Computationally-Guided Thermostabilization of the Primary Endoglucanase from *Hypocrea jecorina* for Cellulosic Biofuel Production

Thesis by

Toni M. Lee

In Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2014

(Defended October 28, 2013)

ACKNOWLEDGEMENTS

One decision can dramatically change life’s trajectory. Six years ago during medical school orientation, I sat in a lecture hall listening to the precise details of how my life would unfold. I would memorize. I would test. Barring some catastrophic error, I would remain employed as long as I wished. Most would find such predictable stability desirable, but I yearned for something more unscripted. That day, I left the hall with graduate school in my sights. Some time later, I visited Caltech as a prospective student. Everywhere I turned I saw friendly faces filled with the excitement of discovery. I knew I was in the right place with the right people.

Thank you Steve Mayo for accepting me into your laboratory and keeping in mind my well-being. I appreciate the times we spent discussing my projects, your interest in advancing my professional career, and your willingness to stand up on my behalf in adverse situations. As your student, I’ve had the chance to sit next to generals at a DARPA meeting, travel to Michigan to meet Dow executives, and present my research at the 2013 Protein Society Symposium and Society for Industrial Microbiology and Biotechnology Annual Meeting. Thank you so much for these opportunities.

Thank you Frances Arnold for treating me as if I were one of your students, providing astoundingly accurate guidance and brilliant insight when I was stuck in a rut. I view my collaboration with you and your group one of the highlights of my Caltech experience and will cherish it forever. If I can be half as successful as you, I will consider myself very lucky.

Thanks to my committee, Doug Rees, David Tirrell, and Shu-ou Shan for your guidance and constructive criticism. You have been there unconditionally whenever I needed advice or perspective.

Thanks to Michael Hoffmann, Mike Day, Carol Carmichael, and Xiangyun Wang, for the wisdom you have imparted through your excellent classes. I’ve enjoyed viewing the world from a broader perspective due to your instruction. Thanks as well to Pamela Bjorkman, Bil Clemons, and Alice Huang for providing opportunities to serve as a teaching assistant and a biology course instructor.

In my service as the Chair and Sustainability Advocate of the Caltech Graduate Student Council, I've met so many wonderful people dedicated to improving life at Caltech. To John Onderdonk, Kristen Van Abel, and Michelle McFadden, thank you for inviting me to the Sustainability Council Meetings. I had a marvelous time working with you in organizing the Earth Week and sustainability segment of graduate orientation. Thanks to Felicia Hunt for your service to graduate students. I have learned so much from you about leading a team, accomplishing goals in a timely manner, and having fun while doing so. Thanks as well to Joseph Shepherd, Anneila Sargent, and every graduate student who served with me on the Graduate Student Council, especially Richard Chen, for making my last year at Caltech memorable and meaningful.

Several of my peers have helped me navigate the maze that is graduate school. Phil Romero and Ophelia Venturelli, thank you for encouraging me to attend Caltech when I was in the decision phase. Thanks to my cohort of fellow Biochemistry and Molecular Biophysics students, David Akopian, Aileen Ariosa, Fay Bi, Maja Bialecka-Fornal, Chineye Idigo, Haejin Kang, Katie Schaefer. Although we saw each other infrequently once classes ended, when we did meet up, we had many commiserating stories to share.

Big thanks to the Mayo lab members. Heidi Privett, Ben Allen, and Jennifer Keefe, thank you for sharing your expertise in protein engineering while I was still finding my footing. Matthew Moore, Seth Lieblich, Alex Nisthal, Kurt Mou, Emzo de los Santos, Jackson Cahn, Gene Kym, Samy Hamdouche, Bernardo Sosa Padilla Araujo, Alexandria Berry, Mohsen Chitsaz, and Tim Wannier, it has been a pleasure interacting with you on a daily basis. I will sincerely miss our heated political debates and general lab antics. Please take care of the tape ball for me.

I would also like to extend my gratitude toward my collaborators in the Arnold lab. Thank you to Mary Farrow and especially Devin Trudeau for working with me on the Cel5A crystallization and engineering projects. Finishing these projects seemed impossible at times, but somehow we completed all of our goals. That we had fun in the process attests to the strength of our partnership. Thanks also to fellow cellulase engineers Matt Smith and Indira Wu for your suggestions and encouragement.

I would also like to thank Jens Kaiser, Pavle Nikolovski, and Julie Hoy for their help with the crystallization and structure determination of *Hj*Cel5A.

Beyond the laboratory, I've had the good fortune of nurturing new and old friendships. To my friends from UCLA, Shiho Tanaka, Morgan Beeby, Stuart Sievers, and Jason Forse, I enjoyed our reunion lunches over the years. To Douglas Tham, Pia Ghosh, Greg Kimball, Samuel Lee, and Alexandria Berry, Zakary Singer, Emzo de los Santos, Jackson Cahn, Devin Trudeau, Helen Yu, Betty Wong, Gloria Sheng, Naeem Hussain, and Elizabeth Jensen, and Yashodhan Bhawe, thanks for all the great moments. I'm looking forward to many more.

Last but not least, I want to thank my parents William and Nancy Lee, my sister Keri Lee, my grandma Lai Wah Lee, and my fiancé Jeff LeHew, a fellow Caltech graduate. They have always offered their support and encouragement in any of my endeavors, successful or not. I am blessed to have such kind, caring people in my life.

As my time at Caltech comes to a close, I am again faced with the uncertainty decision-making brings. While I will never know all the nuanced ways life might differ had I not traded a white coat for a lab coat those six years ago, I certainly would have missed an opportunity to interact with some of the most brilliant and colorful people I may ever encounter. To everyone I've known at Caltech, whether you have provided academic advice, examples to emulate, or shoulders to cry on, thank you from the bottom of my heart.

ABSTRACT

The creation of thermostable enzymes has wide-ranging applications in industrial, scientific, and pharmaceutical settings. As various stabilization techniques exist, it is often unclear how to best proceed. To this end, we have redesigned Cel5A (*Hj*Cel5A) from *Hypocrea jecorina* (anamorph *Trichoderma reesei*) to comparatively evaluate several significantly divergent stabilization methods: 1) consensus design, 2) core repacking, 3) helix dipole stabilization, 4) FoldX $\Delta\Delta G$ approximations, 5) Triad $\Delta\Delta G$ approximations, and 6) entropy reduction through backbone stabilization. As several of these techniques require structural data, we initially solved the first crystal structure of *Hj*Cel5A to 2.05 Å. Results from the stabilization experiments demonstrate that consensus design works best at accurately predicting highly stabilizing and active mutations. FoldX and helix dipole stabilization, however, also performed well. Both methods rely on structural data and can reveal non-conserved, structure-dependent mutations with high fidelity. *Hj*Cel5A is a prime target for stabilization. Capable of cleaving cellulose strands from agricultural waste into fermentable sugars, this protein functions as the primary endoglucanase in an organism commonly used in the sustainable biofuels industry. Creating a long-lived, highly active thermostable *Hj*Cel5A would allow cellulose hydrolysis to proceed more efficiently, lowering production expenses. We employed information gleaned during the survey of stabilization techniques to generate *Hj*Cel5A variants demonstrating a 12-15 °C increase in T_{50} (T_{50} = 84-86 °C), an 11-14 °C increase in optimal temperature (T_{opt} = 75-78 °C) and a 60% increase over the maximal amount of hydrolysis achievable using the WT enzyme. We anticipate that our comparative analysis of stabilization methods will prove useful in future thermostabilization experiments.

TABLE OF CONTENTS

LIST OF TABLES

Number *Page*

LIST OF FIGURES

Number *Page*

NOMENCLATURE

AUC	=	area under the curve (of an ROC curve)
*Ba*Cel5A	=	*Bacillus agaradhaerens* Cel5A
BSA	=	bovine serum albumin
CBD	=	cellulose binding domain
CBM	=	cellulose binding module
$\Delta\Delta G$	=	-RTln (f_{mut}/f_{WT}) (Chapter 3), $\Delta G_{mut} - \Delta G_{WT}$ (Chapter 5)
ΔT_{50}	=	the change in T_{50} relative to WT ($\Delta T_{50} = T_{50,mut}$ - $T_{50,WT}$)
DNP2Fcell	=	2,4 dinitrophenyl-2-deoxy-2-fluoro-β-D-cellobioside
HG-1	=	first generation Kemp eliminase design
HG-2	=	second generation Kemp eliminase
HG-3	=	third generation Kemp eliminase
*Hj*Cel5a	=	*Hypocrea jecorina* Cel5A
*Hj*Cel6a	=	*Hypocrea jecorina* Cel6A
*Hj*Cel7a	=	*Hypocrea jecorina* Cel7A
H. jecorina	=	*Hypocrea jecorina*
KE	=	Kemp elimination
MD	=	molecular dynamics
MI	=	mutual information
MSA	=	multiple sequence alignment
PDB	=	Protein Data Bank
ppmv	=	parts per million by volume
RE	=	relative entropy
ROC	=	receiver operator characteristic
rosetta	=	the forcefield employed in Rosetta
Rosetta	=	computational protein design software
s13pt1-4	=	thermostable *Hj*Cel5A variants
SSRL	=	Stanford Synchrotron Radiation Lightsource
T_{50}	=	temperature at which half of the total enzyme is active
*Ta*Cel5A	=	*Thermoascus aurantiacus* Cel5A
TIM	=	triosephosephate isomerase
T_m	=	temperature at which half the total protein is folded
*Tm*Cel5A	=	Thermotoga maritima Cel5A
T_{opt}	=	optimal operating temperature
Triad	=	computational protein design software
TS	=	transition state
TSA	=	transition state analog
WT	=	wild type
X_{AA}	=	any amino acid

CHAPTER 1

Introduction

1.1 Motivation

Life is delicate. In the history of the earth, at least five mass extinctions have occurred, rendering the vast majority of species that have ever lived extinct [1, 2]. The present era is characterized by a sixth mass extinction driven by the diversion of resources towards human demands [3, 4]. Surveys of many known groups of plants and animals reveal rates of extinction at least several hundred times that expected based on the geological record. In addition, history has documented the anthropogenic demise of 10% of the world's bird species, largely due to habitat loss on islands during colonization [5]. Most extinctions arise from habitat loss in species-rich areas called hotspots [6]. Although these areas cover a mere 12% of available land, the majority of desirable locales fall within these boundaries. As of 1995, nearly 20% of the world's population lived within these hotspots [7]. Further development in these areas appears inevitable as human populations continue to rise.

Habitat loss from climate change is perhaps the most pressing threat to biodiversity. With the capacity to simultaneously alter conditions in nearly all environments, this phenomenon is projected to dramatically accelerate the current extinction trend [8-11]. Moreover, climate change cannot discriminate between humans and other species. Impacts of the current warming period will likely manifest as decreased crop yields, more prevalent vector-based disease, redistribution of freshwater, and an increase in natural disasters [11-14].

Energy needs underlie the current warming period. Greenhouse gasses, such as carbon dioxide, absorb infrared radiation, trapping heat in the atmosphere [15]. Since industrialization, the burning of fossil fuels has increased carbon dioxide levels 31% from 280 parts per million by volume (ppmv) to more than 370 ppmv as of 2003 [16]. Even if carbon dioxide levels are reduced to 2000 levels, the planet is projected to continue warming in decades and centuries to come [17, 18]. Mitigation practices, however, may still limit the severity of the problems humanity will encounter.

Reevaluating current transportation practices constitutes an effective means of reducing greenhouse gas emissions and abating climate change. As of 2011, 28% of U.S. greenhouse gas emissions originated from transportation needs with 90% of fuel derived from petroleum [19]. Recognizing that limiting the use of fossil fuels not only reduces levels of harmful pollutants, but also shrinks dependence on foreign sources, the Federal government has supported the development of biofuels as an alternative energy source. Projections estimate that substituting cellulosic ethanol for gasoline can reduce transportation-related greenhouse gas emissions by 85% [20], demonstrating the potential for this technology to counter climate change.

Since the establishment of the first national renewable fuel standard in 2005, the United States has steadily increased biofuel production [21]. Ethanol in particular has risen as a popular alternative fuel. As the sugars in food sources predominantly exist as easily fermentable starches, most fuel-grade ethanol originates from corn feedstock. This practice, however, has created conflict between the food and fuel industries, increasing the price of corn [21, 22]. Furthermore, elevated grain prices have prompted habitat destruction through encouraging farmers to convert rainforests, peatlands, savannas, and grasslands into grain farms [23]. This increased agricultural intensification also leads to increased water, pesticide, and fertilizer use, causing additional ecological strain [22, 24, 25]. Clearly, using grain feedstocks to meet liquid fuel demands is unsustainable.

Deriving biofuels from inedible photosynthetic waste may reduce many of the environmental problems stemming from corn-based production [22]. Popular feedstocks include algae grown in waste water, corn stover, and other agricultural byproducts [26]. Unlike food sources, these materials primarily store sugars as cellulose wrapped in strands of lignin and hemicellulose [27]. The primary component of plant cell walls, cellulose is not only the most abundant biopolymer on the planet, but also a completely renewable resource [28]. The compound consists of glucosyl units linked by β-1,4 glycosidic bonds, facilitating the formation of intra and intermolecular hydrogen bonds between the hydroxyl groups and the pyranose oxygen [29]. Due to the crystalline nature of cellulose and heterogeneous character of lignin and hemicellulose, this “lignocelulosic” material is highly recalcitrant to degradation with a half-life of over four million years [30]. Hydrolysis often requires a harsh chemical [31] or temperature-based pretreatment to remove the lignin and hemicellulose, then digestion with cocktails of relatively expensive cellulose-digesting enzymes called cellulases. Once released, the free glucose can be fermented into liquid fuel.

At \$0.10-\$1.47 gal^{-1} of ethanol, enzyme costs remain the primary barrier to creating profitable lignocellulosic fuels [32-36]. Current U.S. renewable fuel standards mandate the use of 36 billion gallons (140 x 10^{6} m^{3}) by 2022 with at least 16 billion gallons originating from cellulosic biofuels, and a cap of 15 billion gallons for corn-starch ethanol [21]. While this mandate, in conjunction with other federal regulations, allows current lignocellulosic biofuel production to thrive, enzyme production costs must drop to achieve true economic feasibility.

One method of achieving this goal entails designing thermostable cellulase variants. The ability to retain function at high temperatures benefits enzymes in several ways. Thermostable proteins tend to exhibit stability during all stages of their production, storage, and use [37]. This quality extends product lifetime and subsequently reduces costs. For cellulases in particular, higher stability allows reactions to occur under harsh conditions remaining from feedstock pretreatment [38]. Typically, lignocellulosic material is heated to ~200 °C [39] to expose crystalline cellulose. Performing hydrolysis reactions at temperatures higher than the current industry standard (50 °C) [38, 40] would reduce the amount of energy necessary to cool the pretreated substrate. Higher reaction temperatures also reduce solution viscosity, lowering the energy of mixing. Finally, elevated temperatures reduce microbial contamination and increase reaction rates [41, 42].

1.2 *Hypocrea jecorina* Cellulases

Nature has equipped today's bioengineer with a wide array of cellulolytic enzyme templates. With the sheer amount of cellulolytic systems in existence, choosing one in particular becomes a difficult task. Both bacterial and fungal organisms have been evaluated for enzyme cocktail production. *Clostridium, Cellulomonas, Bacillus, Thermomonospora, Ruminococcus, Bacteriodes, Erwinia, Acetovibria, Microscora,* and *Streptomyces* bacteria are known to produce cellulases. In particular, *Cellulomonas fimi, Bacteroides cellulosolvens*, and *Thermomonospora fusca* have shown promise for cellulase production [38]. These organisms produce cellulases with high specific activity, but with low enzyme titers. In addition, many bacteria exhibit slow growth rates and require anaerobic growth conditions. For these reasons, most commercial cellulase production research has focused on fungi [43]. Fungal organisms with cellulolytic capabilities include *Sclerotium rolfsii, P. chrysosporium,* and species of *Trichoderma, Aspergillus, Schizophyllum,* and *Penicillium* [43, 44]. The current industrial favorite is *Hypocrea jecorina* (anamorph *Trichoderma reesei*). Capable of secreting native proteins with yields of 100 g L^{-1} [45], this filamentous fungus has earned the title of "workhorse" for the lignocellulosic biofuels industry [46].

H. jecorina digests cellulosic material using a cocktail of secreted cellulases. For efficient, synergistic cellulose degradation, three classes are required: 1) exoglucanases which processively remove two-unit glucosides called cellobiose from free chain-ends, 2) endoglucanases which target regions of low crystallinity in the middle of cellulose fibers, and 3) β-glucosidases which hydrolyze cellobiose into glucose monomers [47]. Known cellulases in *H. jecorina* include two exoglucanases (Cel6A (CBHII) and Cel7A (CBHI)), eight endoglucanases (Cel5A (EGII), Cel5B, Cel7B (EGI), Cel12A (EGIII), Cel45A (EGV), Cel61A (EGIV), Cel61B, and Cel74A (EGVI)), and seven β-glucosidases (Cel1A (BGLII), Cel1B, Cel3A (BGLI), Cel3B, Cel3C, Cel3D, and Cel3E) [48]. Several putative β-glucosidases have also been identified in the CAZy database. While this list of cellulases appears extensive, much of the cellulolytic activity can be attributed to four principle enzymes: 1) Cel7A, Cel6A, Cel5A, and Cel7B [49, 50]. In particular, deletion of Cel7A, Cel6A, and Cel5A reduced activity on filter paper by 70, 33, and 12%, respectively [49]. Most of these cellulases consist of an O-glycosylated linker tethering two domains: 1) a catalytic domain and 2) a cellulose binding domain (CBD). The CBD among *H. jecorina* cellulases share ~70% sequence identity and exhibit high thermostability [51]. In addition, recently obtained genomic sequence data for *H. jecorina* strain QM6a indicates that this organism has the fewest cellulases out of all surveyed species capable of hydrolytically degrading plant cell walls [48]. Consequently, creating thermostable cocktails requires relatively little engineering.

Efforts to improve *H. jecorina* cellulase thermostability have met with great success. In 2012, Komor et al. reported the design of a chimeric Cel7A variant 9.2 °C more thermostable than the most thermostable parent [52]. The temperature at which this enzyme exhibited half maximal activity (T_{50}) was 72.1 °C. This improved variant also demonstrated a 10 °C increase in the optimal reaction temperature to 65 °C and a 50% increase in total sugar release from crystalline cellulose. Earlier this year, Wu and Arnold reported the design of a chimeric Cel6A variant 15 °C more thermostable than the most stable parent, *Humicola insolens* (*Hi*Cel6A) (T_{50} = 80.1 °C) [53]. The optimal temperature of this enzyme is 75°C, 15 °C higher than that of the most thermostable parent. This improved thermostability allows for the release of 2.4 times more cellobiose equivalents at its optimum temperature compared with the maximum amount achievable with *Hi*Cel6A. To date, no reports of thermostable *Hj*Cel5A variants exist.

1.3 Thermostabilization Techniques

When it comes to protein stabilization, few concrete rules apply. Almost any method can generate thermostable variants with some efficiency [54]. In fact, analysis of highly thermostable proteins from hyperthermophilic organisms demonstrates that the only commonality between these structures is the presence of increased salt bridges, especially in networks [55]. The placement of these ion pairs, however, often heavily depends on subtleties in the protein structure. All other features including increased hydrogen bonding, improved secondary structure formation, presence of additional disulfide bonds, strengthened hydrophobic packing, decreased surface to volume ratio, more abundant hydrophobic residues in the core, and improved rigidity appear in some, but not all, thermostable proteins. As it is difficult to maximize all of these qualities, certain techniques have arisen to address one or more at a time with varying efficacy.

Computationally-driven methods have a lengthy history of producing thermostable protein variants. As early as 1997, the protein design software ORBIT was employed to generate thermostable streptococcal protein G variants with repacked cores [56]. This software optimizes sidechain rotamers using molecular mechanics forcefields tuned with empirical data. ORBIT has also assisted in the creation of thermostable variants of engrailed homeodomain through improving surface electrostatics [57]. Modifications of the Rosetta protein design software have found use in detecting stabilizing core residues in λ repressor [58]. In addition, computational methods can theoretically be adapted to identify mutations that increase backbone rigidity through redesigning loops, targeting areas with high B-factors, or mutating residues away from highly flexible glycine and toward inflexible proline [59]. As computational capabilities have increased, these methods have expanded to analyze proteins in entirety. Already, Rosetta calculations considering the majority of the protein have successfully contributed to the creation of thermostable antibody scaffolds [60]. These methods are not flawless, however. Rotamer-based computational methods, require extensive training, significant computational resources, and the pre-existence of a high-resolution molecular model.

Predictions based on ΔΔG values provide a more expedient way to probe mutations throughout the entire protein. The computational simplicity of software such as FoldX [61], Dmutant [62], and CUPSAT [63] allows one to calculate energies for all possible mutations and WT. The difference in energy between the mutation and WT residue can then be used to rank all possible mutations within a protein. As demonstrated in the design of Komor et al.'s thermostable Cel7A variants, the rapidity and ease of these calculations renders the technique suitable for combination with other stabilization strategies [52]. In addition, comparative studies have shown that FoldX and Dmutant can predict mutations with a relatively high accuracy of 60% [64]. As is the case with more complex computational methods, this technique also requires a suitable high-resolution molecular model.

Homology-based methods provide a means of creating thermostable protein variants using protein sequence data. These techniques employ multiple sequence alignments (MSAs) to identify mutable regions. In consensus design, residues sampled more frequently at a specific position relative to a background metric (often codon or wild type (WT) frequency) are classified as putatively stabilizing. This approach has proved exceedingly successful in generating thermostable variants of numerous proteins including immunoglobulin domains [65], tetratricopeptide repeats [66], and p53 [67]. A consensus approach also contributed to the creation of the aforementioned thermostable Cel7A variant [52]. In an alternative strategy, sequence information can guide the creation of thermostable chimeric proteins. Correlated mutations in MSAs allow one to calculate residue contact maps. In turn, this information can be used to choose crossover points that minimize the number of disrupted contacts. Through recombining sectors from different proteins, one can generate diversity in numerous characteristics including thermostability. Effective implementation of this technique has contributed to creating both the thermostable Cel7A [68] and Cel6A mutants [69]. These methods require homologous sequence data, a prerequisite that can be limiting for proteins with little homology to current sequences.

Directed evolution can build thermostable mutants with no informational requirements beyond the sequence of the gene of interest [70]. The technique involves generating mutations through error-prone PCR, then performing extensive screening to uncover useful mutations. Many of the mutations in the thermostable Cel6A were detected using directed evolution [53]. This technique, however, is not only time-consuming, but also requires a suitable screen [71].

Each stabilization technique offers advantages countered with shortcomings. In cases where available information meets the prerequisites of more than one strategy, it is often unclear how to proceed. At least one study comparing multiple protein stabilization strategies applied to a single protein exists [60]. Using a combination of Rosetta design, disulfide engineering, consensus design, and domain grafting, the authors raised the T_m of an antibody above 90 °C. These experiments were carried out in series, successively adding mutations to the final construct alongside their discovery. However, this approach obfuscates any improvements generated through each individual method. A more comprehensive comparison of protein strategies evaluated within a single protein system might prove useful for future stabilization projects.

1.4 Thesis Summary

This work documents efforts to comparatively test themostabilization techniques applied to a single protein, the primary endoglucanase in *H. jecorina*. *Hj*Cel5A represents one of the last pieces in the thermostabilized cellulase cocktail puzzle. In addition to creating highly stabilized variants of this protein, we also provide foundational biochemical work crucial to structurally and functionally understanding *Hj*Cel5A. Furthermore, the work here provides recommendations for conducting future thermostabilization projects.

In Chapter 2, I describe the first reported *Hj*Cel5A crystal structure. Thus far, this structure provides the only accurate source of high-resolution structural data for this cellulase. Coming 17, 21, 14, and 10 years after release of the Cel7A [72], Cel6A [73], Cel7B [74], and Cel12A [75] crystal structures, respectively, the Cel5A crystal structure completes the crystallographic survey of core *H. jecorina* cellulases. We use this structural information to computationally detect stabilizing mutations in Chapters 4 and 5.

Chapter 3 discusses consensus design applied to *Hj*Cel5A. In this section, we vary several parameters: 1) the number of sequences incorporated into the alignment, 2) the level of characterization of incorporated sequences, 3) the measures used to assess conservation, and 4) the application of additional covariance criteria, and 5) the numerical thresholds used to classify mutations as stabilizing. Several recommendations for optimal parameters emerge from this study.

Chapter 4 compares computational design targeted to the core or surface/boundary region. These experiments use the protein design software Triad. In one calculation, we attempt to identify stabilizing mutations that improve hydrophobic packing in the core of the protein. The second calculation seeks to identify residues that stabilize the protein through neutralizing the natural dipole in α-helices. In addition to revealing numerous stabilizing mutations, we also provide an analysis to discern whether stabilizing *Hj*Cel5A mutations generally reside in the core, boundary, or surface regions.

In Chapter 5, we compare two methods of determining ΔΔG values: FoldX and Triad. We additionally employ Triad approach to explore how backbone rigidity affects stability and activity. Through mutating glycines to residues containing a Cβ and introducing prolines we uncover several additional stabilizing mutations and discuss the relationship between increased rigidity and activity. Finally, we attempt to introduce new disulfide bridges using a modified version of Triad and Disulfide by Design.

Chapter 6 summarizes the findings of Chapters 3-5 and uses the mutations identified throughout these studies to create more thermostable, more active, and better expressing *Hj*Cel5A variants. We hope that the direct comparisons of the major methods employed in this work will prove useful for future enzymatic stabilization projects. To this end, Appendix A and the attached files contain values from the FoldX, Triad ΔΔG, and MSA calculations for all possible *Hj*Cel5A mutations. These sections also contain extensive information on all 262 single mutants cloned, experimentally characterized, and analyzed during this work.

Finally, Appendix B describes the structural characterization of *de novo* designed Kemp eliminases designed in the laboratory of Dr. Stephen Mayo. The crystal structures demonstrate that while *de novo* enzyme design through computational methods is achievable, there exists much room for improvement. These efforts may one day allow industrially relevant reactions to occur under mild conditions, reducing the ecological footprint of the chemical industry.

1.5 Concluding Remarks

Out of the surveyed stabilization strategies, consensus design was shown to identify highly stabilizing and active mutations with the greatest accuracy. As previously discussed, this technique relies on the preexistence of many homologous sequences. Although improved sequencing technologies have shrunk the cost of surveying whole genomes from millions to thousands of dollars [76], probing all organisms likely requires significant time and financial resources. Humanity has only discovered a fraction of the species on our planet [77] and the current extinction rate suggests that many of these organisms may remain unknown scientifically. Moreover, organisms are much more than their DNA. Full comprehension of even the smallest bacterium requires examining the organism from the tiniest biochemical nuances to its ecology. As humanity continues to divert resources away from potentially useful organisms, extinction shrinks the amount of biological capital available to bioengineers [78, 79]. Perhaps redefining the term progress is in good order.

.

1.1 References

1. Raup, D.M. and Sepkoski Jr J.J. (1982) Mass extinctions in the marine fossil record. Science 215:1501-1503.
2. Benton, M.J. (1995) Diversification and extinction in the history of life. Science (Washington) 268:52-58.
3. Leakey, R. and Lewin R. (1996) The sixth extinction: biodiversity and its survival.
4. Thomas, M.B. (2000) The sixth extinction: How large, where, and when? Nature and Human Society: the quest for a sustainable world:46.
5. Pimm, S.L., Moulton M.P., Justice L.J., Collar N.J., Bowman D. and Bond W.J. (1994) Bird extinctions in the Central Pacific [and discussion]. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 344:27-33.
6. Brooks, T.M., Mittermeier R.A., Mittermeier C.G., Da Fonseca G.A.B., Rylands A.B., Konstant W.R., Flick P., Pilgrim J., Oldfield S. and Magin G. (2002) Habitat loss and extinction in the hotspots of biodiversity. Conservation biology 16:909-923.
7. Cincotta, R.P., Wisnewski J. and Engelman R. (2000) Human population in the biodiversity hotspots. Nature 404:990-992.
8. Hampe, A. and Petit R.J. (2005) Conserving biodiversity under climate change: the rear edge matters. Ecology letters 8:461-467.
9. Thomas, C.D., Cameron A., Green R.E., Bakkenes M., Beaumont L.J., Collingham Y.C., Erasmus B.F.N., De Siqueira M.F., Grainger A. and Hannah L. (2004) Extinction risk from climate change. Nature 427:145-148.
10. Heller, N.E. and Zavaleta E.S. (2009) Biodiversity management in the face of climate change: a review of 22 years of recommendations. Biological conservation 142:14-32.
11. Hughes, T.P., Baird A.H., Bellwood D.R., Card M., Connolly S.R., Folke C., Grosberg R., Hoegh-Guldberg O., Jackson J.B.C. and Kleypas J. (2003) Climate change, human impacts, and the resilience of coral reefs. Science 301:929-933.
12. Patz, J.A., Campbell-Lendrum D., Holloway T. and Foley J.A. (2005) Impact of regional climate change on human health. Nature 438:310-317.
13. Lobell, D.B. and Field C.B. (2007) Global scale climate–crop yield relationships and the impacts of recent warming. Environmental Research Letters 2:014002.
14. Vörösmarty, C.J., Green P., Salisbury J. and Lammers R.B. (2000) Global water resources: vulnerability from climate change and population growth. Science 289:284.
15. Rodhe, H. (1990) A comparison of the contribution of various gases to the greenhouse effect. Science 248:1217-1219.
16. Karl, T.R. and Trenberth K.E. (2003) Modern Global Climate Change. Science 302:1719-1723.
17. Wigley, T.M.L. and Raper S.C.B. (2001) Interpretation of High Projections for Global-Mean Warming. Science 293:451-454.

18. Nakicenovic, N., Alcamo J., Davis G., de Vries B., Fenhann J., Gaffin S., Gregory K., Grubler A., Jung T.Y. and Kram T. Special report on emissions scenarios: a special report of Working Group III of the Intergovernmental Panel on Climate Change. (2000). Pacific Northwest National Laboratory, Richland, WA (US), Environmental Molecular Sciences Laboratory (US).

19. Agency, U.S.E.P. Inventory of U.S. Greenhouse Gas Emissions and Sinks. (2013). Washington, D.C.

20. Wang, M. Updated energy and greenhouse gas emission results of fuel ethanol. (2005) 15th International Symposium on Alcohol Fuels. San Diego, CA. Sept. pp. 26-28.

21. Schnepf, R. and Yacobucci B.D. Renewable Fuel Standard (RFS): Overview and Issues. (2013) Congressional Research Service Report for Congress.

22. Fargione, J., Hill J., Tilman D., Polasky S. and Hawthorne P. (2008) Land clearing and the biofuel carbon debt. Science 319:1235-1238.

23. Secchi, S. and Babcock B.A. 2007. Impact of high crop prices on environmental quality: A case of Iowa and the Conservation Reserve Program, Center for Agricultural and Rural Development, Iowa State University Ames, IA.

24. De Fraiture, C., Giordano M. and Liao Y. (2008) Biofuels and implications for agricultural water use: blue impacts of green energy. Water Policy 10:67.

25. Headey, D. and Fan S. (2008) Anatomy of a crisis: the causes and consequences of surging food prices. Agricultural Economics 39:375-391.

26. Somerville, C., Youngs H., Taylor C., Davis S.C. and Long S.P. (2010) Feedstocks for lignocellulosic biofuels. Science (Washington) 329:790-792.

27. Rubin, E.M. (2008) Genomics of cellulosic biofuels. Nature 454:841-845.

28. Ioelovich, M. (2008) Cellulose as a nanostructured polymer: a short review. BioResources 3:1403-1418.

29. O'Sullivan, A.C. (1997) Cellulose: the structure slowly unravels. Cellulose 4:173-207.

30. Wolfenden, R., Lu X. and Young G. (1998) Spontaneous hydrolysis of glycosides. Journal of the American Chemical Society 120:6814-6815.

31. Kosan, B., Michels C. and Meister F. (2008) Dissolution and forming of cellulose with ionic liquids. Cellulose 15:59-66.

32. Klein ‐ Marcuschamer, D., Oleskowicz ‐ Popiel P., Simmons B.A. and Blanch H.W. (2012) The challenge of enzyme cost in the production of lignocellulosic biofuels. Biotechnology and Bioengineering 109:1083-1087.

33. Jordan, D.B., Bowman M.J., Braker J.D., Dien B.S., Hector R.E., Lee C.C., Mertens J.A. and Wagschal K. (2012) Plant cell walls to ethanol. Biochemical Journal 442:241-252.

34. Kazi, F.K., Fortman J.A., Anex R.P., Hsu D.D., Aden A., Dutta A. and Kothandaraman G. (2010) Techno-economic comparison of process technologies for biochemical ethanol production from corn stover. Fuel 89, Supplement 1:S20-S28.

35. Aden, A. and Foust T. (2009) Technoeconomic analysis of the dilute sulfuric acid and enzymatic hydrolysis process for the conversion of corn stover to ethanol. Cellulose 16:535-545.

36. Timilsina, G.R. and Shrestha A. (2011) How much hope should we have for biofuels? Energy 36:2055-2069.

37. Mingardon, F., Bagert J.D., Maisonnier C., Trudeau D.L. and Arnold F.H. (2011) Comparison of Family 9 Cellulases from Mesophilic and Thermophilic Bacteria. Applied and Environmental Microbiology 77:1436-1442.

38. Sun, Y. and Cheng J. (2002) Hydrolysis of lignocellulosic materials for ethanol production: a review. Bioresource Technology 83:1-11.

39. Liu, C. and Wyman C.E. (2005) Partial flow of compressed-hot water through corn stover to enhance hemicellulose sugar recovery and enzymatic digestibility of cellulose. Bioresource Technology 96:1978-1985.

40. Viikari, L., Alapuranen M., Puranen T., Vehmaanperä J. and Siika-aho M. Thermostable Enzymes in Lignocellulose Hydrolysis. In: Olsson L, Ed. (2007) Biofuels. Springer Berlin Heidelberg, pp. 121-145.

41. Shaw, A.J., Podkaminer K.K., Desai S.G., Bardsley J.S., Rogers S.R., Thorne P.G., Hogsett D.A. and Lynd L.R. (2008) Metabolic engineering of a thermophilic bacterium to produce ethanol at high yield. Proceedings of the National Academy of Sciences 105:13769-13774.

42. Abdel-Banat, B.M.A., Hoshida H., Ano A., Nonklang S. and Akada R. (2010) High-temperature fermentation: how can processes for ethanol production at high temperatures become superior to the traditional process using mesophilic yeast? Applied Microbiology and Biotechnology 85:861-867.

43. Duff, S.J.B. and Murray W.D. (1996) Bioconversion of forest products industry waste cellulosics to fuel ethanol: a review. Bioresource Technology 55:1-33.

44. Sternberg, D. Production of cellulase by *Trichoderma*. (1976) Biotechnology and bioengineering symposium. pp. 35.

45. Saloheimo, M. and Pakula T.M. (2012) The cargo and transport system: secreted proteins and protein secretion in *Trichoderma reesei* (*Hypocrea jecorina*). Microbiology 158:46-47.

46. Seidl, V., Seibel C., Kubicek C.P. and Schmoll M. (2009) Sexual development in the industrial workhorse *Trichoderma reesei*. Proceedings of the National Academy of Sciences.

47. Coughlan, M.P. and Ljungdahl L.G. (1988) Comparative biochemistry of fungal and bacterial cellulolytic enzyme systems. FEMS symposium - Federation of European Microbiological Societies.:11-30.

48. Seiboth, B., Ivanova C. and Seidl-Seiboth V. Trichodera reesei: A Fungal Enzyme Producer for Cellulosic Biofuels. In: Bernardes MADS, Ed. (2011) Biofuel Production-Recent Developments and Prospects. pp. 310-340.

49. Suominen, P.L., Mantyla A.L., Karhunen T., Hakola S. and Nevalainen H. (1993) High frequency one-step gene replacement in *Trichoderma reesei*. II. Effects of deletions of individual cellulase genes. Molecular Genetics and Genomics 241:523-530.

50. Lantz, S., Goedegebuur F., Hommes R., Kaper T., Kelemen B., Mitchinson C., Wallace L., Stahlberg J. and Larenas E. (2010) *Hypocrea jecorina* CEL6A protein engineering. Biotechnology for Biofuels 3:20.

51. Kraulis, P.J., Clore G.M., Nilges M., Jones T.A., Pettersson G., Knowles J. and Gronenborn A.M. (1989) Determination of the three-dimensional solution

structure of the C-terminal domain of cellobiohydrolase I from *Trichoderma reesei*. A study using nuclear magnetic resonance and hybrid distance geometry-dynamical simulated annealing. Biochemistry 28:7241-7257.

52. Komor, R.S., Romero P.A., Xie C.B. and Arnold F.H. (2012) Highly thermostable fungal cellobiohydrolase I (Cel7A) engineered using predictive methods. Protein Engineering Design and Selection 25:827-833.
53. Wu, I. and Arnold F.H. (2013) Engineered thermostable fungal Cel6A and Cel7A cellobiohydrolases hydrolyze cellulose efficiently at elevated temperatures. Biotechnology and Bioengineering 110:1874-1883.
54. Daniel, R.M., Dines M. and Petach H.H. (1996) The denaturation and degradation of stable enzymes at high temperatures. Biochemical Journal 317:1-11.
55. Petsko, G.A. (2001) Structural basis of thermostability in hyperthermophilic proteins, or "There's more than one way to skin a cat". Methods in Enzymology 334:469-478.
56. Dahiyat, B.I. and Mayo S.L. (1997) Probing the role of packing specificity in protein design. Proceedings of the National Academy of Sciences 94:10172-10177.
57. Marshall, S.A., Morgan C.S. and Mayo S.L. (2002) Electrostatics significantly affect the stability of designed homeodomain variants. Journal of Molecular Biology 316:189-199.
58. Borgo, B. and Havranek J.J. (2012) Automated selection of stabilizing mutations in designed and natural proteins. Proceedings of the National Academy of Sciences 109:1494-1499.
59. Matthews, B.W., Nicholson H. and Becktel W.J. (1987) Enhanced protein thermostability from site-directed mutations that decrease the entropy of unfolding. Proceedings of the National Academy of Sciences 84:6663-6667.
60. McConnell, A.D., Spasojevich V., Macomber J.L., Krapf I.P., Chen A., Sheffer J.C., Berkebile A., Horlick R.A., Neben S., King D.J. and Bowers P.M. (2012) An integrated approach to extreme thermostabilization and affinity maturation of an antibody. Protein Engineering Design and Selection.
61. Schymkowitz, J., Borg J., Stricher F., Nys R., Rousseau F. and Serrano L. (2005) The FoldX web server: an online force field. Nucleic Acids Research 33:392-388.
62. Zhou, H. and Zhou Y. (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Science 11:2714-2726.
63. Parthiban, V., Gromiha M.M. and Schomburg D. (2006) CUPSAT: prediction of protein stability upon point mutations. Nucleic Acids Research 34:W239-W242.
64. Khan, S. and Vihinen M. (2010) Performance of protein stability predictors. Human Mutation 31:675-684.
65. Ohage, E. and Steipe B. (1999) Intrabody construction and expression. I. The critical role of VL domain stability. Journal of Molecular Biology 291:1119-1128.
66. Main, E.R.G., Xiong Y., Cocco M.J., D'Andrea L. and Regan L. (2003) Design of Stable α-Helical Arrays from an Idealized TPR Motif. Structure 11:497-508.
67. Nikolova, P.V., Henckel J., Lane D.P. and Fersht A.R. (1998) Semirational design of active tumor suppressor p53 DNA binding domain with enhanced stability. Proceedings of the National Academy of Sciences 95:14675-14680.

68. Smith, M.A., Bedbrook C.N., Wu T. and Arnold F.H. (2013) *Hypocrea jecorina* cellobiohydrolase I stabilizing mutations identified using noncontiguous recombination. ACS Synthetic Biology.

69. Heinzelman, P., Snow C.D., Wu I., Nguyen C., Villalobos A., Govindarajan S., Minshull J. and Arnold F.H. (2009) A family of thermostable fungal cellulases created by structure-guided recombination. Proceedings of the National Academy of Sciences 106:5610-5615.

70. Romero, P.A. and Arnold F.H. (2009) Exploring protein fitness landscapes by directed evolution. Nature Reviews Molecular Cell Biology 10:866-876.

71. Arnold, F.H. and Georgiou G. 2003. Directed enzyme evolution: screening and selection methods, Springer.

72. Divne, C., Stahlberg J., Reinikainen T., Ruohonen L., Pettersson G., Knowles J.K., Teeri T.T. and Jones T.A. (1994) The three-dimensional crystal structure of the catalytic core of cellobiohydrolase I from *Trichoderma reesei*. Science 265:524-528.

73. Rouvinen, J., Bergfors T., Teeri T., Knowles J.K. and Jones T.A. (1990) Three-dimensional structure of cellobiohydrolase II from *Trichoderma reesei*. Science 249:380-386.

74. Kleywegt, G.J., Zou J.-Y., Divne C., Davies G.J., Sinning I., Ståhlberg J., Reinikainen T., Srisodsuk M., Teeri T.T. and Jones T.A. (1997) The crystal structure of the catalytic core domain of endoglucanase I from *Trichoderma reesei* at 3.6 Å resolution, and a comparison with related enzymes. Journal of Molecular Biology 272:383-397.

75. Sandgren, M., Shaw A., Ropp T.H., Wu S., Bott R., Cameron A.D., Ståhlberg J., Mitchinson C. and Jones T.A. (2001) The X-ray crystal structure of the *Trichoderma reesei* family 12 endoglucanase 3, Cel12A, at 1.9 Å resolution. Journal of Molecular Biology 308:295-310.

76. Margulies, M., Egholm M., Altman W.E., Attiya S., Bader J.S., Bemben L.A., Berka J., Braverman M.S., Chen Y.-J. and Chen Z. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376-380.

77. Costello, M.J., May R.M. and Stork N.E. (2013) Response to comments on “Can we name Earth’s species before they go extinct?”. Science 341:237.

78. Ceballos, G. and Ehrlich P.R. (2002) Mammal population losses and the extinction crisis. Science 296:904-907.

79. Costanza, R., d'Arge R., De Groot R., Farber S., Grasso M., Hannon B., Limburg K., Naeem S., O'Neill R.V. and Paruelo J. (1997) The value of the world's ecosystem services and natural capital. Nature 387:253-260.

CHAPTER 2

A Structural Study of *Hypocrea jecorina* Cel5A

A version of this chapter has been published as [1].

2.1 Abstract

Interest in generating lignocellulosic biofuels through enzymatic hydrolysis continues to rise as non-renewable fossil fuels are depleted. The high cost of producing cellulases, hydrolytic enzymes that cleave cellulose into fermentable sugars, currently hinders economically viable biofuel production. Here we report the crystal structure of a prevalent endoglucanase in the biofuels industry, Cel5A from the filamentous fungus *Hypocrea jecorina*. The structure reveals a general fold resembling that of the closest homolog with a high-resolution structure, Cel5A from *Thermoascus aurantiacus*. Consistent with previously described endoglucanase structures, the *H. jecorina* Cel5A active site contains a primarily hydrophobic substrate binding groove and a series of hydrogen bond networks surrounding two catalytic glutamates. The reported structure, however, demonstrates stark differences between side-chain identity, loop regions, and the number of disulfides. Such structural information may aid efforts to improve the stability of this protein for industrial use while maintaining enzymatic activity through revealing non-essential and immutable regions.

2.2 Introduction

Lignocellulosic biofuels have enjoyed recent popularity as sustainable energy alternatives to fossil fuels. In current enzymatic conversion schemes, a pretreatment step with high temperatures or extreme pH conditions removes indigestible lignin from feedstock materials. Cellulase cocktails then break cellulose polymers into component sugars suitable for fermentative fuel production. To achieve efficient digestion, three types of cellulases must exist in the preparation: (1) exoglucanases to cleave cellobiose molecules from cellulose strand termini, (2) endoglucanases to cleave strands internally, and (3) β-glucosidases to cleave cellobiose into glucose monomers [2]. Few known organisms adequately produce cellulases from all three classes. Consequently, the filamentous fungus *Hypocrea jecorina* (*Trichoderma reesei*), a prodigious source of each cellulase class, enjoys wide-spread use in the biofuels industry [3]. Enzyme production costs, however, still constitute a limiting factor to wide-scale bioethanol synthesis. Although advances in all areas of enzyme production have decreased costs to 20 to 30 cents per gallon of ethanol, less-sustainable, corn-derived fuel remains the cheaper alternative at 3 to 4 cents per gallon [4]. One strategy for further reducing enzymatic costs involves extending cellulase lifetimes through enhanced stability. As some protein engineering strategies utilize atomic-resolution models to guide the design process, obtaining crystal structures of each cellulase may significantly aid such endeavors. Thus far, efforts to crystallize *H. jecorina* cellulases have resulted in catalytic domain structures of exoglucanases Cel6A (CBHII) [5] and Cel7A (CBHI) [6] and endoglucanases Cel7B (EGI) [7] and Cel12A (EGIII) [8]. Cel5A (EGII), however, accounts for as much as 55% of *H. jecorina* endoglucanase activity [9], yet has resisted previous crystallographic solution. Here we provide the crystal structure of *H. jecorina* Cel5A (*Hj*Cel5A) resolved to 2.05 Å.

2.3 Results

With the exception of Cel12A, most *H. jecorina* cellulases consist of a heavily O-glycosylated linker tethering a small cellulose binding domain (CBD) to a larger catalytic domain. CBDs of this organism share ~70% sequence identity [10] and a solution structure of the Cel7A CBD has been solved [11]. To minimize sample inhomogeneity resulting from glycosylation, the isolated *H. jecorina* Cel5A catalytic core was expressed in *Escherichia coli* BL21 (DE3) cells. The protein was crystallized, data were collected to 2.05 Å, and the structure solved and refined with an R_{work}/R_{free} of 16.3/20.5% (Table I and Supporting Information Fig. S1).

*Hj*Cel5A adopts a $(\alpha/\beta)_8$ TIM-barrel fold common to other family 5 glycoside hydrolases (Figure 1A). The general topology bears a striking resemblance to Cel5A from *Thermoascus aurantiacus* (*Ta*Cel5A, RMSD of 1.4 Å [12]) (Figure 1B) with 29% sequence identity and 65% sequence similarity (Supporting Information Fig. S2). While both proteins demonstrate similar placement of most secondary structure elements, the *H. jecorina* homolog exhibits extensions in the β1-α1, β3-α3, and α5-β6 loops (see Supporting Information Fig. S3 for secondary structure numbering). The β1-α1 loop projects towards the active site, forming a relatively shallow substrate binding groove. In addition to eight canonical β-strands, the structure also contains a protruding β-hairpin consisting of residues 308 to 315. Sidechain densities along the tip of the loop could not be resolved, suggesting flexibility of the region. Tryptophan 314, however, appears to anchor the C-terminal region of the hairpin to the face of the protein as it rejoins the globular region to form a truncated α8 helix. Although similar β-hairpins appear in the structures of *Thermotoga maritima* Cel5A [13] (*Tm*Cel5A) (3MMW, residues 295-302) and *Clostridium cellulovorans* endoglucanase D (3NDY, residues 324-331), it remains unclear whether this hairpin assumes a functional role. A series of hydrophobic residues (F4, Y98, W142, F177, I214, L287) shields the active site from solvent rather than a short 2-3 β-strand [14] and/or the small N-terminal α-helix plug observed in homologous structures [13].

Glycosylation

Mass spectrometry studies demonstrate that *Hj*Cel5A contains a single GlcNAc N33-linked glycosylation when expressed in the organism of origin [15]. The structure contains no discernable density compatible with such a modification, as expected for a bacterially-expressed protein. N33 is, however, solvent exposed and does not preclude previous findings.

Active site architecture

Consistent with structural studies of other GH5 endoglucanases, the substrate binding pocket consists of a deep catalytic cleft within a shallow binding groove. The deeper cleft contains a hydrophobic patch (F14, V27, Y28, Y40, F34, W292, A294, F297, Y301) surrounded by the β1-α1 loop (residues 15-22), the sidechain of W185, residues 104-107, residues 146-150, and the β6-α6 loop (residues 225-229) (Figure 1C). A short α-helical ledge (residues 183-187) abruptly terminates this hydrophobic groove in a manner that superficially appears incompatible with endoglucanase function—internal cellulose cleavage might require that the substrate thread through the deep cleft to access the active site. The ledge itself, however, forms a shallower hydrophilic groove. This architecture suggests that an extended cellulose chain initially binds to the shallow groove in a non-catalytic manner. Crystallographic studies of the *Bacillus agaradhaerens* Cel5A suggest that the Michaelis complex subsequently forms as the +1 site sugar adopts a $^1S_3$ skew-boat conformation [16]. W185 facilitates formation of this catalytic conformation through stacking with the −1 site sugar ring [Fig. 1(C)]. The resulting ~110°-115° kink allows the substrate to pass over the helical ledge into solvent allowing for the internal cleavage of long cellulose strands. Previous studies characterize *Hj*Cel5A as a promiscuous enzyme that generates a wide range of products including glucose, cellobiose, and cellotriose [17]. The non-catalytic binding groove appears more hydrophilic and shallower than that of *Ta*Cel5A. Further testing may reveal whether product inhomogeneity results from scant interaction between *Hj*Cel5A and the reducing end of the chain beyond the active site.

The obtained *Hj*Cel5A structure depicts an active enzyme as determined by comparison to homologous structures. Like other retaining cellulases, *Hj*Cel5A hydrolyzes internal β-1,4-glycosidic cellulosic bonds through a double-displacement mechanism involving two carboxylates [16]. First, a general acid/base catalyst protonates the glycosidic bond to promote cleavage. A second carboxylate then forms a covalent glucosyl-enzyme intermediate through an oxocarbonium ion transition state, displacing a newly-generated non-reducing cellulose terminus. The apo enzyme finally forms through a second oxocarbonium ion transition state. In *Hj*Cel5A, the terminal oxygen atoms of the general base (E148) and nucleophile (E259) are separated by ~5 Å, typical of retaining β-glycosidases [18]. These residues were identified through homology with *Ta*Cel5A and confirmed as necessary to catalysis through site-directed mutagenesis (Supporting Information Fig. S4). Residues T258, H218, and E148 form a type A catalytic triad involved in raising the pK_a of the donor carboxylate to promote more efficient substrate protonation [19] (Figure 1D). A hydrogen-bonding network around E259 also exists. R60 and Y220 position the nucleophilic glutamate for catalysis through contacting OE2 and OE1, respectively. N147 in turn tethers R60 in place. Although H104 and W292 are conserved across GH5 cellulases and reside near the active site, these residues appear to assist with substrate binding rather than influence the catalytic machinery [12].

Disulfide bonds

*Hj*Cel5A contains eight cysteines, all of which are involved in the formation of disulfide bridges (Figures 2A and B). The covalent link between C16 and C22 tethers the C- and N-terminal regions of the β1-α1 loop that forms one wall of the substrate binding pocket. Near the C-terminal region, residues 273 and 323 anchor the final α-helical segment to the adjacent α7 helix. *Hj*Cel5A exhibits a relatively high apparent T_m of 72°C (Supporting Information Figure S5) that may be due in part to stability conferred by disulfide bonding. The hyperthermostable *Ta*Cel5A exhibits two higher melting transitions at 77°C and 81°C [20], yet contains a single disulfide bond at a location homologous to the linkage between C232 and C268. Observations from homologous structures, however, suggest that the thermostability of *Ta*Cel5A may largely arise due to the truncation of loops, a highly pronounced feature in the *Ta*Cel5A homolog [13]. Our attempts to mutate several disulfide-bonded cysteines to serines resulted in insoluble protein expression (data not shown).

2.4 Discussion

*Hj*Cel5A constitutes only 1-10% of the total cellulase protein in *H. jecorina*, yet accounts for 55% of the total endoglucanase activity [9, 21]. The structural data presented here shows that the protein differs in sidechain identity and loop placement from its most similar crystallographically-probed homolog, *Ta*Cel5A. Additionally, the structure reveals four disulfide bonds, in direct contrast with a previous report suggesting the absence of such elements [22]. While an attempt to engineer *Hj*Cel5A for optimum catalytic efficiency at a particular pH has met with some success, this effort relied on a highly inaccurate homology model built from *Ta*Cel5A coordinates [23]. The information presented here may better inform future efforts to rationally engineer *Hj*Cel5A for various needs, as well as understand the wild-type activity of the protein.

2.5 Materials and Methods

Protein expression and purification

The catalytic domain of *Hj*Cel5A (Genbank JN172972) was expressed in BL21 (DE3) cells and purified as described in the Supporting Information. Cultures were grown at 37°C to an optical density of ~0.5 in LB, induced, then allowed to express protein at 16°C for 24 hours. Purification was achieved through His-tag affinity chromatography and proteins were buffer exchanged into storage buffer (10 m*M* acetate pH 4.8, 100 m*M* NaCl) at a final concentration of 5.3 mg/mL.

Crystallization, data collection and structure determination

Hexagonal plate crystals grew in 21 days by the sitting-drop vapor diffusion method in 0.1 *M* sodium citrate, 1 *M* magnesium sulfate, and 1 m*M* cellobiose. Crystals were flash frozen in cryoprotectant and shipped to beamline 12-2 at the Stanford Synchrotron Radiation Lightsource (SSRL) where a 2.1 Å data set was obtained. Phases were obtained through molecular replacement using a 1H1N mixed model generated with SCWRL [24]. Following molecular replacement, model building and refinement were accomplished with the AutoBuild Wizard in PHENIX [25]/COOT [26] and PHENIX [27] respectively. NCS restraints were applied to all refinement steps. Final coordinates were deposited in the Protein Data Bank with the code 3QR3. Data collection and refinement statistics are listed in Table I.

2.6 Supplementary

Protein expression and purification

The catalytic domain of *Hj*Cel5A was expressed in *Escherichia coli*. Existing constructs were obtained from the laboratory of Frances Arnold in which only the sequence corresponding to the catalytic domain of the protein was cloned into the NcoI/XhoI sites of pET22b+. The protein sequence of the coding region is identical to that of an EGII sequence recently deposited to www.ncbi.nlm.nih.gov (accession number: JF340120.1) with the following two exceptions: the first 10 residues (TSSSTPPTSS) were substituted with methionine and a GGSGSG linker and a C-terminal His6 tag were added through QuikChange mutagenesis (Stratagene) for affinity purification. Clones were sequence verified and transformed into BL21(DE3) cells.

Cultures were grown at 37°C to an optical density of ~0.5 in LB. Induction was achieved by adding isopropyl β-D-1-thiogalactopyranoside (IPTG) to a final concentration of 1 m*M* and allowing cells to shake at 220 rpm for 24 hours at 16°C. Cells were collected through centrifugation at 5000 g, 4°C for 15 min. The resulting pellets were resuspended in lysis buffer (12.5 m*M* Tris pH 7.5, 12.5 m*M* MOPS, 0.1% Tween 20) spiked with a small amount of lysozyme and 10 μL benzonase per liter of culture. Full lysis was achieved through sonication followed by a 30 min incubation at 4°C with rocking. The lysate was cleared through centrifugation at 15,000 g, 4°C for 30 min. Supernatant was nutated with Ni-NTA agarose slurry (Qiagen) for 1 hour and loaded onto a gravity column for affinity chromatography. The column was washed once with lysis buffer, once with wash buffer (50 m*M* NaH_2PO_4 pH 8.0, 300 m*M* NaCl, 20 m*M* imidazole) and eluted in elution buffer (50 m*M* NaH_2PO4 pH 8.0, 300 m*M* NaCl, 250 m*M* imidazole). Eluted protein was buffer exchanged into storage buffer (10 m*M* acetate pH 4.8, 100 m*M* NaCl) and run over a Superdex 75 size exclusion column. Fractions were assessed for purity by gel electrophoresis and solutions deemed pure were combined in PES-membrane spin concentrators (Sartorium Stedim). Since the expressed protein precipitates at concentrations exceeding 7 mg/mL, samples were processed to a final concentration of 5.3 mg/mL and stored at 4°C for crystallization assays.

Crystallization and data collection

Crystals were obtained through sitting-drop vapor diffusion in drops containing 60% 5.3 mg/mL protein solution and 30% mother liquor (0.1 *M* sodium citrate, 1 *M* magnesium sulfate, 1 m*M* cellobiose). Although cellobiose was present in the mother liquor, no corresponding density appeared in the final map. Small hexagonal crystals appeared after a week and ceased growth after 21 days. The resulting thick hexagonal plates were harvested and flash frozen in liquid nitrogen using a 30% glycerol solution as a cryoprotectant. Frozen crystals were shipped to beamline 12-2 at the Stanford Synchrotron Radiation Lightsource (SSRL) and diffraction data were collected to 2.05 Å at a temperature of 100 K. The crystals were found to belong to space group $P2_12_12_1$ with unit cell parameters a = 82.95 Å, b = 84.593 Å, c = 90.11 Å, and $\alpha = \beta = \gamma = 90°$ and to contain two *Hj*Cel5A monomers in each asymmetric unit.

Structure determination

The major endoglucanase from *Thermoascus aurantiacus, Ta*Cel5A (PDB ID 1H1N), shares only ~30% protein sequence identity with the *H. jecorina* homolog, yet demonstrates the greatest similarity to the target protein among homologues with solved structures as identified through the FFAS03 server [28]. Attempts to determine phases using the program PHASER [29] and coordinates for a monomeric 1H1N as a search model failed. A molecular replacement solution was, however, obtained in PHASER using a 1H1N mixed model generated using SCWRL [24], wherein all non-conserved residues are replaced with serines. Following molecular replacement, the resulting solution was entered as an initial model for automated model building the AutoBuild Wizard in Phenix [25]. A near complete model was obtained with two molecules in the asymmetric unit having an initial R_{work}/R_{free} of 21.5/24.7%. Additional refinement proceeded using PHENIX [27] in conjunction with the model building program COOT [26]. All refinement steps were performed using chain A to chain B NCS restraints. Ten rounds of refinement in PHENIX were necessary to achieve an R_{work}/R_{free} of 16.3/20.5%. The final model contains residues 0-328 in both chains A and B of the protein, 503 water molecules, nine sulfate molecules, and four magnesium ions (one magnesium has been modeled with occupancy split among two sets of coordinates) (Supporting Information Figure S1). Although backbone density for residues 310-312 clearly exists, sidechains could not be resolved and accordingly do not appear in the model. The final model is calculated to have an overall RMS bond length deviation of 0.011 Å and a covalent angle deviation of 1.2° with 87.2% of residues falling in the most favored regions of Ramachandran space, 12.8% falling within additional allowed regions, 0% in generously allowed regions, and 0% outliers.

Analysis of active site structure

Homologous structures were superimposed using the program Align [30] implemented in PyMOL [31]. Approximate substrate positioning was modeled through aligning the 1.6 Å resolution structure of the *Bacillus agaradhaerens* Cel5A complexed with the slowly-hydrolyzable cellulose analogue 2,4 dinitrophenyl-2-deoxy-2-fluoro-β-D-cellobioside (DNP2Fcell) (PDB ID code 4A3H) [16].

Active site point mutant generation

Mutations E148A, H218A, T258A, and E259A were generated through QuikChange site-directed mutagenesis (Stratagene) and verified through sequencing. Proteins were expressed and batch purified through affinity chromatography as described above.

Enzymatic activity assay

The enzyme assay was performed as described by Park and Johnson [32]. Enzyme-substrate mixtures containing 0.2 μ*M* protein, 0.15% carboxymethyl cellulose, and 10-20 m*M* acetate buffer pH 5.6 were incubated at 42°C for 2 hours and stored at 4°C before developing the solution with a colorimetric reagent. To develop the solution, 300 uL of reagent A (potassium ferrocyanide, 0.5 g/L, dipotassium phosphate, 34.84 g/L, pH 6) was premixed with 150 uL of reagent B (sodium carbonate, 5.3 g/L, potassium cyanide, 0.65g/L) then immediately added to the incubated protein solution. After boiling for 15 min at 95°C, 300 uL of reagent C (ferric chloride, 2.5 g/L, polyvinylpyrrolidone, 10 g/L, sulfuric acid 2N) were added to the mixture to elicit a yellow to blue color change. Experiments were performed in triplicate and the absorbance at 600 nm was measured using a TECAN Infinite M200 96-well plate reader. H218A and T258A failed to express solubly and enzymatic activity data subsequently could not be acquired. *Hj*Cel5A mutants E148A and E259A demonstrate no activity relative to the background reaction (Supporting Information Figure S4).

Circular dichroism

Circular dichroism scans were performed with protein in acetate buffer at a concentration of 5 μ*M* using a 1 mm cuvette. Wavelength scans were performed at 25°C scanning through the 200-250 nm range (Supporting Information Figure S5). The experiment was performed in triplicate with an averaging time of 5 s and a wavelength step of 1.0 nm.

Circular dichroism signal at 220 nm was also employed to monitor thermal denaturation. Protein at 5 μ*M* was monitored from 1-99°C in steps of 1°C. The sample was subjected to an equilibration period of 2 min per each step before collecting measurements. *Hj*Cel5A was found to unfold irreversibly with an apparent T_m of 71.5°C (Supporting Information Fig. S6).

2.7 Tables and Figures

Table I. *Data collection and refinement statistics*

	*Hj*Cel5A
Data collection	
Space group	$P2_12_12_1$
Cell dimensions	
a, *b*, *c* (Å)	82.95, 84.593, 90.11
a, b, g (°)	90.00, 90.00, 90.00
Resolution (Å)	39-2.05(2.16-2.05)
R_{sym} or R_{merge}	0.081(0.268)
I / sI	19.2(2.8)
Completeness (%)	98.8(92.4)
Redundancy	12.5(9.8)
Refinement	
Resolution (Å)	40-2.05
No. reflections	39858
$R_{\mathrm{work}}/R_{\mathrm{free}}$	0.163/0.205
No. atoms	
Protein	4966
Ligand/ion	74
Water	503
B-factors	22.9
Protein	21.9
Ligand/ion	39.5
Water	29.9
R.m.s. deviations	
Bond lengths (Å)	0.011
Bond angles (°)	1.2
Ramachandran map analysis	
Most favored regions	87.2
Additional allowed regions	12.8
Generously allowed regions	0
Disallowed regions	0

Data were collected from one crystal.
Values in parentheses are for highest-resolution shell.

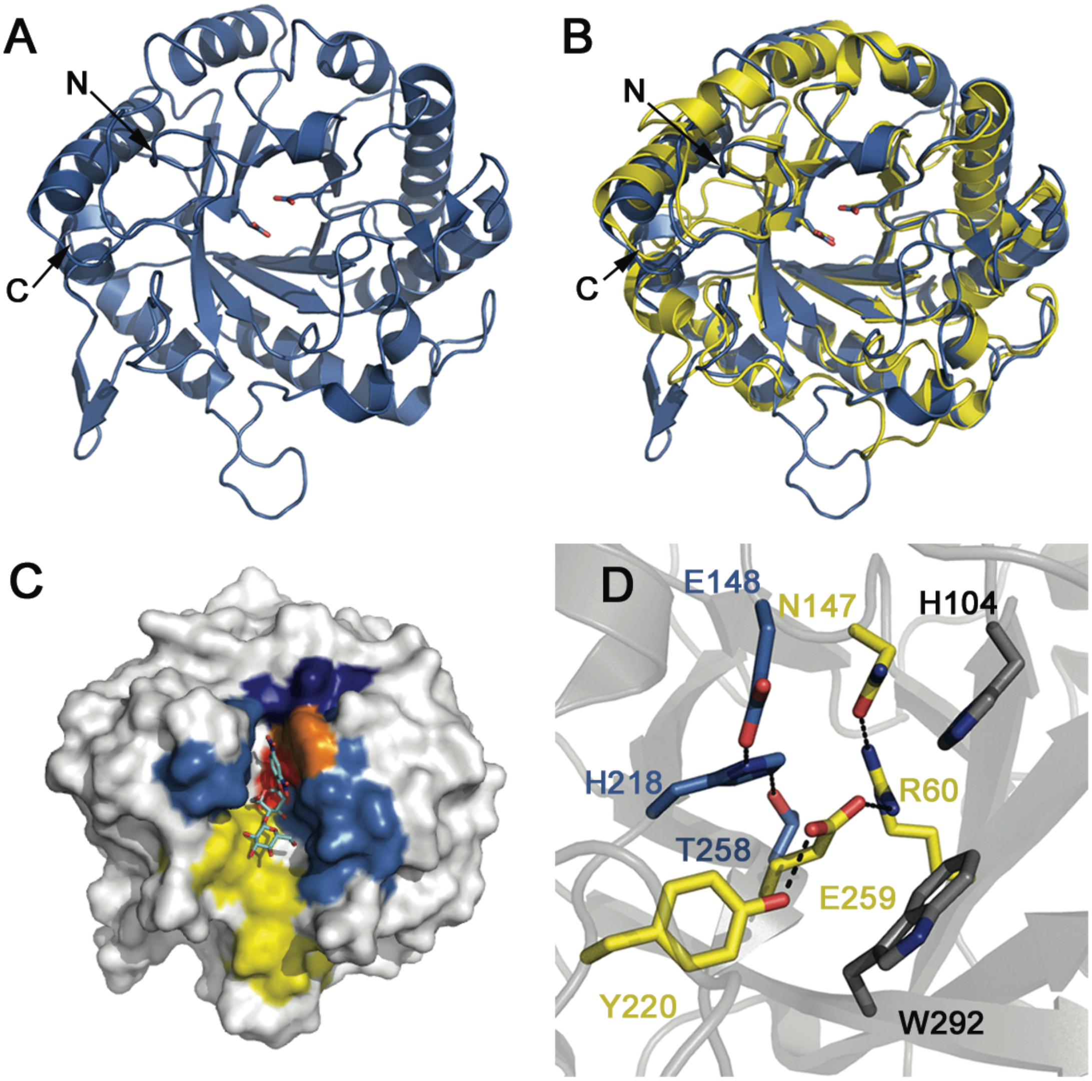


Figure 1. *Structure of HjCel5A.* (A) *Hj*Cel5A shown in cartoon representation with catalytic glutamates shown as sticks. (B) Superposition of *Hj*Cel5A (blue) and *Ta*Cel5A (yellow) generated in PyMOL using the align function. (C) *Hj*Cel5A in surface representation highlighting the hydrophobic substrate docking patch (yellow), sugar-stacking base W185 at site +1 (orange), active site (red), substrate binding groove walls (light blue), and helical ridge composed of residues 183 to 187 (dark blue). The protein is modeled in complex with substrate mimic 2,4-dinitrophenyl-2-deoxy-2-fluoro-β-D-cellobioside from the structure of the *Bacillus agaradhaerens* Cel5A (PDB 4A3H). Sugar superpositioning was achieved through aligning *Ba*Cel5A to *Hj*Cel5A in PyMOL. (D) The active site of *Hj*Cel5A depicting hydrogen bonding networks between the catalytic base (E148) and nucleophile (E259), as well as other conserved residues (gray).

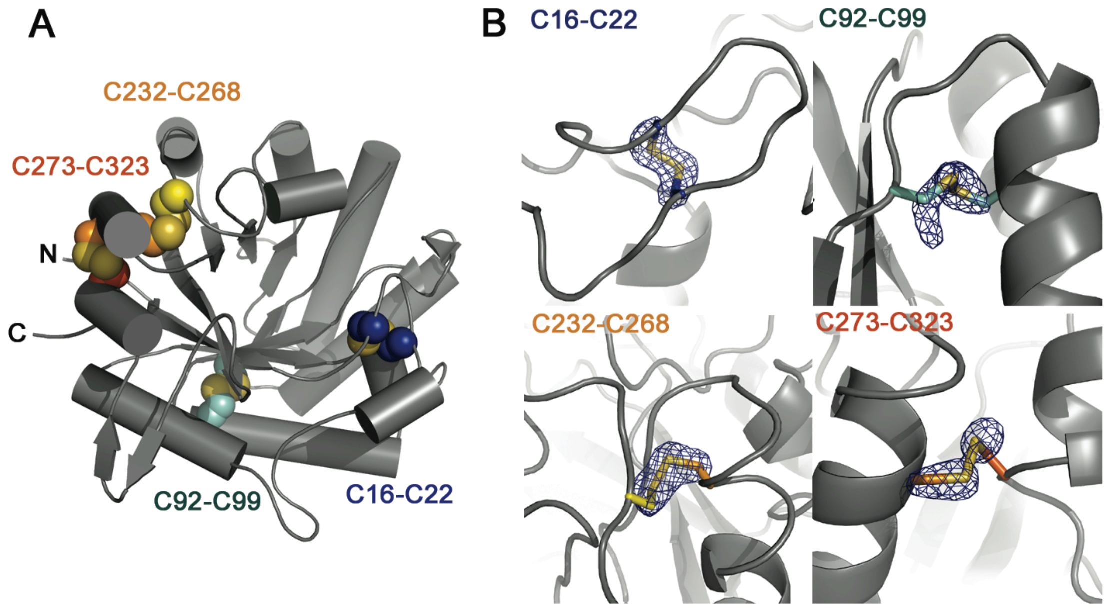


Figure 2. *Disulfide bonding patterns in HjCel5A*. (A) Cartoon representation of the protein highlighting positions of the four intramolecular disulfide bonds detected in the electron density. (B) F_o-F_c cysteine sidechain omit maps contoured to 5 σ. Sidechain atoms from the Cβ to the end of the sidechain were deleted from the model prior to map generation.

2.8 Supplementary Figures

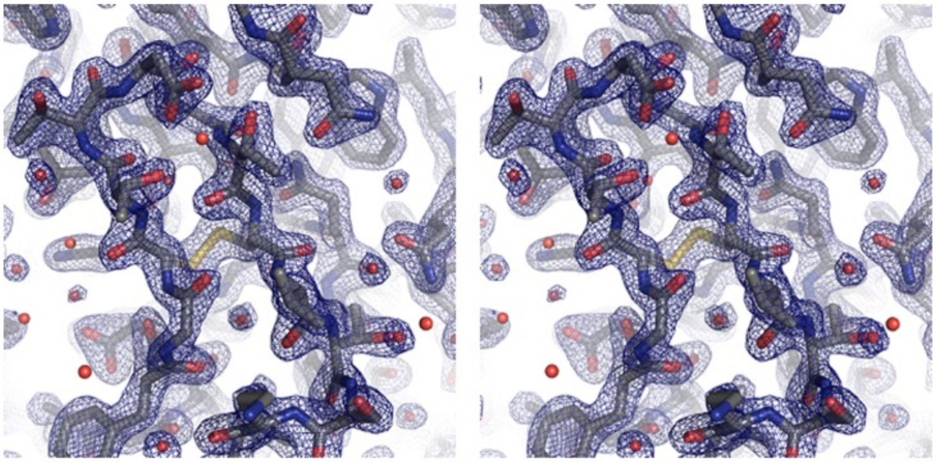
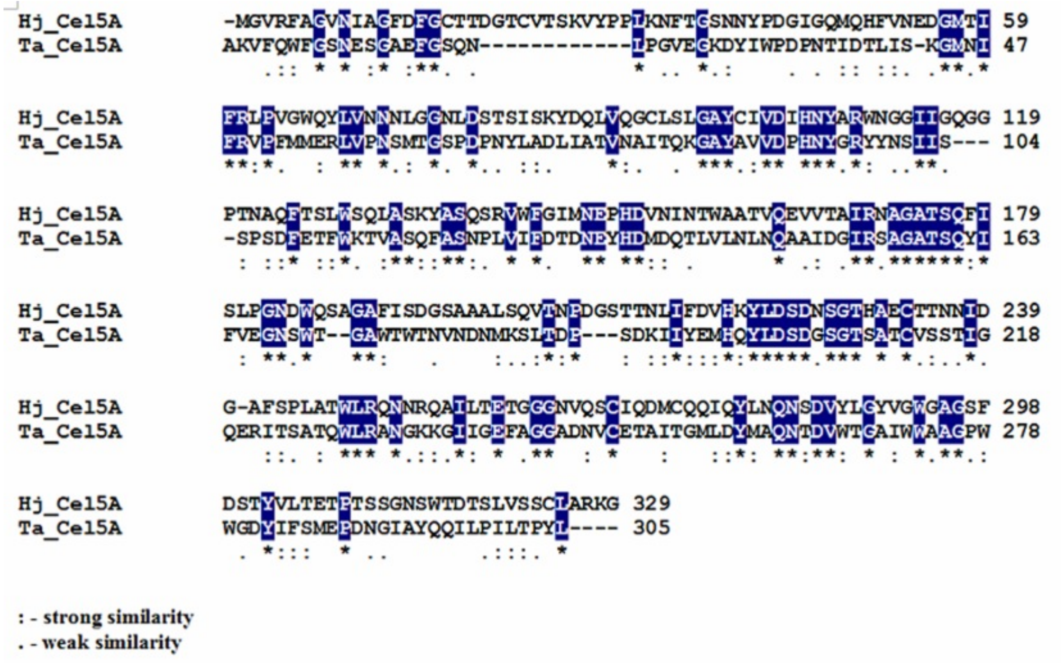


Figure S1. *HjCel5A electron density shown in wall-eyed stereo.* The $2F_o$-F_c electron density map contoured to 1.5 σ clearly shows well defined density for backbone and sidechain atoms for a loop spanning residues 12 to 23, a disulfide bond connecting C16 and C22, and the surrounding protein and solvent structure.



```
Hj_Cel5A        -MGVRFAGVNIAGFDFGCTTDGTCVTSKVYPPLKNFTGSNNYPDGIGQMQHFVNEDGMTI 59
Ta_Cel5A        AKVFQWFGSNESGAEFGSQN------------LPGVEGKDYIWPDPNTIDTLIS-KGMNI 47
                   .:: * * :* :**. .            * .. *.:    . . :: ::. .**.*

Hj_Cel5A        FRLPVGWQYLVNNNLGGNLDSTSISKYDQLVQGCLSLGAYCIVDIHNYARWNGGIIGQGG 119
Ta_Cel5A        FRVPFMMERLVPNSMTGSPDPNYLADLIATVNAITQKGAYAVVDPHNYGRYYNSIIS--- 104
                **:*.  : ** *.: *. *.. ::.    *:.  . ***.:** ***.*: ..**.

Hj_Cel5A        PTNAQFTSLWSQLASKYASQSRVWFGIMNEPHDVNINTWAATVQEVVTAIRNAGATSQFI 179
Ta_Cel5A        -SPSDFETFWKTVASQFASNPLVIFDTDNEYHDMDQTLVLNLNQAAIDGIRSAGATSQYI 163
                 : ::* ::*. :**::**:. * *.  ** **:: .      * .: .**.******:*

Hj_Cel5A        SLPGNDWQSAGAFISDGSAAALSQVTNPDGSTTNLIFDVHKYLDSDNSGTHAECTTNNID 239
Ta_Cel5A        FVEGNSWT--GAWTWTNVNDNMKSLTDP---SDKIIYEMHQYLDSDGSGTSATCVSSTIG 218
                 : **.*   **:    .   :..:*:*   : ::*:::*:*****.*** * *.:..*.

Hj_Cel5A        G-AFSPLATWLRQNNRQAILTETGGGNVQSCIQDMCQQIQYLNQNSDVYLGYVGWGAGSF 298
Ta_Cel5A        QERITSATQWLRANGKKGIIGEFAGGADNVCETAITGMLDYMAQNTDVWTGAIWWAAGPW 278
                   ::. : *** *.::.*: * .**  : *   :   ::*: **:**: * : *.**.:

Hj_Cel5A        DSTYVLTETPTSSGNSWTDTSLVSSCLARKG 329
Ta_Cel5A        WGDYIFSMEPDNGIAYQQILPILTPYL---- 305
                 . *:::  * ..       .:::. *
```

: - strong similarity
. - weak similarity

Figure S2. *Alignment of HjCel5A with the homologous sequence from TaCel5A.* The alignment was generated in CLUSTAL W [33] from the sequences of the crystallized proteins lacking expression and purification tags. Stars and blue highlighted regions indicate conserved regions. Strongly conserved regions are indicated with two marks and weakly conserved regions are indicated with a single mark.

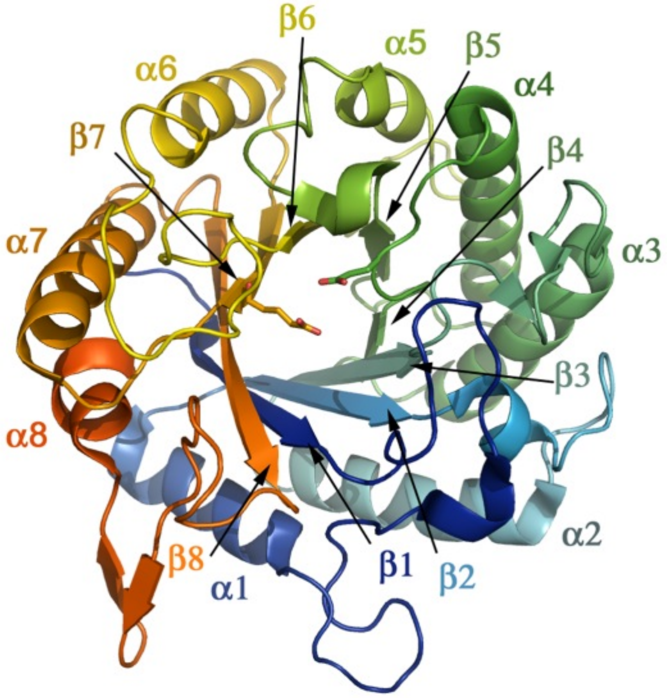


Figure S3. *HjCel5A secondary structure numbering.* A cartoon representation of the protein colored in chainbows by position along the main chain (N-terminus in blue, C-terminus in red). All α-helices and β-strands referred to in the main text are labeled with their corresponding abbreviations.

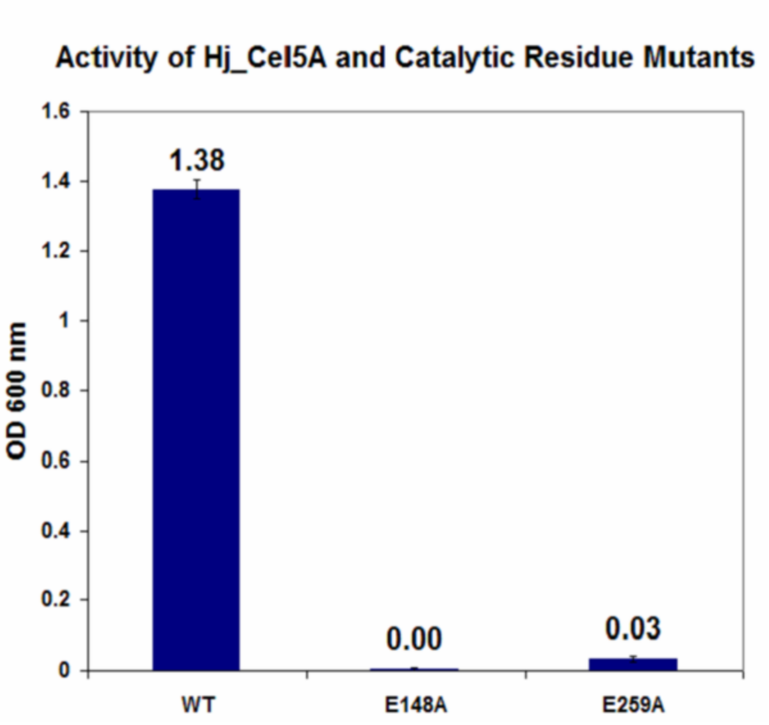


Figure S4. *Enzymatic activities of HjCel5A and catalytic residue mutants.* Activity data measured by OD_{600} are displayed for the wild-type protein and alanine mutations of the two catalytic glutamates, E148 and E259.

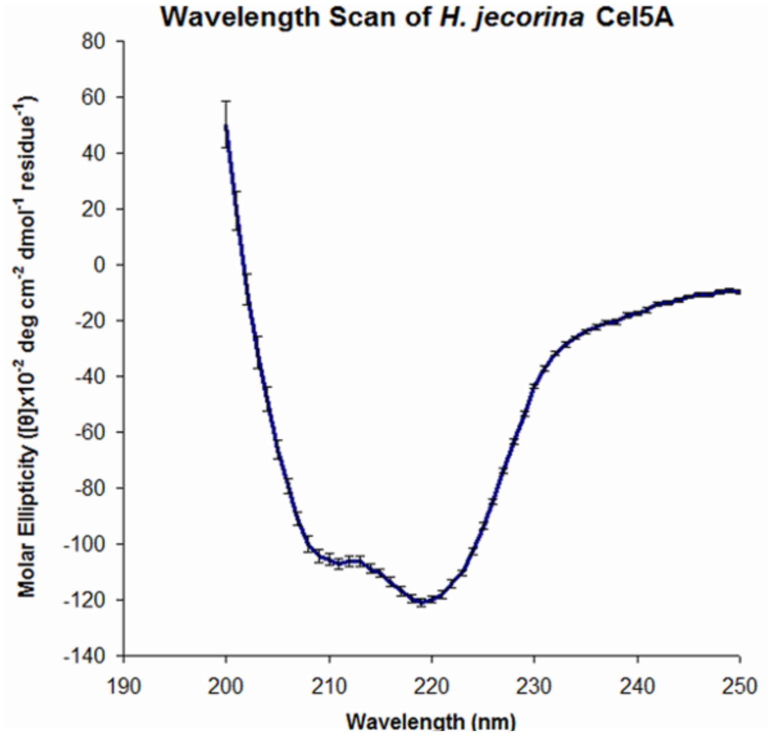


Figure S5. *Circular dichroism wavelength scan of HjCel5A*.

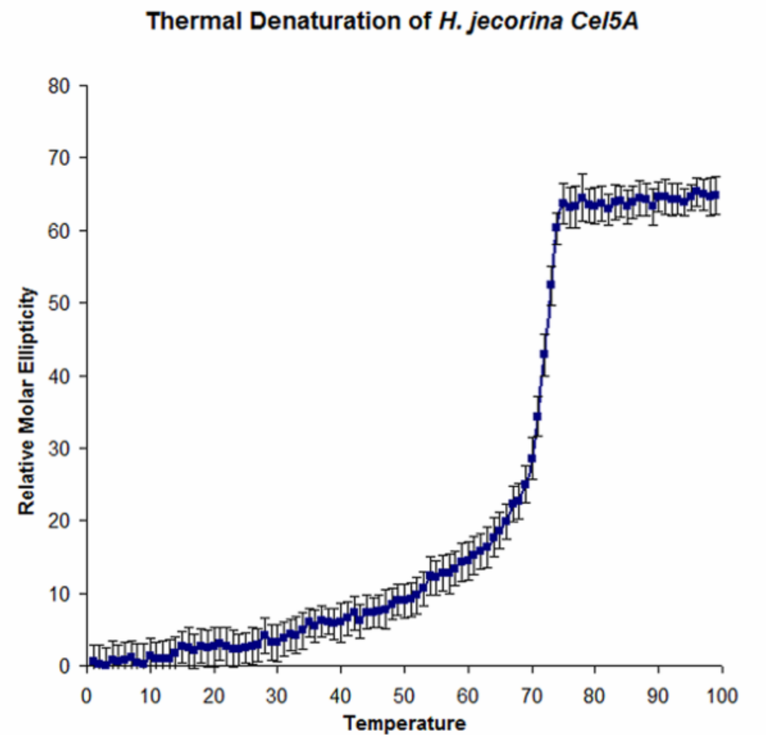


Figure S6. *Thermal denaturation scan of HjCel5A*. Thermal denaturation was monitored at 220 nm from 0 to 99 °C. Significant denaturation becomes detectable starting at approximately 65 °C. The apparent T_m is 71.5 °C.

2.9 References

1. Lee, T.M., Farrow M.F., Arnold F.H. and Mayo S.L. (2011) A structural study of *Hypocrea jecorina* Cel5A. Protein Science 20:1935-1940.
2. Kumar, R., Singh S. and Singh O. (2008) Bioconversion of lignocellulosic biomass: biochemical and molecular perspectives. Journal of Industrial Microbiology & Biotechnology 35:377-391.
3. Bisaria, V.S., Ghose, T.K. (1981) Biodegradation of cellulosic materials: substrate, microorganisms, enzymes and products. Enzyme and Microbial Technology 3:90-104.
4. Stephanopoulous, G. (2007) Challenges in engineering microbes for biofuels production. Science 315:801-804.
5. Rouvinen, J., Bergfors, T., Teeri, T., Knowles, J. K. C., Jones, T. A. (1990) Three-dimensional structure of cellobiohydrolase II from *Trichoderma reesei*. Science 249:380-386.
6. Divne, C., Stahlberg, J., Reinikainen, T., Ruohonen, L., Pettersson, G., Knowles, J.K., Teeri, TT., Jones, T.A. (1994) The three-dimensional crystal structure of the catalytic core of cellobiohydrolase I from *Trichoderma reesei*. Science 265:524-528.
7. Kleywegt, G.J., Zou J.-Y., Divne C., Davies G.J., Sinning I., Ståhlberg J., Reinikainen T., Srisodsuk M., Teeri T.T. and Jones T.A. (1997) The crystal structure of the catalytic core domain of endoglucanase I from *Trichoderma reesei* at 3.6 Å resolution, and a comparison with related enzymes. Journal of Molecular Biology 272:383-397.
8. Sandgren, M., Ståhlberg J. and Mitchinson C. (2005) Structural and biochemical studies of GH family 12 cellulases: improved thermal stability, and ligand complexes. Progress in Biophysics and Molecular biology 89:246-291.
9. Suominen, P.L., Mäntylä A.L., Karhunen T., Hakola S. and Nevalainen H. (1993) High frequency one-step gene replacement in *Trichoderma reesei*. II. Effects of deletions of individual cellulase genes. Molecular and General Genetics MGG 241:523-530.
10. Teeri, T.T., Lehtovaara P., Kauppinen S., Salovuori I. and Knowles J. (1987) Homologous domains in *Trichoderma reesei* cellulolytic enzymes: Gene sequence and expression of cellobiohydrolase II. Gene 51:43-52.
11. Kraulis, P.J., Clore G.M., Nilges M., Jones T.A., Pettersson G., Knowles J. and Gronenborn A.M. (1989) Determination of the three-dimensional solution structure of the C-terminal domain of cellobiohydrolase I from *Trichoderma reesei*. A study using nuclear magnetic resonance and hybrid distance geometry-dynamical simulated annealing. Biochemistry 28:7241-7257.
12. Van Petegem, F., Vandenberghe I., Bhat M.K. and Van Beeumen J. (2002) Atomic resolution structure of the major endoglucanase from *Thermoascus aurantiacus*. Biochemical and Biophysical Research Communications 296:161-166.
13. Pereira, J.H., Chen Z., McAndrew R.P., Sapra R., Chhabra S.R., Sale K.L., Simmons B.A. and Adams P.D. (2010) Biochemical characterization and crystal

structure of endoglucanase Cel5A from the hyperthermophilic *Thermotoga maritima*. Journal of Structural Biology 172:372-379.

14. Sakon, J., Adney W.S., Himmel M.E., Thomas S.R. and Karplus P.A. (1996) Crystal structure of thermostable family 5 endocellulase E1 from *Acidothermus cellulolyticus* in complex with cellotetraose. Biochemistry 35:10648-10660.
15. Hui, J.P.M., White T.C. and Thibault P. (2002) Identification of glycan structure and glycosylation sites in cellobiohydrolase II and endoglucanases I and II from *Trichoderma reesei*. Glycobiology 12:837-849.
16. Davies, G.J., Mackenzie L., Varrot A., Dauter M., Brzozowski A.M., Schülein M. and Withers S.G. (1998) Snapshots along an enzymatic reaction coordinate: Analysis of a retaining β-glycoside hydrolase. Biochemistry 37:11707-11713.
17. Medve, J., Karlsson J., Lee D. and Tjerneld F. (1998) Hydrolysis of microcrystalline cellulose by cellobiohydrolase I and endoglucanase II from *Trichoderma reesei*: Adsorption, sugar production pattern, and synergism of the enzymes. Biotechnology and Bioengineering 59:621-634.
18. Wang, Q., Graham R.W., Trimbur D., Warren R.A.J. and Withers S.G. (1994) Changing enzymic reaction mechanisms by mutagenesis: Conversion of a retaining glucosidase to an inverting enzyme. Journal of the American Chemical Society 116:11594-11595.
19. Shaw, A., Bott R., Vonrhein C., Bricogne G., Power S. and Day A.G. (2002) A novel combination of two classic catalytic schemes. Journal of Molecular Biology 320:303-309.
20. Parry, N.J., Beever, D.E., Owen, E., Vandenberghe, I., Van Beeumen, J., Bhat, M. (2001) Biochemical characterization and mechanism of action of a thermostable Beta-glucosidase purified from *Thermoascus aurantiacus*. Biochemical Journal 353:117-127.
21. Rosgaard, L., Pedersen S., Langston J., Akerhielm D., Cherry J.R. and Meyer A.S. (2007) Evaluation of minimal *Trichoderma reesei* cellulase mixtures on differently pretreated barley straw substrates. Biotechnology Progress 23:1270-1276.
22. Nakazawa, H., Okada, K., Kobayashi, R., Kubota, T., Onodera, T., Ochiai, N., Omata, N., Ogasawara, W., Okada, H., Morikawa, Y. (2008) Characterization of the catalytic domains of *Trichoderma reesei* endoglucanase I, II, and III expressed in *Escherichia coli*. Applied Microbiology and Biotechnology 81:681-689.
23. Qin, Y., Wei X., Song X. and Qu Y. (2008) Engineering endoglucanase II from *Trichoderma reesei* to improve the catalytic efficiency at a higher pH optimum. Journal of Biotechnology 135:190-195.
24. Canutescu, A.A., Shelenkov A.A. and Dunbrack R.L. (2003) A graph-theory algorithm for rapid protein side-chain prediction. Protein Science 12:2001-2014.
25. Terwilliger, T.C., Grosse-Kunstleve R.W., Afonine P.V., Moriarty N.W., Zwart P.H., Hung L.-W., Read R.J. and Adams P.D. (2008) Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. Acta Crystallographica Section D 64:61-69.
26. Emsley, P. and Cowtan K. (2004) Coot: model-building tools for molecular graphics. Acta Crystallographica Section D 60:2126-2132.

27. Adams, P.D., Afonine P.V., Bunkoczi G., Chen V.B., Davis I.W., Echols N., Headd J.J., Hung L.-W., Kapral G.J., Grosse-Kunstleve R.W., McCoy A.J., Moriarty N.W., Oeffner R., Read R.J., Richardson D.C., Richardson J.S., Terwilliger T.C. and Zwart P.H. (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallographica Section D 66:213-221.

28. Rychlewski, L., Jaroszewski L., Li W. and Godzik A. (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Science 9:232-241.

29. McCoy, A. (2007) Solving structures of protein complexes by molecular replacement with Phaser. Acta Crystallographica Section D 63:32-41.

30. Cohen, G. (1997) ALIGN: a program to superimpose protein coordinates, accounting for insertions and deletions. Journal of Applied Crystallography 30:1160-1161.

31. The PyMOL Molecular Graphics System, V.S., LLC.

32. Park, J.T. and J J.M. (1949) A submicrodetermination of glucose. Journal of Biological Chemistry 181:149-151.

33. Thompson, J.D., Higgins, D.G., Gibson, T. J. (1994) Nucleic Acids Research 22:4673-4680.

CHAPTER 3

Identifying Stabilizing Mutations in *Hypocrea jecorina* Cel5A through Examining Residue Conservation and Covariance

This chapter is formatted for submission to the Journal of Molecular Biology.

3.1 Abstract

Consensus design is a canonical method of enhancing protein thermostability, but its efficacy may depend on the quality and quantity of available sequence data. We sought to uncover stabilizing consensus mutations in the primary endoglucanase Cel5A from *Hypocrea jecorina* (*Hj*Cel5A), a molecule with ~400 homologous sequences in the NCBI non-redundant protein database. Using this data, we constructed six multiple sequence alignments (MSAs) varying in the number, level of characterization, and percent identity to the query of the aligned sequences. The alignments were filtered with numerical thresholds to reveal highly conserved residues (high relative entropy) at positions able to mutate independently of other protein sites (low mutual information). Using this method, we identified five stabilizing point mutations, D13E (+3.0 °C), E53D (+2.7 °C), T57N (+1.1 °C), G189A (+0.4 °C), and G293A (+3.6 °C). Catalytic activity is either enhanced or maintained, suggesting that conserved stabilizing residues may be less deleterious to activity than stabilizing mutations identified through other means. Each alignment predicted a different subset of these stabilizing mutations. Thus, employing several alignments in the initial calculation may constitute a useful strategy for future engineering efforts. We also discuss alternative strategies for selecting residues based on conservation and covariance that may improve the methods showcased here.

3.2 Introduction

Developing reliable protein thermostabilization techniques constitutes a longstanding goal of the scientific community. Benefits of heat tolerance include extended enzyme lifetime, improved reaction kinetics, and reduced protein loading [1]. Many proteins with scientific [2], industrial [3-5], or pharmaceutical [6, 7] potential, however, denature at temperatures below that required for their desired application. With the goal of rendering protein products more useful, much attention has focused not only on developing new stabilization strategies, but also improving upon existing methods.

Consensus design is a commonly employed means of detecting stabilizing protein mutations. This semi-rational method entails assembling homologous sequences into a multiple sequence alignment (MSA) and mutating a protein of interest to the most prevalent amino acid at each position [8]. Thus far, a variety of proteins including immunoglobulin domains [9], tetratricopeptide repeats [10], an SH3 domain of a tyrosine kinase binding domain [11], GroEL minichaperones, a glucose dehydrogenase [5], p53 [12], a WW domain [13], and many others have achieved enhanced stability using variations of this strategy. Moreover, consensus design has successfully contributed to stabilizing targets with real world applications; the technique has already aided the stabilization of two cellulases, enzymes employed in the biofuels industry [3, 14].

Despite its widespread use, consensus design bears limitations. Generally, only about half of the mutations predicted from alignments are stabilizing [5, 8, 15], with many of the remaining half requiring compensatory modifications to maintain protein stability and/or function. In 2012, Sullivan et al. applied this concept to the consensus stabilization of triosphosphate isomerase [15]. Through pursuing conserved residues at positions able to mutate independently of other sites, the authors dramatically improved their predictive accuracy, successfully forecasting nine out of ten mutations as stabilizing. This task was accomplished using the information theoretic estimates relative entropy (RE) and mutual information (MI) to assess conservation and covariance, respectively, between positions in a protein sequence alignment. As this study was performed on a triose phosphate isomerase (TIM), a highly-characterized model protein system, it remains unclear whether the method is effective on an enzyme with real-world applications and non-ideal parameters.

We examined whether applying this strategy to *Hj*Cel5A, a key cellulase from the prodigious cellulase producer *Hypocrea jecorina* (anamorph *Trichoderma reesei*) [16], would yield stabilizing mutations with high accuracy. Along with related cellulases [1, 3, 4, 17], *Hj*Cel5A is a biofuels industry target for thermostabilization. In order to synergistically degrade crystalline cellulose into sugars suitable for fermentation into liquid fuels, three classes of cellulases are necessary: 1) exoglucanases such as *Hj*Cel7A and *Hj*Cel6A that cleave two glucose unit sugars called cellobiose from the end of cellulose strands, 2) endoglucanases such as *Hj*Cel5A that cleave in the middle of cellulose strands at amorphous sites in the crystalline lattice, and 3) β-glucosidases such as Cel3A that cleave cellobiose into glucose monomers [18]. Recent efforts have yielded thermostable variants of HjCel7A and HjCel6A capable of functioning at 70 °C with activity that exceeds wild type (WT) [3, 4]. In addition, previous work has demonstrated that the Cel3A from *Talaromyces emersonii* can be expressed in *H. jecorina* and exhibits an optimal temperature of 71.5 °C. The WT *Hj*Cel5A holoenzyme, however, functions optimally at 60 °C and thermally denatures with half of the enzyme remaining folded (T_m) at 69.5 °C as measured through circular dichroism [19]. Thus, the need for highly-active, thermostable *Hj*Cel5A variants is clear.

Here we demonstrate that applying conservation (RE) and correlation (MI) filters to several alignments with a wide range of properties can successfully predict stabilizing mutations in Cel5A. We report five stabilizing mutations that either preserve or enhance activity in the target protein. In addition, we discuss variations on the method and identify parameters that may improve prediction accuracy for future experiments.

3.3 Results and Discussion

Multiple Sequence Alignment Construction

Assembling sequences into an alignment requires numerous subjective decisions. Variables include the number of sequences incorporated, the acceptable percent identity of chosen sequences to the query, and the treatment of truncated sequences that only align with a portion of the target sequence. As MSA content can dramatically alter predictions, these variable factors should be considered during alignment construction.

Performing consensus design on *Hj*Cel5A faces three hurdles. First, the 444 homologous sequences with a percent identity between 30-90% to *Hj*Cel5A is small compared with more characterized proteins like TIM with homologous sequences numbering in the thousands. Examining background noise across MSAs of variable sizes suggests that a minimum of approximately 200 and 125 sequences is necessary to produce consistent RE and MI values, respectively [20, 21]. Several studies, however, have demonstrated that consensus design alone can identify stabilizing mutations from a handful of sequences [2, 3, 8, 14]. As such, it is unclear whether the available *Hj*Cel5A sequence data is sufficient for consensus/covariance design. Second, larger *Hj*Cel5A alignments contain numerous gapped regions. The protein of interest contains two domains: 1) a thermostable cellulose binding module (CBM) that adheres to the substrate [22-24] and 2) a catalytic $(\alpha/\beta)_8$ TIM-barrel [19]. Variable placement of these domains, as well as large non-conserved loop regions, can produce gapped regions that may either shift alignments out of register or reduce the amount of sequence data available at the gapped site. Both possibilities may potentially skew RE and MI calculations, reducing predictive accuracy. Finally, many of the retrieved sequences originate from uncharacterized proteins with little homology to *Hj*Cel5A. With a low average percent identity of ~40% to the query across the 444 available sequences, alignments may be phylogenetically biased to predict mutations that are well-suited for distantly-related homologs, but incompatible with *Hj*Cel5A. Consensus design relies on the assumption that the frequency of a residue correlates with its contribution to protein stability. Alignments heavily biased by phylogenetic relationships disrupt this correlation leading to inaccurate predictions.

We constructed six *Hj*Cel5A alignments differing in the number and characterization level of incorporated sequences to capture the tradeoff between the number and the quality of aligned sequences. The largest MSA contains all 444 sequences retrieved through PSI-BLAST with 30-90% identity to the catalytic domain of *Hj*Cel5A. Sequences with less than 30% identity to the query are difficult to align with sufficient accuracy and were eliminated [25, 26]. This ensemble contains many sequences that poorly align with the query and was culled to remove putative cellulases/endoglucanases, precursors, and heavily gapped sequences yielding a 29-member alignment. Three additional MSAs containing 323, 233, and 195 varying in sequence identity and characterization of incorporated sequences were constructed to provide further MSA diversity. The smallest alignment contains 10 sequences either experimentally confirmed or reasonably expected to function as endoglucanases. Features of each MSA are summarized in Table I.

Evaluating Conservation and Covariance

Measures for conservation and covariance were applied to each position in the alignments as described in Sullivan et al. [15]. To probe for conservation, relative entropy (RE) was calculated using Eq. 1,

$$RE = \sum p_x \ln \frac{p_x}{f_x} \qquad (1)$$

where p_x is the frequency of residue x appearing at a particular position and f_x is the frequency of residue x based on codon usage. In broad terms, the relative entropy is a value that measures how much the frequency of an observed occurrence diverges from the frequency expected if derived randomly from a neutral reference state [21]. To assess covariance, mutual information (MI), the relative entropy between the joint frequency of observing particular residues at two positions in a sequence and the expected frequency based on the separate probability of finding each residue at their respective sites, was calculated using Eq. 2 [27],

$$MI(i,j) = \sum_i \sum_j p_{x,y} \ln \frac{p_{x,y}}{p_x p_y} \qquad (2)$$

where p_x is the frequency of sidechain x appearing at position i, p_y is the frequency of sidechain y at position j, and $p_x p_y$ is the frequency of sidechain x and y appearing at i and j simultaneously. Thresholds for acceptable mutations were based on those used by Sullivan et al. [15]. Highly conserved residues (RE > 1.42) at uncoupled sites (maximum MI ≤ 0.5) were chosen for further scrutiny. Mutations with an RE greater than 3 were also discarded. These relatively invariant sites may superficially exhibit low covariance, but may still require compensatory mutations to preserve protein folding and function.

Experimental Screening and Validation of Thermostability Enhancing Mutations

After applying the RE and MI constraints, a total of 21 unique mutations were predicted as stabilizing from the six alignments (Table II). Point mutants were constructed, proteins were secreted from *Saccharomyces cerevisiae*, and supernatants were screened for activity on Avicel, a microcrystalline cellulose powder, at a temperature two degrees higher (73 °C) than the WT T_m (Figure 1A). As a result of screening supernatant, high signal can indicate greater thermostability, activity, and/or expression, all desirable traits. Mutations D13E, E53D, T57N, I82L, V101L, G189A, and G293A demonstrated greater activity than WT and were selected for further characterization.

The seven candidate *Hj*Cel5A point mutants were purified and pre-incubated at a gradient of temperatures for 10 min before adding Avicel to assess activity over 2 hours at 60 °C. Five mutations exhibited a T_{50}, the temperature at which half of the total enzyme remains active, greater than WT ($\Delta T_{50,\text{ D13E}} = 3.0$, $\Delta T_{50,\text{ E53D}} = 2.7$, $\Delta T_{50,\text{ T57N}} = 1.1$, $\Delta T_{50,\text{ G189A}} = 0.4$, and $\Delta T_{50,\text{ G239A}} = 3.6$ °C) (Table III, Figure 1B, Figure S1). The two remaining mutations exhibited slightly lower stabilities than WT ($\Delta T_{50,\text{ I82L}} = -0.3$, $\Delta T_{50,\text{ V101L}} = -0.4$) and likely exhibit high activity on the screen due to increases in expression level (Table III).

Activity of Stabilizing Mutations

To ensure that the stabilizing mutations do not adversely affect enzymatic activity, the five point mutants were tested for hydrolysis on Avicel after 2 hours at 60 °C (Figure 1C). The T57N and G293A mutants show significantly elevated activity, while the activities of the remaining mutants are comparable to WT (Table III).

Structural Analysis of Stabilizing Mutations

The five stabilizing mutations are dispersed throughout the protein with two, one, and two in the core, surface, and boundary, respectively. Clues outlining the mechanisms through which the five consensus mutations stabilize *Hj*Cel5A emerge upon examining homologous structures (Table IV). In the following discussion, all residue numbering refers to that employed in the HjCel5A crystal structure (PDB ID 3QR3 [19]).

D13E: The equivalent to D13E appears in the *Thermoascus aurantiacus* (*Ta*Cel5A) Cel5A (PDB ID 1GZJ [28]) and Rbcel1 (PDB ID 4EE9 [29]) structures (Figure 2A). In *Ta*Cel5A, the glutamate at position 13 maintains a preexisting hydrogen bond to the backbone nitrogen of G11, but additionally forms a hydrogen bond to the sidechain of a threonine at position 10. In *Hj*Cel5a and Rbcel1, alanine occupies position 10. The Rbcel1 structure, however, still contains E13, demonstrating that its stabilizing effect might arise simply through adding a carbon to the protein interior and improving core packing. Although efforts to mutate position 10 to serine resulted in highly reduced activity on the initial screen (Figure 1A), the site may be more accommodating to threonine.

E53D: The region around position 53 differs dramatically in many *Hj*Cel5A homologs. In both the *Bacillus agaradhaerens* (*Ba*Cel5A) (PDB ID 7A3H [30]) and *Bacillus subtilis* (PDB ID 3PZT [31]) endoglucanase structures, an aspartate at this position forms a salt bridge with an arginine on a neighboring helix (Figure 2B). *Hj*Cel5A, however, does not contain a suitable electrostatic contact partner for an aspartate in this region. It is possible that introducing the E53D mutation may elicit a rearrangement, bringing the sidechain of K327 close enough to fill this role. Alternatively, the E53D mutation may function to increase the distance between the carboxylate sidechain and neighboring residues D54 and D316, reducing electrostatic repulsion.

T57N: In *Ta*Cel5a (PDB ID 1GZJ [28]) and *Ba*Cel5A (PDB ID 7A3H [30]), the N57 sidechain forms hydrogen bonds to the backbone oxygen of position 3 and the backbone nitrogen of position 5 on an adjacent β-strand (Figure 2C). The terminal oxygen of T57 in *Hj*Cel5A falls slightly short of making either of these contacts. Introducing the T57N mutation likely leads to greater stability in *Hj*Cel5A due to the addition of these two hydrogen bonds.

G189A: In both *Hj*Cel5A and Rbcel1 (PDB ID 4EE9 [29]), the G189A mutation and its equivalent position appear at a short, solvent-exposed loop (Figure 2D). As such, the G189A mutation likely enhances stability through reducing backbone entropy.

G293A: The G293A mutation lies behind W292, a residue involved in substrate binding [19] (Figure 2E). A293 appears in the *Ta*Cel5A structure, but the orientation of W292 remains identical to that seen in the *Hj*Cel5A structure. The G293A mutation most likely allows activity to persist at high temperatures through forcing W292 to adopt a catalytically-relevant conformation.

Prediction Accuracy

When examined individually, each MSA successfully predicts at least one stabilizing mutation from eleven or fewer candidates (Table II). Out of seven mutations predicted from the 444-member MSA, two mutations were stabilizing (29% accuracy). The 323-, 233-, and 195-member MSAs predicted eleven candidates each, with three stabilizing mutations predicted from the 323- and 233-sequence MSAs (27%) and two mutations predicted from the 195 member MSA (18%). Only one mutation out of four predicted from the 29-sequence MSA was ultimately stabilizing (25%). Finally, both candidates retrieved from the 10-member MSA provided benefits in stability (100% accuracy). In total, only five mutations out of 21 tested were found to be stabilizing (24%), a lower accuracy than that achieved by Sullivan et al. These results, however, still demonstrate a dramatic improvement over random mutagenesis, which typically yields one stabilizing mutation out of every 10^3-10^4 constructs [8]. Moreover, the activity screen was performed on unpurified protein secreted into supernatant and fails to detect poorly expressing or inactive thermostable proteins. Direct assessment of thermostability may improve the accuracy of this method, a promising prospect for proteins lacking properties amenable to quick screening.

Effects of Multiple Sequence Alignment Size

MSA composition dramatically impacts the list of predicted mutations. As shown in Table I, the average percent identity to the query among incorporated sequences drops to 34.5-39.5% among the larger alignments. The genetic diversity of these larger alignments reduces the amount of false consensus predictions that arise simply through shared evolutionary history (phylogenetic bias). We might surmise that the larger alignments would prove more effective at identifying stabilizing mutations. This presupposition is supported by evidence from the work of Jäckel et al. [32], which demonstrates that improvements to thermostability increase as phylogenetic bias in starting sequences decreases. However, unique stabilizing mutations appear as the number of incorporated sequences shrinks and the average percent identity of incorporated sequences to the query increases. Mutations E53D and G189A only appear when examining the 10-member endoglucanase alignment. One explanation for why certain stabilizing mutations might only appear in prediction lists generated from small MSAs with closely-related sequences is that these alignments may include useful data for hypervariable regions that would otherwise not appear in alignments of less related homologs. The two stabilizing mutations predicted from the 10-member alignment, however, appear at sites with relatively conserved secondary structure within the protein scaffold. It is possible that these mutations, while stabilizing, are not indicative of any trend and that the small number of sequences used in the alignment may lead to a high level of noise in conservation predictions.

Although consensus design seeks to minimize phylogenetic information, Bloom and Glassman have demonstrated that examining evolutionary history can improve protein stabilization efforts [33]. Forming predictions using phylogenies instead of sequence alignments includes a layer of information absent in consensus design. While consensus design relies on information of final evolved sequences, design by phylogeny also includes information concerning selective pressures embodied in substitution probabilities. Future studies may benefit from incorporating information from both sequence alignments and phylogenies to inform predictions.

Predictive Efficacy of Relative Entropy

To further investigate the efficiency of using RE to predict desirable stabilizing mutations, we constructed receiver operator characteristic (ROC) curves. ROC curves are routinely used in psychology, medicine, and increasingly data mining to illustrate the performance of a binary classifier system as its discrimination threshold varies. Curves are generated through plotting the fraction of true positives from the predicted positives versus the fraction of false positives from the true negatives over a range of binary threshold settings [34]. To determine whether applying a threshold for a particular value improves a prediction, a metric called the area under the curve (AUC) is calculated from the area between the curve and the diagonal (maximum AUC = 0.5). Any metric capable of discriminating between desired and unwanted members of a set with some level of accuracy will display an AUC greater than zero, allowing one to easily determine the efficacy of a particular forecasting method.

Using a collection of 262 unique *Hj*Cel5A point mutations derived from various experiments (See Chapters 4-6), we calculated RE and MI for each mutation and pair of positions in the protein and used the information to construct ROC curves. The dataset flags mutations as true positives if they exhibit enhanced stability with adequate activity and expression as determined through the screening and testing methods described in this study. The ROC curves demonstrate that RE thresholds are capable of predicting stable, active mutations (Table V, Figure 3A). We do not expect MI to predict stabilizing mutations as it merely filters datasets to remove highly covering residues. This filtering ability may be evident for larger MSAs in that their ROC curves demonstrate a positive AUC only over less stringent thresholds (Table V, Figure 3B).

Further ROC curve analysis indicates that applying an MI threshold to a dataset improves prediction accuracy (Figure 3D and F, Table V). Over all MSA sizes, RE ROC curves computed from a dataset purged of mutations with an MI greater than 0.5 demonstrate an average increase in AUC of 0.18 units when compared to a similar curve lacking the MI constraint. Adding the MI constraint dramatically reduces the number of predicted mutations. While eliminating mutations using the RE threshold employed in this study yields six to eighteen mutations per alignment, removing mutations that did not meet both the RE and MI thresholds used in this study resulted in only two to eleven predicted mutations per MSA. While the smaller group contains a higher fraction of desirable mutations, only two or three stabilizing mutations appear per MSA. Slightly relaxing the covariance constraints to obtain a candidate pool size suitable for the available screening or testing method may increase the number of candidate mutations.

Alternative Methods of Identifying Consensus Mutations

Multiple means of assessing residue conservation exist. We have recapitulated the study performed by Sullivan *et al.* on *Hj*Cel5A using relative entropy as a metric for conservation. A recent *Hj*Cel7A engineering effort by Komor et al., evaluates conservation through assuming that the most probable distribution of amino acids can be modeled with Boltzmann's law [3]. In this classic approach developed by Steipe et al.[8], the statistical free energy is derived from mutational frequencies:

$$\Delta\Delta G = -RT \ln \frac{p_x}{f_x} \tag{3}$$

where p_x is the frequency of the mutation and f_x is the frequency of the original amino acid at a particular position. We computed ΔΔG values for the 262 *Hj*Cel5A point mutation dataset and evaluated the predictive properties of this metric with ROC curves (Figure 3C). In generating these curves, we accepted residues with ΔΔG values lower than the selected threshold. ΔΔG is highly predictive, generating an average AUC of 0.23 versus 0.18 achieved with RE (Table V). Moreover, computing the ROC curve using the culled dataset (MI ≤ 0.5) increases predictive accuracy (Figure 3E and F). Under these conditions, the ΔΔG values yield an average AUC of 0.37 versus 0.26 for RE. Thus, using either ΔΔG values or RE as a conservation constraint can be used as measure of conservation in the thermostabilization strategy pioneered by Sullivan et al. [15]

We surmise that ΔΔG thresholds are slightly more predictive than RE in part because the protein of interest originates from a filamentous fungus rather than a model organism. RE uses codon frequency as a reference state. We employed the *S. cerevisiae* codon table due to the small number of observations used to create a *H. jecorina* table [35]. While our results might improve through using the *H. jecorina* data, efforts to stabilize proteins from organisms without adequate sequence data may suffer from similar problems.

Optimal Consensus/Correlation Thresholds

Using the 262-mutation dataset, we determined optimal RE, MI, and ΔΔG cutoffs for the conditions presented in this study (Table V). After excluding noisy data from the 10 and 29-sequence MSAs, the average optimal cutoffs, values that maximize the number of predicted true positives while minimizing false positives, for RE and ΔΔG in isolation are 0.01 and 0.8 kcal mol^{-1}, respectively. For this dataset, the average RE is ~0.2 and the average ΔΔG is ~3.5 kcal mol^{-1}. Optimal RE and ΔΔG thresholds shift to more stringent cutoffs when the curves are computed on datasets filtered by MI. Additionally, the optimal RE and ΔΔG thresholds appear to vary with MSA size upon MI prefiltering. For this system, the optimal MI cutoff yielding the largest AUC for RE and ΔΔG curves generated from any size MSA is 0.3-0.5 (Table V, Supplementary Table I), although a cutoff anywhere between 0.3 and 0.7 improves accuracy (Figure 4 A and B). As previously discussed, a tradeoff exists between accuracy and the number of mutations predicted. Although applying the MI constraint reduces the number of false positives in the predicted set to a considerable degree, the total quantity of stabilizing mutations present in this pool is low. After applying the MI ≤ 0.5 constraint to the 444-sequence MSA, only 4 stabilizing mutations remained in the pool of candidates. Thus, the ideal thresholds will vary depending on the desired number of stabilizing mutations. In addition, future experiments on systems beyond *Hj*Cel5A are necessary to determine if these values are universal or vary when applied to different proteins.

3.4 Summary and Conclusions

We have provided further evidence that filtering alignments for consensus mutations at non-covarying sites can rapidly identify stabilizing mutations in a protein of industrial significance. Complete site-saturation mutagenesis of *Hj*Cel5A would require constructing and screening 6232 mutants, an intractable task without robotic assistance. Upon application of both RE and MI constraints, however, only 21 unique mutations were predicted across all six examined MSAs with five experimentally verified as improving stability while maintaining activity. Additionally, sequence data for *Hj*Cel5A, while abundant, demonstrates a low average identity to the protein of interest and contains large gapped regions that frustrate alignment attempts. These results reinforce the robustness of the technique beyond ideal conditions.

In addition to revealing five stabilizing mutations, we have also determined optimal parameters for several variables inherent in the process. We demonstrate that:

1) $\Delta\Delta G$ values can be substituted for RE to assess conservation in the method pioneered by Sullivan et al.

2) The highest accuracy is achieved using an MI threshold of 0.4 in combination with filters for conservation. To increase the number of discovered mutations, this value can be relaxed to about 0.7 or 0.9 and without dramatically compromising effectiveness.

This study seeks to answer some of the questions Sullivan et al. could not address due to a limited dataset, namely ideal limits for conservation and correlation and whether varying degrees of taxonomic bias in the MSA can change the list of predicted mutations. Although these results are valid for the protein and methods used in this study, additional tests are necessary to determine whether these trends hold beyond the *Hj*Cel5A system.

3.5 Materials and Methods

MSA Construction and Analysis

Sequences homologous to the catalytic domain of *Hj*Cel5A (from GVR to CLARKG) were retrieved using the Position-Specific Iterated BLAST (PSI-BLAST) [36] database search applied to the non-redundant protein sequences National Center for Biotechnology Information (NCBI) database. Constraints on the percent identity of the sequences to the query were introduced using the formatting options feature within the BLAST tool. Relative entropy was calculated using the yeast codon probabilities from Sullivan et al. [15]. Other considerations necessary to determine RE, MI, and ΔΔG are described in the results section.

To determine the background level of noise in MI calculations, residues in were scrambled within alignment columns to eliminate true covariance. MI values were then recalculated to determine the amount of covariance at each site attributable to random chance. The number of observations exceeding the noise threshold at each site appears in Table II and in the supplementary files associated with this thesis.

Cel5A Plasmid Construction

The Cel5A gene was synthesized by DNA 2.0 (Menlo Park, CA, USA) with codon frequency optimized for *S. cerevisiae*. The construct consists of an αMFpp8 secretory leader sequence (GenBank BK006949 193648-194145) followed by a region coding for the CBM from the *H. jecorina* CBM (GenBank ABA64553.1) preceded by an extra 'AR' introduced during cloning. This DNA sequence is:

5'-
GGCTAGACAACAAACAGTATGGGGTCAATGTGGTGGTATTGGATGGTCTGGT
CCGACAAACTGTGCTCCAGGCTCGGCATGTTCGACACTAAATCCATATTACG
CTCAATGTATCCCTGGCGCTACCACTATAACAACTTCTACTAGACCACCTTCT
GGTCCGACGACAACTACAAGGGCTACCTCAACCTCTTCCTCTACACCCCCTAC
TTCCAGC – 3'

The additional ‘AR’ sequence does not significantly affect any protein properties. The CBM region is then followed by an *Hj*Cel5A catalytic domain sequence identical to GenBank entry JN172972.1. This construct contains a short linker and an N-terminal His-tag. The assembled gene was cloned into the yeast expression vector YEp352/PGK91-1-αss between the BglII and MboI restriction sites using the Gibson assembly method [37]. Point mutations were introduced using the QuikChange Lightning Site-Directed Mutagenesis Kit from Agilent Technologies (Santa Clara, CA, USA) using primers designed with the online tool provided by Agilent:

www.genomics.agilent.com/primerDesignProgram.jsp.

Following sequence verification, clones were transformed into YDR483W BY4742 ΔKre2 *S. cerevisiae* cells using the method outlined in [38].

Thermostability/Activity Screen

S. cerevisiae carrying the *Hj*Cel5A plasmid were inoculated into 1 mL SD-Ura media in 24-well plates and allowed to grow overnight at 30 °C with shaking at 200 rpm. 4 mL of YPD were added and the cells were allowed to shake at 30 °C for an additional 48 hours before harvesting the supernatant through centrifugation. 5 µL of supernatant, 45 µL of YPD, and 60 µL of a 1.5% Avicel PH-101 (Sigma-Aldrich) slurry in 50 mL sodium acetate, pH 5.0 (cellulase buffer) were combined in a 96-well PCR plate and incubated for 1.5 hours to allow the CBM to bind to the substrate. The bound enzymes were washed three times with cellulase buffer and incubated at 73 °C for 2 hours. Following hydrolysis, 50 µL of the reaction supernatants were tested for reducing sugar concentrations via a modified Park-Johnson assay [39]. All screen samples were run in duplicate.

Park-Johnson Assay

To detect reducing end release, 50 µL of sample were combined with 100 µL of reagent A (0.5 g L^{-1} $K_3Fe_3(CN)_6$, 34.84 g L^{-1} PO_4K_2H, pH 6.0) and 50 µL of reagent B (5.3 g L^{-1}

Na_2CO_3, 0.65 g L^{-1} KCN). In experiments resulting in high amounts of reducing ends, 25 µL of sample is combined with 175 µL of the 2A:1B mixture. After incubating the mixture at 95 °C for 15 minutes in a PCR block, the plate is cooled on ice for five minutes. In a flat well plate, 90 µL of reagent C (2.5 g L^{-1} $FeCl_3$, 10 g L^{-1} polyvinyl pyrrolidone, 2 N H_2SO_4) is combined with 180 µL of the heat treated sample. The sample is then allowed to incubate for five minutes before measuring absorbance at 595 nm.

Enzyme Purification

Yeast colonies carrying the *Hj*Cel5A plasmid were inoculated into 6 mL of SD-Ura media and grown at 30 °C with shaking at 200 rpm. The preculture was then added to YPD and incubated for 48 hours. Following centrifugation, the supernatant was subjected to an 80% ammonium sulfate precipitation. The mixture was spun for 20 minutes at 8 kg and the pellet resuspended in 20 mL of lysis buffer (50 mM NaH_2PO_4 pH 7.4, 300 mM NaCl, 10 mM imidazole). Following a pH adjustment to 7.4, the protein was nutated at 4°C with 1 mL of Ni-NTA resin (Qiagen) conditioned with lysis buffer for 1 hour. The mixture was loaded into a gravity column, washed with 20 mL of lysis buffer, 20 mL of wash buffer (50 mM NaH_2PO_4 pH 7.4, 300 mM NaCl, 20 mM imidazole), and eluted with 6 mL of elution buffer (50 mM NaH_2PO_4 pH 7.4, 300 mM NaCl, 250 mM imidazole). After concentrating the elution to 0.5 mL, the protein was further purified and buffer exchanged into cellulase buffer through size exclusion chromatography. Protein concentrations were determined through measuring absorbance at 280 nm (ε_{280} = 81950 cm^{-1} M^{-1}).

T_{50} Assay

To assess thermostability via enzymatic activity, 40 µL of protein at a concentration of 0.25 µM was added to a PCR plate in triplicate for each of 12 temperatures. Enzyme was pretreated from 60-80 °C for ten minutes, then allowed to cool for an additional five minutes. 60 µL of a 1.5% Avicel slurry in cellulase buffer was added to each well and the plates were incubated at 60 °C for an hour. The plates were promptly cooled for 5 minutes on ice then centrifuged for 5 minutes to pellet the Avicel. Activity assessment with the Park-Johnson assay immediately followed using a 50 µL sample volume. To compare T_{50} values, the data were scaled from 0 to 1 using the following equation:

$$\text{Fraction Active} = \frac{(A_T - A_{\min})}{(A_{\max} - A_{\min})}$$

In this equation, A_T is the activity as measured by A_{595} at a particular temperature, $A_{\min}$ is the lowest observed activity, and $A_{\max}$ is the highest observed activity for a particular protein. T_{50} values were derived from generating curve fits using the Hill equation:

$$\text{Curve Fit} = \frac{T^n}{T^n + m^n}$$

Here n is the Hill coefficient, m is the T_{50}, and T is the temperature. Values for n and m were solved using the curve fit tool in MATLAB [40]. Because the T_{50} can fluctuate by approximately 1 °C depending on fluctuations in Avicel milling, subtle changes in cooling time, and PCR plate edge effects, all samples were run simultaneously with a WT standard. The ΔT_{50} values are calculated as $T_{50,\text{ mut}}$ - $T_{50,\text{ WT}}$.

Single-Point Activity Assay

To rigorously determine enzyme activity, 40 μL of enzyme at 0.5 μM was combined with 60 μL of 1.5% Avicel in a PCR plate. The mixture was incubated at 60 °C for two hours to allow hydrolysis to proceed. After cooling the plate on ice for 5 minutes, 100 μL of 0, 50, 100, 150, 200, 250, 300, and 350 μM cellobiose standards were added to the plate in triplicate. The plate was centrifuged to pellet the Avicel and 25 μL of the samples were extracted to perform a Park Johnson activity assay. All samples were tested in triplicate.

3.6 Tables and Figures

Table I. *Multiple sequence alignment characteristics*[a]

MSA Size	Accepted Identity (%)	Average Identity (%)	Level of Characterization Required	Gaps Removed
444	30-90	39.8	All Accepted	No
323	30-60	34.6	All Accepted	No
233	30-60	35.8	Partial/Hypothetical Removed	No
195	30-90	34.5	Partial/Hypothetical Removed	Yes
29	30-90	35.5	Precursors/Putatives Removed	Yes
10	30-100	74.0	Marked as Endoglucanase	Yes

[a] For alignment data in FASTA format, see attached files accompanying this thesis

Table II. *RE and MI values for mutations predicted as stabilizing*

Predicted Mutation	RE	Max MI	#MI>Noise[b]	ΔΔG (kcal mol^{-1})[c]
		444 Sequences		
T57N[a]	2.36	0.24	141	-2.7
I82L	1.45	0.48	241	-3.5
V101L	1.48	0.29	199	-1.0
Y135F	2.46	0.49	206	-0.8
W142I	1.61	0.48	241	-3.0
Q186T	1.61	0.48	221	-3.1
G293A[a]	1.95	0.31	221	-2.0
		323 Sequences		
D13E[a]	2.12	0.46	208	-1.5
T57N[a]	2.41	0.23	115	-2.5
N70P	2.34	0.45	224	-2.6
I82L	1.52	0.48	221	-3.9
V101L	1.46	0.33	180	-0.9
Y135F	2.51	0.45	186	-1.1
V164A	1.86	0.45	220	-1.7
V165I	2.06	0.34	3	-0.9
Q186T	1.66	0.34	201	-5.2
A255G	2.16	0.46	213	-1.0
G293A[a]	2.04	0.24	154	-2.0
		233 Sequences		
A10S	1.77	0.38	194	-1.0
D13E[a]	2.19	0.50	189	-1.9
T57N[a]	2.4	0.31	98	-2.4
V101L	1.48	0.43	154	-0.8
Y135F	2.49	0.48	178	-1.1
V165I	2.15	0.38	147	-1.1
Q186T	1.61	0.43	200	-4.9
A255G	2.2	0.44	185	-1.1
G293A[a]	2.13	0.24	106	-2.2
V302Y	2.59	0.49	192	-13.6
T308P	2.59	0.39	162	-4.6
		195 Sequences		
A10S	1.86	0.39	178	-1.1
N33P	2.43	0.48	158	-13.6
T57N[a]	2.39	0.34	118	-2.5
V101L	1.49	0.49	163	-0.6
Y135F	2.52	0.46	184	-1.2
V165I	2.17	0.42	147	-1.1
Q186T	1.61	0.48	201	-4.7
A255G	2.29	0.43	181	-1.3
G293A[a]	2.2	0.23	74	-2.3
V302Y	2.82	0.35	164	-13.7
T308P	2.86	0.20	46	-5.2

Table II Cont'd. *RE and MI values for mutations predicted as stabilizing*

Mutation	RE	Max MI	#MI>Noise[b]	ΔΔG (kcal mol^{-1})[c]
29 Sequences				
K32P	2.94	0.38	5	-13.6
T57N[a]	2.65	0.17	0	-3.3
N205D	2.48	0.41	26	-1.8
I276L	1.77	0.5	24	-13.7
10 Sequences				
E53D[a]	2.53	0.06	93	-13.7
G189A[a]	2.54	0.06	88	-13.7

[a] Indicates a thermostabilizing mutation. For values of all possible mutations, please consult the supplemental files attached to this thesis.

[b] Indicates the number of MI values greater than the background noise calculated for each position (see materials and methods).

[a] The value calculated as described by equation 3 (see page 50).

Table III. *Stabilizing Mutations*

Mutations	$T_{50,WT}$[a] (°C)	$T_{50,mut}$ (°C)	ΔT_{50} (°C)	Activity (µM Cellobiose Equivalents)	ΔActivity (µM Cellobiose Equivalents)	Expression Level Relative to WT	Location
WT	-	-	-	193.7±12.23	0.0	-	-
D13E	68.6±0.3	71.5±0.4	3.0±0.5	184.4±1.15	-9.3	1.6	Core
E53D	68.7±0.3	71.4±0.6	2.7±0.7	201.9±4.91	8.2	0.9	Boundary
T57N	71.0±0.0	72.1±0.0	1.1±0.0	240.7±4.08	47.0	0.3	Surface
G189A	70.8±0.3	71.2±0.3	0.4±0.4	190.3±3.37	-3.4	1.2	Boundary
G293A	70.7±0.1	74.3±0.1	3.9±0.2	221.0±1.75	27.3	0.8	Core
I82L	69.1±0.4	68.9±0.3	-0.2±0.5	N/A	N/A	1.6	Core
V101L	68.8±0.1	68.3±0.3	-0.5±0.3	N/A	N/A	2.3	Core

[a] The T_{50} of WT *Hj*Cel5A fluctuates by 1 °C due to variables described in the materials and methods section. All mutants were assayed simultaneously with a WT standard.

Table IV. *HjCel5A homologous crystal structures*

Originating Organism	Protein Name	PDB ID	Percent Identity (%)
Hypocrea jecorina	*Hj*Cel5A	3QR3 [19]	100
Thermoascus aurantiacus	*Ta*Cel5A	1H1N [41],1GZJ [28]	34
Uncultured Bacterium	RBcel1	4EE9 [29]	24
Piromyces rhinzinflatus	Eg1A	3AYR [42]	17
Pyrococcus horikoshii	Endocellulase	3QHO[43]	15
Acidothermus cellulolyticus	Endocellulase E1	1ECE [44]	17
Thermotoga maritima	*Tm*Cel5A	3MMU [45], 3AOF [46], 3AZR[46]	20, 16
Clostridium cellulovorans	Endoglucanase D	3NDY[a], 3NDZ[a]	20, 19
Fervidobacterium nodosum	*Fn*Cel5A	3NCO[a]	17
Prevotella bryantii	Endoglucanase	3VDH[a]	25
Clostridium cellulolyticum	celCCA	1EDG [47]	21
Paenisbacillus pabuli	GH5 Xyloglucanase	2JEP [48]	17
Bacillus sp.	Alkaline Cellulase K	1G0C [49]	16
Bacillus subtilis	Endoglucanase	3PZT [31], similar to 1LF1	23
Candidia Albicans	Exoglucanase	3N9K [50]	15
Thermobifida fusca	*Tf*Cel5A	2CKS[a], 2CKR[a]	24
Clostridium thermocellum	CelC	1CEC [51], 1CE0 [52]	18
Bacillus agaradhaerens	*Ba*Cel5A	7A3H [30]	21
Thermomonospora fusca	β-Mannase	1BQC [53]	14
Erwinia chrysanthemi	Cel5	1EGZ [54]	16

[a] To be published

Table V. *AUC values and optimal thresholds*

Number of Sequences	**444**	**323**	**233**	**195**	**29**	**10**
Area Under the Curve (AUC)						
RE	0.16	0.22	0.15	0.14	0.19	0.20
MI	0.03	0.02	0.04	0.02	-0.04	-0.13
ΔΔG[g]	0.23	0.25	0.19	0.20	0.25	0.27
RE (MI ≤ 0.5)[a]	0.42	0.39	0.39	0.36	0.44	0.16
ΔΔG (MI ≤ 0.5)[b,g]	0.41	0.42	0.34	0.33	0.39	0.12
Optimal Thresholds						
RE	0.02	-0.02	-0.02	0.03	-0.02	0.04
MI (w/RE)[c]	0.4	0.3	0.3	0.5	0.5	0.4-0.6
MI (w/ ΔΔG)[d]	0.4	0.3	0.3	0.3	0.2-0.3	≥5
ΔΔG	0.50	0.60-0.70	0.40	1.00	0.00-0.20	1.80-2.10
RE (MI ≤ 0.5)[e]	0.03-0.04	0.06-0.07	0.24-0.47	0.18-0.52	2.33-2.67	0.10-2.47
ΔΔG (MI ≤ 0.5)[f,g]	-0.40	-0.40	-0.10- -0.40	-0.10- -0.30	-3.30- -1.90	2.1-0.00

[a] The AUC from an RE ROC curve computed on a dataset filtered with MI.

[b] The AUC from a ΔΔG ROC computed on a dataset filtered with MI.

[c] The MI threshold giving the largest AUC from an RE ROC curve.

[d] The MI threshold giving the largest AUC from a ΔΔG ROC curve.

[e] The RE threshold giving the largest fraction of true positives/false positives while fixing the MI threshold.

[f] The ΔΔG threshold giving the largest fraction of true positives/false positives while fixing the MI threshold.

[g] The ΔΔG units are kcal mol^{-1}.

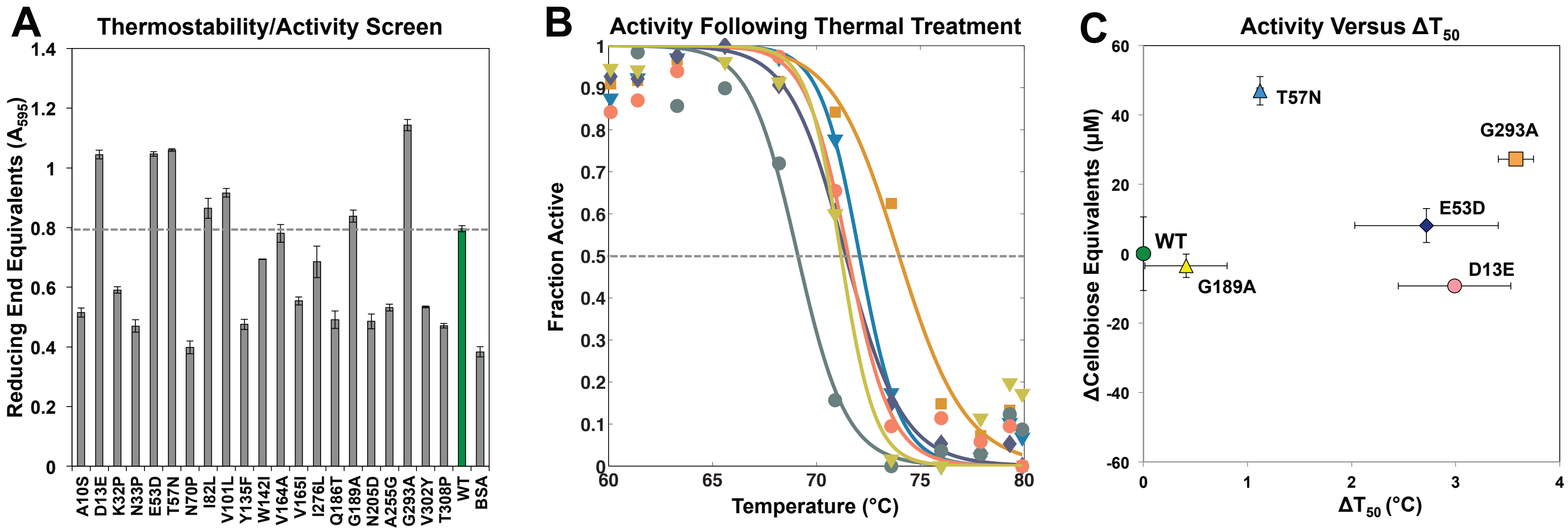


Figure 1. *Identifying stabilizing consensus mutations*. (A) Thermostability/activity screen performed on *Hj*Cel5A point mutants. WT is shown in green in all panels. Variants with activity exceeding that of WT, indicated by the dashed line, were purified and tested for thermostability. (B) Activity of *Hj*Cel5A point mutants after treatment over a range of temperatures. Data are shown for WT (green circles), D13E (pink circles), E53D (dark blue diamonds), T57N (light blue triangles), G189A (yellow triangles), and G293A (orange squares). The dashed line indicates the point at which 50% of the initial activity persists (T_{50}). Although each mutant was tested alongside WT to accurately assess (ΔT_{50}), only one WT trial is displayed for clarity. See Supplementary Figure 1 for additional data. (C) Activity of *Hj*Cel5A point mutants versus ΔT_{50}. Data are shown for WT (green circle), D13E (pink circle), E53D (dark blue diamond), T57N (light blue triangle), G189A (yellow triangle), and G293A (orange square). Values for both the change in activity and ΔT_{50} are reported with respect to WT.

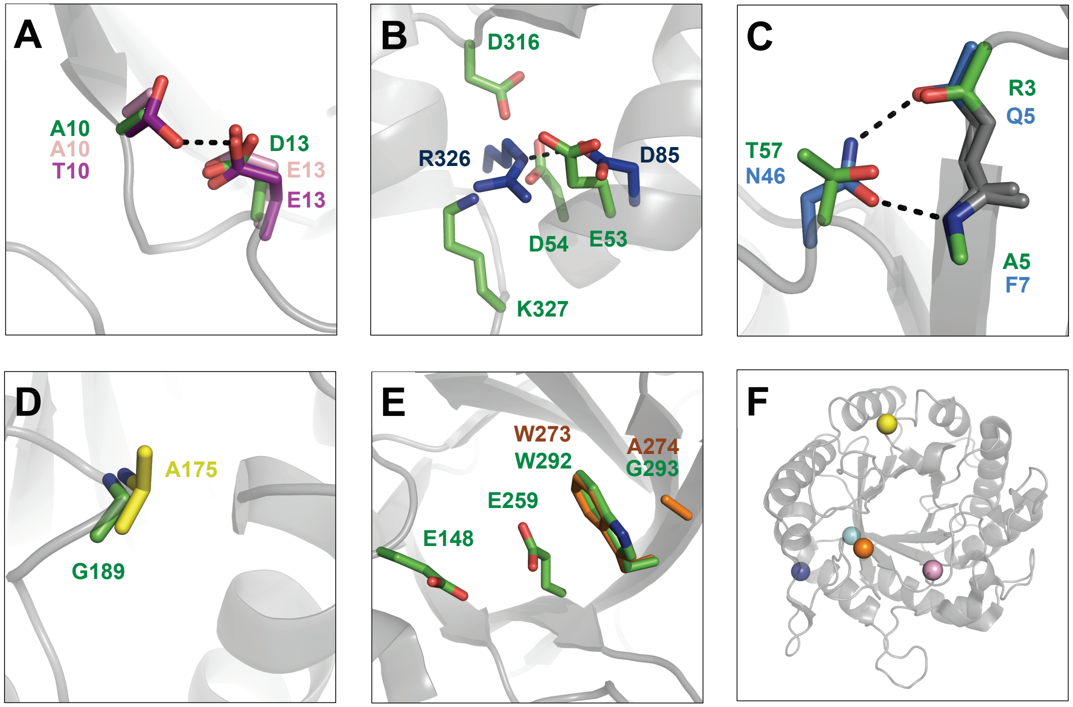


Figure 2. *Structural analysis of stabilizing mutations.* WT is shown in green in all panels. Numbering is retained from the original PDB files. (A) Equivalent residues at the D13E mutation site in *Ta*Cel5A (magenta, 1GZJ [28]) and RBcel1 (pink, 4EE9 [29]). The *Ta*Cel5A structure contains a threonine at position 10 that serves as a hydrogen bonding partner for E13. (B) Equivalent residues at the E53D mutation site in a *B. subtilis* endoglucanase (dark blue, 3PZT [31]). Although D85 and R326 form a salt bridge in the *B. subtilis* structure, *Hj*Cel5A lacks an adjacent arginine. (C) The T57N mutation site compared with *Ta*Cel5A (light blue, 1GZJ [28]). N46 forms two backbone hydrogen bonds in the *Ta*Cel5A structure. (D) A residue equivalent to G189A in Rbcel1 (yellow, 4EE9 [29]). (D) The equivalent residues at the G293A mutation site in *Ta*Cel5A (orange, 1GZJ [28]). W292/W273 serves as a substrate binding residue while E148 and E259 comprise the catalytic machinery of the enzyme. (F) Locations of the five stabilizing mutations, D13E (pink), E53D (dark blue), T57N (light blue), G189A (yellow), and G293A (orange).

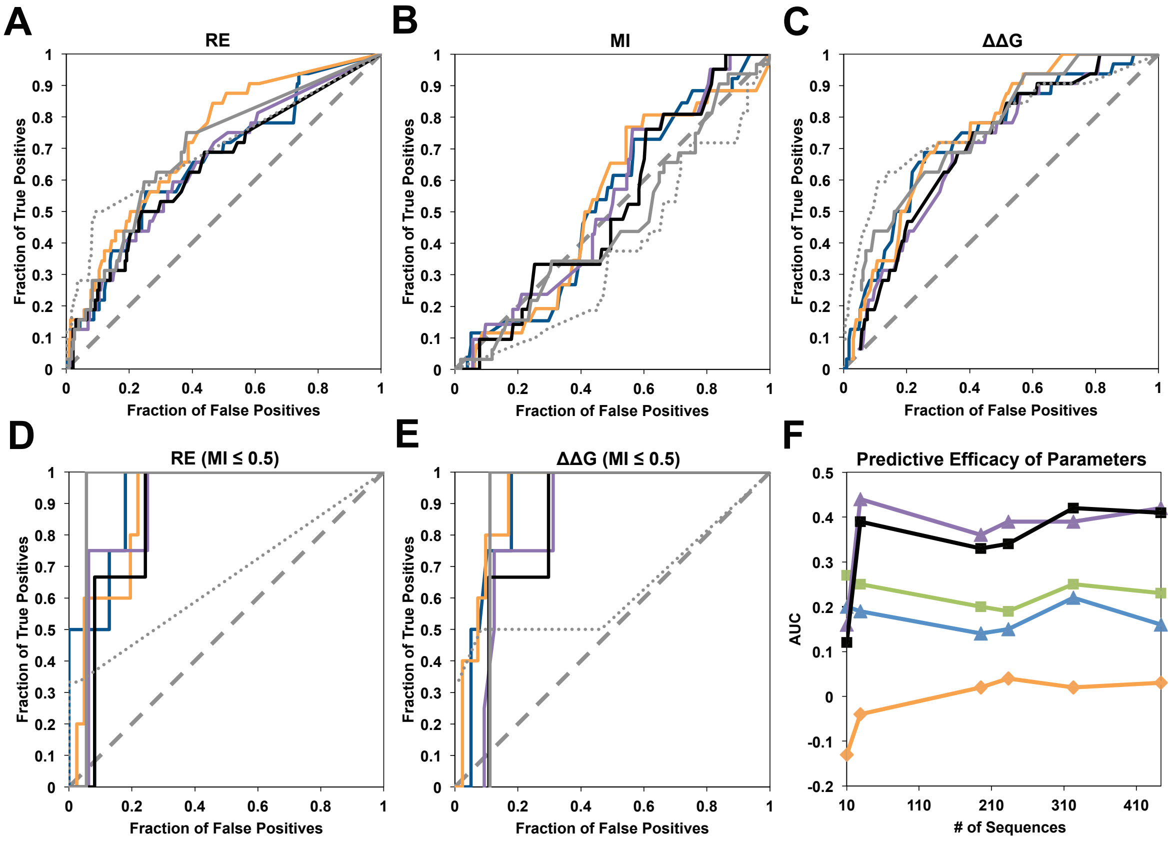


Figure 3. *Receiver Operator Characteristic curves.* (A-E) Receiver operator curves are shown with data for each alignment size: 444 sequences (blue), 323 sequences (orange), 233 sequences (purple), 195 sequences (black), 29 sequences (gray), and 10 sequences (dashed gray). The ROC plots were generated through comparing the number of true positives with the number of false positives while varying thresholds for (A) relative entropy, (B) mutual information, and (C) ΔΔG. ROC curves are also shown for (D) relative entropy and (E) ΔΔG on datasets only containing mutations at non-correlated sites (MI ≤ 0.5). (F) Area under the curve (AUC) versus the number of sequences in the alignment plotted for relative entropy (blue triangles), relative entropy (MI ≤ 0.5) (purple triangles), ΔΔG (green squares), ΔΔG (MI ≤ 0.5) (black squares), and mutual information (orange diamonds).

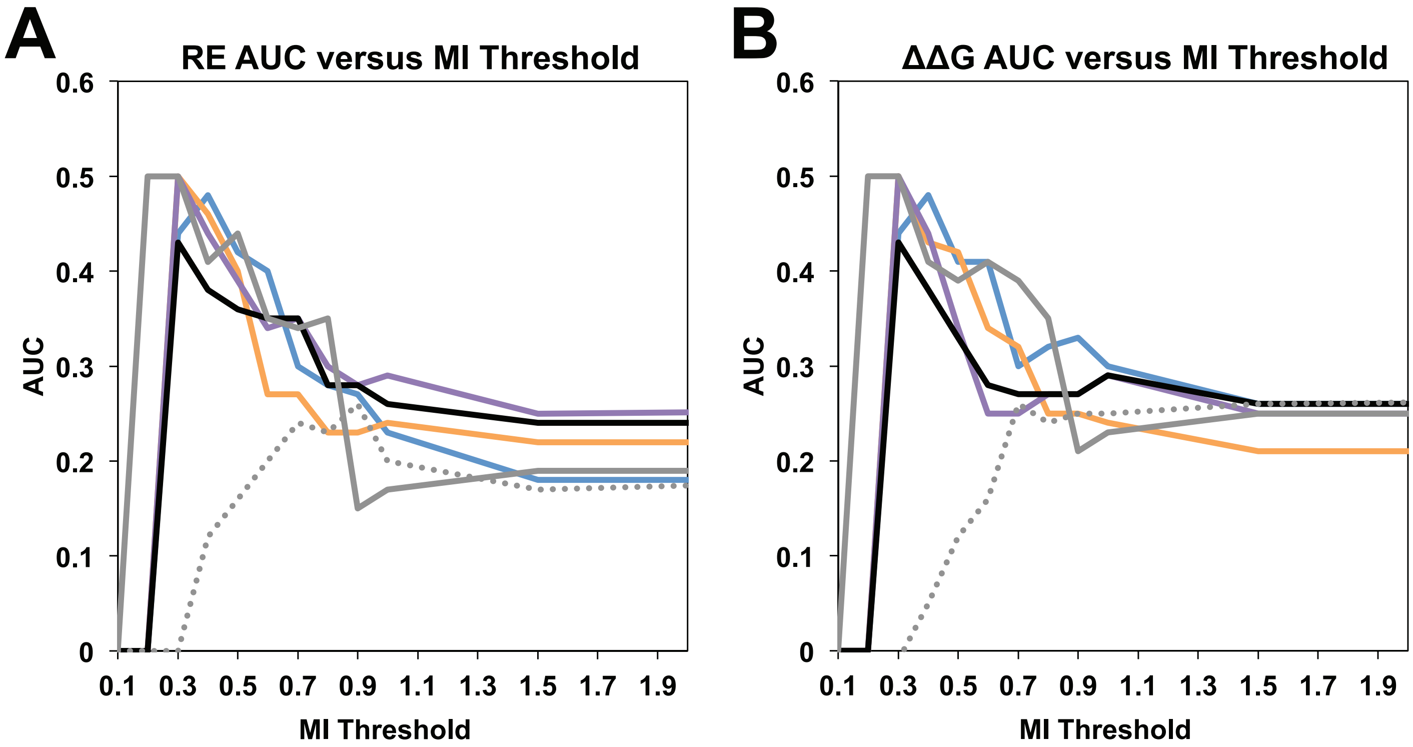


Figure 4. *Determining an optimal MI threshold.* Area under the (A) RE or (B) ΔΔG ROC curve (AUC) versus the MI threshold. Data are shown for each alignment size: 444 sequences (blue), 323 sequences (orange), 233 sequences (purple), 195 sequences (black), 29 sequences (gray), and 10 sequences (dashed gray).

3.7 Supplementary Tables and Figures

Supplementary Table I. *RE and ΔΔG AUC values for various MI thresholds*

Number of Sequences	**444**	**323**	**233**	**195**	**29**	**10**
Area Under the Curve (AUC)						
Mutual Information	0.03	0.02	0.04	0.02	-0.04	-0.13
Relative Entropy	0.16	0.22	0.15	0.14	0.19	0.20
RE (MI ≤ 0.1)	-	-	-	-	-	-
RE (MI ≤ 0.2)	-	-	-	-	0.50	-
RE (MI ≤ 0.3)	0.44	0.50	0.50	0.43	0.50	-
RE (MI ≤ 0.4)	0.48	0.46	0.44	0.38	0.41	0.12
RE (MI ≤ 0.5)	0.42	0.40	0.39	0.36	0.44	0.16
RE (MI ≤ 0.6)	0.40	0.27	0.34	0.35	0.35	0.20
RE (MI ≤ 0.7)	0.30	0.27	0.35	0.35	0.34	0.24
RE (MI ≤ 0.8)	0.28	0.23	0.30	0.28	0.35	0.23
RE (MI ≤ 0.9)	0.27	0.23	0.28	0.28	0.15	0.26
RE (MI ≤ 1.0)	0.23	0.24	0.29	0.26	0.17	0.20
RE (MI ≤ 1.5)	0.18	0.22	0.25	0.24	0.19	0.17
RE (MI ≤ 5.0)	0.18	0.22	0.26	0.24	0.19	0.20
RE (MI ≤ 10.0)	0.18	0.22	0.26	0.24	0.19	0.20
ΔΔG	0.23	0.25	0.19	0.20	0.25	0.27
ΔΔG (MI ≤ 0.1)	-	-	-	-	-	-0.01
ΔΔG (MI ≤ 0.2)	-	-	-	-	0.5	-0.01
ΔΔG (MI ≤ 0.3)	0.44	0.50	0.50	0.43	0.5	-0.01
ΔΔG (MI ≤ 0.4)	0.48	0.43	0.44	0.38	0.41	0.05
ΔΔG (MI ≤ 0.5)	0.41	0.42	0.34	0.33	0.39	0.12
ΔΔG (MI ≤ 0.6)	0.41	0.34	0.25	0.28	0.41	0.16
ΔΔG (MI ≤ 0.7)	0.30	0.32	0.25	0.27	0.39	0.26
ΔΔG (MI ≤ 0.8)	0.32	0.25	0.27	0.27	0.35	0.24
ΔΔG (MI ≤ 0.9)	0.33	0.25	0.27	0.27	0.21	0.25
ΔΔG (MI ≤ 1.0)	0.30	0.24	0.29	0.29	0.23	0.25
ΔΔG (MI ≤ 1.5)	0.26	0.21	0.25	0.26	0.25	0.26
ΔΔG (MI ≤ 5.0)	0.26	0.21	0.25	0.26	0.25	0.27
ΔΔG (MI ≤ 10.0)	0.26	0.21	0.25	0.26	0.25	0.27

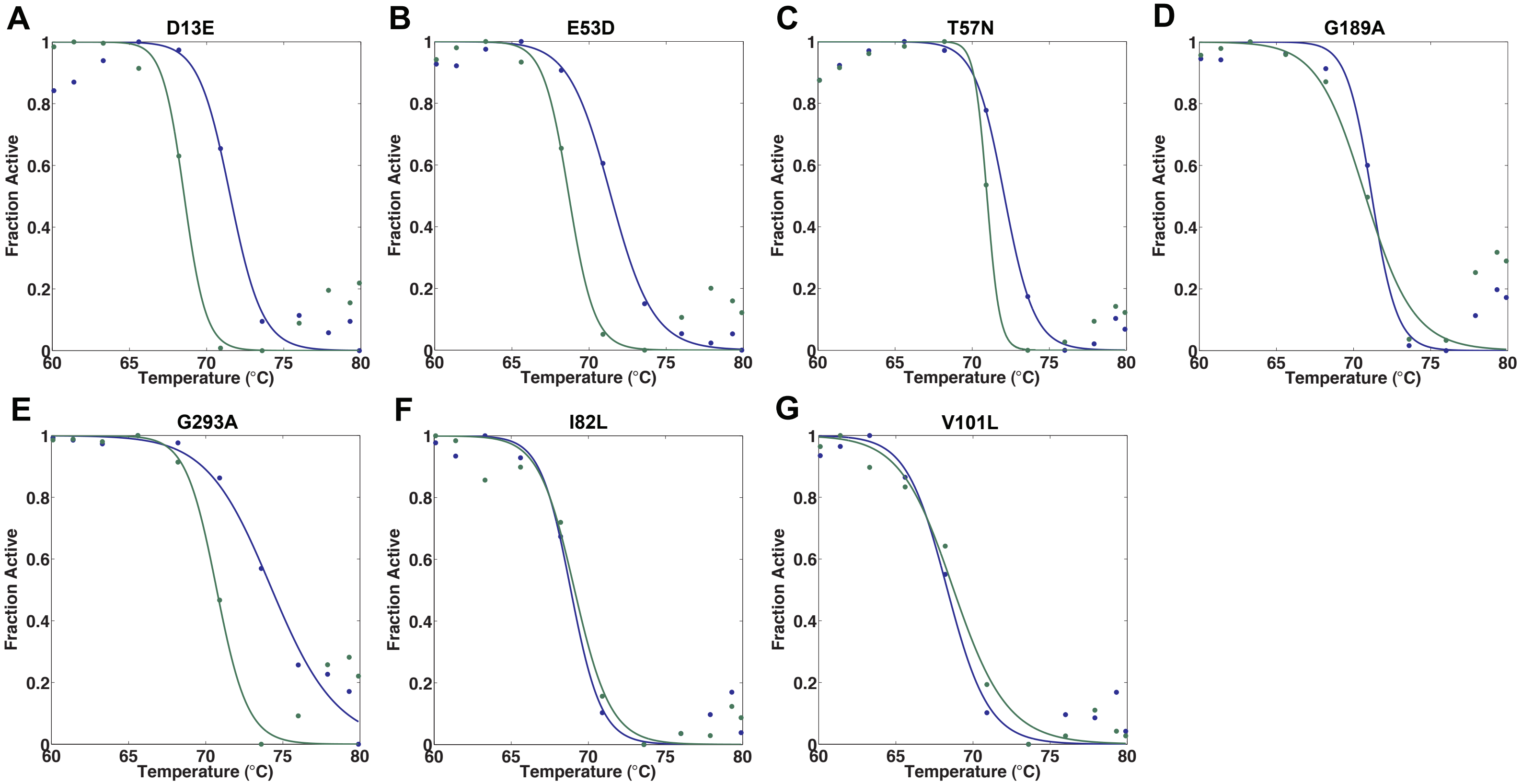


Figure S1. *T_{50} plots of all tested point mutants.* The fraction active after a 10 minute heat treatment from 60-80 °C is shown for mutants (A) D13E, (B) E53D, (C) T57N, (D) G189A, (E) G293A, (F) 182L, and (G) V101L. WT and the point mutant are shown in green and blue, respectively.

3.8 References

1. Viikari, L., Alapuranen M., Puranen T., Vehmaanperä J. and Siika-aho M. Thermostable enzymes in lignocellulose hydrolysis. In: Olsson L, Ed. (2007) Biofuels. Springer Berlin Heidelberg, pp. 121-145.
2. Amin, N., Liu A.D., Ramer S., Aehle W., Meijer D., Metin M., Wong S., Gualfetti P. and Schellenberger V. (2004) Construction of stabilized proteins by combinatorial consensus mutagenesis. Protein Engineering Design and Selection 17:787-793.
3. Komor, R.S., Romero P.A., Xie C.B. and Arnold F.H. (2012) Highly thermostable fungal cellobiohydrolase I (Cel7A) engineered using predictive methods. Protein Engineering Design and Selection 25:827-833.
4. Wu, I. and Arnold F.H. (2013) Engineered thermostable fungal Cel6A and Cel7A cellobiohydrolases hydrolyze cellulose efficiently at elevated temperatures. Biotechnology and Bioengineering 110:1874-1883.
5. Vázquez-Figueroa , E., Chaparro-Riggers J. and Bommarius A.S. (2007) Development of a thermostable glucose dehydrogenase by a structure-guided consensus concept. ChemBioChem 8:2295-2301.
6. Wang, W. (1999) Instability, stabilization, and formulation of liquid protein pharmaceuticals. International Journal of Pharmaceutics 185:129-188.
7. Solá, R.J. and Griebenow K. (2009) Effects of glycosylation on the stability of protein pharmaceuticals. Journal of Pharmaceutical Sciences 98:1223-1245.
8. Steipe, B., Schiller B., Plückthun A. and Steinbacher S. (1994) Sequence statistics reliably predict stabilizing mutations in a protein domain. Journal of Molecular Biology 240:188-192.
9. Ohage, E. and Steipe B. (1999) Intrabody construction and expression. I. The critical role of VL domain stability. Journal of Molecular Biology 291:1119-1128.
10. Main, E.R.G., Xiong Y., Cocco M.J., D'Andrea L. and Regan L. (2003) Design of stable α-helical arrays from an idealized TPR motif. Structure 11:497-508.
11. Maxwell, K.L. and Davidson A.R. (1998) Mutagenesis of a buried polar interaction in an SH3 domain: Sequence conservation provides the best prediction of stability effects. Biochemistry 37:16172-16182.
12. Nikolova, P.V., Henckel J., Lane D.P. and Fersht A.R. (1998) Semirational design of active tumor suppressor p53 DNA binding domain with enhanced stability. Proceedings of the National Academy of Sciences 95:14675-14680.
13. Lehmann, M., Loch C., Middendorf A., Studer D., Lassen S.r.F., Pasamontes L., van Loon A.P.G.M. and Wyss M. (2002) The consensus concept for thermostability engineering of proteins: further proof of concept. Protein Engineering 15:403-411.
14. Anbar, M., Gul O., Lamed R., Sezerman U.O. and Bayer E.A. (2012) Improved thermostability of *Clostridium thermocellum* endoglucanase Cel8A by using consensus-guided mutagenesis. Applied and Environmental Microbiology 78:3458-3464.
15. Sullivan, B.J., Nguyen T., Durani V., Mathur D., Rojas S., Thomas M., Syu T. and Magliery T.J. (2012) Stabilizing proteins from sequence statistics: the

Interplay of conservation and correlation in triosephosphate isomerase stability. Journal of Molecular Biology 420:384-399.

16. Suominen, P.L., Mantyla A.L., Karhunen T., Hakola S. and Nevalainen H. (1993) High frequency one-step gene replacement in *Trichoderma reesei*. II. Effects of deletions of individual cellulase genes. Molecular Genetics and Genomics 241:523-530.

17. Heinzelman, P., Snow C.D., Wu I., Nguyen C., Villalobos A., Govindarajan S., Minshull J. and Arnold F.H. (2009) A family of thermostable fungal cellulases created by structure-guided recombination. Proceedings of the National Academy of Sciences 106:5610-5615.

18. Kumar, R., Singh S. and Singh O. (2008) Bioconversion of lignocellulosic biomass: biochemical and molecular perspectives. Journal of Industrial Microbiology & Biotechnology 35:377-391.

19. Lee, T.M., Farrow M.F., Arnold F.H. and Mayo S.L. (2011) A structural study of *Hypocrea jecorina* Cel5A. Protein Science 20:1935-1940.

20. Martin, L.C., Gloor G.B., Dunn S.D. and Wahl L.M. (2005) Using information theory to search for co-evolving residues in proteins. Bioinformatics 21:4116-4124.

21. Magliery, T. and Regan L. (2005) Sequence variation in ligand binding sites in proteins. BMC Bioinformatics 6:240.

22. Seiboth, B., Ivanova C. and Seidl-Seiboth V. *Trichoderma reesei*: a fungal enzyme producer for cellulosic biofuels. In: Bernardes MADS, Ed. (2011) Biofuel Production-Recent Developments and Prospects. pp. 310-340.

23. Karlsson, J., Siika-aho M., Tenkanen M. and Tjerneld F. (2002) Enzymatic properties of the low molecular mass endoglucanases Cel12A (EG III) and Cel45A (EG V) of *Trichoderma reese*i. Journal of Biotechnology 99:63-78.

24. Ouyang, J., Yan M., Kong D. and Xu L. (2006) A complete protein pattern of cellulase and hemicellulase genes in the filamentous fungus *Trichoderma reesei*. Biotechnology Journal 1:1266-1274.

25. Rost, B. (1999) Twilight zone of protein sequence alignments. Protein Engineering 12:85-94.

26. Sauder, J.M., Arthur J.W. and Dunbrack Jr R.L. (2000) Large-scale comparison of protein sequence alignment algorithms with structure alignments. Proteins: Structure, Function, and Bioinformatics 40:6-22.

27. Cronbach, L.J. On the non-rational application of information measures in psychology. In: Quastler H, Ed. (1954) Information Theory in Psychology: Problems and Methods. Free Press, Glencoe, Illinois, pp. 14-30.

28. Lo Leggio, L. and Larsen S. (2002) The 1.62 Å structure of *Thermoascus aurantiacus* endoglucanase: completing the structural picture of subfamilies in glycoside hydrolase family 5. FEBS Letters 523:103-108.

29. Delsaute, M., Berlemont R., Dehareng D., Van Elder D., Galleni M. and Bauvois C. (2013) Three-dimensional structure of RBcel1, a metagenome-derived psychrotolerant family GH5 endoglucanase. Acta Crystallographica Section F 69:828-833.

30. Davies, G.J., Mackenzie L., Varrot A., Dauter M., Brzozowski A.M., Schülein M. and Withers S.G. (1998) Snapshots along an enzymatic reaction coordinate: analysis of a retaining β-glycoside hydrolase. Biochemistry 37:11707-11713.

31. Santos, C.R., Paiva J.H., Sforça M.L., Neves J.L., Navarro R.Z., Cota J., Akao P.K., Hoffmam Z.B., Meza A.N., Smetana J.H., Nogueira M.L., Polikarpov I., Xavier-Neto J., Squina F.M., Ward R.J., Ruller R., Zeri A.C. and Murakami M.T. (2012) Dissecting structure–function–stability relationships of a thermostable GH5-CBM3 cellulase from *Bacillus subtilis* 168. Biochemical Journal 441:95-104.

32. Jäckel, C., Bloom J.D., Kast P., Arnold F.H. and Hilvert D. (2010) Consensus Protein design without phylogenetic bias. Journal of Molecular Biology 399:541-546.

33. Bloom, J.D. and Glassman M.J. (2009) Inferring stabilizing mutations from protein phylogenies: application to influenza hemagglutinin. PLoS Comput Biol 5.

34. Mason, S.J. and Graham N.E. (2002) Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: statistical significance and interpretation. Quarterly Journal of the Royal Meteorological Society 128:2145-2166.

35. Nakamura, Y., Gojobori T. and Ikemura T. (2000) Codon usage tabulated from the international DNA sequence databases: status for the year 2000. Nucleic Acids Research 28:292.

36. Altschul, S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W. and Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25:3389-3402.

37. Gibson, D.G., Lei Y., Ray-Yuan C., Venter J.C., Hutchinson C.A. and Smith H.O. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. Nature Methods 6:343-345.

38. Schiestl, R.H., Manivasakam P., Woods R.A. and Gietz R.D. Introducing DNA into yeast by transformation. In: Johnston M,Fields S, Eds. (1993) Methods; a companion to Methods in Mezymology. Academic Press, Inc. , pp. 79-85.

39. Park, J.T. and Johnson M.J. (1949) A submicrodetermination of glucose. Journal of Biological Chemistry 181:149-151.

40. MATLAB and Statistics Toolbox Release 2011b, The MathWorks, Inc., Natick, Massachusetts, United States.

41. Van Petegem, F., Vandenberghe I., Bhat M.K. and Van Beeumen J. (2002) Atomic resolution structure of the major endoglucanase from *Thermoascus aurantiacus*. Biochemical and Biophysical Research Communications 296:161-166.

42. Tseng, C.-W., Ko T.-P., Guo R.-T., Huang J.-W., Wang H.-C., Huang C.-H., Cheng Y.-S., Wang A.H.J. and Liu J.-R. (2011) Substrate binding of a GH5 endoglucanase from the ruminal fungus *Piromyces rhizinflata*. Acta Crystallographica Section F 67:1189-1194.

43. Kim, H.W. and Ishikawa K. (2011) Functional analysis of hyperthermophilic endocellulase from *Pyrococcus horikoshii* by crystallographic snapshots. Biochemical Journal 437:223-230.

44. Sakon, J., Adney W.S., Himmel M.E., Thomas S.R. and Karplus P.A. (1996) Crystal structure of thermostable family 5 endocellulase E1 from A*cidothermus cellulolyticus* in complex with cellotetraose. Biochemistry 35:10648-10660.

45. Pereira, J.H., Chen Z., McAndrew R.P., Sapra R., Chhabra S.R., Sale K.L., Simmons B.A. and Adams P.D. (2010) Biochemical characterization and crystal structure of endoglucanase Cel5A from the hyperthermophilic *Thermotoga maritima*. Journal of Structural Biology 172:372-379.

46. Wu, T.-H., Huang C.-H., Ko T.-P., Lai H.-L., Ma Y., Chen C.-C., Cheng Y.-S., Liu J.-R. and Guo R.-T. (2011) Diverse substrate recognition mechanism revealed by *Thermotoga maritima* Cel5A structures in complex with cellotetraose, cellobiose and mannotriose. Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics 1814:1832-1840.

47. Ducros, V., Czjzek M., Belaich A., Gaudin C., Fierobe H.P., Davies G.J. and Haser R. (1995) Crystal structure of the catalytic domain of a bacterial cellulase belonging to family 5. Structure 3:939-949.

48. Gloster, T.M., Ibatullin F.M., Macauley K., Eklöf J.M., Roberts S., Turkenburg J.P., Bjørnvad M.E., Jørgensen P.L., Danielsen S., Johansen K.S., Borchert T.V., Wilson K.S., Brumer H. and Davies G.J. (2007) Characterization and three-dimensional structures of two distinct bacterial xyloglucanases from families GH5 and GH12. Journal of Biological Chemistry 282:19177-19189.

49. Shirai, T., Ishida H., Noda J.-i., Yamane T., Ozaki K., Hakamada Y. and Ito S. (2001) Crystal structure of alkaline cellulase K: insight into the alkaline adaptation of an industrial enzyme. Journal of Molecular Biology 310:1079-1087.

50. Patrick, W.M., Nakatani Y., Cutfield S.M., Sharpe M.L., Ramsay R.J. and Cutfield J.F. (2010) Carbohydrate binding sites in *Candida albicans* exo-β-1,3-glucanase and the role of the Phe-Phe 'clamp' at the active site entrance. FEBS Journal 277:4549-4561.

51. Dominguez, R., Souchon H., Spinelli S., Dauter Z., Wilson K.S., Chauvaux S., Béguin P. and Alzari P.M. (1995) A common protein fold and similar active site in two distinct families of beta-glycanases. Nature Structural & Molecular Biology 2:569-576.

52. Domínguez, R., Souchon H., Lascombe M.-B. and Alzari P.M. (1996) The Crystal structure of a family 5 endoglucanase mutant in complexed and uncomplexed forms reveals an induced fit activation mechanism. Journal of Molecular Biology 257:1042-1051.

53. Hilge, M., Gloor S.M., Rypniewski W., Sauer O., Heightman T.D., Zimmerman W., Winterhalter K. and Piontek K. (1998) High-resolution native and complex structures of thermostable beta-mannase from *Thermomonospora fusca* - substrate specificity in glycosyl hydrolyase family 5. Structure 6:1433-1444.

54. Chapon, V., Czjzek M., El Hassouni M., Py B., Juy M. and Barras F. (2001) Type II protein secretion in gram-negative pathogenic bacteria: the study of the structure/secretion relationships of the cellulase cel5 (formerly EGZ) from *Erwinia chrysanthemi*. Journal of Molecular Biology 310:1055-1066.

CHAPTER 4

Identifying Stabilizing Mutations in *Hypocrea jecorina* Cel5A Through Computational Methods: Core Repacking and Helix Dipole Surface Stabilization

This chapter is formatted for submission to the Journal of Molecular Biology.

4.1 Abstract

Canonical methods of computational protein thermostabilization often seek to stabilize specific structural regions. These methods include, but are not limited to, core stabilization through hydrophobic repacking or engineering more stable protein surfaces through methods such as helix dipole stabilization. While several studies have attempted to incrementally improve these methods, little attention has focused on directly comparing their effectiveness. Here we identify stabilizing mutations in the primary endoglucanase from *Hypocrea jecorina* (*Hj*Cel5A) using two computational methods: 1) core repacking and 2) helix dipole stabilization. We identify two and nine stabilizing mutations from the core repacking and helix dipole stabilization strategies, respectively, that may be useful for industrial or future research purposes. While the helix dipole stabilization strategy revealed more stabilizing mutations than the core repacking method, many of these mutations only marginally improved protein thermostability. We demonstrate that these mutations can further improve thermostability and protein expression when incorporated into a combination construct. Finally, analysis of a 262-member *Hj*Cel5A point mutation database suggests that the helix termini, and the surface/boundary region in general, appear more amenable to mutation than the core of the protein. Highly stabilizing mutations, however, appear to evenly fall between the core and boundary.

4.2 Introduction

The thermostabilization of useful proteins is a longstanding goal in the field of biochemistry. Improving resistance to thermal degradation not only preserves enzymatic activity at elevated temperatures, but may also confer increased resistance to proteolysis [1] and increase half-life across a thermal range [2]. When used in industrial applications, higher operating temperatures may also reduce bacterial contamination and diminish solution viscosity, resulting in lower operating costs [3]. Despite these numerous benefits, a universal strategy for rapidly generating thermostable protein variants remains elusive. Studying protein thermostability may provide key insights towards improving current stabilization methods, ultimately rendering protein products more suitable for real-world applications.

In the past two decades, computational protein design has become a canonical means of generating thermostable protein variants. This design strategy attempts to stabilize the folded structure while discouraging the unfolded state. It has long been accepted that the hydrophobic effect is the principle driver of protein folding [4], i.e., proteins primarily fold to bury hydrophobic groups in a solvent-shielded "core," thereby minimizing the unfavorable disruption of aqueous polar contacts. As such, many rotamer optimization algorithms focus on mutating the protein core to increase hydrophobic sidechain content or improve packing. This strategy has led to the successful stabilization of RNase HI [5], λ repressor [6], and streptococcal protein G β1 [7]. Complete protein redesign projects have demonstrated that thermostable variants display an enrichment of nonpolar residues in RNA-binding U1A and procarboxypeptidase (activity was not assessed) [8], lending further credence to this concept. Moreover, core repacking studies on *Bacillus subtilis* lipase A [9] demonstrates that the technique can simultaneously improve stability and function.

Despite these findings, several studies suggest that better packing does not necessarily lead to higher thermostability. Comparisons of certain highly stable proteins (the glyceraldehyde-3-phosphate dehydrogenase from *Sulfolobus solfataricus* [10], isopropylmalate dehydrogenase [11], and *Sulfolobus acidocaldarius* superoxide dismutase [12]) with their mesostable counterparts demonstrate no change in packing volume [13]. Furthermore, creation of a 32 Å^3 cavity in *Thermus thermophilus* isopropylmalate dehydrogenase had no effect on thermostability [11].

Recent studies increasingly indicate that many stabilizing mutations reside beyond the core [14]. Malakauskas et al. were able to design a variant streptococcal protein G β1 with a T_m in excess of 100 °C through targeting the boundary region, the area between the core and surface of the protein [15]. Marshall et al. improved the thermostability of the *Drosophila* engrailed homeodomain through considering N-capping and helix dipole effects, strategies that target the protein surface [16]. In addition, Joo et al. generated thermostable variants of a *Bacillus circulans* xylanase with activity similar to wild type (WT) by applying a cavity-filling method to surface pockets [17]. While these studies demonstrate that thermostable mutations exist beyond the core, they do not comparatively examine whether probing specific protein regions over others will yield more fruitful results.

Here, we utilize computational rotamer optimization to compare two methods of protein stabilization that target the core (core repacking) and the surface/boundary (N-capping and helix dipole stabilization) within the same protein system. Experiments were conducted on the primary endoglucanase (*Hj*Cel5A) from *Hypocrea jecorina* (anamorph *Trichoderma reesei*) [18]. This cellulase, an enzyme capable of hydrolyzing cellulose into smaller components, is a thermostabilization target for cellulosic biofuel production [19]. In addition to revealing two core and nine helix-dipole-stabilizing mutations in *Hj*Cel5A, we also investigate whether these mutations diminish catalytic activity and expression. Furthermore, we use the results generated in this study and stability information in a database of 262 *Hj*Cel5A point mutations to discern patterns in the spatial distribution of stabilizing mutations.

4.3 Results and Discussion

Residue Classification

In preparation for computational design, we first classified residues in the 2.1 Å *Hj*Cel5A crystal structure (PDB ID 3QR3 [19]) according to their proximity to the solvent-exposed surface. In total, 131, 118, and 80 residues were identified as part of the core, boundary, and surface, respectively. As expected, most of the core residues reside in the interior of the α/β barrel while the boundary and surface residues primarily decorate the solvent-exposed helical faces (Figure 1).

Core Repacking Calculation

Our strategy for identifying stabilizing core mutations relies on computationally mutating core positions to improve hydrophobic packing. Before performing a repacking calculation, one must chose positions within the protein for examination. Core residues forming contacts with the protein backbone (8, 46, 81, 85, 105, 128, 157, 160, 169, 196, 219, 222, 241, 253, 285, 286, 289, 296, 304, 305, 306, 317, 326) were discarded. Residues important for catalysis and their nearby neighbors (102, 259, 288, 293), positions near disulfide-bonded cysteines (64), prolines (41, 149,181, 307), and glycines with glycine-specific Φ and Ψ angles were also removed (11). In total, 57 design positions (5, 6, 7, 9, 10, 12, 20, 31, 47, 56, 58, 59, 61, 63, 65, 69, 82, 88, 89, 100, 101, 103, 107, 124, 127, 132, 141, 143, 144, 145, 158, 161, 165, 178, 180, 182, 188, 191, 213, 214, 215, 217, 221, 255, 256, 257, 261, 262, 272, 275, 276, 290, 291, 319, 320, 324, 325) were chosen from the initial pool of 131 core residues.

Our computational design algorithm functions through iteratively generating series of alternative sidechain conformations and identities at design positions and preserving those that provide an energy benefit. These sidechain conformations can be generated using ideal bond angles (rotamers) or modeled from sidechains observed from structures in the Protein Data Bank (conformers). We performed the core repacking calculation using two rotamer libraries based on those developed by Dunbrack and Karplus [20] and a conformer library described by Lassila et al. [21]. The rotamer libraries include a backbone-independent set containing the most probable χ angles optimized for polar residues (bbind02.May.e0) and a backbone-independent set containing rotamers with mean χ values and mean $\chi \pm 1$ standard deviation for χ_1 and χ_2 (bbind02.May.e2). This rotamer library is suggested for use on aromatic residues. A midsized backbone-independent conformer library was also used for calculations (bda-bbind_1.0.cpdslib). The use of larger or backbone dependent libraries proved too computationally costly and was not further pursued.

We performed two sets of calculations, allowing design positions to sample any amino acid identity in one case and restricting the allowed residues to hydrophobics (Ala, Val, Leu, Ile, Phe, Tyr, Trp) in the other (Table I). While the identity of predicted mutations differed among the six sets of calculations (Table II), no discernable trend was observed from these distinctions. Collectively, these calculations predicted 32 mutations as stabilizing.

Detection of Stabilizing Core Mutations

Previous studies have redesigned protein cores in batch, introducing several mutations into a single construct simultaneously. Our earlier efforts to stabilize *Hj*Cel5A demonstrate that the inclusion of even one highly deleterious mutation may result in an unfolded, or inactive protein. This observation is consistent with literature reporting that most mutations are destabilizing and that introducing a highly destabilizing point mutation can “completely collapse” the structure [22-24]. Creating and screening point mutations as opposed to composite constructs containing multiple mutations is relatively fast and inexpensive. Screens performed on individual mutations also provide clear evidence concerning whether the mutation should pass to the next round of characterization or be discarded. Moreover, a recent study utilizing a core repacking algorithm optimized to select point mutations achieved a 17.6 °C thermostability increase from three highly stabilizing core mutations [6]. Thus, even if the majority of core point mutants require a compensatory mutation to avoid adverse effects, detecting a handful of highly stabilizing point mutations might provide sufficient stability for the intended purpose.

To determine whether our set of predicted stabilizing core mutations could be made as point mutants, average values for maximal mutual information (MI) were tabulated for each position. This value employs protein sequence data to measure covariance between a pair of sites and increases as correlation levels rise (see Chapter 2). The average maximum MI for all positions in HjCel5A (0.70) was higher than the average for the core mutation set (0.62), demonstrating that mutations at the selected positions did not exhibit a particularly high need for compensatory mutations.

The 32 predicted *Hj*Cel5A mutations were individually cloned as point mutations and screened for stability and adequate activity using the methods outlined in Chapter 3 (Figure 2A). Supernatants harboring secreted protein were screened for activity after incubation for 1 hour at 73 °C, 3.5 degrees higher than the WT T_m (69.5 °C). Only two mutations with activities exceeding WT were detected from the screening step, I82M and V101I. Two variants, each carrying one of these stabilizing mutations were expressed and purified to assess thermostability more directly. Both mutants slight increases in thermostability with I82M conferring a 0.3 and V101I a 0.5 °C increase in $T_{50,}$ the temperature at which half of the maximal activity persists (Table V, Supplementary Figure 1A and B).

Helix Dipole Stabilization Calculation

In the absence of external contacts, all α helices contain a natural dipole arising from three unsatisfied hydrogen bonds at each terminus [25-27]. Several studies have demonstrated that stabilizing this dipole can confer stability to the protein in entirety. This has been achieved through either introducing N-capping interactions, hydrogen bonds between the side-chain of the residue immediately preceding the helix or through mutating residues at the ends of the helix to counter the partial electrostatic charges.

To identify helix dipole stabilizing mutations in *Hj*Cel5A, we have adopted the strategy developed by Marshall et al. for the *Drosophila* engrailed homeodomain [16]. In this design scheme, N-capping positions sample amino acid identities with the highest N-capping propensity (Ser, Thr, Asn, and Asp) [28]. The three most N-terminal residues are prohibited from mutating to positively charged amino acids (His, Lys, and Arg), while the three most C-terminal residues are barred from sampling negatively charged sidechains (Asp and Glu).

Using the general Marshall strategy, we performed parallel calculations allowing two sets of residues at either terminus. In the first scheme, the N- and C-terminal residues may adopt any identity except for those that violate the aforementioned rules. Thus, the three most N-terminal positions of a helix may remain as the WT residue or mutate to any other sidechain except His, Lys, or Arg. Likewise, the three most C-terminal positions of a helix may sample WT or any other identity other than Asp or Glu. The second "strict" scheme allows residues at the N- and C-terminal residues to mutate only if the charge of the introduced sidechain will counter the dipole. The three most N-terminal positions of a helix, for example, may only mutate to an Asp or Glu when favorable. WT was included as an option for all design positions in both calculation schemes. The architecture of *Hj*Cel5A is a TIM barrel fold containing eight major helices [19], many of which have ambiguous termini. During the selection of design positions (Table III), some exceptions to the provided rules were allowed to accommodate these eccentricities. To reduce the computational load, each helix was redesigned separately using the bda-bbind_1.0.cpdslib conformer library employed in the core repacking calculation. Collectively, these computations predicted 44 mutations with 55% and 34% appearing in the boundary and surface, respectively (Table IV).

Detection of Stabilizing Helix Dipole Mutations

For the reasons outlined above, the 44 predicted helix mutations were constructed as point mutants and screened for activity with the same procedure used to probe core mutations (Figure 2B). In the activity screen, 14 constructs demonstrated greater activity than WT. Following purification and activity screening at a gradient of temperatures, nine constructs demonstrated a positive ΔT_{50} (T80E, S133R, N155E, N155Q, T165E, G239E, Y278F, S318E, and S318Q), four showed a decrease in thermostability (S79E, T80Q, A122E, and G239Q), and one behaved similarly to WT (S79Q) (Table V, Supplementary

Figure 1C-P). As is the case with the core mutations, the helix dipole stabilizing mutations provide modest stability benefits ($\Delta T_{50} \leq 1$ °C). Only five of the nine stabilizing mutations exhibit a $\Delta T_{50} \geq 0.5$ °C.

Given the relatively low enhancements in thermostability observed for the helix dipole stabilizing mutations, it remained unclear whether these mutations would provide any tangible benefit. We created a combination construct containing the T80E, S133R, N155E, G239E, Y278F, and S318Q mutations and determined its ΔT_{50} (2.4 °C) and optimal reaction temperature ($T_{opt\,,helix\ combo}$= 66 °C, $T_{opt\,,WT}$ = 63.5 °C) (Figure 5A and B). Both values show modest increases of ~2.5 °C. While the six stabilizing mutations did not additively increase T_{50} and T_{opt}, the combination mutant still demonstrates improved thermostability compared to the most beneficial helix point mutant. In addition, the combination mutant shows a 4.5 fold improvement in expression over WT (Table V).

Structural Analysis of Stabilizing Mutations

Structural analysis suggests that the stabilizing core mutations primarily fulfill their roles through expected mechanisms. I82M and V101I both fill voids within the protein core (Figures 4A and B). Analysis of homologous structures shows that the equivalent to I82M appears in the structures of endoglucanase D from *Clostridium cellulovorans* (3NDY) and celCCA from *Clostridium cellulolyticus* (1EDG). The equivalent to V101I appears in almost every homologous structure examined with most of the remaining structures alternatively containing a leucine (V101L greatly reduces activity in *Hj*Cel5A, see Chapter 3). In all of these structures, these bulkier sidechains occupy more space within the protein core and contribute to better hydrophobic packing.

Eight of the nine stabilizing helix mutations appear to reduce the inherent dipole. T80E, N155E, and N155Q may possibly form bonds to the unsatisfied hydrogen bond donors at the N-terminus of the helix (Figure 4C, E, and F). Homologous structures lack equivalents to T80E and N155Q. N155E, however, appears in the structures of *Thermotoga maritima* Cel5A (*Tm*Cel5A) (PDB ID 3MMW [29]), *C. cellulovorans* endoglucanase D, an endoglucanase from *Prevotella byrantii* (PDB ID 3VDH, to be published), and a *Bacillus sp.* alkaline cellulase K (PDB ID 1G0C [30]). Hydrogen bonding to an N-terminal amine is only observed in alkaline cellulase K. The residue adopts a solvent-exposed conformation in the remaining structures. S133R, T156E, G239E, S318E, and S318Q likely adopt solvent-exposed conformations (Figure 4D, G, H, J, and K). A residue shifted by one position in *Tm*Cel5A (PDB ID 3MMW [29]) resembles the S318E mutation. While T156E and G239E lack homologous counterparts, equivalents to the remaining residues do not form contacts with the protein. These negatively charged sidechains likely confer stability by improving the global charge balance along the helix rather than forming specific contacts. Finally, Y278F eliminates the unsatisfied OH at the tip of the sidechain (Figure 4I). This mutation arose as an artifact of the calculation and does not appear to alter the electrostatics of the helix. Interestingly, only one stabilizing mutation was recovered from the C-terminal end of the helix, supporting observations from previous studies demonstrating that the N-terminus is a more fruitful target for stabilization efforts [31].

Activity of Stabilized Mutants

Useful enzyme mutations not only confer stability, but also elevate or preserve activity. We tested the activity of each stabilizing point mutation at 60 °C for two hours on Avicel, a crystalline cellulose powder (Table V). While the two stabilizing core mutants have activities comparable to WT (Figure 3B), the helix mutations show an even distribution between lower and higher activities (Figure 3C). Previous studies suggest that extremely rigid core structure near active site residues can dramatically reduce enzymatic activity [32, 33]. The low number of mutations identified in this study, however, precludes attempts to concretely discern any such patterns.

Expression

In addition to preserving activity, desirable mutations will also maintain or enhance protein expression levels. Five of the helix dipole stabilizing hits on the activity screen were either neutral or destabilizing. Four of these mutations confer greater protein expression than WT in *S. cerevisiae*, the probable cause for their high activity on the screen. In general, the point mutants and helix combination mutant demonstrated large increases in expression level (Table V). We expressed the catalytic domain of *Hj*Cel5A

V101I in *Escherichia coli* and found a three-fold increase in protein yield. This result suggests that, in at least one case, expression may increase due to a structural changes rather than DNA level improvements (i.e. codon optimization). In the case of the helix dipole stabilizing mutations, we surmise that the mutations assist helix folding, reducing the time required for protein synthesis and resulting in higher protein yield. This hypothesis, however, remains untested.

Comparing Strategies

As implemented in this study, helix dipole stabilization appears to outperform the core repacking strategy. In stabilizing the helices, nine positive mutations were retrieved from 44 predictions yielding an accuracy of 20%. Thirty-two predictions generated from the core repacking calculation, however, revealed only two stabilizing mutations for an accuracy of 6%. Although some mutations recovered from the helix dipole method decrease enzymatic activity, the remaining mutations are more numerous and stabilizing than the core mutations. Moreover, many of the helix constructs labeled as negatives on the initial activity screen retained some degree of activity over a BSA standard. This observation stands in stark contrast with the core mutations, most of which dramatically reduce activity on the screen. Thus, it appears that successful stabilization strategies should preferentially target helices, not the core of the protein.

Location of Stabilizing Mutations

We hypothesized that helix dipole stabilizing positions are more amenable to mutation due to their location in boundary and surface positions. Analysis of a 262-member *Hj*Cel5A point mutation database described in Chapter 6 reveals that most of the stabilizing mutations appear in the boundary and surface regions (Figures 6A and B). Core positions occupy 47% of the positions in *Hj*Cel5A, yet only 19% of the stabilizing mutations from the database appear in this region. Meanwhile, surface and boundary regions contain 53% of the positions in the protein, yet house 81% of the stabilizing mutations. The average ΔT_{50} of mutations in the core, boundary, and surface regions is 2.0, 1.3, and 0.9 °C, respectively. While it appears that core mutations are more stabilizing on average, both the core and boundary in *Hj*Cel5A contain four highly stabilizing mutations ($\Delta T_{50} > 2$ °C). The surface contains one highly stabilizing mutation, S318P, yet this mutation provided the second largest thermostability benefit ($\Delta T_{50} = 3.4$ °C) within the 262-member dataset. We surmise that the set of stabilizing core mutations is small and overrepresented with highly stabilizing mutations because most core mutations are destabilizing. The surface and boundary regions, however, contain a higher number of moderately stabilizing mutations. These results indicate that more prudent thermostabilization strategies should attempt to uncover mutations from all protein regions.

4.4 Summary and Conclusions

In the course of this study, we sought to detect stabilizing *Hj*Cel5A mutations using two computational methods: 1) core repacking and 2) helix dipole stabilization. These efforts revealed a total of eleven weakly stabilizing mutations. Neither strategy proved exceptionally effective in producing a highly thermostable variant of our target cellulase. For example, combining six of these mutations into a single molecule produced only a minimal increase in ΔT_{50} of 2.4 °C. This improvement, however, is similar to the modest elevations reported in some cellulase stabilization projects [34, 35].

Our experiments and analysis of the 262-point mutant database additionally show that most stabilizing mutations occur within the surface or boundary regions of the protein. Although the core repacking and helix dipole stabilization calculations target specific areas within *Hj*Cel5a, the other calculations used to predict database mutations evaluate all protein regions. Additionally, the computational methods here poorly model solvation and generally perform better at modeling hydrophobic interactions. As such, we do not believe the 262-point mutant database is biased towards predicting surface mutations.

It is possible that stabilizing core mutations, although rare, confer a considerable level of stabilization to the protein. The results of our screen demonstrate that most core mutations affect activity in a detrimental manner. However, several highly stabilizing mutations in the 262-member point mutant database reside within the core region. Many of these mutations did not appear in our predictions as their corresponding design positions were excluded from the calculation. In most of these cases, these residues sat too close to critical catalytic residues or formed sidechain contacts with the protein backbone. Assuming sufficient computational resources, future calculations should strive to include all core positions in the design process.

Despite the modest increases in thermostability observed for the combination mutant, the strategies explored here may benefit future protein engineering efforts. Recent protein stabilization efforts have noted a marked decrease in expression and solubility among designed enzymes [8]. Conversely, many of the mutations identified within this study improve protein yield. Twelve of the tested mutations increased expression levels over WT. Moreover, the combination mutant exhibits a 4.5 fold improvement in yield. While the mechanisms underlying these elevated expression levels remain unexplored, the design strategies provided here may prove useful for rescuing yield following more successful stabilization efforts. The design methods showcased in this study may also supplement stabilization efforts using homologous sequence data. As the computational methods solely rely on structural information to detect stabilizing mutations, the set of predicted mutations may dramatically differ from those identified through consensus design. Indeed, four of the stabilizing mutations recovered in this study have no homologous counterparts among currently solved crystal structures and all of the mutations (with the exception of Y278F) show extremely low conservation scores (see Chapter 6). In addition, none of the stabilizing mutations identified in this study were predicted from examining homologous sequences (see Chapter 3).

The results presented in this study demonstrate that most of the currently known stabilizing mutations in *Hj*Cel5A reside beyond the relatively immutable core with many of the highly stabilizing mutations evenly dispersed between the boundary and the core. In addition, the helix dipole stabilization method identified seven more stabilizing mutations than the core repacking strategy. With a relatively high WT T_m of 69.5 °C and a highly packed interior housing a hydrogen-bond rich active site, *Hj*Cel5A may have already evolved near-optimal core stability. Future experiments using less thermostable and more loosely packed proteins cores may reveal whether this trend is universally applicable.

4.5 Materials and Methods

Classification of Residues

Residue classification as core, boundary, or surface was performed through first drawing a solvent-accessible surface around the protein structure, then calculating residue-surface distances. In this commonly used method, a Connolly dot surface [36, 37] is drawn by rolling a spherical probe with an 8 Å radius along the van der Waals spheres of the accessible C_α atoms. A vector following the trajectory along the C_α-C_β bond is then extended toward the surface of the protein. The C_α-surface and C_β-surface distances determine the residue classification:

Core: C_α-surface ≥ 5 Å and C_β-surface ≥ 2 Å
Surface: C_α-surface + C_β-surface ≤ 2.7 Å
Boundary: all other residues

The classification calculation was performed using chain A the 2.1 Å *Hj*Cel5A structure (PDB ID 3QR3 [19]) optimized with 50 steps of gradient-based energy minimization using the Rosetta forcefield. Although the experiments detailed in this chapter employ an energy function similar to the DREIDING forcefield, a larger subset of mutations in the 262-member database of *Hj*Cel5A point mutations were predicted using the Rosetta forcefield.

Structure preparation

Designs were performed on chain A of the *Hj*Cel5A crystal structure (PDB ID 3QR3 [19]). After removing water molecules and ions, hydrogens were added to the structure using the protein process application within the design software TRIAD [38]. This application was additionally employed to optimize the structure through 50 steps of gradient-based energy minimization using the energy function described in the computational design section.

Computational Design

Computational design parameters outlined in this section were kept consistent between the core repacking and helix dipole stabilization calculations. All calculations were executed using an energy function based on the DREIDING forcefield [39] that includes terms for van der Waals [7], hydrogen bonding [40], electrostatics [40], implicit solvation, and phi-psi propensities. In calculating the implicit solvation term, an occlusion-based solvation potential was applied with scale factors of 0.05 for nonpolar burial, 2.5 for nonpolar exposure, and 1.0 for polar burial [41]. Sequence optimization was performed with FASTER [42, 43] and a Monte Carlo-based algorithm was used to sample sequences near the minimum energy sequence [44, 45].

Cel5A Plasmid Construction

See equivalent section in Chapter 3.

Thermostability/Activity Screen

See equivalent section in Chapter 3.

T_{opt} Assay

To assess the optimal operating temperature of *Hj*Cel5A constructs, 40 µL of protein at a concentration of 0.25 µM was combined with 60 µL of a 1.5% Avicel slurry in cellulase buffer in a PCR plate in triplicate for each of 12 temperatures. The plates were incubated at 60 °C for two hours and promptly cooled for 5 minutes on ice. After centrifugation for 5 minutes to pellet the insoluble substrate, activity was assessed with the Park-Johnson assay using a 25 µL sample volume. Bovine serum albumin (BSA) at a final concentration identical to the protein of interest served as a negative control.

Park-Johnson Assay

See equivalent section in Chapter 3.

Enzyme Purification

See equivalent section in Chapter 3.

T_{50} Assay

See equivalent section in Chapter 3.

Single-Point Activity Assay

See equivalent section in Chapter 3.

4.6 Tables and Figures

Table I. *Calculation energies*

Calculation	WT	Design
Core, bda-bbind 1.0, All residues	-2800.42	-2940.74
Core, bda-bbind 1.0, Hydrophobic	-2863.85	-2930.73
Core, bbind02.May.e0, All residues	-2816.35	-2951.88
Core, bbind02.May.e0, Hydrophobic	-2875.19	-2924.27
Core, bbind02.May.e2, All residues	-2793.61	-3003.28
Core, bbind02.May.e2, Hydvrophobic	-2946.75	-2849.25
Helix 1, All residues	-2630.10	-2817.06
Helix 1, Strict scheme	-2630.10	-2819.48
Helix 2, All residues	-2630.10	-2815.90
Helix 2, Strict scheme	-2630.10	-2781.00
Helix 3, All residues	-2630.10	-2813.01
Helix 3, Strict scheme	-2630.10	-2803.47
Helix 4, All residues	-2630.10	-2830.32
Helix 4, Strict scheme	-2630.10	-2811.18
Helix 5, All residues	-2630.10	-2773.96
Helix 5, Strict scheme	-2630.10	-2754.44
Helix 6, All residues	-2630.10	-2782.16
Helix 6, Strict scheme	-2630.10	-2800.89
Helix 7, All residues	-2630.10	-2844.91
Helix 7, Strict scheme	-2630.10	-2835.69
Helix 8, All residues	-2630.10	-2789.91
Helix 8, Strict scheme	-2630.10	-2784.59

Table II. *Predicted core repacking mutation energy differences from WT*

		All Residue Types			Hydrophobic Only		
Mutation	**Max MI**	**bda-bbind 1.0**	**bbind02.May.e0**	**bbind02.May.e2**	**bda-bbind 1.0**	**bbind02.May.e0**	**bbind02.May.e2**
V7T[a]	0.37	-	-	N/A[a]	-	-	-
I9L	1.01	-13.87	-	-	-15.09	-	-14.88
A10S	0.51	-8.45	-8.68	-8.40	-	-	-
L31I	-[b]	-5.44	-5.92	-5.95	-5.88	-5.92	-6.30
L61C[a]	0.56	N/A[a]	-	-	-	-	-
V69L	0.86	-	-	-	-6.14	-	-
V69M	0.86	-	-	-10.06	-	-	-
V69N	0.86	-	-12.56	-	-	-	-
I82M	0.48	-6.14	-6.68	-	-	-	-
I82Q[a]	0.48	-	-	N/A[a]	-	-	-
V89M[a]	0.36	-	-	-	-	-	-
V89L	0.36	-	-	-	-	-	-0.30
V101I	0.29	-2.99	-4.04	-	-3.31	-3.28	-1.58
A107N	0.69	-13.79	-13.61	-13.62	-	-	-
F143M	0.13	-16.20	-11.43	-14.12	-	-	-
I145V[a]	0.74	-	-	-	-	-	-
V161L[a]	1.11	N/A[a]	-	-	-	-	-
V161I	1.11	-	-	-	-0.13	-	-
V165I	0.54	-	-	-	-0.91	-	-
A188C[a]	0.45	N/A[a]	-	-	-	-	-
F191W[a]	0.62	-	-	N/A[a]	-	-	-
V217I	0.94	-	-	-	-	-3.74	-3.54
V217L	0.94	-	-	-	-7.37	-	-
L221N	0.32	-8.43	-9.56	-9.63	-	-	-
A255C	0.53	-13.89	-	-	-	-	-
A255T	0.53	-	-14.72	-14.12	-	-	-
I256M[a]	0.61	-	-	N/A[a]	-	-	-
L257I	0.41	-	-4.22	-3.33	-	-4.42	-4.45
I276M	0.66	-7.30	-7.37	-	-	-	-
L319M	-[b]	-0.83	-	-	-	-	-
L324M	-[b]	-	-	-3.20	-	-	-
L324F	-[b]	-	-	-	-2.35	-	-

[a] Mutations were not predicted from a second pass calculation meant to calculate energies for individual mutations.

[b] Insufficient homologous sequence data.

Table III. *Helix dipole stabilization design positions*

Helix Number	WT N-Cap	Potential N-Cap	Disallow Positive	Disallow Negative	Float
1	42	-	43, 44, 45	51, 52, 53	30, 31, 32, 40, 41, 46, 47, 48, 49, 50, 54, 55, 56, 59, 87, 88, 95, 97, 314, 316, 326, 327
2	78	-	79, 80, 81	92, 93, 94	47, 59, 61, 67, 68, 69, 70, 76, 77, 82, 83, 84, 85, 86, 88, 89, 90, 91, 95, 96, 97, 98, 99, 138, 140, 141
3	120	-	121, 122, 123	132, 133, 134	65, 73, 74, 82, 86, 114, 115, 118, 119, 124, 125, 126, 127, 128, 129, 130, 131, 135, 136, 138, 143, 160, 163, 167, 168, 171, 173
4	153	-	154, 155, 156	168, 169, 170	115, 116, 128, 132, 143, 145, 151, 152, 157, 158, 159, 160, 164, 165, 166, 167, 171, 172, 173, 174, 175, 176, 178,183, 194, 195, 196, 199, 205, 206, 211, 212, 213
5	194	-	196, 197, 198	190, 199, 200, 201	154, 157, 158, 161, 180, 182, 183, 184, 190, 191, 192, 193, 195, 196, 197, 202, 203, 210, 215, 253
6	-	234, 236	237, 238, 239, 241, 242, 243	248, 249, 250, 251	188, 192, 215, 217, 219, 235, 237, 238, 239, 240, 244, 245, 246, 247, 257, 275, 278, 279, 252, 253, 254, 255, 282, 284, 285, 286
7	-	265	266, 267	278, 279, 280, 281	0, 1, 2, 230, 231, 232, 237, 241, 242, 245, 257, 262, 263, 264, 269, 270, 271, 272, 273, 274, 275, 276, 277, 282, 283, 286, 288, 289, 318, 319, 323, 326
8	317	-	318, 319, 320	320, 321, 322	50, 54, 56, 261, 262, 263, 264, 265, 268, 269, 272, 304, 305, 306, 307, 315, 316, 322, 323, 324, 325, 319, 269, 316, 317, 318, 327-

Table IV. *Predicted helix mutation energy differences from WT*

Mutations	Max MI	All Residues	Strict Scheme	Location
V51R[a]	0.93	-	-12.02	Core
N52R	1.11	-15.84	-15.56	Surface
E53R	0.75	-20.89	-16.64	Boundary
S79E	0.61	-	-10.86	Surface
S79Q	0.61	-13.51	-	Surface
T80E[c]	0.73	-	-3.46	Surface
T80Q	0.73	-9.85	-	Surface
S94R	0.72	-11.82	-11.69	Surface
T120S	-[b]	-3.11	-3.10	Surface
N121E	1.00	-	-9.82	Boundary
A122E	0.51	-	-2.84	Surface
A122Q	0.51	-9.08	-	Surface
S133R[c]	0.88	-9.95	-13.79	Surface
K134R	0.90	-10.00	-10.02	Boundary
I154M	-[b]	-	-	Boundary
N155E[c]	0.69	-	-3.01	Surface
N155Q[c]	0.69	-9.76	-	Surface
T156E[c]	0.98	-	-3.74	Boundary
I168H	0.23	-	-0.61	Boundary
N170R	0.64	-13.75	-13.79	Surface
A197M	1.00	-	-	Core
A197F	1.00	-7.14	-	Core
A199V	1.08	-	-	Boundary
S201Q	-[b]	-9.28	-	Boundary
S201K	-[b]		-	Boundary
D238E	0.92	-	-3.85	Surface
D238Q	0.92	-9.78	-	Surface
G239E[c]	0.96	-	-8.79	Boundary
G239Q	0.96	-13.89	-	Boundary
S242D	0.65		-9.81	Boundary
S242Q	0.65		-	Boundary
P243E	0.90		-	Boundary
P243Q	0.90	-11.04	-	Boundary
Q250R	0.86	-8.11	-11.81	Surface
S267Q	0.98	-6.64	-	Boundary
Y278F[c]	0.64	-14.62	-	Boundary
Y278L	0.64		-	Boundary
N280R	0.98	-5.95	-9.28	Boundary
Q281R	0.82	-14.19	-12.12	Surface
S318E[c]	-[b]		-7.30	Boundary
S318Q[c]	-[b]	-8.65	-	Boundary
S321K	1.02		-13.17	Boundary
S321R	1.02	-9.70	-	Boundary
S322R	0.96	-13.41	-15.68	Boundary

[a] Mutations were not predicted from a second pass calculation meant to calculate energies for individual mutations.

[b] Insufficient homologous sequence data.

[c] Stabilizing mutation

Table V. *Characterization of stabilizing mutations*

Construct	$T_{50,WT}$ (°C)	$T_{50,mut}$ (°C)	ΔT_{50} (°C)	Activity (µM Cellobiose Equivalents)	ΔActivity (µM Cellobiose Equivalents)	ΔΔG[b] (kcal mol^{-1})	Site MI[b]	Expression Level Relative to WT
WT	-	-	-	193.7±12.2	0	-	-	-
Core Mutations								
I82M	69.5±0.2	69.8±0.5	0.3±0.5	188.5±2.4	-5.2	-0.9	0.48	1.3
V101I	69.6±0.3	70.1±0.2	0.5±0.4	210.1±1.8	16.4	-0.4	0.29	1.7
Helix Mutations								
T80E	69.3±0.2	69.8±0.1	0.5±0.2	203.6±9.2	9.8	2.7	0.73	2.3
S133R	68.9±0.1	69.4±0.1	0.4±0.2	197.1±2.6	3.4	1.8	0.88	1.8
N155E	69.5±0.3	70.0±0.1	0.5±0.3	199.4±1.1	5.6	0.5	0.69	4.9
N155Q	68.4±0.1	68.5±0.1	0.1±0.1	172.8±4.0	-20.9	0.1	0.69	1.1
T156E	69.5±0.2	69.7±0.3	0.2±0.3	217.6±7.9	23.9	1.2	0.98	4.9
G239E	69.7±0.1	70.0±0.3	0.2±0.3	216.9±6.5	23.2	-0.2	0.96	1.0
Y278F	69.2±0.2	70.2±0.4	1.0±0.5	174.7±2.7	-19.1	1.3	0.64	0.4
S318E	69.7±0.2	70.5±0.1	0.9±0.2	244.1±4.3	50.4	-0.5	N/A[a]	0.6
S318Q	68.9±0.1	69.4±0.2	0.5±0.2	196.0±2.4	2.3	-0.3	N/A[a]	0.9
S79Q	69.4±0.2	69.5±0.2	0.0±0.3	174.4±5.1	-19.3	0.4	0.61	1.4
S79E	69.9±0.2	69.7±0.0	-0.1±0.2	N/A	N/A	-0.3	0.61	5.5
T80Q	69.2±0.2	69.1±0.1	-0.1±0.2	N/A	N/A	3.2	0.73	2.0
A122E	69.0±0.5	68.8±0.2	-0.2±0.5	N/A	N/A	1.0	0.51	3.2
G239Q	69.1±0.1	68.5±0.1	-0.9±0.2	N/A	N/A	-1.6	0.96	0.9
Helix Combo	69.5±0.5	71.9±0.3	2.4±0.5	N/A	N/A	N/A	N/A	4.5

[a] Insufficient homologous sequence data.

[b] Values calculated from the 444-sequence multiple sequence alignment described in Chapter 3.

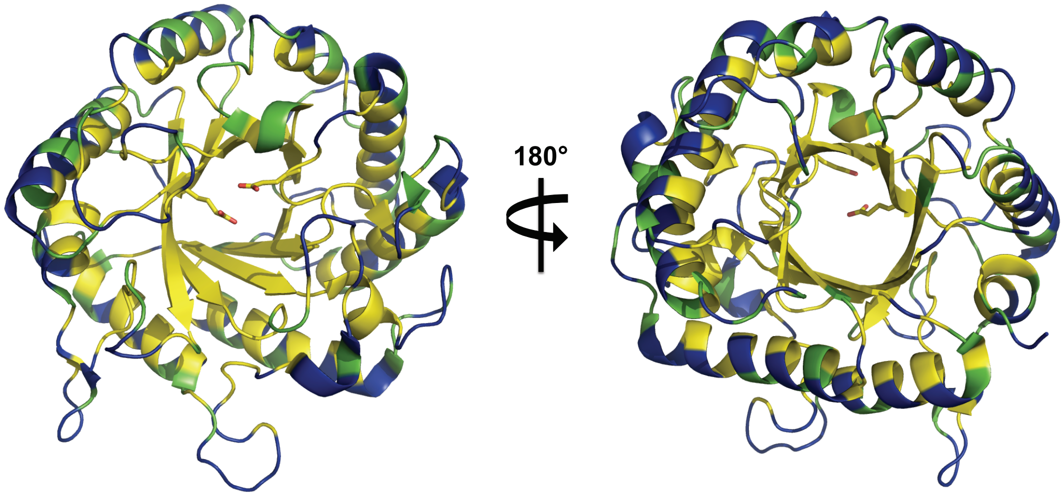


Figure 1. *Residue classification in HjCel5A.* Areas within the crystal structure of HjCel5A are color coded to highlight the core (yellow), boundary (green), and surface (blue). The active site (left) and a 180° rotation to display the non-catalytic face (right) are both displayed. The two catalytic carboxylates E148 and E259 appear as sticks.

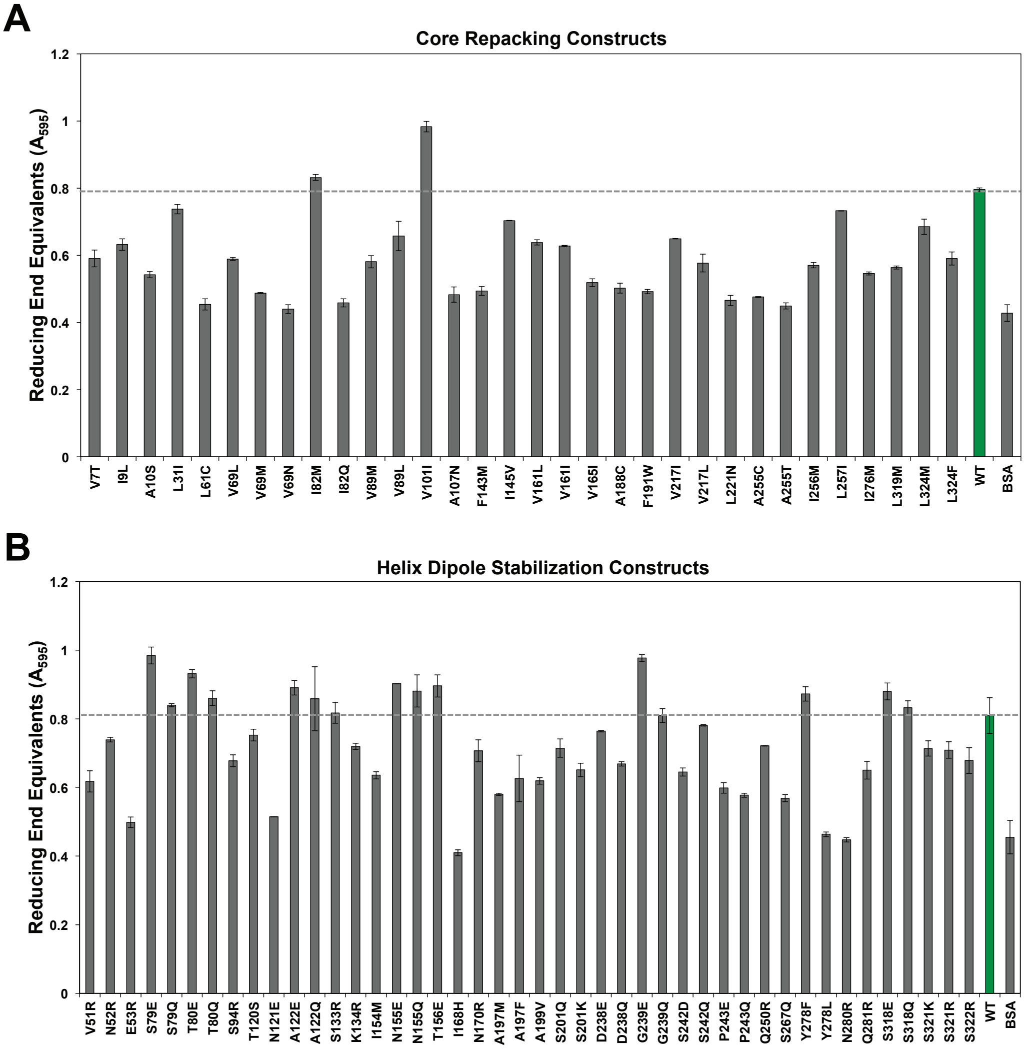


Figure 2. *Activity screens.* Screens performed on *Hj*Cel5A point mutants identified from the (A) core repacking and (B) helix dipole stabilization calculations. WT is highlighted in green. Variants with activity exceeding that of WT, indicated by the dashed line, were purified and tested for thermostability.

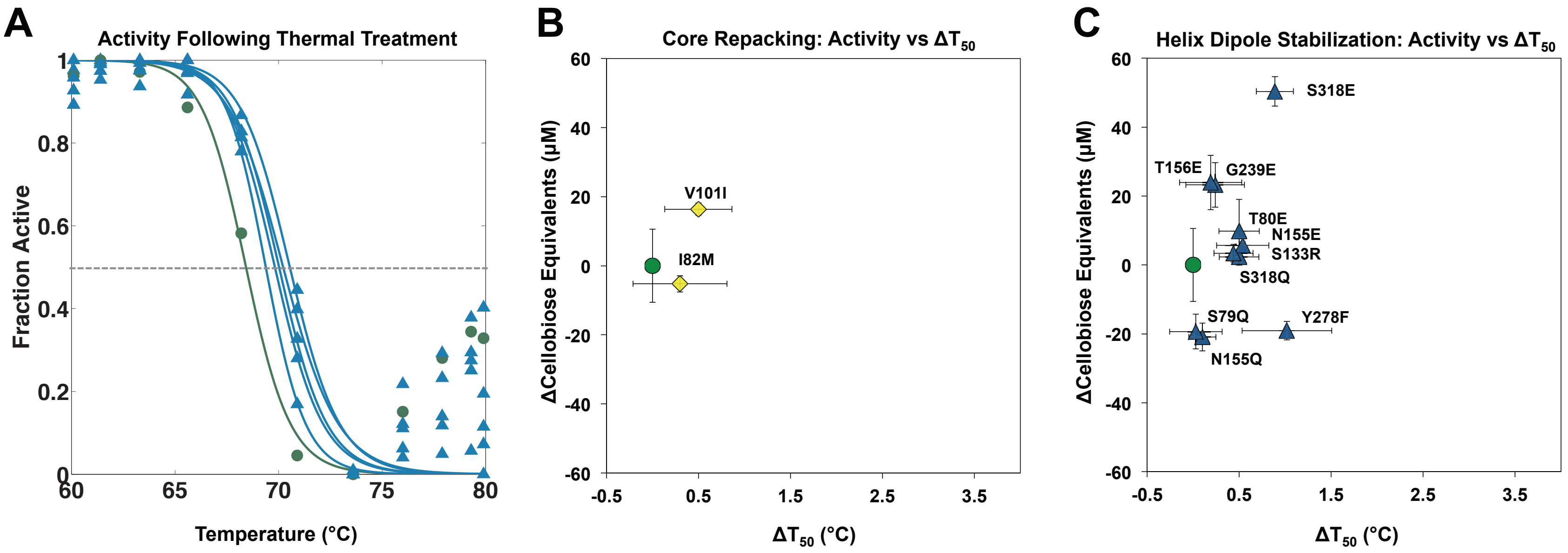


Figure 3. *Thermostability and activity of point mutations.* (A) Activity of *Hj*Cel5A point mutants after treatment over a range of temperatures. Data are plotted for WT (green circles) and mutations conferring a $\Delta T_{50} > 0.5$ °C (blue diamonds). All plotted mutations were identified from the helix dipole stabilization calculation. The dashed line indicates the point at which 50% of the initial activity persists (T_{50}). Although each mutant was tested alongside WT to accurately assess (ΔT_{50}), only one WT trial is displayed for clarity. The recovery of activity at higher temperatures is due to refolding caused by PCR plate edge effects. See Supplementary Figure 1 for additional data. (B,C) Activity versus ΔT_{50} for the core (yellow diamonds) and the helix (blue triangles) mutations.

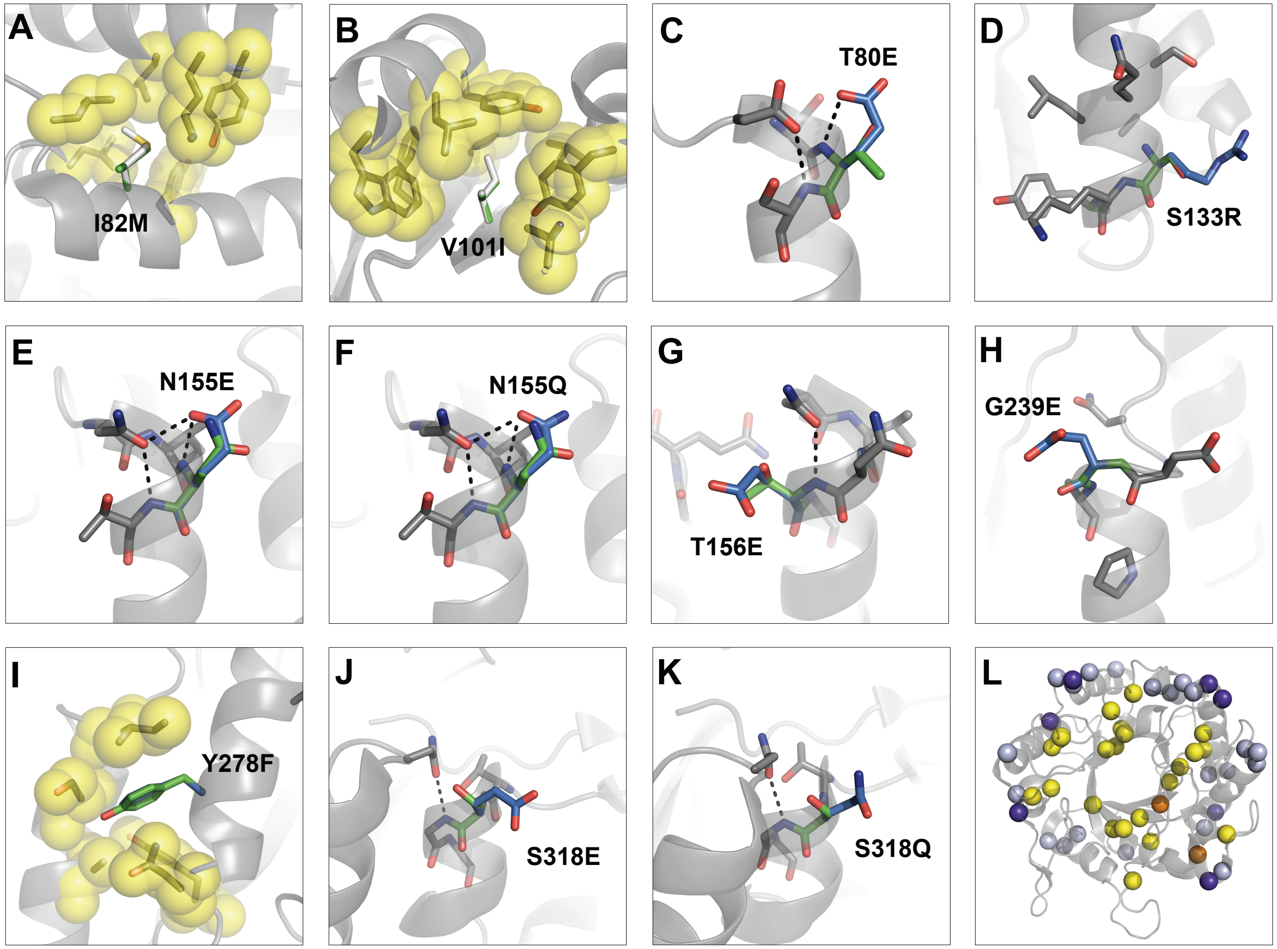


Figure 4. *Structural analysis of stabilizing point mutations.* Panels A-K show each of the stabilizing mutations along with the WT residue (green) Core mutations and helix mutations are shown as white and blue sticks, respectively. Residues around space filling mutations are depicted as yellow spheres. (L) The location of core (yellow) and helix mutations (light blue) predicted as stabilizing. Mutations confirmed as stabilizing are shown in orange (core) and dark blue (helix).

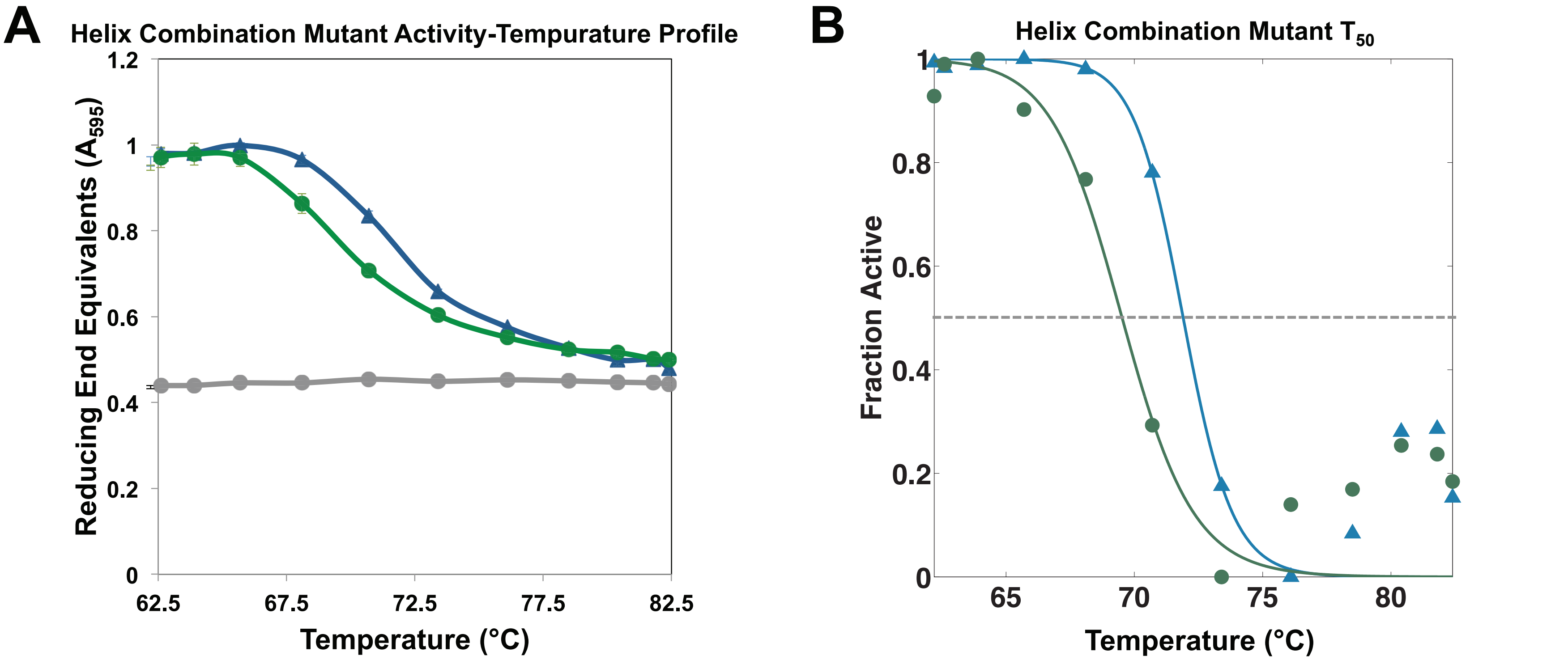


Figure 5. *T_{opt} and T_{50} of the helix combination mutant.* (A) The activity of the helix combination mutant from a 2 hour incubation across a temperature gradient from 62.5 to 82.5 °C. WT is shown as green circles, the combination mutant as blue triangles and a BSA standard as gray circles. (B) The activity from a 1 hour incubation at 60 °C following a 10 minute preincubation at a gradient of temperatures from 60 to 80 °C. Curves for WT (green circles) and the helix combination mutant (blue triangles) are displayed. The dotted line marks the point at which half maximal activity persists (T_{50}).

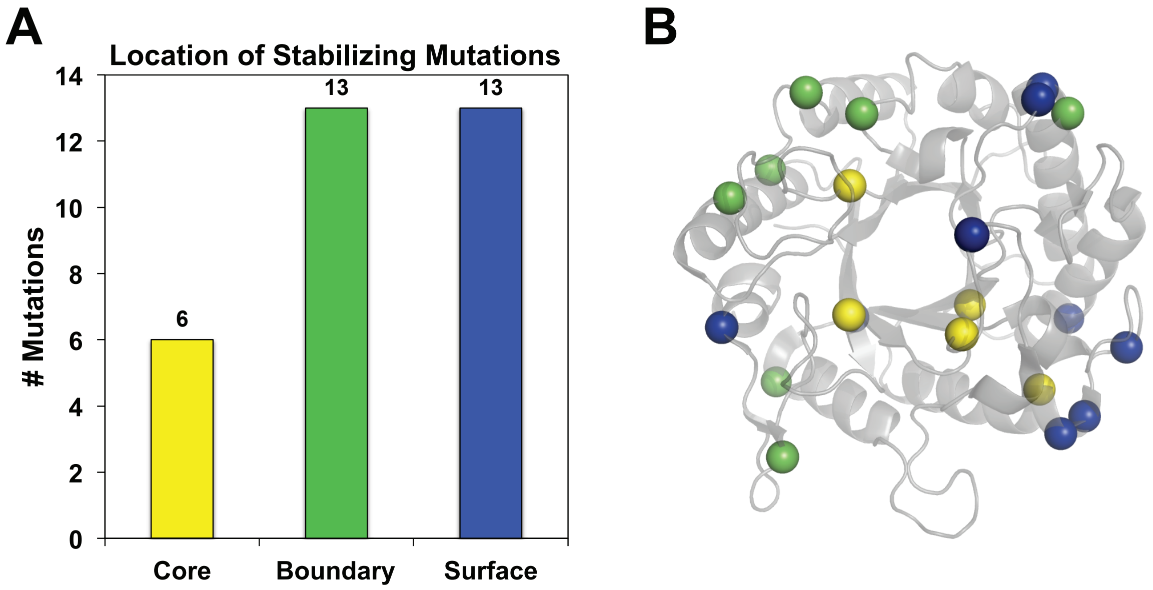


Figure 6. *Location of stabilizing point mutations.* (A) The distribution of stabilizing mutations from a 262 *Hj*Cel5A point mutation database. (B) Placement of stabilizing core (yellow), boundary (green), and surface (blue) mutations in the *Hj*Cel5A structure. Several stabilizing mutations occur at the same position.

4.7 Supplementary Figure

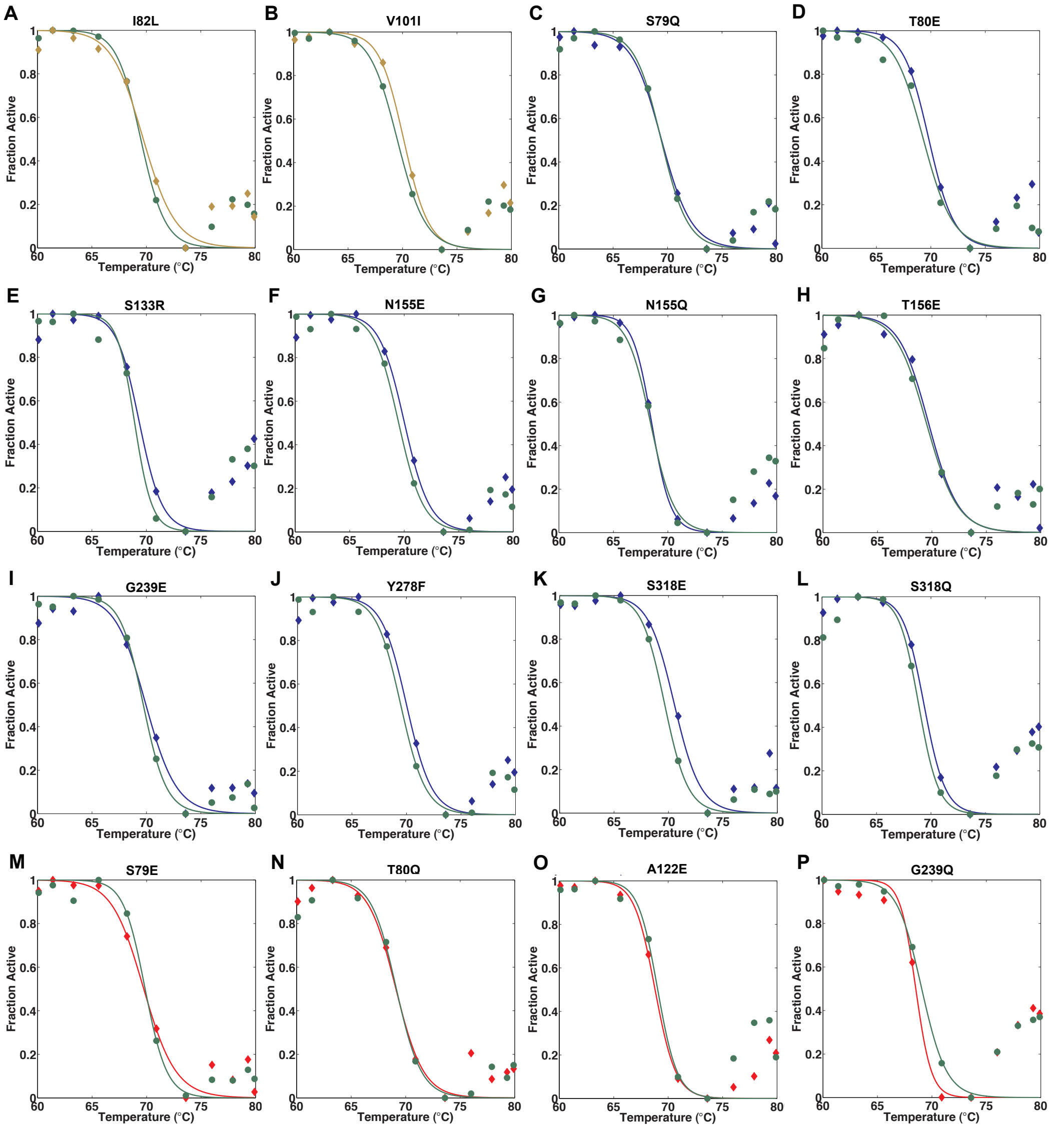


Figure S1. *T$_{50}$ plots of all tested point mutations.* Panels show the activity of *Hj*Cel5A point mutants after treatment over a range of temperatures. WT is represented as green circles in all panels. (A-B) Data for core mutations are depicted as yellow diamonds. (C-L) Data for stabilizing helix dipole mutants are shown in blue. Panel C shows data for a mutation with stability similar to WT. (M-P) Data for destabilizing helix dipole mutants.

4.8 References

1. Daniel, R.M., Cowan D., A., Morgan H.W. and Curran M.P. (1982) A correlation between protein thermostability and resistance to proteolysis. Biochemical Journal 207:641-644.
2. Wu, I. and Arnold F.H. (2013) Engineered thermostable fungal Cel6A and Cel7A cellobiohydrolases hydrolyze cellulose efficiently at elevated temperatures. Biotechnology and Bioengineering 110:1874-1883.
3. Viikari, L., Alapuranen M., Puranen T., Vehmaanperä J. and Siika-aho M. Thermostable enzymes in lignocellulose hydrolysis. In: Olsson L, Ed. (2007) Biofuels. Springer Berlin Heidelberg, pp. 121-145.
4. Pace, C.N. (1990) Conformational stability of globular proteins. Trends in biochemical sciences 15:14-17.
5. Akasako, A., Haruki M., Oobatake M. and Kanaya S. (1997) Conformational Stabilities of *Escherichia coli* RNase HI variants with a series of amino acid substitutions at a cavity within the hydrophobic core. Journal of Biological Chemistry 272:18686-18693.
6. Borgo, B. and Havranek J.J. (2012) Automated selection of stabilizing mutations in designed and natural proteins. Proceedings of the National Academy of Sciences 109:1494-1499.
7. Dahiyat, B.I. and Mayo S.L. (1997) Probing the role of packing specificity in protein design. Proceedings of the National Academy of Sciences 94:10172-10177.
8. Dantas, G., Kuhlman B., Callender D., Wong M. and Baker D. (2003) A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. Journal of Molecular Biology 332:449-460.
9. Yun, H., Park H., Joo J. and Yoo Y. (2013) Thermostabilization of *Bacillus subtilis* lipase A by minimizing the structural deformation caused by packing enhancement. Journal of Industrial Microbiology & Biotechnology 40:1223-1229.
10. Isupov, M.N., Fleming T.M., Dalby A.R., Crowhurst G.S., Bourne P.C. and Littlechild J.A. (1999) Crystal structure of the glyceraldehyde-3-phosphate dehydrogenase from the hyperthermophilic archaeon *Sulfolobus solfataricus.* Journal of Molecular Biology 291:651-660.
11. Wallon, G., Kryger G., Lovett S.T., Oshima T., Ringe D. and Petsko G.A. (1997) Crystal structures of *Escherichia coli* and *Salmonella typhimurium* 3-isopropylmalate dehydrogenase and comparison with their thermophilic counterpart from *Thermus thermophilus*. Journal of Molecular Biology 266:1016-1031.
12. Knapp, S., Kardinahl S., Hellgren N., Tibbelin G., Schäfer G. and Ladenstein R. (1999) Refined crystal structure of a superoxide dismutase from the hyperthermophilic archaeon *Sulfolobus acidocaldarius* at 2.2 Å resolution. Journal of Molecular Biology 285:689-702.
13. Petsko, G.A. (2001) Structural basis of thermostability in hyperthermophilic proteins, or "There's more than one way to skin a cat". Methods in Enzymology 334:469-478.

14. Eijsink, V.G.H., Bjørk A., Gåseidnes S., Sirevåg R., Synstad B., Burg B.v.d. and Vriend G. (2004) Rational engineering of enzyme stability. Journal of Biotechnology 113:105-120.

15. Malakauskas, S.M. and Mayo S.L. (1998) Design, structure and stability of a hyperthermophilic protein variant. Nature Strucutural Biology 5:470-475.

16. Marshall, S.A., Morgan C.S. and Mayo S.L. (2002) Electrostatics significantly affect the stability of designed homeodomain variants. Journal of Molecular Biology 316:189-199.

17. Joo, J.C., Pohkrel S., Pack S.P. and Yoo Y.J. (2010) Thermostabilization of *Bacillus circulans* xylanase via computational design of a flexible surface cavity. Journal of Biotechnology 146:31-39.

18. Suominen, P.L., Mantyla A.L., Karhunen T., Hakola S. and Nevalainen H. (1993) High frequency one-step gene replacement in *Trichoderma reesei*. II. Effects of deletions of individual cellulase genes. Molecular Genetics and Genomics 241:523-530.

19. Lee, T.M., Farrow M.F., Arnold F.H. and Mayo S.L. (2011) A structural study of Hypocrea jecorina Cel5A. Protein Science 20:1935-1940.

20. Dunbrack Jr, R.L. and Karplus M. (1993) Backbone-dependent rotamer library for proteins application to side-chain prediction. Journal of Molecular Biology 230:543-574.

21. Lassila, J.K., Privett H.K., Allen B.D. and Mayo S.L. (2006) Combinatorial methods for small-molecule placement in computational enzyme design. Proceedings of the National Academy of Sciences 103:16710-16715.

22. Shortle, D., Stites W.E. and Meeker A.K. (1990) Contributions of the large hydrophobic amino acids to the stability of staphylococcal nuclease. Biochemistry 29:8033-8041.

23. Matthews, B.W. (1993) Structural and Genetic Analysis of Protein Stability. Annual Review of Biochemistry 62:139-160.

24. Tokuriki, N., Stricher F., Schymkowitz J., Serrano L. and Tawfik D.S. (2007) The stability effects of protein mutations appear to be universally distributed. Journal of Molecular Biology 369:1318-1332.

25. Walter, S., Hubner B., Hahn U. and Schmid F.X. (1995) Destabilization of a protein helix by electrostatic interactions. Journal of Molecular Biology 252:133-143.

26. Hol, W.G.J., Van Duijnen P.T. and Berendsen H.J.C. (1978) The α-helix dipole and the properties of proteins. Nature 273:443-446.

27. Lockhart, D.J. and Kim P.S. (1992) Internal stark effect measurement of the electric field at the amino terminus of an α-helix. Science 260.

28. Aurora, R. and Rose G.D. (1998) Helix capping. Protein Science 7:21-38.

29. Pereira, J.H., Chen Z., McAndrew R.P., Sapra R., Chhabra S.R., Sale K.L., Simmons B.A. and Adams P.D. (2010) Biochemical characterization and crystal structure of endoglucanase Cel5A from the hyperthermophilic *Thermotoga maritima*. Journal of Structural Biology 172:372-379.

30. Shirai, T., Ishida H., Noda J.-i., Yamane T., Ozaki K., Hakamada Y. and Ito S. (2001) Crystal structure of alkaline cellulase K: insight into the alkaline adaptation of an industrial enzyme. Journal of Molecular Biology 310:1079-1087.

31. Doig, A.J. and Baldwin R.L. (1995) N and C-capping preferences for all 20 amino acids in alpha-helical peptides. Protein Science 4:1325-1336.

32. Yoshida, Y., Ohkuri T., Kino S., Ueda T. and Imoto T. (2005) Elucidation of the relationship between enzyme activity and internal motion using a lysozyme stabilized by cavity-filling mutations. Cellular and Molecular Life Sciences CMLS 62:1047-1055.

33. Lee, C., Park S.-H., Lee M.-Y. and Yu M.-H. (2000) Regulation of protein function by native metastability. Proceedings of the National Academy of Sciences 97:7727-7731.

34. Németh, A., Kamondi S., Szilágyi A., Magyar C., Kovári Z. and Závodszky P. (2002) Increasing the thermal stability of cellulase C using rules learned from thermophilic proteins: a pilot study. Biophysical Chemistry 96:229-241.

35. Heinzelman, P., Komor R., Kanaan A., Romero P., Yu X., Mohler S., Snow C. and Arnold F. (2010) Efficient screening of fungal cellobiohydrolase class I enzymes for thermostabilizing sequence blocks by SCHEMA structure-guided recombination. Protein Engineering Design and Selection 23:871-880.

36. Connolly, M. (1983) Analytical molecular surface calculation. Journal of Applied Crystallography 16:548-558.

37. Connolly, M.L. (1993) The molecular surface package. Journal of molecular graphics 11:139-141.

38. Triad. (2012). Protabit LLC, Pasadena, CA.

39. Mayo, S.L., Olafson B.D. and Goddard W.A. (1990) DREIDING: a generic force field for molecular simulations. The Journal of Physical Chemistry 94:8897-8909.

40. Dahiyat, B.I., Gordon D.B. and Mayo S.L. (1997) Automated design of the surface positions of protein helices. Protein Science 6:1333-1337.

41. Chica, R.A., Moore M.M., Allen B.D. and Mayo S.L. (2010) Generation of longer emission wavelength red fluorescent proteins using computationally designed libraries. Proceedings of the National Academy of Sciences 107:20257-20262.

42. Allen, B.D. and Mayo S.L. (2006) Dramatic performance enhancements for the FASTER optimization algorithm. Journal of Computational Chemistry 27:1071-1075.

43. Desmet, J., Spriet J. and Lasters I. (2002) Fast and accurate side-chain topology and energy refinement (FASTER) as a new method for protein structure optimization. Proteins: Structure, Function, and Bioinformatics 48:31-43.

44. Metropolis, N., Rosenbluth A.W., Rosenbluth M.N., Teller A.H. and Teller E. (1953) Equation of state calculations by fast computing machines. Journal of Chemical Physics 21:1087-1092.

45. Voigt, C.A., Gordon D.B. and Mayo S.L. (2000) Trading accuracy for speed: a quantitative comparison of search algorithms in protein sequence design. Journal of Molecular Biology 299:789-803.

CHAPTER 5

Identifying Stabilizing Mutations in *Hypocrea jecorina* Cel5A Through ΔΔG Approximations (FoldX, Triad) and Backbone Stabilization

This chapter is formatted for submission to the Journal of Molecular Biology.

5.1 Abstract

Computational methods of detecting stabilizing mutations fall into three main categories: 1) physical, 2) knowledge-based, and 3) empirical. Physical methods rely on molecular and/or quantum mechanical calculations, requiring considerable computational resources. Likewise, knowledge-based methods utilize large amounts of pre-existing information in databases. In recent years, empirical methods have gained popularity due to their ease of use, rapid calculations, and broad applicability. Typically, the procedure only requires a molecular model of sufficient resolution to forecast accurate predictions. Here, we use two empirical software packages, FoldX and Triad, to identify eleven stabilizing mutations in the primary endoglucanase from *Hypocrea jecorina* (*Hj*Cel5A), an industrial target for thermostabilization. These results and analysis of a 262-point mutation database demonstrate that while FoldX outperforms the Rosetta-based method, each method yields unique stabilizing mutations. The computational protein design program Triad with the rosetta forcefield was additionally used to explore the possibility of stabilizing *Hj*Cel5A through reducing its entropy of unfolding, i.e., stabilizing the protein backbone. Restricting calculations to detect Gly → X_{AA} and X_{AA} → Pro mutations, uncovered eight additional stabilizing mutations. Many of the Triad ΔΔG, glycine, and proline mutations decreased activity due to altering residues near a substrate binding groove or possibly restricting flexibility necessary for function. These results lead to the recommendation that future stabilization efforts primarily use FoldX supplemented with Triad if additional mutations are necessary.

5.2 Introduction

Thermostable protein variants provide a multitude of benefits over less stable alternatives. In addition to demonstrating improved resistance to various forms of degradation [1, 2], thermostable proteins allow experiments or industrial processes to operate at otherwise intractable conditions. For example, reactions proceeding at elevated temperatures exhibit lower levels of microbial contamination and reduced solution viscosity [2]. Provided they remain folded, some thermostable enzymes also often demonstrate an increase in activity with rising temperatures [3-5] as modeled by the Arrhenius equation. As such, much interest exists in developing reliable methods of engineering thermostable versions of desired proteins.

Generally, three classes of computational methods for predicting thermostable protein mutations exist: 1) physical, knowledge-based, and empirical [6]. In the physical scheme, predictions are calculated based on molecular and/or quantum mechanical approximations of the free energy of unfolding. These methods require considerable computational resources and a high level of user expertise, but generally yield highly accurate results. Knowledge based methods generate predictions using information from databases filtered by selection criteria. This material may include DNA sequences [7], protein sequences (see Chapter 3) [8, 9], and protein thermostability data (ProTherm) [10, 11]. Such methods are rapid and require little user training, yet rely on the preexistence of large amounts of data. The empirical approach combines information gleaned from databases to tune molecular mechanics potentials. In this strategy, calculations proceed relatively quickly and the empirical data need not originate from the system of interest, broadening the applicability of the technique.

The wide appeal of empirical methods has resulted in the emergence of multiple competing techniques. Popular empirical prediction software packages include Dmutant [12], CUPSAT [13], PopMuSic-2.0 [14], I-Mutant2.0 [15], and FoldX [16]. Among these tools, FoldX (and Dmutant) was shown to perform the most reliably with an accuracy of ~60% [17]. FoldX employs an energy function with terms for van der Waals effects, hydrogen bonding, electrostatics, and solvation, calibrated to closely approximate experimentally derived ΔΔG values [16, 18]. This software module has played a role in constructing thermostable variants of tumor necrosis factor-related apoptosis-inducing ligand [19], the YtvA LOV domain [20], and Cel7A from *Hypocrea jecorina* [5]. Recently, some groups have also employed the Rosetta protein design software to predict ΔΔG values of point mutations [21, 22]. The forcefield from Rosetta (rosetta) includes terms for van der Waals interactions and solvation effects, with hydrogen bonding terms weighted by evidence from high-resolution protein structures [23].

While empirical methods can identify stabilizing mutations throughout a protein structure, they may also target specific regions. One popular strategy involves designing molecules with more rigid backbones to reduce the entropy of unfolding, often through mutating glycines to residues containing a Cβ [24], introducing prolines [24], adding disulfides [25, 26], or targeting residues with high B-factors [27]. These structural alterations restrict the area of Ramachandran space available to each residue, reducing the entropy of unfolding. The observation that thermostable proteins are often enriched with prolines [4, 28, 29] lends further credence to this stabilization strategy. Moreover, application of this method has already uncovered thermostable mutations in a methyl parathion hydrolase [30] and bacteriophage T4 lysozyme [24].

In this chapter, we use FoldX and the computational protein design software Triad (using the Rosetta forcefield) to predict stabilizing mutations in an industrial target, the primary endoglucanase (*Hj*Cel5A) from *Hypocrea jecorina* (anamorph *Trichoderma reesei*) [31]. This enzyme is used in the alternative biofuels industry to hydrolyze cellulosic material into fermentable sugars [32]. In addition to identifying ten unique stabilizing mutations, our efforts also compare the efficacy of the FoldX and Triad ΔΔG stabilization methods. Finally, we employ the Rosetta forcefield (referred to here as rosetta) to predict mutations that stabilize the protein backbone through Gly → X_{AA} and X_{AA} → Pro mutations, revealing eight additional stabilizing mutations.

5.3 Results and Discussion

FoldX Designs

We performed *in silico* site saturated mutagenesis on chain A from the *Hj*Cel5A crystal structure (PDB ID 3QR3). From 6232 possible point mutations, FoldX predicted 1008 as stabilizing ($\Delta\Delta G < 0$ kcal mol^{-1}). As this sizeable pool of candidate mutations was too large to screen using available resources, we chose to examine mutations with a $\Delta\Delta G \leq -1.75$ kcal mol^{-1}. In addition, a study seeking to stabilize an exoglucanase from *H. jecorina* using FoldX-guided mutagenesis determined that tightening the energy cutoff from -0.75 to -1.75 kcal mol^{-1} improved reliability by 45% [5]. Applying this stringent criterion reduced the candidate pool to a manageable 43 mutations.

Triad ΔΔG Designs

In parallel to the FoldX designs, we performed site-saturated mutagenesis using Triad with the Rosetta forcefield and the *Hj*Cel5A crystal structure. Out of 6232 possible point mutations, 789 were predicted as stabilizing ($\Delta\Delta G < 0$). To reduce the number of candidate mutations to a manageable quantity, an arbitrary -1.75 kcal mol^{-1} cutoff was applied leaving 47 mutations.

Detecting Stabilizing Mutations

Point mutants for each of the 43 FoldX and 44 Triad mutations were constructed for secretion from *Saccharomyces cerevisiae*. Proteins were expressed and the supernatant was screened for activity at 73 °C, 3.5 degrees higher than the melting temperature (T_m) of the native protein. Six FoldX and five Triad mutants demonstrated higher activity than wild type (WT) and were selected for more rigorous characterization.

As the screen was performed on unpurified protein in supernatant, hits may indicate a variant with improved thermostability, yield, activity, or a combination thereof. To determine the source of their improvement, the mutants and WT were simultaneously assayed for activity at 60 °C following a 10 minute incubation at a gradient of temperatures ranging from 60 – 80 °C. The T_{50}, the temperature at which half of the maximum activity persists, of the mutation was computed and compared to the WT value. All of the mutations proved more thermostable than WT ($\Delta T_{50} > 0$ °C) (Table I). FoldX predicted three highly stabilizing mutations, D271F (ΔT_{50} = 3.1 °C), D271Y (ΔT_{50} = 2.7 °C), and S318P (ΔT_{50} = 3.2 °C). Similarly, three mutations among the Triad predictions, K219A (ΔT_{50} = 2.0 °C), K219Q (ΔT_{50} = 2.8 °C), and S309F (ΔT_{50} = 2.7 °C), conferred significant increases in stability. The mutant expression levels were similar or lower than WT.

Structural Analysis of Stabilizing FoldX and Triad Mutations

The eleven stabilizing mutations detected using FoldX or Triad appear to enhance structural integrity through several means. The two mutations to proline, S79P (Figure 2A) and S318P (Figure 2F), appear at the N-terminus of helices after the N-capping position (Ncap + 1). Prolines in this position lead to more stable structures for three reasons [33]: 1) proline's rigid pyrrolidine sidechain locks the N-terminus in a helical conformation through restricting backbone flexibility, 2) Ncap + 1 prolines reduce the number of unpaired hydrogen bond acceptors at the N-terminus, and 3) the reduced requirement for hydrogen bonding partners to the N-terminus facilitates favorable interactions between the two residues preceding proline. Two mutations appear to improve electrostatic contacts within the protein. The N153D mutation strengthens the pre-existing Asn N-cap by increasing the negative charge around the partially positive N-terminal helix region (Figure 2B). The shorter side chain introduced through the K219Q mutation likely reduces conformational entropy while maintaining hydrogen bonds to both N255 and N236 (Figure 2H). Interestingly enough, the K219A mutation confers a large thermostability benefit, but lacks these two hydrogen bonds (Figure 2G). This observation suggests that the K219 sidechain may be unsuited for its environment and that the reduction in entropy achieved through mutating the site to an alanine is sufficiently beneficial to overcome the loss of these two contacts. The remaining mutations provide stability through filling surface pockets. D217F and Y fill a cavity created by a surface-exposed loop (Figures 2C and D), while S309L, F, and W improve packing between a disordered loop [34] at the end of a β-hairpin and an α-helix (Figures 2E, I, and J).

Six out of the eleven stabilizing mutations contain no equivalents within homologous structures currently deposited in the Protein Data Bank (PDB). Residues equivalent to S79P, N153D, and K219Q appear in *Thermoascus aurantiacus* structures (PDB ID 1GZJ [35] and 1H1N [36]). The RBcel1 [37] and *Bacillus agaradhaerens* Cel5A (PDB ID 7A3H [38]) structures contain a proline in a similar position to S318P in *Hj*Cel5A. In addition, an alanine at the position equivalent to 219 in *Hj*Cel5A appears in the *Piromyces rhinzinflatus* Eg1A structure (PDB ID 3AYR [39]). The remaining mutations (D271F/Y, S309L/F/W) appear at regions with significant structural differences in available homologous structures. The β-hairpin harboring position 309 and the loop near position 271 appear to be unique to *Hj*Cel5A.

Activity of Stabilizing FoldX and Triad Mutants

Thermostable enzymes provide little benefit if activity decreases. To determine whether the stabilizing FoldX and Triad mutations impact activity, purified protein was assayed for activity after 2 hours at 60 °C (Table III, Figures 1C and 1F). While the Rosetta mutations show decreased activity compared to WT, the FoldX mutations exhibit more diversity in activity. Notably, the FoldX mutations S318P and D271Y elevate the WT activity by 12.8 and 16.8%.

Predictive Efficacy of FoldX and Triad

Based on the thermostability data presented in this section, FoldX appears slightly more efficient at recovering stabilizing and sufficiently active mutations than the Triad-based strategy. Six stabilizing FoldX mutations were recovered from 43 candidates for a predictive accuracy of 14.0%. In comparison, five stabilizing Triad mutations emerged from 47 candidates (10.6% accuracy). Moreover, the FoldX mutations not only show larger thermostability benefits, but also seem more adept at preserving activity than the Triad mutations. These results, however, originate from small datasets and may not apply to more general cases.

To more thoroughly compare the performance of FoldX and Triad, receiver operator characteristic (ROC) curves calculated from a 262 *Hj*Cel5A point mutation dataset (See

Chapter 6) were generated (Figure 3). ROC curves plot the fraction of true positives from the predicted positives against the fraction of false positives from the true negatives across a range of acceptance thresholds [40]. Each point on the curve corresponds to a single cutoff value ranging from loose (all mutations are predicted, corresponding to the upper right corner of the graph) to extremely stringent (no mutations are predicted, corresponding to the lower left corner). If a metric can discriminate between desired and unwanted members of a set with some level of accuracy, the area under the curve (AUC) will exceed zero. For the purposes of this study, we define the AUC as the area between the curve and the diagonal, setting the maximum AUC possible to 0.5. The FoldX and Rosetta curves yield AUCs of 0.16 and 0.09, respectively, demonstrating that while both measures can identify thermostable mutations with some accuracy, FoldX provides a slight advantage. The thresholds generating the greatest ratio of true positives to false positives are 0.3 kcal mol^{-1} for FoldX and 0.5 kcal mol^{-1} for Triad.

Although FoldX performs with the greatest accuracy, Triad provides additional highly stabilizing mutations. Only one mutual prediction, S309L, appeared both calculations, indicative of low redundancy between the two methods. Additionally, many of the Triad candidates have high FoldX ΔΔG scores. FoldX predicts K219Q to be highly destabilizing (ΔΔG of 1.28 kcal mol^{-1}), yet this mutation proved most stabilizing out of all of the Triad predictions. Given this evidence, future thermostabilization projects might consider utilizing both methods to maximize the number and diversity of positive hits.

Backbone Stabilization: Removing Glycines and Adding Prolines

Adapting empirical methods to improve backbone stability may provide additional stabilizing mutations. To test this hypothesis, all Gly → X_{AA} and X_{AA} → Pro mutations were fetched and ranked by ΔΔG value. All of the mutations with ΔΔG values ≤ 0 kcal mol^{-1} were designated as potentially stabilizing. This relaxed cutoff allowed for the prediction of mutations that did not pass the -1.75 kcal mol^{-1} threshold enforced in the general Triad ΔΔG calculation. Only five mutual members appear in both the general Triad and glycine mutation lists. In addition, all predicted proline mutations were previously uncharacterized.

In total, 51 glycine and 46 proline mutations were predicted as stabilizing with four (G64P, G144P, G239P, and D316P) appearing in both lists (Tables IV and V). Due to screening constraints, only the top 44 glycine (Figure 4A) and 46 proline (Figure 4D) mutants were constructed and screened using the setup described for the FoldX and Rosetta constructs. Nine glycine and three proline mutants demonstrated higher activity than WT on the screen. After purifying these enzymes and determining their ΔT_{50}s, five glycine (G189A ΔT_{50} = 0.4 °C, G189S ΔT_{50} = 1.2 °C, G239D ΔT_{50} = 0.4 °C, G239N ΔT_{50} = 0.7 °C, and G293A ΔT_{50} = 3.5 °C) and three proline mutations (T18P ΔT_{50} = 2.0 °C, N76P ΔT_{50} = 2.0 °C, and S139P ΔT_{50} = 2.0 °C) demonstrated thermostability benefits (Table VI, Figures 4B and E).

Structural analysis supports the notion that these stabilizing mutations improve thermostability through restricting backbone movement. With the exception of G293A, all of the stabilizing glycine and proline mutations sit on loop regions (Figures 5A-D and 5F-I), areas that tend to exhibit higher conformational flexibility compared with well-ordered secondary structure elements. G293A appears to improve activity at high temperatures by fixing W292, a residue necessary for substrate binding [34], in a catalytically competent configuration (Figure 5E). T18P and S139P both appear in slightly distorted type I β-turns at position i+1, the most commonly observed location for prolines [41].

As was observed for the FoldX and Triad mutations, several of the glycine and proline mutations have no equivalents in homologous structures. The *T. aurantiacus* Cel5A structures contain residues corresponding to the G189A and S139P mutations. Residues equivalent to G239N appear in the *Ba*Cel5A and *Thermobifida fusca* Cel5A (PDB ID 2CKS, unpublished data) structures. The remaining mutations show no homology to currently available crystal structures, supporting the assertion that this empirical computational method may provide additional stabilizing mutations to supplement homology studies.

Although eight stabilizing mutations were identified, most caused decreases in enzymatic activity (Table VI, Figures 4C and F). Only G189S (+14.5 µM) and G293A (+27.3 µM) showed an improvement in hydrolysis. While the majority of mutations marginally impacted catalytic performance, S139P reduced the amount of cellobiose released by 41.6 µM, a decrease of 21% from the WT output. It is possible that increasing protein rigidity may improve thermostability with a tradeoff in enzymatic activity. Recent studies have documented functional impairments in stabilized or more inflexible variants of 3-isopropylmalate dehydrogenase [42], HIV-1 protease [43], snake venom metalloproteases [44], rendering this explanation a possibility. Finally, results from this section and the more general survey reveal that mutations at positions 189, 219, and 239 generally decrease activity. These positions reside near the substrate-binding pocket and may adversely affect catalytic function (Figure 6).

Backbone Stabilization: Disulfide Engineering

While removing glycines and introducing prolines provides modest stability benefits, incorporating stabilizing disulfide linkages constitutes one of the most effective means of reducing the entropy of protein unfolding. Engineered disulfide bonds improved the thermostability of several proteins by several degrees including *Drosophila melanogaster* acetylcholinesterase [45], a thermolysin-like protease [46], T4 lysozyme [47], and *Clostridium thermocellum* cellulase C [48]. We attempted to engineer disulfides into *HjCel5A* using two prediction methods: 1) Triad using the Rosetta forcefield and 2) the program Disulfide by Design [49]. The Triad relies on rotamer optimization to design new disulfides. Lenient disulfide bond geometries are employed that allow the program to recapture preexisting disulfides. Disulfide by Design identifies potential disulfide linkages through a simple geometry-based algorithm and is freely available through a web server. Three and fifteen mutations were predicted from Triad and Disulfide by Design calculations, respectively (Table VII). Only one construct (I44C-G91C) performed better than WT on the activity screen (Figure 7A). However, more rigorous analysis revealed that this construct demonstrated a ΔT_{50} of -0.5 °C (Table VI, Figure 7B).

Ample evidence exists that the addition of disulfide bonds between flexible regions can radically improve protein thermostability [45, 47, 50]. Yet, even studies reporting successful stabilization through disulfide engineering demonstrate that most attempts negatively impact folding and stability. This destabilization may occur even when disulfides form correctly, suggesting that future attempts should require a broad assessment of local structure that cannot be captured using the methods examined in this study. Though our efforts did not reveal any stabilizing disulfide bonds beyond the four present in the WT structure, more rigorous methods might provide better results.

5.4 Summary and Conclusion

This study explores the efficiency of employing computational ΔΔG approximations to recover active, stabilizing mutations in an industrially-relevant endoglucanase. When applied to *Hj*Cel5A, FoldX performs slightly better than Triad at predicting stabilizing mutations. These findings are reminiscent of Khan et al.'s comprehensive survey of ΔΔG-based methods wherein FoldX was deemed one of the most reliable predictors of stabilizing mutations [17]. In addition, the mutations recovered with FoldX show better retention of function than those predicted with Triad. Two of the Triad mutations fall in a cleft important to substrate binding and subsequently reduce activity. Our calculations employed purely structural data with no special considerations for catalytically-important residues. As such, the difference in activity may stem from error caused by small datasets and may not represent a general trend.

The majority of mutations were predicted using a single strategy. In general, Triad appears to recover more mutations from the protein core while FoldX performs well on surfaces. This observation may stem from differences in the employed forcefields. The FoldX energy function contains a term to explicitly model the extra stabilizing free energy provided by a water molecule making more than one hydrogen bond to the protein (ΔG_{wb}) [16]. The rosetta forcefield used in Triad, however, represents solvent implicitly as a continuous medium using the method of Lazaridis and Karplus [51]. This difference may explain why mutations from each method skew towards different sectors of the protine and why little overlap between mutation pools predicted from FoldX and Triad exists. In future experiments, these strategies may be used in a complementary fashion to increase the number of stabilizing mutations recovered.

The accuracies achieved here (FoldX 14.0%, Rosetta 10.6%) do not approach the 60% value reported in the Khan study. Although the screen used in our study cannot identify stabilizing mutations that significantly lower expression or activity, it is unclear how useful such mutations might prove in a real-world application. Consequently, only stabilizing mutations that do not dramatically reduce expression or activity were classified as true positives. The low accuracies reported in this study may originate from the omission of these undesirable stabilizing mutations.

In addition to comparing the efficiency of FoldX and Triad, this study reports numerous unique stabilizing mutations in *Hj*Cel5A. The calculations recovered primarily non-overlapping sets of mutations, allowing for the discovery of six and five stabilizing mutations from FoldX and Triad, respectively. Seven of the mutations improve thermostability by ≥ 1 °C. Such increases are large when compared to previous studies (see Chapter 4). In addition, efforts to improve backbone stability revealed five Gly → X_{AA} and three X_{AA} → Pro stabilizing mutations. Many of these mutations slightly decrease enzymatic activity. With a handful of notable outliers such as G293A, most of the mutations also provide only modest stability improvements. Mirroring these results, an attempt to rationally introduce prolines into the i+1 position of type I β-turns within *Hj*Cel6A recovered one marginally stabilizing mutation out of ten candidates [4]. Previous reports have also noted that proline mutations improve stability in a manner that is highly dependent on local structure [52]. The methods employed in this study may fail to capture some structural feature common to stabilizing proline mutations. Taken together, these studies suggest that strictly focusing on backbone stabilization through decreasing the entropy of unfolding is not the most effective stabilization protocol. Nevertheless, the information provided in this report may be of use in an industrial setting.

Attempts to engineer new disulfide bonds in *Hj*Cel5A met with little success. As the native molecule already contains four well-formed bridges, it is possible that suitable positions for adding additional linkages do not exist. Adding more cysteines may also increase the likelihood of protein aggregation as free thiols may form unfavorable inter and intramolecular bonds. More likely, better search algorithms are needed to engineer stabilizing disulfides. Although all predicted constructs were experimentally characterized for completeness, many of the constructs predicted by Disulfide by Design and Triad could have been discarded based on chemical intuition.

Finally, several cautionary lessons relevant to applying ΔΔG-based stabilization methods to enzymes emerge from this study. While rendering the backbone more rigid may enhance stability, activity may consequently suffer. In addition, procedures that rely solely on structural information may alter areas necessary for function. Mutations at positions 189, 219, and 239 elevated stability, but appear near a binding pocket and detrimentally affect catalysis. Rational analysis of the *Hj*Cel5A crystal structure may have suggested against testing mutations in this region. In this study, however, all computationally predicted mutations were tested to remove experimenter bias. Future experiments need not take such precautions, possibly leading to improvements in engineering efficiency.

5.5 Materials and Methods

FoldX Calculation

All FoldX calculations were performed with version 3.0 [16]. After removing waters and ligands, Chain A from the *Hj*Cel5A crystal structure (PDB ID: 3QR3) was prepared using the optimize and repairPDB functions within the software. A position scan was performed to compute energy values for WT and mutations to all other 19 amino acids. To compute ΔΔG values, each mutation was compared to WT using the following equation:

$$\Delta\Delta G_{Mut} = \Delta G_{Mut} - \Delta G_{WT}$$

where ΔG_{mut} is the energy computed for the mutation and ΔG_{WT} is the energy computed for the WT residue at the same position. All calculations were performed using default parameters unless otherwise specified.

Triad ΔΔG Calculation

All Rosetta calculations were performed using a modified version of the rosetta energy function described by Rohl et al. [23] and implemented within the protein design software Triad [53]. The version of rosetta implemented in Triad employs a softer Lennard-Jones potential, a different set of amino-acid reference energies, and modified hydrogen bond and amino acid propensity weights. The energy function also lacks terms unnecessary for point mutation calculations including those for disulfide bonding, Ramachandran, proline closure, and omega tethering. Designs were performed on chain A of the *Hj*Cel5A crystal structure (PDB ID 3QR3 [34]). After removing water molecules and ions, hydrogens were added to the structure using the protein process application within the design software Triad [53]. Triad was additionally employed to optimize the structure through 50 steps of gradient-based energy minimization using the rosetta forcefield.

Glycine Scan

The glycine scan was performed in Triad using the modified version of the rosetta forcefield described in the Triad ΔΔG Calculation section (above). The scan mutates glycine in the native structure to each of the other 19 amino acids. All other conditions addressed in the rosetta calculation section apply to the glycine scan. The information calculated here is simply a subset of the data retrieved from the more comprehensive Triad-rosetta ΔΔG scan and is reformatted to facilitate data analysis.

Proline Scan

The proline scans were performed with a restricted version of the algorithm used for the original Triad ΔΔG scan calculation. This scheme calculates ΔΔG values for mutating every position in the protein to proline and provides a ranked list of mutations. As in the glycine calculation, the generated information is a reformatted subset of the data retrieved from the Triad ΔΔG scan.

Disulfide Bond Engineering Calculations

Disulfide bond engineering calculations were performed using the ssdesign application in the protein design software Triad [53]. In this application, if two cysteine (CSS) rotamers come in close contact, the program adopts smaller values for force constants and barriers for DREIDING bonds, angles, and torsions. This leniency has been optimized to detect native disulfides as many disulfide geometries show slight deviations from canonical values. To design disulfides, the rotamer optimization algorithm simultaneously switches a pair of residues to CSS rotamers. Pair moves are biased towards those with good pairwise energies, i.e., those likely to form disulfides. Calculations were performed with 7 trajectories, a rotamer pair factor of 10, an iterations multiplier of 5, a disulfide force constant of 15, a disulfide max benefit of 35, and a CSS penalty of 15. All calculations employed a version of the rosetta forcefield, as described in the Triad ΔΔG Calculation section (above), modified to employ the DREIDING disulfide bonding energy terms. The term used for bonds within disulfides is:

$$E_{bond} = 12700.0B_{1,2}(r_{1,2} - R_{1,2})^2$$

where B is the bond order (1, 1.5, 2, or 3), r is the Cartesian distance between atoms 1 and 2, and R is the equilibrium bond distance between atoms 1 and 2. The term for disulfide angles is as follows:

$$E_{angle} = 12100.0(\angle_{1,2,3} - \theta_{1,2,3})^2$$

where $\angle_{1,2,3}$ is the observed angle between atoms 1, 2, and 3 and $\theta_{1,2,3}$ is the equilibrium angle. Disulfide torsion angles are defined as:

$$E_{torsion} = 12K_{1,2,3,4}N(1 - d_{1,2,3,4}\cos(n_{1,2,3,4}\chi_{1,2,3,4})$$

where $K_{1,2,3,4}$ is the energy barrier, N is the number of torsion terms where atoms 2 and 3 are placed in the center, $d_{1,2,3,4}$ is the phase factor (1 for cis, -1 for trans), $n_{1,2,3,4}$ is the periodicity, and $\chi_{1,2,3,4}$ is the torsion angle.

Additional disulfide bond engineering calculations were performed using the online server for Disulfide by Design version 2.11 [49]. Calculations were executed on both chains in the *Hj*Cel5A crystal structure (PDB ID 3QR3). Disulfide bonds principally contain four atoms linked in a linear fashion: C_β-S_γ-S_γ-C_β. In this calculation, a disulfide model is generated with fixed C_β-S_γ (1.81 Å) and S_γ-S_γ (2.04 Å) bond lengths and C_β-S_γ-S_γ (104.15°) bond angles. To initiate the calculation, a pair of residues is chosen. The $\chi3$ torsion angle, formed through rotating the C_β about the S_γ-S_γ bond, is allowed to vary until the C_β-C_β distance matches that observed in the crystal structure. Energies (E_{ij}) are then calculated using the following equations:

$$E_{ij} = E\left(\chi_{1,i}\right) + E\left(\chi_{1,j}\right) + E\left(\theta_i\right) + E\left(\theta_j\right) \quad (1)$$

$$E\left(\chi_1\right) = 1.4[1 + \cos(3\chi_1)] \quad (2)$$

$$E\left(\chi_3\right) = 4.0[1 - \cos(2\chi_3 + 160)] \ (3)$$

$$E\left(\theta\right) = 55.0[\theta - \theta_0]^2 \, (4)$$

Where *i* and *j* are residue positions, θ is the C_α-C_β-S_γ angle, and $\theta_0 = 114.6°$. Energies are computed in kcal mol^{-1} with higher values corresponding to more favorable mutations. All calculations were performed with default settings.

Cel5A Plasmid Construction

Double mutants for disulfide engineering were constructed using a modified version of the Quikchange method in which two primer pairs are added to a single reaction. See equivalent section in Chapter 3 for additional details.

Thermostability/Activity Screen

See equivalent section in Chapter 3.

Park-Johnson Assay

See equivalent section in Chapter 3.

Enzyme Purification

See equivalent section in Chapter 3.

T_{50} Assay

See equivalent section in Chapter 3.

Single-Point Activity Assay

See equivalent section in Chapter 3

5.6 Tables and Figures

Table I. *Predicted FoldX mutations*

Mutation	ΔΔG (kcal mol^{-1})
R3P	-2.64
K32P	-2.50
D54A	-2.76
D54L	-5.29
D54R	-3.03
D54C	-3.14
D54M	-3.85
D54K	-2.54
D54N	-3.89
S79P	-1.84
T120D	-2.16
N153D	-2.91
Q186G	-2.95
A230P	-1.80
V265D	-1.87
S267P	-2.75
D271Y	-1.83
D271F	-2.23
S283P	-1.77
E305G	-2.24
E305A	-2.46
E305V	-2.18
E305I	-2.62
E305S	-1.77
E305C	-2.86
E305M	-4.09
E305K	-4.04
E305Q	-2.13
E305N	-1.88
E305F	-3.77
E305H	-2.91
S309L	-1.78
D316A	-1.91
D316P	-2.36
D316S	-2.20
D316C	-2.44
D316Q	-1.81
D316A	-1.91
S318P	-2.32
S318M	-2.09
S318W	-1.81
S318F	-1.94
S322L	-1.87

Table II. *Predicted Triad ΔΔG mutations*

Mutation	ΔΔG (kcal mol^{-1})
N8A	-2.53
N8V	-2.64
M56F	-2.3
R60V	-1.75
G112F	-2.04
G112L	-2.13
G112R	-2.03
G112W	-1.76
G112Y	-2.05
Q116D	-2.33
Q116N	-2.92
Q116W	-1.93
W142E	-2.06
W142F	-3.76
W142H	-3.38
W142I	-4.9
W142L	-2.62
W142M	-3.65
W142T	-2.6
W142V	-4.29
W142Y	-2.91
T156G	-1.77
Q186D	-3.13
Q186E	-2.99
Q186N	-2.05
K219A	-3.2
K219E	-1.82
K219S	-3.65
K219Q	-1.92
N236G	-1.89
I237F	-1.99
I237W	-2.17
I237Y	-2.06
R253Q	-2.11
I276H	-1.75
N282R	-1.93
N282Q	-1.97
E305F	-2.24
E305H	-3.07
E305L	-1.95
E305T	-2.14
E305Y	-2.13
S309F	-1.94
S309L	-2.26
S309W	-2.24
L324F	-2.35
L324H	-1.86

Table III. *Characterization of FoldX and Triad ΔΔG mutations*

Construct	$T_{50,WT}$ (°C)	$T_{50,mut}$ (°C)	ΔT_{50} (°C)	Activity (µM Cellobiose Equivalents)	ΔActivity (µM Cellobiose Equivalents)	FoldX ΔΔG (kcal mol^{-1})	Triad ΔΔG (kcal mol^{-1})	Expression Level (Mut/ WT)	Location
WT	-	-	-	193.7±12.2	0.0	-	-	-	
FoldX									
S79P	69.9±0.3	70.2±0.5	0.3±0.5	205.1±6.7	11.4	-1.84	0.97	1.0	Surface
N153D	70.3±0.7	70.7±0.6	0.5±0.9	196.2±1.5	2.5	-2.91	-0.25	0.2	Surface
D271F	70.5±0.9	73.6±0.6	3.1±1.1	167.2±6.3	-26.5	-2.23	-1.08	0.1	Boundary
D271Y	71.3±0.3	73.9±0.1	2.7±0.4	209.5±3.9	32.5	-1.83	-1.07	0.4	Boundary
S309L	71.3±0.3	72.7±0.3	1.5±0.3	196.1±1.9	2.4	-1.78	-2.26	0.8	Boundary
S318P	69.9±0.6	73.1±0.6	3.2±0.9	218.5±3.4	24.8	-2.32	3.27	0.3	Surface
Triad									
K219A	68.1±0.7	70.1±0.1	2.0±0.7	167.0±6.2	-26.7	2.07	-3.20	1.4	Core
K219Q	68.5±0.0	71.3±0.1	2.8±0.1	162.0±1.4	-31.7	1.28	-1.92	1.2	Core
S309F	68.3±0.1	71.0±0.1	2.7±0.1	164.1±8.6	-29.7	-0.57	-1.94	0.8	Boundary
S309L	71.3±0.3	72.7±0.3	1.5±0.3	196.1±1.9	2.4	-1.78	-2.26	0.8	Boundary
S309W	68.2±0.0	68.6±0.1	0.4±0.1	202.8±6.8	9.1	0.19	-2.24	0.5	Boundary

Table IV. *Predicted glycine mutations*

Mutation	ΔΔG (kcal mol^{-1})
G64P	-2.33
G64A	-0.78
G112K	-1.14
G112Q	-1.14
G112V	-1.05
G112T	-1.04
G112C	-1.04
G112N	-1.03
G112A	-0.84
G112I	-0.75
G112M	-0.60
G112E	-0.51
G112H	-0.21
G112S	-0.16
G112D	-0.09
G144A	-0.80
G144P	-0.31
G144D	-0.16
G144N	-0.09
G189A	-0.18
G189E	-0.99
G189H	-0.28
G189K	-0.52
G189N	-0.26
G189Q	-0.52
G189R	-0.34
G189S	-0.04
G239A	-0.80
G239C	-0.34
G239D	-0.46
G239I	-0.45
G239K	-0.56
G239L	-0.87
G239M	-0.40
G239N	-0.97
G239P	-0.18
G239R	-0.02
G239S	-1.05
G239T	-1.22
G239V	-0.36
G293A	-0.50
G311D	-0.55
G311N	-1.41
G328T	-0.31

Table V. *Predicted proline mutations*

Mutation	ΔΔG (kcal mol^{-1})
V2P	-0.44
N8P	-2.80
A10P	-0.52
F14P	-2.02
T18P	-0.58
V27P	-2.29
Y40P	-1.75
G64P	-2.33
N76P	-1.26
D78P	-1.97
D86P	-3.26
S94P	-0.48
A97P	-0.97
D102P	-1.27
H104P	-1.36
I114P	-1.91
T125P	-0.14
S126P	-0.51
S129P	-0.32
A136P	-0.81
S139P	-0.44
G144P	-0.31
N147P	-1.03
E163P	-0.93
R169P	-0.1
N170P	-1.31
Q176P	-1.25
S179P	-1.63
S187P	-1.18
S193P	-0.04
S201P	-1.37
N205P	-2.32
D222P	-0.72
S223P	-0.14
E231P	-0.56
G239P	-0.24
Q250P	-0.6
N264P	-1.18
I269P	-0.05
N280P	-0.73
Q281P	-0.83
T304P	-2.01
E305P	-0.05
D316P	-0.97
T317P	-0.88
A325P	-0.99

Table VI. *Backbone stabilizing mutations*

Construct	$T_{50,WT}$ (°C)	$T_{50,mut}$ (°C)	ΔT_{50} (°C)	Activity (μM Cellobiose Equivalents)	ΔActivity (μM Cellobiose Equivalents)	FoldX ΔΔG (kcal mol^{-1})	Triad ΔΔG (kcal mol^{-1})	Expression Level (Mut/WT)	Location
WT	-	-	-	193.7±12.2	0	-	-	-	
GLY → X_{AA}									
G189A	70.8±0.3	71.2±0.3	0.4±0.4	170.1±3.4	-23.7	-0.93	-0.76	1.2	Boundary
G189S	70.1±0.3	71.3±0.2	1.2±0.4	208.2±6.0	14.5	-0.45	-0.97	1.2	Boundary
G239D	70.8±0.1	71.2±0.1	0.4±0.2	185.6±3.1	-8.1	-0.80	-1.46	1.3	Boundary
G239N	70.4±0.1	71.1±0.0	0.7±0.1	174.9±4.3	-18.8	-0.13	-1.43	1.1	Boundary
G293A	70.3±0.1	73.7±0.1	3.5±0.2	221.0±1.2	27.3	6.66	-0.08	0.8	Core
G189E	70.6±0.1	70.6±0.2	0.0±0.2	N/A	N/A	-0.58	-1.01	2.0	Boundary
G64A	70.3±0.1	69.6±0.1	-0.7±0.1	190.5±5.0	-3.2	2.97	-0.30	2.3	Core
G189K	70.6±0.1	70.5±0.1	-0.1±0.2	N/A	N/A	-0.89	-0.55	1.2	Boundary
G239S	70.8±0.2	69.7±0.3	-1.0±0.3	N/A	N/A	-0.20	-1.04	0.9	Boundary
X_{AA} → PRO									
T18P	70.2±0.1	70.4±0.0	0.2±0.1	187.5±3.8	-6.2	-1.67	0.46	1.1	Surface
N76P	69.7±0.2	70.5±0.4	0.8±0.5	177.3±6.8	-16.4	0.00	0.95	1.7	Surface
S139P	69.6±0.2	71.5±0.6	1.8±0.6	152.1±8.0	-41.6	-1.33	2.14	1.8	Surface
Disulfide Mutations									
I44C/G91C	69.8±0.4	69.3±0.5	-0.5±0.7	N/A	N/A	N/A	N/A	0.8	Surface/ Boundary

Table VII. *Predicted disulfide mutations*

Mutation	ΔΔG (kcal mol^{-1})
Disulfide by Design	
G1C, N280C	1.70
G6C, I58C	0.37
F12C, V63C	0.15
D13C, H104C	1.34
D13C, P62C	1.93
I44C, G91C	2.80
G74C, Q123C	1.69
D78C, S81C	1.96
A132C, A173C	1.61
A136C, G172C	2.73
S139C, S175C	1.54
D184C, A190C	1.58
Q186C, H218C	0.65
N235C, Q274C	2.17
W292C, G293C	2.34
Triad	
P29C, L31C	-16.51
P29C, F34C	-21.47
L31C, F34C	-19.19

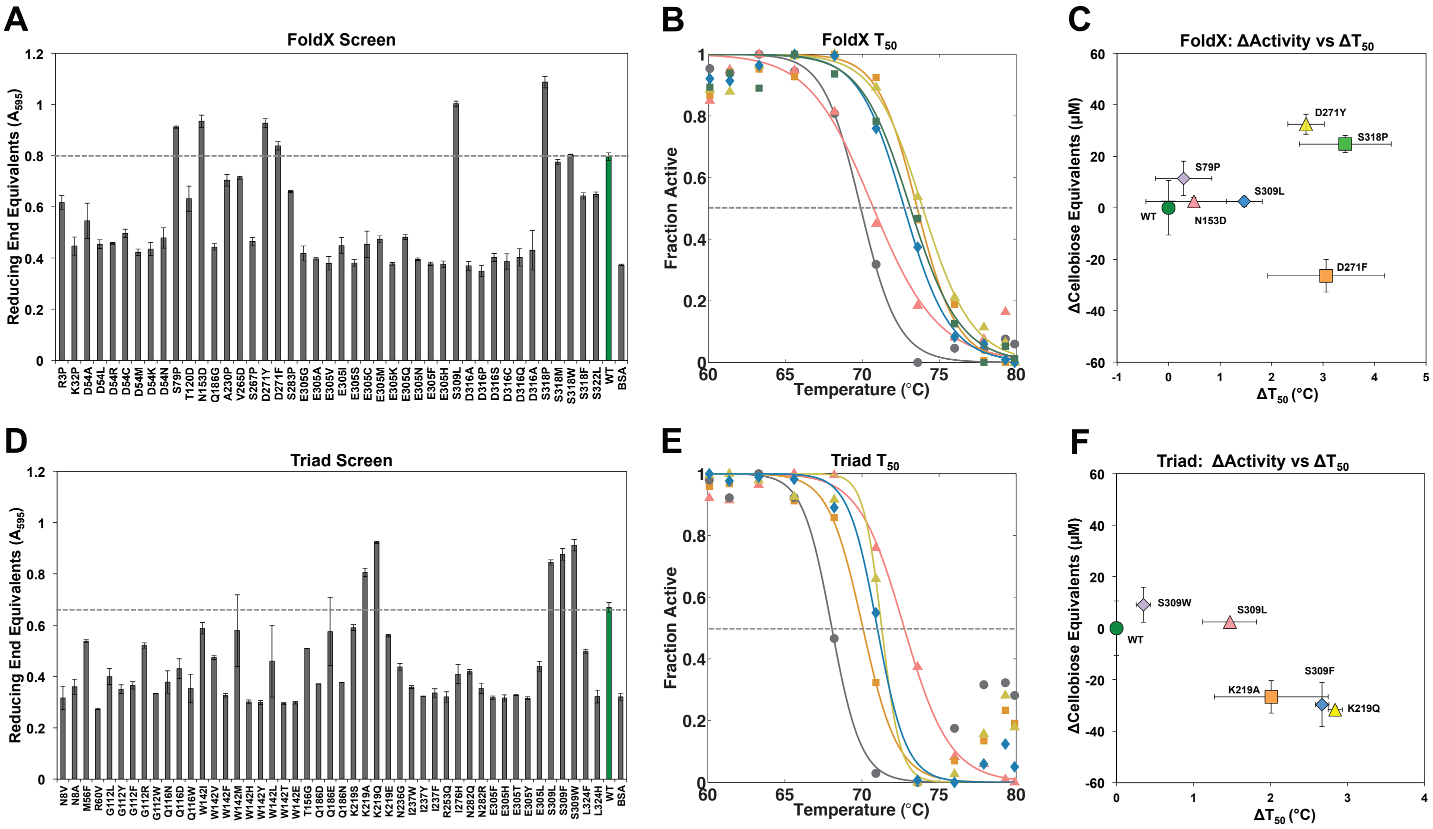


Figure 1. *Screening and characterization of FoldX and Triad mutants.* (A, D) Activity screen on FoldX (A) and Triad (D) mutants. WT is shown in green. The dotted line marks the WT activity level for comparison. (B, E) Activity of *Hj*Cel5A point mutants after treatment over a range of temperatures. The dotted line marks the point at which half of the original activity remains (T_{50}). (B) Data for the FoldX mutants N153D (pink triangles), D271F (orange squares), D271Y (yellow triangles), S309L (blue diamonds), and S319P (green squares) are plotted. (E) Data for the Triad mutants K219Q (orange squares), K219Q (yellow triangles), S309F (blue diamonds), and S309L (pink triangles) are shown. (C, F) Activity versus $ΔT_{50}$ for the FoldX (C) and the Triad (F) mutants.

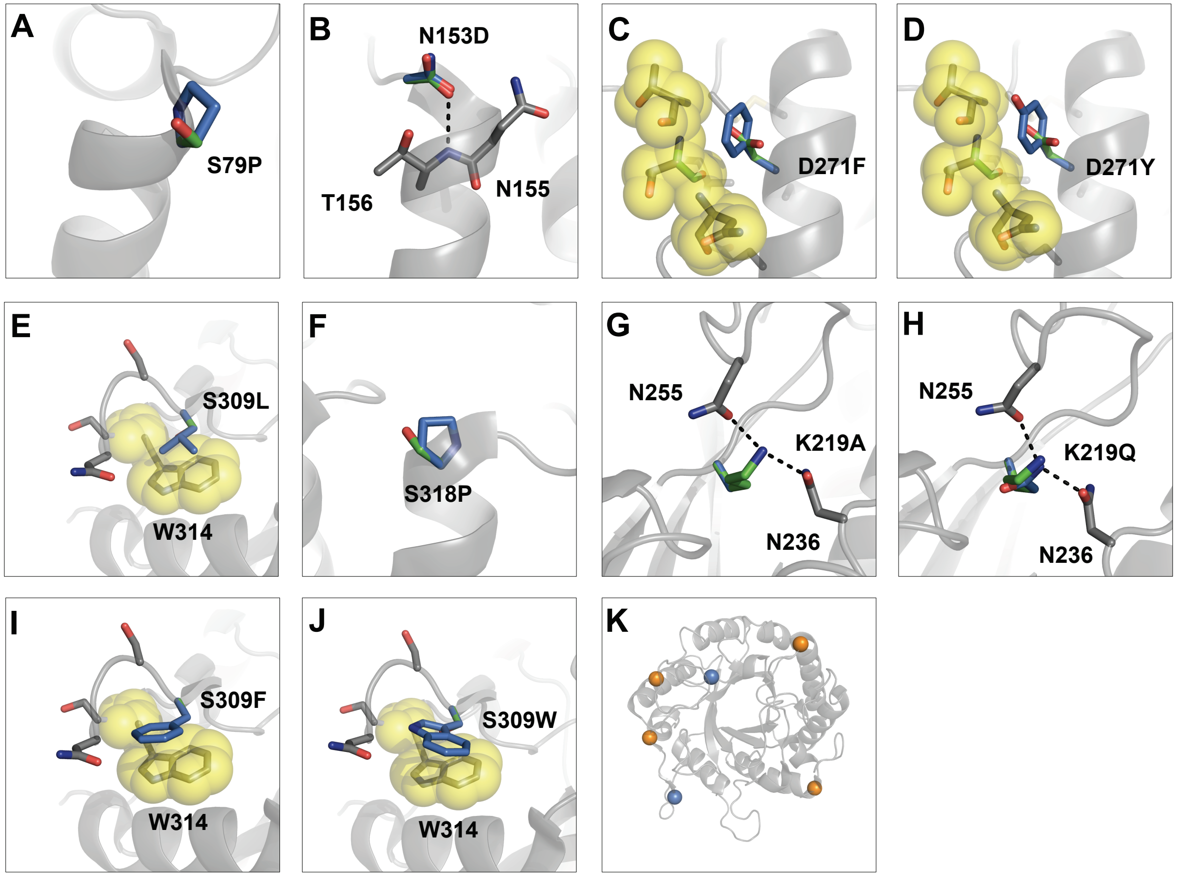


Figure 2. *Structural analysis of stabilizing FoldX and Triad mutations*. FoldX mutations are depicted in panels A-F. Triad mutations appear in panels E and G-J. The WT and mutation sidechains are shown as green and blue sticks, respectively. (K) The location of stabilizing FoldX (orange) and Triad (blue) mutations within the *Hj*Cel5A structure (gray cartoon).

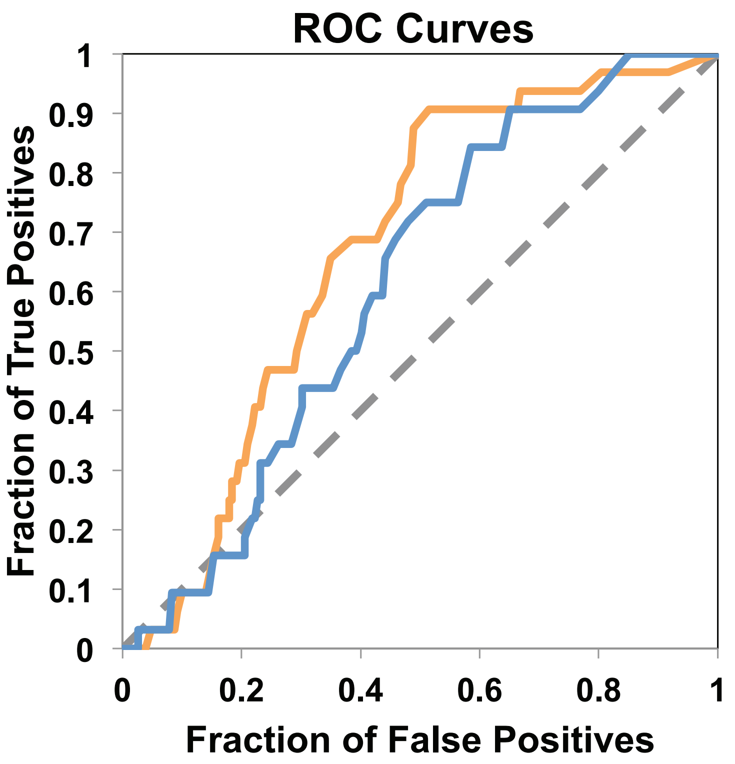


Figure 3. *FoldX and Triad Receiver Operator Characteristic curves*. ROC curves for predicting adequately active and well-expressed thermostable mutations are shown for FoldX (orange) and Triad (blue) predictions. Curves closer to the diagonal (dotted line) represent metrics that provide no predictive benefit over random choice.

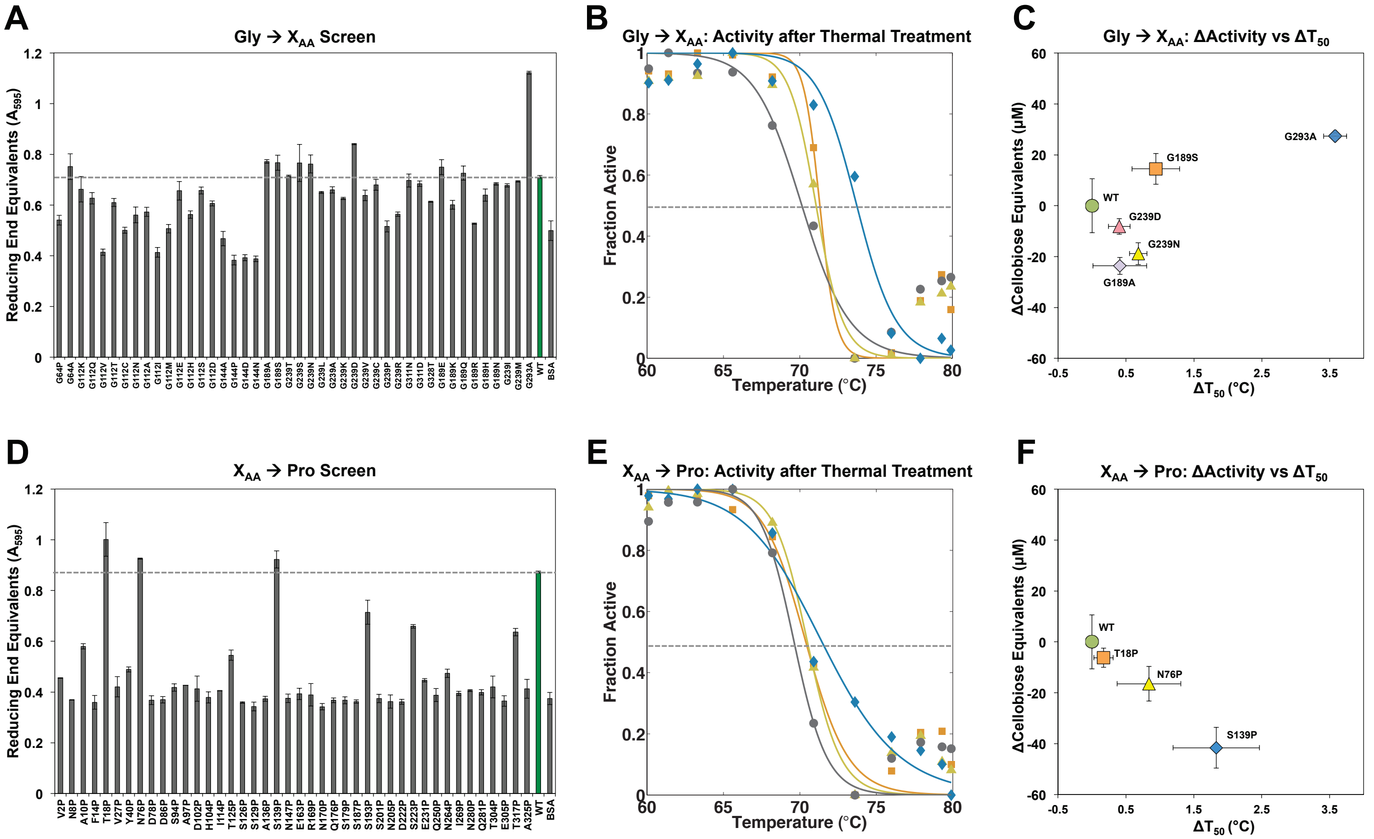


Figure 4. *Screening and characterization of glycine and proline mutations.* (A, D) Activity screen on Glycine (A) and Proline (D) mutants. WT is shown in green. The WT activity level is marked with a dotted line for reference. (B, E) Activity of *Hj*Cel5A point mutants after treatment over a range of temperatures. The dotted line marks the point at which half of the original activity remains (T_{50}). (B) Data for the glycine mutants G189S (orange squares), G239N (yellow triangles), and G293A (blue diamonds) are plotted. (E) Data for the proline mutants T18P (orange squares), N76P (yellow triangles), and S139P (blue diamonds) are shown. (C, F) Activity versus $ΔT_{50}$ for the glycine (C) and the proline (F) mutations.

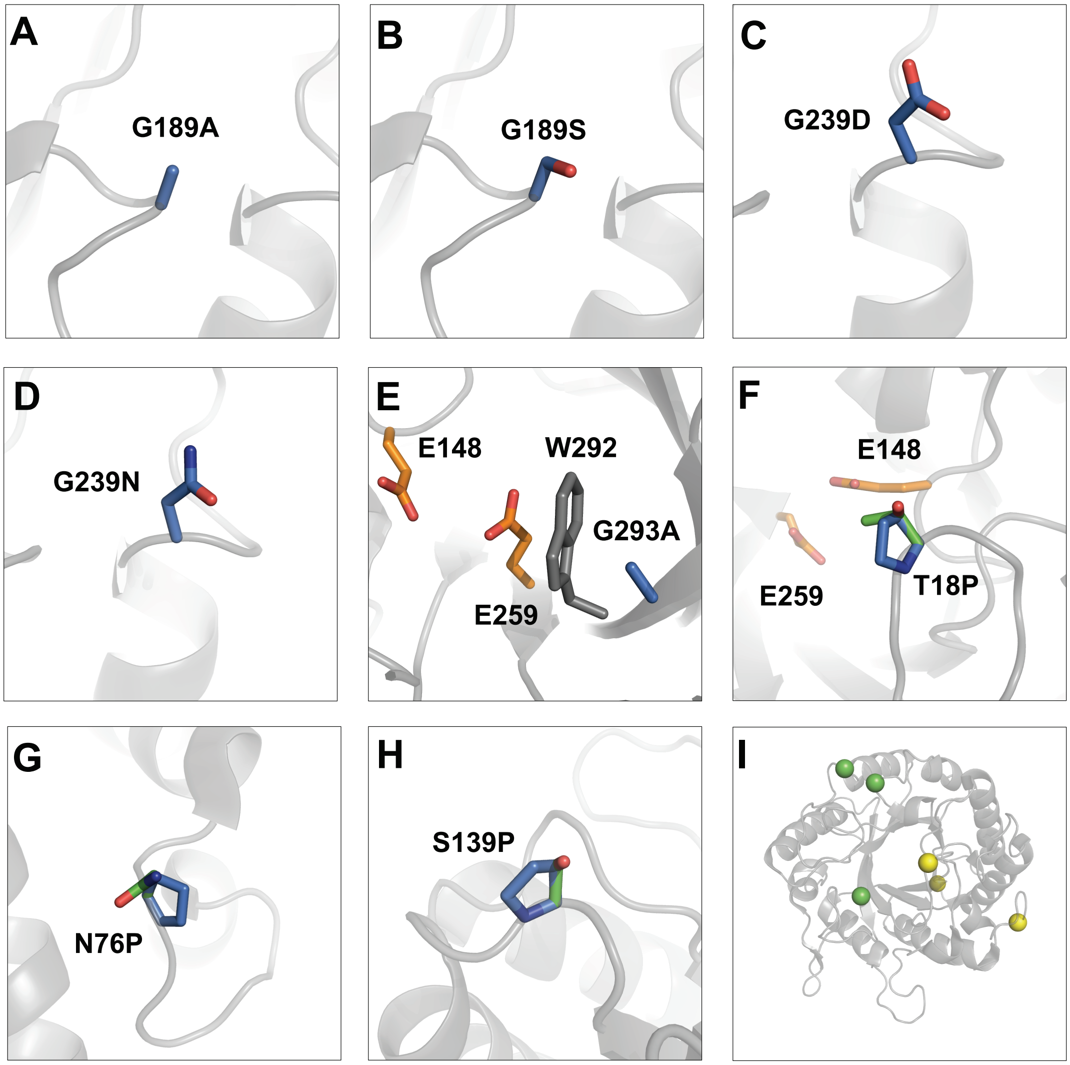


Figure 5. *Structural analysis of stabilizing glycine and proline mutations*. (A-E) Location of stabilizing glycine mutations. (F-H) Location of stabilizing proline mutations. In panels A-H, the WT and mutated sidechains are shown as green and blue sticks, respectively. (I) The location of stabilizing glycine (green) and proline (yellow) mutations within the *Hj*Cel5A structure (gray cartoon).

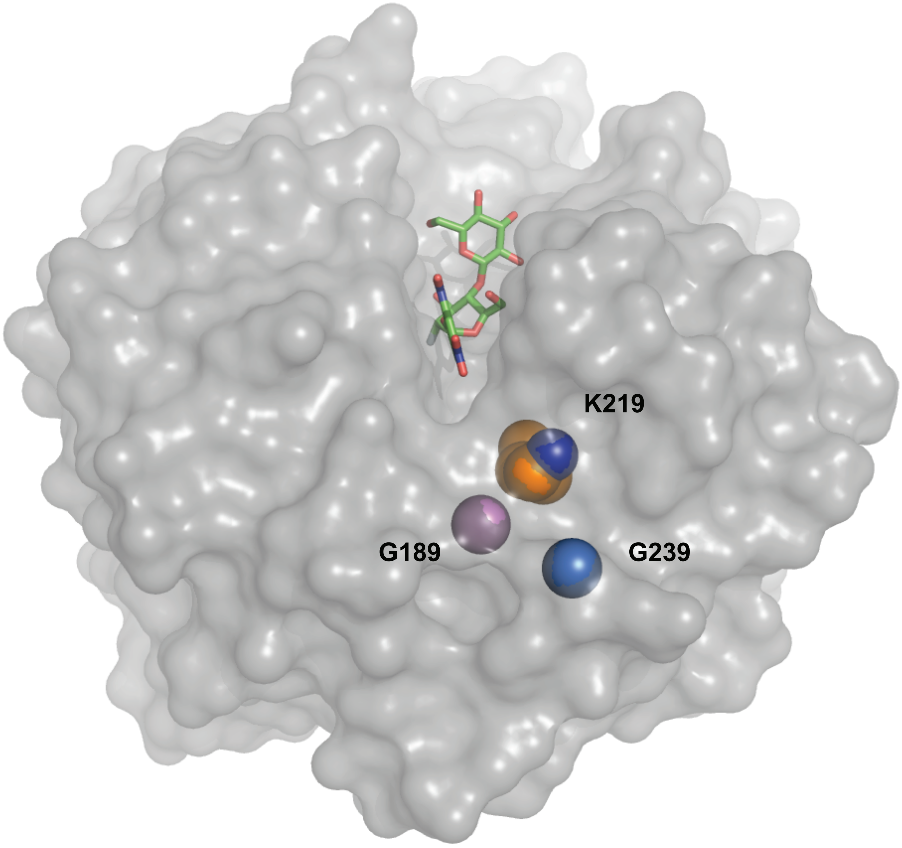


Figure 6. *Location of positions 189, 219 and 239*. Three stabilizing residues sitting near the substrate binding pocket, G189, K219, and G239, are drawn as spheres. To highlight the active site, the substrate analog 2,4-dinitrophenyl-2-deoxy-2-fluoro-β-D-cellobioside is modeled as green sticks. This molecule appears in the *Bacillus agaradhaerens* Cel5A crystal structure (PDB ID 4A3H [38]) and was superimposed onto the *Hj*Cel5A structure using the align command in PyMOL [54].

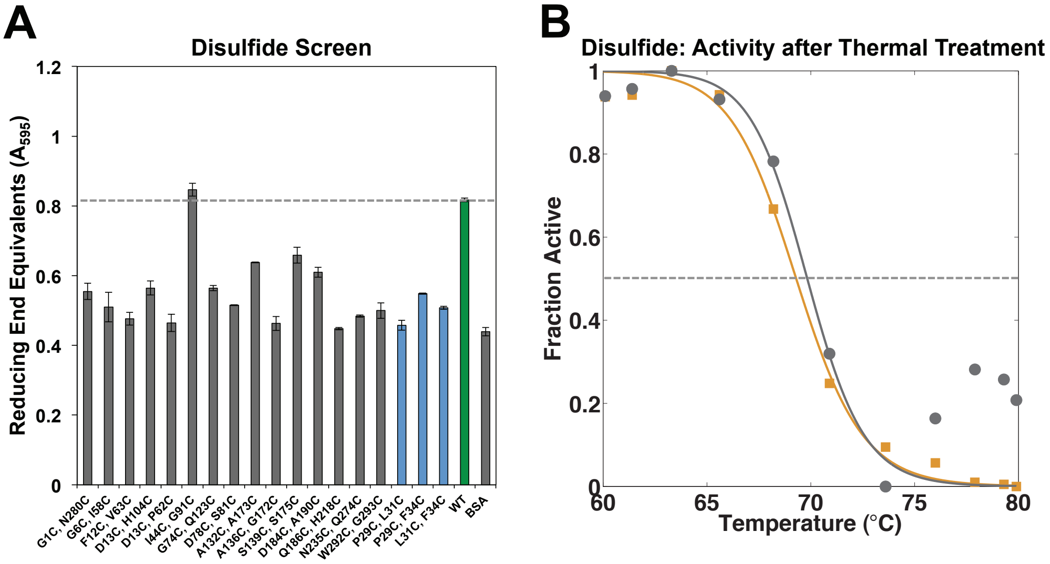


Figure 7. *Screening and characterization of disulfide mutants*. (A) The activity screen for all disulfide constructs. Bars marked in blue correspond to constructs predicted with Triad. WT is shown in green and its activity is marked with a dotted line to facilitate comparison. (B) Activity at 60 °C following a 10 minute incubation across a gradient of temperatures. WT is represented with gray circles. Data for the disulfide bond mutant I44C, G91C is displayed as orange squares. The temperature at which half of the maximal activity lingers (ΔT_{50}) occurs at the point where the curves intersect the dotted line.

5.7 References

1. Daniel, R.M., Cowan D., A., Morgan H.W. and Curran M.P. (1982) A correlation between protein thermostability and resistance to proteolysis. Biochemical Journal 207:641-644.
2. Viikari, L., Alapuranen M., Puranen T., Vehmaanperä J. and Siika-aho M. Thermostable enzymes in lignocellulose hydrolysis. In: Olsson L, Ed. (2007) Biofuels. Springer Berlin Heidelberg, pp. 121-145.
3. Mingardon, F., Bagert J.D., Maisonnier C., Trudeau D.L. and Arnold F.H. (2011) Comparison of family 9 cellulases from mesophilic and thermophilic bacteria. Applied and Environmental Microbiology 77:1436-1442.
4. Wu, I. and Arnold F.H. (2013) Engineered thermostable fungal Cel6A and Cel7A cellobiohydrolases hydrolyze cellulose efficiently at elevated temperatures. Biotechnology and Bioengineering 110:1874-1883.
5. Komor, R.S., Romero P.A., Xie C.B. and Arnold F.H. (2012) Highly thermostable fungal cellobiohydrolase I (Cel7A) engineered using predictive methods. Protein Engineering Design and Selection 25:827-833.
6. Talluri, S. (2011) PLS for prediction of therma stability of protein mutants. Journal of Advanced Bioinformatics Applications and Research 2:155-160.
7. Zeldovich, K.B., Berezovsky I.N. and Shakhnovich E.E. (2007) Protein and DNA Sequence Determinants of Thermophilic Adaptation. . PLoS Comput Biol 3:e5.
8. Montanucci, L., Fariselli P., Martelli P.L. and Casadio R. (2008) Predicting protein thermostability changes from sequence upon multiple mutations. Bioinformatics 24:i190-i195.
9. Lehmann, M., Kostrewa D., Wyss M., Brugger R., D'Arcy A., Pasamontes L. and van Loon A.P.G.M. (2000) From DNA sequence to improved functionality: using protein sequence comparisons to rapidly design a thermostable consensus phytase. Protein Engineering 13:49-57.
10. Capriotti, E., Fariselli P. and Casadio R. (2004) A neural-network-based method for predicting protein stability changes upon single point mutations. Bioinformatics 20:i63-i68.
11. Kumar, M.D.S., Bava K.A., Gromiha M.M., Prabakaran P., Kitajima K., Uedaira H. and Sarai A. (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein–nucleic acid interactions. Nucleic Acids Research 34:D204-D206.
12. Zhou, H. and Zhou Y. (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Science 11:2714-2726.
13. Parthiban, V., Gromiha M.M. and Schomburg D. (2006) CUPSAT: prediction of protein stability upon point mutations. Nucleic Acids Research 34:W239-W242.
14. Dehouck, Y., Grosfils A., Folch B., Gilis D., Bogaerts P. and Rooman M. (2009) Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. Bioinformatics 25:2537-2543.

15. Capriotti, E., Fariselli P. and Casadio R. (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Research 33:W306-W310.

16. Schymkowitz, J., Borg J., Stricher F., Nys R., Rousseau F. and Serrano L. (2005) The FoldX web server: an online force field. Nucleic Acids Research 33:392-388.

17. Khan, S. and Vihinen M. (2010) Performance of protein stability predictors. Human Mutation 31:675-684.

18. Guerois, R., Nielsen J.E. and Serrano L. (2002) Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations. Journal of Molecular Biology 320:369-387.

19. van der Sloot, A.M., Tur V., Szegezdi E., Mullally M.M., Cool R.H., Samali A., Serrano L. and Quax W.J. (2006) Designed tumor necrosis factor-related apoptosis-inducing ligand variants initiating apoptosis exclusively via the DR5 receptor. Proceedings of the National Academy of Sciences 103:8634-8639.

20. Song, X., Wang Y., Shu Z., Hong J., Li T. and Yao L. (2013) Engineering a More Thermostable Blue Light Photo Receptor *Bacillus subtilis* YtvA LOV Domain by a Computer Aided Rational Design Method. PLoS Comput Biol 9:e1003129.

21. Chandrasekaran, P., Doss C.G., Nisha J., Sethumadhavan R., Shanthi V., Ramanathan K. and Rajasekaran R. (2013) *In silico* analysis of detrimental mutations in ADD domain of chromatin remodeling protein ATRX that cause ATR-X syndrome: X-linked disorder. Network Modeling Analysis in Health Informatics and Bioinformatics 2:123-135.

22. Kellogg, E.H., Leaver-Fay A. and Baker D. (2011) Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins: Structure, Function, and Bioinformatics 79:830-838.

23. Rohl, C.A., Strauss C.E.M., Misura K.M.S. and Baker D. Protein Structure Prediction Using Rosetta. In: Ludwig B,Michael LJ, Eds. (2004) Methods in Enzymology. Academic Press, pp. 66-93.

24. Matthews, B.W., Nicholson H. and Becktel W.J. (1987) Enhanced protein thermostability from site-directed mutations that decrease the entropy of unfolding. Proceedings of the National Academy of Sciences 84:6663-6667.

25. Li, Y., Coutinho P.M. and Ford C. (1998) Effect on thermostability and catalytic activity of introducing disulfide bonds into *Aspergillus awamori* glucoamylase. Protein Engineering 11:661-667.

26. Wakarchuk, W.W., Sung W.L., Campbell R.L., Cunningham A., Watson D.C. and Yaguchi M. (1994) Thermostabilization of the *Bacillus circulans* xylanase by the introduction of disulfide bonds. Protein Engineering 7:1379-1386.

27. Reetz, M.T., Carballeira J.D. and Vogel A. (2006) Iterative Saturation Mutagenesis on the Basis of B Factors as a Strategy for Increasing Protein Thermostability. Angewandte Chemie International Edition 45:7745-7751.

28. Suzuki, Y., Oishi K., Nakano H. and Nagayama T. (1987) A strong correlation between the increase in number of proline residues and the rise in thermostability of five *Bacillus* oligo-1,6-glucosidases. Applied Microbiology and Biotechnology 26:546-551.

29. Suzuki, Y., Hatagaki K. and Oda H. (1991) A hyperthermostable pullulanase produced by an extreme thermophile, *Bacillus flavocaldarius* KP 1228, and

evidence for the proline theory of increasing protein thermostability. Applied Microbiology and Biotechnology 34:707-714.

30. Tian, J., Wang P., Gao S., Chu X., Wu N. and Fan Y. (2010) Enhanced thermostability of methyl parathion hydrolase from *Ochrobactrum sp.* M231 by rational engineering of a glycine to proline mutation. FEBS Journal 277:4901-4908.

31. Saloheimo, M., Lehtovaara P., Penttilä M., Teeri T.T., Ståhlberg J., Johansson G., Pettersson G., Claeyssens M., Tomme P. and Knowles J.K.C. (1988) EGIII, a new endoglucanase from *Trichoderma reesei*: the characterization of both gene and enzyme. Gene 63:11-21.

32. Bisaria, V.S. and Ghose T.K. (1981) Biodegradation of cellulosic materials: Substrates, microorganisms, enzymes and products. Enzyme and Microbial Technology 3:90-104.

33. Kim, M.K. and Kang Y.K. (1999) Positional preference of proline in alpha-helices. Protein Science 8:1492-1499.

34. Lee, T.M., Farrow M.F., Arnold F.H. and Mayo S.L. (2011) A structural study of *Hypocrea jecorina* Cel5A. Protein Science 20:1935-1940.

35. Lo Leggio, L. and Larsen S. (2002) The 1.62 Å structure of *Thermoascus aurantiacus* endoglucanase: completing the structural picture of subfamilies in glycoside hydrolase family 5. FEBS Letters 523:103-108.

36. Van Petegem, F., Vandenberghe I., Bhat M.K. and Van Beeumen J. (2002) Atomic resolution structure of the major endoglucanase from *Thermoascus aurantiacus*. Biochemical and Biophysical Research Communications 296:161-166.

37. Delsaute, M., Berlemont R., Dehareng D., Van Elder D., Galleni M. and Bauvois C. (2013) Three-dimensional structure of RBcel1, a metagenome-derived psychrotolerant family GH5 endoglucanase. Acta Crystallographica Section F 69:828-833.

38. Davies, G.J., Mackenzie L., Varrot A., Dauter M., Brzozowski A.M., Schülein M. and Withers S.G. (1998) Snapshots along an Enzymatic Reaction Coordinate: Analysis of a Retaining β-Glycoside Hydrolase. Biochemistry 37:11707-11713.

39. Tseng, C.-W., Ko T.-P., Guo R.-T., Huang J.-W., Wang H.-C., Huang C.-H., Cheng Y.-S., Wang A.H.J. and Liu J.-R. (2011) Substrate binding of a GH5 endoglucanase from the ruminal fungus *Piromyces rhizinflata*. Acta Crystallographica Section F 67:1189-1194.

40. Mason, S.J. and Graham N.E. (2002) Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. Quarterly Journal of the Royal Meteorological Society 128:2145-2166.

41. Hutchinson, E.G. and Thornton J.M. (1994) A revised set of potentials for β-turn formation in proteins. Protein Science 3:2207-2216.

42. Závodszky, P., Kardos J., Svingor Á. and Petsko G.A. (1998) Adjustment of conformational flexibility is a key event in the thermal adaptation of proteins. Proceedings of the National Academy of Sciences 95:7406-7411.

43. Mittal, S., Cai Y., Nalam M.N., Bolon D.A. and A. S.C. (2012) Hydrophobic core flexibility modulates enzyme activity in HIV-1 protease. Journal of the American Chemical Society 134:4163-4168.
44. Wallnoefer, H.G., Lingott T., Gutiérrez J.M., Merfort I. and Liedl K.R. (2010) Backbone flexibility controls the activity and specificity of a protein−protein interface: specificity in snake venom metalloproteases. Journal of the American Chemical Society 132:10330-10337.
45. Siadat, O., Lougarre A., Lamouroux L., Ladurantie C. and Fournier D. (2006) The effect of engineered disulfide bonds on the stability of *Drosophila melanogaster* acetylcholinesterase. BMC Biochemistry 7:12.
46. Mansfeld, J., Vriend G., Dijkstra B.W., Veltman O.R., Van den Burg B., Venema G., Ulbrich-Hofmann R. and Eijsink V.G.H. (1997) Extreme Stabilization of a Thermolysin-like Protease by an Engineered Disulfide Bond. Journal of Biological Chemistry 272:11152-11156.
47. Matsumura, M., Becktel W.J., Levitt M. and Matthews B.W. (1989) Stabilization of phage T4 lysozyme by engineered disulfide bonds. Proceedings of the National Academy of Sciences 86:6562-6566.
48. Németh, A., Kamondi S., Szilágyi A., Magyar C., Kovári Z. and Závodszky P. (2002) Increasing the thermal stability of cellulase C using rules learned from thermophilic proteins: a pilot study. Biophysical Chemistry 96:229-241.
49. Dombkowski, A.A. (2003) Disulfide by Design™: a computational method for the rational design of disulfide bonds in proteins. Bioinformatics 19:1852-1853.
50. Betz, S.F. (1993) Disulfide bonds and the stability of globular proteins. Protein Science 2:1551-1558.
51. Lazaridis, T. and Karplus M. (1999) Effective energy function for proteins in solution. Proteins: Structure, Function, and Bioinformatics 35:133-152.
52. Daniel, R.M., Dines M. and Petach H.H. (1996) The denaturation and degradation of stable enzymes at high temperatures. Biochemical Journal 317:1-11.
53. Triad. (2012). Protabit LLC, Pasadena, CA.
54. The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC.

CHAPTER 6

A Comparison of Stabilization Techniques Applied to *Hypocrea jecorina* Cel5A

This chapter is partially formatted for submission to the Proceedings of the National Academy of Sciences.

6.1 Abstract

Producing liquid fuel and polymer products from cellulosic material may mitigate the detrimental environmental and social effects of grain-based production [1-3]. Sustainable hydrolysis of cellulosic feedstocks into fermentable sugars, however, prohibitively requires relatively expensive enzymes [4]. One proposed strategy for alleviating this problem involves engineering thermostable variants of industrial cellulases, enzymes that cleave cellulose chains into smaller sugars. Such molecules would ideally function above current operating temperatures (~50 °C), providing more efficient glucose release and lowering production costs.

Using the primary endoglucanase from the industrial workhorse *Hypocrea jecorina* (*Hj*Cel5A) as a molecular guinea pig, we compare a plethora of disparate methods designed to improve protein stability. These methods include consensus design, core repacking, helix dipole stabilization, ΔΔG-based methods (FoldX, Triad), and backbone stabilization. For the examined system, consensus design not only provides the largest improvements in stability, but also preserves or even elevates activity. FoldX ΔΔG approximations also revealed several highly stabilizing and active mutations.

An initial combination mutant containing highly stabilizing mutations with enzymatic activity as a secondary consideration showed high thermostability, but poor performance in long-term hydrolysis assays. Mutations reducing activity were substituted for those conferring smaller stability improvements with less adverse effects on enzymatic function. The resulting combination mutants demonstrate a 12-15 °C increase in T_{50} (T_{50} = 84-86 °C), an 11-14 °C increase in optimal temperature (T_{opt} = 75-78 °C) and a 60%

increase over the maximal amount of hydrolysis achievable using the WT enzyme. These studies highlight the importance of maintaining enzymatic function when searching for highly stable variants.

6.2 Significance Statement

Thermostable protein variants have broad applications in industrial and scientific realms. In this study, several highly divergent stabilization strategies are applied to a single enzyme, *Hypocrea jecorina* (anamorph *Trichoderma reesei*) Cel5A (*Hj*Cel5A). This enzyme serves as the primary endoglucanase in the workhorse of the biofuels industry and is a target for thermostabilization. In providing a comprehensive, experimental survey of currently popular stabilization techniques, we demonstrate that consensus design provides the most stabilizing and active mutations from the employed methods. Using the mutations revealed in this survey, we additionally constructed a combination mutant that showed a 14 °C improvement in thermostability as measured by the temperature at which half of the maximal activity persists (T_{50}). After removing or further altering stabilizing mutations that decreased activity, subsequent combination mutants demonstrated a 12-15 °C increase in T_{50} (T_{50} = 84-86 °C), an 11-14 °C increase in optimal temperature (T_{opt} = 75-78 °C) and a 60% elevation in hydrolysis over the maximum amount achievable with WT.

6.3 Introduction

A need for alternative liquid fuels exists. As a non-renewable resource, oil will eventually grow increasingly scarce and difficult to extract [5, 6]. Its use not only emits pollution [7, 8], but also fosters economic dependence on fossil fuel exporters [9]. Recognizing this looming hurdle, the Federal government has invested considerable resources in promoting the use of biofuels, mainly those originating from corn. However, diverting food for fuel creates competition between the energy and agricultural sectors, leading to agricultural intensification and elevated food prices [1-3]. Creating fuel from inedible cellulosic feedstocks from waste streams can alleviate this problem.

One method of generating liquid fuel precursors from cellulosic material involves enzymatically digesting feedstock into fermentable sugars. Generally, three major classes of cellulases are necessary to synergistically hydrolyze crystalline cellulose into glucose monomers: 1) exoglucanases which release two sugar-unit molecules called cellobiose from either the reducing or non-reducing ends of cellulose chains, 2) endoglucanases which cleave strands internally at amorphous kinks in the crystalline lattice, and 3) β-glucosidases which cleave smaller fragments into glucose monomers [10]. As the cost of producing these enzymes remains prohibitively high in the range of $0.10 - $1.47/gal [4, 11-13], considerable effort has been invested in streamlining this process [14-18].

Engineering thermostable variants constitutes one means of dramatically reducing enzyme production costs. Reactions typically proceed at 50 °C, the temperature at which the major cellulases remain optimally active [19]. Increasing the reaction temperature minimizes microbial contamination and saves energy through reducing the amount of cooling necessary after pretreatment at temperatures around 200 °C [20]. Improving protein stability also protects against degradation during storage, production, and hydrolysis [21]. Furthermore, provided the enzymes remain folded, reaction rates typically increase as temperatures rise [22, 23].

Previous attempts to engineer thermostable protein variants have collectively determined that no “general rational rule” exists that will provide maximal stability benefits by optimizing certain protein characteristics [24-28]. Typically, methods targeting a single characteristic fail to recognize many beneficial mutations easily captured through examining other features. Although acknowledgement of this trend dates to over a decade ago, only a handful of studies have attempted to detect stabilizing mutations using multifarious criteria within the same protein system [17]. Many studies focus on comparing improved versions of a technique to their predecessors. Such methods include design by ΔΔG values [29], repacking the hydrophobic core [30], and consensus design [31]. A more comprehensive comparison evaluating the performance of disparate methods, however, may provide further insights.

Here, we describe the construction of stable Cel5A (EGII, *Hj*Cel5A) variants. *Hj*Cel5A is the primary endoglucanase from *Hypocrea jecorina* (anamorph *Trichoderma reesei*) (*Hj*Cel5A) [32]. Capable of producing 100 g L^{-1} of native enzymes, *H. jecorina* serves as the source for many of the enzymes used for cellulosic biofuel production [33]. To gain a comprehensive understanding of how to best thermostabilize a protein, we employed several highly diverse stabilization strategies and analyzed each for effectiveness. This comprehensive approach includes consensus design, core repacking, helix dipole stabilization, ΔΔG-based methods (FoldX, Triad), and backbone stabilization. Mutations recovered from these studies were combined to create highly thermostable *Hj*Cel5A variants with elevated activity and expression level in *Saccharomyces cerevisiae.* These results may find use in an industrial setting.

6.4 Results

Identification and Characterization of Stabilizing Mutations

The first step in constructing a thermostable enzyme variant involves identifying the changes necessary to foster stabilization. To facilitate cloning and testing, seven methods were chosen or modified to produce point mutations. These methods are: 1) consensus design, 2) core repacking, 3) helix dipole stabilization, design with 4) FoldX and 5) Triad ΔΔG approximations, and backbone stabilization through 6) mutating glycines to residues with a $C_β$ and 7) introducing prolines.

Consensus design employs multiple sequence alignments to determine the most prevalent residue at a given position. This approach selects residues as potentially stabilizing when they appear more frequently in a multiple sequence alignment of homologs to the protein of interest than is expected based on a reference state (e.g. codon frequency or the frequency of the wild type (WT) residue) as potentially stabilizing [31, 34].

The core repacking strategy seeks to stabilize the folded state over the unfolded state through improving hydrophobicity in the interior of the protein. Protein folding is largely driven by the hydrophobic effect, i.e. polypeptide chains fold to bury hydrophobic residues and minimize disruption of hydrogen bonds among solvent molecules [35]. To heighten this effect, we employed computational design software to fill voids and increasing the prevalence of hydrophobic sidechains in the protein core.

The helix dipole stabilization method rests on the principle that changing the electrostatic properties of residues at the ends of helices may confer stability [36]. Every α-helix contains an inherent dipole originating from three potentially unsatisfied backbone hydrogen bonds at the N- and C-terminal residues of the protein. Adding an N-capping residue to contact an unpaired amine or mutating the N- and C-terminal residues to counter this dipole may enhance protein, stability. Using the same software employed in the core repacking strategy, we attempted to design thermostable *Hj*Cel5A variants through stabilizing the helix dipole.

We additionally used FoldX [37] and Triad [38] to calculate ΔΔG values for every possible mutation in *Hj*Cel5A. These methods employ molecular mechanics forcefields tuned with empirically-influenced weights to predict differences in energy between the WT and mutated sequences. One can adapt this strategy to search for mutations that decrease the entropy of unfolding through restricting the trajectory of the protein backbone. Gly → X_{AA} and X_{AA} → Pro mutations reduce the allowable Ramachandran space, potentially stabilizing the folded state over the unfolded [39].

A more thorough review the seven design strategies and their implementation appears in chapters 3 (consensus design), 4 (core repacking and helix dipole stabilization), and 5 (FoldX, Triad, glycine and proline backbone stabilization) of this thesis. To reduce bias and form accurate comparisons between methods, rational input from the experimenter was purposefully excluded. Thus, the set of characterized mutations contains those clearly providing no benefit (e.g., tryptophan mutations on the surface or mutations disrupting active site networks).

In total, 262 unique point mutations were predicted as stabilizing and subsequently characterized (see Appendix A, Table I). Constructs were secreted from *Saccharomyces cerevisiae* and the expression supernatants were screened for ability to release sugar from Avicel, a crystalline cellulose powder, after two hours at 73 °C. This temperature exceeds that at which half of the WT *Hj*Cel5A remains folded (T_m) by 3.5 °C ($T_{m,WT}$ = 69.5 °C). Under these conditions, only constructs with improved expression, stability, activity, or a mix thereof will demonstrate enough activity to outperform WT. From the 262 constructs tested, 43 mutations showed greater activity than WT in the screen. These mutations were expressed, purified, and assessed for activity at 60 °C for one hour following a 10 minute heat treatment across a gradient ranging from 60 – 80 °C. Thirty two mutations showed an improvement in T_{50}, the temperature at which half of the maximal activity persists (Tables I and II).

Database Construction

Appendix A, Table II in this chapter, and the supplementary accessory files submitted with this work summarize the information collected during the prediction and experimental phases. These tables contain all of the measures generated through methods capable of calculating a value for the WT residue and every possible mutation in the protein.

First Generation Combination Mutants

Using the information gleaned from the various stabilization strategies, we assembled a series of combination constructs containing highly stabilizing mutations. Mutations demonstrating a $\Delta T_{50} \geq 0.5$ °C were selected for incorporation. If several stabilizing mutations appeared in the same region within the *Hj*Cel5A structure, the mutations with the highest ΔT_{50} values were generally retained. In ambiguous situations where interaction between two sites could not clearly be ascertained, several alternative constructs were tested. The final chosen set contains 13 possible mutations: T57N, N76P, T80E, S139P, N155E, G189S, K219Q, G239N, D271F, Y278F, G293A, S309F, and S318P.

These mutations were incrementally added to the WT sequence with the least and most mutated constructs containing 1 and 13 mutations, respectively (Table III). Following cloning, mutants were tested for activity with the same screen employed to detect point mutations (Figure 1A). All of the constructs showed improvements in activity over WT.

Two constructs predicted to have either high activity or thermostability were chosen for further characterization. With nine mutations (T57N, N76P, T80E, S139P, N155E, G189S, D271F, Y278F, and G293A), construct 16 demonstrated the greatest performance on the activity screen. Construct 20 contains all 13 possible mutations and was projected to demonstrate the greatest thermostability. These two combination mutants were expressed, purified and tested to assess their hydrolytic capabilities on Avicel over a gradient of temperatures (Figure 1B), obtain their T_{50} values (Figure 1C), and determine their activity at 60 °C after 2 hours (Figure 1D). Table IV provides a summary of these results. The 9-point mutant (construct 16) shows a 10 °C increase in both the optimal operating temperature (T_{opt}) and in T_{50} relative to WT. As expected, the 13-point mutant demonstrates even greater improvement in thermostability with a 16 °C increase in T_{opt} and a 14°C increase in T_{50} relative to WT. These increases, however, were accompanied by decreases in activity of 4% for the 9-point mutant and 17% for the 13-point mutant. Analysis revealed that the 13-point mutant contains more mutations that are known from previous experiments to decrease activity than the 9-point mutant. As such, the drop in activity likely originates from the collective effects of point mutations rather than incompatibilities between the mutations.

To ascertain whether additional mutagenesis could counter these activity decreases, we assembled a 20-point mutant containing mutations with $\Delta T_{50} \geq 0$ °C. This combination mutant contains all 13 of the previously incorporated mutations plus T18P, G64A, S79P, V101I, S133R, D13E, and E53D. Individually, many of the less stabilizing mutations show improvements in activity; summing the changes in activity measured for these point mutations gives a net activity increase of 20.7 μM cellobiose equivalents. As such, we tested the possibility that including these less stabilizing mutations might boost thermostability while rescuing activity. While the 20-point mutant shows a ΔT_{50} of 16.8 °C (Figure 1C), its activity at 60 °C decreases even further to 47.1 μM cellobiose equivalents below WT (22% of WT activity) (Figure 1D). This result suggests that activity losses from individual mutations permanently accumulate and cannot be rescued by adding mutations that improve activity in isolation.

Although these initial combination mutants exhibit diminished activity, we hypothesized that their enhanced thermostability might improve performance in longer assays. Due to its relatively modest decrease in activity, the 9-point mutant (construct 16) was chosen for 60-hour activity tests on Avicel at 60 and 70 °C (Figures 1E and F). This construct demonstrates a nine-fold activity improvement over WT at the elevated temperature. When compared to WT hydrolysis at 60 °C, the total product yield improves by about 134 μM of cellobiose equivalents (24% increase, 1.2 fold improvement).

Second Generation Combination Mutants

Given that several alternate stabilizing, highly-active mutations appear in the pool of tested constructs, we surmised that further improvements were possible. Using the 9-point mutant as a template, we created second-generation combination mutants by excluding all mutations detrimentally affecting activity. The process involved reverting mutations N76P, S139P, K219Q, G239N, D271F, Y278F, and S309F back to the WT residue or, if available, a less stable, more active alternate. Several changes occurred in an area of the protein adjacent to the active site (Figures 4A-C). As this region putatively serves as a substrate-binding channel, several mutations appear in this region that highly modulate activity. Mutations appearing in the four final second generation 13-point combination mutants (s13pt 1-4) are summarized in Table III.

The second-generation combination mutants perform as well or better than WT on all tested metrics. All four constructs show enhancements in thermostability with s13pt2 demonstrating the highest increase (ΔT_{50} = 15.4 °C) (Table IV, Figures 2A and B). Activity at 60 °C improved over WT for all constructs except s13pt1 (Figure 2C). In this case, activity declined by a mere 1.0 μM of cellobiose, and insignificant value. The combination mutants exhibit dramatic improvements in T_{opt}. s13pt1/2 and s13pt3/4 optimally function at 78 (ΔT_{opt} = 14 °C) and 75 degrees (ΔT_{opt} = 11 °C), respectively. Finally, all of the mutants show ~4-6 fold increases in expression level over WT.

To explore whether these improvements would translate to conditions approximating industrial reactions, we conducted 60-hour hydrolysis experiments on the constructs with the highest activity (s13pt4) and thermostability (s13pt2). At 60 °C, both combination mutants performed similarly to WT (Figure 3A). This improvement dramatically increased at 70 °C (Figure 3B). Compared to the WT performance at the same temperature, s13pt4 and s13pt2 exhibit ~9.5-10 fold increases in activity. Moreover, the combination mutants improve yield by 358-414 μM cellobiose equivalents (~60% increase) over the maximal amount possible using the WT enzyme.

As the second-generation combination mutants optimally perform at 75-78 °C in the 2-hour hydrolysis experiments, we tested whether long-term hydrolysis improves at higher temperatures. After 60-hours of hydrolysis, less activity was observed at 75 or 78 °C than at 70 °C (Figures 3C and D). This decrease in activity likely occurs due to gradual thermal degradation triggered by elevated temperatures. For example, trials wherein WT enzyme was pre-incubated at 50 °C for several hours before performing activity tests show that even long exposure to temperatures below the T_m of the protein reduces activity (data not shown). Thus, the optimal temperature for long-term hydrolysis represents a compromise between increased activity due to the Arrhenius effect and decreased activity due to slow thermal degradation.

Comparison of Stabilization Strategies

Although information from each design strategy contributed to constructing the final mutants, some methods proved more effective than others. Constructs s13pt2 and s13pt4 contain five mutations predicted from the helix dipole stabilization strategy, four from FoldX, three from consensus design, two from mutating glycines, one from redesigning the core, and one from Rosetta ΔΔG predictions (Figure 4D). While this distribution of mutations seems to suggest helix dipole stabilization as the most effective measure, more rigorous analysis reveals that each method provides a different combination of benefits.

Number of stabilizing mutations detected: Targeting the helix dipole successfully predicted the highest number of stabilizing mutations from all probed methods (Figure 5A). On average, these mutations, however, provided minimal thermostabilizing effects. With the exception of core repacking and backbone stabilization through adding prolines, the remaining methods predicted a fair number of mutations considering the size of each candidate pool.

Accuracy: Consensus design provided the highest prediction accuracy with helix dipole stabilization placing second (Figure 5B). Compared with third ranked method, FoldX, consensus design is 1.7 times more accurate, demonstrating the effectiveness of this approach. Moreover, it is highly possible that the accuracy of consensus design could increase dramatically using design parameters outlined in Chapter 3.

Expression: In general, most methods preserved *Hj*Cel5A expression in the heterologous host *S. cerevisiae* (Figure 5C). Only FoldX and the Rosetta ΔΔG method reduced the average expression below WT levels. Notably, the average expression level of the FoldX mutants was approximately half that of the WT protein. Conversely, many of the helix stabilizing mutations dramatically elevated expression levels, raising the average expression level to twice that of WT. It is unclear whether this improvement would persist during endogenous production from *H. jecorina*. Also, as manipulations on the DNA level (e.g., codon optimization, promoter engineering) might rescue low expression levels, this finding may prove inconsequential.

Thermostability: The greatest observed stability benefits originate from consensus, Triad ΔΔG, and FoldX mutations (Figure 5D). Mutations from core repacking and helix dipole stabilization calculations stabilize *Hj*Cel5A to a marginal degree. These differences may originate from the properties of targeted sectors within the protein. Techniques that produce more mutations in the boundary and core regions tend to yield highly stabilizing mutations, provided that these rare mutations are even detected. Interestingly, the core mutations appear to provide little benefit. As the core repacking calculation only identified two stabilizing mutations, the sample size may be insufficiently large to form any concrete conclusions.

Activity: Design through FoldX, consensus, and helix dipole stabilization appears most effective at identifying highly active mutations (Figure 5E). Consensus design appears to perform particularly well, producing an average increase in activity over WT 1.6 times higher than that achieved using the second most effective method.

This analysis demonstrates that while no clearly superior stabilization method exists, consensus design appears best at predicting stabilizing mutations with the qualities desired for the purposes of this study. Here, consensus mutations not only improve thermostability by a high average of 2.2 °C, but also enhance activity by an average of 14.0 μmol cellobiose equivalents (Table I, Figure 6A). FoldX mutations also appear to simultaneously enhance thermostability and activity, but with lower prediction accuracy (Table I, Figure 6D).

6.5 Discussion

In this study, we sought to create *Hj*Cel5A variants capable of providing enhanced hydrolysis at temperatures higher than the current industrial standard. Constructs s13pt2 and s13pt4 exhibit dramatically improved thermostability ($\Delta T_{50,s13pt2}$ = 15.4 °C, $\Delta T_{50,s13pt4}$ = 12.2 °C), activity ($\Delta \text{Activity}_{s13pt2}$ = 27.4 μmol cellobiose equivalents, $\Delta \text{Activity}_{s13pt4}$ = 68.8 μmol cellobiose equivalents), and yield (4.1 and 3.7 fold increase over WT). These mutants improve long-term hydrolysis of crystalline cellulose by approximately 60%. In addition, the detailed information concerning each point and combination mutant may prove useful for future stabilization efforts.

The results presented in this work suggest that consensus design is the most effective method for identifying *Hj*Cel5A mutations that enhance thermostability while maintaining or improving expression and activity. Consensus design also demonstrates the greatest predictive accuracy, correctly identifying five highly-stabilizing, highly-active mutations out of a mere 21 candidates. Given that natural selection typically eliminates detrimental mutations, the pool of homologous mutations has already been "prescreened" for members that do not adversely affect folding and/or function. Conversely, methods that rely strictly on structural information may provide highly stabilizing mutations that perturb activity or dramatically reduce expression levels.

Although design by consensus performed best out of the surveyed techniques, all of the examined methods proved useful in stabilizing *Hj*Cel5A. Homology-based design cannot detect mutations appearing in the vicinity of unique structural features. For example, ΔΔG approximations with FoldX or Triad revealed several highly stabilizing mutations around solvent-exposed loops absent in homologous structures. Supplemental methods may also enhance features unaltered through consensus design. Incorporation of subtly stabilizing helix dipole mutations radically improved the expression levels of our constructs. Additionally, many of the stabilizing mutations were predicted using a single strategy. Consequently, simply testing mutations mutually predicted by two methods would discard most of the highly stabilizing mutations recovered in this study. We suggest using consensus design supplemented with FoldX, helix dipole stabilization, or Triad ΔΔG for future stabilization efforts.

Regardless of the employed design strategy, rankings did not correlate with the degree of stabilization. Plotting ΔT_{50} or activity versus the various metrics used in this study reveals no significant trends. As addressed by Potapov et al. [40], FoldX and similar computational methods can generate lists of mutations enriched in stabilizing members, but lack the resolution in accuracy to reliably rank individual mutations. Improving prediction accuracy will require improvements to existing algorithms or alternative design strategies.

In this study, several stabilization techniques were performed on a single protein using identical characterization methods to directly compare methods. Future studies are necessary, however, to test whether these results apply to proteins beyond *Hj*Cel5a. Thus far, stabilization experiments performed on an SH3 domain demonstrate that designs based on multiple sequence alignment data provided greater fidelity in prediction accuracy than a structure-based approach [41]. Previous studies also note that comprehensive methods targeting multiple protein features tend to improve prediction accuracy and performance of the final molecule [24, 25]. In the design of a thermostable *Hj*Cel7A, Komor *et al.* employed both FoldX and a consensus approach [17]. These results demonstrate that both methods proved moderately effective in selecting a pool of mutations to test, but cannot predict whether individual mutations will be stabilizing. In 2012, a combinatorial approach using computational (Rosetta) design, disulfide engineering, consensus design, and rational design by homology was employed to produce an antibody with a T_m over 90 °C [42]. While, the study showed that all methods contributed to the final design, experiments were performed in sequence, rendering comparisons of each method difficult.

Taken together, the results of this study emphasize the importance of considering function in designing thermostable enzyme variants. Many studies relegate this parameter to a secondary position, choosing to evaluate activity after achieving considerable gains in stability. Analysis of the most thermostable mutant created in this study (20-point mutant) shows that blindly incorporating stabilizing mutations into a combination variant may decrease activity enough to obviate any gains achieved through the enhanced stability. As the fundamental goal in many stabilization studies is the creation of enzyme variants that perform similar to or better than the WT catalyst at higher temperatures, activity should remain a paramount consideration in all design steps. The screens used in this study only identify mutations that exhibit improved thermostability, activity, or expression. This practice reduces the reported accuracy of the technique as highly stabilizing mutations that dramatically decrease activity or expression will escape detection. As a result, the 14% accuracy we report for FoldX is significantly lower than the previously reported 60% value [29]. The utility of detecting stabilizing mutations that cripple other essential aspects of protein function, however, remains unclear. We propose a realignment of priorities towards improving function at a particular condition rather than focusing primarily on thermostability.

6.6 Materials and Methods

Cel5A Plasmid Construction

See equivalent section in Chapter 3.

Thermostability/Activity Screen

See equivalent section in Chapter 3.

Park-Johnson Assay

See equivalent section in Chapter 3.

Enzyme Purification

See equivalent section in Chapter 3.

T_{50} Assay

See equivalent section in Chapter 3.

T_{opt} Assay

Determination of the temperature yielding the maximum activity proceeded through incubating enzyme with Avicel for two hours at a gradient of temperatures, then determining sugar release with the Park-Johnson assay. In a 96-well PCR plate, 40 μL of purified enzyme were combined with 60 μL of 1.5% Avicel suspended in cellulase buffer (100 mL 50 mM sodium acetate, pH 5.0). Samples were incubated for 2 hours across a 20 °C gradient centered around a temperature projected to capture the peak of activity based on T_{50} values. Activity was assessed from 25 μL of supernantant using the Park-Johnson assay.

Single-Point Activity Assay

See equivalent section in Chapter 3.

60-Hour Activity Assay

To assess activity over a constant temperature for 60 hours, enzyme and substrate mixtures were combined in individual PCR tubes and frozen to arrest hydrolysis. In each tube, 40 μL of purified enzyme at a concentration of 0.5 μM was combined with 60 μL of 1.67% Avicel suspended in cellulase buffer (100 mL 50 mM sodium acetate, pH 5.0). Incubation occurred in a PCR block preheated before adding samples to prevent background activity. Time points were collected at 0, 4, 8, 16, 24, 36, 48, and 60 hours. Following hydrolysis, the reactions were thawed and moved to a 96-well plate to facilitate centrifugation. Supernantants were robotically collected. Cellobiose standards containing 0.0, 166.6, 333.3, 500.0, 833,3, 1000.0, 1500.0, and 2000.0 μM of cellobiose and 50 μL of the reaction supernantants were assessed for reducing sugar concentrations via the Nelson-Somogyi assay [43, 44]. All experiments were performed in triplicate.

6.7 Tables and Figures

Table I. *Summary of design strategies*

Strategy	**# Tested Mutations**	**# Mutations (ΔT_{50} > 0°C)**	**% Accuracy**	**Average ΔT_{50} (°C)**	**Average ΔActivity (µmol cellobiose equivalents)**	**Average Expression (Mut/WT)**
Consensus Design	21	5	23.8	2.2±0.6	14.0±10.4	1.0±0.2
Core Repacking	32	2	6.3	0.3±0.1	5.6±10.8	1.5±0.2
Helix Dipole	44	9	20.5	0.6±0.1	8.7±7.4	2.0±0.6
FoldX ΔΔG	43	6	14.0	1.9±0.5	7.9±8.5	0.5±0.1
Triad ΔΔG	47	5	10.6	1.9±0.5	-15.3±8.7	0.9±0.2
Glycine	51	5	9.8	1.2±0.6	-1.8±9.8	1.1±0.1
Proline	46	3	6.5	0.9±0.5	-21.4±10.5	1.3±0.2

Table II. *Evaluated mutations by design strategy*

		Mutation	ΔT$_{50}$ (°C)	ΔActivity (µM Cellobiose Equivalents)	Expression	Location
Experiment	**Consensus**	G293A	3.9±0.2	27.3	0.8	Core
		D13E	3.0±0.5	-9.3	1.6	Core
		E53D	2.7±0.7	8.2	0.9	Boundary
		T57N	1.1±0.0	47.0	0.3	Surface
		G189A	0.4±0.4	-3.4	1.2	Boundary
		I82L	-0.2±0.5	N/A	1.6	Core
		V101L	-0.5±0.3	N/A	2.3	Core
	Core	I82M	0.3±0.5	-5.2	1.3	Core
		V101I	0.5±0.4	16.4	1.7	Core
	Helix Dipole Stabilization	S318Q	0.5±0.2	2.3	0.9	Surface
		Y278F	1.0±0.5	-19.1	0.4	Boundary
		S318E	0.9±0.2	50.4	0.6	Surface
		N155E	0.5±0.3	5.6	4.9	Surface
		T80E	0.5±0.2	9.8	2.3	Surface
		S133R	0.4±0.2	3.4	1.8	Surface
		G239E	0.2±0.3	23.2	1.0	Boundary
		T156E	0.2±0.3	23.9	4.9	Boundary
		N155Q	0.1±0.1	-20.9	1.1	Surface
		S79Q	0.0±0.3	-19.3	1.4	Surface
		T80Q	-0.1±0.2	N/A	2.0	Surface
		S79E	-0.1±0.2	N/A	5.5	Surface
		A122E	-0.2±0.5	N/A	3.2	Surface
		G239Q	-0.9±0.2	N/A	0.9	Boundary
	FoldX	S318P	3.2±0.9	24.8	0.3	Surface
		D271F	3.1±1.1	-26.5	0.1	Boundary
		D271Y	2.7±0.4	32.5	0.4	Boundary
		S309L	1.5±0.3	2.4	0.8	Boundary
		N153D	0.5±0.9	2.5	0.2	Surface
		S79P	0.3±0.5	11.4	1.0	Surface
	Triad ΔΔG	K219Q	2.8±0.1	-31.7	1.2	Core
		S309F	2.7±0.1	-29.7	0.8	Boundary
		K219A	2.0±0.7	-26.7	1.4	Core
		S309L	1.5±0.3	2.4	0.8	Boundary
		S309W	0.4±0.1	9.1	0.5	Boundary
	Backbone Entropy Reduction	G293A	3.5±0.2	27.3	0.8	Core
		G189S	1.2±0.4	14.5	1.2	Boundary
		G239N	0.7±0.1	-18.8	1.1	Boundary
		G189A	0.4±0.4	-23.7	1.2	Boundary
		G239D	0.4±0.2	-8.1	1.3	Boundary
		G189E	0.0±0.2	N/A	2.0	Boundary
		G189K	-0.1±0.2	N/A	1.2	Boundary
		G64A	-0.1±0.2	-3.2	2.3	Core
		G239S	-0.7±0.1	N/A	0.9	Boundary
		S139P	1.8±0.6	-41.6	1.8	Surface
		N76P	0.8±0.5	-16.4	1.7	Surface
		T18P	0.2±0.1	-6.2	1.1	Surface
		I44C, G91C	-0.5±0.7	N/A	0.8	-

Table III. *Composition of combination constructs*

Construct	# of Mutations	T57N	N76P	T80E	S139P	N155E	G189S	K219Q	G239N	D271F	Y278F	G293A	S309F	S318P
1	1		x											
2	1				x									
3	1					x								
4	2		x		x									
5	4	x	x		x	x								
6	5	x	x		x	x	x							
7	5	x	x		x	x		x						
8	5	x	x		x	x			x					
9	5	x	x		x	x				x				
10	6	x	x		x	x				x	x			
11	6	x	x		x	x	x	x						
12	6	x	x		x	x	x		x					
13	6	x	x		x	x	x	x	x					
14	7	x	x		x	x	x			x	x			
15	8	x	x		x	x	x			x	x	x		
16	9	x	x	x	x	x	x			x	x	x		
17	10	x	x		x	x	x	x	x	x	x	x		
18	11	x	x	x	x	x	x	x	x	x	x	x		
19	12	x	x	x	x	x	x	x	x	x	x	x		x
20	13	x	x	x	x	x	x	x	x	x	x	x	x	x
Construct	**# of Mutations**	**T57N**	**T80E**	**N155E**	**G189S**	**G239E**	**D271Y**	**G293A**	**S309L/W**	**S318E/P**	**S79P**	**V101I**	**S133R**	**E53D**
s13pt1	13	x	x	x	x	x	x	x	x(L)	x(P)	x	x	x	x
s13pt2	13	x	x	x	x	x	x	x	x(W)	x(P)	x	x	x	x
s13pt3	13	x	x	x	x	x	x	x	x(L)	x(E)	x	x	x	x
s13pt4	13	x	x	x	x	x	x	x	x(W)	x(E)	x	x	x	x

Table IV. *Characterization of combination mutants*

Construct	# of Mutations	$T_{50,mut}$ (°C)	ΔT_{50} (°C)	T_{opt} (°C)	Activity (µM Cellobiose Equivalents)	ΔActivity (µM Cellobiose Equivalents)	Expression Level Relative to WT
WT	0	69.6±1.0	-	64	216.7±10.6	-	-
9 pt	9	81.5±0.4	10.0±0.8	74	184.2±5.7	-9.6	3.5
13 pt	13	86.8±0.6	13.8±0.9	80	157.4±1.4	-36.3	4.1
20 pt	20	90.0±0.1	16.8±0.7	82	146.6±9.8	-47.1	2.7
s13pt1	13	85.3±0.2	14.9±0.7	78	192.7±6.9	-1.0	4.1
s13pt2	13	85.6±0.4	15.4±0.7	78	221.1±5.9	27.4	4.1
s13pt3	13	83.0±0.4	12.0±0.8	75	231.2±5.6	37.5	5.9
s13pt4	13	83.9±0.2	12.2±0.7	75	262.6±11.5	68.8	3.7

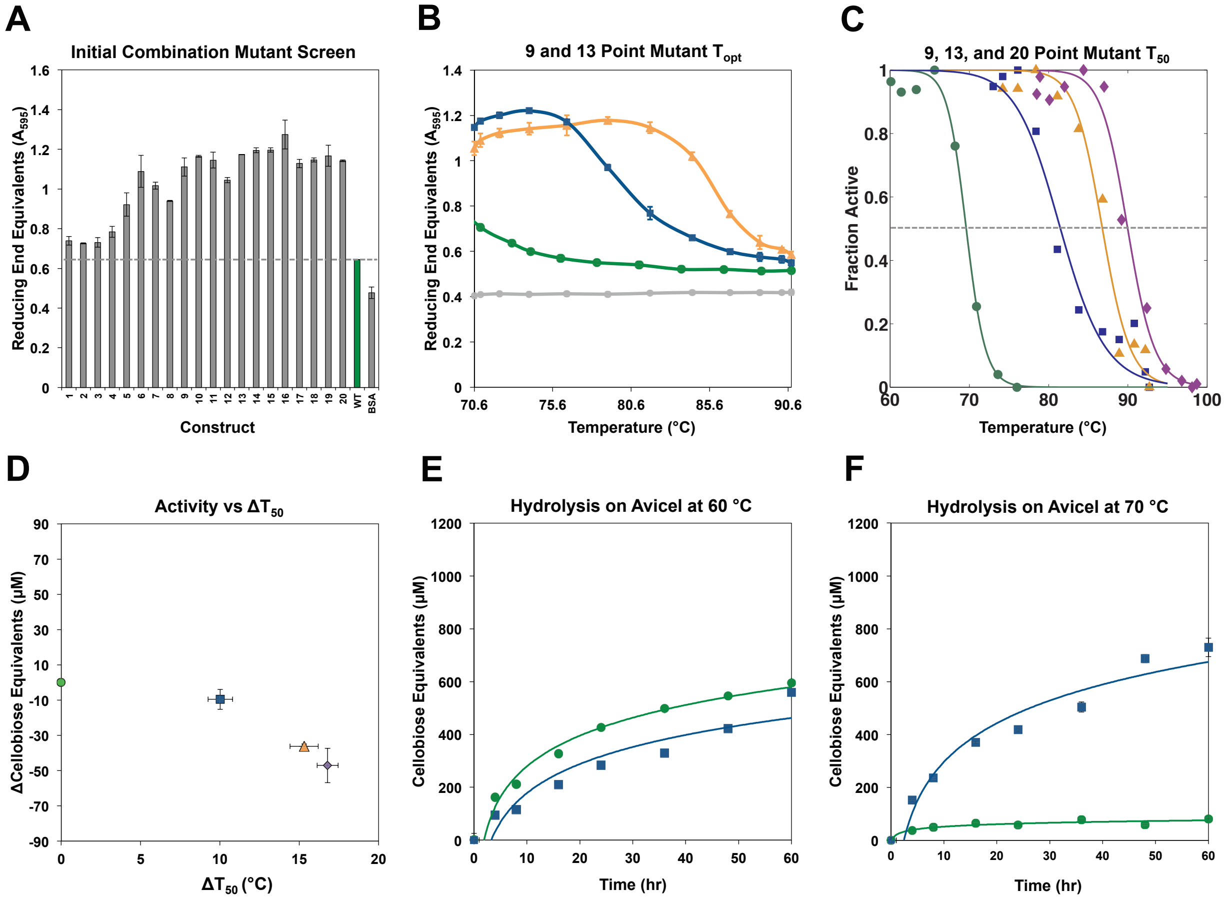


Figure 1. *Characterization of initial combination mutants.* (A) An activity screen performed on first generation combination mutants. WT is highlighted in green. The dashed line marking the WT activity is provided for reference. (B) The activity of the 9- and 13-point mutants across a gradient of temperatures. Activity curves for WT and BSA (gray circles) are provided for reference. (C) T_{50} curves for WT and the initial combination mutants. (D) Single point activity at 60 °C for the initial combination mutants. (E-F) Sixty-hour hydrolysis on Avicel at 60 or 70 °C. In all figures, WT is presented as green circles, the 9-point mutant as blue squares, the 13-point mutant as yellow triangles, and the 20-point mutant as purple diamonds.

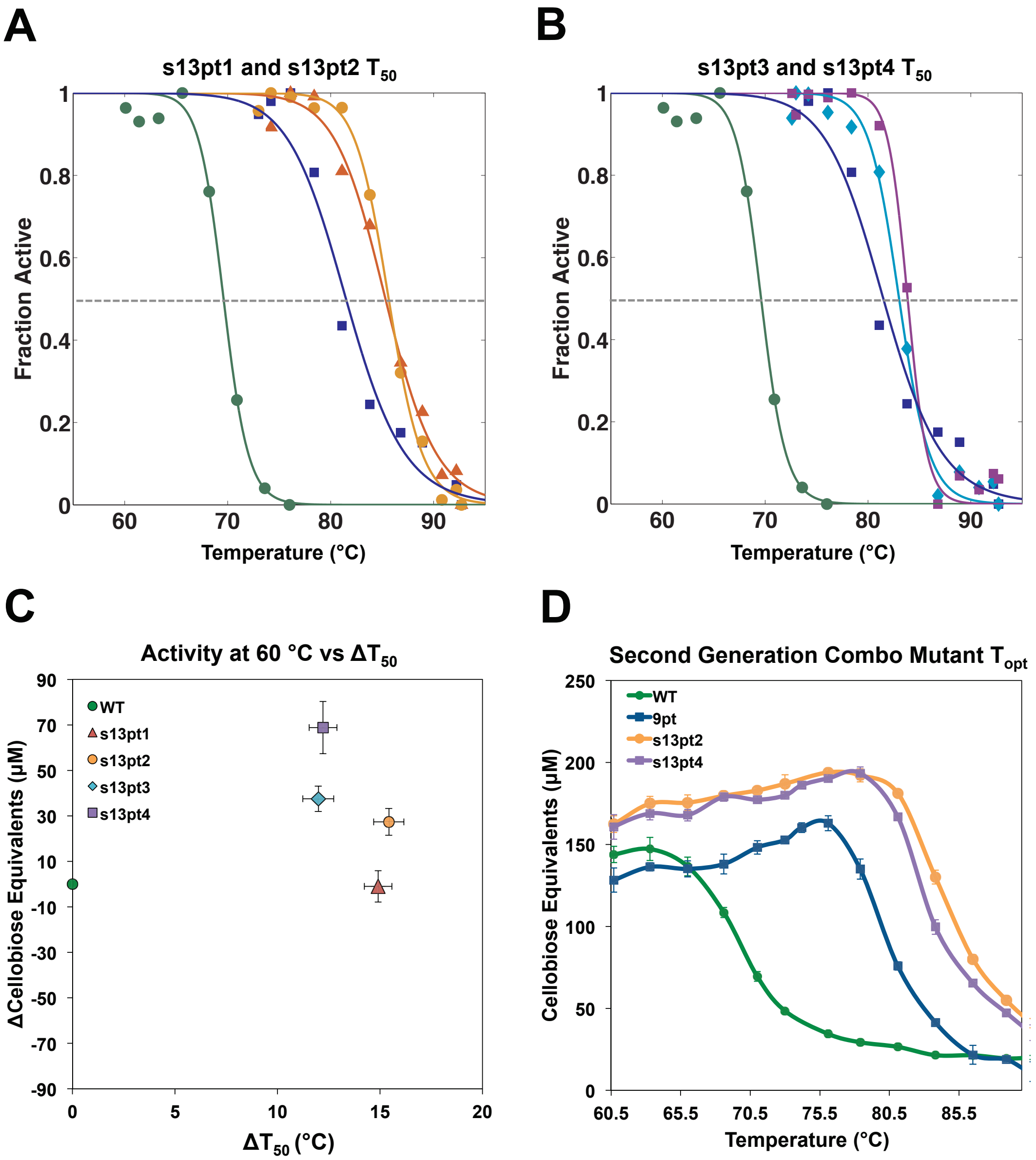


Figure 2. *Characterization of second-generation combination mutants.* (A-B) T_{50} curves for WT and the second-generation combination mutants. (C) Single point activity at 60 °C for the second-generation combination mutants. (D) The activity of WT, the 9-point mutant, s13pt2, and s13pt4 across a gradient of temperatures. In all figures, WT is presented as green circles, the 9-point mutant as blue squares, s13pt1 as red triangles, s13pt2 as yellow circles, s13pt3 as turquoise diamonds, and s13pt4 as purple squares.

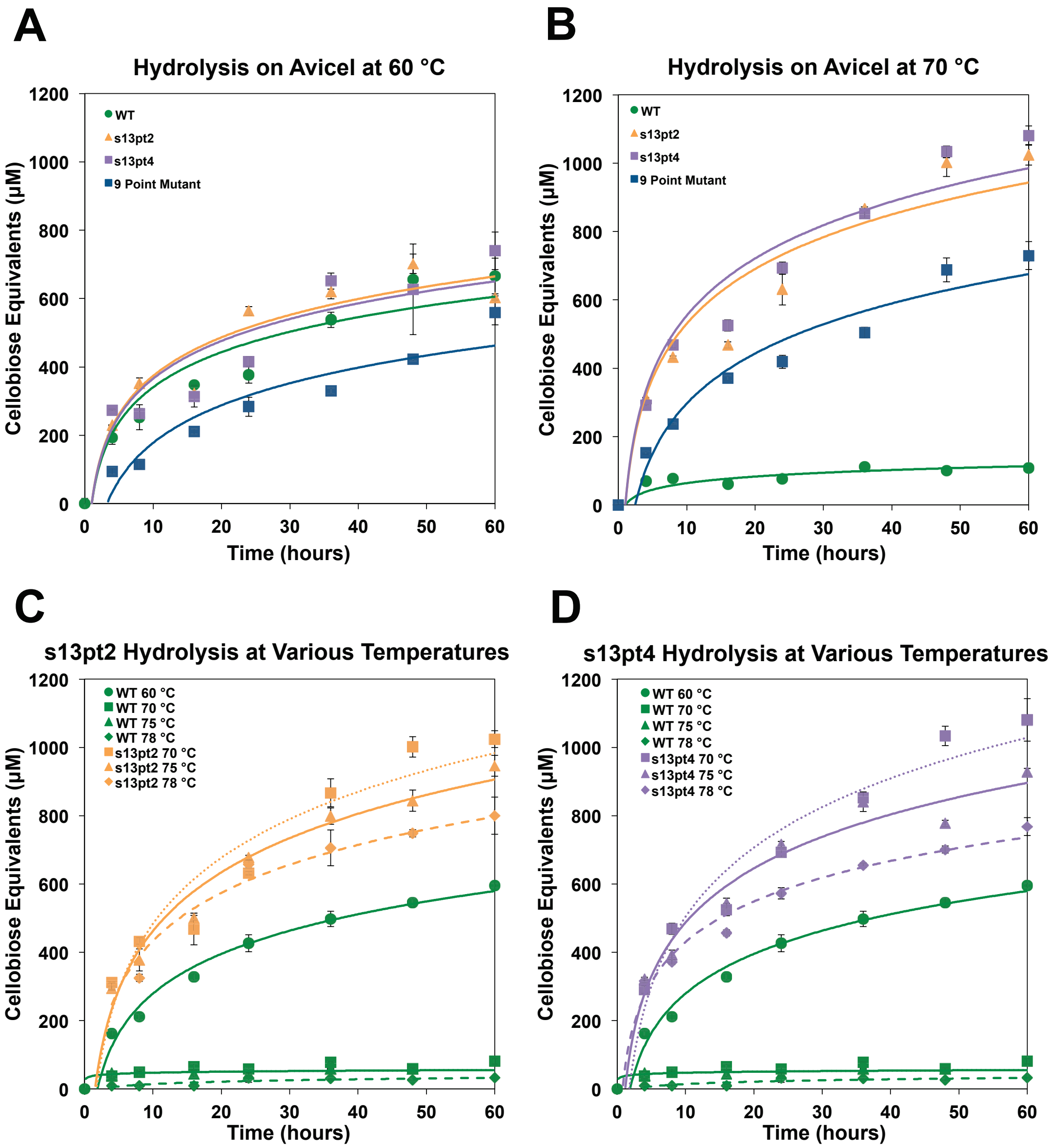


Figure 3. *Second-generation combination mutant 60-hour hydrolysis on Avicel.* Long-term hydrolysis experiments were performed on Avicel at (A) 60 and (B) 70 °C and compared to WT and the 9-point mutant. Panels C and D compare hydrolysis of WT, s13pt2, and s13pt4 at various temperatures. In all panels, WT is presented in green, s13pt2 in orange, and s13pt4 in purple.

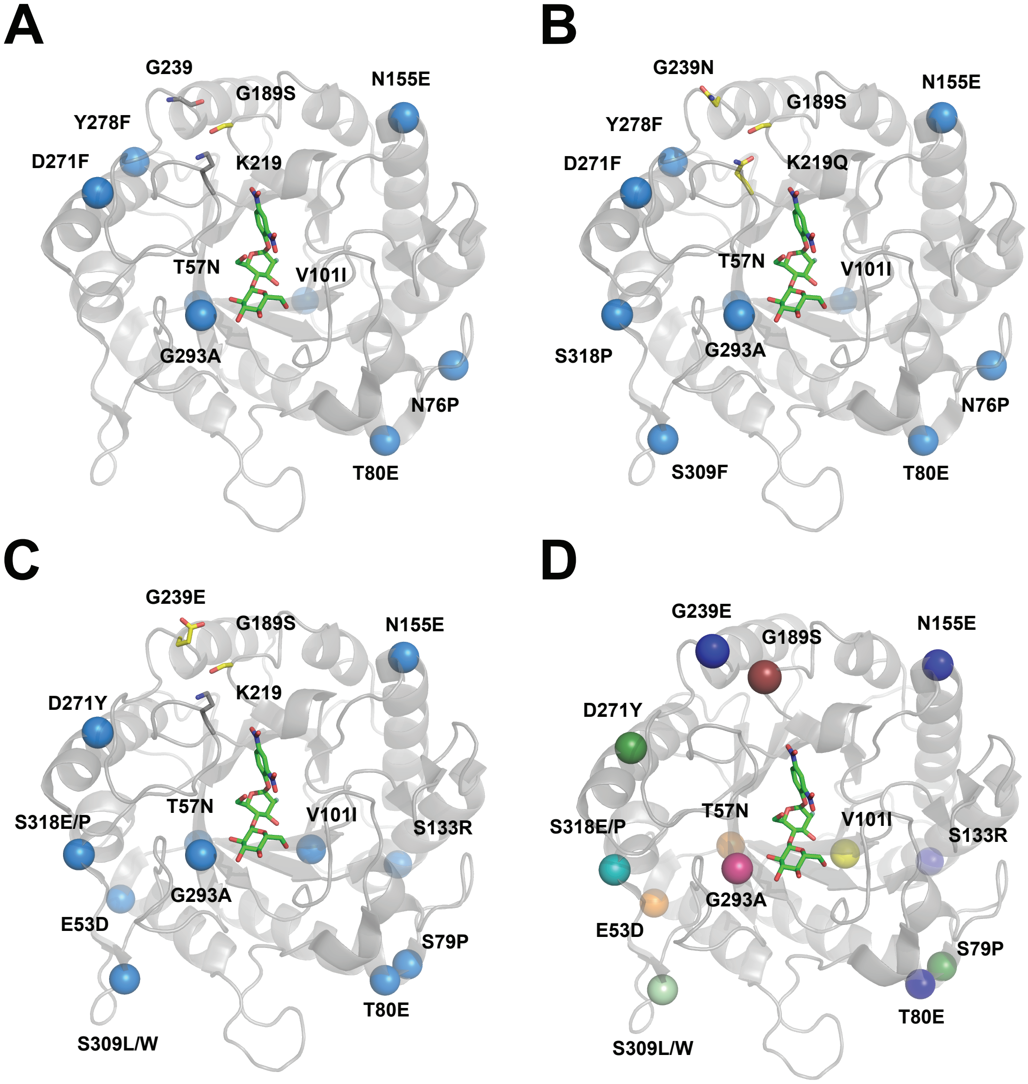


Figure 4. *Structural analysis of combination mutants***.** Figures outlining the position of mutations in (A) the 9-point mutant, (B) the 13-point mutant, and (C) s13pt2/4 are shown. Mutations appearing in a cleft potentially involved in substrate binding are shown as yellow sticks. (D) Mutations in s13pt2/4 are shown as spheres color coded by the experiment in which they were predicted: consensus design (orange), core repacking (yellow), helix dipole stabilization (dark blue), FoldX (dark green), glycine design (G189S), FoldX/Triad ΔΔG (light green), consensus/glycine design (hot pink), and FoldX/helix dipole stabilization (teal).

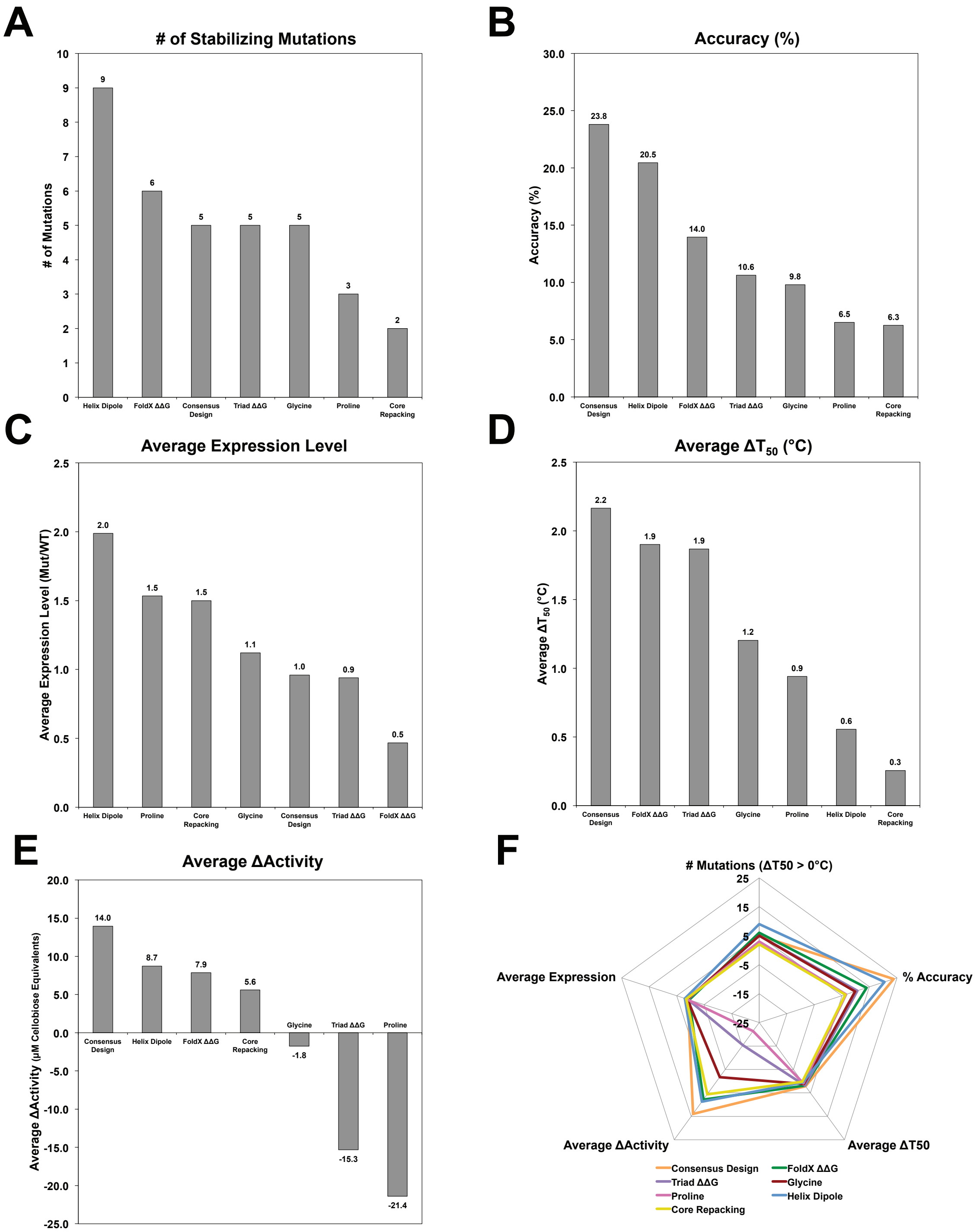


Figure 5. *Analysis of stabilization techniques.* The seven methods explored in this work are graphically ranked by (A) number of stabilizing mutations detected, (B) accuracy, (C) average expression level, (D) average ΔT_{50}, and (E) average change in activity. (F) Radar chart summarizing the performance of all design methods

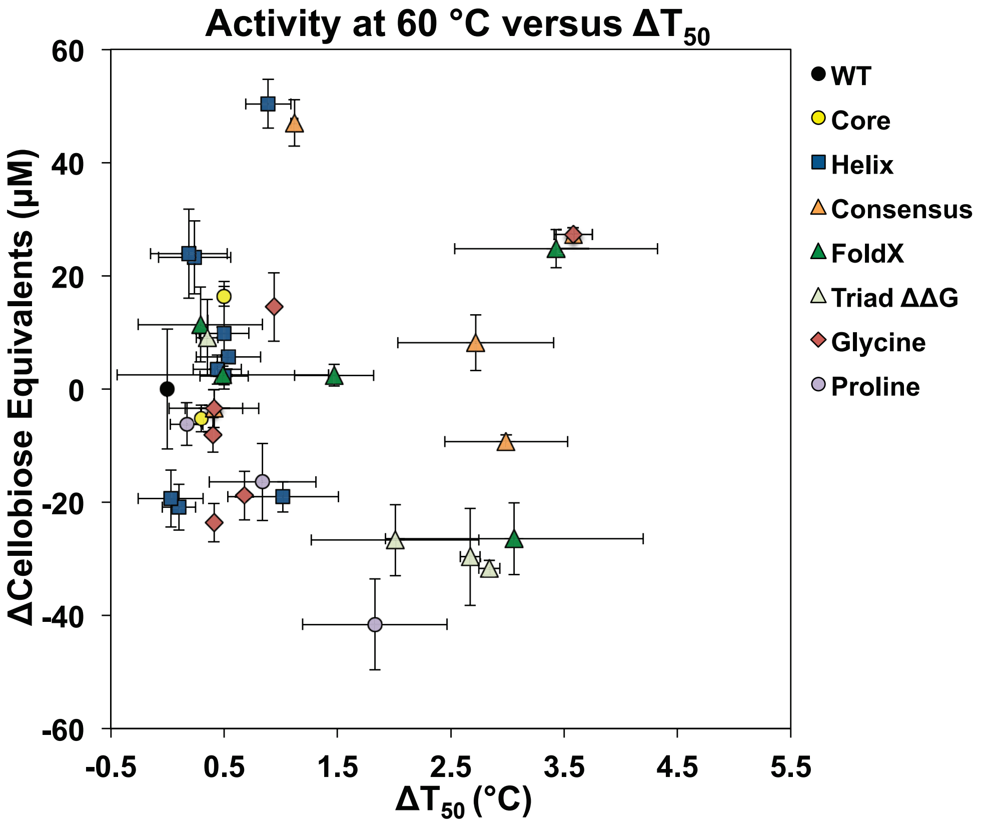


Figure 6. *Activity versus ΔT_{50}.* Plots are generated for consensus design (orange triangles), core repacking (yellow circles), helix dipole stabilization (blue squares), FoldX (dark green triangles), Triad ΔΔG (light green triangles), Gly → X_{AA} mutations (red diamonds), and X_{AA} → Pro mutations (purple circles). WT is shown in black at the 0,0 mark.

6.8 References

1. Naylor, R.L., Liska A.J., Burke M.B., Falcon W.P., Gaskell J.C., Rozelle S.D. and Cassman K.G. (2007) The ripple effect: biofuels, food security, and the environment. Environment: Science and Policy for Sustainable Development 49:30-43.
2. Headey, D. and Fan S. (2008) Anatomy of a crisis: the causes and consequences of surging food prices. Agricultural Economics 39:375-391.
3. De Fraiture, C., Giordano M. and Liao Y. (2008) Biofuels and implications for agricultural water use: blue impacts of green energy. Water Policy 10:67.
4. Klein ‐ Marcuschamer, D., Oleskowicz ‐ Popiel P., Simmons B.A. and Blanch H.W. (2012) The challenge of enzyme cost in the production of lignocellulosic biofuels. Biotechnology and Bioengineering 109:1083-1087.
5. Shafiee, S. and Topal E. (2009) When will fossil fuel reserves be diminished? Energy Policy 37:181-189.
6. Olah, G.A. (2005) Beyond Oil and Gas: The Methanol Economy. Angewandte Chemie International Edition 44:2636-2639.
7. Jacobson, M.Z. (2010) Short-term effects of controlling fossil-fuel soot, biofuel soot and gases, and methane on climate, Arctic ice, and air pollution health. Journal of Geophysical Research: Atmospheres 115:D14209.
8. Atlas, R.M. (1995) Bioremediation of petroleum pollutants. International Biodeterioration & Biodegradation 35:317-327.
9. Bang, G. (2010) Energy security and climate change concerns: Triggers for energy policy change in the United States? Energy Policy 38:1645-1653.
10. Kumar, R., Singh S. and Singh O. (2008) Bioconversion of lignocellulosic biomass: biochemical and molecular perspectives. Journal of Industrial Microbiology & Biotechnology 35:377-391.
11. Jordan, D.B., Bowman M.J., Braker J.D., Dien B.S., Hector R.E., Lee C.C., Mertens J.A. and Wagschal K. (2012) Plant cell walls to ethanol. Biochemical Journal 442:241-252.
12. Kazi, F.K., Fortman J.A., Anex R.P., Hsu D.D., Aden A., Dutta A. and Kothandaraman G. (2010) Techno-economic comparison of process technologies for biochemical ethanol production from corn stover. Fuel 89, Supplement 1:S20-S28.
13. Aden, A. and Foust T. (2009) Technoeconomic analysis of the dilute sulfuric acid and enzymatic hydrolysis process for the conversion of corn stover to ethanol. Cellulose 16:535-545.
14. Wu, I. and Arnold F.H. (2013) Engineered thermostable fungal Cel6A and Cel7A cellobiohydrolases hydrolyze cellulose efficiently at elevated temperatures. Biotechnology and Bioengineering 110:1874-1883.
15. Heinzelman, P., Snow C.D., Wu I., Nguyen C., Villalobos A., Govindarajan S., Minshull J. and Arnold F.H. (2009) A family of thermostable fungal cellulases created by structure-guided recombination. Proceedings of the National Academy of Sciences 106:5610-5615.
16. Heinzelman, P., Komor R., Kanaan A., Romero P., Yu X., Mohler S., Snow C. and Arnold F. (2010) Efficient screening of fungal cellobiohydrolase class I

enzymes for thermostabilizing sequence blocks by SCHEMA structure-guided recombination. Protein Engineering Design and Selection 23:871-880.

17. Komor, R.S., Romero P.A., Xie C.B. and Arnold F.H. (2012) Highly thermostable fungal cellobiohydrolase I (Cel7A) engineered using predictive methods. Protein Engineering Design and Selection 25:827-833.

18. Smith, M.A., Rentmeister A., Snow C.D., Wu T., Farrow M.F., Mingardon F. and Arnold F.H. (2012) A diverse set of family 48 bacterial glycoside hydrolase cellulases created by structure-guided recombination. FEBS Journal 279:4453-4465.

19. Viikari, L., Alapuranen M., Puranen T., Vehmaanperä J. and Siika-aho M. Thermostable Enzymes in Lignocellulose Hydrolysis. In: Olsson L, Ed. (2007) Biofuels. Springer Berlin Heidelberg, pp. 121-145.

20. Liu, C. and Wyman C.E. (2005) Partial flow of compressed-hot water through corn stover to enhance hemicellulose sugar recovery and enzymatic digestibility of cellulose. Bioresource Technology 96:1978-1985.

21. Ferdjani, S., Ionita M., Roy B., Dion M., Djeghaba Z., Rabiller C. and Tellier C. (2011) Correlation between thermostability and stability of glycosidases in ionic liquid. Biotechnology letters 33:1215-1219.

22. Mingardon, F., Bagert J.D., Maisonnier C., Trudeau D.L. and Arnold F.H. (2011) Comparison of family 9 cellulases from mesophilic and thermophilic bacteria. Applied and Environmental Microbiology 77:1436-1442.

23. Wolfenden, R., Snider M., Ridgway C. and Miller B. (1999) The temperature dependence of enzyme rate enhancements. Journal of the American Chemical Society 121:7419-7420.

24. Eijsink, V.G.H., Bjørk A., Gåseidnes S., Sirevåg R., Synstad B., Burg B.v.d. and Vriend G. (2004) Rational engineering of enzyme stability. Journal of Biotechnology 113:105-120.

25. Joo, J.C., Pohkrel S., Pack S.P. and Yoo Y.J. (2010) Thermostabilization of Bacillus circulans xylanase via computational design of a flexible surface cavity. Journal of Biotechnology 146:31-39.

26. Petsko, G.A. (2001) Structural basis of thermostability in hyperthermophilic proteins, or "There's more than one way to skin a cat". Methods in Enzymology 334:469-478.

27. Jaenicke, R. and Böhm G. (1998) The stability of proteins in extreme environments. Current Opinion in Structural Biology 8:738-748.

28. Daniel, R.M., Dines M. and Petach H.H. (1996) The denaturation and degradation of stable enzymes at high temperatures. Biochemical Journal 317:1-11.

29. Khan, S. and Vihinen M. (2010) Performance of protein stability predictors. Human Mutation 31:675-684.

30. Borgo, B. and Havranek J.J. (2012) Automated selection of stabilizing mutations in designed and natural proteins. Proceedings of the National Academy of Sciences 109:1494-1499.

31. Sullivan, B.J., Nguyen T., Durani V., Mathur D., Rojas S., Thomas M., Syu T. and Magliery T.J. (2012) Stabilizing proteins from sequence statistics: The interplay of conservation and correlation in triosephosphate isomerase stability. Journal of Molecular Biology 420:384-399.

32. Suominen, P.L., Mäntylä A.L., Karhunen T., Hakola S. and Nevalainen H. (1993) High frequency one-step gene replacement in *Trichoderma reesei*. II. Effects of deletions of individual cellulase genes. Molecular Genetics and Genomics 241:523-530.

33. Saloheimo, M. and Pakula T.M. (2012) The cargo and transport system: secreted proteins and protein secretion in *Trichoderma reesei* (*Hypocrea jecorina*). Microbiology 158:46-47.

34. Steipe, B., Schiller B., Pluckthun A. and Steinbacher S. (1994) Sequence statistics reliably predict stabilizing mutations in a protein domain. Journal of Molecular Biology 240:188-192.

35. Pace, C.N. (1990) Conformational stability of globular proteins. Trends in biochemical sciences 15:14-17.

36. Marshall, S.A., Morgan C.S. and Mayo S.L. (2002) Electrostatics significantly affect the stability of designed homeodomain variants. Journal of Molecular Biology 316:189-199.

37. Schymkowitz, J., Borg J., Stricher F., Nys R., Rousseau F. and Serrano L. (2005) The FoldX web server: an online force field. Nucleic Acids Research 33:392-388.

38. Rohl, C.A., Strauss C.E.M., Misura K.M.S. and Baker D. Protein Structure Prediction Using Rosetta. In: Ludwig B,Michael LJ, Eds. (2004) Methods in Enzymology. Academic Press, pp. 66-93.

39. Matthews, B.W., Nicholson H. and Becktel W.J. (1987) Enhanced protein thermostability from site-directed mutations that decrease the entropy of unfolding. Proceedings of the National Academy of Sciences 84:6663-6667.

40. Potapov, V., Cohen M. and Schreiber G. (2009) Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. Protein Engineering Design and Selection 22:553-560.

41. Maxwell, K.L. and Davidson A.R. (1998) Mutagenesis of a buried polar interaction in an SH3 Domain: sequence conservation provides the best prediction of stability effects. Biochemistry 37:16172-16182.

42. McConnell, A.D., Spasojevich V., Macomber J.L., Krapf I.P., Chen A., Sheffer J.C., Berkebile A., Horlick R.A., Neben S., King D.J. and Bowers P.M. (2012) An integrated approach to extreme thermostabilization and affinity maturation of an antibody. Protein Engineering Design and Selection.

43. Nelson, N. (1944) A photometric adaptation of the Somogyi method for the determination of glucose. J. biol. Chem 153:375-379.

44. Somogyi, M. (1952) Notes on sugar determination. Journal of Biological Chemistry 195:19-23.

APPENDIX A

262 *Hj*Cel5A Point Mutation Database

This appendix contains a series of tables that compile information from Chapters 2-5. The data presented here only address the 262 mutations that were constructed and experimentally tested for activity. Relative entropy, mutual information, and ΔΔG data for all possible mutations are available in the associated electronic files submitted with this work. For further information on this database, please consult Chapter 6.

Table I. *All predicted mutations*

Mutation	MSA	Core	Helix	FoldX	Rosetta ΔΔG	Glycine	Proline
V2P							x
R3P				x			
V7T		x					
N8V					x		
N8A					x		
N8P							x
I9L		x					
A10P							x
A10S	x	x					
D13E	x						
F14P							x
T18P							x
V27P							x
L31I		x					
K32P	x			x			
N33P	x						
Y40P							x
V51R			x				
N52R			x				
E53D	x						
E53R			x				
D54A				x			
D54C				x			
D54K				x			
D54L				x			
D54M				x			
D54N				x			
D54R				x			
M56F					x		
T57N	x						
R60V					x		
L61C		x					
G64P						x	x
G64A						x	
V69L							
V69M		x					
V69N		x					
N70P	x						
N76P							x
S79E			x				

Table I Cont'd. *All predicted mutations*

Mutation	MSA	Core	Helix	FoldX	Rosetta ΔΔG	Glycine	Proline
S79P				x			
S79Q			x				
T80E			x				
T80Q			x				
I82L	x						
I82M		x					
I82Q		x					
D86P							x
V89L		x					
V89M		x					
S94P							x
S94R			x				
A97P							x
V101I		x					
V101L	x						
D102P							x
H104P							x
V107N		x					
G112A						x	
G112E						x	
G112C						x	
G112D						x	
G112H						x	
G112I						x	
G112K						x	
G112L					x	x	
G112F					x	x	
G112M						x	
G112N						x	
G112Q						x	
G112R					x	x	
G112S						x	
G112T						x	
G112V						x	
G112W					x	x	
G112Y					x	x	
I114P							x
Q116N					x		
Q116D					x		
Q116W					x		

Table I Cont'd. *All predicted mutations*

Mutation	MSA	Core	Helix	FoldX	Rosetta ΔΔG	Glycine	Proline
T120D				x			
T120S			x				
N121E			x				
A122E			x				
A122Q			x				
T125P							x
S126P							x
S129P							x
S133R			x				
S134K			x				
Y135F	x						
A136P							x
S139P							x
W142I					x		
W142V					x		
W142F					x		
W142M					x		
W142H					x		
W142Y					x		
W142L					x		
W142T					x		
W142E					x		
F143M		x					
G144A						x	
G144P						x	x
G144D						x	
G144N						x	
I145V		x					
N147P							x
N153D				x			
I154M			x				
N155E			x				
N155Q			x				
T156E			x				
T156G					x		
V161I		x					
V161L		x					
E163P							x
V164A	x						
V165I	x	x					
I168H			x				

Table I Cont'd. *All predicted mutations*

Mutation	MSA	Core	Helix	FoldX	Rosetta ΔΔG	Glycine	Proline
R169P							x
N170P							x
N170R			x				
Q176P							x
Q186D					x		
Q186G				x			
Q186E					x		
Q186N					x		
Q186T	x						
S187P							x
A188C		x					
G189A	x					x	
G189E						x	
G189H						x	
G189K						x	
G189N						x	
G189Q						x	
G189R						x	
G189S						x	
F191W		x					
S193P							x
A197F			x				
A197M			x				
A199V			x				
S201K			x				
S201P							x
S201Q			x				
N205D	x						
N205P							x
V217I		x					
V217L		x					
K219S					x		
K219A					x		
K219Q					x		
K219E					x		
L221N		x					
D222P							x
S223P							x
A230P				x			
E231P							x
N236G					x		

Table I Cont'd. *All predicted mutations*

Mutation	MSA	Core	Helix	FoldX	Rosetta ΔΔG	Glycine	Proline
I237W					x		
I237Y					x		
I237F					x		
D238E			x				
D238Q			x				
G239A						x	
G239C						x	
G239D						x	
G239E			x			x	
G239I						x	
G239K						x	
G239L						x	
G239M						x	
G239N						x	
G239P						x	x
G239Q			x			x	
G239R						x	
G239S						x	
G239T						x	
G239V						x	
S242D			x				
S242Q			x				
P243E			x				
P243Q			x				
Q250P							x
Q250R			x				
R253Q					x		
A255G	x						
A255C		x					
A255T		x					
I256M		x					
L257I		x					
N264P							x
V265D				x			
S267P				x			
S267Q			x				
I269P							x
D271F				x			
D271Y				x			
I276H					x		
I276L	x						

Table I Cont'd. *All predicted mutations*

Mutation	MSA	Core	Helix	FoldX	Rosetta ΔΔG	Glycine	Proline
I276M		x					
Y278F			x				
Y278L			x				
N280P							x
N280R			x				
Q281P							x
Q281R			x				
N282Q					x		
N282R					x		
S283P				x			
G293A	x					x	
V302Y	x						
T304P							x
E305A				x			
E305C				x			
E305F				x	x		
E305G				x			
E305H				x	x		
E305I				x			
E305K				x			
E305L					x		
E305M				x			
E305N				x			
E305P							x
E305Q				x			
E305S				x			
E305T					x		
E305V				x			
E305Y					x		
T308P	x						
S309F					x		
S309L				x	x		
S309W					x		
G311N						x	
G311D						x	
D316A				x			
D316C				x			
D316G				x			
D316P				x			x
D316Q				x			
D316S				x			

Table I Cont'd. *All predicted mutations*

Mutation	MSA	Core	Helix	FoldX	Rosetta ΔΔG	Glycine	Proline
T317P							x
S318E			x				
S318F				x			
S318L				x			
S318M				x			
S318P				x			
S318Q			x				
S318W				x			
L319M		x					
S321K			x				
S321R			x				
S322R			x				
S322L				x			
L324F		x			x		
L324H					x		
L324M		x					
A325P							x
G328T						x	

Table II. *Multiple sequence alignment ΔΔG values for predicted mutations*

	444	323	233	195	29	10
Mutation	Sequence	Sequence	Sequence	Sequence	Sequence	Sequence
	ΔΔG	ΔΔG	ΔΔG	ΔΔG	ΔΔG	ΔΔG
V2P	1.0	1.1	-0.5	-1.1	10.4	13.1
R3P	2.0	1.8	1.0	0.6	12.8	13.3
V7T	3.0	2.5	2.5	2.4	1.7	13.8
N8V	13.3	4.7	13.3	13.4	13.7	13.8
N8A	2.9	2.8	3.0	13.4	13.7	13.8
N8P	13.3	13.4	13.3	13.4	13.7	13.8
I9L	-1.0	-1.7	-1.8	-1.8	-0.6	13.7
A10P	4.0	3.0	2.4	2.4	2.0	13.8
A10S	-0.1	-0.8	-1.0	-1.1	0.2	13.8
D13E	-1.2	-1.5	-1.9	-1.9	-1.3	13.8
F14P	5.8	13.6	13.5	13.6	13.7	13.8
T18P	12.2	3.7	9.8	9.6	12.2	13.6
V27P	1.6	1.5	10.4	10.5	10.4	12.9
L31I	9.8	0.0	-1.5	-0.2	1.0	12.2
K32P	9.8	0.0	-9.1	-9.2	-13.6	12.6
N33P	10.1	8.0	-4.0	-13.6	0.0	12.2
Y40P	2.1	-	-	-	-	12.9
V51R	2.4	11.5	10.6	8.5	12.1	13.7
N52R	-0.2	-1.6	1.1	9.2	1.4	13.5
E53D	-1.0	-1.2	-2.8	-11.0	-2.1	-13.7
E53R	3.6	2.8	8.4	0.0	10.4	0.0
D54A	1.0	0.4	0.3	0.1	0.9	13.8
D54C	12.0	11.6	11.7	11.7	12.1	13.8
D54K	-0.4	-1.1	-1.2	-1.3	-0.5	13.8
D54L	2.7	2.0	2.2	2.4	1.6	13.8
D54M	2.7	2.2	2.2	2.4	12.1	13.8
D54N	2.0	1.2	1.3	3.1	12.1	13.8
D54R	2.5	2.0	2.2	2.0	1.6	13.8
M56F	4.3	1.8	11.7	2.0	1.4	-0.9
T57N	-2.7	-2.5	-2.4	-2.5	-3.3	-1.4
R60V	13.7	13.8	13.8	13.8	13.8	13.8
L61C	4.7	12.9	13.0	13.0	13.2	13.8
G64P	3.3	2.7	2.4	2.4	11.8	13.7
G64A	0.2	-1.1	-0.9	-0.8	0.3	13.7
V69L	2.0	2.5	2.7	2.6	12.5	13.5
V69M	12.1	12.2	12.2	12.2	12.5	13.5
V69N	12.1	12.2	3.8	12.2	12.5	13.5

All values are reported in kcal mol^{-1}

Table II Cont'd. *Multiple sequence alignment ΔΔG values for predicted mutations*

	444	323	233	195	29	10
Mutation	**Sequence**	**Sequence**	**Sequence**	**Sequence**	**Sequence**	**Sequence**
	ΔΔG	**ΔΔG**	**ΔΔG**	**ΔΔG**	**ΔΔG**	**ΔΔG**
N70P	-1.4	-2.6	-1.6	8.5	-1.4	13.6
N76P	-2.3	-2.7	-2.7	-2.9	-12.5	-11.5
S79E	-0.3	-0.6	-0.7	-0.5	0.8	1.1
S79P	-0.2	-0.4	-0.6	-0.6	0.8	1.8
S79Q	0.4	0.1	-0.1	-0.2	0.8	13.3
T80E	2.7	2.5	2.5	2.2	0.9	13.1
T80Q	3.2	2.8	2.5	2.4	1.6	13.1
I82L	0.0	-3.9	-3.9	-4.2	-13.5	0.7
I82M	-0.9	-1.6	-1.6	-1.8	-10.4	12.2
I82Q	2.1	1.4	9.5	9.2	0.0	12.2
D86P	3.9	4.0	12.1	12.1	11.8	13.7
V89L	2.7	2.4	2.3	2.1	3.1	13.7
V89M	3.5	3.1	3.1	2.9	13.5	13.7
S94P	3.6	3.4	3.7	3.5	12.2	13.1
S94R	2.9	3.0	3.0	3.5	1.8	13.1
A97P	3.2	-0.3	10.9	9.6	13.0	13.6
V101I	-0.4	-0.5	-0.4	-0.3	-0.5	0.0
V101L	-1.0	-0.9	-0.8	-0.6	-1.1	13.1
D102P	13.7	13.7	13.7	13.7	13.7	13.8
H104P	13.7	13.8	13.8	13.8	13.8	13.8
V107N	13.2	13.0	13.0	13.0	13.2	13.8
G112A	0.1	0.0	0.1	0.2	0.4	0.7
G112E	0.1	0.0	-0.1	0.1	-0.3	0.7
G112C	11.3	11.2	11.3	11.4	11.5	12.9
G112D	1.8	1.4	1.6	1.4	11.5	12.9
G112H	1.2	0.8	0.6	0.5	1.1	12.9
G112I	1.9	1.6	1.6	1.4	11.5	12.9
G112K	-0.3	-0.5	-0.4	-0.4	0.4	1.4
G112L	11.3	11.2	11.3	11.4	11.5	12.9
G112F	11.3	11.2	11.3	11.4	11.5	12.9
G112M	11.3	11.2	11.3	11.4	11.5	12.9
G112N	-1.1	-0.8	-0.6	-0.6	-0.8	12.9
G112Q	-0.3	-0.5	-0.4	-0.3	-0.5	12.9
G112R	1.9	1.6	2.9	2.8	1.1	12.9
G112S	0.3	0.0	-0.5	-0.2	1.1	12.9
G112T	1.0	0.7	0.6	0.5	1.1	12.9
G112V	1.5	1.1	1.0	1.7	11.5	12.9

All values are reported in kcal mol^{-1}

Table II Cont'd. *Multiple sequence alignment ΔΔG values for predicted mutations*

Mutation	444 Sequence ΔΔG	323 Sequence ΔΔG	233 Sequence ΔΔG	195 Sequence ΔΔG	29 Sequence ΔΔG	10 Sequence ΔΔG
G112W	2.9	11.2	11.3	11.4	11.5	12.9
G112Y	0.6	0.2	0.1	0.3	1.1	12.9
I114P	13.5	13.6	13.6	13.6	13.6	13.7
Q116N	3.1	12.2	1.8	3.1	12.2	13.8
Q116D	0.8	3.1	0.6	0.6	0.4	13.8
Q116W	12.4	12.2	12.3	12.3	12.2	13.8
T120D	4.6	0.5	12.4	12.2	0.7	13.6
T120S	1.3	-0.2	1.1	1.0	0.3	1.4
N121E	4.1	3.7	3.3	3.0	-0.7	13.3
A122E	1.0	1.1	1.1	1.3	2.6	0.7
A122Q	3.0	3.0	3.7	3.5	13.0	12.9
T125P	12.5	11.3	11.5	11.5	11.8	12.6
S126P	11.8	11.6	11.6	11.7	12.1	13.5
S129P	11.6	11.4	11.1	11.1	11.5	12.9
S133R	2.5	2.7	2.4	2.5	0.0	12.2
S134K	0.0	0.0	0.0	0.0	0.0	0.0
Y135F	-0.8	-1.1	-1.1	-1.2	-1.3	13.8
A136P	4.2	3.9	3.6	3.5	1.9	13.1
S139P	0.1	0.0	0.1	0.2	1.6	13.6
W142I	-3.0	-4.3	-4.4	-4.2	-2.9	-0.3
W142V	-0.9	-2.3	-2.4	-2.2	-1.4	1.1
W142F	1.9	1.1	0.7	0.7	10.4	12.6
W142M	-0.5	-1.9	-1.9	-1.7	-0.7	1.1
W142H	10.4	9.1	9.1	9.2	10.4	12.6
W142Y	10.4	9.1	9.1	9.2	10.4	12.6
W142L	-0.8	-2.3	-2.5	-2.4	0.0	12.6
W142T	1.0	-0.5	-0.7	-0.7	10.4	12.6
W142E	10.4	9.1	9.1	9.2	10.4	12.6
F143M	13.7	13.8	13.8	13.8	13.8	13.8
G144A	4.5	4.5	13.3	13.2	13.3	0.0
G144P	13.3	13.2	13.3	13.2	13.3	13.8
G144D	0.7	0.3	0.5	0.4	0.6	13.8
G144N	3.8	3.8	3.8	3.6	13.3	13.8
I145V	0.0	0.7	0.9	0.7	0.9	0.0
N147P	13.8	13.8	13.8	13.8	13.8	13.8
N153D	-1.9	-2.6	-2.7	-2.9	-2.3	-0.3
I154M	4.3	3.4	2.9	2.6	12.4	13.8

All values are reported in kcal mol^{-1}

Table II Cont'd. *Multiple sequence alignment ΔΔG values for predicted mutations*

	444	323	233	195	29	10
Mutation	Sequence	Sequence	Sequence	Sequence	Sequence	Sequence
	ΔΔG	ΔΔG	ΔΔG	ΔΔG	ΔΔG	ΔΔG
N155E	0.5	-0.4	-0.8	-0.7	-0.3	13.3
N155Q	0.1	-0.5	-1.0	-1.2	0.0	13.3
T156E	1.2	0.6	0.4	0.4	1.6	13.3
T156G	2.5	1.9	2.0	2.6	12.1	13.3
V161I	3.1	2.7	2.7	3.3	12.4	13.8
V161L	1.6	1.2	1.3	2.2	12.4	13.8
E163P	0.4	9.1	9.1	9.2	0.0	12.2
V164A	-1.4	-1.7	-1.8	-1.7	-2.1	0.2
V165I	-0.5	-0.9	-1.1	-1.1	-0.5	13.8
I168H	13.6	13.7	13.7	13.7	13.7	13.8
R169P	13.7	13.8	13.8	13.8	13.8	13.8
N170P	10.8	9.1	9.5	8.5	11.8	13.6
N170R	0.3	-1.7	-1.5	-2.4	1.4	13.6
Q176P	13.1	9.4	8.4	0.0	13.0	13.8
Q186D	-0.3	-2.5	-2.2	-2.1	-11.1	12.6
Q186G	10.0	8.0	8.4	8.5	0.0	12.6
Q186E	2.3	0.0	0.0	0.0	0.0	12.6
Q186N	0.7	-1.4	-1.4	-1.4	-10.4	12.6
Q186T	-3.1	-5.2	-4.9	-4.7	-13.2	-0.8
S187P	12.3	12.0	12.1	12.1	12.2	13.8
A188C	13.5	5.6	13.7	13.8	13.7	13.8
G189A	-0.2	0.1	0.2	0.0	-11.8	-13.7
G189E	-0.1	0.1	0.0	0.1	0.0	-11.5
G189H	-1.5	-1.9	-2.0	-1.9	-12.8	0.0
G189K	1.1	0.7	2.5	2.3	0.0	0.0
G189N	2.1	1.7	1.8	2.3	0.0	0.0
G189Q	-0.1	-0.4	-0.3	-0.3	-11.5	0.0
G189R	10.9	10.8	10.9	10.8	-10.4	0.0
G189S	1.8	2.8	2.5	2.3	0.0	0.0
F191W	-1.1	-1.5	-1.8	-1.8	-1.0	13.7
S193P	3.0	3.6	11.7	11.7	1.7	13.3
A197F	-0.1	12.6	11.5	11.4	12.8	13.3
A197M	-0.8	3.9	11.5	11.4	12.8	13.3
A199V	3.4	1.1	11.4	11.4	1.8	13.8
S201K	-0.8	-1.2	-1.4	-1.9	-1.1	1.4
S201P	3.1	11.0	0.3	0.3	11.1	12.9
S201Q	1.0	0.2	0.3	0.0	11.1	12.9

All values are reported in kcal mol^{-1}

Table II Cont'd. *Multiple sequence alignment ΔΔG values for predicted mutations*

Mutation	444 Sequence ΔΔG	323 Sequence ΔΔG	233 Sequence ΔΔG	195 Sequence ΔΔG	29 Sequence ΔΔG	10 Sequence ΔΔG
N205D	0.6	-0.9	0.3	0.2	-1.8	13.8
N205P	11.9	2.6	2.7	2.5	11.8	13.8
V217I	1.6	1.3	1.5	2.1	2.8	13.7
V217L	2.8	2.9	2.6	2.8	13.3	2.2
K219S	12.2	11.8	11.8	11.9	12.2	13.8
K219A	3.7	3.0	3.4	3.3	12.2	13.8
K219Q	-1.1	-1.7	-1.8	-1.7	-1.2	13.8
K219E	1.3	0.6	0.7	1.9	1.1	13.8
L221N	4.7	5.0	5.3	5.2	13.6	13.8
D222P	5.9	13.8	13.8	13.8	13.7	13.8
S223P	3.9	3.6	3.5	3.7	2.7	13.7
A230P	0.4	0.2	0.0	-0.2	0.4	13.1
E231P	2.8	2.3	2.1	2.0	12.2	13.7
N236G	1.9	-1.4	-2.0	-2.1	12.5	13.8
I237W	5.3	-	-	-	13.1	13.3
I237Y	4.6	-	-	-	13.1	13.3
I237F	5.3	-	-	-	13.1	13.3
D238E	1.4	-1.9	-1.6	-1.4	11.5	13.6
D238Q	1.8	-3.8	-2.9	-2.8	1.1	13.6
G239A	-2.4	0.0	-9.1	-8.5	-11.1	12.2
G239C	9.9	8.7	0.0	0.0	0.0	12.2
G239D	-0.7	0.0	0.0	0.0	-11.1	-0.4
G239E	-0.2	8.7	0.0	0.0	-13.2	0.7
G239I	0.6	8.7	0.0	0.0	0.0	12.2
G239K	-0.8	8.7	0.0	0.0	0.0	12.2
G239L	0.0	8.7	0.0	0.0	0.0	12.2
G239M	9.9	8.7	0.0	0.0	0.0	12.2
G239N	0.3	8.7	0.0	0.0	-10.4	0.7
G239P	-0.3	8.7	0.0	0.0	-10.4	12.2
G239Q	-1.6	8.7	0.0	0.0	-11.1	12.2
G239R	1.1	8.7	0.0	0.0	0.0	12.2
G239S	-1.4	0.7	0.0	0.0	-10.4	12.2
G239T	-0.6	0.7	0.0	0.0	0.0	12.2
G239V	-1.9	8.7	0.0	0.0	0.0	12.2
S242D	1.4	0.9	9.1	8.5	10.4	11.5
S242Q	-1.5	-1.7	9.1	8.5	-1.4	11.5
P243E	1.2	-0.2	-0.2	0.3	10.4	13.5

All values are reported in kcal mol^{-1}

Table II Cont'd. *Multiple sequence alignment ΔΔG values for predicted mutations*

	444	323	233	195	29	10
Mutation	Sequence	Sequence	Sequence	Sequence	Sequence	Sequence
	ΔΔG	ΔΔG	ΔΔG	ΔΔG	ΔΔG	ΔΔG
P243Q	11.7	10.4	10.0	9.9	10.4	13.5
Q250P	3.8	10.0	11.8	11.9	11.1	12.9
Q250R	1.3	0.6	1.5	1.8	11.1	12.9
R253Q	1.5	0.0	0.8	0.6	1.1	13.8
A255G	-0.6	-1.0	-1.1	-1.3	-1.0	13.8
A255C	4.8	4.4	12.4	12.3	12.4	13.8
A255T	4.8	12.4	12.4	12.3	12.4	13.8
I256M	1.5	12.4	1.6	2.8	1.1	1.6
L257I	1.3	1.2	1.2	1.5	1.5	13.3
N264P	2.4	13.3	2.4	2.4	1.9	13.8
V265D	-0.9	-2.3	-2.8	-2.5	0.4	13.5
S267P	4.6	8.7	3.8	8.5	1.8	13.8
S267Q	0.4	8.7	0.0	8.5	-0.2	13.8
I269P	2.4	-9.1	8.4	0.0	10.4	12.2
D271F	0.1	-11.0	-11.0	-11.1	-10.4	0.0
D271Y	-1.3	0.0	-8.4	0.0	-11.8	-2.1
I276H	3.1	10.1	10.0	9.6	0.0	11.5
I276L	-2.5	-3.4	-3.6	-4.0	-13.7	11.5
I276M	1.5	0.5	0.2	1.1	0.0	11.5
Y278F	1.3	1.5	1.7	1.9	2.1	1.1
Y278L	3.9	4.0	4.9	4.7	2.1	13.3
N280P	3.2	2.4	2.9	2.8	12.1	13.6
N280R	2.1	1.4	1.3	1.2	0.9	13.6
Q281P	12.1	11.9	11.8	11.9	11.5	12.9
Q281R	2.3	1.8	1.7	2.2	1.1	12.9
N282Q	1.6	4.4	4.5	4.4	13.6	13.7
N282R	13.4	13.5	13.6	13.6	13.6	13.7
S283P	3.8	3.8	3.6	4.6	12.6	13.5
G293A	-2.0	-2.0	-2.2	-2.3	-2.0	0.0
V302Y	3.0	2.3	-13.6	-13.7	1.6	13.1
T304P	11.6	10.9	9.5	9.6	12.4	12.9
E305A	3.3	2.2	0.7	1.1	13.3	13.7
E305C	12.1	2.2	-1.0	1.1	13.3	13.7
E305F	2.6	10.2	9.8	9.6	13.3	13.7
E305G	4.4	10.2	0.0	0.0	13.3	13.7
E305H	12.1	10.2	9.8	9.6	1.7	13.7
E305I	2.0	0.6	9.8	9.6	13.3	13.7

All values are reported in kcal mol^{-1}

Table II Cont'd. *Multiple sequence alignment ΔΔG values for predicted mutations*

	444	323	233	195	29	10
Mutation	**Sequence**	**Sequence**	**Sequence**	**Sequence**	**Sequence**	**Sequence**
	ΔΔG	**ΔΔG**	**ΔΔG**	**ΔΔG**	**ΔΔG**	**ΔΔG**
E305K	12.1	10.2	9.8	9.6	13.3	13.7
E305L	0.0	-1.1	1.4	1.1	13.3	13.7
E305M	0.3	1.1	9.8	9.6	13.3	13.7
E305N	3.7	10.2	-2.7	-2.9	2.1	13.7
E305P	12.1	10.2	9.8	9.6	13.3	13.7
E305Q	3.0	10.2	9.8	9.6	1.2	2.2
E305S	4.4	10.2	-3.3	-3.5	13.3	13.7
E305T	4.4	10.2	-2.0	-2.2	13.3	13.7
E305V	-0.2	-0.7	0.7	0.4	2.1	13.7
E305Y	12.1	10.2	0.7	9.6	13.3	13.7
T308P	0.6	2.6	-4.6	-5.2	12.4	13.5
S309F	3.9	0.3	0.3	-0.4	10.4	0.0
S309L	2.5	1.5	0.0	-0.7	10.4	0.0
S309W	11.6	-3.3	1.4	0.7	10.4	0.0
G311N	1.3	2.6	2.9	2.1	11.5	1.6
G311D	-0.8	3.6	1.5	0.7	-1.2	13.1
D316A	0.7	12.0	0.3	-0.7	12.2	13.7
D316C	12.2	12.0	10.4	9.6	12.2	13.7
D316G	2.8	12.0	1.0	0.0	12.2	13.7
D316P	4.4	12.0	2.1	1.1	12.2	13.7
D316Q	1.6	12.0	10.4	9.6	12.2	13.7
D316S	0.6	12.0	0.3	-0.7	1.8	13.7
T317P	1.0	9.1	1.6	0.9	0.2	13.6
S318E	-0.5	0.0	11.0	11.1	0.0	1.1
S318F	10.4	9.1	11.0	11.1	10.4	12.6
S318L	1.3	0.0	1.9	2.6	-2.1	1.1
S318M	2.6	9.1	2.6	2.6	10.4	12.6
S318P	0.1	-1.8	1.0	1.0	10.4	1.1
S318Q	-0.3	-0.5	11.0	11.1	10.4	12.6
S318W	2.6	9.1	11.0	11.1	10.4	12.6
L319M	11.9	3.9	9.5	9.2	13.5	12.6
S321K	-1.3	-0.9	1.2	1.4	-0.5	0.9
S321R	0.2	9.6	1.1	1.0	1.1	1.6
S322R	-0.2	0.3	9.1	0.0	-0.7	13.3
S322L	2.0	1.4	9.1	0.0	10.4	13.3
L324F	0.8	8.0	0.0	0.0	-1.1	0.7
L324H	3.3	8.0	-8.4	-8.5	10.4	13.3

All values are reported in kcal mol^{-1}

Table II Cont'd. *Multiple sequence alignment ΔΔG values for predicted mutations*

Mutation	444 Sequence ΔΔG	323 Sequence ΔΔG	233 Sequence ΔΔG	195 Sequence ΔΔG	29 Sequence ΔΔG	10 Sequence ΔΔG
L324M	2.4	0.0	0.0	0.0	10.4	13.3
A325P	-0.9	1.1	1.9	10.9	0.0	13.3
G328T	0.0	-9.1	-1.7	-2.2	10.4	0.0

All values are reported in kcal mol^{-1}

Table III. *Relative entropy values for predicted mutations*

Mutation	444 Sequence RE	323 Sequence RE	233 Sequence RE	195 Sequence RE	29 Sequence RE	10 Sequence RE
V2P	0.0000	0.0194	0.0340	0.0505	0.0000	0.0000
R3P	-0.0049	-0.0132	-0.0117	-0.0131	0.0000	0.0000
V7T	-0.0183	-0.0077	-0.0028	-0.0024	0.0584	0.0000
N8V	0.0000	-0.0149	0.0000	0.0000	0.0000	0.0000
N8A	-0.0135	-0.0083	-0.0110	0.0000	0.0000	0.0000
N8P	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
I9L	0.5600	0.4619	0.7163	0.7648	0.5243	0.0000
A10P	-0.0138	-0.0160	-0.0142	-0.0151	0.0310	0.0000
A10S	0.8519	1.0903	1.4179	1.5246	0.6359	0.0000
D13E	1.7627	1.9733	2.1475	2.1600	1.7463	0.0000
F14P	-0.0072	0.0000	0.0000	0.0000	0.0000	0.0000
T18P	0.0000	-0.0087	0.0000	0.0000	0.0000	0.0000
V27P	-0.0162	-0.0151	0.0000	0.0000	0.0000	0.0000
L31I	0.0000	0.0000	0.1875	0.2244	0.1566	0.0000
K32P	0.0000	0.0000	-0.0148	-0.0153	2.9526	0.0000
N33P	0.0000	0.0000	2.2320	2.4889	0.0000	0.0000
Y40P	-0.0150	0.0000	0.0000	0.0000	0.0000	0.0000
V51R	-0.0163	0.0000	0.0000	0.0000	0.0000	0.0000
N52R	0.0460	0.0966	0.0341	0.0000	-0.0077	0.0000
E53D	0.3212	0.1966	1.2326	1.9206	0.4305	2.4688
E53R	-0.0071	-0.0085	0.0000	0.0000	0.0000	0.0000
D54A	0.0406	0.0260	0.0394	0.0623	0.0140	0.0000
D54C	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
D54K	0.5391	0.5823	0.6836	0.7431	0.3673	0.0000
D54L	-0.0284	-0.0288	-0.0261	-0.0229	-0.0350	0.0000
D54M	-0.0045	-0.0062	-0.0062	-0.0074	0.0000	0.0000
D54N	-0.0205	-0.0188	-0.0191	-0.0127	0.0000	0.0000
D54R	-0.0161	-0.0163	-0.0160	-0.0163	-0.0086	0.0000
M56F	0.1251	0.0709	0.0521	0.0584	0.1557	1.2084
T57N	2.4459	2.4495	2.4246	2.4184	2.6724	2.0639
R60V	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
L61C	-0.0048	0.0000	0.0000	0.0000	0.0000	0.0000
G64P	-0.0116	-0.0085	-0.0101	-0.0111	0.0000	0.0000
G64A	0.1022	0.1348	0.0948	0.0973	0.0689	0.0000
V69L	-0.0339	-0.0287	-0.0259	-0.0280	0.0000	0.0000
V69M	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
V69N	0.0000	0.0000	-0.0115	0.0000	0.0000	0.0000

Table III Cont'd. *Relative entropy values for predicted mutations*

Mutation	444 Sequence RE	323 Sequence RE	233 Sequence RE	195 Sequence RE	29 Sequence RE	10 Sequence RE
N70P	1.9534	2.4510	1.6585	0.0000	2.0282	0.0000
N76P	0.1797	0.2738	0.2812	0.2900	0.5064	0.0821
S79E	0.2066	0.2718	0.2843	0.2055	0.0482	0.2250
S79P	0.2061	0.2753	0.3272	0.3450	0.0884	0.0821
S79Q	0.0734	0.1142	0.1476	0.1753	0.0996	0.0000
T80E	-0.0235	-0.0238	-0.0238	-0.0230	0.0041	0.0000
T80Q	-0.0143	-0.0138	-0.0108	-0.0111	-0.0047	0.0000
I82L	1.2095	1.2689	1.3189	1.4042	1.5745	0.0049
I82M	0.0366	0.0659	0.0737	0.0660	0.0170	0.0000
I82Q	-0.0068	-0.0081	0.0000	0.0000	0.0000	0.0000
D86P	-0.0107	-0.0085	0.0000	0.0000	0.0000	0.0000
V89L	-0.0310	-0.0231	-0.0156	-0.0029	-0.0350	0.0000
V89M	0.0033	0.0138	0.0172	0.0274	0.0000	0.0000
S94P	-0.0107	-0.0125	-0.0101	-0.0110	0.0000	0.0000
S94R	-0.0147	-0.0148	-0.0142	-0.0111	-0.0086	0.0000
A97P	-0.0161	0.3337	0.0000	0.0000	0.0000	0.0000
V101I	0.4448	0.4937	0.4692	0.5174	0.3982	1.0190
V101L	0.8293	0.7653	0.7493	0.6502	0.8754	0.0000
D102P	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
H104P	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
V107N	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
G112A	0.0282	0.0204	0.0205	0.0174	0.0140	0.2535
G112E	0.0168	0.0136	0.0259	0.0192	0.1039	0.2250
G112C	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
G112D	-0.0201	-0.0212	-0.0209	-0.0213	0.0000	0.0000
G112H	0.0058	0.0179	0.0310	0.0447	0.0164	0.0000
G112I	-0.0204	-0.0225	-0.0230	-0.0237	0.0000	0.0000
G112K	0.0442	0.0687	0.0633	0.0724	-0.0038	0.0317
G112L	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
G112F	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
G112M	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
G112N	0.3838	0.1905	0.1248	0.1634	0.3335	0.0000
G112Q	0.1160	0.1388	0.1290	0.1399	0.2541	0.0000
G112R	-0.0157	-0.0163	-0.0101	-0.0111	-0.0086	0.0000
G112S	-0.0217	-0.0079	0.0506	0.0205	-0.0327	0.0000
G112T	-0.0197	-0.0162	-0.0130	-0.0070	-0.0184	0.0000
G112V	-0.0207	-0.0202	-0.0188	-0.0200	0.0000	0.0000

Table III Cont'd. *Relative entropy values for predicted mutations*

	444	323	233	195	29	10
Mutation	**Sequence**	**Sequence**	**Sequence**	**Sequence**	**Sequence**	**Sequence**
	RE	**RE**	**RE**	**RE**	**RE**	**RE**
G112W	-0.0037	0.0000	0.0000	0.0000	0.0000	0.0000
G112Y	0.0178	0.0420	0.0666	0.0455	0.0008	0.0000
I114P	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Q116N	-0.0195	0.0000	-0.0193	-0.0182	0.0000	0.0000
Q116D	0.0804	-0.0205	0.0904	0.0927	0.2717	0.0000
Q116W	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
T120D	-0.0096	0.1272	0.0000	0.0000	0.1197	0.0000
T120S	0.0101	0.3503	0.0380	0.0468	0.1745	0.1619
N121E	-0.0125	-0.0097	-0.0153	-0.0173	0.1039	0.0000
A122E	0.0488	0.0521	0.0619	0.0257	-0.0218	0.2250
A122Q	-0.0145	-0.0144	-0.0131	-0.0138	0.0000	0.0000
T125P	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
S126P	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
S129P	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
S133R	-0.0147	-0.0147	-0.0159	-0.0150	0.0305	0.0000
S134K	0.1456	0.0485	-0.0033	-0.0107	-0.0038	1.5838
Y135F	1.8479	2.0614	2.0753	2.1415	2.2827	0.0000
A136P	-0.0107	-0.0124	-0.0141	-0.0150	-0.0084	0.0000
S139P	0.1842	0.1593	0.1239	0.0888	0.0310	0.0000
W142I	1.6046	1.5853	1.5984	1.5847	1.5124	0.7260
W142V	0.0261	0.0432	0.0426	0.0453	0.1231	0.0571
W142F	-0.0105	-0.0083	-0.0101	-0.0111	0.0000	0.0000
W142M	0.0539	0.0751	0.0636	0.0556	0.0818	0.1558
W142H	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
W142Y	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
W142L	-0.0186	-0.0042	0.0086	0.0191	-0.0350	0.0000
W142T	-0.0188	-0.0207	-0.0212	-0.0216	0.0000	0.0000
W142E	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
F143M	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
G144A	-0.0145	-0.0138	0.0000	0.0000	0.0000	0.0000
G144P	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
G144D	0.5391	0.7400	0.6151	0.7281	0.6151	0.0000
G144N	-0.0205	-0.0198	-0.0200	-0.0211	0.0000	0.0000
I145V	0.1352	0.0247	0.0050	0.0133	0.0137	1.0900
N147P	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
N153D	0.8655	0.9601	0.9145	0.9970	0.8221	0.7729
I154M	-0.0065	-0.0060	-0.0069	-0.0072	0.0000	0.0000

Table III Cont'd. *Relative entropy values for predicted mutations*

Mutation	444 Sequence RE	323 Sequence RE	233 Sequence RE	195 Sequence RE	29 Sequence RE	10 Sequence RE
N155E	-0.0154	-0.0051	-0.0042	-0.0121	0.1039	0.0000
N155Q	0.0360	0.0336	0.0458	0.0600	0.0996	0.0000
T156E	-0.0073	0.0041	0.0254	0.0331	-0.0218	0.0000
T156G	-0.0184	-0.0187	-0.0186	-0.0164	0.0000	0.0000
V161I	-0.0191	-0.0181	-0.0174	-0.0130	0.0000	0.0000
V161L	-0.0339	-0.0340	-0.0350	-0.0280	0.0000	0.0000
E163P	-0.0150	0.0000	0.0000	0.0000	0.0000	0.0000
V164A	1.6565	1.8003	1.8036	1.7731	1.3733	0.7842
V165I	1.3439	1.6234	1.8030	1.8106	1.2880	0.0000
I168H	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
R169P	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
N170P	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
N170R	-0.0001	0.0059	0.0187	0.0136	-0.0086	0.0000
Q176P	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Q186D	-0.0181	-0.0162	-0.0156	-0.0142	0.0120	0.0000
Q186G	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Q186E	-0.0084	-0.0095	-0.0117	-0.0130	0.0000	0.0000
Q186N	-0.0201	-0.0198	-0.0217	-0.0222	-0.0195	0.0000
Q186T	1.4144	1.3478	1.3053	1.2419	1.2351	1.7336
S187P	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
A188C	0.0000	-0.0044	0.0000	0.0000	0.0000	0.0000
G189A	0.0272	-0.0109	-0.0115	-0.0048	0.1235	2.4942
G189E	0.0068	-0.0172	-0.0118	-0.0158	0.0000	0.0432
G189H	0.7588	0.9341	1.0730	0.9337	1.0896	0.0000
G189K	-0.0262	-0.0267	-0.0122	-0.0136	0.0000	0.0000
G189N	-0.0161	-0.0175	-0.0168	-0.0127	0.0000	0.0000
G189Q	0.0423	0.0485	0.0385	0.0429	0.0996	0.0000
G189R	0.0000	0.0000	0.0000	0.0000	-0.0086	0.0000
G189S	-0.0225	-0.0105	-0.0131	-0.0146	0.0000	0.0000
F191W	2.8631	3.2197	3.4351	3.4166	3.0713	0.0000
S193P	-0.0160	-0.0100	0.0000	0.0000	0.0310	0.0000
A197F	0.0822	0.0000	0.0000	0.0000	0.0000	0.0000
A197M	0.5257	-0.0077	0.0000	0.0000	0.0000	0.0000
A199V	-0.0177	0.0921	0.0000	0.0000	0.0137	0.0000
S201K	0.1487	0.1893	0.0633	0.0999	0.2159	0.0317
S201P	-0.0089	0.0000	-0.0153	-0.0162	0.0000	0.0000
S201Q	-0.0095	0.0128	-0.0129	-0.0134	0.0000	0.0000

Table III Cont'd. *Relative entropy values for predicted mutations*

Mutation	444 Sequence RE	323 Sequence RE	233 Sequence RE	195 Sequence RE	29 Sequence RE	10 Sequence RE
N205D	0.4365	1.6287	0.1278	0.1402	2.3277	0.0000
N205P	0.0000	-0.0158	-0.0158	-0.0162	0.0000	0.0000
V217I	0.0142	0.0265	0.0187	-0.0190	-0.0219	0.0000
V217L	-0.0329	-0.0305	-0.0337	-0.0315	0.0000	0.0049
K219S	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
K219A	-0.0124	-0.0138	-0.0111	-0.0123	0.0000	0.0000
K219Q	1.8246	2.1894	2.2762	2.4069	1.9725	0.0000
K219E	-0.0024	0.0044	-0.0003	-0.0236	0.0041	0.0000
L221N	-0.0160	-0.0142	-0.0114	-0.0127	0.0000	0.0000
D222P	-0.0074	0.0000	0.0000	0.0000	0.0000	0.0000
S223P	-0.0150	-0.0158	-0.0159	-0.0149	-0.0084	0.0000
A230P	0.2568	0.3139	0.3815	0.4662	0.3203	0.0000
E231P	-0.0159	-0.0162	-0.0153	-0.0157	0.0000	0.0000
N236G	-0.0105	1.5782	1.5945	1.6808	0.0000	0.0000
I237W	-0.0037	0.0000	0.0000	0.0000	0.0000	0.0000
I237Y	-0.0103	0.0000	0.0000	0.0000	0.0000	0.0000
I237F	-0.0080	0.0000	0.0000	0.0000	0.0000	0.0000
D238E	-0.0228	-0.0238	-0.0176	-0.0187	0.0000	0.0000
D238Q	-0.0132	0.1942	0.2139	0.2484	-0.0047	0.0000
G239A	0.4490	0.3726	2.8767	2.8767	0.0326	0.0000
G239C	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
G239D	-0.0061	0.3656	0.0000	0.0000	0.0303	0.8577
G239E	-0.0228	0.0000	0.0000	0.0000	1.4152	0.1126
G239I	-0.0217	0.0000	0.0000	0.0000	0.0000	0.0000
G239K	-0.0139	0.0000	0.0000	0.0000	0.0000	0.0000
G239L	-0.0337	0.0000	0.0000	0.0000	0.0000	0.0000
G239M	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
G239N	-0.0222	0.0000	0.0000	0.0000	-0.0156	0.1224
G239P	-0.0086	0.0000	0.0000	0.0000	-0.0023	0.0000
G239Q	0.1580	0.0000	0.0000	0.0000	0.0622	0.0000
G239R	-0.0141	0.0000	0.0000	0.0000	0.0000	0.0000
G239S	0.0117	0.0425	0.0000	0.0000	-0.0316	0.0000
G239T	-0.0093	0.0942	0.0000	0.0000	0.0000	0.0000
G239V	0.1902	0.0000	0.0000	0.0000	0.0000	0.0000
S242D	-0.0180	-0.0193	0.0000	0.0000	0.0000	0.0000
S242Q	0.2917	0.2740	0.0000	0.0000	0.1837	0.0000
P243E	-0.0198	-0.0171	-0.0238	-0.0222	0.0000	0.0000

Table III Cont'd. *Relative entropy values for predicted mutations*

Mutation	444 Sequence RE	323 Sequence RE	233 Sequence RE	195 Sequence RE	29 Sequence RE	10 Sequence RE
P243Q	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Q250P	-0.0074	0.0000	0.0000	0.0000	0.0000	0.0000
Q250R	-0.0107	0.0814	-0.0116	-0.0140	0.0000	0.0000
R253Q	0.0299	0.0489	0.0758	0.1080	0.0385	0.0000
A255G	1.5339	1.8306	1.9096	2.0537	1.6764	0.0000
A255C	-0.0042	-0.0044	0.0000	0.0000	0.0000	0.0000
A255T	-0.0082	0.0000	0.0000	0.0000	0.0000	0.0000
I256M	0.0592	0.0560	0.0309	-0.0048	0.0818	0.1558
L257I	0.2044	0.2542	0.2391	0.1717	0.1678	0.0000
N264P	-0.0063	0.0000	-0.0084	-0.0110	0.0310	0.0000
V265D	0.0554	0.1560	0.1197	0.1018	0.0120	0.0000
S267P	-0.0075	0.0000	-0.0100	0.0000	-0.0084	0.0000
S267Q	0.2607	0.0000	0.2886	0.0000	0.4370	0.0000
I269P	-0.0159	-0.0145	-0.0100	0.0000	0.0000	0.0000
D271F	-0.0071	0.0175	0.0186	0.0268	-0.0089	0.0807
D271Y	0.2166	0.0000	-0.0089	0.0000	0.1945	2.5344
I276H	-0.0055	0.0000	0.0000	0.0000	0.0000	0.0000
I276L	1.4146	1.6012	1.6410	1.7350	1.8994	0.3443
I276M	-0.0062	-0.0046	-0.0034	-0.0072	0.0000	0.0000
Y278F	0.1639	0.1170	0.0951	0.0584	0.0301	0.3001
Y278L	-0.0231	-0.0218	-0.0133	-0.0150	-0.0222	0.0000
N280P	-0.0112	-0.0122	-0.0100	-0.0110	0.0000	0.0000
N280R	-0.0163	-0.0163	-0.0155	-0.0140	0.0305	0.0000
Q281P	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Q281R	-0.0157	-0.0142	-0.0139	-0.0163	-0.0054	0.0000
N282Q	-0.0145	-0.0135	-0.0131	-0.0138	0.0000	0.0000
N282R	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
S283P	-0.0155	-0.0148	-0.0159	-0.0113	0.0000	0.0000
G293A	1.9012	1.9987	2.1132	2.1568	2.0977	0.5018
V302Y	-0.0106	0.0111	2.6834	2.8923	0.0008	0.0000
T304P	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
E305A	-0.0162	-0.0199	-0.0164	-0.0124	0.0000	0.0000
E305C	0.0000	0.0024	0.0645	-0.0047	0.0000	0.0000
E305F	-0.0164	0.0000	0.0000	0.0000	0.0000	0.0000
E305G	-0.0083	0.0000	-0.0186	-0.0184	0.0000	0.0000
E305H	0.0000	0.0000	0.0000	0.0000	0.1628	0.0000
E305I	-0.0228	0.0114	0.0000	0.0000	0.0000	0.0000

Table III Cont'd. *Relative entropy values for predicted mutations*

	444	323	233	195	29	10
Mutation	**Sequence**	**Sequence**	**Sequence**	**Sequence**	**Sequence**	**Sequence**
	RE	**RE**	**RE**	**RE**	**RE**	**RE**
E305K	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
E305L	0.2039	0.5964	-0.0136	-0.0151	0.0000	0.0000
E305M	0.3552	0.0350	0.0000	0.0000	0.0000	0.0000
E305N	-0.0136	0.0000	0.3811	0.4054	0.0089	0.0000
E305P	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
E305Q	-0.0142	0.0000	0.0000	0.0000	0.2541	0.0929
E305S	-0.0099	0.0000	0.8509	0.9491	0.0000	0.0000
E305T	-0.0087	0.0000	0.1001	0.1226	0.0000	0.0000
E305V	0.4394	0.4292	-0.0164	-0.0176	0.0137	0.0000
E305Y	0.0000	0.0000	-0.0118	0.0000	0.0000	0.0000
T308P	0.1426	-0.0162	2.7432	2.9526	0.0000	0.0000
S309F	-0.0094	-0.0153	-0.0146	-0.0111	0.0000	0.0000
S309L	-0.0279	-0.0175	-0.0350	-0.0344	0.0000	0.0000
S309W	0.0000	3.2718	-0.0020	0.0000	0.0000	0.0000
G311N	-0.0087	-0.0197	-0.0222	-0.0189	0.0000	0.1224
G311D	0.8417	-0.0197	0.0509	0.1289	0.7281	0.0000
D316A	0.0896	0.0000	0.1546	0.3255	0.0000	0.0000
D316C	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
D316G	-0.0181	0.0000	0.0321	0.0949	0.0000	0.0000
D316P	-0.0078	0.0000	-0.0138	-0.0052	0.0000	0.0000
D316Q	0.0097	0.0000	0.0000	0.0000	0.0000	0.0000
D316S	0.0502	0.0000	0.0842	0.2199	-0.0221	0.0000
T317P	0.0182	0.0000	0.0202	0.0149	0.1576	0.0000
S318E	0.0557	-0.0153	0.0000	0.0000	-0.0214	0.0432
S318F	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
S318L	-0.0305	-0.0332	-0.0347	-0.0342	0.3140	0.0049
S318M	-0.0071	0.0000	-0.0010	0.0077	0.0000	0.0000
S318P	0.0204	0.5668	0.0821	0.1597	0.0000	0.0821
S318Q	0.0735	0.0533	0.0000	0.0000	0.0000	0.0000
S318W	-0.0037	0.0000	0.0000	0.0000	0.0000	0.0000
L319M	0.0000	-0.0047	0.0000	0.0000	0.0000	0.0000
S321K	0.2819	0.1716	0.0084	-0.0014	0.1601	0.2020
S321R	0.0107	0.0000	0.0735	0.0946	-0.0077	0.0814
S322R	0.0570	0.0019	0.0000	0.0000	0.0341	0.0000
S322L	-0.0220	-0.0280	0.0000	0.0000	0.0000	0.0000
L324F	0.1606	0.0000	0.0000	0.0000	0.1996	0.5718
L324H	-0.0073	0.0000	-0.0056	-0.0010	0.0000	0.0000

Table III Cont'd. *Relative entropy values for predicted mutations*

Mutation	444 Sequence RE	323 Sequence RE	233 Sequence RE	195 Sequence RE	29 Sequence RE	10 Sequence RE
L324M	0.0053	0.0189	0.0000	0.0000	0.0000	0.0000
A325P	0.7880	0.0592	-0.0100	0.0000	0.0277	0.0000
G328T	0.0000	0.2174	0.2084	0.2351	0.0000	0.0000

Table IV. *Mutual information values for predicted mutations*

Mutation	444 Sequence MI	323 Sequence MI	233 Sequence MI	195 Sequence MI	29 Sequence MI	10 Sequence MI
V2P	0.9432	-	1.1062	1.0761	-	1.0670
R3P	0.9864	-	1.1263	1.0696	1.0254	0.8979
V7T	0.3710	0.4587	-	0.6530	0.8191	0.0000
N8V	0.4752	0.4265	-	0.4269	0.2848	0.0000
N8A	0.4752	0.4265	-	0.4269	0.2848	0.0000
N8P	0.4752	0.4265	-	0.4269	0.2848	0.0000
I9L	1.0085	0.9571	-	1.0066	1.2085	0.3251
A10P	0.5122	0.6317	0.3803	0.3875	0.7244	0.0000
A10S	0.5122	0.6317	0.3803	0.3875	0.7244	0.0000
D13E	0.6420	0.4641	0.4980	0.5209	0.6509	0.0000
F14P	0.3851	0.4075	0.4929	0.5061	0.4635	0.0000
T18P	1.0041	1.0267	-	-	1.1885	0.6390
V27P	0.9303	0.8067	-	-	-	1.4658
L31I	-	-	1.1868	1.1728	1.1835	-
K32P	-	-	1.1171	1.0718	0.3756	-
N33P	-	-	0.5718	0.4774	0.5525	-
Y40P	-	-	-	-	-	1.2376
V51R	0.9287	0.8547	-	-	1.2485	0.3495
N52R	1.1055	0.9423	-	-	1.5338	0.6109
E53D	0.7504	0.6218	-	-	1.1486	0.3495
E53R	0.7504	0.6218	-	-	1.1486	0.3495
D54A	1.1039	0.8223	1.0118	1.0728	1.2536	0.0000
D54C	1.1039	0.8223	1.0118	1.0728	1.2536	0.0000
D54K	1.1039	0.8223	1.0118	1.0728	1.2536	0.0000
D54L	1.1039	0.8223	1.0118	1.0728	1.2536	0.0000
D54M	1.1039	0.8223	1.0118	1.0728	1.2536	0.0000
D54N	1.1039	0.8223	1.0118	1.0728	1.2536	0.0000
D54R	1.1039	0.8223	1.0118	1.0728	1.2536	0.0000
M56F	0.6783	0.7096	0.7283	0.8038	0.9080	1.2206
T57N	0.2428	0.2291	0.3056	0.3367	0.1739	0.6192
R60V	0.0782	0.0085	0.0062	0.0052	0.0351	0.0000
L61C	0.5632	0.4809	0.6282	0.6247	0.9686	0.0000
G64P	1.0061	0.8340	0.9421	1.0587	1.1398	0.3495
G64A	1.0061	0.8340	0.9421	1.0587	1.1398	0.3495
V69L	0.8565	0.8893	0.9300	0.9165	1.2155	0.7198
V69M	0.8565	0.8893	0.9300	0.9165	1.2155	0.7198
V69N	0.8565	0.8893	0.9300	0.9165	1.2155	0.7198

Table IV Cont'd. *Mutual information values for predicted mutations*

Mutation	444 Sequence MI	323 Sequence MI	233 Sequence MI	195 Sequence MI	29 Sequence MI	10 Sequence MI
N70P	0.8432	0.4485	-	-	0.8603	0.8173
N76P	0.8101	0.8115	0.8554	0.9314	1.1700	1.2459
S79E	0.6087	0.6485	0.7860	0.8790	1.1577	0.8640
S79P	0.6087	0.6485	0.7860	0.8790	1.1577	0.8640
S79Q	0.6087	0.6485	0.7860	0.8790	1.1577	0.8640
T80E	0.7331	0.7029	0.8166	0.9384	1.1547	1.1004
T80Q	0.7331	0.7029	0.8166	0.9384	1.1547	1.1004
I82L	0.4754	0.4843	0.5638	0.5503	0.5969	1.1386
I82M	0.4754	0.4843	0.5638	0.5503	0.5969	1.1386
I82Q	0.4754	0.4843	0.5638	0.5503	0.5969	1.1386
D86P	0.9013	0.8952	1.0127	1.0972	1.5265	0.3495
V89L	0.3552	0.3632	0.4699	0.4972	0.5683	0.3495
V89M	0.3552	0.3632	0.4699	0.4972	0.5683	0.3495
S94P	0.7186	0.7601	0.8499	0.8698	1.2650	1.2206
S94R	0.7186	0.7601	0.8499	0.8698	1.2650	1.2206
A97P	0.8484	-	-	-	1.0780	0.6192
V101I	0.2887	0.3251	0.4283	0.4945	0.6814	0.5545
V101L	0.2887	0.3251	0.4283	0.4945	0.6814	0.5545
D102P	0.1990	0.2101	0.2030	0.1962	0.5002	0.0000
H104P	0.0375	0.0420	0.0323	0.0052	0.0351	0.0000
V107N	0.6919	0.6276	0.7105	0.7190	0.6837	0.0000
G112A	0.9863	0.9914	1.2758	1.3827	1.4023	1.3366
G112E	0.9863	0.9914	1.2758	1.3827	1.4023	1.3366
G112C	0.9863	0.9914	1.2758	1.3827	1.4023	1.3366
G112D	0.9863	0.9914	1.2758	1.3827	1.4023	1.3366
G112H	0.9863	0.9914	1.2758	1.3827	1.4023	1.3366
G112I	0.9863	0.9914	1.2758	1.3827	1.4023	1.3366
G112K	0.9863	0.9914	1.2758	1.3827	1.4023	1.3366
G112L	0.9863	0.9914	1.2758	1.3827	1.4023	1.3366
G112F	0.9863	0.9914	1.2758	1.3827	1.4023	1.3366
G112M	0.9863	0.9914	1.2758	1.3827	1.4023	1.3366
G112N	0.9863	0.9914	1.2758	1.3827	1.4023	1.3366
G112Q	0.9863	0.9914	1.2758	1.3827	1.4023	1.3366
G112R	0.9863	0.9914	1.2758	1.3827	1.4023	1.3366
G112S	0.9863	0.9914	1.2758	1.3827	1.4023	1.3366
G112T	0.9863	0.9914	1.2758	1.3827	1.4023	1.3366
G112V	0.9863	0.9914	1.2758	1.3827	1.4023	1.3366

Table IV Cont'd. *Mutual information values for predicted mutations*

Mutation	444 Sequence MI	323 Sequence MI	233 Sequence MI	195 Sequence MI	29 Sequence MI	10 Sequence MI
G112W	0.9863	0.9914	1.2758	1.3827	1.4023	1.3366
G112Y	0.9863	0.9914	1.2758	1.3827	1.4023	1.3366
I114P	0.3480	0.3268	0.2037	0.2223	0.6416	0.4192
Q116N	0.9036	-	0.9343	0.9778	-	0.0000
Q116D	0.9036	-	0.9343	0.9778	-	0.0000
Q116W	0.9036	-	0.9343	0.9778	-	0.0000
T120D	-	0.7999	-	-	1.1555	0.5004
T120S	-	0.7999	-	-	1.1555	0.5004
N121E	0.9967	0.8702	-	-	1.4469	0.6846
A122E	0.5053	0.4719	0.6345	0.7112	0.9308	0.8415
A122Q	0.5053	0.4719	0.6345	0.7112	0.9308	0.8415
T125P	0.6930	0.7337	0.9200	1.0229	1.2906	1.2040
S126P	0.9499	1.0025	0.9599	0.9005	1.0909	0.9404
S129P	0.8529	0.8456	0.9488	0.9321	1.2417	1.1935
S133R	0.8766	0.8966	1.2206	1.2778	1.4123	0.9503
S134K	0.9013	0.8952	0.9225	1.0857	1.3495	0.8513
Y135F	0.4885	0.4529	0.4784	0.4629	0.5548	0.0000
A136P	0.5164	0.5509	0.7410	0.8253	1.2293	0.6931
S139P	0.6982	0.7023	0.9600	1.0230	1.3744	0.6390
W142I	0.4837	0.5182	0.5925	0.6178	0.8881	1.2799
W142V	0.4837	0.5182	0.5925	0.6178	0.8881	1.2799
W142F	0.4837	0.5182	0.5925	0.6178	0.8881	1.2799
W142M	0.4837	0.5182	0.5925	0.6178	0.8881	1.2799
W142H	0.4837	0.5182	0.5925	0.6178	0.8881	1.2799
W142Y	0.4837	0.5182	0.5925	0.6178	0.8881	1.2799
W142L	0.4837	0.5182	0.5925	0.6178	0.8881	1.2799
W142T	0.4837	0.5182	0.5925	0.6178	0.8881	1.2799
W142E	0.4837	0.5182	0.5925	0.6178	0.8881	1.2799
F143M	0.1318	0.1195	0.1230	0.1413	0.1892	0.0000
G144A	0.7373	0.7174	0.7452	0.7631	0.8142	0.0000
G144P	0.7373	0.7174	0.7452	0.7631	0.8142	0.0000
G144D	0.7373	0.7174	0.7452	0.7631	0.8142	0.0000
G144N	0.7373	0.7174	0.7452	0.7631	0.8142	0.0000
I145V	0.7356	0.6748	0.6814	0.7630	1.1486	0.6931
N147P	0.0598	0.0095	0.0060	0.0052	0.0351	0.0000
N153D	0.7128	0.6918	0.8783	0.9210	1.2438	1.1412
I154M	-	0.9805	1.2037	1.2526	1.5201	0.0000

Table IV Cont'd. *Mutual information values for predicted mutations*

Mutation	444 Sequence MI	323 Sequence MI	233 Sequence MI	195 Sequence MI	29 Sequence MI	10 Sequence MI
N155E	0.6907	0.6288	0.9175	0.9638	1.2675	0.9503
N155Q	0.6907	0.6288	0.9175	0.9638	1.2675	0.9503
T156E	0.9803	0.9340	1.0606	1.1664	1.5455	1.0889
T156G	0.9803	0.9340	1.0606	1.1664	1.5455	1.0889
V161I	1.1055	0.9985	1.0228	1.0875	1.1935	0.0000
V161L	1.1055	0.9985	1.0228	1.0875	1.1935	0.0000
E163P	0.5793	0.5134	0.6164	0.6860	1.0531	1.2799
V164A	0.5027	0.4550	0.5458	0.6331	0.9309	0.8171
V165I	0.5374	0.3428	0.3775	0.4158	0.6995	0.0000
I168H	0.2329	0.1664	0.1533	0.1431	0.3406	0.0000
R169P	0.1486	0.0212	0.0044	0.0052	0.0351	0.0000
N170P	0.6410	0.5573	0.6485	0.7015	1.1374	0.5004
N170R	0.6410	0.5573	0.6485	0.7015	1.1374	0.5004
Q176P	0.5892	-	-	-	0.9178	0.0000
Q186D	0.4775	0.3354	0.4310	0.4794	0.7640	0.7198
Q186G	0.4775	0.3354	0.4310	0.4794	0.7640	0.7198
Q186E	0.4775	0.3354	0.4310	0.4794	0.7640	0.7198
Q186N	0.4775	0.3354	0.4310	0.4794	0.7640	0.7198
Q186T	0.4775	0.3354	0.4310	0.4794	0.7640	0.7198
S187P	0.6879	0.5978	0.5571	0.6216	0.6651	0.0000
A188C	0.4478	0.3998	0.3175	0.1526	0.5446	0.0000
G189A	1.0804	1.0200	1.0302	1.1271	1.2849	0.4192
G189E	1.0804	1.0200	1.0302	1.1271	1.2849	0.4192
G189H	1.0804	1.0200	1.0302	1.1271	1.2849	0.4192
G189K	1.0804	1.0200	1.0302	1.1271	1.2849	0.4192
G189N	1.0804	1.0200	1.0302	1.1271	1.2849	0.4192
G189Q	1.0804	1.0200	1.0302	1.1271	1.2849	0.4192
G189R	1.0804	1.0200	1.0302	1.1271	1.2849	0.4192
G189S	1.0804	1.0200	1.0302	1.1271	1.2849	0.4192
F191W	0.6151	0.5497	0.5463	0.5848	0.5890	0.3495
S193P	-	1.1856	0.8660	0.8789	1.3229	1.0639
A197F	0.9969	0.9233	1.6886	1.8145	-	0.6846
A197M	0.9969	0.9233	1.6886	1.8145	-	0.6846
A199V	1.0818	0.9025	0.9983	-	1.2545	0.0000
S201K	-	0.8285	1.1982	1.2958	1.1476	1.4114
S201P	-	0.8285	1.1982	1.2958	1.1476	1.4114
S201Q	-	0.8285	1.1982	1.2958	1.1476	1.4114

Table IV Cont'd. *Mutual information values for predicted mutation*

Mutation	444 Sequence MI	323 Sequence MI	233 Sequence MI	195 Sequence MI	29 Sequence MI	10 Sequence MI
N205D	-	-	1.4381	1.4223	0.4416	0.0000
N205P	-	-	1.4381	1.4223	0.4416	0.0000
V217I	0.9378	0.8482	0.8543	0.8961	0.9109	0.3251
V217L	0.9378	0.8482	0.8543	0.8961	0.9109	0.3251
K219S	0.6963	0.5775	0.6251	0.5793	0.8350	0.0000
K219A	0.6963	0.5775	0.6251	0.5793	0.8350	0.0000
K219Q	0.6963	0.5775	0.6251	0.5793	0.8350	0.0000
K219E	0.6963	0.5775	0.6251	0.5793	0.8350	0.0000
L221N	0.3243	0.2421	0.2390	0.2167	0.5985	0.0000
D222P	0.2578	0.0424	0.0555	0.0385	0.2643	0.0000
S223P	0.8114	0.8122	1.1191	1.1864	1.2064	0.3495
A230P	0.7936	0.8483	0.9581	0.9795	1.4012	1.3719
E231P	0.7215	0.7260	0.9146	1.0206	1.0051	0.4192
N236G	0.8610	0.8619	0.9315	0.9384	1.0379	0.0000
I237W	0.7122	-	-	-	1.1041	0.9503
I237Y	0.7122	-	-	-	1.1041	0.9503
I237F	0.7122	-	-	-	1.1041	0.9503
D238E	0.9238	0.7019	0.8208	0.8425	1.1466	0.6852
D238Q	0.9238	0.7019	0.8208	0.8425	1.1466	0.6852
G239A	0.9604	-	-	-	1.0965	1.6096
G239C	0.9604	-	-	-	1.0965	1.6096
G239D	0.9604	-	-	-	1.0965	1.6096
G239E	0.9604	-	-	-	1.0965	1.6096
G239I	0.9604	-	-	-	1.0965	1.6096
G239K	0.9604	-	-	-	1.0965	1.6096
G239L	0.9604	-	-	-	1.0965	1.6096
G239M	0.9604	-	-	-	1.0965	1.6096
G239N	0.9604	-	-	-	1.0965	1.6096
G239P	0.9604	-	-	-	1.0965	1.6096
G239Q	0.9604	-	-	-	1.0965	1.6096
G239R	0.9604	-	-	-	1.0965	1.6096
G239S	0.9604	-	-	-	1.0965	1.6096
G239T	0.9604	-	-	-	1.0965	1.6096
G239V	0.9604	-	-	-	1.0965	1.6096
S242D	0.6473	0.7033	-	-	1.4154	1.3124
S242Q	0.6473	0.7033	-	-	1.4154	1.3124
P243E	0.8971	0.8315	1.0284	1.1729	1.1523	0.8018

Table IV Cont'd. *Mutual information values for predicted mutations*

Mutation	444 Sequence MI	323 Sequence MI	233 Sequence MI	195 Sequence MI	29 Sequence MI	10 Sequence MI
P243Q	0.8971	0.8315	1.0284	1.1729	1.1523	0.8018
Q250P	0.8561	-	0.9478	1.0889	1.1228	1.4708
Q250R	0.8561	-	0.9478	1.0889	1.1228	1.4708
R253Q	0.9792	0.8779	0.9839	1.0704	1.2313	0.0000
A255G	0.5350	0.4583	0.4377	0.4305	0.8512	0.0000
A255C	0.5350	0.4583	0.4377	0.4305	0.8512	0.0000
A255T	0.5350	0.4583	0.4377	0.4305	0.8512	0.0000
I256M	0.6102	0.6139	0.6397	0.6787	1.0480	1.2206
L257I	0.4072	0.3248	0.3463	0.4110	0.5168	0.7425
N264P	0.8018	0.6162	0.9334	1.0101	1.1070	0.0000
V265D	0.7214	0.6624	0.7241	0.7959	1.0951	0.6109
S267P	0.9756	0.0686	1.1042	0.1066	1.5201	0.0000
S267Q	0.9756	0.0686	1.1042	0.1066	1.5201	0.0000
I269P	1.1679	0.7694	0.8384	0.9623	1.3144	1.4768
D271F	0.8647	0.5501	0.7026	0.7835	1.1545	0.6192
D271Y	0.8647	0.5501	0.7026	0.7835	1.1545	0.6192
I276H	0.6591	0.5718	0.5811	0.5480	0.4110	0.9648
I276L	0.6591	0.5718	0.5811	0.5480	0.4110	0.9648
I276M	0.6591	0.5718	0.5811	0.5480	0.4110	0.9648
Y278F	0.6392	0.6284	0.7155	0.7507	0.9017	0.8387
Y278L	0.6392	0.6284	0.7155	0.7507	0.9017	0.8387
N280P	0.9812	0.9213	0.9669	1.0615	1.3480	0.6192
N280R	0.9812	0.9213	0.9669	1.0615	1.3480	0.6192
Q281P	0.8161	0.6948	1.0196	1.1976	1.3329	0.8047
Q281R	0.8161	0.6948	1.0196	1.1976	1.3329	0.8047
N282Q	0.4504	0.4644	0.4356	0.4977	0.6509	0.4192
N282R	0.4504	0.4644	0.4356	0.4977	0.6509	0.4192
S283P	0.6073	0.7380	0.9520	0.9993	1.2797	0.6109
G293A	0.3148	0.2401	0.2403	0.2348	0.5752	1.1629
V302Y	-	-	0.4893	0.3524	1.1357	0.8570
T304P	-	-	0.9830	0.9876	0.8245	1.3366
E305A	-	-	0.6291	0.5959	0.9807	0.4192
E305C	-	-	0.6291	0.5959	0.9807	0.4192
E305F	-	-	0.6291	0.5959	0.9807	0.4192
E305G	-	-	0.6291	0.5959	0.9807	0.4192
E305H	-	-	0.6291	0.5959	0.9807	0.4192
E305I	-	-	0.6291	0.5959	0.9807	0.4192

Table IV Cont'd. *Mutual information values for predicted mutations*

Mutation	444 Sequence MI	323 Sequence MI	233 Sequence MI	195 Sequence MI	29 Sequence MI	10 Sequence MI
E305K	-	-	0.6291	0.5959	0.9807	0.4192
E305L	-	-	0.6291	0.5959	0.9807	0.4192
E305M	-	-	0.6291	0.5959	0.9807	0.4192
E305N	-	-	0.6291	0.5959	0.9807	0.4192
E305P	-	-	0.6291	0.5959	0.9807	0.4192
E305Q	-	-	0.6291	0.5959	0.9807	0.4192
E305S	-	-	0.6291	0.5959	0.9807	0.4192
E305T	-	-	0.6291	0.5959	0.9807	0.4192
E305V	-	-	0.6291	0.5959	0.9807	0.4192
E305Y	-	-	0.6291	0.5959	0.9807	0.4192
T308P	-	-	0.3887	0.1950	1.1317	0.8018
S309F	-	0.7413	-	-	0.8871	1.4272
S309L	-	0.7413	-	-	0.8871	1.4272
S309W	-	0.7413	-	-	0.8871	1.4272
G311N	-	0.7861	-	-	1.1843	1.2984
G311D	-	0.7861	-	-	1.1843	1.2984
D316A	-	-	-	-	-	0.4192
D316C	-	-	-	-	-	0.4192
D316G	-	-	-	-	-	0.4192
D316P	-	-	-	-	-	0.4192
D316Q	-	-	-	-	-	0.4192
D316S	-	-	-	-	-	0.4192
T317P	-	-	-	-	1.2732	0.5004
S318E	-	-	-	-	1.3111	1.3662
S318F	-	-	-	-	1.3111	1.3662
S318L	-	-	-	-	1.3111	1.3662
S318M	-	-	-	-	1.3111	1.3662
S318P	-	-	-	-	1.3111	1.3662
S318Q	-	-	-	-	1.3111	1.3662
S318W	-	-	-	-	1.3111	1.3662
L319M	-	-	-	-	0.4853	0.6390
S321K	-	-	-	-	1.2908	1.0820
S321R	-	-	-	-	1.2908	1.0820
S322R	-	-	-	-	1.0711	0.9410
S322L	-	-	-	-	1.0711	0.9410
L324F	-	-	-	-	-	0.9410
L324H	-	-	-	-	-	0.9410

Table IV Cont'd. *Mutual information values for predicted mutations*

Mutation	444 Sequence MI	323 Sequence MI	233 Sequence MI	195 Sequence MI	29 Sequence MI	10 Sequence MI
L324M	-	-	-	-	-	0.9410
A325P	-	-	-	-	-	1.0639
G328T	-	-	-	-	-	-

Table V. *ΔΔG and T_{50} values for predicted mutations*

Mutation	FoldX (kcal mol^{-1})	Rosetta (kcal mol^{-1})	ΔT_{50} (°C)
V2P	2.65	1.54	-
R3P	-2.64	2.44	-
V7T	2.00	-0.15	-
N8V	1.39	-2.64	-
N8A	3.30	-2.53	-
N8P	7.75	3.43	-
I9L	0.39	1.39	-
A10P	3.44	3.39	-
A10S	-0.16	0.12	-
D13E	0.30	2.85	3.0±0.5
F14P	7.72	6.23	-
T18P	-1.67	0.46	0.2±0.1
V27P	2.72	2.29	-
L31I	0.02	-0.60	-
K32P	-2.50	3.74	-
N33P	-0.71	19.43	-
Y40P	3.42	6.63	-
V51R	-0.73	0.40	-
N52R	-0.27	-0.04	-
E53D	0.06	-0.77	2.7±0.7
E53R	-0.70	-0.41	-
D54A	-2.76	0.51	-
D54C	-3.14	1.31	-
D54K	-2.54	3.37	-
D54L	-5.29	0.92	-
D54M	-3.84	1.89	-
D54N	-3.89	-0.78	-
D54R	-3.03	4.25	-
M56F	-0.67	-2.30	-
T57N	-0.30	0.43	1.1±0.0
R60V	3.41	-1.75	-
L61C	2.11	2.75	-
G64P	4.06	1.03	-
G64A	2.97	-0.30	-0.1±0.2
V69L	-0.48	-0.16	-
V69M	0.34	0.93	-
V69N	1.34	1.26	-
N70P	6.86	212.76	-

Table V Cont'd. *ΔΔG and T_{50} values for predicted mutations*

Mutation	FoldX (kcal mol^{-1})	Rosetta (kcal mol^{-1})	ΔT$_{50}$ (°C)
N76P	0.00	0.95	0.8±0.5
S79E	-1.07	-0.19	-0.1±0.2
S79P	-1.84	0.97	0.3±0.5
S79Q	-0.66	-0.31	0.0±0.3
T80E	-0.44	-0.06	0.5±0.2
T80Q	-0.30	0.12	-0.1±0.2
I82L	1.42	-0.12	-0.2±0.5
I82M	-1.01	0.44	0.3±0.5
I82Q	1.97	-0.58	-
D86P	7.94	5.72	-
V89L	-0.71	1.67	-
V89M	0.24	1.57	-
S94P	5.50	7.04	-
S94R	0.84	0.25	-
A97P	0.94	3.19	-
V101I	-0.83	0.11	0.5±0.4
V101L	-0.62	1.62	-0.5±0.3
D102P	6.90	32.81	-
H104P	6.50	6.66	-
V107N	2.10	4.53	-
G112A	-0.06	-0.95	-
G112E	0.17	-1.17	-
G112C	0.28	0.57	-
G112D	0.42	-0.96	-
G112H	0.58	-1.69	-
G112I	0.56	2.20	-
G112K	-0.33	-1.50	-
G112L	-0.45	-2.13	-
G112F	-0.23	-2.04	-
G112M	-0.41	-1.73	-
G112N	0.08	-1.28	-
G112Q	-0.07	-1.63	-
G112R	0.22	-2.03	-
G112S	0.56	-0.57	-
G112T	0.56	0.19	-
G112V	0.92	1.28	-
G112W	-0.11	-1.76	-
G112Y	-0.20	-2.05	-

Table V Cont'd. *ΔΔG and T_{50} values for predicted mutations*

Mutation	FoldX (kcal mol^{-1})	Rosetta (kcal mol^{-1})	ΔT$_{50}$ (°C)
I114P	7.75	1.78	-
Q116N	1.42	-2.92	-
Q116D	1.54	-2.33	-
Q116W	1.11	-1.93	-
T120D	-2.16	0.36	-
T120S	-0.81	-0.10	-
N121E	-0.19	0.82	-
A122E	-0.70	0.00	-0.2±0.5
A122Q	0.08	-0.10	-
T125P	4.23	5.31	-
S126P	4.39	6.31	-
S129P	4.56	6.77	-
S133R	-0.95	-0.65	0.4±0.2
S134K	0.00	0.83	-
Y135F	0.77	0.25	-
A136P	2.04	6.38	-
S139P	-1.33	2.14	1.8±0.6
W142I	3.55	-4.90	-
W142V	4.72	-4.29	-
W142F	4.64	-3.76	-
W142M	3.64	-3.65	-
W142H	5.31	-3.38	-
W142Y	5.17	-2.91	-
W142L	3.66	-2.62	-
W142T	6.86	-2.60	-
W142E	6.80	-2.06	-
F143M	1.48	3.10	-
G144A	1.43	-0.53	-
G144P	8.48	9.19	-
G144D	5.68	-1.44	-
G144N	0.21	0.40	-
I145V	0.89	0.02	-
N147P	9.31	10.37	-
N153D	-2.91	-0.25	0.5±0.9
I154M	-0.57	0.80	-
N155E	-1.45	0.06	0.5±0.3
N155Q	-0.42	-0.02	0.1±0.1
T156E	-0.62	-0.18	0.2±0.3

Table V Cont'd. *ΔΔG and T_{50} values for predicted mutations*

Mutation	FoldX (kcal mol^{-1})	Rosetta (kcal mol^{-1})	ΔT$_{50}$ (°C)
T156G	0.61	-1.77	-
V161I	0.57	0.53	-
V161L	-0.33	1.61	-
E163P	4.31	5.79	-
V164A	2.26	0.59	-
V165I	0.94	2.03	-
I168H	13.83	3.94	-
R169P	7.92	6.32	-
N170P	2.36	4.67	-
N170R	-0.32	0.13	-
Q176P	-0.53	2.06	-
Q186D	0.81	-3.13	-
Q186G	-2.95	-1.40	-
Q186E	1.97	-2.99	-
Q186N	-0.33	-2.05	-
Q186T	2.38	1.23	-
S187P	6.16	5.17	-
A188C	0.33	4.31	-
G189A	-0.93	-0.76	0.4±0.4
G189E	-0.58	-1.01	0.0±0.2
G189H	-0.08	-0.26	-
G189K	-0.89	-0.55	-0.1±0.2
G189N	-0.33	-0.20	-
G189Q	-0.65	-0.58	-
G189R	-0.58	-0.37	-
G189S	-0.45	-0.97	1.2±0.4
F191W	2.84	0.47	-
S193P	-1.26	0.32	-
A197F	3.66	1.86	-
A197M	1.22	4.09	-
A199V	0.27	1.28	-
S201K	-0.22	-0.42	-
S201P	3.46	5.42	-
S201Q	0.08	0.12	-
N205D	1.94	-1.51	-
N205P	3.19	3.48	-
V217I	-0.22	2.27	-
V217L	-0.47	2.80	-

Table V Cont'd. *ΔΔG and T_{50} values for predicted mutations*

Mutation	FoldX (kcal mol^{-1})	Rosetta (kcal mol^{-1})	ΔT_{50} (°C)
K219S	1.98	-3.65	-
K219A	2.07	-3.20	2.0±0.7
K219Q	1.28	-1.92	2.8±0.1
K219E	1.77	-1.82	-
L221N	1.86	-1.41	-
D222P	-1.11	7.87	-
S223P	-1.26	1.04	-
A230P	-1.80	2.84	-
E231P	0.33	0.82	-
N236G	0.77	-1.89	-
I237W	2.73	-2.17	-
I237Y	2.10	-2.06	-
I237F	1.67	-1.99	-
D238E	-0.12	0.24	-
D238Q	-0.06	0.12	-
G239A	-0.14	-0.56	-
G239C	-0.53	1.35	-
G239D	-0.08	-1.46	0.4±0.2
G239E	-0.55	-1.20	0.2±0.3
G239I	-0.78	0.57	-
G239K	-0.78	-0.89	-
G239L	-1.23	-0.79	-
G239M	-0.94	-0.39	-
G239N	-0.13	-1.43	0.7±0.1
G239P	-0.56	0.62	-
G239Q	-0.44	-0.94	-0.9±0.2
G239R	-0.73	-0.55	-
G239S	-0.20	-1.04	-0.7±0.1
G239T	-0.09	-0.84	-
G239V	-0.31	0.10	-
S242D	-1.00	-0.09	-
S242Q	-0.86	-0.03	-
P243E	1.86	0.45	-
P243Q	1.62	0.22	-
Q250P	2.36	4.98	-
Q250R	-0.41	0.06	-
R253Q	0.36	-2.11	-
A255G	2.00	1.92	-

Table V Cont'd. *ΔΔG and T_{50} values for predicted mutations*

Mutation	FoldX (kcal mol^{-1})	Rosetta (kcal mol^{-1})	ΔT$_{50}$ (°C)
A255C	0.62	2.99	-
A255T	1.71	1.40	-
I256M	0.48	0.36	-
L257I	-0.25	0.32	-
N264P	1.32	3.92	-
V265D	-1.87	2.41	-
S267P	-2.75	1.46	-
S267Q	-0.67	-0.13	-
I269P	2.95	7.40	-
D271F	-2.23	-1.08	3.1±1.1
D271Y	-1.83	-1.07	2.7±0.4
I276H	5.24	-1.75	-
I276L	0.63	0.55	-
I276M	-0.31	0.29	-
Y278F	0.19	-0.41	1.0±0.5
Y278L	1.60	2.77	-
N280P	3.56	6.67	-
N280R	-0.46	-0.93	-
Q281P	2.84	8.74	-
Q281R	0.28	0.09	-
N282Q	1.00	-1.97	-
N282R	0.75	-1.93	-
S283P	-1.77	1.32	-
G293A	6.66	-0.08	3.5±0.2
V302Y	0.81	1.03	-
T304P	0.61	2.88	-
E305A	-2.46	-0.78	-
E305C	-2.86	0.14	-
E305F	-3.77	-2.24	-
E305G	-2.23	0.42	-
E305H	-2.91	-3.07	-
E305I	-2.62	-0.89	-
E305K	-4.04	0.47	-
E305L	-4.82	-1.95	-
E305M	-4.09	-1.03	-
E305N	-1.88	-0.97	-
E305P	1.45	54.88	-
E305Q	-2.12	0.66	-

Table V Cont'd. *ΔΔG and T_{50} values for predicted mutations*

Mutation	FoldX (kcal mol^{-1})	Rosetta (kcal mol^{-1})	ΔT_{50} (°C)
E305S	-1.77	-0.79	-
E305T	-1.61	-2.14	-
E305V	-2.18	-0.47	-
E305Y	-1.60	-2.13	-
T308P	1.45	29.64	-
S309F	-0.57	-1.94	2.7±0.1
S309L	-1.78	-2.26	1.5±0.3
S309W	0.19	-2.24	0.4±0.1
G311N	0.82	-1.23	-
G311D	0.78	-0.58	-
D316A	-1.91	-0.45	-
D316C	-2.44	1.43	-
D316G	-0.98	0.09	-
D316P	-2.36	1.07	-
D316Q	-1.81	-0.37	-
D316S	-2.20	-0.39	-
T317P	-1.41	0.50	-
S318E	-1.20	-0.21	0.9±0.2
S318F	-1.94	0.94	-
S318L	-1.54	-0.41	-
S318M	-2.09	-0.01	-
S318P	-2.31	3.27	3.2±0.9
S318Q	-0.75	-0.38	0.5±0.2
S318W	-1.81	1.07	-
L319M	0.55	1.01	-
S321K	0.36	-0.97	-
S321R	0.50	-0.56	-
S322R	-0.98	-0.26	-
S322L	-1.87	1.15	-
L324F	-1.58	-2.35	-
L324H	0.77	-1.86	-
L324M	-0.19	0.26	-
A325P	3.97	4.67	-
G328T	N/A	0.03	-

APPENDIX B

Crystallographic Analysis of Designed Kemp Eliminases

This appendix contains relevant excerpts from [1]. The author contributed three crystal structures (HG-2 holo, 1A53-2 apo, and 1A53-2 holo), Figure 2B-F, Figure 6, and data for the HG-2 holo, 1A53-2 apo, 1A43-2 holo structures in Figure S3. Please refer to the original publication for further information.

B.1 Abstract

A general approach for the computational design of enzymes to catalyze arbitrary reactions is a goal at the forefront of the field of protein design. Recently, computationally designed enzymes have been produced for three chemical reactions through the synthesis and screening of a large number of variants. Here, we present an iterative approach that has led to the development of the most catalytically efficient computationally designed enzyme for the Kemp elimination to date. Previously established computational techniques were used to generate an initial design, HG-1, which was catalytically inactive. Analysis of HG-1 with molecular dynamics simulations (MD) and X-ray crystallography indicated that the inactivity might be due to bound waters and high flexibility of residues within the active site. This analysis guided changes to our design procedure, moved the design deeper into the interior of the protein, and resulted in an active Kemp eliminase, HG-2. The cocrystal structure of this enzyme with a transition state analog (TSA) revealed that the TSA was bound in the active site, interacted with the intended catalytic base in a catalytically relevant manner, but was flipped relative to the design model. MD analysis of HG-2 led to an additional point mutation, HG-3, that produced a further threefold improvement in activity. This iterative approach to computational enzyme design, including detailed MD and structural analysis of both active and inactive designs, promises a more complete understanding of the underlying principles of enzymatic catalysis and furthers progress toward reliably producing active enzymes.

B.2 Introduction

The high efficiency, chemoselectivity, regio- and stereospecificity, and biodegradability of enzymes make them extremely attractive catalysts. However, the finite repertoire of naturally occurring enzymes limits their applicability to broad problems in biotechnology. A general method for the computational design of enzymes that can efficiently catalyze arbitrary chemical reactions would allow the benefits of enzymatic catalysis to be applied to chemical transformations of interest that are currently inaccessible via natural enzymes. Bolon and Mayo provided important early evidence that such an approach is feasible [2], which motivated significant progress toward this goal in recent years. Using quantum mechanics-based active site design and the Rosetta software suite, Baker, Houk, and coworkers designed enzymes for three chemically unrelated nonnatural reactions in a variety of catalytically inert scaffolds [3-5]. In early incarnations of computational protein design, a strategy for methods development was put forth in terms of the so-called "protein design cycle" in which experimental evaluation of an initial design is used to inform adjustments to the design process for subsequent rounds of design [6, 7]. Ideally, these steps would be continued iteratively until the protein sequences predicted by the algorithm exhibit the desired characteristics. However, there is little evidence that this strategy has been used for purposes other than force-field parameterization [6, 8-10]. Proteins from failed computational design efforts are typically discarded without comment or investigation into the cause of failure. This situation is unfortunate, because valuable information is lost when only successful designs are reported. Without detailed computational and/or experimental analysis of failed designs, flaws in the design procedure cannot be identified and remedied to produce proteins with the desired characteristics [11, 12]. In addition, a focus on reporting only successful designs can lead to the impression that current computational protein design methods are errorless.

The recent successes in designing enzymes show that the field is well on the way to its goal of developing a general method for designing protein catalysts [3-5]. However, the catalytic rate enhancements of computationally designed enzymes are still well below those of natural enzymes, and the methods are dominated by false positives. In the case of the Kemp eliminase enzymes designed by Röthlisberger et al., 59 of the many individual sequences predicted to be active by their protein design methods were selected for experimental screening, and only eight of these turned out to be active. Although active enzymes were in fact produced, the need for a "shotgun" approach suggests an incomplete understanding of the details of the enzymatic system and/or inaccurate modeling by the protein design algorithm [13].

In this work, we focus on the development of a single designed enzyme to test our understanding of enzymatic catalysis and the applicability of the protein design cycle to computational enzyme design problems. We targeted our efforts on the Kemp elimination (KE) (Fig. 1), a well-studied model system for the deprotonation of carbon [14]. The KE was selected as a model reaction for this study because catalysts for it have been reported in multiple protein scaffolds [3, 15-17]. In addition, from a computational design perspective, the use of the KE allows a direct comparison to the eight enzymes that were computationally designed for this reaction by Röthlisberger et al. [3].

Our approach to KE enzyme design consisted of three steps, which are described in detail by Lassila et al. [18]. First, we designed an idealized active site for the KE that included an ab initio calculated transition state (TS) and contacting catalytic residues oriented to facilitate binding and catalysis (Fig. 2A). Next, targeted ligand placement was used to simultaneously sample TS poses and catalytic amino acid positions and orientations within a poly-alanine–substituted binding pocket of a protein scaffold that does not naturally catalyze the KE. Active site configurations that fulfill all of the required catalytic contacts were identified. Finally, one of these active site configurations was selected, and the remaining binding pocket residues were designed to support the TS pose and the geometry of the catalytic residues. Our initial design, HG-1, showed no measurable KE activity. To identify deficiencies in the design procedure, we investigated possible causes of inactivity by using X-ray crystallography and molecular dynamics (MD) simulations. Two problems were identified: The active site was overly exposed to solvent, and critical active site residues showed a high degree of flexibility and orientations inconsistent with the design objectives. Iterating on the protein design process, we corrected these problems in subsequent rounds of computational design using the same protein scaffold. The design with the highest activity, HG-3, was found to have a k_{cat}/K_m of 430 M^{-1} s^{-1}.

B.3 Results

The following section only contains text relevant to data obtained by the author of this thesis. For a full description of results, please consult [1].

Second-Generation Design

A key observation from the crystal structure and the MD simulations is that a significant number of water molecules are present in the active site of the first-generation HG-1 design. This finding suggests a substantial desolvation barrier for substrate binding and a bulk, solvent-like pK_a of the base (E237). The high flexibility of the active site side chains and low degree of preorganization may further add to the observed inactivity. On the basis of early work by Kemp and coworkers, who showed that a nonpolar environment is best suited for the base-catalyzed KE [19, 20], increasing the hydrophobic character of the HG-1 active site is expected to facilitate the binding of the hydrophobic 5-nitrobenzisoxazole substrate and also elevate the pK_a of the base. We therefore sought a more embedded active site pocket in order to maximize these effects. Manual inspection identified native D127 as a promising candidate for the catalytic base. This aspartate forms a salt bridge with R81 and defines the bottom of a well-packed, narrow solvent-accessible pocket in the core of the $(\alpha/\beta)_8$ barrel, well removed from the native TAX binding pocket. Using a computational approach, we sought to increase the size of this pocket to accommodate the substrate and the additional catalytic residues. This area also contains polar and charged residues, which do not provide the ideal environment for the KE. Substantial modifications of R81, N130, N172, T236, and E237 would be necessary to allow the substrate access to the base and to form a hydrophobic binding pocket to facilitate proton abstraction.

By focusing the design on the native D127 as the general base, an active site search was carried out in a manner similar to that for HG-1 using identical geometric constraints. Compared to the HG-1 calculation, the active site search for this design was shifted 7 Å further into the barrel of the scaffold (Figure 4A, see [1]).

The final catalytic configuration consisted of D127 as the general base, T44W as the π-stacking residue, and T265S as the hydrogen bond donor (Figure 2C). The isoxazole ring of the TS points into the back of the active site pocket and is well shielded from solvent. Active site repacking produced the second-generation design HG-2, whose sequence differs by 12 mutations from wild-type TAX (SI Appendix, Table S2, see [1]) and 19 mutations from HG-1. As expected, the design model shows major changes in the size and hydrophobicity of the active site residues relative to wildtype TAX. Fig. 4 demonstrates the variation of the active sites basis on this scaffold. Of note, R81, which forms a buried salt bridge with D127 in TAX, was mutated to a glycine in the design, making room for the substrate to access the base. Nearby H83 and N130 were also mutated to glycine to further open up space in the active site for the substrate and the catalytic residues. Q42, T84, N172, T236, and E237 were mutated to large hydrophobic residues, which increases the overall hydrophobicity of the active site and promotes packing around the TS and catalytic residues.

Characterization of Second-Generation Design

A 1.2-Å resolution X-ray crystal structure of HG-2 with the transition state analog (TSA) 5-nitrobenzotriazole (5-NBT) bound in the active site provides direct evidence of catalytically competent substrate interaction with the putative base (Figures 2D–F). The protein crystallized with two molecules in the asymmetric unit, which allows for observation of two active sites. Ligand density in chain A was modeled in two orientations (Figures 2D and E). The dual orientations may reflect the conformational flexibility of the engineered active site, some of which was observed in the MD simulations. Unambiguous density for a single TSA orientation appears in chain B (Figure 2F). This orientation (O2) differs from that of the design (O1) in that the TSA is flipped from the designed position, which places the nitro group in contact with S265 rather than K50. In both O1 and O2, the TSA contacts the putative base (D127) in a catalytically relevant manner.

Recapitulation of Previous KE Designs

We also tested the ability of our computational design methods to recapitulate the active sites of three functional enzymes from Röthlisberger et al. [3]. KE59 was based on the *Sulfolobus solfataricus* indole-3-glycerolphosphate synthase scaffold [21]; KE07 and

KE10 were based on the Thermotoga maritima imidazoleglycerolphosphate synthase scaffold [22].

Starting with the base positions and scaffolds from the active KE07, KE10, and KE59 enzymes, TS poses and catalytic residue positions that satisfied the catalytic contacts specified in the HG-1 and HG-2 designs were retained and stabilized through packing of the surrounding amino acid side chains. We generated five designs: 1THF-1, 1THF-2, 1A53-1, 1A53-2, and 1A53-3 (SI Appendix, Table S2, see [1]). Despite using the same base position as in the Röthlisberger designs, our 1THF- and 1A53-based designs differ by eight to ten mutations and give rise to active site geometries that are distinct from the Röthlisberger designs (SI Appendix, Figures S8 and S9, see [1]). These differences can be attributed to variations in the geometries used to define the active site as well as differences in the ligand pose sampling methods and force field used by Rosetta and our method. Three of the five designs showed significant activity over background (SI Appendix, Figure S10, see [1]), which indicates that multiple, geometrically unique active sites for KE catalysis can be generated from the same scaffold.

Crystallographic Analysis of 1A53-2

X-ray crystal structures of 1A53-2 were determined in the apo and 5-NBT-bound forms to 1.6- and 1.5-Å resolution, respectively. The full protein rmsd for the ligand-bound crystal structure with the design model is 0.51 Å, which indicates that the overall fold is maintained. Active site side-chain conformations in the cocrystal structure are in general agreement with the design (Figure 6A). As in the case of the HG-2 cocrystal structure, the position of the TSA is flipped from the designed orientation. Importantly, however, the ligand maintains a catalytically competent contact with the putative base (E178). The apo structure shows that the W210 side chain rotates from the catalytically relevant stacking position seen in the cocrystal structure to fill the substrate binding pocket (Figure 6B). The data collection and refinement statistics for these structures are summarized in SI Appendix, Table S3.

B.4 Conclusions

The iterative approach to computational enzyme design described here has led to the most active computationally designed enzyme catalyst for the KE to date. Inactive designs were probed by X-ray crystallography and MD simulations to learn the likely causes of inactivity. These data informed the next round of design and led to active enzymes. In this way, computational methods and crystallography were used, rather than combinatorial experimental approaches, to create effective enzyme catalysts. We believe that this iterative approach constitutes a significant advance in enzyme design methodology that, in addition to leading to improved designs, should contribute to a more complete understanding of the mechanisms of enzymatic activity. The relocation of the active site into the core of the HG-2 scaffold is a departure from previous enzyme design procedures, which focus designs solely in natural binding pockets of the scaffold [3-5]. Although the site of the catalytic base in the HG-2 active site was manually selected, a subsequent broader computational search for possible active sites also identified D127 among a large list of potential base positions outside of the natural binding pocket. The possibility of expanded active site searches suggests an opportunity for the improvement of computational design methodology to more efficiently carry out these large searches and to rank identified active site possibilities by their likelihood of supporting catalysis.

As with previous computationally designed enzymes, the activity levels reported here are low compared to many natural enzymes. Directed evolution has been shown to be an effective strategy to increase the activity of designed enzymes [3, 23, 24] and may offer insight into the deficiencies in the design. All-atom explicit solvent MD simulations have previously been shown to be effective at recapitulating the activity of computationally designed KE enzymes [12]. Here, MD was carried out prior to experimentation for all cases except HG-1, and the integration of MD into the iterative design process proved to be useful for identifying underlying problems in the structure and dynamics of HG-1 and in guiding the improvement of HG-2. The recent design of enzymes that stereoselectively promote a Diels-Alder reaction demonstrates the applicability of MD to more complicated chemistries [5].

The discrepancy between the ligand orientation in the modeled structures and in the crystal structures of HG-2 and 1A53-2 may be due to the inaccurate modeling of the TS ligand and/or inadequate sampling of possible ligand positions within the active site. Improvements to the force field may be necessary for accurate modeling of the ligand's nitro group in a hydrophobic environment. In addition, the utility of combining computational protein design with MD simulations suggests that future inclusion of full backbone flexibility, loop modeling, and MD move sets directly into computational design procedures may lead to more accurate predictions of ligand positions and improved de novo designed enzymes.

B.5 Materials and Methods

Crystals of HG-2/NBT, 1A53-2/NBT, and apo 1A53-2 were obtained through sitting-drop vapor diffusion carried out at room temperature with a protein concentration of 9.5 mg/mL. Co-crystallization for HG-2/NBT and 1A53-2/NBT was achieved through pre-incubation of the protein with 5 mM 5-nitrobenzotriazole (5-NBT, Ryan Scientific) prior to crystallization trials. A 100 mM stock solution of 5-NBT was prepared in DMSO before combining with the protein. Reservoir solutions for HG-2/NBT (0.1 M sodium acetate pH 4.6, 2 M ammonium sulfate), 1A53-2/NBT (0.1 M sodium citrate/citric acid pH 5.6, 0.2 M potassium sodium tartrate, 2 M ammonium sulfate), and apo 1A53-2 (0.1 M Bis-Tris pH 5.5, 0.2 M ammonium acetate, and 25% PEG 3350) were combined with protein in a 1:1 ratio. A single HG-2/NBT, multiple cube-like 1A53-2/NBT, and several plate-like apo 1A53-2 crystals developed with a minimum growth time of one month. The crystals were cryo-protected with paraffin oil and shipped to the Stanford Synchrotron Radiation Lightsource, beamline 12-2 for remote data collection. Diffraction data were processed with the program MOSFLM using the interface iMOSFLM [25].

Data were scaled using the program SCALA [26]. Molecular replacement was carried out with PHASER [26, 27]. The coordinates for Thermoascus aurantiacus xylanase I (PDB code 1GOR) [28] and Sulfolobus solfataricus (PDB code 1A53) [21] were modified to contain alanine at all point mutations in the designs and were subsequently used as the molecular replacement starting models for HG-2 and 1A53-2, respectively. Model building was carried out using COOT [29]. The structure was refined using REFMAC [30] and PHENIX [31]. Backbone density for the HG-2 structure appeared in two distinct backbone conformations in chain B, similar to the dual backbone conformation found in the structure of red fluorescent protein variant FP611 (PDB code 3E5T) [32]. The apo structure of 1A53-2 was processed with a twinning fraction of 0.13 towards the end of refinement. Crystallographic data statistics are summarized in Table S2.

B.6 Tables and Figures

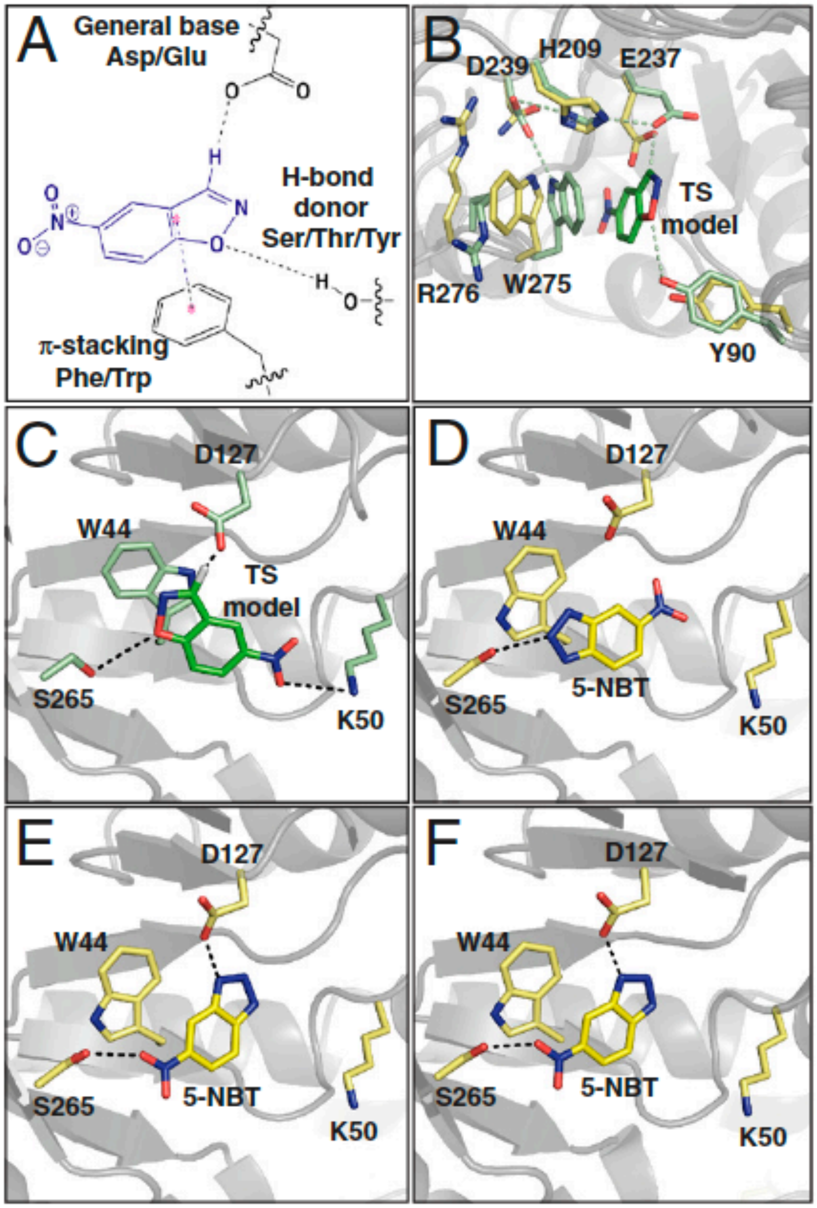


Figure 2. *KE enzyme design models and crystal structures.* (A) KE idealized active site. (B) Overlay of HG-1 crystal structure active site residues (yellow) with design model (green). (C) The HG-2 design model. (D) and (E) Crystal structure of HG-2 active site, chain A. The two conformations of the TSA 5-NBT are shown separately for clarity. (F) Crystal structure of HG-2 active site, chain B with the single observed conformation of the TSA.

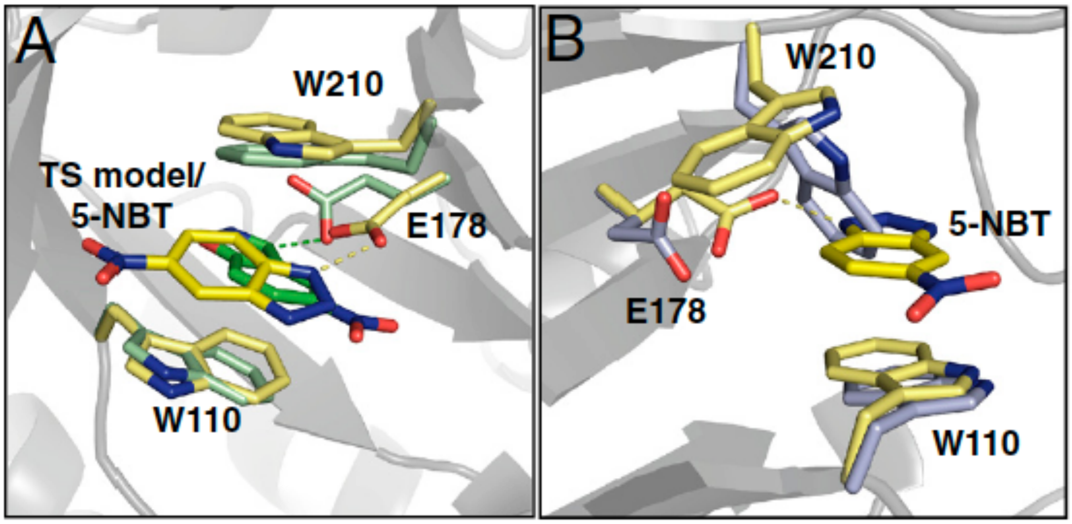


Figure 6. *Crystal structures of 1A53-2.* (A) Overlay of 1A53-2 holostructure (yellow) and the design model (green). (B) Overlay of 1A53-2 apo crystal structure (lavender) and holostructure (yellow).

B.8 Supplementary Table

Table S3. Data collection and refinement statistics for HG-1, HG-2, and 1A53-2 crystal structures

	HG-1 (PDB: 3O2L)	HG-2/NBT (PDB: 3NYD)	1A53-2/NBT (PDB: 3NZ1)	1A53-2 (PDB: 3NYZ)
Data Collection				
Space group	$P2_12_12_1$	$P2_12_12_1$	$P3_121$	$P2_1$
Cell dimensions				
a, b, c, Å	48.3, 72.5, 74.6	75.8, 78.1, 98.2	60.7, 60.7, 120.2	38.0, 46.3, 127.0
α, β, γ, °	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 120.0	90.0, 92.0, 90.0
Resolution, Å	2.0	1.2	1.6	1.5
R_{sym} , %*	2.5 (4.4)	4.6 (33.8)	6.6 (50.4)	11.6 (45.6)
I/σI*	30.7 (21.1)	10.3 (1.9)	8.8 (17.6)	4.8 (2.0)
Completeness, %*	99.2 (98.2)	98.8 (96.4)	99.8 (95.8)	99.4 (99.9)
Redundancy*	2.7 (2.6)	3.5 (2.3)	4.0 (3.7)	2.9 (2.9)
Wavelength, Å	1.54	0.99	1.01	1.01
Refinement				
Resolution (Å)*	35.4-2.0 (2.1-2.0)	34.1-1.23 (1.3-1.23)	52.6-1.6 (1.6-1.5)	36.7-1.5 (1.6-1.5)
Number of reflections				
Working set	17130	158059	37189	67105
Test set	874	8337	1856	2041
R_{work}/R_{free} ,%	16.7/22.9	15.7/19.6	17.2/21.0	20.2/25.1
Number of atoms				
Protein	2326	4591	2007	4089
Ligand/Ion	0	50	72	5
Water	218	670	181	237
RMS deviations				
Bond lengths, Å	0.024	0.028	0.027	0.010
Bond angles, °	1.80	2.35	2.40	1.24
Twin fraction	N/A	N/A	N/A	0.13

*Parentheses indicate statistics for outer shell of data.
N/A, not applicable.

B.9 References

1. Privett, H.K., Kiss G., Lee T.M., Blomberg R., Chica R.A., Thomas L.M., Hilvert D., Houk K.N. and Mayo S.L. (2012) Iterative approach to computational enzyme design. Proceedings of the National Academy of Sciences 109:3790-3795.
2. Bolon, D.N. and Mayo S.L. (2001) Enzyme-like proteins by computational design. Proceedings of the National Academy of Sciences 98:14274-14279.
3. Röthlisberger, D., Khersonsky O., Wollacott A.M., Jiang L., DeChancie J., Betker J., Gallaher J.L., Althoff E.A., Zanghellini A. and Dym O. (2008) Kemp elimination catalysts by computational enzyme design. Nature 453:190-195.
4. Jiang, L., Althoff E.A., Clemente F.R., Doyle L., Röthlisberger D., Zanghellini A., Gallaher J.L., Betker J.L., Tanaka F. and Barbas C.F. (2008) De novo computational design of retro-aldol enzymes. Science 319:1387-1391.
5. Siegel, J.B., Zanghellini A., Lovick H.M., Kiss G., Lambert A.R., Clair J.L.S., Gallaher J.L., Hilvert D., Gelb M.H. and Stoddard B.L. (2010) Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. Science 329:309-313.
6. Dahiyat, B.I. and Mayo S.L. (1996) Protein design automation. Protein Science 5:895-903.
7. Street, A.G. and Mayo S.L. (1999) Computational protein design. Structure 7:R105-R109.
8. Dahiyat, B.I. and Mayo S.L. (1997) Probing the role of packing specificity in protein design. Proceedings of the National Academy of Sciences 94:10172-10177.
9. Dahiyat, B.I., Benjamin Gordon D. and Mayo S.L. (1997) Automated design of the surface positions of protein helices. Protein Science 6:1333-1337.
10. Zollars, E.S., Marshall S.A. and Mayo S.L. (2006) Simple electrostatic model improves designed protein sequences. Protein Science 15:2014-2018.
11. Morin, A., Kaufmann K.W., Fortenberry C., Harp J.M., Mizoue L.S. and Meiler J. (2011) Computational design of an endo-1, 4-β-xylanase ligand binding site. Protein Engineering Design and Selection 24:503-516.
12. Kiss, G., Röthlisberger D., Baker D. and Houk K.N. (2010) Evaluation and ranking of enzyme designs. Protein Science 19:1760-1773.
13. Frushicheva, M.P., Cao J., Chu Z.T. and Warshel A. (2010) Exploring challenges in rational enzyme design by simulating the catalysis in artificial kemp eliminase. Proceedings of the National Academy of Sciences 107:16869-16874.
14. Kemp, D.S. and Casey M.L. (1973) Physical organic chemistry of benzisoxazoles. II. Linearity of the Broensted free energy relation for the base-catalyzed decomposition of benzisoxazoles. Journal of the American Chemical Society 95:6670-6680.
15. Thorn, S.N., Daniels R.G., Auditor M.-T.M. and Hilvert D. (1995) Large rate accelerations in antibody catalysis by strategic use of haptenic charge.
16. Hollfelder, F., Kirby A.J. and Tawfik D.S. (1996) Off-the-shelf proteins that rival tailor-made antibodies as catalysts. Nature 383:60-63.

17. Korendovych, I.V., Kulp D.W., Wu Y., Cheng H., Roder H. and DeGrado W.F. (2011) Design of a switchable eliminase. Proceedings of the National Academy of Sciences 108:6823-6827.

18. Lassila, J.K., Privett H.K., Allen B.D. and Mayo S.L. (2006) Combinatorial methods for small-molecule placement in computational enzyme design. Proceedings of the National Academy of Sciences 103:16710-16715.

19. Kemp, D.S. and Paul K.G. (1975) Physical organic chemistry of benzisoxazoles. III. Mechanism and the effects of solvents on rates of decarboxylation of benzisoxazole-3-carboxylic acids. Journal of the American Chemical Society 97:7305-7312.

20. Casey, M.L., Kemp D.S., Paul K.G. and Cox D.D. (1973) Physical organic chemistry of benzisoxazoles. I. Mechanism of the base-catalyzed decomposition of benzisoxazoles. The Journal of Organic Chemistry 38:2294-2301.

21. Hennig, M., Darimont B.D., Jansonius J.N. and Kirschner K. (2002) The catalytic mechanism of indole-3-glycerol phosphate synthase: crystal structures of complexes of the enzyme from *Sulfolobus solfataricus* with substrate analogue, cubstrate, and product. Journal of Molecular Biology 319:757-766.

22. Lang, D., Thoma R., Henn-Sax M., Sterner R. and Wilmanns M. (2000) Structural evidence for evolution of the β/α barrel scaffold by gene duplication and fusion. Science 289:1546-1550.

23. Khersonsky, O., Röthlisberger D., Wollacott A.M., Murphy P., Dym O., Albeck S., Kiss G., Houk K.N., Baker D. and Tawfik D.S. (2011) Optimization of the *in-silico* designed Kemp eliminase KE70 by computational design and directed evolution. Journal of Molecular Biology 407:391-412.

24. Khersonsky, O., Röthlisberger D., Dym O., Albeck S., Jackson C.J., Baker D. and Tawfik D.S. (2010) Evolutionary optimization of computationally designed enzymes: Kemp eliminases of the KE07 series. Journal of Molecular Biology 396:1025-1042.

25. Leslie, A.G.W. (1992) Recent changes to the MOSFLM package for processing film and image plate data.

26. Storoni, L.C., McCoy A.J. and Read R.J. (2004) Likelihood-enhanced fast rotation functions. Acta Crystallographica Section D: Biological Crystallography 60:432-438.

27. McCoy, A.J., Grosse-Kunstleve R.W., Storoni L.C. and Read R.J. (2005) Likelihood-enhanced fast translation functions. Acta Crystallographica Section D: Biological Crystallography 61:458-464.

28. Lo Leggio, L., Kalogiannis S., Eckert K., Teixeira S., Bhat M.K., Andrei C., Pickersgill R.W. and Larsen S. (2001) Substrate specificity and subsite mobility in *T. aurantiacus* xylanase 10A. FEBS Letters 509:303-308.

29. Emsley, P. and Cowtan K. (2004) Coot: model-building tools for molecular graphics. Acta Crystallographica Section D: Biological Crystallography 60:2126-2132.

30. Murshudov, G.N., Vagin A.A. and Dodson E.J. (1997) Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallographica Section D: Biological Crystallography 53:240-255.

31. Adams, P.D., Grosse-Kunstleve R.W., Hung L.W., Ioerger T.R., McCoy A.J., Moriarty N.W., Read R.J., Sacchettini J.C., Sauter N.K. and Terwilliger T.C. (2002) PHENIX: building new software for automated crystallographic structure determination. Acta Crystallographica Section D: Biological Crystallography 58:1948-1954.

32. Nienhaus, K., Nar H., Heilker R., Wiedenmann J.r. and Nienhaus G.U. (2008) Trans− Cis Isomerization is Responsible for the Red-Shifted Fluorescence in Variants of the Red Fluorescent Protein eqFP611. Journal of the American Chemical Society 130:12578-12579.