Title

Selected Topics in Metabolic and Protein Engineering: Identifying the Bottleneck Step in Triazine Degradation, Characterization of Various Supercharging Methods on Protein Stability and Expression, And Assessment of Tools for Prediction of Impacts of...

Author

Connolly, Morgan

Publication Date

2022

Peer reviewed|Thesis/dissertation

Selected Topics in Metabolic and Protein Engineering: Identifying the Bottleneck Step in Triazine Degradation, Characterization of Various Supercharging Methods on Protein Stability and Expression, And Assessment of Tools for Prediction of Impacts of Point Mutation on Protein Stability

By

MORGAN P. CONNOLLY
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in
Microbiology
in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

_____
Justin Siegel, Chair

_____
Shota Atsumi

_____
Patrick Shih

Committee in Charge
2022

Table of Contents

Abstract

Biotechnology has the potential to deliver solutions to many global problems in medicine, materials science, nutrition, agriculture, natural resource preservation, and energy. Engineered cells and enzymes can perform chemical transformations that are rare or unknown in nature, and even catalyze reactions not accessible to traditional synthetic chemistry while also operating at gentler, more environmentally friendly conditions. Decreases in the price of DNA sequencing and synthesis has led to generation of vast databases that can be screened for any imaginable function. These sequence databases are even more powerful now due to the development of software to enable rapid generation of 3D protein structures, like AlphaFold2. However, tools to predict the function of these proteins or their performance in engineered cells are not yet robust, leading to long development times and limited successful applications to date. New tools and methods must be developed for the true potential of biotechnology to be unlocked.

For my thesis, I explore metabolic pathway construction and screening, protein design, and protein sequence-structure-function relationships across broad contexts with the objective of tool and knowledge development for future efforts in biotechnology. My first chapter discusses the introduction of the triazine degradation pathway, of interest for remediation of contaminated sites, into *E. coli* and methods for characterizing pathway flux and identifying bottlenecks to guide engineering efforts and limit accumulation of metabolic intermediates. My second chapter focuses on methods for the design of supercharged proteins, which have many interesting potential applications, and parameters that increase the likelihood of successful design of these proteins. My third chapter regards the generation and characterization of a library of single point mutations in the enzyme β-glucosidase B and use of kinetic data to predict the effects of changes

in sequence on enzyme function. These seemingly disparate topics all serve to improve tools for protein screening, production, functional prediction, and application, addressing several gaps toward improved development timelines and success rates for biocatalysts.

Acknowledgements

This work was made possible by the many kind and supportive friends, family members, and coworkers who have helped me get to this point. Thank you to my parents for the support in everything I've done to date. A special thanks to my partner, Jesse Mattson, who has been with me through the roller coaster that has been graduate school. For their support and, at times, commiseration, I thank all of my labmates past and present and in particular Dr. Pamela Denish, Dr. Kathryn Guggenheim, Dr. Terrence O'Brien, Ryan Caster, Tram Nguyen, Alexander Kehl, and Emma Luu. Thank you to Akanksha Majumdar for her efforts on the supercharging project discussed in Chapter 2. Chapter 3 would not be possible without the efforts of Simon Kit Sang Chu, Peishan Huang, Ashley Vater, and many undergraduate students of the Siegel lab. Thank you to my friends Kaitlyn Mulligan, Nicole Hart, Kristy Northrup, Julianna Kluger, Monica Serrano, and Gavin Kinder for their encouragement and support. Lastly I would like to thank my grandparents, Mary and Chuck Drapeau, who did not get the chance to see me finish graduate school, but believed in me beyond compar

Chapter 1:

Characterization of Bottleneck Steps in Triazine Degradation Illuminates the Path Toward Pathway Improvement

Abstract

Bioremediation has long been a desirable approach for removal of contaminants from affected sites. However, the application of this method is often restricted by low rates of degradation and the accumulation of intermediate metabolites. The triazine degradation pathway from *Pseudomonas* sp. ADP has been characterized to identify the enzymes responsible and the rate of atrazine hydrolysis to the first intermediate hydroxyatrazine has been determined. However, there has not been detailed characterization of the pathway flux through each intermediate, so the rate-limiting step for atrazine degradation is not known. Through *in vitro* and *in vivo* assays, it was determined that the dechlorination of atrazine to hydroxyatrazine was rate limiting in *Escherichia coli* expressing either an AtzABC or TrzNAtzBC operon. This knowledge is important to guide engineering strategies to improve pathway flux and reduce accumulation of intermediates.

Introduction

Environmental contamination by man-made chemicals is a widespread problem, with 1,333 sites in the United States on the Superfund National Priorities List as of July 2022(1). Bioremediation is a promising approach for removal of contaminants from such sites due to potentially lower costs and environmental impact relative to other remediation methods. However, the application of bioremediation at contaminated sites is often limited by slow

turnover rates and the accumulation of metabolic intermediates, which may pose their own risks. Therefore, strategies for increasing flux through degradative pathways while limiting the buildup of intermediate products will be critical for the expansion of bioremediation.

Triazine herbicides are used for weed control in commercial production of various crops. Due to their high usage they are common ground and stream water contaminants in the United States and levels of these compounds can exceed the U.S. EPA maximum contaminant levels in waters draining areas of high usage (2). A robust system for microbial degradation of triazines could be a useful tool in mitigating risks posed by exposure to triazine contaminants.

The bacteria *Pseudomonas* sp. ADP (PADP) has been extensively studied for its ability to degrade the triazine herbicide atrazine. This strain is able to fully mineralize the triazine ring of atrazine to ammonia and carbon dioxide, making it a promising system for bioremediation due to the non-toxic nature of the products (3). However, field studies using PADP have shown mixed results in terms of the performance of this strain outside of laboratory conditions, making its applicability in bioremediation dubious (4–7). The genetic and phenotypic instability of PADP also limits the relevance of this organism in bioremediation outside of laboratory environments ((8–10). Due to the limited range of substrates catalyzed by the atrazine chlorohydrolase (AtzA) enzyme of PADP, many commercially and environmentally relevant triazines are not degraded by this organism (11). In this study we have therefore chosen to focus on the triazine degradation pathway expressed heterologously in the model system *Escherichia coli* rather than the native PADP. Field studies using *E. coli* expressing AtzA, catalyzing the first reaction of atrazine degradation, have shown similar rates of atrazine degradation to those of PADP (12). The vast array of genetic tools available in *E. coli* presents the opportunity to improve pathway flux to

improve bioremediation of triazine herbicides including the ability to utilize enzymes with broader substrate ranges than AtzA from PADP, such as TrzN from *Arthrobacter aurescens* (13).

Many studies on triazine degradation have focused on the initial degradation of atrazine to hydroxyatrazine, but as the risks associated with hydroxyatrazine are not well studied, a more comprehensive approach to degradation would focus on the degradation to either the final products ammonia and carbon dioxide, or the intermediate cyanuric acid, as environmental bacteria able to degrade cyanuric acid are widespread (14, 15). It should also be noted that the EPA guideline for cyanuric acid is over 10,000 times higher (40 mg/L) than the maximum contaminant level for atrazine of 0.003 mg/L so cyanuric acid produced as a result of atrazine degradation at relevant concentrations is unlikely to present a health concern (16, 17). This study will therefore focus on improving degradation of atrazine to cyanuric acid using *E. coli.*

Materials and Methods

Bacterial strains and plasmids

A sample of *Pseudomonas* sp. ADP was graciously provided by the laboratory of Dr. Lawrence Wackett at University of Minnesota. This strain was grown on minimal media with 100 ppm atrazine as the sole nitrogen source as previously described (18).

Coding sequences of TrzN, AtzA, AtzB, and AtzC were codon optimized and cloned under control of a T7 promoter into the pET29b(+) vector by Twist Biosciences. The TrzNAtzBCDEF operon was synthesized by Life Technologies and cloned into pASK-IBA63a+ under control of the tetracycline promoter by restriction digest and ligation. The coding sequences for AtzDEF were removed by restriction and Gibson cloning. The TrzN coding

sequence was replaced with AtzA by Gibson cloning to produce the pASK-AtzABC plasmid. DNA sequences used can be found in Supplementary Material S1-1.

Quantitation of pathway metabolites

All assay materials were assessed using high performance liquid chromatography-mass spectrometry (LC-MS). A Synergi 4 μm Fusion-RP 80 Å column was used with water with 0.1% formic and acid and acetonitrile with 0.1% formic acid as mobile phases. Atrazine and hydrozyatrazine were quantified based on the area of the M+1 ion in single ion monitoring (SIM) positive mode. N-isopropylammelide was quantified based on the area of M-1 ion in SIM negative mode. Atrazine, hydroxyatrazine, and cyanuric acid were quantified compared to commercial chemical standards. N-isopropylammelide standards were prepared by incubation of hydroxyatrazine with purified AtzB for 72 hours, after which no hydroxyatrazine was detected. Cyanuric acid was not robustly detected by this method, so cyanuric acid produced in both the *in vitro* and *in vivo* assays was calculated by the molar deficit after quantitation of the other three metabolites.

In vitro enzyme assays

For enzyme deficit assays, the "Balanced" enzyme mix represents a mixture of 100 nM of each enzyme specified (i.e. TrzN with AtzB and AtzC) with 150 uM atrazine. Reactions were quenched with acetonitrile and analyzed by the given LC-MS method. Enzyme deficit assays were incubated for 1 hour at room temperature. An enzyme deficit denotes a 10-fold deficit of the specified enzyme relative to the other enzymes in the mixture, resulting in a concentration of 10 nM in the reaction mixture.

For enzyme enrichment assays, the "Balanced" enzyme mix represents a mixture of 10 nM of each enzyme specified (i.e. TrzN with AtzB and AtzC) with 150 uM atrazine. Reactions were quenched with acetonitrile and analyzed by the given LC-MS method. Enzyme enrichment assays were incubated for 4 hours at room temperature. An enzyme enrichment denotes a 10-fold enrichment of the specified enzyme relative to the other enzymes in the mixture, resulting in a concentration of 100 nM in the reaction mixture. Both enzyme deficit and enrichment assays were conducted as technical duplicates. Purified green fluorescent protein (GFP) at the same concentration as the enzymes was incubated with 150 uM atrazine as a negative control.

In vivo atrazine degradation assays

To prepare cells for the *in vivo* assay, *E. coli* strains with pET-29b+ plasmids were induced with 1 mM IPTG and *E. coli* strains with pASK-IBA63a+ were induced with 200 µg/L anhydrotetracycline. Induced cells were incubated at 18^{o}C for 20 hours, harvested by centrifugation, and resuspended in phosphate buffered saline with metals added to an optical density of 1.0. PADP were cultured in minimal media with atrazine as the sole nitrogen source as previously described, harvested by centrifugation, and resuspended in phosphate buffered saline with metals added to an optical density of 1.0. The resulting cell suspensions were incubated with 150 µM atrazine or hydroxyatrazine then quenched with acetonitrile and analyzed by the given LC-MS method. Both enzyme deficit and enrichment assays were conducted as technical duplicates. *E. coli* expressing either pET-29b+-GFP or pASK-IBA63a+-GFP were incubated with 150 uM atrazine as a negative control.

Results

In vitro enzyme enrichment and depletion

Enrichment of the first enzyme of the triazine degradation pathway, TrzN, increased flux to cyanuric acid. The 10-fold enrichment of TrzN relative to AtzB and AtzC led to complete hydrolysis of atrazine within 4 hours, with only hydroxyatrazine and N-isopropylammelide present at measurable levels. In comparison, when TrzN when stoichiometrically based with AtzB and AtzC with each enzyme at 10 nM, only 23.3% of the available atrazine was hydrolyzed within 4 hours (Figure 1-1A). Interestingly, a small amount (7.8 uM) of cyanuric acid was calculated in the Balanced enzyme mixture, but no cyanuric acid was calculated in the TrzN-enriched sample. This may suggest an over-estimation in the calculations used to quantify atrazine, hydroxyatrazine, and N-isopropylammelide relative to the chemical standards. This is further suggested by the total metabolite concentration calculated in the AtzB-enriched sample being greater than the 150 uM which was initially input to the assay.

A deficit of TrzN greatly hindered flux toward cyanuric acid. While the stoichiometrically balanced mixture of TrzN, AtzB, and AtzC at 100 nM of each enzyme achieved complete atrazine hydrolysis and produced a calculated 62.9 uM cyanuric acid, the 10-fold depletion of TrzN led to only 19.5% atrazine hydrolysis and produced only a calculated 20.9 uM cyanuric acid (Figure 1-1B). No atrazine loss was observed for reactions incubated with GFP.

In vivo degradation of triazines in E. coli and PADP

Rates of atrazine hydrolysis were substantially lower in *E. coli* expressing both pASK-AtzABC and pASK-TrzNAtzBC than in PADP (Table 1-1). Interestingly, in *E. coli* strains fed

atrazine, a maximum concentration of 4 uM hydroxyatrazine and no N-isopropylammelide were detected during the time course, suggesting that hydrolysis of atrazine to hydroxyatrazine is rate limiting in these strains. To further this point, hydrolysis of hydroxyatrazine was much higher than that of atrazine in both pASK-TrzNAtzBC and pASK-AtzABC, reaching 86% of hydroxyatrazine degraded, suggesting that the pathway can operate at higher fluxes after the initial transformation from atrazine to hydroxyatrazine is complete (Table 1-1).

It is possible to achieve high rates of atrazine hydrolysis in *E. coli*. High levels of AtzA were produced by IPTG-induced expression via T7 promoter of pET29b+ whereas achieving high expression of TrzN required both the T7 expression system and used of the mutant TrzN with higher expression (heTrzN) identified by Jackson et al. (19) When AtzA or TrzN are highly expressed in *E. coli*, rates of atrazine hydrolysis approached that of PADP (Figure 1-2). In addition to demonstrating that high rates of atrazine hydrolysis are possible in *E. coli*, similarity of these strains to PADP suggests that differential transport of atrazine is not responsible for the difference in atrazine hydrolysis between the engineered *E. coli* and PADP.

Discussion

Hydrolysis of atrazine to hydroxyatrazine limits flux through the atrazine degradation pathway. The system is likely secondarily constrained by hydrolysis of hydroxyatrazine to N-isopropyl ammelide. Given the performance of *in vitro* mixtures with deficits of AtzC, limited availability of this enzyme does not have substantial impacts on performance, as mixtures with a 10-fold deficit of AtzC produce similar amount of cyanuric acid as when AtzC is in stoichiometric balance with the other pathway enzymes.

Enzyme expression, rather than activity, seems to most restrict the rate of atrazine hydrolysis and thus limit progress of the pathway toward later metabolites. This is shown by the stark impact of a 10-fold stoichiometric deficit of TrzN in *in vitro* mixtures of pathway enzymes. Deficit of TrzN caused reduced production of cyanuric acid in 4 hours by 66.8% compared to the balanced 100 nM mixture. This is further supported by the complete loss of atrazine when TrzN was enriched within 4 hours. The balanced 10 nM enzyme mixture achieved only 23.3% degradation of atrazine in the same amount of time. This shows the critical role that the level of available atrazine chlorohydrolase plays in the performance of the triazine degradative pathway. Strong expression of AtzA or TrzN in *E. coli* was sufficient to achieve improve atrazine degradation, further suggesting that enzyme expression rather than activity is the key constraint on the current triazine degradation operon system in *E. coli*.

The assertion that enzyme expression is the factor most limiting flux through the triazine degradative pathway agrees with prior data regarding kinetics of the pathway enzymes. Published kinetic constants for TrzN, AtzA, AtzB, and AtzC fall within 10-fold of each other on their respective substrates (20–23). Thus, no enzyme within the pathway has a large difference in activity on its substrate compared to other pathway enzymes so tuning of expression levels would be the main determinant of flux to the final product and accumulation of metabolites.

Knowledge of the factors most limiting atrazine hydrolysis in the recombinant *E. coli* system described here can be used to target engineering strategies to design an improved biocatalyst for remediation of triazine herbicides. As expression of the TrzN or AtzA enzyme seems to be most constraining in this system, investigation of alternative promoters, ribosome binding sites, and/or protein variants could be useful in overcoming this obstacle. Strains with improved rates of atrazine degradation and reduced accumulation of intermediates could be a

useful tool in removing triazine contaminants from the environment and helping to mitigate the risks associated with these compounds.

Figures

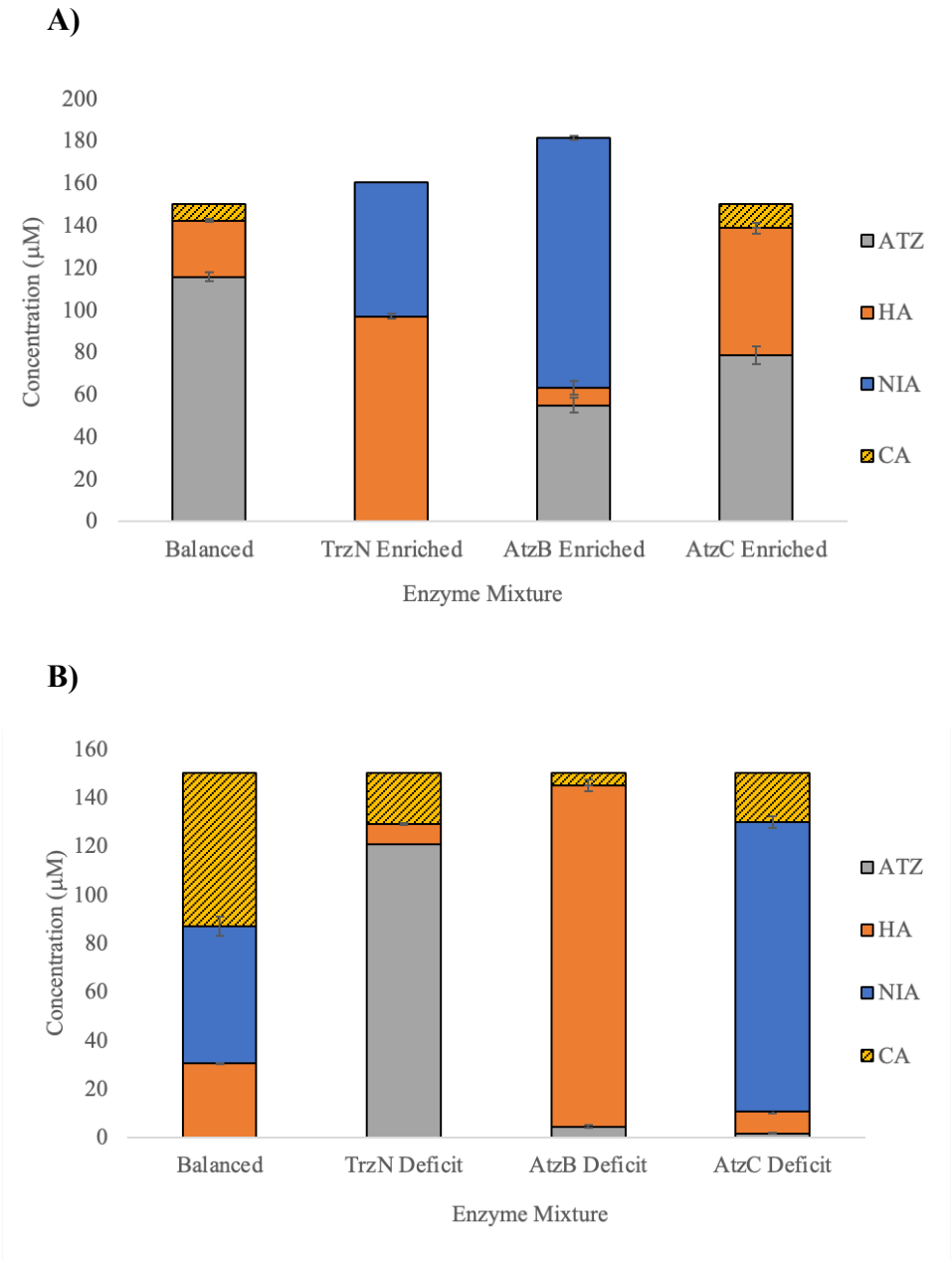


Figure 1-1. Concentrations of atrazine and its metabolites after incubation with the designated enzyme mixtures. "Enriched" enzymes are 100 nM with other enzymes at 10 nM, while "Deficit" enzymes are 10 nM with other enzymes at 100 nM. Deficit reactions were incubated for 1 hour while enrichment reactions were incubated for 4 hours. Compounds are abbreviated as

follows: ATZ = atrazine, HA = hydroxyatrazine, NIA = N-isopropylammelide, CA = cyanuric acid. Cyanuric acid (striped) was calculated based on the remaining stoichiometric amount after quantitation of atrazine, hydroxyatrazine, and N-isopropylammelide. Values shown represent the mean of technical duplicates, with error bars indicating +/- 1 standard deviation.

Table 1-1. Average and standard deviation substrate consumption for triazine degrading strains incubated with 150 μM atrazine or hydroxyatrazine for 24 hours

Strain	Substrate	Change in Concentration (μM)	St. Dev. (μM)
pASK-NBC	Atrazine	10.0	27.4
	Hydroxyatrazine	129.7	3.0
pASK-ABC	Atrazine	0.0	52.0
	Hydroxyatrazine	129.6	7.4
PADP	Atrazine	150.0	0.0
	Hydroxyatrazine	87.1	7.6

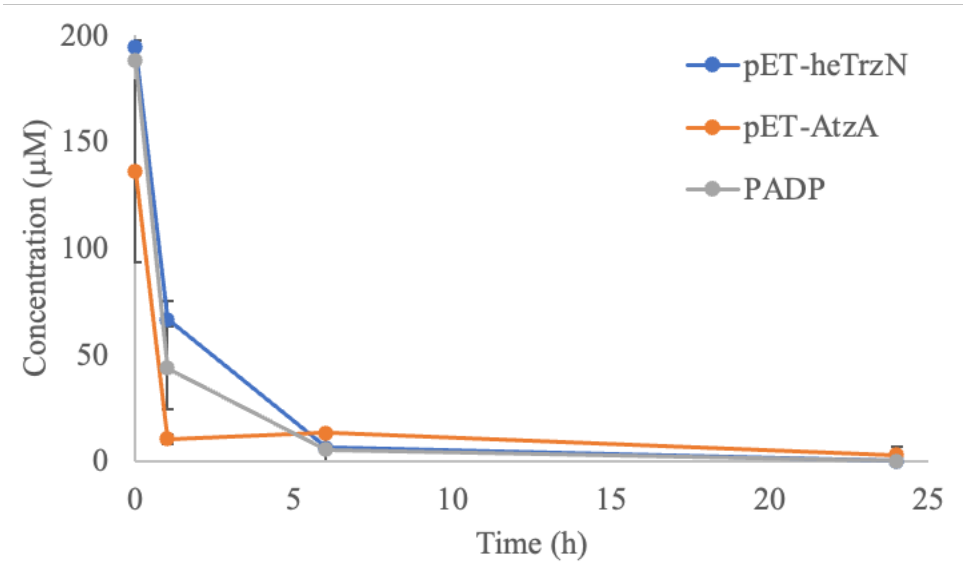


Figure 1-2. Atrazine concentration during incubation of *E. coli* expressing AtzA or high-expressing TrzN under the strong T7 promoter compared to *Pseudomonas* PADP. Points represent the mean of technical duplicates, with error bars indicating +/- 1 standard deviation.

References

1. Superfund: National Priorities List (NPL) | US EPA. https://www.epa.gov/superfund/superfund-national-priorities-list-npl. Retrieved 6 July 2022.
2. Robert J. Gilliom JEB, Jeffrey D. Martin NN, Paul E. Stackelberg GPT. 2007. The Quality of our Nation's Waters Pesticides in the Nation's Streams and Ground Water, 1992-2001. National Water-Quality Assessment Program. https://pubs.usgs.gov/circ/2005/1291/pdf/circ1291.pdf. Retrieved 17 September 2022.
3. Mandelbaum RT, Allan DL, Wackett LP. 1995. Isolation and Characterization of a Pseudomonas sp. That Mineralizes the s-Triazine Herbicide Atrazine. Appl Environ Microbiol 61:1451–1457.
4. Govantes F, Porrúa O, García-González V, Santero E. 2009. Atrazine biodegradation in the lab and in the field: Enzymatic activities and gene regulation. Microb Biotechnol 2:178–185.

5. Wackett L, Sadowsky M, Martinez B, Shapir N. 2002. Biodegradation of atrazine and related s-triazine compounds: From enzymes to field studies. Appl Microbiol Biotechnol 58:39–45.
6. Topp E. 2001. A comparison of three atrazine-degrading bacteria for soil bioremediation. Biology and Fertility of Soils 2001 33:6 33:529–534.
7. Chelinho S, Moreira-Santos M, Lima D, Silva C, Viana P, André S, Lopes I, Ribeiro R, Fialho AM, Viegas CA, Sousa JP. 2010. Cleanup of atrazine-contaminated soils: Ecotoxicological study on the efficacy of a bioremediation tool with Pseudomonas sp. ADP. J Soils Sediments 10:568–578.
8. de Souza ML, Wackett LP, Sadowsky MJ. 1998. The atzABC genes encoding atrazine catabolism are located on a self-transmissible plasmid in Pseudomonas sp. strain ADP. Appl Environ Microbiol 64:2323–2326.
9. Govantes F, García-González V, Porrúa O, Platero AI, Jiménez-Fernández A, Santero E. 2010. Regulation of the atrazine-degradative genes in Pseudomonas sp. strain ADP. FEMS Microbiol Lett 310:1–8.
10. Martinez B, Tomkins J, Wackett LP, Wing R, Sadowsky MJ. 2001. Complete Nucleotide Sequence and Organization of the Atrazine Catabolic Plasmid pADP-1 from Pseudomonas sp. Strain ADP. J Bacteriol 183:5684.
11. Seffernick JL, Johnson G, Sadowsky MJ, Wackett LP. 2000. Substrate specificity of atrazine chlorohydrolase and atrazine-catabolizing bacteria. Appl Environ Microbiol 66:4247–4252.
12. Strong LC, McTavish H, Sadowsky MJ, Wackett LP. 2000. Field-scale remediation of atrazine-contaminated soil using recombinant Escherichia coli expressing atrazine chlorohydrolase. Environ Microbiol 2:91–98.

13. Shapir N, Rosendahl C, Johnson G, Andreina M, Sadowsky MJ, Wackett LP. 2005. Substrate specificity and colorimetric assay for recombinant TrzN derived from Arthrobacter aurescens TC1. Appl Environ Microbiol 71:2214–2220.

14. Aukema KG, Tassoulas LJ, Robinson SL, Konopatski JF, Bygd MD, Wackett LP. 2020. Cyanuric Acid Biodegradation via Biuret: Physiology, Taxonomy, and Geospatial Distribution. Appl Environ Microbiol 86.

15. Seffernick JL, Wackett LP. 2016. Ancient Evolution and Recent Evolution Converge for the Biodegradation of Cyanuric Acid and Related Triazines. Appl Environ Microbiol 82:1638.

16. National Primary Drinking Water Regulations | US EPA. https://www.epa.gov/ground-water-and-drinking-water/national-primary-drinking-water-regulations#one. Retrieved 6 July 2022.

17. Usepa, Region, Foia. 2008. Guidelines for Drinking-water Quality THIRD EDITION INCORPORATING THE FIRST AND SECOND ADDENDA Volume 1 Recommendations Geneva 2008 WHO Library Cataloguing-in-Publication Data.

18. Mandelbaum RT, Wackett LP, Allan DL. 1993. Mineralization of the s-triazine ring of atrazine by stable bacterial mixed cultures. Appl Environ Microbiol 59:1695.

19. Jackson CJ, Coppin CW, Carr PD, Aleksandrov A, Wilding M, Sugrue E, Ubels J, Paks M, Newman J, Peat TS, Russell RJ, Field M, Weik M, Oakeshott JG, Scott C. 2014. 300-fold increase in production of the Zn2+-dependent dechlorinase trzN in soluble form via apoenzyme stabilization. Appl Environ Microbiol 80:4003–4011.

20. Zhou N, Wang J, Wang W, Wu X. 2021. Purification, characterization, and catalytic mechanism of N-Isopropylammelide isopropylaminohydrolase (AtzC) involved in the degradation of s-triazine herbicides. Environmental Pollution 268:115803.

21. Seffernick JL, Aleem A, Osborne JP, Johnson G, Sadowsky MJ, Wackett LP. 2007. Hydroxyatrazine N-Ethylaminohydrolase (AtzB): an Amidohydrolase Superfamily Enzyme Catalyzing Deamination and Dechlorination. J Bacteriol 189:6989.

22. Seffernick JL, Reynolds E, Fedorov AA, Fedorov E, Almo SC, Sadowsky MJ, Wackett LP. 2010. X-ray structure and mutational analysis of the atrazine chlorohydrolase TrzN. Journal of Biological Chemistry 285:30606–30614.

23. Scott C, Jackson CJ, Coppin CW, Mourant RG, Hilton ME, Sutherland TD, Russell RJ, Oakeshott JG. 2009. Catalytic Improvement and Evolution of Atrazine Chlorohydrolase. Appl Environ Microbiol 75:2184.

Supplementary Material

Figure S1-1. DNA sequences used in this study

>pET29-b+-heTrzN Insert

ATGATCCTGATTCGTGGCGCTCGCCGTGTAATTACGTTTGATGACCAAGACCGCGAG

CTCGAAGATGCCGATATCCTGATCGATGGCCCGAAAATTGTGGCTGTAGGTAAGAA

CCTGCCAGATGAAGACGTGGATCGCGTGATTGATGGTCGTGGGTGTATTGCTCTGCC

TGGTCTGATCAACACCCATCATCATCTGTACGAGGGCGCAATGCGCGCCATTCCACA

GCTCGAGCGCGTGACCATGTTTGAATGGCTCCGCGGTGTCTATGAACTGAATGCGCA

ATGGTGGCGCGACGGCAAATTCGGCCCTGATGTTGTGCGTGAGGTCGCGCGCGCGG

CGCTCCTGGAACTGCTGTTAGGTGGGTGTACGACGGTATCGGACCAGCACCCGATCT

TCCCCGGTGGGACCCCAGAACGTTATATTGATGCGACGATTGAAGCCGCACGCGAC

CTGGGCATCCGGTTTCATGCCGTGCGTGGCTCCATGACACTGGGTAAATCACAAGGC

GGCTTCTGCCCGGACGAGTTTGTCGAACCCGTGGACGCCGTGGTTAAGCACTGTCAA

CGCCTGATCGATAAATACCATGACCCGTCGCCGTACGCTATGGTTCGTATCGCGTTG

GGTCCGTGCTCACCTCCATACGATACGCCGGAATTATTTCGCGAATTTGCGCAAATG

GCACGCGACTACGATGTCCGCCTGCATACTCATTTCTATGAACCATTGGACGCGCGC

TACAGCCTTGAACTGTATGGTATGACGCCGTGGCGTTTTCTTGAGAAACACGGTTGG

GCGGGTGACCGTGTGTGGTTTGCGCATGCGGTGAAGCCGCCTGATGATGAAATTCCG

GAATTTGCCCGTGCGGGTACAGGCATCGCACACTGTATTGCGTCCGACCTTCGTATG

GGTTGGGGCCTTGCACCGATTCGTGAATACCTCGATGCCGGTATCCCTGTTGGGTTT

GGCACCACCGGTAGCGCCAGCAATGATGGTGGCAATCTGCTGGGCGATCTGCGCCT

GGCGATGCTGGCCCATCGCCCGGCTAATCCCAATGAACCGGAAAAGTGGTTGAGTG

CCCGTGAATTGCTGCGCATGGCTACTCGTGGTGGCGCCGAATGCTTGGGTCGTCCGG

ACCTTGGGGTCCTGGAGCCGGGTAAAGCCGCTGACATCGCGTGCTGGCGTTTGGACG

GGGTCGATCGGGTGGGTGTGCACGACCCGGCAATCGGGCTGATTATGACAGGCCTG

AGCGATCGTGCGCACCTGGTAATCGTGAATGGCCAGGTCCTGGTCGAAAATGAGCG

CCCTGTCACCGCCGATCTGGAGCGTATCGTTGCGGAAACTACAGCCCTCATCCCAAA

AAATTTG

>pET29b+-AtzA-Insert

ATGCAAACACTCAGCATCCAGCATGGTACGCTGGTCACCATGGATCAGTATCGTCGG

GTCCTGGGTGACAGCTGGGTACACGTTCAGGATGGCCGCATTGTAGCTCTGGGTGTG

CATGCGGAAAGCGTACCACCGCCCGCCGACCGGGTGATTGACGCCCGTGGCAAAGT

TGTCTTACCAGGCTTCATTAACGCCCATACGCACGTAAATCAAATCCTGCTGCGCGG

GGGACCGTCACATGGGCGTCAGTTCTACGACTGGCTGTTTAATGTGGTTTACCCAGG

CCAGAAGGCGATGCGTCCAGAGGATGTGGCCGTCGCTGTCCGCCTCTACTGCGCGG

AGGCGGTTCGTTCGGGGATTACCACCATCAACGAAAACGCTGATTCTGCCATCTACC

CGGGTAATATTGAGGCAGCAATGGCGGTGTATGGCGAAGTAGGCGTGCGCGTTGTT

TATGCACGTATGTTCTTCGATCGTATGGATGGCCGGATTCAGGGCTATGTGGATGCC

CTTAAAGCGCGCTCACCTCAGGTTGAGCTGTGTTCGATTATGGAAGAGACCGCAGTG

GCAAAAGACCGTATCACCGCACTGTCGGACCAGTACCACGGCACCGCCGGCGGTCG

CATCTCTGTGTGGCCGGCACCAGCAACGACCACCGCTGTGACAGTTGAAGGTATGC

GTTGGGCGCAGGCATTCGCACGTGATCGCGCCGTCATGTGGACCCTCCACATGGCGG

AGAGCGACCACGACGAACGTATCCATGGTATGAGCCCTGCGGAGTATATGGAGTGT

TATGGCCTGCTGGATGAACGTCTGCAGGTCGCCCATTGCGTGTACTTCGACCGTAAA

GATGTCCGCCTGCTGCATCGTCACAACGTAAAAGTTGCTAGCCAGGTTGTATCCAAC

GCCTACTTGGGATCGGGCGTCGCACCGGTGCCGGAAATGGTCGAGCGTGGCATGGC

CGTGGGAATCGGCACGGACAACGGCAATAGCAACGACAGCGTGAATATGATCGGTG

ACATGAAATTCATGGCCCATATTCATCGTGCGGTTCATCGCGACGCGGATGTGCTGA

CCCCAGAAAAAATTCTGGAGATGGCCACTATCGACGGTGCCCGTTCCTTAGGAATGG

ACCACGAAATTGGCTCCATCGAGACGGGCAAACGCGCGGATCTGATCCTGCTGGAT

TTGCGTCATCCCCAAACCACTCCGCATCACCACCTCGCGGCTACTATTGTGTTTCAG

GCTTATGGGAACGAAGTAGACACGGTTCTTATCGATGGGAATGTTGTCATGGAAAA

CCGCCGGCTGAGCTTCCTGCCGCCGGAACGTGAGCTGGCCTTTCTCGAAGAGGCGCA

GTCCCGCGCCACCGCGATCCTGCAGCGTGCTAACATGGTGGCCAACCCGGCTTGGCG

CAGCCTC

>pASK-IBA63a+-AtzA-Insert

GAAGTGCCATTCCGCCTGACCTGTGAAATGAATAGTTCGACAAAAATCTAGAAATA

ATTTTGTTTAACTTTAAGAAGGAGATATACCATGCAGACGCTGCTGATTCGTCACGG

CACGGTGGTTACCATGGACGATGACCGTCGCGTCCTGGAAGACGGATGGGTCCATG

TCCAGGACGGTCGCATTGTGGCGCTTGGCGTGCATGCTACGTCGGTCCCGCCACCTG

CAGATCGCGTGATTGATGCGCGCGGAAAAGTAGTGCTCCCGGGCTTTATTAACGCCC

ATACGCATGTCAACCAGATTCTGCTTCGTGGAGGCCCGAGCCACGGTCGTCAGTTTT

ACGATTGGCTTTATAATGTGGTGTATCCAGGCCTGAAGGCGATGCGGCCGGAAGAT

GTTGCCGTTGCCGTCCGTCTGTATTGCGCGGAAGCAGTGCGCAGTGGTATCACTACG

ATCAACGAAAACGCGGATTCCGCTATTTATCCGGGAAATATCGAAGCCGCCATGGC

CGTCTATGGTGAAGTGGGTGTACGTGTTGTTTACGCCCGTATGTTCTTCGATCGCGTC

GACGGCCGGCTGCAGGAATACGTGGACGCAATTTTTGCAAAAGCCCCGCAGGTGGA

GCTGTGCTCGATTTTTGAACCGACAGATAAAGCCAAAAAAGACATTGAACGCCTTGC

GGATAAATGGCACGGCACCGCTAACGGTCGTATTCGGGTTTGGCCGGCGCCTGCAA

CCCCGACGACCGTATCTGAAGAAGGTATGCGGTGGGCTCAGGAGTTTGCACGCGAC

CGTGGCGTCATGTGGACCCTGCATATGGCAGAAAGTCCGCACGATGAACGCGTGCA

TGGAATGAGCCCAGCAGAATATCTGGAGAAGTATGGTTTACTCGATGAACGTCTGCT

GGTCGCGCACTGTGTTTATCTCGATGACAAAGATATTGAACTTCTCGCACGCCATGA

TGTGAAAGTCGCACACTGCCCGGTTAGCAATGCATACCTGGGATCCGGCGTTGCGCC

GGTGCCGGAAATGGTTGAACGTGGTATCGCCGTCGGCATTGGCACAGATAACGGGG

CAAGCAACGACAGCGTTAATATGATCGAAGATATGAAATTCGCCGCGCACATCCAC

CGTGCCGTGCATCGTGATGCAGACGTCCTGACGCCTGAAAAAGTTCTGGAAATGGCC

ACTATTGATGGGGCGCGCGCTCTGGGCATGGAGGACGAGATCGGTTCCATCGAACC

GGGCAAGCGCGCCGATCTGATCCTGGTGGATCTGCGCCATCCGCAGACCACGCCGC

ATCATCATCTGGCCGCTACGATCGTGTTCCAAGCTTATGGTAACGAAGTTGATACGG

TCTTAATTGATGGGAATGTGGTGATGGAAAATCGCCGTCTGTCGTTCCTGCCTCCGG

AACGCGAGTTGGAATTTCTGGAAGAAGCTCAACGTCGCGCCACGGAAATCCTGCAA

CGTGCCAACATGGATGCTAACCCAGCGTGGCGTAGTCTCCTCGAGTGATAGATTCGA

GACTCGAATATAAGAGAGGCTAGGTGGAGGCTCAGTG

Chapter 2:

Comparison of the Impacts of Various Computational Protein Supercharging Methods on Protein Expression and Stability

Abstract

Supercharged proteins are of interest both for their unique stability properties and applications based on electrostatic interactions. However, the literature to date on protein supercharging is heavily skewed to studies on green fluorescent protein (GFP), with few other published successful use cases. A more comprehensive assessment of different computational methods for the design of supercharged proteins on a structurally diverse set of benchmarking proteins could be instrumental in determining the circumstances in which protein supercharging is most likely to succeed. In this work, we test two established and a newly developed supercharging method on a panel of five structurally diverse *de novo* designed proteins to characterize the effects of these methods on protein stability.

Introduction

Supercharged (SC) proteins have been demonstrated to have many interesting properties such as high thermostability, reversible denaturation, and the ability to penetrate cell membranes.[1–4] high surface charge also creates opportunities for interactions with materials or other macromolecules based on electrostatic interactions. However, despite the many potential uses of SC proteins, few successful use cases of this design approach exist in the literature and those that are published only cover a small portion of the structural diversity available across known protein sequences.[2,3,5,6]

Many SC variants of green fluorescent protein (GFP) have been designed and characterized to demonstrate their high stability.[2,3] Fusions of SC GFP variants with protein cargo have been demonstrated to penetrate mammalian cells.[4] SC GFP variants have also been shown to increase transfection of DNA and RNA cargo into mammalian cells.[7] Anti-MS scFv antibody fragments were supercharged, resulting in increased thermostability and binding affinity.[6] Cellulase domains were supercharged to prevent non-specific lignin binding, but the resultant proteins showed an unintended reduction in catalytic activity.[5] These represent most of the published cases of successful design of folded supercharged proteins. However, they do not show an effective sampling of the structural diversity available in known proteins. Notably, most of these proteins are composed of entirely or predominantly beta strand secondary structures, with the exception of the cellulase (PDB ID 4IM4).[2,3,5,6] A systematically chosen set of proteins with diverse secondary structure elements would more effectively evaluate the success of supercharging design methods to produce successfully folded protein.

Two main methods for the design of supercharged proteins have been developed: AvNAPSA (Average Neighboring Atoms Per Side Chain Atom) and Rosetta Supercharge. AvNAPSA uses an energy independent approach in which selected amino acid residues are ranked in order of surface exposure, determined by the number of atoms in other residues within 10 A, and deterministically mutated to lysine in positive mode and glutamate in negative (except in the case of asparagine residues which are mutated to aspartate).[2] Rosetta Supercharge expands the mutational possibilities by allowing mutation of neutral polar, hydrophobic, and oppositely charged residues to either lysine or arginine in positive mode and aspartate or glutamate in negative mode.[3] Here we present a third method, Supercharging Incorporating Neutralizing Mutations (ScIN). ScIN uses the residue selection of AvNAPSA then passes these potential

mutation sites to an expanded version of Rosetta Supercharge that includes mutations that neutralize oppositely charged residues to small nonpolar residues (Figure 2-1). These three methods were benchmarked on a set of structurally diverse proteins for their impact on soluble expression and thermostability.

A set of hyperstable *de novo* proteins was chosen for benchmarking of the various supercharging methods. Two proteins with only beta secondary structure (Protein DataBank (PDB) ID 6D0T and 6E5C), one with only alpha secondary structure (PDB ID 1P68), and two with mixed alpha/beta secondary structure (PDB ID 2LN3 and 2LV8) were chosen to represent a sampling of potential structural diversity available in known proteins. Each of these proteins has been shown to have high stability in the presence of chemical and thermal stressors.[8–11] Therefore, this study characterizes a set of proteins most likely to tolerate the potentially destabilizing effects of high net charge, and thus a set of proteins most likely to tolerate supercharging. Analysis of the products of each supercharging method on these hyperstable scaffolds could enable the identification of parameters that determine the likelihood of success for each supercharging method based on protein secondary structure.

Methods

Computational Design of Supercharged Proteins

AvNAPSA and Rosetta Supercharge designs for each scaffold were run using the default parameters in the Rosetta Online Server that Includes Everyone (ROSIE) server.[12] When the default parameters produced a protein with a net charge of less than +/- 0.1 per residue, the scaffold was run again with an additional parameter of target net charge set at the boundary of

+/- 0.1 net charge per residue to obtain designs with sufficient net charge for consideration. ScIN designs for each scaffold were produced using a PyRosetta script available at https://github.com/mpconnolly/supercharging/blob/main/ScIN_Supercharging.ipynb. Naming conventions for these designs will be as follows: PDB ID, design method (R= Rosetta Supercharge, A=AvNAPSA, S = ScIN), and charging mode (+ or -). As an example, the 1P68 scaffold charged in negative mode with Rosetta will be denoted 1P68 R+.

Cloning, Expression, and Characterization of Supercharged Proteins

The coding sequence for each protein was codon optimized and synthesized under control of the T7 promoter and terminator from pET29b+ in the pTwist Kan Medium Copy vector by Twist Biosciences. These sequences can be obtained in Supplementary Information S1. DNA encoding the 2LN3 scaffold designed with AvNAPSA in positive mode was not able to be synthesized and was omitted from further analysis. The resulting plasmids were transformed into *E. coli* BLR(λDE3) cells, cultured in 50 mL of Terrific Broth at 37^{o}C to OD 0.7 then induced with 1 mM IPTG. Cells were harvested, sonicated to lyse, clarified, and purified on Ni-NTA beads as previously described. Purified proteins were diluted 2-fold and quantified compared to a standard curve of bovine serum albumin using the Pierce BCA Assay Kit.

Results and Discussion

ScIN Designed Proteins are Distinct from Those Designed by Existing Methods

All design methods were able to design proteins with net charge greater than +/- 0.1 per residue for all scaffolds. AvNAPSA produced the highest net charge in both positive and

negative modes for all scaffolds. Rosetta Supercharge typically resulted in the lowest absolute net charges, with the net charge of ScIN designed proteins in between those of the previous methods, with the exception of 2LV8 in positive mode where the net charge resulting from ScIN and Rosetta Supercharge are equal (Table 2-1). This validates that ScIN is capable of designing proteins of comparable net charge to those produced by previous methods.

Soluble Expression of Supercharged Proteins (Preliminary Data)

Two rounds of expression and purification have been completed on the library of designed supercharged proteins. For three designs (1P68 A+, 1P68 S+, and 6D0T R+), no colonies were obtained in multiple attempts of transformation. Further investigation is needed to determine if this is the result of growth inhibition due to toxicity of the proteins. While a quantifiable amount of soluble protein can be measured in all samples of successfully transformed supercharged proteins, both the absolute and relative amounts of each protein vary substantially between rounds of purification (Figure 2-4A). This must be rectified to obtain a more reliable characterization of the performance of each supercharging method on each of the scaffolds.

However, when the maximum concentration of each protein is extracted from the data, some preliminary patterns start to emerge, though more robust results are needed before drawing definitive conclusions (Figure 2-4B). In all tested scaffolds, Rosetta Supercharge produced the highest amount of soluble protein in positive, though it should be noted that 6D0T R+ was not tested due to the absence of viable colonies discussed above. In negative mode, ScIN produced the highest amount of soluble protein in three of the scaffolds (1P68, 2LN3, and 6E5C) and Rosetta Supercharge produced the highest amount of soluble protein in the remaining two

scaffolds (2LV8 and 6D0T). Thus, for all scaffolds in both positive and negative modes the energy-dependent Rosetta Supercharge and ScIN produced higher amounts of soluble protein than the energy-independent AvNAPSA, suggesting the importance of energetic calculations in selection of mutations for supercharging. Rosetta Supercharge and ScIN also produce proteins with lower absolute net charges, reiterating the previous observation that soluble expression of supercharged proteins decreases at high net charge.[3] The low levels of soluble expression of the alpha helical protein 1P68 may suggest the importance of protein secondary structure elements in determining the success of supercharging efforts, with beta strands potentially more tolerant to supercharging.

The location and residue identities of mutations produced by ScIN were distinct from those produced by AvNAPSA or Rosetta Supercharge. Across all scaffolds, 28% of mutations in positive mode and 41% of mutations in negative were mutations of opposing charged residues to neutral residues, an option not available in the previous design methods (Figure 2-1). This is also reflected in the presence of a substantial amount of unique mutation sites selected by ScIN that were not selected by AvNAPSA or Rosetta Supercharge (Figure 2-2). Thus, the supercharged proteins designed are distinct from those designed by previous methods and ScIN represents an orthogonal supercharging strategy.

Figures

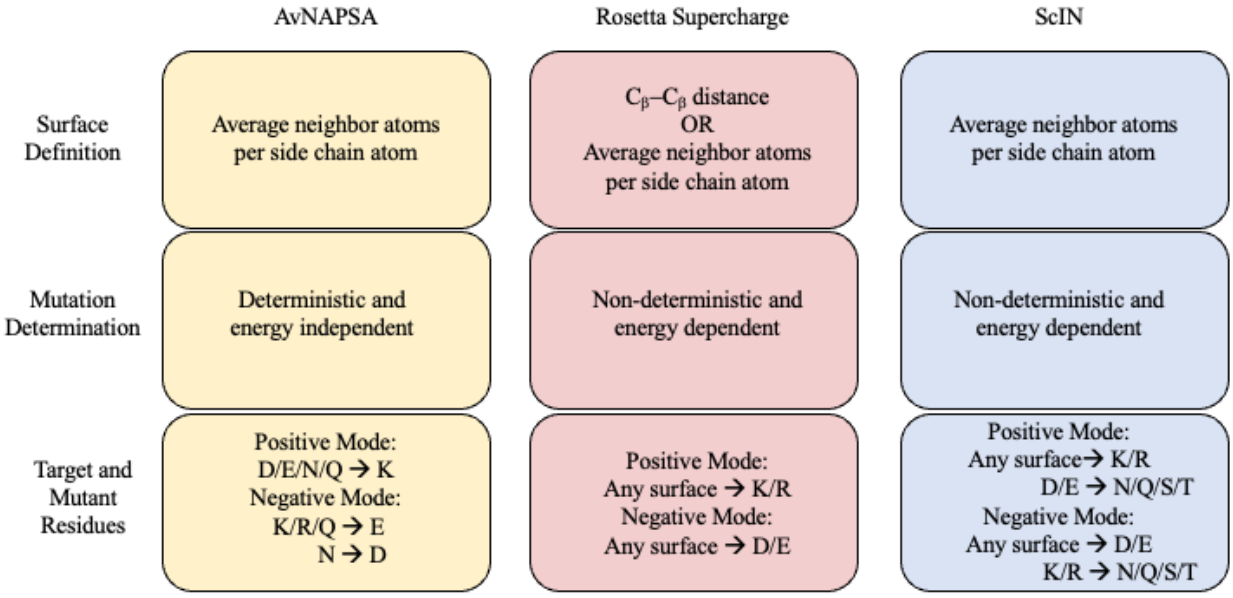


Figure 2-3. Summary of key features differentiating the new method ScIN from the previous methods, AvNAPSA and Rosetta Supercharge

Table 2-1. Net charges for *de novo* scaffolds designed with each of the three computational design methods evaluated.

Scaffold	Length	Native	Rosetta SC		AvNAPSA		ScIN	
			Pos	Neg	Pos	Neg	Pos	Neg
1P68	102	-9	13	-26	27	-37	36	-27
2LN3	83	-5	9	-17	23	-27	20	-18
2LV8	110	-3	29	-16	42	-38	29	-25
6D0T	111	-2	24	-17	32	-38	25	-19
6E5C	79	0	9	-14	24	-28	18	-17

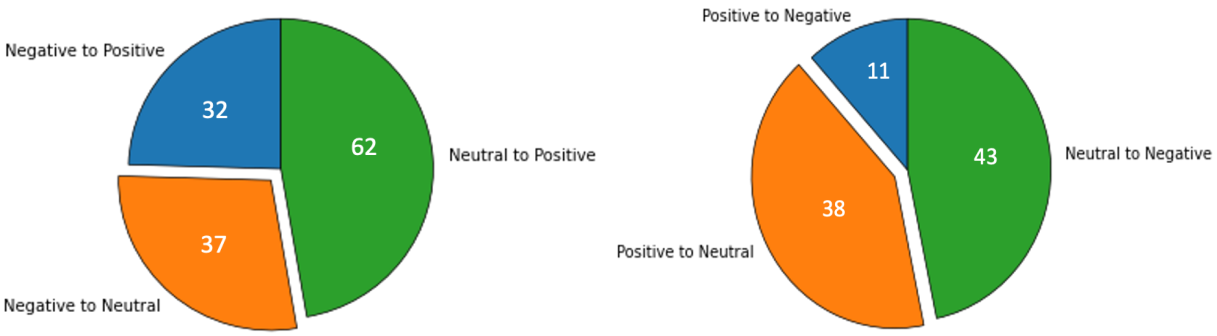


Figure 2-2. Categorization of parent and mutant residue types aggregated across the five selected protein scaffolds

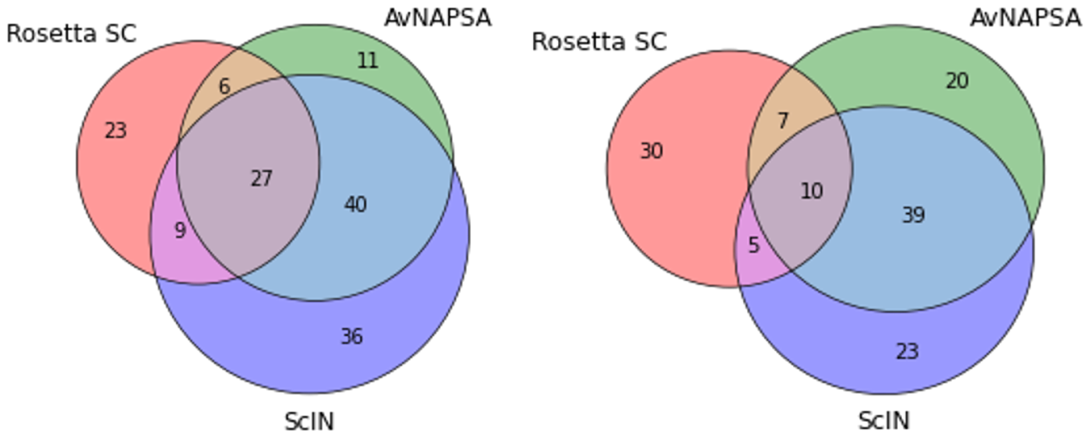


Figure 2-3. Comparison of mutation sites resulting from design of 5 scaffolds with AvNAPSA, Rosetta Supercharge, or ScIN aggregated across the five protein scaffolds

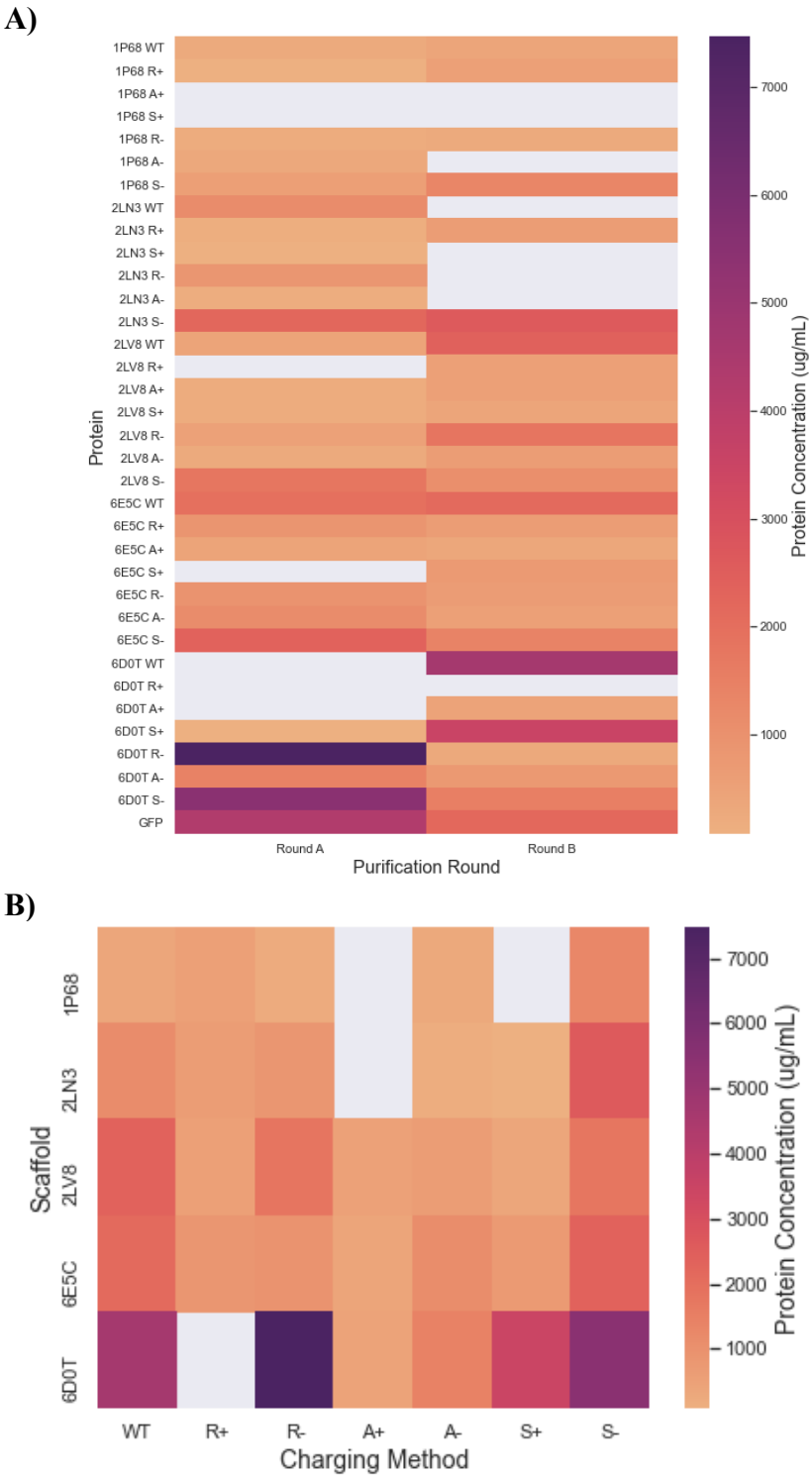


Figure 2-4. A) Heatmap of the soluble concentrations measured in each of two rounds of purification of each given protein. B) Maximum values extracted from 4A, representing the highest concentration achieved for each protein in each design strategy.

Conclusion

A novel and distinct method of designing supercharged proteins, Supercharging Incorporating Neutralizing Mutations (ScIN) was developed and validated. This method is undergoing benchmarking against the previously developed methods AvNAPSA and Rosetta Supercharge using systematically chosen, hyperstable *de novo* protein scaffolds. The resulting dataset could enable a more comprehensive understanding of the factors influencing the success of supercharging efforts and thus enable better access to the potential applications of supercharged proteins.

References

(1) Ma, C.; Malessa, A.; Boersma, A. J.; Liu, K.; Herrmann, A. Supercharged Proteins and Polypeptides. **2020**. https://doi.org/10.1002/adma.201905309.

(2) Lawrence, M. S.; Phillips, K. J.; Liu, D. R. Supercharging Proteins Can Impart Unusual Resilience. *J Am Chem Soc* **2007**, *129* (33), 10110–10112. https://doi.org/10.1021/JA071641Y/SUPPL_FILE/JA071641YSI20070628_045815.PDF.

(3) Der, B. S.; Kluwe, C.; Miklos, A. E.; Jacak, R.; Lyskov, S.; Gray, J. J.; Georgiou, G.; Ellington, A. D.; Kuhlman, B. Alternative Computational Protocols for Supercharging Protein Surfaces for Reversible Unfolding and Retention of Stability. *PLOS ONE* **2013**, *8* (5), e64363. https://doi.org/10.1371/JOURNAL.PONE.0064363.

(4) Cronican, J. J.; Thompson, D. B.; Beier, K. T.; Mcnaughton, B. R.; Cepko, C. L.; Liu, D. R. Potent Delivery of Functional Proteins into Mammalian Cells in Vitro and in Vivo Using a Supercharged Protein. **2010**. https://doi.org/10.1021/cb1001153.

(5) Whitehead, T. A.; Bandi, C. K.; Berger, M.; Park, J.; Chundawat, S. P. S. Negatively Supercharging Cellulases Render Them Lignin-Resistant. *ACS Sustainable Chemistry and Engineering* **2017**, *5* (7), 6247–6252. https://doi.org/10.1021/ACSSUSCHEMENG.7B01202/SUPPL_FILE/SC7B01202_SI_001.PDF.

(6) Miklos, A. E.; Kluwe, C.; Der, B. S.; Pai, S.; Sircar, A.; Hughes, R. A.; Berrondo, M.; Xu, J.; Codrea, V.; Buckley, P. E.; Calm, A. M.; Welsh, H. S.; Warner, C. R.; Zacharko, M. A.; Carney, J. P.; Gray, J. J.; Georgiou, G.; Kuhlman, B.; Ellington, A. D. Structure-Based Design of Supercharged, Highly Thermoresistant Antibodies. *Chemistry & Biology* **2012**, *19* (4), 449–455. https://doi.org/10.1016/J.CHEMBIOL.2012.01.018.

(7) McNaughton, B. R.; Cronican, J. J.; Thompson, D. B.; Liu, D. R. Mammalian Cell Penetration, SiRNA Transfection, and DNA Transfection by Supercharged Proteins. *Proc Natl Acad Sci U S A* **2009**, *106* (15), 6111–6116. https://doi.org/10.1073/PNAS.0807883106/SUPPL_FILE/APPENDIX_PDF.PDF.

(8) Wei, Y.; Kim, S.; Fela, D.; Baum, J.; Hecht, M. H. Solution Structure of a de Novo Protein from a Designed Combinatorial Library. *Proc Natl Acad Sci U S A* **2003**, *100* (23), 13270–13273. https://doi.org/10.1073/PNAS.1835644100/ASSET/D37C6874-CB12-454A-B50D-CCDDA11D7097/ASSETS/GRAPHIC/PQ2235644002.JPEG.

(9) Dou, J.; Vorobieva, A. A.; Sheffler, W.; Doyle, L. A.; Park, H.; Bick, M. J.; Mao, B.; Foight, G. W.; Lee, M. Y.; Gagnon, L. A.; Carter, L.; Sankaran, B.; Ovchinnikov, S.; Marcos, E.; Huang, P. S.; Vaughan, J. C.; Stoddard, B. L.; Baker, D. De Novo Design of a Fluorescence-Activating β-Barrel. *Nature* **2018**, *561* (7724), 485–491. https://doi.org/10.1038/S41586-018-0509-0.

(10) Marcos, E.; Chidyausiku, T. M.; McShan, A. C.; Evangelidis, T.; Nerli, S.; Carter, L.; Nivón, L. G.; Davis, A.; Oberdorfer, G.; Tripsianes, K.; Sgourakis, N. G.; Baker, D. De Novo Design of a

Non-Local β-Sheet Protein with High Stability and Accuracy. *Nat Struct Mol Biol* **2018**, *25* (11), 1028–1034. https://doi.org/10.1038/S41594-018-0141-6.

(11) Koga, N.; Tatsumi-Koga, R.; Liu, G.; Xiao, R.; Acton, T. B.; Montelione, G. T.; Baker, D. Principles for Designing Ideal Protein Structures. *Nature* **2012**, *491* (7423), 222–227. https://doi.org/10.1038/NATURE11600.

(12) Lyskov, S.; Chou, F. C.; Conchúir, S. Ó.; Der, B. S.; Drew, K.; Kuroda, D.; Xu, J.; Weitzner, B. D.; Renfrew, P. D.; Sripakdeevong, P.; Borgo, B.; Havranek, J. J.; Kuhlman, B.; Kortemme, T.; Bonneau, R.; Gray, J. J.; Das, R. Serverification of Molecular Modeling Applications: The Rosetta Online Server That Includes Everyone (ROSIE). *PLOS ONE* **2013**, *8* (5), e63906. https://doi.org/10.1371/JOURNAL.PONE.0063906.

Supplementary Information

Figure S1. FASTA sequences of DNA inserts cloned into pTwist Kan Medium Copy for use in this study

>1P68_A-

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTTT
AACTTTAAGAAGGAGATATACATATGTACGGTGAGCTTAACGACCTGCTGGAAGATTTGCAGGA
GGTCCTTGAAAACCTTCACGAAAATTGGCATGGTGGTGAGGATGACCTCCATGATGTCGATGACC
ATTTAGAAGACGTTATTGAGGATATCCATGATTTTATGGAAGGCGGCGGCAGCGGTGGCGAACTC
GAAGAAATGATGGAAGAGTTTCAGGAAGTGCTGGATGAACTGAACGACCACCTTGAGGGTGGCG
AACATACCGTGCATCACATTGAGGAAAACATTGAGGAAATCTTTCATCACCTGGAAGAACTGGTG
CATGAGCTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAG
CTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTC
TTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

>1P68_A+

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTTT
AACTTTAAGAAGGAGATATACATATGTATGGCAAATTAAATAAACTCTTGAAAAAGTTGCAGAA
GGTGCTGAAGAACTTACACAAGAATTGGCATGGCGGCAAGAAGAAACTGCATAAGGTGGACAAA
CACCTGAAAAAAGTGATTAAGGACATCCACAAGTTCATGAAAGGCGGCGGCAGCGGTGGGAAGT
TGAAAAAAATGATGAAAGAATTTCAAAAGGTCCTGAAGGAACTGAATAAACATCTGAAAGGAGG
AAAACATACTGTGCACCATATTGAAAAAAATATCAAAAAAATCTTTCACCATCTGAAAAAGCTTG
TACATCGCCTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGA
AGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGG
TCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

>1P68_S-

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTTT

AACTTTAAGAAGGAGATATACATATGTATGGGGAACTGAATGAGTTGCTGGAAGAACTCCAGGA

AGAACTGAGCGAACTGCATAACAACTGGCATGGGGACTCGGATGAACTCCATGATGTGGATAAC

CATCTGCAAGAAGTGATTGAAGATATCCATGATTTTATGCAAGGCGACGGTAGTGGGGGGAAAC

TGCAAGAGATGATGTCGGAGTTCGAACAGGTGCTGGAGGAGCTGAACAACCACTTAGATGGCGG

CGAGGAGACGGTTCACCATATCGAACAGAATATTAACGAAATTTTTCACCATCTGGAGGAACTGG

TGCATCGTCTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGA

AGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGG

TCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

>1P68_S+

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTTT

AACTTTAAGAAGGAGATATACATATGTATGGTAAATTGAAAAAGCTGTTGAAAAAATTGCAAAA

AGTCCTCAAGAAATTACATAAGAAATGGCATGGAGGCAAAAAAAACCTTAAGAAAGTTGACAAA

CACTTGCAAAAAGTTATTAAAGATATCAAAAAATTCATGCAGGGGAAAGGCTCGGGCGGCAAGC

TGAAGAAAATGATGAAAAAGTTCCAGAAAGTTCTCAAGGAACTGAAGAAACACCTTAAAGGTGG

GAAAAAGACGGTTCGCCACATCAAACAAAATATTAAAGAAATCTTCAAACATCTGAAAAAATTA

GTGAAACGTCTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGG

AAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGG

GTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

>1P68_R-

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTTT

AACTTTAAGAAGGAGATATACATATGGATGGCAAACTTGACGATCTGTTAGAAGATTTAGAAGA

AGTGGATAAAAACTTGGAAAAAAACGATGAAGGTGGCAAAGACAACCTGCATGACGTTGACGA

AGATCTGCAAAACGTTATTGAGGACATCCACGATGAAGATCAGGGCGG

TGGTTCCGGCGGTAAACTGCAGGAAATGATGAAAGAATTCCAGCAGGTCGATGATGAATTGAAT

AATCATCTTCAAGGCGGTAAAGACACCGTTCATGATATTGAGGACGAAATTAAGGAGGAATTTC

ATCACCTGGAAGAACTGGTACACCGTCTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCT

AACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCC

TTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

>1P68_R+

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTTT

AACTTTAAGAAGGAGATATACATATGAAAGGCAAACTTCGTGATCTGTTAGAGGATTTGCGTGAG

GTGCTTAAGAATCTGAAGAAAAATTGGCGCGGCGGGAAAGATCGTTTGCACGATGTAGACCGTC

GCCTGCAGAATGTGATCGAAAAAATCCATAAAAAGATGCAGGGCGGTGGTTCCGGTGGCAAGTT

GCAAGAAATGAAGAAAGAGTTCCAAAAGGTGCGTGACGAATTGAACAACCATCTTCAGGGCGGG

AAAAAAACCGTGCATCGCATCGAACAACGTATTAAACGCCGTTTCCACCATCTGGAAGAACTCGT

ACATCGTCTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAA

GCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGT

CTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

>1P68_WT

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTTT

AACTTTAAGAAGGAGATATACATATGTACGGGAAACTCAATGATCTGCTGGAAGACTTGCAGGA

AGTGCTGAAAAACCTTCATAAAAACTGGCATGGCGGTAAAGATAACTTACATGATGTCGACAAC

CACCTTCAGAACGTAATCGAAGACATCCATGATTTTATGCAAGGGGGCGGCTCCGGTGGTAAACT

CCAAGAAATGATGAAGGAGTTCCAACAGGTGCTGGACGAACTGAACAACCATCTTCAGGGGGGC

AAGCACACGGTGCATCATATCGAGCAAAACATTAAAGAAATCTTTCACCACCTGGAAGAATTGG

TCCACCGCCTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGA

AGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGG

TCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

>2LN3_AV-

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTTT

AACTTTAAGAAGGAGATATACATATGGGCTTGACCGAAACTATCACGTCTGAAGATAAAGAAGA

ATTGCTGGAAATCGCCCTCGAATTCATCAGCGAGGGCCTGGACCTGGAGGTGGAGTTTGATAGCA

CGGATGATGAAGAAATTGAAGAATTTGAAGAAGATATGGAAGATCTGGCGGAAGAAACGGGCG

TGGAAATTGAAAAAGAGTGGGAGGGCGATAAGCTCCGCATTCGTCTTGAGGGCTCCCTCGAACA

CCACCATCACCACCACCTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCC

GAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCT

AAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

>2LN3_AV+

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTTT

AACTTTAAGAAGGAGATATACATATGGGTTTAACCCGCACAATCACGAGCAAGAAGAAGAAGAA

GCTCCTGAAAATCGCATTGAAGTTCATTAGCAAAGGGCTGGATCTTGAAGTCGAGTTCAAGTCGA

CGAAGAAGAAGGAAATTAAGAAGTTTGAGCGTGATATGGAGAAGTTGGCAAAGAAGACCGGAG

TCAAGATCAAGAAGAAGTGGAAGGGGAAGAAGCTGCGCATTCGCCTTAAGGGCAGCTTAAAGCA

CCACCACCACCACCACCTTGAACATCATCATCATCATCATTAAGATCCGGCTGCTAACAAAGCCC

GAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCT

AAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

>2LN3_S-

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTTT

AACTTTAAGAAGGAGATATACATATGGGTGAGACCACGACCATTACCAGTCAAGATAAGGAAGA

GCTGCTGGAGATTGCGCTGCAATTTATCTCGCAGGGATTAGATCTTGAAGTTGAATTCGATAGCA

CCGATGAGACCGAAATCGAAGAATTCGAGCAGGATATGGAAGAGCTGGCGACGCAGACAGGCG

TGGAAATCCAGCAGCAGTGGCAGGGTAACACGCTGCGTATCCGTCTGAAAGGAAGCCTGGACCA

CCATCACGATGAGCACCTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCC

CGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTC

TAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

>2LN3_S+

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTTT

AACTTTAAGAAGGAGATATACATATGCGCCAGACCCGCGTTACAGTTAAACCGGGGCGTACCTA

CCAAGTAAAGGTGAAACCCGGAAAACGCGTGGAGATTCAGGCGAAAGGCCCGGCGGAATTCCA

GGGTGGCGGTACCAAAACCCGCCTGAAACCCGGACAGAGCTATAAATTTAAAAATAAAACGTCA

CAACCGTTGAAAATCAAATTGCGGAACCTGTCCAAAAAACCGATTACCTTTCGTATTAAAGAAGA

GCTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAG

TTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAG

GGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

>2LN3_R-

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTTT

AACTTTAAGAAGGAGATATACATATGGGTTTGGACCGCACCATTACCTCAGATAATAAAGAAGA

ACTGCTGGAAATCGCCGAAAAATTCATTGAACAGGGGCTGGATCTGGAGGTTGAGTTCGATTCGG

ATGATGATAAAGAAGAAGAGGAATTTGAACGTGACATGGAAGACCTCGCTAAAAAAACCGGTGT

GCAAATTGATAAACAATGGCAAGGCAATGATCTCCGTATCCGGCTGAAAGGTGATGACGAAGAT

CACCACCATCATCACCTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCG

AAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTA

AACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

>2LN3_R+

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTTTGTTT

AACTTTAAGAAGGAGATATACATATGGGTTTGACCCGCACCATTAAGAGTCAGAATAAGGAAGA

ATTATTGGAGATCGCAAAAAAATTTATCAAAAAAGGGCTGGACCTCGAGGTGGAATTCCGTAGT

ACCGACGATAAGAAAATTGAAGAATTTGAGCGTAAAATGGAAGATCTGGCCAAGAAAACCGGTC

GTCAAATCCAGAAACAGTGGCAGGGTAATAAATTGCGTATCCGCCTGAAAGGCTCGAAAAAACA

TCATCACCATCATCATCTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCC

GAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGA

GCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAA

AGGAGGAACTATATCCGGAT

>2LN3_WT

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTT

TGTTTAACTTTAAGAAGGAGATATACATATGGGCCTTACACGTACCATTACGAGCCAGAA

CAAAGAGGAGCTGTTAGAAATCGCCCTGAAATTTATCTCTCAGGGTCTTGATTTAGAAGT

TGAATTTGATAGCACCGACGACAAAGAAATTGAGGAGTTCGAGCGCGATATGGAAGACTT

AGCGAAAAAAACCGGCGTGCAAATTCAGAAACAGTGGCAGGGCAATAAATTGCGCATCCG

TTTGAAAGGCAGTCTGGAACATCATCATCATCATCACCTCGAGCACCACCACCACCACCA

CTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGA

GCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAA

AGGAGGAACTATATCCGGAT

>2LV8_A-

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTT

TGTTTAACTTTAAGAAGGAGATATACATATGCTGCTCTACGTCCTGATTATTAGTGATGA

TGAGGAACTTATCGAGGAAGCGGAAGAAATGGCCGAAGAAGCCGATCTGGAATTAGAGAC

TGTTGAAACCGAGGATGAACTGGAGGAATATCTGGAAGAATTCGAAGAAGAATCAGAAGA

TATTAAGGTGCTGATCTTGGTTTCTGATGATGAAGAACTGGATAAAGCGAAAGAATTAGC

GCAGGAAATGGAAATTGATGTTCGCACACGCGAGGTAACCAGCCCGGACGAAGCCAAGGA

ATGGATCGAGGAATTCAGCGAAGAAGGGGGTAGCCTGGAACATCACCATCATCATCATCT

CGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGA

GTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGT

CTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

>2LV8_A+

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTT

TGTTTAACTTTAAGAAGGAGATATACATATGCTGCTTTACGTATTGATTATTAGCAAAAA

AAAGAAACTCATTAAAAAAGCACGTAAGATGGCGAAAAAAGCTAAACTCAAACTGCGCAC

GGTGAAAACTAAGAAAAAACTTAAGAAGTATCTCAAAAAATTTCGCAAAAAATCGAAAAA

AATTAAAGTGCTGATTCTGGTTTCGAAAAAAAAAGAACTGAAAAAAGCTAAAAAGTTAGC

CCAGAAGATGAAAATCGACGTCCGTACTCGTAAAGTCACGTCGCCGAAAAAAGCCAAGCG

TTGGATCAAAGAATTTAGCAAGAAAGGGGGTTCGCTGGAACATCACCACCACCACCACCT

CGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGA

GTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGT

CTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

>2LV8_S-

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTT

TGTTTAACTTTAAGAAGGAGATATACATATGCTGCTGTATGTGCTCATCATTAGTAATGA

CGAAGATCTGATTGATGAAGCACGTGAAATGGCCGAGCAGGCGAATCTGGACTTGCGCAC

AGTTACCACCGAAGAGGAGTTAAAAACGTACCTGGAAGAATTTCAGAACGAAAGCGATGA

TATTAAGGTTCTCATCCTGGTCAGTGAAGATGAAGAATTAGAAAAGGCAAAAGAACTTGC

CCAGCAAATGGAAATTGACGTTCGGACGCGGCAAGTCACTGATCCAGAAGAGGCTAAAAC

ATGGATCAAAGAATTCTCTGAGGAAGGGGGCTCTGAAGAACACGACGATCACGATCACCT

CGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGA

GTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGT

CTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

>2LV8_S+

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTT

TGTTTAACTTTAAGAAGGAGATATACATATGTTGCTTTACGTGCTGATTATTTCGAATGA

TAAAAAACTGATTAAGGAAGCACGCAAAATGGCAAAAAAAGCAAACCTGCAGCTGCGCAC

TGTGCGGACCAAAAAACAATTGAAAAAATATCTGAAACAGTTTAAGAAAAATAAACGCAA

CATTAAGGTCCTGATTCTGGTATCCCGGAATAAAGAACTGAAAAAGGCTAAAACCCTGGC

TCAGCGCATGAATATCGATGTGCGTACTCGCAAGGTCACAAGCCCAAATGAAGCCAAACG

CTGGATTAAACAGTTTAGTCAGCAGGGTGGATCCAAACAGCATCACCACAAGAAGCATCT

CGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGA

GTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGT

CTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

>2LV8_R-

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTT

TGTTTAACTTTAAGAAGGAGATATACATATGCTGCTGTACGTTTTAATCATTAGCGACGA

CAAAGACCTGATCGAGGAGGCGCGTAAGATGGCGGAGAAAGCAAACCTGGAGCTGCGTAC

CGTCAAAACGGAGGATGAACTCAAAAAGTACTTGGAGGAGTTCGAAGAAGAAGACGATAA

TATCAAGGTACTCATTCTGGTCAGTAATGATGAAGAACTTGACAAAGCAAAAGAACTGGC

GCAAAAGATGGAAATCGACGTGCGCACACGCAAAGTGACCAGCCCGGATGAAGCAAAACG

CTGGATTAAAGAATTCTCAGAAGAGGGTGGCTCAGAAGAACACGAAGACGACCACCATCT

CGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGA

GTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGT

CTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

>2LV8_R+

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTT

TGTTTAACTTTAAGAAGGAGATATACATATGTTGCTGTATGTTTTAATCATTTCGAAAAA

GAAGAAACTGATCGAAGAAGCGCGTAAAATGGCAGAGAAAGCGAACTTAGAGCTGCGTAC

GGTCAAAACCGAGAAAGAACTGAAAAAATACTTGAAGAAATTCCGCAAACGGCGTAAAAA

CATCAAGGTGTTGATCCTGGTGTCGAAAAAGAAGGAACTGAAGAAAGCAAAAAAACTGGC

GCAGAAAATGAAAATTGACGTGCGCACCCGTAAAGTGACGAAACCTGACAAAGCTAAACG

TTGGATCAAGGAATTTAGTGAAAAAGGCGGGTCGAAGGAGCATAAAAAACATCATCACCT

CGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGA

GTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGT

CTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

>2LV8_WT

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTT

TGTTTAACTTTAAGAAGGAGATATACATATGCTGTTGTATGTTCTGATTATTAGCAACGA

TAAGAAACTGATCGAAGAAGCGCGCAAGATGGCCGAAAAAGCCAACCTGGAACTGCGCAC

CGTCAAAACCGAAGATGAGCTGAAAAAATATCTGGAAGAGTTTCGTAAGGAGTCACAAAA

CATCAAGGTGCTGATCCTGGTCAGTAACGATGAGGAACTGGATAAAGCGAAAGAACTGGC

GCAGAAAATGGAAATCGATGTTCGCACCCGGAAAGTTACTAGCCCCGACGAGGCTAAACG

CTGGATTAAAGAATTTTCCGAGGAAGGTGGGAGCCTGGAACATCATCATCACCACCACCT

CGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGA

GTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGT

CTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

>6D0T_A-

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTT

TGTTTAACTTTAAGAAGGAGATATACATATGGTTGATGCGGCGGAATATTTTCCTGGTAC

CTGGGAGTTTGAATTCGAAAGCAGTGACGGAGAAGAGTATGAAGGTACAGTGGAAATGGA

ACCTGAGACCCCGACCGAGATCGAAATTGAATTTGAAGGGGAGTCTAGTGACGGTGAACC

GGTGGAAGGCGAAGGCTCTATCGAAGTAGAATCTCCCTACGAGTACGAATTTGAAATGGA

ATCGAGCGATGGAGCGGAGTGGGAGGGTACCCTGGAAGTTGAGTCCCCGGATTCTGTGGA

AGTTGAATTTGAGTCCAGTGATGGCCGCGAATATTCCGGCGAATTCCGCCGTGAAGAGGG

TCTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGC

TGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACG

GGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

>6D0T_A+

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTT

TGTTTAACTTTAAGAAGGAGATATACATATGGTTAAAGCGGCGAAATACTTCCCTGGAAC

GTGGAAGTTCCGTTTCCGCAGCTCTAAAGGCAAGGAATACCGGGGGACCGTCAAAATGAA

GCCCCGTACGCCGACCAAAATTGAAATCCGGTTTAAAGGTAAATCCAGCAAAGGTCGCCC

GGTGAAAGGCCGTGGCAGTATTGAAGTACGTTCCCCGTATAAGTACCGTTTCGAAATGAA

ATCTAGTGATGGAGCTCGGTGGAAGGGGACGCTTAAAGTCCGTTCGCCCAAATCCGTCAA

AGTGCGTTTCAAAAGCTCAAAGGGACGCAAATATAGTGGCGAATTTCGCCGCAAAAAGGG

GCTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGC

TGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACG

GGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

>6D0T_S-

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTT

TGTTTAACTTTAAGAAGGAGATATACATATGGTTGACGCGGCCCAATACTTTGAAGGCAC

CTGGGAATTCCGTTTTCGTTCCTCAGACGGCAAGGAATACGAGGGTACGGTGGAAATGCA

ACCGACGACCCCCACGGAAATCGAAATTCAGTTCCAGGGGCAGTCCAGCGACGGGGAACC

GGTAGAGGGCTCAGGATCGATCGAAGTCACCAGTCCAGAAGAGTACCGTTTCGAAATGCA

GTCTAGTGACGGCGCGACGTGGGAGGGCACGCTGCAGGTCCAATCACCGGAGAGCGTCGA

GGTCCAATTCGAATCATCGGATGGACGCGAATATAGCGGCGAGTTTCAACGTCAGGAAGG

CCTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGC

TGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACG

GGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

>6D0T_S+

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTT

TGTTTAACTTTAAGAAGGAGATATACATATGGTCAACGCGGCGAAATATTTCAAAGGCAC

CTGGAGCTTTCGTTTCCGGAGCAAACAAGGACGCCAATATAAAGGCACCGTTGAGATGCG

TCCTAAAACCCCCACACAGATTGAAATTCGTTTCAAGGGTAAGTCGAGCAGCGGGAAACC

GGTTACCGGTCGCGGCTCCATCGAAGTGCGCAGCCCTAAGCAGTATCGGTTTAAAATGCA

GAGCAGCCAAGGGGCGAAATGGAAAGGAACGCTGCAGGTCCGCAGCCCGAAAAAAGTACA

GGTTAAATTCAAATCCAGCACCGGTCGCACGTATAGCGGAGAATTCAAACGTCAGAATAA

ACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGC

TGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACG

GGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

>6D0T_R-

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTT

TGTTTAACTTTAAGAAGGAGATATACATATGGTGGACGCAGCAGAATATTTTCCGGGTAC

CTGGGAATTCCGCTTTCGTAGTGAGGACGGCAAAGAGTACCGCGGGACGGTTGAGATGGA

ACCTGAGACGCCGACCGAAATTGAAATTCGGTTCGAAGGCGAAGACTCTGATGGCCGTCC

AGTCGAAGGAGAAGGAAGCATCGAGGACCGCAGTCCGGACGAATATCGCTTTGAAATGGA

AAGCAGTGATGGGGCCCGCTGGGAGGGTACTCTGCAAGTCCGCAGTCCGGACAGCGTGGA

AGTGCGGTTCAAAGAAAGCGATGGCCGTGAATATTCTGGTGAATTTCGTCGGCAGGAAGG

CCTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGC

TGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACG

GGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

>6D0T_R+

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTT

TGTTTAACTTTAAGAAGGAGATATACATATGGTCAAAGCCGCGAAATATTTTCCCGGCAC

CTGGGAGTTCCGCTTTCGTTCAAAAAAAGGGAAACGCTATCGCGGCACCGTAGAAATGCG

TCCACGTCGGCCGACCGAAATCGAGATTCGCTTCAAAGGTCGTCGCTCACGGGGTCGCCC

GGTAGAAGGCCGCGGATCGATCGAAAAGCGCTCGCCGCGGGAATATCGCTTTCGCATGCG

TAGTTCCGATGGTCGTCGCTGGCGCGGTACTTTGCAAGTTCGGAGTCCGCGTTCGGTGCG

CGTTCGGTTCAAAAGCTCCGACGGGCGGGAATACAGCGGTGAATTCCGTCGTCAAGAAGG

CCTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGC

TGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACG

GGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

>6D0T_WT

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTT

TGTTTAACTTTAAGAAGGAGATATACATATGGTTGACGCTGCGCAGTATTTCCCGGGAAC

GTGGGAATTCCGTTTTCGGAGTAGCGATGGCAAAGAATACCGTGGCACGGTGGAAATGCA

GCCCCGCACGCCAACGGAAATTGAGATCCGTTTTAAGGGCCAATCGAGTGACGGCCGTCC

TGTTGAGGGTCGCGGGTCTATTGAAGTGCGTAGTCCGTACGAGTATCGCTTTGAAATGCA

GAGCTCGGACGGGGCGCGCTGGGAAGGTACGCTGCAAGTTCGTTCGCCGGATTCCGTCGA

AGTCCGCTTTAAAAGCAGCGATGGCCGCGAATATAGCGGGGAATTCCGTCGCCAAGAGGG

CCTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGC

TGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACG

GGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

>6E5C_A-

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTT

TGTTTAACTTTAAGAAGGAGATATACATATGACCGAGGAAACAGAAGTCACGGTCGATCC

AGGGGAAGAATACGAAGTTGAAGTAGACCCGGGCACCGAAGTCGAGATTCAGGCGGAAGG

TCCTGCAGAGTTTGAGGGTGGTGGCACCGAAACTGAGCTGGATCCGGGCGAATCGTACGA

ATTCGAGAACTTAACCTCAGAGCCGTTGGAAATTGAACTCCGGAACCTGTCTGATACTCC

GATTGAGTTTGAGATCGAAGAAGAACTCGAGCACCACCACCACCACCACTGAGATCCGGC

TGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGC

ATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTAT

ATCCGGAT

>6E5C_A+

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTT

TGTTTAACTTTAAGAAGGAGATATACATATGACCCGCAAAACCAAAGTGACCGTAAAACC

GGGCAAGAAATATAAGGTCAAAGTAAAACCCGGAACACGTGTGGAAATCCAGGCGAAAGG

GCCGGCGGAATTCAAAGGAGGGGGGACGCGTACACGTCTGAAGCCGGGCAAAAGCTATAA

ATTTAAGAACCTGACGTCAAAGCCGCTTCGCATCCGCCTCCGCAACCTCTCCAAAACGCC

GATCAAGTTCCGCATTCGCGAGAAGCTCGAGCACCACCACCACCACCACTGAGATCCGGC

TGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGC

ATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTAT

ATCCGGAT

>6E5C_S-

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTT

TGTTTAACTTTAAGAAGGAGATATACATATGACGCAGGAAACCAGCGTTACAGTTGAACC

CGGCGATGAATACGAAGTTGAGGTAGAACCAGGTACCCAGGTGGAAATTCAGGCAAAGGG

CCCGGCCGAATTCGAGGGTGGTGGGACCACCGACCAATTAAATCCGGGTGAAAGTTACAC

GTTTGAAAATTTAACGGATGAACCGCTGACCATTACATTACGCAATCTTTCCGAAACTCC

GATTGAGTTCACCATCACTGAAGACCTCGAGCACCACCACCACCACCACTGAGATCCGGC

TGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGC

ATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTAT

ATCCGGAT

>6E5C_S+

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTT

TGTTTAACTTTAAGAAGGAGATATACATATGACGCGTCAAACACGCGTTACTGTCAAACC

TGGGCGTACTTATCAGGTGAAAGTCAAGCCGGGCAAACGTGTAGAGATTCAGGCAAAAGG

GCCAGCCGAATTTCAGGGCGGCGGGACAAAAACCCGTCTGAAACCTGGGCAAAGCTATAA

ATTTAAAAATAAAACCTCCCAGCCGTTGAAGATTAAATTACGCAATTTATCCAAGAAACC

TATTACCTTCCGCATTAAAGAGGAACTCGAGCACCACCACCACCACCACTGAGATCCGGC

TGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGC

ATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTAT

ATCCGGAT

>6E5C_R-

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTT

TGTTTAACTTTAAGAAGGAGATATACATATGGAACGTGAAACCAAAGTCGAGGTAGAACC

TGGTGAGGAGTACGAGGTTAAAGAGGAGCCGGGGACCCGCGTTGAGATTCAGGCGAAAGG

CCCGGCAGAGTTCGAGGGTGGTGGTGAACGTACCCGTGAAAACCCAGGTGAATCATATGA

AGAAGAGAATCTTACTAGCCAACCTCTGCGCGATCGTGAACGTAATCTGTCAGATGAGCC

CGAGGAGTTTCGGATTCGTGAAGAACTCGAGCACCACCACCACCACCACTGAGATCCGGC

TGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGC

ATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTAT

ATCCGGAT

>6E5C_R+

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTT

TGTTTAACTTTAAGAAGGAGATATACATATGCGTCGTGAAACTAAAGTGAAAGTTAAGCC

CGGCAAAGAAAAGGAAGTTAAAGTAAAACCGGGTACACGCGTCGAGATTCAGGCAAAGGG

ACCGGCCGAGTTTGAGGGTGGCGGTAAACGCACCCGCTTGAACCCGGGCGAAAGCTACAA

GTTCGAAAATCTGACCTCGCAGCCATTGCGCATCCGTCTCCGTAACCTCTCCGATAAGCC

GATTGAATTCCGCATCCGTGAAGAGCTCGAGCACCACCACCACCACCACTGAGATCCGGC

TGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGC

ATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTAT

ATCCGGAT

>6E5C_WT

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTT

TGTTTAACTTTAAGAAGGAGATATACATATGACCCGCGAAACCAAGGTGACAGTGAATCC

GGGTGAAGAGTATGAAGTTAAAGTAAACCCGGGTACCCGTGTGGAAATTCAGGCTAAGGG

CCCTGCTGAGTTTGAAGGCGGTGGAACGCGTACCCGTCTTAACCCGGGCGAATCTTATAA

ATTTGAAAACCTGACTTCCCAGCCACTGCGTATTCGTTTACGCAACCTGTCGGATACGCC

AATTGAATTTCGTATTCGTGAAGAGCTCGAGCACCACCACCACCACCACTGAGATCCGGC

TGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCAATAACTAGC

ATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTAT

ATCCGGAT

>GFP

TAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGAAATAATTT

TGTTTAACTTTAAGAAGGAGATATACATATGGGTGGCAGCAAAGGTGAAGAACTGTTTAC

CGGTGTTGTTCCGATTCTGGTTGAACTGGATGGTGATGTTAATGGCCACAAATTTTCAGT

TCGTGGTGAAGGCGAAGGTGATGCAACCAATGGTAAACTGACCCTGAAATTTATCTGTAC

CACCGGCAAACTGCCGGTTCCGTGGCCGACCCTGGTTACCACCCTGACCTATGGTGTTCA

GTGTTTTAGCCGTTATCCGGATCACATGAAACGCCACGATTTTTTCAAAAGCGCAATGCC

GGAAGGTTATGTTCAAGAACGTACCATCTCCTTTAAAGATGATGGCACCTATAAAACCCG

TGCCGAAGTTAAATTTGAAGGTGATACCCTGGTGAATCGCATTGAACTGAAAGGCATCGA

TTTCAAAGAAGATGGTAATATCCTGGGCCACAAACTGGAATATAATTTCAATAGCCACAA

CGTGTATATCACCGCAGACAAACAGAAAAATGGCATCAAAGCCAACTTTAAAATCCGGCA

TAATGTTGAAGATGGCAGCGTTCAGCTGGCAGATCATTATCAGCAGAATACCCCGATTGG

TGATGGTCCGGTTCTGCTGCCGGATAATCATTATCTGAGCACCCAGAGCGTTCTGAGCAA

AGATCCGAATGAAAAACGTGATCACATGGTGCTGCTGGAATTTGTTACCGCAGCAGGTAT

TACCCATGGTATGGATGAACTGTATAAAGGTAGCCTCGAGCACCACCACCACCACCACTG

AGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCTGCTGCCACCGCTGAGCA

ATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGG

AGGAACTATATCCGGAT

Chapter 3:

Evaluating Protein Engineering Thermostability Prediction Tools Using an Independently Generated Dataset

Abstract

Engineering proteins to enhance thermal stability is a widely utilized approach for creating industrially relevant biocatalysts. The development of new experimental datasets and computational tools to guide these engineering efforts remains an active area of research. Thus, to complement the previously reported measures of T_{50} and kinetic constants, we are reporting an expansion of our previously published dataset of mutants for β-glucosidase to include both measures of T_M and $\Delta\Delta G$. For a set of 51 mutants, we found that T_{50} and T_M are moderately correlated, with a Pearson correlation coefficient and Spearman's rank coefficient of 0.58 and 0.47, respectively, indicating that the two methods capture different physical features. The performance of predicted stability using nine computational tools was also evaluated on the dataset of 51 mutants, none of which are found to be strong predictors of the observed changes in T_{50}, T_M, or $\Delta\Delta G$. Furthermore, the ability of the nine algorithms to predict the production of isolatable soluble protein was examined, which revealed that Rosetta $\Delta\Delta G$, FoldX, DeepDDG, PoPMuSiC, and SDM were capable of predicting if a mutant could be produced and isolated as a soluble protein. These results further highlight the need for new algorithms for predicting modest, yet important, changes in thermal stability as well as a new utility for current algorithms for prescreening designs for the production of mutants that maintain fold and soluble production properties.

Introduction

A common goal of enzyme engineering is the enhancement of thermal stability [1]. For industrial applications, improving a protein's robustness to thermal challenges or half-life at elevated temperatures can often be the deciding factor for the commercialization of a biocatalyst [2−5]. Currently, the most common approach for improving thermal stability is through directed evolution methodologies [6,7], which can be time consuming, costly, and limited in the ability to search sequence space. Computational design tools to predictably identify single and combinatorial mutations that enhance thermal stability are rapidly developing and growing in popularity [8−14]. However, accurate predictions using computational tools to guide protein stability design remain an active area of research and is not always successful. The use of large datasets on the mutational effect on protein stability, such as ProTherm [15] now maintained by ProtaBank [16], is often used to train computational methods for predicting thermal stability. The datasets utilized generally consist of the equilibrium constant of unfolding (K_u) or the melting temperature of an enzyme (T_M) [17]. In our previous study, we determined the thermal stability of 79 β-glucosidase B (BglB) variants by finding T50, a type of kinetic stability that is determined by the temperature at which a mutant's residual activity is reduced by 50% after a heat challenge over a defined time [4,17,18]. When analyzing this set of mutants using two established computational programs (Rosetta ΔΔG and FoldX PSSM) for predicting thermal stability, we found that there was no significant correlation between the predictions and the observed T50 [19].

One hypothesis explaining the poor predictive performance of the algorithms with the BglB dataset is that the algorithms are evaluated on T_M, a direct measure of structural thermal stability. However, the algorithms were being used to predict T_{50}, which is an indirect measure of

the protein's thermal stability [17]. Alternatively, the poor performance could have come from the narrow T_{50} range (extreme variants are +6.06 and −5.02 °C from the wild type (WT)) as the algorithms are generally benchmarked on larger changes in thermal stability and ±3 °C may be within the error of the currently developed algorithms. In this study, we evaluated both hypotheses. To assess if there was a significant difference in T_M and T_{50}, we developed a dataset of 51 BglB mutants (Figure 1) in which both thermal stability measurements, T_{50} and T_M, were measured. Interestingly, for the set of 51 measurements, there was only a modest correlation between T_{50} and T_M, with a Pearson coefficient correlation (PCC) and Spearman's rank correlation (SRC) of 0.58 and 0.47, respectively. This highlights the difference in the physical properties being measured using these two techniques, T_M being the thermal stability of the protein's structural elements and T_{50} reporting on the thermal stability to irreversible denaturation. However, similar to the previous study [19], the relationship between the predicted stability with the experimental T_M only results in a weak correlation not only with the previous algorithms evaluated (Rosetta ΔΔG and FoldX PSSM) but also with five other commonly used methods: ELASPIC, DeepDDG, PoPMuSiC, SDM, and AUTO-MUTE. This result suggests that while the two measurements are reporting on different physical properties, this is not the key factor that led to the low predictive accuracy of established algorithms on this dataset.

To evaluate the second hypothesis, that the changes in thermal stability of the BglB dataset are too small for current algorithms, we investigated the ability of the algorithms to predict if a mutation reduced thermal stability to the point that the protein could no longer be produced and isolated in a soluble form. Analysis of the computational algorithms to predict destabilization to the point where no soluble protein could be isolated showed a significant enrichment based on the calculated energetics of the mutants for several algorithms, the most

significant of which is for Rosetta $\Delta\Delta G$. This supports the hypothesis that the lack of performance on the BglB dataset is due to the narrow range in changes observed for thermal stability. These slight molecular changes, especially interactions that are less than 1 kcal/mol, are challenging to accurately model. This highlights the need for new algorithms for predicting modest, yet important, changes in thermal stability as well as a new utility for current algorithms for prescreening designs for the production of mutants likely to maintain protein structure and be produced as a soluble protein.

Methods

Mutant Selection, Protein Expression, and Purification

Out of 79 mutants of BglB that were previously characterized with T50 data [19], 51 variants with plasmid readily available were transformed into chemically competent Escherichia coli BLR (DE3) cells. The variants were produced and purified, as previously described [14]. Expression was carried out by growing a 5 mL overnight culture in a 50 mL falcon tube with a breathable seal in Terrific broth (TB) medium with kanamycin while shaking at 250 rpm at 37°C. After the initial overnight culture, cells were spun down and resuspended in fresh TB with kanamycin with 1 mM isopropyl β-d-1-thiogalactopyranoside in a 50 mL falcon tube with a breathable seal and incubated while shaking at 250 rpm at 18°C for 24 h. Then, the cells were spun down, lysed, and purified using immobilized metal ion affinity chromatography, as previously described [19]. The purity of the protein samples was analyzed using 12–14% SDS-PAGE (Figure SI 3-1), and the yield was assessed based on the A280 for proteins that appeared >75% pure in the SDS-PAGE analysis. Protein samples were considered expressed if they were

detectable in the SDS-PAGE analysis and greater than 0.10 mg/mL using A280, as previously described [19].

Melting Temperature Assay

The melting temperature (T_M) of BglB was determined using the Protein Thermal Shift (PTS) kit (Applied Biosystems, from Thermo Fisher Scientific). Standard protocols provided by the manufacturer were used. Protein concentrations ranged from 0.1–0.5 mg/mL, and fluorescence reading was monitored with a QuantaStudio 3 system from 20 to 90 °C. The temperature melting curve was first smoothed with a 20 step sliding window average (Script SI 3-2). T_M was determined from the average of three to four replicates at which the derivative was largest, and all melting curves can be found in Figure SI 3-3.

ΔG Calculations from T_M

Calculations were conducted, as previously described [21]. First, we assumed that the protein follows the two-state folding mechanism, a binary conversion of native state to full denaturation. Second, to derive $\Delta G^\circ_{unfolding}$, the fluorescence intensity was first translated into fractions of folded (P_f) and unfolded (P_u) proteins of the linear portion of the graph at different temperatures starting from the minimum fluorescence (F_{min}) to the maximum fluorescence (F_{max}) shown in Equation 1.

$$P_f = 1 - \frac{F - F_{min}}{F_{max} - F_{min}}$$

Equation 1

By taking a two-state folding–unfolding model, the equilibrium constant of unfolding (K_u) at different temperatures is then given by Equation 2.

$$K_u = \frac{P_f}{P_u}$$

Equation 2

We plotted ln K_u against $1/T$ using the van't Hoff method shown in Equation 3 (Script SI 3-2), the $-\Delta H/R$ is defined by the slope, $\Delta S/R$ is the *y*-intercept, *T* is temperature, and *R* is the ideal gas constant.

$$\ln K_u = -\frac{\Delta H}{RT} + \frac{\Delta S}{R}$$

Equation 3

The Gibbs free energy of protein-unfolding can then be determined using Equation 4, where $\Delta G°_{unfolding}$ is the unfolding energy at a T_{RT} of 298 K. All calculations can be found in Script SI 3-2.

$$\Delta G°_{unfolding} = \Delta H - T_{RT}\Delta S$$

Equation 4

Molecular Modeling

Seven popular, readily accessible, and recently developed molecular modeling methods, many of them force-field and machine-learning-based, were evaluated for their ability to recapitulate the experimental data: Rosetta ΔΔ*G* [22], FoldX [23], ELASPIC [24], DeepDDG [25], PoPMuSiC [26], SDM[27], and AUTO-MUTE (DDG) [28]. The crystal structure of BglB (PDB ID: 2JIE) was used across seven different algorithms. First, using a previously described method [19] the 2JIE structure was used as input to the Rosetta ΔΔ*G* application and run, as previously described (Script SI 3-5). Briefly, 50 poses of the WT and the mutant were generated for which 15 energy terms were reported from the score function

used. [22] The three lowest system energy scores out of 50 from WT and the mutant were averaged to give the final Rosetta $\Delta\Delta G$ score. Second, for the FoldX position-specific scoring metric (PSSM) protocol, the 2JIE structure was first minimized for any potential inaccurate rotamer assignment using the RepairPDB application [23]. The repaired PDB structure was mutated with single-point mutants and then modeled using FoldX PSSM. The model was scored based on 17 terms within the FoldX force-field [23]. Third, the ELASPIC protocol first constructed a homology model of the WT using the crystal structure, sequence, molecular, and energetics information. Using the standard procedure described, the FoldX algorithm was used to construct the mutant model. The final mutational change is predicted using Stochastic Gradient Boosted Decision Trees based on the energetic, chemical, and structural features from FoldX [24,29]. Fourth, using a curated dataset derived from the Protherm database [15], DeepDDG used their previously described shared residue pair neural network structure to make a prediction of stability [25]. The DeepDDG output indicated that >0 kcal/mol could be considered stable, whereas <0 kcal/mol could be considered unstable. Fifth, PoPMuSiC estimated the stability of the WT structure and mutants using 13 statistically potential terms, and an additional two terms that account for the volume differences of the residues between WT and the mutant [26]. Sixth, the SDM method evaluated mutational changes using a statistical potential energy function based on environment-specific substitution tables. These tables consisted of data such as structural information, solvent accessibility of the sidechain, and hydrogen bonding [27]. Lastly, similar to SDM, the seventh method, AUTO-MUTE, which predicts for ΔT_M and $\Delta\Delta G$, utilized a statistical potential to calculate the environmental changes of the residue compared to the WT [28]. The protocol was performed using tree regression at 23 °C and pH 7.5.

Apart from predicting $\Delta\Delta G$, two additional methods were used to evaluate the algorithms' ability to predict ΔT_M changes. As mentioned above, the AUTO-MUTE prediction using the Stability Changes (ΔT_M) protocol was performed with tree regression. Also, HotMuSiC was used to evaluate the mutational effect with the temperature-dependent potential and other statistical potential terms such as solvent accessibility, structural, and sequence-based information [30].

Pearson correlation coefficient (PCC) and Spearman's rank correlation (SRC) analyses were performed between their respective $\Delta\Delta G$ ($\Delta\Delta G = \Delta G_{mutants} - \Delta G_{WT}$) or the change in total system energy (ΔTSE) of the nine computational methods. Additionally, the available individual features within the Rosetta $\Delta\Delta G$ and FoldX PSSM force field were further evaluated against the T_M dataset for correlation.

Finally, ΔTSE was evaluated against mutants that could be isolated as a soluble protein and those that lost structural integrity and either precipitated or were degraded and therefore could no longer be isolated as a soluble protein (nonisolated). A Student's *t*-test was used to obtain *p*-values for the nine computational methods. The two categories between isolated and nonisolated protein were treated as an independent sample using an unequal variance.

Results

Evaluating the Relationship between T_M and T_{50}

To the best of our knowledge, there has not been a large dataset (>50 data points) directly comparing the T_M and T_{50} relationship for a single set of protein mutants uniformly produced and characterized. It is important to distinguish both T_M and T_{50} methods since the measurements are quantifying and reporting different structural and functional properties. T_M is the temperature at

which half the enzyme is found in the unfolded state over the folded state [17,31]. This is often evaluated through denaturation assays, where the thermodynamic measurements ($\Delta G_{unfolding}$) can be obtained [31]. This method is generally a lower throughput method as purified protein is required to obtain an accurate measurement for the structural properties for the mutant being evaluated. T_{50} measures the temperature of half-inactivation that leads to irreversible unfolding [11,32], and it is determined by the reduction of half of the enzymatic activity due heat challenges [17]. This is a very common assay for protein engineering due to its compatibility with high-throughput assays and the ability to use cell lysates to evaluate function.

To complement our previously measured dataset of T_{50}, 51 of the 92 expressed proteins with available plasmids [19] were selected and evaluated for T_M using the Protein Thermal Shift assay to compare T_{50} and T_M. The WT BglB T_M was determined to be 45.97 ± 1.03 °C, while the previously determined T_{50} was 39.9 ± 0.1 °C [19]. When evaluating the entire dataset, the T_M ranged between 37.1 and 54.3 °C, slightly larger than what was observed for T_{50}, which was between 34.9 and 46.0 °C (Figure SI 3-1-2). The variant that had the highest T_M in this dataset was E167A, with a ΔT_M of 54.3 °C (+8.33 °C), which was also observed to have a similar increase in T_{50} compared with the WT (+6.06 °C) [19]. The variant that had the lowest T_M in this dataset was found to be E225A, with a ΔT_M of −8.9 °C, which had a corresponding T_{50} of −3.1 °C.

The relationship between T_{50} and T_M is plotted in Figure 3-2A. The PCC and SRC of 0.58 and 0.47, respectively, indicate that the two methods are moderately positively correlated. Correlation between methods increased in cases where mutations resulted in extremely stable and unstable products, for example, E167A and E225A, respectively. This is an expected result for small changes (<3 °C) in thermal stability; the differences in measurement methods would be

expected to play a more significant role than for larger changes (>5 °C). The evaluation of ΔT_M and ΔT_{50} with experimentally derived $\Delta\Delta G$ is also plotted in Figure 3- 2B and 3-2C, respectively. The PCC and SRC show that the T_M method and experimentally derived $\Delta\Delta G$ are strongly correlated (PCC and SRC of −0.76), compared to those between $\Delta\Delta G$ and T_{50} (PCC of −0.35 and SRC of −0.24).

Evaluating Computational Stability Tools Using the BglB T_M Dataset

The computational evaluation of protein stability of the current experimental T_M dataset was analyzed in the same manner as our previous study on T_{50} [19]. An energetically evaluated model for each mutant was generated using established computational methods and subsequently plotted as a function of T_M to evaluate the calculated energies related to the observed T_M. The PCC and SRC for the most commonly assessed term, the ΔTSE, was found to be highest for FoldX PSSM (PCC of −0.34 and SRC of −0.35) with ΔT_M (Figure 3-3). Similarly, the FoldX PSSM correlations with experimentally derived ΔT_{50} data were found to be −0.21 and −0.16 for PCC and SRC, respectively. The overall relationship between the ΔTSE and the ΔT_M thermal stability dataset slightly improved for FoldX, DeepDDG, PoPMuSiC, and AUTO-MUTE (Figure SI 3-1-3), while Rosetta $\Delta\Delta G$ and ELASPIC remained relatively unchanged with no significant correlation. Interestingly, SDM was the only method where the correlation with ΔT_{50} is stronger than that of ΔT_M (Figure SI 3-1-3).

An analysis of individual energetic term from Rosetta $\Delta\Delta G$ and FoldX PSSM did not uncover any specific feature in either method's energetic evaluation that was strongly correlated with the T_M dataset, as was previously observed for the T_{50} dataset [19] (Figure SI 3-4). The strongest PCC for T_M against any of the available energetic terms was 0.39 for the Δbackbone

clash term from FoldX PSSM and −0.31 for the Omega energy term from Rosetta $\Delta\Delta G$. To be consistent with the previous performance assessment, we also evaluated the algorithms on experimentally derived $\Delta\Delta G$ in this dataset (Figure 3-3). The PCC and SRC of 0.39 and 0.36, respectively, between experimental $\Delta\Delta G$ and ΔTSE for FoldX PSSM outperformed six other algorithms that were compared. The correlation between experimental $\Delta\Delta G$ with ΔTSE was not unexpected as T_M showed a correlation with $\Delta\Delta G$ with a PCC and SRC of −0.76 (Figure 3-2B). Analysis of AUTO-MUTE and HotMuSiC to predict for ΔT_M revealed no significant correlation with the experimental ΔT_M (Figure SI 3-1-3).

Based on this analysis, it is apparent that the general performance of all given methods at best only weakly correlates with the experimentally determined effects of the mutations. This data fails to support the hypothesis that the lack of a previously observed correlation of these established computational tools with observed changes in thermal stability in the BglB dataset is due to the difference in the physical property being measured.

Prediction of Mutant Soluble Expression

The current dataset consists primarily of modest changes in thermal stability of <5 °C, calculated to be ±4 kcal/mol of the WT, and therefore may be challenging for current computational methods to predict. However, this change has only been analyzed in a fraction of the 129 mutants tested in the overall BglB dataset. Of the 129 mutants, only 92 were found to be produced and isolated in a soluble form. All purification procedures are conducted at ~20 °C. Since the WT has a T_{50} of 39.9 °C, any reduction in T_{50} of >18 °C would result in a loss of structural stability from which insoluble aggregates or proteolytic degradation would readily occur during production and purification. In this case, the proteins would no longer be able to be

isolated in a soluble form similar to the WT protein. Therefore, it seemed pertinent to evaluate if any of the nine algorithms could differentiate variants in this dataset that could be isolated as a soluble protein versus those that were not able to be separated as a soluble protein.

For this evaluation, all of the previously reported 129 mutants were assessed using the nine algorithms following the same methods used for T_{50} and T_M. A mutant was generally considered soluble if it was observed on an SDS-PAGE analysis and had an A280 >0.1 mg/mL. The WT protein produced using the methods described generally resulted in an average A280 of 1.5 mg/mL, which would provide a >10-fold change in yield for mutants having an A280 less than 0.1 mg/mL. While factors other than thermal stability can affect production and isolation of soluble protein, in this case, it is assumed that the primary factor that decreases soluble protein yield is from denaturation of the mutant protein either during expression or purification. The results of this analysis are presented in Figure 3-4.

Of the nine algorithms evaluated, Rosetta $\Delta\Delta G$, FoldX, DeepDDG, PoPMuSiC, and SDM can capture the enrichment of mutants isolated as a soluble protein. The differences were evaluated for statistical significance using the Student's *t*-test, and the highest among the top five methods was shown for Rosetta $\Delta\Delta G$ with a *p*-value of 1.0×10^{-5}. In contrast, enrichment was lower for ELASPIC, AUTO-MUTE (DDG), AUTO-MUTE (ΔT_M), and HotMuSiC with *p*-values of 0.06, 0.38, 0.65, and 0.07, respectively.

A few outliers were observed in all methods, except for ELASPIC (Figure SI 3-1-4). For example, the mutant G15N for both Rosetta $\Delta\Delta G$ and FoldX PSSM was identified as severely energetically unfavorable, which is consistent with the observation that this variant was not able to be isolated as a soluble protein.

Discussion

Both T_M and T_{50} are methods commonly used to quantify different physical aspects of protein thermal stability; however, to date, there has been relatively little experimental data collected to empirically evaluate the relationship of these two measurements. Using a dataset of 51 protein mutants, we observed that there is a moderate positive correlation (PCC of 0.58 and SRC of 0.47) between these two properties. The theory comparing two methods has been extensively described in the work of Hei and Clark [33]. Briefly, T_{50} can only be used to assess the temperature at which half of the protein is irreversibly unfolded. Meanwhile, T_M provides information on the folded state of the protein regardless of whether or not the unfolding events are irreversible. Therefore, it is not surprising that there is only a moderate correlation between the relationship of T_{50} and T_M.

Mutants with extreme stability changes, such as E164A (>6 °C), usually exhibit a similar magnitude of change in T_M and T_{50} results. However, the majority of the mutants show a change of ∼3 °C or less in this T_M and T_{50} dataset being analyzed, a range in which the relationship between T_M and T_{50} appears to be weaker. Therefore, analysis with larger datasets with more extreme stability changes may reveal an even stronger correlation between these two properties. The relationship between ΔT_M and the experimentally derived $\Delta\Delta G$ of this dataset (PCC and SRC of −0.76) is not expected to reach a perfect correlation since it is dependent on the temperature at which $\Delta\Delta G$ was evaluated, as described by Pucci et al. [34] For example, the $\Delta\Delta G$ evaluated at T_M of the WT will yield a correlation closer to −1 and $\Delta\Delta G_{(25°C)}$ will lead to a lower correlation (−0.68). [34]

Consistent with our previous analysis, we found a lack of performance using established computational tools when predicting T_M and T_{50} from the WT for this dataset. According to Jia et

al., stability prediction using the experimentally derived free energy change of unfolding $\Delta\Delta G$ (kcal/mol) outperforms the prediction using ΔT_M (°C). [35] However, in this case, we saw no significant change in the predictive performance for all seven computational tools compared to the experimentally derived free energy change. In addition, we found T_M and $\Delta\Delta G$ to be strongly correlated with this dataset, which may suggest that the improved performance is only relevant for more diverse datasets composed of different proteins as opposed to mutants of a single protein.

While none of the computational methods demonstrated a strong predictive power for the mutants in this study, Rosetta $\Delta\Delta G$, FoldX PSSM, ELASPIC, and PoPMuSiC all have previously been shown to have high correlations with experimental data (PCC between 0.69 to 0.83) [22, 26, 29, 36]. This dataset with an experimental $\Delta\Delta G$ range of ±∼4 kcal/mol is within the majority of the mutants observed in the algorithms that were typically evaluated on (+8 to −5 kcal/mol) [17, 25, 26, 29]. One potential reason for the lack of performance could be that the structure used in this dataset has a ligand bound structure, and often, the structures used in the development of the methods were apoprotein structures. However, using the PDBFlex database, [37] a clustering of five available PDBs of BglB from the bacterium *P. polymyxa* showed an average RMSD of 0.234 and a maximum RMSD of 0.274, thus making BglB a rigid structure. As there are no significant structural changes between the apo-form and holo-form of the protein, it is unlikely that the exact structure used for this study resulted in the low level of performance by the algorithms. Another possibility is that the protein evaluated here (BglB) is an outlier when compared to the proteins used to develop the algorithms in terms of its structure–function relationship. However, a related study to our analysis has been conducted for human superoxide dismutase 1 in which a low correlation is observed between experimental and

predicted stability. [38] This further validates that current algorithms have limited utility for proteins outside of those they were benchmarked on. This limitation hindered by an over-representation of protein families such as lysozyme, tryptophan synthase, and ribonuclease in curated datasets is often utilized in benchmarking. [39] Thus, this highlights the importance of generating high-quality and diverse datasets of more proteins for evaluating and training new computational tools.

This study underlines the need for new computational tools that can more accurately predict modest changes, rather than major changes, in thermal stability. This becomes particularly important because single-point mutants often increase thermal stability by a few degrees at a time, while major changes are more often produced from the synergistic effect of combining multiple mutations [11, 40-42]. Furthermore, as larger datasets of protein mutants with explicitly measured biophysical properties are generated, opportunities to explore combinations of molecular modeling and machine learning methods will become practical. These algorithms and datasets will enable the development of robust predictors of thermal stability.

Figures

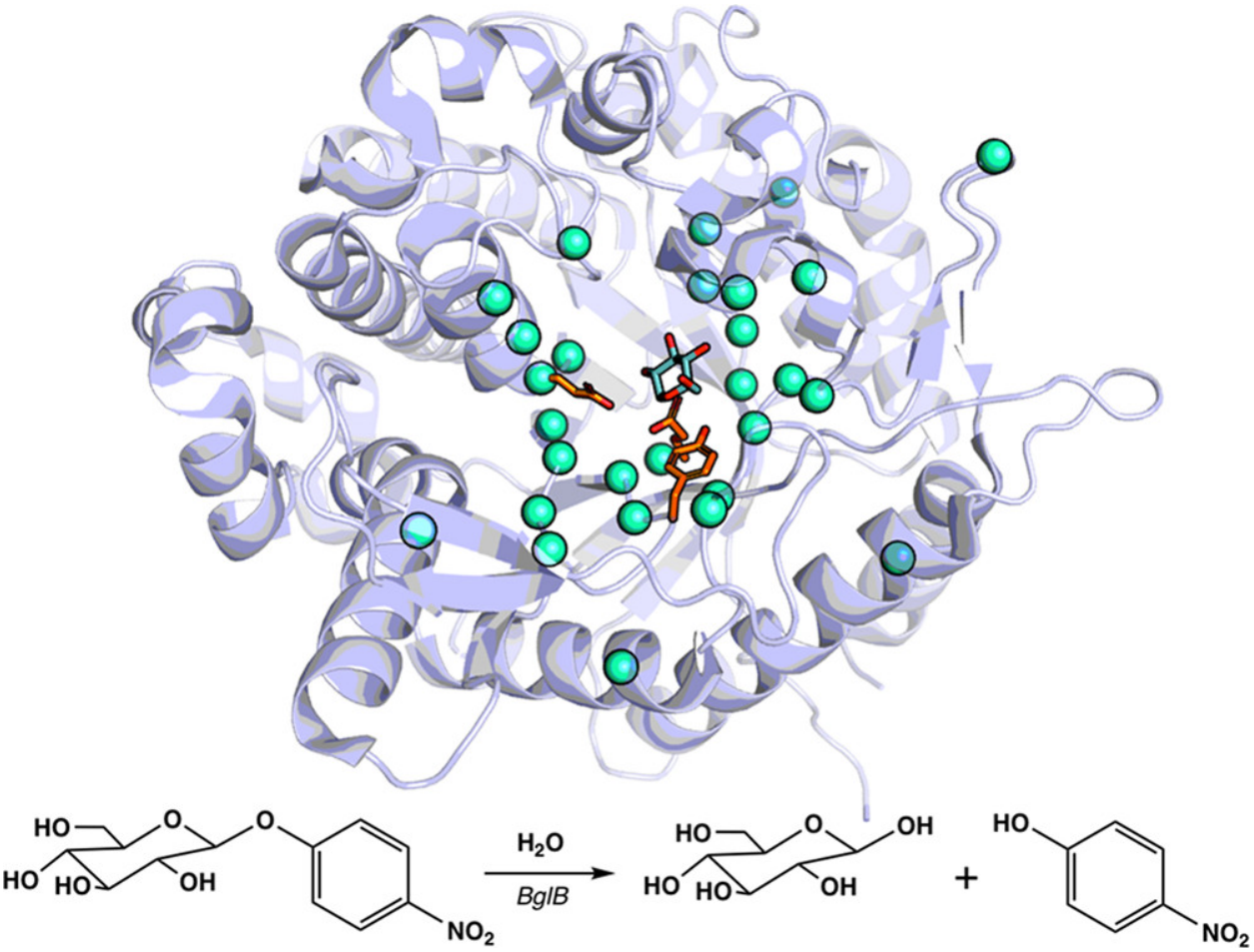


Figure 3-4. Structure of BglB (PDB ID: 2JIE) from the bacterium *Paenibacillus polymyxa*. PyMOL rendering [20] of BglB showing the 28 sequence-positions (teal spheres) of the 51 mutants chosen out of the original 92 previously expressed proteins for the T_M analysis [19]. The reaction scheme of the hydrolysis of 4-nitrophenyl β-D-glucopyranoside by BglB used in the T_{50} study [19].

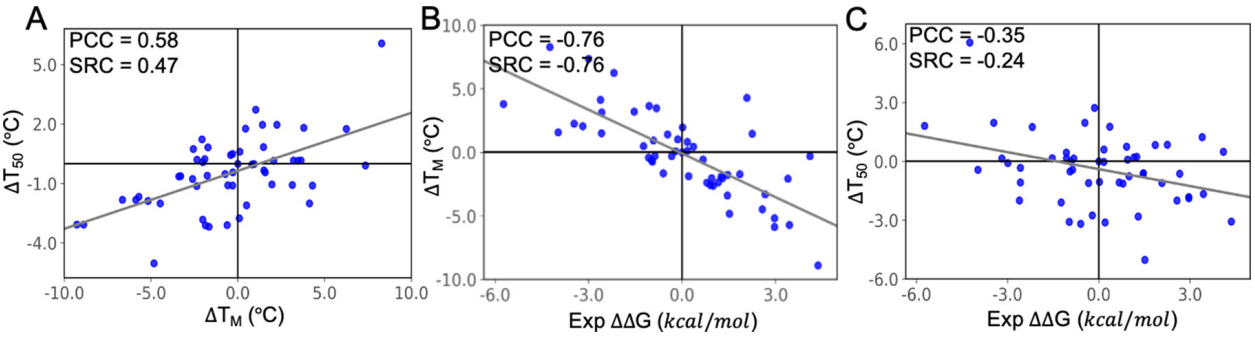


Figure 3-2. Comparison of two different experimental thermal stability datasets and experimentally derived $\Delta\Delta G$. (A) Relationship for each mutant between T_{50} and T_M. The PCC of 0.58 illustrates that the two methods are modestly positively correlated with mutations that are in the extreme ends of the temperature range (±5 °C). (B) Evaluation of ΔT_M with experimentally derived $\Delta\Delta G$ shows the two qualities are highly correlated (PCC = −0.76), unlike (C) where the relationship between ΔT_{50} and experimentally derived $\Delta\Delta G$ has a PCC of −0.35.

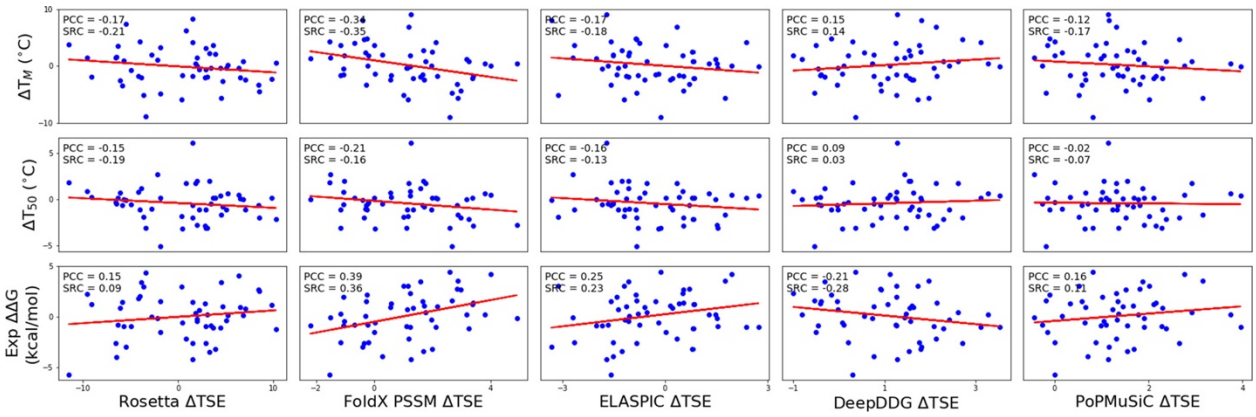


Figure 3-3. Evaluation of the algorithms ΔTSE versus the experimentally derived ΔΔ*G* and the T_M and T_{50} datasets. The Pearson correlation coefficient and Spearman's rank correlation for each performance against three types of experimental data were determined. Five representative comparisons are illustrated above, with four additional algorithms, SDM, AUTO-MUTE (DDG), AUTO-MUTE (ΔT_M), and HoTMuSiC provided in Figure SI 3-1-3. No algorithm resulted in a significant correlation between the calculated energies and the observed T_M, T_{50}, or ΔΔ*G* for this dataset.

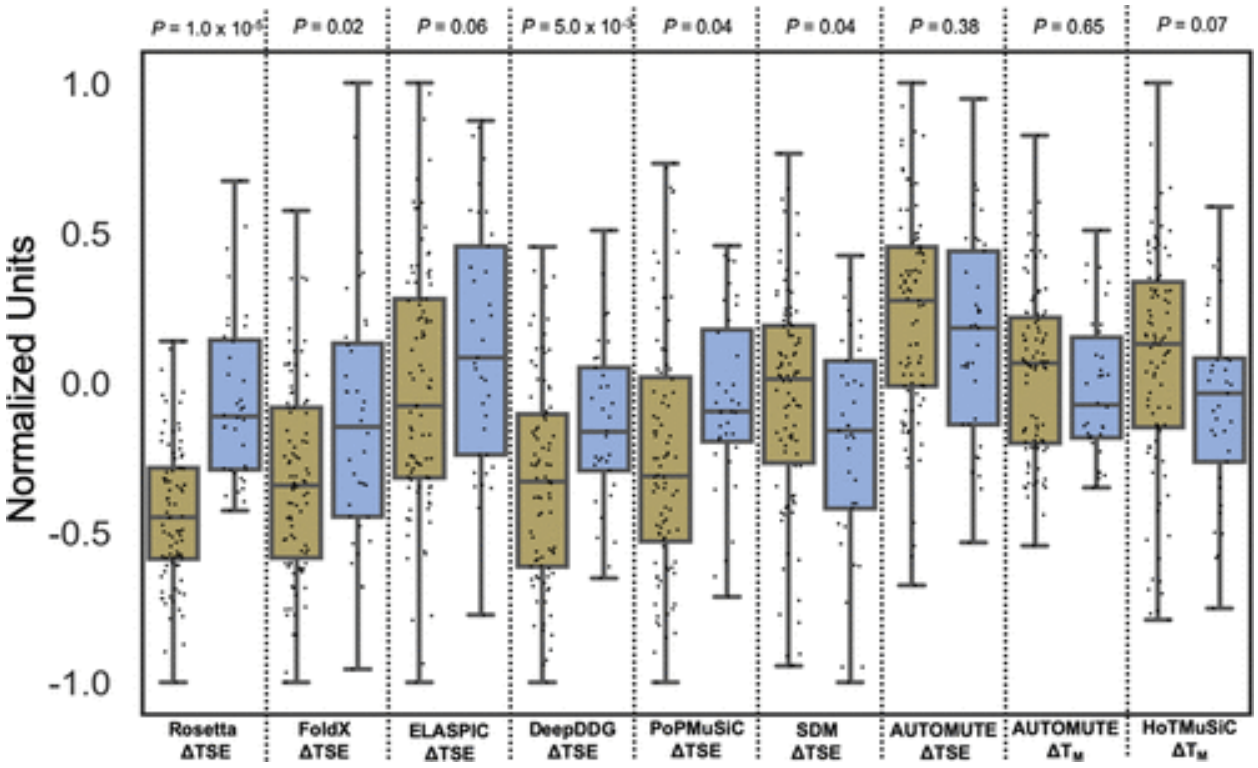


Figure 3-4. Computational prediction for the effect on mutant soluble protein production using nine different algorithms. From left to right: Rosetta ΔΔ*G*, FoldX PSSM, ELASPIC, DeepDDG, PoPMuSiC, SDM, AUTO-MUTE (DDG), AUTO-MUTE (ΔT_M), and HoTMuSiC of soluble (green) and nonisolated protein (blue). In this case, mutants that resulted in a significant (>10-fold) decrease in yield of purified soluble protein are considered nonisolatable. Significance in population differences was determined using a Student's *t*-test. The units (ΔTSE and ΔT_M) of all algorithms are individually normalized between 1 to −1. For visualization purposes, outliers were omitted after normalization. Each graph without normalization and with outliers can be found in Figure SI 3-1-4 and all raw values in Figure SI 3-4.

References

(1) Iyer, P. V.; Ananthanarayan, L. Enzyme Stability and Stabilization-Aqueous and Non-Aqueous Environment. Process Biochem. 2008, 43, 1019−1032.

(2) Turner, P.; Mamo, G.; Karlsson, E. N. Potential and Utilization of Thermophiles and Thermostable Enzymes in Biorefining. Microb. Cell Fact. 2007, 6, 9.

(3) Ferdjani, S.; Ionita, M.; Roy, B.; Dion, M.; Djeghaba, Z.; Rabiller, C.; Tellier, C. Correlation between Thermostability and ACS Omega http://pubs.acs.org/journal/acsodf Article https://dx.doi.org/10.1021/acsomega.9b04105 ACS Omega 2020, 5, 6487−6493 6492 Stability of Glycosidases in Ionic Liquid. Biotechnol. Lett. 2011, 33, 1215−1219.

(4) Xie, Y.; An, J.; Yang, G.; Wu, G.; Zhang, Y.; Cui, L.; Feng, Y. Enhanced Enzyme Kinetic Stability by Increasing Rigidity within the Active Site. J. Biol. Chem. 2014, 289, 7994−8006.

(5) Wu, I.; Arnold, F. H. Engineered Thermostable Fungal Cel6A and Cel7A Cellobiohydrolases Hydrolyze Cellulose Efficiently at Elevated Temperatures. Biotechnol. Bioeng. 2013, 110, 1874−1883.

(6) Kazlauskas, R. J.; Bornscheuer, U. T. Finding Better Protein Engineering Strategies. Nat. Chem. Biol. 2009, 5, 526.

(7) Denard, C. A.; Ren, H.; Zhao, H. Improving and Repurposing Biocatalysts via Directed Evolution. Curr. Opin. Chem. Biol. 2015, 25, 55−64.

(8) Borgo, B.; Havranek, J. J. Automated Selection of Stabilizing Mutations in Designed and Natural Proteins. Proc. Natl. Acad. Sci. U. S. A. 2012, 109, 1494−1499.

(9) Jacak, R.; Leaver-Fay, A.; Kuhlman, B. Computational Protein Design with Explicit Consideration of Surface Hydrophobic Patches. Proteins: Struct., Funct., Bioinf. 2012, 80, 825−838.

(10) Lehmann, M.; Loch, C.; Middendorf, A.; Studer, D.; Lassen, S. F.; Pasamontes, L.; van Loon, A. P. G. M.; Wyss, M. The Consensus Concept for Thermostability Engineering of Proteins: Further Proof of Concept. Protein Eng. 2002, 15, 403−411.

(11) Goldenzweig, A.; Fleishman, S. J. Principles of Protein Stability and Their Application in Computational Design. Annu. Rev. Biochem. 2018, 87, 105−129.

(12) Cruz, L.; Urbanc, B.; Borreguero, J. M.; Lazo, N. D.; Teplow, D. B.; Stanley, H. E. Solvent and Mutation Effects on the Nucleation of Amyloid β-Protein Folding. Proc. Natl. Acad. Sci. U. S. A. 2005, 102, 18258−18263.

(13) Dehghanpoor, R.; Ricks, E.; Hursh, K.; Gunderson, S.; Farhoodi, R.; Haspel, N.; Hutchinson, B.; Jagodzinski, F. Predicting the Effect of Single and Multiple Mutations on Protein Structural Stability. Molecules 2018, 23, 251.

(14) Tokuriki, N.; Stricher, F.; Schymkowitz, J.; Serrano, L.; Tawfik, D. S. The Stability Effects of Protein Mutations Appear to Be Universally Distributed. J. Mol. Biol. 2007, 369, 1318−1332.

(15) Gromiha, M. M.; An, J.; Kono, H.; Oobatake, M.; Uedaira, H.; Prabakaran, P.; Sarai, A. ProTherm, Version 2.0: Thermodynamic Database for Proteins and Mutants. Nucleic Acids Res. 2000, 28, 283− 285.

(16) Wang, C. Y.; Chang, P. M.; Ary, M. L.; Allen, B. D.; Chica, R. A.; Mayo, S. L.; Olafson, B. D. ProtaBank: A Repository for Protein Design and Engineering Data. Protein Sci. 2018, 27, 1113−1124.

(17) Polizzi, K. M.; Bommarius, A. S.; Broering, J. M.; ChaparroRiggers, J. F. Stability of Biocatalysts. Curr. Opin. Chem. Biol. 2007, 220.

(18) Bloom, J. D.; Labthavikul, S. T.; Otey, C. R.; Arnold, F. H. Protein Stability Promotes Evolvability. Proc. Natl. Acad. Sci. U. S. A. 2006, 103, 5869−5874.

(19) Carlin, D. A.; Hapig-Ward, S.; Chan, B. W.; Damrau, N.; Riley, M.; Caster, R. W.; Bethards, B.; Siegel, J. B. Thermal Stability & Kinetic Constants for 129 Variants of a Family 1 Glycoside Hydrolase Reveal That Enzyme Activity & Stability Can Be Separately Designed. PLoS One 2017, 12, No. e0176255.

(20) Schrödinger, L. The PyMOL Molecular Graphics System http:// www.pymol.org.

(21) Wright, T. A.; Stewart, J. M.; Page, R. C.; Konkolewicz, D. Extraction of Thermodynamic Parameters of Protein Unfolding Using Parallelized Differential Scanning Fluorimetry. J. Phys. Chem. Lett. 2017, 8, 553−558.

(22) Kellogg, E. H.; Leaver-Fay, A.; Baker, D. Role of conformational sampling in computing mutation-induced Changes in Protein Structure and Stability. Proteins: Struct., Funct., Bioinf. 2011, 79, 830− 838.

(23) Schymkowitz, J.; Borg, J.; Stricher, F.; Nys, R.; Rousseau, F.; Serrano, L. The FoldX Web Server: An Online Force Field. Nucleic Acids Res. 2005, 33, W382−W388.

(24) Witvliet, D. K.; Strokach, A.; Giraldo-Forero, A. F.; Teyra, J.; Colak, R.; Kim, P. M. ELASPIC Web-Server: Proteome-Wide Structure-Based Prediction of Mutation Effects on Protein Stability and Binding Affinity. Bioinformatics 2016, 32, 1589−1591.

(25) Cao, H.; Wang, J.; He, L.; Qi, Y.; Zhang, J. Z. DeepDDG: Predicting the Stability Change of Protein Point Mutations Using Neural Networks. J. Chem. Inf. Model. 2019, 59, 1508−1514.

(26) Dehouck, Y.; Kwasigroch, J. M.; Gilis, D.; Rooman, M. PoPMuSiC 2.1: A Web Server for the Estimation of Protein Stability Changes upon Mutation and Sequence Optimality. BMC Bioinf. 2011, 12, 151.

(27) Worth, C. L.; Preissner, R.; Blundell, T. L. SDM - A Server for Predicting Effects of Mutations on Protein Stability and Malfunction. Nucleic Acids Res. 2011, 39, W215−W222.

(28) Masso, M.; Vaisman, I. I. AUTO-MUTE 2.0: A Portable Framework with Enhanced Capabilities for Predicting Protein Functional Consequences upon Mutation. Adv. Bioinf. 2014, 2014, 278385.

(29) Berliner, N.; Teyra, J.; Çolak, R.; Lopez, S. G.; Kim, P. M. Combining Structural Modeling with Ensemble Machine Learning to Accurately Predict Protein Fold Stability and Binding Affinity Effects upon Mutation. PLoS One 2014, 9, No. e107353.

(30) Pucci, F.; Bourgeas, R.; Rooman, M. Predicting Protein Thermal Stability Changes upon Point Mutations Using Statistical Potentials: Introducing HoTMuSiC. Sci. Rep. 2016, 6, 23257.

(31) Musil, M.; Konegger, H.; Hon, J.; Bednar, D.; Damborsky, J. Computational Design of Stable and Soluble Biocatalysts. ACS Catal. 2019, 9, 1033−1054.

(32) Colon, W.; Church, J.; Sen, J.; Thibeault, J.; Trasatti, H.; Xia, K.′ Biological Roles of Protein Kinetic Stability. Biochemistry 2017, 56, 6179−6186.

(33) Hei, D. J.; Clark, D. S. Estimation of Melting Curves from Enzymatic Activity−Temperature Profiles. Biotechnol. Bioeng. 1993, 42, 1245−1251.

(34) Pucci, F.; Bourgeas, R.; Rooman, M. High-Quality Thermodynamic Data on the Stability Changes of Proteins upon Single-Site Mutations. J. Phys. Chem. Ref. Data 2016, 45, No. 023104.

(35) Jia, L.; Yarlagadda, R.; Reed, C. C. Structure Based Thermostability Prediction Models for Protein Single Point Mutations with Machine Learning Tools. PLoS One 2015, 10, e0138022.

(36) Guerois, R.; Nielsen, J. E.; Serrano, L. Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More than 1000 Mutations. J. Mol. Biol. 2002, 320, 369−387.

(37) Hrabe, T.; Li, Z.; Sedova, M.; Rotkiewicz, P.; Jaroszewski, L.; Godzik, A. PDBFlex: Exploring Flexibility in Protein Structures. Nucleic Acids Res. 2016, 44, D423−D428.

(38) Kepp, K. P. Computing Stability Effects of Mutations in Human Superoxide Dismutase 1. J. Phys. Chem. B 2014, 118, 1799− 1812.

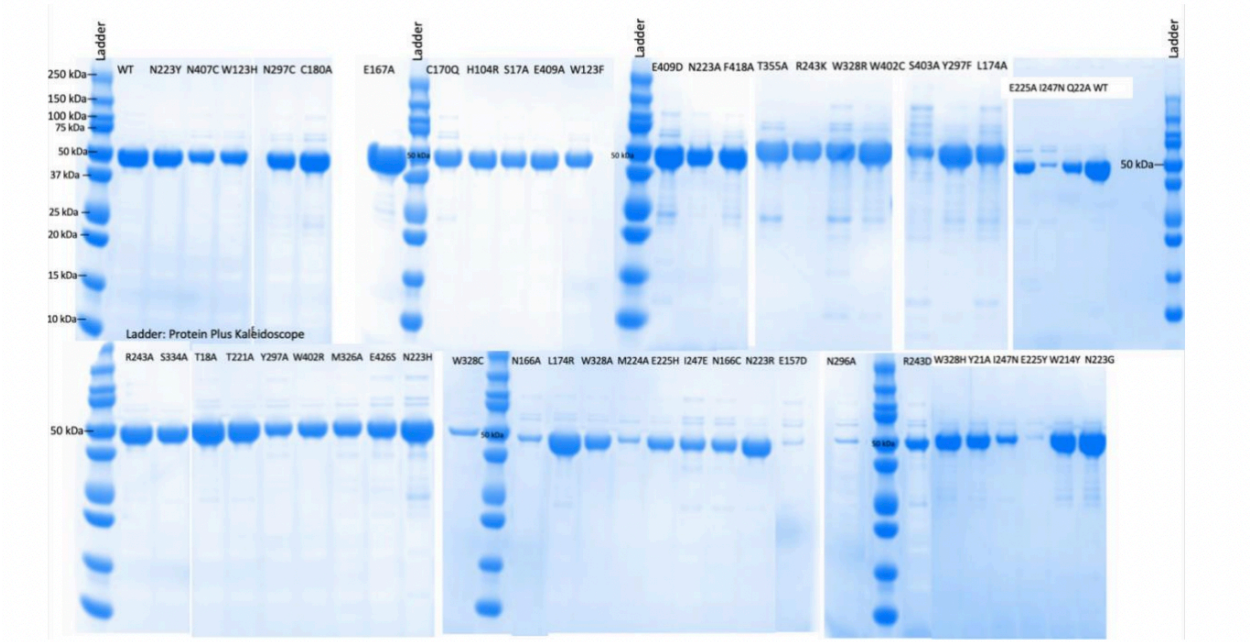
(39) McGuinness, K. N.; Pan, W.; Sheridan, R. P.; Murphy, G.; Crespo, A. Role of Simple Descriptors and Applicability Domain in Predicting Change in Protein Thermostability. PLoS One 2018, 13, e0203819.

(40) Korkegian, A.; Black, M. E.; Baker, D.; Stoddard, B. L. Computational Thermostabilization of an Enzyme. Science 2005, 308, 857−860.

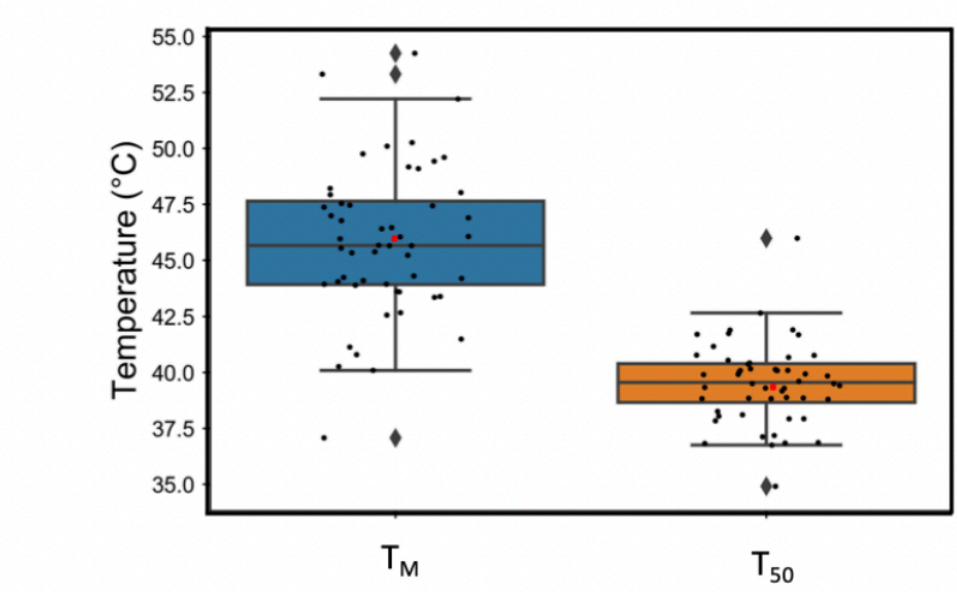
(41) Wakabayashi, H.; Griffiths, A. E.; Fay, P. J. Combining Mutations of Charged Residues at the A2 Domain Interface Enhances Factor VIII Stability over Single Point Mutations. J. Thromb. Haemostasis 2009, 7, 438−444.

(42) Li, G.; Fang, X.; Su, F.; Chen, Y.; Xu, L.; Yan, Y. Enhancing the Thermostability of Rhizomucor Miehei Lipase with a Limited Screening Library by Rational-Design Point Mutations and Disulfide Bonds. Appl. Environ. Microbiol. 2018, 84, No. e02129.

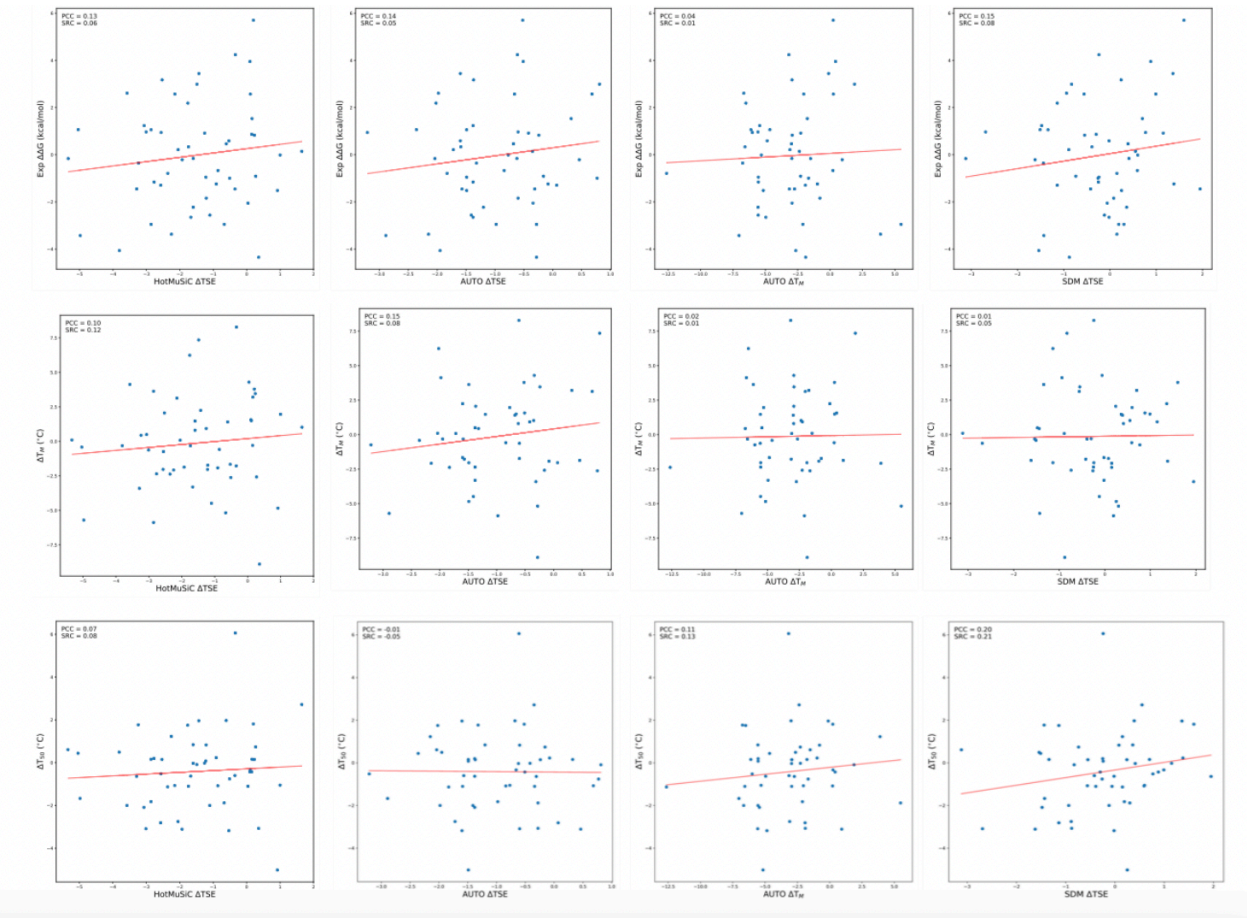
Supplementary Information



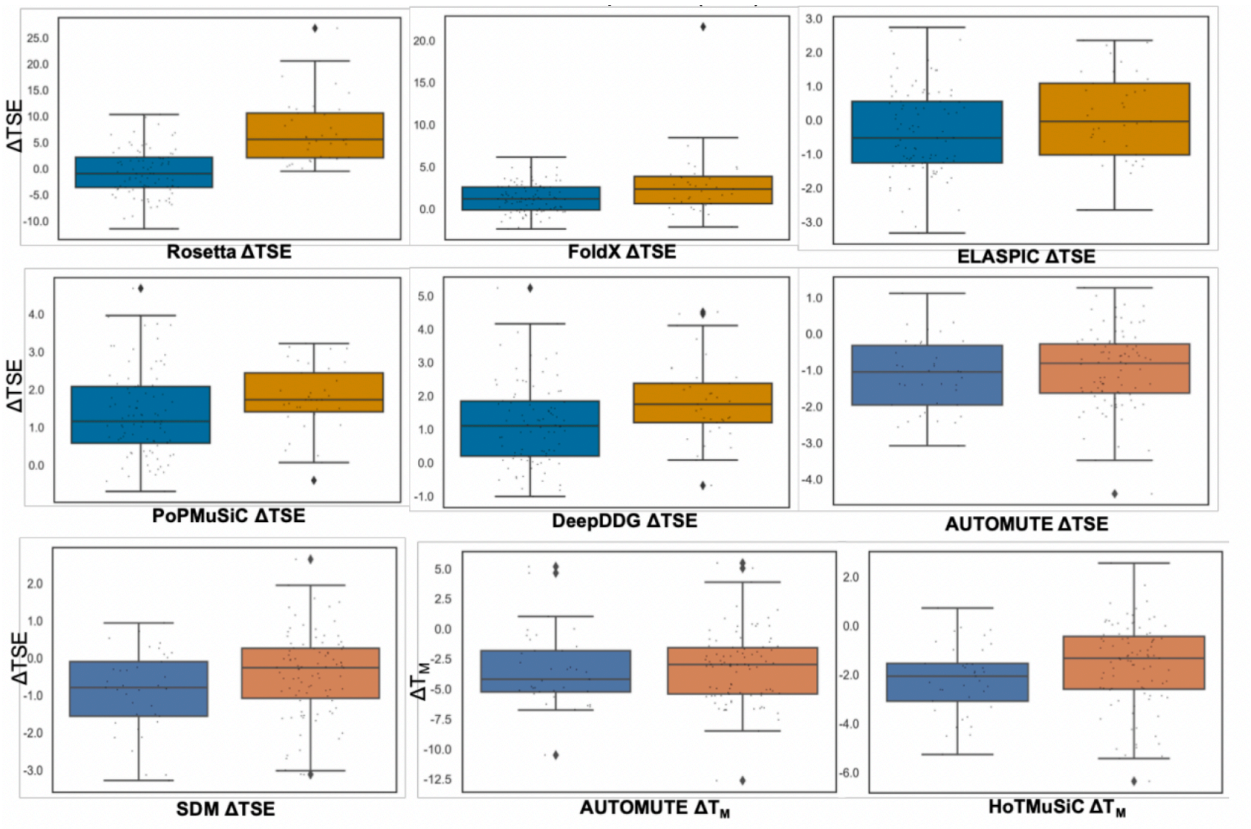
SI 3-1-1. 12-14% SDS-PAGE for 51 BglB mutants and wild type (WT) using Protein Plus Kaleidoscope as ladder.



SI 3-1-2. A distribution analysis of temperatures observed for each method. Black circles represent mutants and the red circle represents the native protein observed TM or T50.



SI 3-1-3. Evaluation of AUTO-MIUTE, SDM, and HoTMuSiC on thermal stability dataset.



SI 3-1-4. Evaluation of nine computational methods on protein expression. Top panel is Rosetta ΔΔG, FoldX, and ELASPIC ΔTSE; middle panel is PoPMuSic,DeepDDG, and AUTOMUTE (ddG) ΔTSE; and bottom panel is SDM ΔTSE, and AUTOMUTE(ΔTM), and HotMuSiC ΔTM of soluble (blue) and non-isolated protein (blue).

SI 3-2. The SI_script.ipynb script contains method used for data acquisitions to generate all the graphs from experimental TM data, as well as methods to obtain all the thermodynamics parameters (ΔΔG, ΔΔH, ΔΔS, and ΔTM). The folder also includes individual .csv files of all raw data used (fluorescence vs temperature) for data acquisitions. Available at https://pubs.acs.org/doi/10.1021/acsomega.9b04105

SI 3-3. The PDF file contains images of fluorescence graphs, 1st derivative graphs, and Van't Hoff plot for 51 mutants in quadruplicates and 6 biological replicates for WT in .pdf format. All graphs were generated using matplotlib found in SI 2. Available at https://pubs.acs.org/doi/10.1021/acsomega.9b04105

SI 3-4. The SI 4 file contains two folders (Rosetta ΔΔG and FoldX PSSM) consisted of PCC graphs of ΔTM with each individual energy term described in the Rosetta ΔΔG and FoldX PSSM energy scoring protocols. The two folder also has a .csv file containing all the raw data from FoldX PSSM and Rosetta ΔΔG for all previously described mutants. An excel files (.csv) containing raw data for DeepDDG, ELASPIC, PoPMuSiC, AUTO-MUTE(ΔTM), AUTO-MUTE(ΔΔG), HoTMuSiC, and SDM are also included. Lastly, the SI 4 file contains a Finalized_exp_dgg.csv file including all the thermodynamic parameters 5 (ΔΔG, ΔΔH, ΔΔS, and ΔTM) derived from the fluorescence melting curve, as well as gel number that corresponds to each of the 51 mutants. Available at https://pubs.acs.org/doi/10.1021/acsomega.9b04105

SI 3-5. This folder contains bglb_apo.pdb, flags, mutant_file, and sub.sh file needed to execute Rosetta_ddg_monomer application that have been previously described in Kellogg 2011.1 (1) Kellogg, E. H.; Leaver-Fay, A.; Baker, D. Changes in Protein Structure and Stability. Proteins 2011, 79 (3), 830–838. https://doi.org/10.1002/prot.22921.Role. Available at https://pubs.acs.org/doi/10.1021/acsomega.9b04105