khudyakov, Y. (2015). Antigenic cooperation among intrahost HCV variants organized into a complex network of cross-immunoreactivity. Proc Natl Acad Sci U S A 112, 6653-6658.

Smith, J.M. (1970). Natural selection and the concept of a protein space. Nature 225, 563-564.

Thyagarajan, B., and Bloom, J.D. (2014). The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. eLife 3.

Tinberg, C.E., Khare, S.D., Dou, J., Doyle, L., Nelson, J.W., Schena, A., Jankowski, W., Kalodimos, C.G., Johnsson, K., Stoddard, B.L., et al. (2013). Computational design of ligand-binding proteins with high affinity and selectivity. Nature 501, 212-216.

Tokuriki, N., Oldfield, C.J., Uversky, V.N., Berezovsky, I.N., and Tawfik, D.S. (2009). Do viral proteins possess unique biophysical features? Trends Biochem Sci 34, 53-59.

Tyndall, J.D., Nall, T., and Fairlie, D.P. (2005). Proteases universally recognize beta strands in their active sites. Chem Rev 105, 973-999.

van Nimwegen, E., Crutchfield, J.P., and Huynen, M. (1999). Neutral evolution of mutational robustness. Proc Natl Acad Sci U S A 96, 9716-9720.

Vila, J.A., Baldoni, H.A., Ripoll, D.R., Ghosh, A., and Scheraga, H.A. (2004). Polyproline II helix conformation in a proline-rich environment: a theoretical study. Biophys J 86, 731-742.

Weinreich, D.M., Delaney, N.F., Depristo, M.A., and Hartl, D.L. (2006). Darwinian evolution can follow only very few mutational paths to fitter proteins. Science 312, 111-114.

Weinreich, D.M., Lan, Y., Wylie, C.S., and Heckendorn, R.B. (2013). Should evolutionary geneticists worry about higher-order epistasis? Curr Opin Genet Dev 23, 700-707.

Whitehead, T.A., Chevalier, A., Song, Y., Dreyfus, C., Fleishman, S.J., De Mattos, C., Myers, C.A., Kamisetty, H., Blair, P., Wilson, I.A., et al. (2012). Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. Nat Biotechnol 30, 543-548.

Wilke, C.O., and Adami, C. (2003). Evolution of mutational robustness. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 522, 3-11.

Wilke, C.O., Wang, J.L., Ofria, C., Lenski, R.E., and Adami, C. (2001). Evolution of digital organisms at high mutation rates leads to survival of the flattest. Nature 412, 331-333.

Wrenbeck, E.E., Azouz, L.R., and Whitehead, T.A. (2017). Single-mutation fitness landscapes for an enzyme on multiple substrates reveal specificity is globally encoded. Nat Commun 8, 15695.

Wright, S. (1931). Evolution in Mendelian Populations. Genetics 16, 97-159.

Wu, N.C., Dai, L., Olson, C.A., Lloyd-Smith, J.O., and Sun, R. (2016). Adaptation in protein fitness landscapes is facilitated by indirect paths. eLife 5.

Yang, J.R., Liao, B.Y., Zhuang, S.M., and Zhang, J. (2012). Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. Proc Natl Acad Sci U S A 109, E831-840.

Yi, L., Gebhard, M.C., Li, Q., Taft, J.M., Georgiou, G., and Iverson, B.L. (2013). Engineering of TEV protease variants by yeast ER sequestration screening (YESS) of combinatorial libraries. Proc Natl Acad Sci U S A 110, 7229-7234.

4.6. Supplementary Methods

Two step screening approach to avoid stop codons:

The LY104 vector was a gift from Y. Li, B. Iverson, and G. Georgiou (University of Texas at Austin). The library was constructed using a two-step screening approach to avoid enrichment of false positives. The first step was an expression screen, which was done by combining the library with a protease inactive vector (LY104 S139A knockout). The recombination was performed by homologous recombination technique in yeast EBY100 cells. We modified an electroporation-based method as described in (Benatuil et al., 2010). The transformed library was allowed to grow for 48 hours at 30 C, up to an OD_{600} of 2.0. Dilutions of 1/10, 1/100 and $1/1000^{th}$ were plated from the initial culture to calculate the transformation efficiency and library size. The double positive cell population was isolated and enriched using a Fluorescence Assisted Cell Sorting technique. The expressible library was then recombined with a vector containing the active protease, using the aforementioned homologous recombination technique. This library of functional variants was allowed to grow up to 48 hours at 30 C and then sorted into three sequence pools – cleaved, partially cleaved and uncleaved. The gates for the FACS were defined using clonal substrates that displayed varying levels of cleavage activities. The three sequence pools were enriched via three rounds of successive selection (using FACS) and growth. The DNA from the three sequence pools was extracted using the Omega E.Z.N.A yeast plasmid kit. Technical duplicates were sequenced to get an estimate of error correction necessary for post processing this data.

Library Generation methodology:

The library was constructed using a PCR amplification based technique using NNK mixed base oligonucleotides (Integrated DNA Technologies). The LY104 vector was linearized using DNA oligonucleotides (IDT). The NNK library insert (~576 bp) and linearized vector (~6000 bp) were combined using Homologous Recombination using electro-competent EBY100 yeast cells. The transformed EBY100 cells were rescued using a YPD medium and allowed to grow in a 250 mL Selective Complete Growth Medium (-UW). The media was supplemented with 250 µL of Ampicillin and Kanamycin to avoid bacterial contamination.

Library Testing and Enrichment:

The transformed library was allowed to grow for ~48 hrs (up to OD_{600} 2.0) and then induced and tested using Flow cytometry. 1.5 x 10^7cells(OD_{600} ~0.5) were pelleted and resuspended in 2 mL induction media (20g/L galactose, 2 g/L glucose) supplemented with 2 µL each of 1000x antibiotics (carbenicillin, kanamycin). The induction cultures were grown overnight at 30 C (225 rpm) to an OD_{600} of 1-1.5. All spins in the protocol were done at 3000 r.c.f for 5 min. The induced cultures were pelleted and washed with 500 µL PBS followed by 500 µL PBS+ 0.5% BSA. 1 µL of each antibody stain (anti-FLAG PE from Prozyme, PJ315 and anti-HA FITC from Genscript, A01621) was incubated with 10^7 cells for 30 min at 4 C. The samples were resuspended by vortexing and incubated at RT for an additional 30 min. The cells were washed with 100µL PBS with 0.5% BSA, pelleted and then resuspended in 500 µL PBS. Samples were diluted to

achieve a final concentration of 10^6 cells/mL and then FITC (anti-HA) and PE (anti-FLAG) intensities were detected using a Flow Cytometer (Beckman Coulter Gallios).

The tested cells were then enriched using a MoFlo XDP Cell Sorter (final cell density 10^7). Up to 10^6 cells were collected and grown in the Selective Complete Growth Media for 48 Hours. Two rounds of sorting and enrichment were carried out to select for clones that were expressed. The selected cells were grown for 48 hours. The DNA from the selected cell population was extracted using E.Z.N.A Zymoprep Kit (Omega).

Cell Sorting into Cleaved, Uncleaved, Partially Cleaved Populations:

The expressible fraction of the library was combined with the active LY104 vector using a second round of Homologous recombination following the same protocol as mentioned above. Using the MoFlo XDP Cell Sorter we defined Cleaved, Uncleaved and Partially cleaved gates for further selection of this population. These gates were defined based on previously experimentally tested sequences.

These cells from the selected population were put through three rounds of enrichment and sorting. In the first round of sorting, cells were collected into two gates – Cleaved and Uncleaved. The Uncleaved sample was further enriched in the second sorting round whereas the Cleaved population was separated into Cleaved and Partially cleaved gates. Cells were collected for each sorting round until a cell count of 10^6 was reached. At the culmination of each sorting round, DNA was collected from each population by using a Zymoprep Kit (Omega).

Preparation for Illumina Sequencing Run:

The DNA samples collected from each of the populations were prepared by 25 cycle amplification (Kowalsky et al. 2015) with inner primers (Supplementary Table 3). The samples were then run on a 1% Agarose gel to confirm the amplification of a single species. These were further amplified using 8 PCR cycles to include the DNA barcode used in the deep sequencing protocol and checked for quality using a Bioanalyzer 2100. The Deep sequencing was performed on a NextSeq 500 (Illumina) giving a 75 bp paired end read.

I. Expression Protocols:

II. Protease expression:

Expression and purification protocol was a modification of previously published protocols (Wittekind et al. 2001; Gallinari et al. 1998; Romano et al. 2012). Transformed BL21 (DE3) E. coli cells were grown at 37°C and induced at an optical density of 0.6 by adding 1 mM IPTG. Cells were harvested after 5 hours of expression, pelleted, and frozen at −80°C for storage. Cell pellets were thawed, resuspended in 5 mL/g of resuspension buffer (50 mM phosphate buffer, 500 mM NaCl, 10% glycerol, 30 mM imidazole, 2 mM β-ME, pH 7.5) and lysed with a sonicator. The soluble fraction was retained, applied to a nickel column (Qiagen), washed with resuspension buffer, and eluted with resuspension buffer supplemented with 200 mM imidazole. The eluent was dialyzed overnight (MWCO 10 kD) into a protease storage buffer (20mM Tris.HCl,pH

8.0, glycerol 20%, 100 mM KCl, 1mM DTT, 0.2 mM EDTA) to remove the imidazole. The purified protein was then flash frozen and stored at -80 C.

Substrate (MBP-GST construct) expression: The transformed BL21(DE3) cells were grown at 37 C to an optical density of 0.6 and induced using 0.2 mM IPTG. Upon induction the cells were grown overnight at 18 C. the cells were harvested and the cell pellet was resuspended in a resuspension buffer (50 mM Tris.HCl, pH8.0, 500 mM NaCl, 30 mM immidazole). The resuspended cells were lysed via sonication and the soluble fraction was applied to a Nickel column (Qiagen). The column was washed using the resuspension buffer and then the protein eluted using an Elution buffer- 50 mM Tris.HCl, pH8.0, 150 mM NaCl, 300 mM imidazole. The protein was dialyzed overnight to remove the imidazole and frozen until use.

Gel based validation assay: The frozen aliquots of substrate solutions were thawed and dialyzed overnight into the reaction buffer (50mM HEPES (pH 7.5), 150 mM NaCl, 0.1% Triton X-100, 15% glycerol, 10mM DTT). 28.5 nM substrate was incubated overnight with 500nM, 700nM, 1µM, 2µM, 3µM and 4 µM protease. The resultant reactions were run on a SDS PAGE gel to check for cleavage activity.

III. Sequence Processing

A. Sequence Alignment and Trimming

Data was received oriented in the correct orientation and filtered for quality of 20. Each sequence was searched for the presence and location of “TCTTTATAA”, a unique

string within the WT sequence, to align the sequences. If the index of "TCTTTATAA" in sequence *a* is less than the index of "TCTTTATAA" in the WT sequence, the beginning of sequence *a* is padded to match the beginning of the WT sequence. If the index of "TCTTTATAA" in sequence *a* is greater than the index of "TCTTTATAA" in the WT sequence, the beginning of sequence *a* is truncated to match the beginning of the WT sequence. If "TCTTTATAA" is not found in sequence *a*, it is discarded. If the padded or truncated sequence *a* is shorter than the index of the library region in the WT sequence, sequence *a* is discarded. If sequence *a* is longer than the index of the library region but shorter than the WT sequence, the end of sequence *a* is padded to match the WT sequence. Finally, we check that the padded or truncated sequence *a* matches the WT sequence entirely except for the library region. If it does not match the WT sequence, we discard sequence *a*.

B. Threshold Determination

After aligning and trimming sequences, we calculate a normalized count of each sequence so that the sum of the normalized counts in each population is equal. This is achieved by multiplying each sequence count in population *a* by a normalization factor that is equal to the number of sequences in the largest library divided by the number of sequences in library *a*. Then, to determine the minimum frequency of each sequence in the population above which we are confident of the validity of its representation in the library, we used several methods:

1) **Overlap between cleaved and uncleaved sequences**:

We expect little overlap between the populations of cleaved and uncleaved sequences. However, at low counts, there is some overlap between the two populations. For each threshold, we calculated the number of sequences that overlapped between cleaved and uncleaved sequences, and normalized by the count of unique translated cleaved DNA sequences at that threshold. We determined the amount of overlap as a percentage of the initial overlap between the populations at a threshold of 1, and then found the threshold that gave <= 10% of the initial overlap (see Figure 3.2). We repeated this analysis for all four variant populations. The threshold was less than or slightly greater than 11 for all variants.

2) **Duplicate population error:**

We sampled technical duplicates for the third round of enrichment for cleaved, uncleaved and partially cleaved sequence pools. As a post - processing step in the pipeline, we introduced duplicate population error correction, by plotting the difference in counts for common sequences of the technical duplicate samples and plotting against the counts in the first sample.

3) **SVM Convergence**:

In order to select for the threshold that gave us the most distinct populations, we generated cleaved and uncleaved sequence sets for thresholds 5,10,11,12,13,14, 15, 16, 25, 50, 75 and 100. Using an SVM based technique described previously (Chapter 2) we calculated the auROC for all cleaved and uncleaved sequence populations for the listed thresholds. This enabled us to identify which threshold increases the distinction between the two populations.

We decided upon a frequency threshold of 11 as one that satisfies all categories of threshold determination.

C. Enrich Software

We used a modified version of the Enrich software (Fowler et al. 2010) to find an enrichment ratio (ER) for each sequence. We only included sequences that had a normalized count (as defined above) of greater than or equal to eleven for both the unselected and selected populations. The enrichment ratio of sequence *v* in population *X* is defined using Equation 1. $F_{v,X}$ is the frequency of sequence *v* in population *X*.

$$ER_{v,X} = \log_2 \frac{F_{v,X}}{F_{v,input}} \quad (1)$$

D. Population Categorization

Sequences were sorted into one of three pools (cleaved, uncleaved and partially cleaved), based on the following criteria. Sequences that had a positive ER for more than one pool were discarded. Sequences that had a positive ER for either or both replicates for one pool only were assigned to that pool. Negative ERs were ignored.

We also sorted a second set with more stringent criteria, which was then used for training the SVM. For this set, if a sequence was found in more than one pool (even if it had a negative ER in the second pool), it was discarded. Additionally, only sequences with a positive ER > 2.0 were considered.

Computational

Graph Generation

Graph generation was done using Gephi 0.9.1 (Bastian et al. 2009). Nodes were assigned a fitness of 2.0 for cleaved nodes, 1.5 for partially cleaved nodes, and 1.0 for uncleaved nodes. Edge directionality was determined by distance from DEMEE, the starting sequence for library generation; in the case of edge *a* connecting nodes *b* and *c*, the node with a smaller hamming distance from DEMEE was chosen as the source for edge *a*. Edge weight was defined as the ratio of the starting sequence fitness to the ending sequence fitness. The graph layout was run in two steps, starting with a Fruchterman-Reingold layout to separate the nodes and then ending with the ForceAtlas2 layout to generate a force-directed graph. All statistics were run with Gephi default settings.

Random Graph

The edges in the wild-type HCV graph were randomized using the following process. The source of each edge was kept and a population (cleaved, partially cleaved, or uncleaved) was randomly chosen for the target of the edge. The target of the edge was then randomly chosen from among that population.

SVM Sequence Features

We used an encoding scheme that included twenty binary features per amino acid residue, where one of those features was a one and the rest were zeroes. The placement

of the one was dependent on the identity of the amino acid. With five amino acid residues per sequence, this resulted in 100 total sequence features.

Mutual Information

Correlation between residues at different positions was calculated using a mutual-information based metric (Equation 2), with modifications based on Buslje et al. (Equation 3) (Buslje et al. 2009) and Gouveia-Oliveira and Pedersen (Equation 4) (Gouveia-Oliveira & Pedersen 2007). We begin with MI between amino acid *a* at position *i* and amino acid *b* at position *j* defined as:

$$MI_{a_i b_j} = \log \frac{P(a_i b_j)}{P(a_i) \cdot P(b_j)} \tag{2}$$

$P(a_i)$ and $P(a_i b_j)$ are defined with a pseudocount to correct for MSAs with low counts.

$$P(a_i) = \frac{\lambda + N(a_i)}{\sum_x \lambda + N(x_i)} \tag{3}$$

$N(a_i)$ is the count of amino acid *a* appearing at position *i*. λ is equal to the length of sequences in the MSA divided by 20 for single-amino acid counts ($N(a_i)$) and 400 for double-amino acid counts $N(a_i b_j)$. We also modified MI to include row-column weighting:

$$MI_{rcw} = \frac{MI_{a_i b_j}}{(\sum_x MI_{x_i b_j} + \sum_y MI_{a_i y_j} - MI_{a_i b_j})/19} \quad (4)$$

Obtaining viral genomes from patient populations: The list of complete viral polyprotein genomes was accessed and downloaded from NCBI. These genomes were checked to ensure that the sequence covered all NS3 substrate regions. We translated the DNA sequence that we downloaded from NCBI into a protein sequence and compared the five substrate regions "DLEVVTST", "DEMEECASHL", "EDVVCCSM", ECTTPCSGS" and "ALVTPCASH" to discover the diversity found in the substrate region for the patient genomes.

The dataset of aligned genomes utilized in Cuypers et al. was used for dN/dS measurements and for the mapping of predicted cleaved and uncleaved sequences within the genome (Cuypers et al. 2015).

Supplementary Tables:

1. **Genes:**

Gene	DNA sequence
HCV protease (PDB ID: 3SV6)	**CGGATAACAA TTCCCCTCTA GAAATAATTT TGTTTAACTT TAAGAAGGAG ATATACATATG**GGC AGT CAC ATG GCC TCG ATG AAA AAG AAA GGC TCT GTG GTG ATC GTG GGG CGC ATC AAC CTG TCT GGC GAT ACC GCG TAC GCG CAA CAG ACG CGG GGT GAG GAA GGC TGT CAG GAG ACC TCG CAA ACG GGT CGT GAT AAA AAC CAG GTA GAG GGT GAA GTG CAG ATT GTG AGT ACA GCG ACG CAG ACC TTT CTG GCC ACC TCG ATC AAT GGT GTA CTG TGG ACG GTA TAT CAT GGT GCT GGC ACA CGT ACT ATT GCG TCG CCG AAA GGC CCT GTG ACG CAG ATG TAC ACA AAT GTG GAC AAA GAT TTG GTG GGA TGG CAG GCT CCG CAA GGT AGC CGC AGT TTG ACT CCT TGT ACG TGC GGT TCG TCA GAT CTG TAT CTT GTG ACT CGC CAC GCG GAT GTC ATC CCG GTA CGC CGC

	CGT GGC GAT TCC CGT GGT TCT CTG CTT TCT CCG CGC CCT ATC TCA TAT CTT AAA GGT TCA AGT GGA GGA CCA CTG TTA TGT CCG GCG GGG CAC GCA GTC GGA ATT TTT CGT GCG GCG GTT TCT ACT CGG GGA GTT GCA AAA GCT GTT GAC TTC ATT CCG GTT GAA TCT TTG GAA ACA ACC ATG CGG TCG CCG **CTCGAGCAC CATCACCACC ACCACTGA**

2. Cell sorting statistics:

	Functional pool	Sort Round	Cell #
1	CLEAVED UNCLEAVED	1	420 K 109 K
	CLEAVED MIDDLE UNCLEAVED	2	1.05M 105K 775K + 295K
	CLEAVED MIDDLE UNCLEAVED	3	1.55M 89K 675K
2	CLEAVED UNCLEAVED	1	1 M 205 K
	CLEAVED MIDDLE UNCLEAVED	2	1.15M 300K 1.05 M
	CLEAVED MIDDLE UNCLEAVED	3	2M 262K 707K
9	CLEAVED UNCLEAVED	1	812K + 2.65 M 359K
	CLEAVED MIDDLE UNCLEAVED	2	1.4 M 94 K 1.02 M
	CLEAVED MIDDLE UNCLEAVED	3	1.77 M 324 K 1.5 M
10	CLEAVED UNCLEAVED	1	2.7 M 646 K
	CLEAVED MIDDLE UNCLEAVED	2	1.04M 183K 1.06 M
	CLEAVED MIDDLE UNCLEAVED	3	1.59M 1.16M 1.5M

3. List of oligomers for next - sequencing library generation

Primer	DNA Sequence
NNK library reverse primer	TTTCACTGCCTTTATCATCATCATCTTTATAATCACTGCC CAAATGAGAAGCACAMNNMNNMNNMNNMNNCGACCC TCCGCCTCCGCTACCGCCTCCACC
Library insert forward primer	CTGGGGTAATTAATCAGCGAAGCGATGATTTTTGATCTA TTAACAGATATATAAATGC
Vector forward primer	GGCAGTGATTATAAAGATGATGATGATAAAGGCAGTGA AA
Vector reverse primer	GCATTTATATATCTGTTAATAGATCAAAAATCATCGCTT CGCTGATTAATTACCCCAG
Insert amplification post library generation	TTTCACTGCCTTTATCATCATCATCTTTATAATCACTGCC

4. List of oligos for Illumina sample prep and sequencing

Primers	Sequence
Illumina Insert Amplification Forward	CGT TCC AGA CTA CGC TCT GCA GGC TA
Illumina Insert Amplification Reverse	GGC AGT GAT TAT AAA GAT GAT GAT GAT AAA GGC AGT G
Sequencing LYSeq_114	GCC GGA CAG GAT GAT TCT GCC TAC GAT TAC TAC TGA GCC
Sequencing P104	GGATATTACATGGGAAAACATGTTGTTTACGGAG

5. Deep sequencing processing statistics

Variant	Population	Initial		Post-thresholding		Post-categorization
		Unique Counts	Unique Ratios	Unique Counts	Unique Ratios	Unique Sequences
WT	Background	379361		74575		
	Cleaved-Rep1	216254	84773	30328	23550	7472
	Cleaved-Rep2	260764	95730	29238	23692	
	Partial-Rep1	219369	89830	32355	22690	8737
	Partial-Rep2	354253	123573	28631	21986	
	Uncleaved-Rep1	587740	183536	39235	32298	14702

	Uncleaved-Rep2	473115	160980	39115	32053	
R155K/ A156T/ D168A	Background	339049		64406		
	Cleaved	139722	50895	16374	10948	3135
	Partial	270663	108408	40602	29624	11562
	Uncleaved	209209	75869	23432	10425	3703
A156T	Background	367896		68199		
	Cleaved	140479	52199	18718	9911	3644
	Partial	251274	95066	26348	17151	8461
	Uncleaved	277994	109684	29935	17594	9564
D168A	Background	314942		65787		
	Cleaved	197578	65957	19018	10348	4350
	Partial	336654	108567	30535	16929	5780
	Uncleaved	286784	96578	26993	15155	7514

6. List of oligos for testing substrates in yeast surface display

Primers	DNA Sequence
TLIIPCASHL forward	CGGTAGCGGAGGCGGAGGGTCGACATTGATTATTCCTTGTGC
TLIIPCASHL reverse	CTTTATAATCACTGCCCAAATGAGAAGCACAAGGAATAATCAATGTCGAC
ASIIPCASHL forward	CGGTAGCGGAGGCGGAGGGTCGGCGTCAATTATTCCTTGTG
ASIIPCASHL reverse	CTTTATAATCACTGCCCAAATGAGAAGCACAAGGAATAATTGACGCCGA
TATTA forward	CGGTAGCGGAGGCGGAGGGTCGACAGCGACAACAGCGT
TATTA reverse	CTTTATAATCACTGCCCAAATGAGAAGCACACGCTGTTGTCGCTGT
LHTNI forward	GGTAGCGGAGGCGGAGGGTCGTTGCAT ACAAATATT TGTGCTTCTCATTTG
LHTNI reverse	TTATCATCATCATCTTTATAATCACTGCCCAAATGAGAAG CACAAATATTTGTATGCAA
HNTSN forward	GGTAGCGGAGGCGGAGGGTCGCAT AAT ACA TCA AAT TGTGCTTCTCATTTG
HNTSN reverse	TATCATCATCATCTTTATAATCACTGCCCAAATGAGAAGC ACAATTTGATGTATTATG
SQTGQ forward	GGTAGCGGAGGCGGAGGGTCGTCA CAA ACA GGT CAA TGTGCTTCTCATTTG
SQTGQ reverse	TATCATCATCATCTTTATAATCACTGCCCAAATGAGAAGC ACATTGACCTGTTTGTGA
PSTVL forward	GGTAGCGGAGGCGGAGGGTCGCCT TCA ACA GTG TTG TGTGCTTCTCATTTG
PSTVL reverse	TATCATCATCATCTTTATAATCACTGCCCAAATGAGAAGC

	ACACAACACTGTTGAAGG
PSTTL forward	GGTAGCGGAGGCGGAGGGTCGCCT TCA ACA ACA TTG TGTGCTTCTCATTTG
PSTTL reverse	TATCATCATCATCTTTATAATCACTGCCCAAATGAGAAGC ACACAATGTTGTTGAAGG
PSTVF forward	GGTAGCGGAGGCGGAGGGTCGCCT TCA ACA GTG TTC TGTGCTTCTCATTTG
PSTVF reverse	TATCATCATCATCTTTATAATCACTGCCCAAATGAGAAGC ACAGAACACTGTTGAAGG
PSTTF forward	GGTAGCGGAGGCGGAGGGTCGCCT TCA ACA ACA TTC TGTGCTTCTCATTTG
PSTTF reverse	TATCATCATCATCTTTATAATCACTGCCCAAATGAGAAGC ACAGAATGTTGTTGAAGG
LSLQP forward	GGTAGCGGAGGCGGAGGGTCGTTG TCA TTG CAA CCT TGTGCTTCTCATTTG
LSLQP reverse	TATCATCATCATCTTTATAATCACTGCCCAAATGAGAAGC ACAAGGTTGCAATGACAA
LSPQP forward	GGTAGCGGAGGCGGAGGGTCG TTG TCA CCT CAA CCT TGTGCTTCTCATTTG
LSPQP reverse	TATCATCATCATCTTTATAATCACTGCCCAAATGAGAAGC ACAAGGTTGAGGTGACAA
LSLIP forward	GGTAGCGGAGGCGGAGGGTCG TTG TCA TTG ATT CCT TGTGCTTCTCATTTG
LSLIP reverse	TATCATCATCATCTTTATAATCACTGCCCAAATGAGAAGC ACAAGGAATCAATGACAA
LSPIP forward	GGTAGCGGAGGCGGAGGGTCG TTG TCA CCT ATT CCT TGTGCTTCTCATTTG
LSPIP reverse	TATCATCATCATCTTTATAATCACTGCCCAAATGAGAAGC ACAAGGAATAGGTGACAA
LTTQA forward	GGTAGCGGAGGCGGAGGGTCG TTG ACA ACA CAA GCG TGTGCTTCTCATTTG
LTTQA reverse	TATCATCATCATCTTTATAATCACTGCCCAAATGAGAAGC ACACGCTTGTGTTGTCAA
LTTKA forward	GGTAGCGGAGGCGGAGGGTCG TTG ACA ACA AAG GCG TGTGCTTCTCATTTG
LTTKA reverse	TATCATCATCATCTTTATAATCACTGCCCAAATGAGAAGC ACACGCCTTTGTTGTCAA
LTTQL forward	GGTAGCGGAGGCGGAGGGTCG TTG ACA ACA CAA TTG TGTGCTTCTCATTTG
LTTQL reverse	TATCATCATCATCTTTATAATCACTGCCCAAATGAGAAGC ACACAATTGTGTTGTCAA
LTTKL forward	GGTAGCGGAGGCGGAGGGTCG TTG ACA ACA AAG TTG TGTGCTTCTCATTTG
LTTKL reverse	TATCATCATCATCTTTATAATCACTGCCCAAATGAGAAGC ACACAACTTTGTTGTCAA
ECTIP forward	GGTAGCGGAGGCGGAGGGTCG GAA TGT ACA ATT

	CCTTGTGCTTCTCATTTG
ECTIP reverse	TATCATCATCATCTTTATAATCACTGCCCAAATGAGAAGC ACAAGGAATTGTACATTC
DTMEE forward	GGTAGCGGAGGCGGAGGGTCG GAT ACA ATG GAA GAATGTGCTTCTCATTTG
DTMEE reverse	TATCATCATCATCTTTATAATCACTGCCCAAATGAGAAGC ACATTCTTCCATTGTATC
DEMIE forward	GGTAGCGGAGGCGGAGGGTCG GAT GAA ATGATT GAA TGTGCTTCTCATTTG
DEMIE reverse	TATCATCATCATCTTTATAATCACTGCCCAAATGAGAAGC ACATTCAATCATTTCATC
ALGTG forward	GGTAGCGGAGGCGGAGGGTCG GCG TTG GGT ACA GGT TGTGCTTCTCATTTG
ALGTG reverse	TATCATCATCATCTTTATAATCACTGCCCAAATGAGAAGC ACAACCTGTACCCAACGC
RPGPG forward	GGTAGCGGAGGCGGAGGGTCG CGC CCT GGT CCT GGT TGTGCTTCTCATTTG
RPGPG reverse	TATCATCATCATCTTTATAATCACTGCCCAAATGAGAAGC ACAACCAGGACCAGGGCG
ALVTG forward	GGTAGCGGAGGCGGAGGGTCG GCG TTG GTG ACA GGT TGTGCTTCTCATTTG
ALVTG reverse	TATCATCATCATCTTTATAATCACTGCCCAAATGAGAAGC ACAACCTGTCACCAACGC
EEMIQ forward	GGTAGCGGAGGCGGAGGGTCG GAA GAA ATG ATT CAA TGTGCTTCTCATTTG
EEMIQ reverse	TATCATCATCATCTTTATAATCACTGCCCAAATGAGAAGC ACATTGAATCATTTCTTC
QTSEM forward	GGTAGCGGAGGCGGAGGGTCG CAA ACA TCA GAA ATG TGTGCTTCTCATTTG
QTSEM reverse	TATCATCATCATCTTTATAATCACTGCCCAAATGAGAAGC ACACATTTCTGATGTTTG
WSAIP forward	GGTAGCGGAGGCGGAGGGTCG TGG TCA GCG ATT CCT TGTGCTTCTCATTTG
WSAIP reverse	TATCATCATCATCTTTATAATCACTGCCCAAATGAGAAGC ACAAGGAATCGCTGACCA
STPNK forward	GGTAGCGGAGGCGGAGGGTCG TCA ACA CCT AAT AAG TGTGCTTCTCATTTG
STPNK reverse	TATCATCATCATCTTTATAATCACTGCCCAAATGAGAAGC ACACTTATTAGGTGTTGA
GTTIP forward	GGTAGCGGAGGCGGAGGGTCG GGT ACA ACA ATT CCT TGTGCTTCTCATTTG
GTTIP reverse	TATCATCATCATCTTTATAATCACTGCCCAAATGAGAAGC ACAAGGAATTGTTGTACC
HNLAP forward	GGTAGCGGAGGCGGAGGGTCG CAT AAT TTG GCG CCT TGTGCTTCTCATTTG
HNLAP reverse	TATCATCATCATCTTTATAATCACTGCCCAAATGAGAAGC

	ACAAGGCGCCAAATTATG
FDTLN forward	GGTAGCGGAGGCGGAGGGTCG TTC GAT ACA TTG AAT TGTGCTTCTCATTTG
FDTLN reverse	TATCATCATCATCTTTATAATCACTGCCCAAATGAGAAGC ACAATTCAATGTATCGAA
SDYDL forward	GGTAGCGGAGGCGGAGGGTCG TCA GAT TAT GAT TTG TGTGCTTCTCATTTG
SDYDL reverse	TATCATCATCATCTTTATAATCACTGCCCAAATGAGAAGC ACACAAATCATAATCTGA

7. Primers to generate Drug Resistant Mutants

Primer	Sequence
A156T forward	CGTGGGCATATTTAGGACAGCGGTGTGCACCCG
A156T reverse	CGGGTGCACACCGCTGTCCTAAATATGCCCACG
D168A forward	CTAAGGCGGTGGCGTTTATCCCTGTGGAGAAC
D168A reverse	GTTCTCCACAGGGATAAACGCCACCGCCTTAG
Triple Mutant forward	CGTGGGCATATTTAAGACAGCGGTGTGCACCCG
Triple Mutant reverse	CGGGTGCACACCGCTGTCTTAAATATGCCCACG

8. Vector amplification primers for YESS assay

Primers	DNA Sequence
Vector amplification LY104 for-Gibson	CGACCCTCCGCCTCCGCTACC
Vector amplification LY104 rev-Gibson	TGTGCTTCTCATTTGGGCAGTGATTATAAAGATGATGATGATA A

9. SVM parameter tuning: grid search for optimal boxconstraint and rbfsigma parameters. Average AUC is for each set of parameters run with an 80:20 split on the WT experimental full data set for 100 iterations. N/A is shown if the SVM did not converge with these parameters. A boxconstraint of 1 and rbfsigma of 10 was decided on for future calculations.

		boxconstraint					
	AUC	**0.01**	**0.1**	**1**	**10**	**100**	**1000**
rbfsigma	**0.01**	0.5	0.5	0.5	0.5	0.5	0.5
	0.1	0.5	0.5	0.5	0.5	0.5	0.5
	1	0.8715	0.8718	0.872	0.872	0.8723	0.8721
	10	0.9549	0.9811	0.9839	0.9829	0.9809	0.981

	100	0.9695	0.9696	0.975	0.919	0.9825	N/A
	1000	0.9691	0.9691	0.9693	0.9691	0.9748	0.9819

Supplementary References:

Bastian, M., Heymann, S. & Jacomy, M., 2009. Gephi: An Open Source Software for Exploring and Manipulating Networks. Third International AAAI Conference on Weblogs and Social Media, pp.361–362.

Benatuil, L. et al., 2010. An improved yeast transformation method for the generation of very large human antibody libraries. Protein Eng. Des. Sel., 23(4), pp.155-9.

Buslje, C.M. et al., 2009. Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. Bioinformatics, 25(9), pp.1125–1131.

Cuypers, L. et al., 2015. Genetic diversity and selective pressure in hepatitis C virus genotypes 1–6: Significance for direct-acting antiviral treatment and drug resistance. Viruses, 7(9), pp.5018–5039.

Fowler, D.M. et al., 2010. High-resolution mapping of protein sequence-function relationships. Nature Methods, 7(9), pp.741--U108.

Gallinari, P. et al., 1998. Multiple enzymatic activities associated with recombinant NS3 protein of hepatitis C virus. J Virol, 72(8), pp.6758–6769.

Gouveia-Oliveira, R. & Pedersen, A.G., 2007. Finding coevolving amino acid residues using row and column weighting of mutual information and multi-dimensional amino acid representation. Algorithms for Molecular Biology : AMB, 2, p.12.

Kowalsky, C.A. et al., 2015. High-resolution sequence-function mapping of full-length proteins. PLoS ONE, 10(3), pp.1–23.

Romano, K.P. et al., 2012. The molecular basis of drug resistance against hepatitis C virus NS3/4A protease inhibitors. PLoS Pathog, 8(7), p.e1002832.

Wittekind, M. et al., 2001. Modified forms of hepatitis C NS3 protease for facilitating inhibitor screening and structural studies of protease:inhibitor complexes. US6333186 B1

Chapter 5: Conclusion

5.1. Summary

Proteases are ubiquitous to the process of biological signaling. Uncovering the biophysical basis of protease specificity may not only lead to the possibility of designing rational therapies and designing new synthetic biology tools, but also would be important for understanding the general principles of biological information transfer. This dissertation aimed at deepening the understanding of the biophysical basis of protease specificity through the development of generalizable methods for recapitulation and prediction of specificity for varied classes of proteases. We successfully recapitulated the specificity profile for the substrate- protease recognition of the HCV NS3 protease and uncovered biophysical rules that underlie the process of protease – substrate coevolution.

We demonstrated successful use of a generalizable; structure based biophysical approach for protease specificity recapitulation and prediction(Pethe et al. 2017). In this study, we tested that a near attack, pre transition model of substrate acylation was a good static model representative of substrate selectivity in the mechanism of proteolysis. We constructed thousands of high - resolution models of protease- substrate interactions from high resolution crystal structures accessed from the Protein Data Bank as well as cleaved and uncleaved substrate sets available through literature databases. Rosetta and AMBER were used to calculate energies that describe the protease – substrate interaction interface. We uncovered that a linear combination of these energies robustly recapitulated

specificity profiles across our protease datasets. While our structure guided, energy based method outperformed available sequence based methods when tested with an SVM based approach, we noted that adding sequence information to the SVM added orthologous information increasing the SVMs discriminatory power. We further used the SVM to predict novel specificities for the Hepatitis C NS3 viral protease and successfully validated these novel sequences using a yeast based cell surface display assay(Yi et al. 2015; Pethe et al. 2017).

Structure based prediction methods are relatively slow and thus impede the process of predicting and designing multispecificity. We developed a rapid, flexible-backbone self-consistent mean field theory-based technique, MFPred(Rubenstein et al. 2017), for multispecificity modeling at protein-peptide interfaces. Recapitulating specificities for a range of receptors benchmarked the method: protease and kinase enzymes, and protein recognition modules including SH2, SH3, MHC Class I and PDZ domains. We observed robust recapitulation of peptide specificity as well as ~10-1000-fold decrease in computational expense.

Hepatitis C NS3 protease is multispecific, and a key functional player in the viral replication and maturation process. Viral replication operates via a polyprotein that is translated containing core, non - structural and structural proteins that assemble to form a mature, functional virus. The polyprotein is selectively and specifically cleaved by the multispecific viral protease. We mapped the specificity landscape of the HCV NS3/4A protease to obtain a comprehensive understanding of the protease substrate interaction network. Using an in vivo yeast surface display assay, Fluorescence Assisted Cell

Sorting, Next Generation Sequencing technology we were able to experimentally explore ~ 30% of the interaction landscape. To reconstruct the entire landscape, we used the aforementioned SVM based approach using sequence-based information as well as calculated interaction energies. We find extensive clustering of cleavable and uncleavable motifs in sequence space indicating mutational robustness, and thereby providing a plausible molecular mechanism to buffer the effects of low replicative fidelity(Cuypers et al. 2015) of this RNA virus. Specificity landscapes of known drug-resistant(Romano et al. 2010; Romano et al. 2012; Li et al. 2017) variants are similarly clustered indicating that substrates that are recognized by several mutant proteases are not only robust to changes on the protease but also the to changes in substrate residue identity. Our results highlight the key and constraining role of molecular-level energetics in shaping plateau-like fitness landscapes from quasi-species theory.

5.2. Future Directions & Implications

Interrogating the interaction landscape of HCV NS3/4A protease – substrates in our study, brings to light the fact that traversing across the substrate landscape is not constrained by biophysical barriers. This suggests that there are other factors involved in viral evolution that are contributing to purifying selection of canonical substrate sites to preserve residue identity in these regions. So far, we have witnessed viral evolution in nature that increases diversity(Romano et al. 2010; Romano et al. 2012; Li et al. 2017) on the protease interface to compete with antiviral drugs that are introduced to combat Hepatitis C infection. These mutations work via allosteric effects or by introducing clashes in the binding site making it harder for the drug to bind in the active site of the

protease causing weaker binding and reduced effect. Our studies hint at a population of substrates that are biophysically unhindered (on trajectories starting from canonical substrates) and thus able to be sampled in nature, on the substrate landscape and are efficient substrates for not only the drug resistant mutant but also the wild type protease. Through the study we identify the possibility of protease substrate co-evolution of the virus that would evade antiviral therapies aimed at the hepatitis C protease.

The yeast based assay that was chosen for this study has a few limitations – first, being a cell - based study, the inherent bias towards enrichment in each generation is towards faster growing clones as opposed to functional clones. We aim to balance this effect by introducing structure - based features into the SVM classifier, which solely accounts for function. Some of the enrichment profiles are reflective of the yeast codon bias and not of humans (the host where the canonical substrates are evolving). The scheme of the assay involves testing of one substrate per protease and thus other constraining factors such as DNA, RNA secondary structure effects, order of substrate cleavage cannot be tested. Through this assay we do not have sufficient information to predict whether the novel cleaved sequences would indeed produce a functional virus. This would then reduce barriers to resistance associated substrate variants and thus explain the narrow diversity of substrates found in nature.

Our study brings up several avenues to further explore the molecular evolution landscape of the virus as well as questions regarding the emerging drug resistance in the quasispecies of the hepatitis C virus infective population. The yeast based assay followed by a FACS screen and NGS technology coupled with structure based and SVM learning

tools enable several exploratory experiments that could provide answers to these questions. Our experimental scheme is well set up to test the cleavage profiles of proteases in the presence of an antiviral drug. Using this methodology, we could investigate the following questions – does the WT/Drug resistant mutant protease cleavage profile substantially change in the presence of the drug? Can we isolate substrate variants that are partially cleaved/ uncleaved for the WT protease but shift to the cleaved population in the presence of the drug with the drug resistant mutant protease? This assay could also enable the elucidation of the full resistance profile on both the protease as well as the substrate, for a new drug that is designed against Hepatitis C.

A comprehensive exploration of the protease – substrate network, such as our study, strengthens our understanding of the nuances of protease – substrate interface interactions. To enable progress in protease design it is essential to account for protease – substrate covariance data in the design algorithm. Designing smart libraries at the protease interface residues, as well as at substrate positions (P6-P2; known to be specificity determining) and investigating the cleavage profile by comparing and contrasting changes in these protease mutants would give us an understanding of this covariance network. This study of the robustness on the protease side of this interaction network would be the next natural step in generation of designer proteases.

We hope that the dissertation will further provide measures to understand the biophysical basis of protease specificity and further our understanding of the biophysical interaction

landscape of the Hepatitis C NS3 protease, as well as aid in the development in rational design of proteases for therapeutic and synthetic biology use.

5.3. References

Cuypers, L. et al., 2015. Genetic Diversity and Selective Pressure in Hepatitis C Virus Genotypes 1-6: Significance for Direct-Acting Antiviral Treatment and Drug Resistance. Viruses, 7(9), pp.5018–39.

Li, Z. et al., 2017. Naturally occurring drug resistance associated variants to hepatitis C virus direct-acting antiviral agents in treatment-naive HCV genotype 1b-infected patients in China. Medicine, 96(19), p.e6830.

Pethe, M.A., Rubenstein, A.B. & Khare, S.D., 2017. Large-Scale Structure-Based Prediction and Identification of Novel Protease Substrates Using Computational Protein Design. Journal of Molecular Biology, 429(2), pp.220–236.

Romano, K.P. et al., 2010. Drug resistance against HCV NS3/4A inhibitors is defined by the balance of substrate recognition versus inhibitor binding. Proceedings of the National Academy of Sciences of the United States of America, 107(49), pp.20986–20991.

Romano, K.P. et al., 2012. The Molecular Basis of Drug Resistance against Hepatitis C Virus NS3/4A Protease Inhibitors A. Gamarnik, ed. PLoS Pathogens, 8(7), p.e1002832.

Rubenstein, A.B. et al., 2017. MFPred: Rapid and accurate prediction of protein-peptide recognition multispecificity using self-consistent mean field theory A. R. R. Panchenko, ed. PLOS Computational Biology, 13(6), p.e1005614.

Yi, L. et al., 2015. Yeast Endoplasmic Reticulum Sequestration Screening for the Engineering of Proteases from Libraries Expressed in Yeast. Methods in molecular biology (Clifton, N.J.), 1319, pp.81–93.