HIGH-THROUGHPUT ENZYMATICS: COMPARING CASPASES AND ENGINEERING GLYCOSIDE HYDROLASES WITH MICROFLUIDICS

by

Hridindu S. Roychowdhury

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

Doctor of Philosophy

(Integrated Program in Biochemistry)

at the

UNIVERSITY OF WISCONSIN–MADISON

2022

Date of final oral examination: 02/21/2022

The dissertation is approved by the following members of the Final Oral Committee:
Alessandro Senes, Professor, Biochemistry
Andrew Buller, Assistant Professor, Chemistry
Srivatsan Raman, Assistant Professor, Biochemistry
Philip A. Romero, Advisor, Assistant Professor, Biochemistry

ACKNOWLEDGMENTS

This dissertation could not have been completed without the immeasurable support of all the following people. Thank you to all of you.

My dear mother, father, Hiranya and Debasmita who have accompanied me my whole life—You are an inspiration to me and I could not be in this position without your steadfast guidance and support. And my dear sister, Indumati, who has accompanied me her whole life—Thank you for your humor, for your moral guidance, and for making sure my head doesn't get too big. I am glad to be able to go through life with you.

My phenomenal friends and colleagues within and without the Romero Lab: Juan, Leland, Eddie, Delia, Eddie, Taylor, Danro, Aidan, Adam, Dana, N8 Murray, Job, Mark, Jonathan, Zach—thank you for every beer, meal, laugh, and minute we shared.

My advisor, Phil—Thank you for your mentorship, patience, and everything you have taught me about the scientific process, and how to persevere through it, especially when the times get tough.

My comrades in Madison and abroad—I appreciate you keeping me in the fight.

Everyone at The Library—for holding it down.

Finally, my partner, Taryn McGinn Valley, who has accompanied me lockstep through my graduate tenure. Thank you for supporting me through difficult times, sharing in wonderful times, and for catalyzing my personal growth through every one of those. I look forward to the times yet to come.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF SUPPLEMENTARY FIGURES

ABSTRACT

This dissertation is concerned with the development and application of a novel high-throughput microlfuidic screening platform for use in enzyme engineering and deep mutational scanning. By designing the platform to maintain strict control over how long in vitro enzymatic reactions proceed for, we were able to generate detailed enough sequence-function maps of Caspase-3 and Caspase-7 that we could distinguish functionally unique sequence elements between them. We also utilized the platform to engineer a variant of the CMX glycoside hydrolase that has greatly improved resilience to the solvent, γ-valerolactone, which we hope will serve as a useful catalyst in industrial biofuel production. The methods developed here, which allow directly screening enzymes based on their activity, has broad applicability in enzyme engineering and also allows us to compare nearly identical enzymes for distinguishing features that could be leveraged for hyper-specific drug design.

Chapter 1

BACKGROUND AND INTRODUCTION

HSR wrote this chapter.

Portions of this chapter are adapted from the following manuscripts:

Roychowdhury R, Romero P A. *Microfluidic deep mutational scanning of the human executioner caspases reveals differences in structure and regulation.* Nature Cell Death Discovery. **2022**.

Roychowdhury R, Romero P A. *Deep Mutational Scanning of Yak rumen glycoside hydrolase reveals mutations conferring tolerance to industrial solvent γ-valerolactone.* In preparation. **2022**.

A brief overview of enzymes and their engineering

Proteins account for over half the dry weight of a typical *E. coli* cell, and comprise a similarly large fraction of cellular organisms across life. They crucially perform myriad structural functions, serve as signaling and regulatory molecules, and catalyze the chemical reactions required for life to exist.[1] Humans have been unknowingly exploiting and engineering such chemical reactions from before recorded history, using fermentative processes to brew, bake bread, and preserve food.[2,3] The first description of living cells and their biomolecules being central to the process, however, was not until the 19th century when physiologist Wilhelm Kuhne noticed some proteins catalyzing reactions—he named them enzymes.[4]

Our understanding of enzymes, their biophysical properties, physiological roles, and catalytic capabilities have since greatly expanded to the point that we are able to directly engineer enzymes for our own purposes. Altering enzyme substrate specificity, pH dependance, stability, solvent tolerances, and other properties has provided scientists a plethora of new and a adaptable biocatalysts that have paved the way for such famous technologies as polymerase chain reaction (PCR), luminescent biological reporters, and industrial biocatalysts.[5–7] **Figure 1.1** describes a timeline of notable events in protein and enzyme engineering, including the advent of directed evolution and contemporary computational design methods.

Broadly, there are two methodologies for engineering proteins and enzymes: directed evolution and rational design. Typically, researchers utilize both approaches. This disserta-

tion focuses on development and utilization of a droplet microfluidic platform that enables the directed evolution of enzymes in ultra high throughput, and also allows us to map and compare detailed sequence-function landscapes of highly similar enzymes.

Rational Approaches to Enzyme Engineering

Rational approaches to protein engineering requires leveraging an a priori understanding of the protein in question, often based in comprehensive structural knowledge or large phylogenetic data sets to inform site-directed mutagenesis. There are three main approaches to rational protein design: structural analysis, phylogenetic analysis, and, hyper-contemporarily, machine learning based computational approaches.

Structural and biophysical approaches to protein engineering. By analyzing protein structures derived from x-ray crystallography, nuclear magnetic resonance (NMR), or cryo electron microscopy (CryoEM), we gain a detailed and well informed understanding of the protein and the physical roles individual amino acids play. By observing the structures of enzyme active sites and interaction interfaces, protein engineers can make informed hypotheses regarding which mutations will alter protein function.[9] Altering key residues in an enzyme's substrate binding interface, for example, can lead to drastically altered specificity or kinetics. Similarly, by understanding the geometry and properties of an enzyme active site, one may completely alter the chemistry that enzyme is capable of catalyzing. The key shortcoming of structure-guided protein engineering is our own

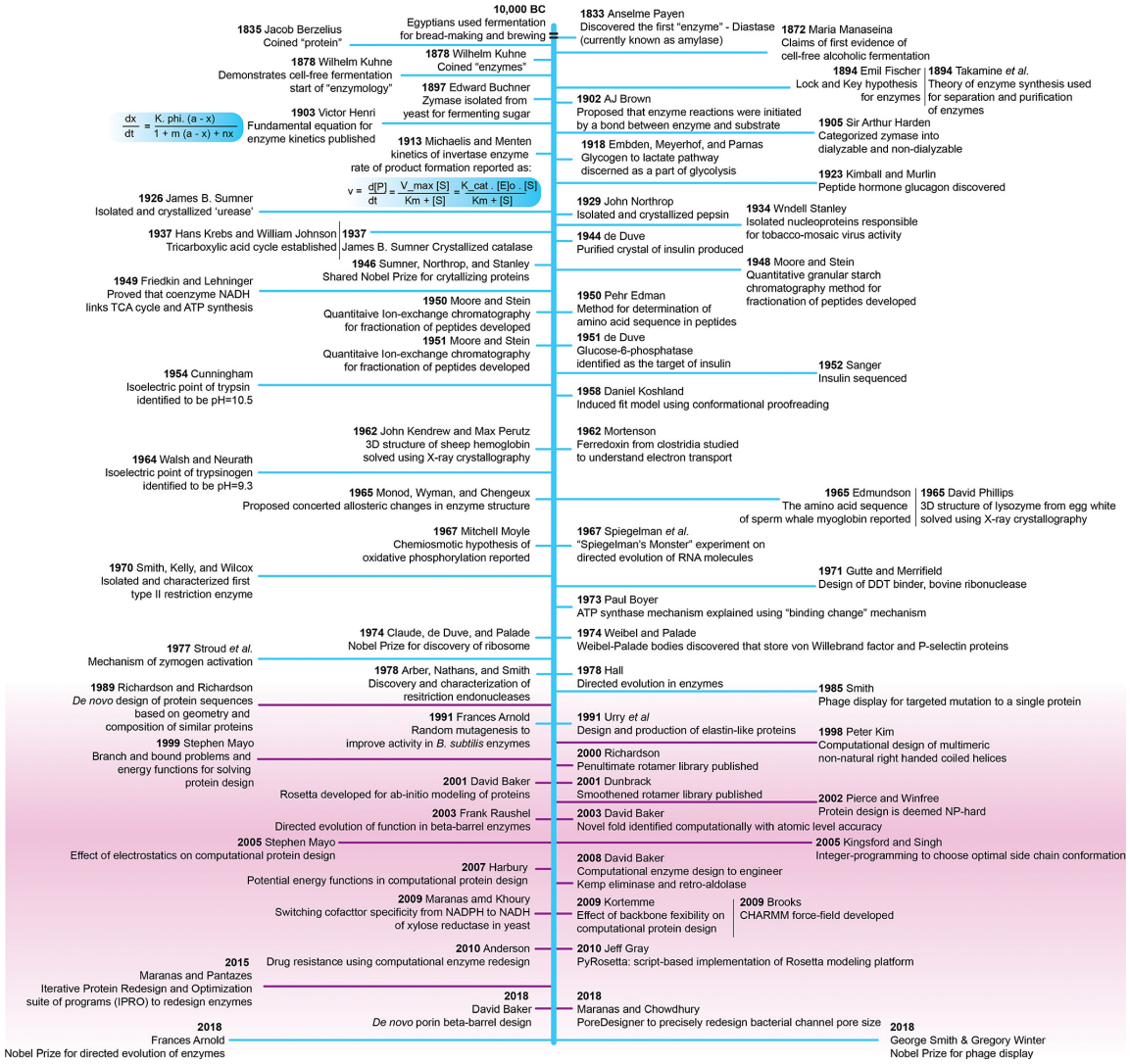


Figure 1.1: Timeline of protein engineering. Adapted from Chowdhury and Maranas' review of enzyme engineering, humanity's history of engineering enzymes from prehistory through Frances Arnold receiving the Nobel Prize for her work in directed evolution.[8] Purple lines represent computational efforts.

naïveté regarding the complex intramolecular interactions that produce key functional characteristics.[10–13]

Sophisticated force-field models, such as those used in Rosetta, have been successful in structure-guided protein engineering efforts, especially in altering ligand binding or improving overall protein stability.[14–17] More ambitious engineering goals, such as altering catalytic specificity or introducing novel catalytic mechanisms, have so far been less successful.[18,19] Rosetta's capabilities have nevertheless displayed constant and consistent improvement, and its prominence in the field will continue to grow.

Phylogenetic approaches to protein engineering. We can glean insights about a protein's sequence-function landscape from phylogenetic and evolutionary data.[20] For example, observing historic conservation in residues can inform our understanding of their contribution to the protein's overall function. With increasingly sophisticated statistical analysis, such as direct coupling analysis, we can even begin to infer not only residues' epistatic dependencies, but full allosteric pathways.[21] These analyses draw from limited subsets of sequence space; Inferences drawn from extant sequence-function relationships are largely relevant only to their native physiological contexts.

Computational approaches to protein engineering. Searching sequence space for functionally improved proteins is incredibly resource and labor intensive. It is in researchers' interest to narrow that space as much as possible using rational design approached to

maximize the likelihood of finding hits.[22–24] While the human brain is exceptionally well suited to pattern recognition, our own development of those faculties are optimized for identifying the juiciest mangoes and evading tigers[25]—not necessarily correlating protein sequence-function relationships.[13,26] By generating and pre-screening genetically diverse libraries in silico, researchers can reduce the labor and resource cost of engineering proteins.

Contemporary computational techniques, particularly Machine Learning (ML), can recognize protein-sequence function patterns to generate models that can ostensibly predict highly functional sequences.[27,28] Supervised learning approaches, where models are trained on labeled data mapping sequence to function, has enjoyed significant success in predicting novel peptides with improved or novel functions.[23,29,30] In supervised ML, an algorithm fits a function f using sequence features (often encoded in a matrix $\mathbf{X}$) to predict a labeled protein feature such as thermal stability, substrate specificity, or other kinetic parameter y. Fitting the function $f(\mathbf{X})$ can be as simple as linear regression, and can increase in complexity to include polynomial regression, logistic regression, support vector machines, all the way up to various neural networks. The performance of the model is estimated by k-fold cross validation. When cross validating a model, data is subdivided into training data, to estimate the parameters of $f(\mathbf{X})$, and test data to evaluate their accuracy. The k subdivisions of training/test data are iterated through and the parameters optimized by minimizing or maximizing a loss or profit function, such as R^2 for linear functions.[30]

Supervised ML approaches have some key drawbacks. The quality of the input training data has massive impacts on models. ML algorithms fundamentally compress data into

their most descriptive latent parameters, thus noise, uncertainty, and the quantity of data significantly affects how predictive the model is.[31] Choosing the correct model is also critical. Overly simple, under-fitted models are too general to usefully predict complex interactions. Overly complex and over-fitted models trained with too little validation data poorly predict novel sequences.[23,30] Typically, these concerns are mitigated by combining post hoc experimental validation, fine-tuning model hyper-parameters, and choosing effective cross validation divisions.[23,30] Whereas ML has been effectively used to predict properties such as activity, ligand-binding, and stability, models have been unable to predict complex global phenomenon such as solvent resistance.[30]

Generating high quality labeled data for supervised learning is challenging. It often relies on decades of research on any given family of proteins, effectively carrying out non-trivial high-throughput screening experiments, or otherwise having access to well-curated sequence-function databases. Gaussian process regression and other active learning methods are a popular strategy to optimize traversing through sequence-space efficiently.[32,33] Researchers have also turned to unsupervised and semi-supervised ML models trained on unlabeled data when engineering proteins. The massive amounts of sequencing data generated by the next-generation sequencing revolution have been used to pretrain models using autoencoders, Long-Short-Term-Memory, and transformers, drawing inspiration from natural language processing (NLP).[27,30,34,35] Such models depict sparse vector representations of proteins (called embeddings) which can be used in conjunction with labeled phylogenetic, structural, or HTS data to predict new libraries of functional proteins. Despite

the relative nascency of unsupervised and semi-supervised learning techniques in protein engineering, they have shown great success.[28]

Hopefully, as the iterative processes of model generation, library validation, and overall benchmarking continue, ML will serve to greatly augment our ability to engineer proteins and explore sequence-space. While the field is indeed young and many physical properties of proteins are yet to be be predicted by ML, we will only benefit from integrating it into our protein engineering pipelines.[28]

Directed Evolution Approaches to Enzyme Engineering

Directed evolution is comparatively straightforward compared to the rational techniques previously described, and does not mandate significant prior knowledge of protein sequence-structure-function relationships.[10,13,26] The strategy is deceptively simple: By mutagenizing a protein to create a library of variants and subjecting that library to a selection or functional screen, one is able to pick out variants of the protein with improved or changed function. By emulating the natural processes of evolution, one can rapidly and effectively alter an enzyme's properties. There are two major considerations when designing directed evolution pipelines: library design and screening techniques.

Comprehensively sampling protein sequence space is impossible. A small 100 residue peptide sequence exists in a combinatorial space of ~10^{130} possible peptides. The most liberal estimates of the number of fundamental particles in the observable universe is ~10^{97}—one would need 10^{33} universes to encode one unique peptide on each single particle.[36] It

seems like a massive waste of time to even try. One must explore the sequence landscape by starting with functional sequences and iteratively sampling sparse spaces surround it to accrue beneficial mutations. One such tactic is random mutagenesis. By using strategies such as error-prone PCR, random mutations are incorporated into the wild-type gene and putatively activating variants can be picked from the population.[37,38]

DNA recombination is another effective strategy to generate enzyme libraries. DNA recombination mimics sexual reproduction by shuffling fragments of functional parent protein. This technique leverages the fact that the parent sequences contribute putatively active sequence fragments to massively extend the assayed sequence space compared to creating libraries of single and double mutants.[39] Contemporarily, library design that maximize sequence diversity and also functionality are also informed by machine learning techniques, some of which I described above.

The major challenge is screening these libraries for functional variants. There are myriad strategies. With critical metabolically linked enzymes, one can apply evolutionary pressure by linking the enzyme's activity to cell survival, directly mirroring evolution. Selection strategies have been used to predict antibiotic resistance, evolve amino acid synthetases, recombinases, polymerases and many other enzymes central to cell viability.[40–43] Selections are typically incredibly stringent, and typically optimizing a protein requires numerous iterations. David Liu's group developed a technique called Phage Assisted Continuous Evolution (PACE), a powerful tool for accelerating the accumulations of activating mutations in turbidostatic bioreactors.[44] It has been used extensively for engineering recombinases

and polymerases; however, the system is not easily generalizable.[44,45]

For engineering enzymes that cannot be readily linked to survival, we must turn to other screening methods. Technologies for monitoring the conversion of substrate to product abound, including chromatography, mass spectrometry, and many spectroscopic techniques.[46–48] These techniques depend on assaying individual library members, and their throughput is usually limited to that of microtiter plates and graduate student perseverance. Employing such strategies in enzyme engineering usually requires small, targeted libraries based on a priori understanding of the structure-function relationships. Limiting the search space in that way excludes many potentially useful mutations.

Flow cytometry, specifically Fluorescence Activated Cell Sorting (FACS), provides an ultra-high-throughput platform capable of screening 10^8 variants per day and has also been extensively used to engineer proteins. For example, surface-display coupled FACS screens have been used to alter antibody specificities, and transcriptional coupling has been used to alter transcription factor specificity.[49,50] Successful and powerful as it is, FACS is limited to monitoring either intracellular reaction or cell-surface display based interactions[51,52]. FACS has a punk-rock cousin: In vitro compartmentalization techniques based on droplet microfluidic devices that allow direct observation of substrate conversion in microreactions at kilohertz frequencies.[53–55]

Microfluidic Enzyme Screening Techniques

The major focus of this dissertation is the development and implementation of a high-throughput droplet microfluidic device capable of generating detailed sequence-function maps to aid enzyme engineering and as a deep mutational scanning platform more broadly. Droplet microfluidics allows us to encapsulate billions of isolated in vitro picoliter-scale chemical reactions in a stable emulsion that is incredibly well suited for high-throughput screening.[56] In contrast to FACS and other methods previously described, our platform is capable of precisely controlling for how long an enzymatic reaction proceeds prior to measurement, effectively providing a snapshot of an enzyme variant's velocity V_0 instead of only measuring the total product formation or being limited to surface-display based techniques.

By screening enzymes based on velocity, we are able to distinguish subtle differences in sequence function relationships, heretofore impossible to capture. Having strict control of reaction time allows us to effectively differentiate significantly attenuated, but still folded, enzymes from highly active enzymes, providing a better picture of mutations that affect catalysis and not structure.[57] We demonstrate our platforms ability to find subtle differences between members of the highly similar caspase family, and demonstrate its ability to engineer solvent-resistant glycoside hydrolases under evolutionary pressure.

Comparing Caspases

Caspases are a ubiquitous family of cysteine proteases that play fundamental roles in programmed cell death and inflammation.[58] These enzymes have numerous ancillary roles in organismal development and homeostasis including cell differentiation, synaptic pruning, and cytokine processing.[59,60] In humans, there are twelve expressed members of the family, with Caspase-3, -6, -7, -8, and -9 primarily involved in apoptosis, and the others involved in pyroptosis and inflammation. All caspases have a conserved core proteolytic domain and a variable N-terminal domain that is involved with regulation of enzyme activity.[60]

Dysregulation of caspase activity is associated with cancer, neurodegeneration, vascular ischemia, and inflammatory diseases.[58,61,62] Consequently, these enzymes represent important therapeutic targets to treat a variety of human diseases.[61] However, despite caspases' central role in human biology and disease, every caspase-targeting drug candidate has failed to pass through clinical trials.[63,64] A key challenge for therapeutic development has been the caspase family's highly conserved proteolytic domain, which makes it difficult to selectively target one particular member and leads to off-target effects.[65] A deeper understanding of caspase structure, function, and regulation may eventually lead to small molecule modulators that selectively target members of the caspase family and open the door for novel therapeutics.[63,66,67]

In Chapter 2, we develop a high-throughput microfluidic platform for caspase screening

and apply it to systematically map sequence-function relationships in the human executioner caspases.[11,68] Our microfluidic system consists of a fully integrated lab-on-a-chip that combines the addition of a fluorogenic substrate, incubation of the enzyme reaction, and fluorescence measurement. Our microfluidic chip can perform kinetics-based screening on millions of caspase variants. We applied our screening system to perform deep mutational scanning (DMS) on caspase-3 (CASP3) and caspase-7 (CASP7). The DMS data displayed known and expected signatures of caspase structure and function, but also revealed important differences between CASP3 and CASP7 that may be related to allosteric regulation and protein stability. Future exploration of the differences between human caspases may lead to more targeted drug design efforts.

Engineering Glycoside Hydrolases

Humanity's destructive impact on global biospheres and ecologies by way of fossil fuel consumption cannot be understated. Lignocellulose biomass, a major byproduct of industrial agriculture, may be utilized as an alternative biofuel and as a precursor to industrially relevant chemicals.[69] Itself a biomass-derived product, the solvent γ-valerolactone (GVL) can be used to pretreat lignocellulosic biomass for further saccharification, wherein biomass is solvated in an 80% GVL and 20% water.[70–72] The pretreatment strips over 80% of the lignin away, maintains near complete cellulose retention, and nearly all the GVL can be recovered and reused. After fractionation, downstream enzymatic hydrolysis turning polysaccharides into sugars is seen to be the logical next step in biofuel production.[70,73]

RuCelA, isolated from the Yak gut microbiome, a promiscuous glycoside hydrolase with cellulase, mannase, and xylanase activity (earning the enzyme the moniker "CMX," which we use in this study) has been considered a strong candidate for enzymatically processing GVL treated products.[74] Predictably, having been evolved and optimized for lignocellulytic activity in the guts of yaks, we have found the enzyme's desired activity in the industrial conditions specified above to be deficient. Particularly, we have found that residual GVL leftover from fractionation halves CMX activity at concentrations as low as 3% volume/volume.

In chapter 3 we take the first steps in engineering CMX to better tolerate non-native conditions. We again use a fully-integrated microfluidic lab-on-a-chip to screen a library of CMX mutants in high-throughput while subject to a 3% GVL challenge. We identify an initial CMX variant with twice the activity of the wild type in the presence of 3% GVL. The microfluidic screening platform allows us to screen millions of in vitro isotemporal enzymatic reactions against the fluorogenic substrate resorufin cellobioside, providing us quantitative information on individual mutations' contributions to the enzyme sequence-function landscape.[68,75] We hope that with additional iterative rounds of evolution, we will soon evolve an enzyme well suited to the industrial decomposition of lignocellulose biomass.

References

[1] Neidhardt, F. C. *Escherichia coli and Salmonella: cellular and molecular biology*. 579.8 ESC (1996).

[2] McGovern, P. & Mondavi, R. Stone age wine. *Ancient Wine: the search for the origins of viniculture. Princeton University Press, Princeton, New Jersey* 1–15 (2003).

[3] McGovern, P. E. *et al.* Fermented beverages of pre-and proto-historic china. *Proceedings of the National Academy of Sciences* **101**, 17593–17598 (2004).

[4] Alba-Lois, L. & Segal-Kischinevzky, C. Yeast fermentation and the making of beer and wine. *Nature Education* **3**, 17 (2010).

[5] Hall, B. G. Number of mutations required to evolve a new lactase function in escherichia coli. *Journal of bacteriology* **129**, 540–543 (1977).

[6] Saiki, R. K. *et al.* Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230**, 1350–1354 (1985).

[7] Eun, H.-M. *Enzymology primer for recombinant DNA technology* (Elsevier, 1996).

[8] Chowdhury, R. & Maranas, C. D. From directed evolution to computational enzyme engineering—A review. *AIChE Journal* **66**, e16847 (2020). URL `https://onlinelibrary.wiley.com/doi/full/10.1002/aic.16847https:`

//onlinelibrary.wiley.com/doi/abs/10.1002/aic.16847https://aiche.onlinelibrary.wiley.com/doi/10.1002/aic.16847.

[9] Bilal, M. & Iqbal, H. Tailoring multipurpose biocatalysts via protein engineering approaches: a review. *Catalysis Letters* **149**, 2204–2217 (2019).

[10] Firnberg, E., Labonte, J. W., Gray, J. J. & Ostermeier, M. A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape. *Molecular Biology and Evolution* **31**, 1581–1592 (2014). URL https://academic.oup.com/mbe/article/31/6/1581/2925654.

[11] Fowler, D. M., Stephany, J. J. & Fields, S. Measuring the activity of protein variants on a large scale using deep mutational scanning (2014). URL http://www.nature.com/doifinder/10.1038/nprot.2014.153.

[12] Fowler, D. M. & Fields, S. Deep mutational scanning: A new style of protein science (2014). URL http://www.nature.com/articles/nmeth.3027.

[13] Francis, J. & Hansche, P. Directed evolution of metabolic pathways in microbial populations. i. modification of the acid phosphatase ph optimum in s. cerevisiae. *Genetics* **70**, 59–73 (1972).

[14] Cui, D. *et al.* A computational strategy for altering an enzyme in its cofactor preference to nad (h) and/or nadp (h). *The FEBS Journal* **282**, 2339–2351 (2015).

[15] García-Guevara, F., Bravo, I., Martínez-Anaya, C. & Segovia, L. Cofactor specificity switch in shikimate dehydrogenase by rational design and consensus engineering. *Protein Engineering, Design and Selection* **30**, 533–541 (2017).

[16] Khoury, G. A. *et al.* Computational design of candida boidinii xylose reductase for altered cofactor specificity. *Protein Science* **18**, 2125–2138 (2009).

[17] Lehmann, A. & Saven, J. G. Computational design of four-helix bundle proteins that bind nonbiological cofactors. *Biotechnology progress* **24**, 74–79 (2008).

[18] Grisewood, M. J. *et al.* Computational redesign of acyl-acp thioesterase with improved selectivity toward medium-chain-length fatty acids. *ACS catalysis* **7**, 3837–3849 (2017).

[19] Chen, C.-Y., Georgiev, I., Anderson, A. C. & Donald, B. R. Computational structure-based redesign of enzyme activity. *Proceedings of the National Academy of Sciences* **106**, 3764–3769 (2009).

[20] Süel, G. M., Lockless, S. W., Wall, M. A. & Ranganathan, R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature structural biology* **10**, 59–69 (2003).

[21] Raman, A. S., White, K. I. & Ranganathan, R. Origins of allostery and evolvability in proteins: a case study. *Cell* **166**, 468–480 (2016).

[22] Fox, R. J. *et al.* Improving catalytic function by prosar-driven enzyme evolution. *Nature biotechnology* **25**, 338–344 (2007).

[23] Yang, K. K., Wu, Z. & Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nature methods* **16**, 687–694 (2019).

[24] Song, H. & Raskutti, G. PUlasso: High-Dimensional Variable Selection With Presence-Only Data. *Journal of the American Statistical Association* **115**, 334–347 (2020). URL `https://www.tandfonline.com/action/journalInformation?journalCode=uasa20`. `1711.08129`.

[25] Voulodimos, A., Doulamis, N., Doulamis, A. & Protopapadakis, E. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience* **2018** (2018).

[26] Romero, P. A. & Arnold, F. H. Exploring protein fitness landscapes by directed evolution. *Nature reviews Molecular cell biology* **10**, 866–876 (2009).

[27] Hie, B. L. & Yang, K. K. Adaptive machine learning for protein engineering. *Current opinion in structural biology* **72**, 145–152 (2022).

[28] Wittmann, B. J., Johnston, K. E., Wu, Z. & Arnold, F. H. Advances in machine learning for directed evolution. *Current opinion in structural biology* **69**, 11–18 (2021).

[29] Li, G., Dong, Y. & Reetz, M. T. Can machine learning revolutionize directed evolution of selective enzymes? *Advanced Synthesis & Catalysis* **361**, 2377–2386 (2019).

[30] Siedhoff, N. E., Schwaneberg, U. & Davari, M. D. Machine learning-assisted enzyme engineering. *Methods in Enzymology* **643**, 281–315 (2020).

[31] Ghahramani, Z. Probabilistic machine learning and artificial intelligence. *Nature* **521**, 452–459 (2015).

[32] Bedbrook, C. N. *et al.* Machine learning-guided channelrhodopsin engineering enables minimally invasive optogenetics. *Nature methods* **16**, 1176–1184 (2019).

[33] Romero, P. A., Krause, A. & Arnold, F. H. Navigating the protein fitness landscape with gaussian processes. *Proceedings of the National Academy of Sciences* **110**, E193–E201 (2013).

[34] Yang, K. K., Wu, Z., Bedbrook, C. N. & Arnold, F. H. Learned protein embeddings for machine learning. *Bioinformatics* **34**, 2642–2648 (2018).

[35] Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118** (2021).

[36] Munafo, R. Notable properties of specific numbers (2019).

[37] Landwehr, M. *et al.* Enantioselective α-hydroxylation of 2-arylacetic acid derivatives and buspirone catalyzed by engineered cytochrome p450 bm-3. *Journal of the American Chemical Society* **128**, 6058–6059 (2006).

[38] Bloom, J. D. *et al.* Evolution favors protein mutational robustness in sufficiently large populations. *BMC Biology* **5**, 1–21 (2007). URL

https://link.springer.com/articles/10.1186/1741-7007-5-29https://link.springer.com/article/10.1186/1741-7007-5-29.

[39] Cirino, P. C. & Qian, S. Protein engineering as an enabling tool for synthetic biology. In *Synthetic Biology*, 23–42 (Elsevier, 2013).

[40] MacBeath, G., Kast, P. & Hilvert, D. Redesigning enzyme topology by directed evolution. *Science* **279**, 1958–1961 (1998).

[41] Orencia, M. C., Yoon, J. S., Ness, J. E., Stemmer, W. P. & Stevens, R. C. Predicting the emergence of antibiotic resistance by directed evolution and structural analysis. *Nature structural biology* **8**, 238–242 (2001).

[42] Wilson, D. S., Keefe, A. D. & Szostak, J. W. The use of mrna display to select high-affinity protein-binding peptides. *Proceedings of the National Academy of Sciences* **98**, 3750–3755 (2001).

[43] Ali, M., Ishqi, H. M. & Husain, Q. Enzyme engineering: Reshaping the biocatalytic functions. *Biotechnology and bioengineering* **117**, 1877–1894 (2020).

[44] Esvelt, K. M., Carlson, J. C. & Liu, D. R. A system for the continuous directed evolution of biomolecules. *Nature* **472**, 499–503 (2011).

[45] Dickinson, B. C., Packer, M. S., Badran, A. H. & Liu, D. R. A system for the continuous directed evolution of proteases rapidly reveals drug-resistance mutations. *Nature communications* **5**, 1–8 (2014).

[46] Longwell, C. K., Labanieh, L. & Cochran, J. R. High-throughput screening technologies for enzyme engineering. *Current opinion in biotechnology* **48**, 196–202 (2017).

[47] Luetz, S., Giver, L. & Lalonde, J. Engineered enzymes for chemical production. *Biotechnology and bioengineering* **101**, 647–653 (2008).

[48] Xiao, H., Bao, Z. & Zhao, H. High throughput screening and selection methods for directed enzyme evolution. *Industrial & engineering chemistry research* **54**, 4011–4020 (2015).

[49] Lauterjung, K., Liu, X., Henderson, K., Raman, V. & Record, T. In vivo effects of discriminator sequences on transcription initiation in e. coli. *Biophysical Journal* **114**, 248a (2018).

[50] Yang, G. & Withers, S. G. Ultrahigh-throughput facs-based screening for directed enzyme evolution. *ChemBioChem* **10**, 2704–2715 (2009).

[51] Stavrakis, S., Holzner, G., Choo, J. & DeMello, A. High-throughput microfluidic imaging flow cytometry. *Current opinion in biotechnology* **55**, 36–43 (2019).

[52] Qin, Y. *et al.* A fluorescence-activated single-droplet dispenser for high accuracy single-droplet and single-cell sorting and dispensing. *Analytical chemistry* **91**, 6815–6819 (2019).

[53] Griffiths, A. D. & Tawfik, D. S. Miniaturising the laboratory in emulsion droplets. *Trends in biotechnology* **24**, 395–402 (2006).

[54] Mair, P., Gielen, F. & Hollfelder, F. Exploring sequence space in search of functional enzymes using microfluidic droplets. *Current Opinion in Chemical Biology* **37**, 137–144 (2017).

[55] Basova, E. Y. & Foret, F. Droplet microfluidics in (bio) chemical analysis. *Analyst* **140**, 22–38 (2015).

[56] Romero, P. A., Tran, T. M. & Abate, A. R. Dissecting enzyme function with microfluidic-based deep mutational scanning. *Proceedings of the National Academy of Sciences* **112**, 7159–7164 (2015).

[57] Roychowdury, H. & Romero, P. A. Microfluidic deep mutational scanning of the human executioner caspases reveals differences in structure and regulation. *Cell Death Discovery* **8**, 1–8 (2022).

[58] Shalini, S., Dorstyn, L., Dawar, S. & Kumar, S. Old, new and emerging functions of caspases (2015). URL `https://www.nature.com/cdd/journal/v22/n4/pdf/cdd2014216a.pdf`.

[59] Graham, R. K. *et al.* Cleavage at the Caspase-6 Site Is Required for Neuronal Dysfunction and Degeneration Due to Mutant Huntingtin. *Cell* **125**, 1179–1191 (2006). URL `https://www.sciencedirect.com/science/article/pii/S0092867406005587`.

[60] Fuentes-Prior, P. & Salvesen, G. S. The protein structures that shape caspase activity,

specificity, activation and inhibition. *Biochem. J* **384**, 201–232 (2004). URL http://www.biochemj.org/content/ppbiochemj/384/2/201.full.pdf.

[61] MacKenzie, S. H., Schipper, J. L. & Clark, A. C. The potential for caspases in drug discovery (2010). URL http://www.ncbi.nlm.nih.gov/pubmed/20812148http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3289102.

[62] McIlwain, D. R., Berger, T. & Mak, T. W. Caspase functions in cell death and disease. *Cold Spring Harbor perspectives in biology* **5**, a008656 (2013). URL http://www.ncbi.nlm.nih.gov/pubmed/23545416http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3683896.

[63] Häcker, H. G., Sisay, M. T. & Gütschow, M. Allosteric modulation of caspases (2011). URL http://www.sciencedirect.com/science/article/pii/S0163725811001604.

[64] Krishna Deepak, R. N., Abdullah, A., Talwar, P., Fan, H. & Ravanan, P. Identification of FDA-approved drugs as novel allosteric inhibitors of human executioner caspases. *Proteins: Structure, Function and Bioinformatics* **86**, 1202–1210 (2018). URL https://onlinelibrary.wiley.com/doi/full/10.1002/prot.25601https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.25601https://onlinelibrary.wiley.com/doi/10.1002/prot.25601.

[65] Agniswamy, J., Fang, B. & Weber, I. T. Conformational similarity in the activation of

caspase-3 and -7 revealed by the unliganded and inhibited structures of caspase-7. *Apoptosis* **14**, 1135–1144 (2009). URL http://www.pymol.orghttp://link.springer.com/10.1007/s10495-009-0388-9.

[66] Kudelova, J., Fleischmannova, J., Adamova, E. & Matalova, E. Pharmacological caspase inhibitors: research towards therapeutic perspectives. *Journal of physiology and pharmacology : an official journal of the Polish Physiological Society* **66**, 473–82 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/26348072.

[67] Hardy, J. A., Lam, J., Nguyen, J. T., O'Brien, T. & Wells, J. A. Discovery of an allosteric site in the caspases. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 12461–6 (2004). URL http://www.ncbi.nlm.nih.gov/pubmed/15314233http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC514654.

[68] Romero, P. A., Tran, T. M. & Abate, A. R. Dissecting enzyme function with microfluidic-based deep mutational scanning. *Proc. Nat. Adad. Sci.* **112**, 7159–7164 (2015). URL http://www.pnas.org/content/112/23/7159.full.pdf.

[69] Serrano-Ruiz, J. C., Luque, R. & Sepúlveda-Escribano, A. Transformations of biomass-derived platform molecules: from high added-value chemicals to fuelsvia aqueous-phase processing. *Chem. Soc. Rev.* **40**, 5266–5281 (2011). URL http://dx.doi.org/10.1039/C1CS15131B.

[70] Shuai, L., Questell-Santiago, Y. M. & Luterbacher, J. S. A mild biomass pretreatment using γ-valerolactone for concentrated sugar production. *Green Chemistry* **18**, 937–943 (2016). URL https://pubs.rsc.org/en/content/articlehtml/2016/gc/c5gc02489ghttps://pubs.rsc.org/en/content/articlelanding/2016/gc/c5gc02489g.

[71] Mellmer, M. A., Martin Alonso, D., Luterbacher, J. S., Gallo, J. M. R. & Dumesic, J. A. Effects of γ-valerolactone in hydrolysis of lignocellulosic biomass to monosaccharides. *Green Chemistry* **16**, 4659–4662 (2014).

[72] Luterbacher, J. S. *et al.* Nonenzymatic sugar production from biomass using biomass-derived γ-valerolactone. *Science* **343**, 277–280 (2014).

[73] Chaturvedi, V. & Verma, P. An overview of key pretreatment processes employed for bioconversion of lignocellulosic biomass into biofuels and value added products. *3 Biotech* **3**, 415–431 (2013). URL https://doi.org/10.1007/s13205-013-0167-8.

[74] Chang, L. *et al.* Characterization of a bifunctional xylanase/endoglucanase from yak rumen microorganisms. *Applied Microbiology and Biotechnology* **90**, 1933–1942 (2011). URL https://link.springer.com/article/10.1007/s00253-011-3182-x.

[75] Song, H., Bremer, B. J., Hinds, E. C., Raskutti, G. & Romero, P. A. Inferring Protein Sequence-Function Relationships with Large-Scale Positive-Unlabeled Learning. *Cell Systems* **12**, 92–101.e8 (2021).

Chapter 2

MICROFLUIDIC DEEP MUTATIONAL SCANNING OF THE HUMAN EXECUTIONER CASPASES REVEALS DIFFERENCES IN STRUCTURE AND REGULATION

HSR designed and conducted experiments, interpreted results, and analyzed data; HSR and PAR wrote the manuscript.

This chapter has been accepted for publication:

Roychowdhury R, RomeroPA. *Microfluidic deep mutational scanning of the human executioner caspases reveals differences in structure and regulation .* Nature Cell Death Discovery. **2022**.

Abstract

The human caspase family comprises 12 cysteine proteases that are centrally involved in cell death and inflammation responses. The members of this family have conserved sequences and structures, highly similar enzymatic activities and substrate preferences, and overlapping physiological roles. In this paper, we present a deep mutational scan of the executioner caspases CASP3 and CASP7 to dissect differences in their structure, function, and regulation. Our approach leverages high-throughput microfluidic screening to analyze hundreds of thousands of caspase variants in tightly controlled *in vitro* reactions. The resulting data provides a large-scale and unbiased view of the impact of amino acid substitutions on the proteolytic activity of CASP3 and CASP7. We use this data to pinpoint key functional differences between CASP3 and CASP7, including a secondary internal cleavage site, CASP7 Q196 that is not present in CASP3. Our results will open avenues for inquiry in caspase function and regulation that could potentially inform the development of future caspase-specific therapeutics.

Introduction

Caspases are a ubiquitous family of cysteine proteases that play fundamental roles in programmed cell death and inflammation.[1] These enzymes have numerous ancillary roles in organismal development and homeostasis including cell differentiation, synaptic pruning, and cytokine processing.[2,3] In humans, there are twelve expressed members of the family,

with Caspase-3, -6, -7, -8, and -9 primarily involved in apoptosis, and the others involved in pyroptosis and inflammation1. All caspases have a conserved core proteolytic domain and a variable N-terminal domain that is involved with regulation of enzyme activity.[3]

Dysregulation of caspase activity is associated with cancer, neurodegeneration, vascular ischemia, and inflammatory diseases.[1,4,5] Consequently, these enzymes represent important therapeutic targets to treat a variety of human diseases.[4] However, despite their central role in human biology and disease, every caspase-targeting drug candidate has failed to pass through clinical trials.[6,7] A key challenge for therapeutic development has been the caspase family's highly conserved proteolytic domain, which makes it difficult to selectively target one particular member and leads to off-target effects.[8] A deeper understanding of caspase structure, function, and regulation may eventually lead to small molecule modulators that selectively target members of the caspase family and open the door for novel therapeutics.[6,9,10]

In this work, we develop a high-throughput microfluidic platform for caspase screening and apply it to systematically map sequence-function relationships in the human executioner caspases.[11,12] Our microfluidic system consists of a fully integrated lab-on-a-chip that combines the addition of a fluorogenic substrate, incubation of the enzyme reaction, and fluorescence measurement. Our microfluidic chip can perform kinetics-based screening on millions of caspase variants. We applied our screening system to perform deep mutational scanning (DMS) on caspase-3 (CASP3) and caspase-7 (CASP7). The DMS data displayed known and expected signatures of caspase structure and function, but also revealed im-

portant differences between CASP3 and CASP7 that may be related to allosteric regulation and protein stability. Future exploration of the differences between human caspases may lead to more targeted drug design efforts.

Results

A microfluidic platform for ultra-high-throughput screening of caspases High-throughput screening is an important tool for studying protein structure and function.[12,13] Caspases are challenging to screen because their activity cannot be readily linked cell growth or cellular fluorescence. Furthermore, caspases' high catalytic rates make any cell-based assay difficult because the proteolytic cleavage reactions occur on significantly faster timescales than cell growth or fluorescent protein production. We developed a droplet microfluidic platform capable of in vitro, kinetics-based screening of millions of caspase variants.

Our microfluidic system encapsulates single *E. coli* cells, each expressing a unique caspase variant, into ~10 picoliter microdroplets that contain cell lysis reagents and a fluorogenic peptide substrate (**Fig 2.1a**). The droplets physically separate each cell and allow enzyme reactions to proceed in isolation. After encapsulation, the cells quickly lyse, releasing the expressed caspase and allowing it to interact with the substrate. The droplets are then incubated in an on-chip continuous flow reactor for ~3 minutes to allow the reaction to proceed. We found these short incubation times were necessary to separate highly active caspases from variants with severely attenuated activity. After incubation, each droplet is scanned with a laser fluorimeter and droplets displaying high fluorescence signals are

sorted for downstream analysis. Our microfluidic platform is capable of screening 360,000 caspase variants per hour, while consuming only ~100 μL of assay reagents.

We tested the ability of our emulsion-based assay to distinguish active CASP3 from an inactive D175A mutant (**Fig 2.1cd**). We encapsulated cells expressing each variant and analyzed the droplets using fluorescence microscopy. Droplets that contained active CASP3 displayed a strong fluorescence signal, while droplets with the inactive mutant had no measurable fluorescence. We next tested our assay on-chip using the integrated laser fluorimeter. The active enzyme was easily distinguished from the inactive mutant, with the average CASP3 droplet signal being at least 5 fold greater than the inactive mutant (**Supp Fig 2.1**). We next evaluated our microfluidic system's to enrich active caspases from a mixed variant population. We performed a mock sorting experiment by combining active CASP3 with a tenfold excess of an inactive empty plasmid control. We ran this mixed control population through our microfluidic system and sorted droplets with high fluorescence values. We then analyzed the proportion of active CASP3 versus empty plasmid by agarose gel electrophoresis (**Supp Fig 2.1c**). We found the initial population contained 9% active CASP3, as expected, and the sorted population contained 95% active CASP3 (**Fig 2.1d**). These results indicate that our system can enrich active caspases by at least 10 fold, which is ample for high-throughput screening.

Deep mutational scanning of the human executioner caspases Caspases 3, 6, and 7 are referred to as the executioner caspases because they perform the large-scale cellular

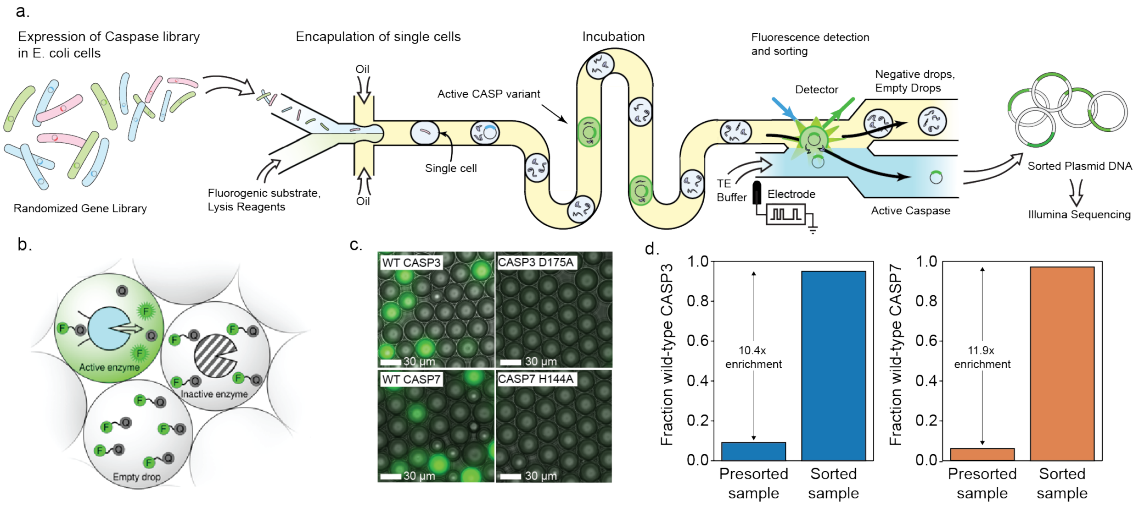


Figure 2.1: A droplet microfluidic platform for screening caspases.. **a**, A schematic of our microfluidic screening system. A dilute suspension of *E. coli* expressing caspase variants are injected into a microfluidic device and individual cells are encapsulated into microdroplets containing lysis reagents and a fluorogenic caspase substrate. The cells are lysed, the enzyme reaction is incubated on-chip, and the fluorescence of each droplet is analyzed using a laser. The fluorescent droplets are then sorted by electrocoalescence with an aqueous stream that collects the sorted plasmids for downstream analysis. **b**,Droplets containing active caspase variants will fluoresce, whereas empty droplets and droplets containing inactive caspases will not. **c**, Microscopy images of droplets containing active WT CASP3 and WT CASP7 display strong green fluorescence, while droplets containing the inactive CASP3 D175A and CASP7 H144A variants remain dark. **d**, Results of a mock screen demonstrate over 10-fold enrichment of active CASP3 and CASP7.

proteolysis that leads to apoptosis. These enzymes share similar in vitro substrate preference, however have been implicated in nonredundant cellular roles that cannot be fully explained by either structural differences or protein expression levels.[14–16] It is likely that subtle differences in their primary amino acid sequence may explain their in vivo and in vitro functional profiles.

We leveraged our microfluidic screening platform to systematically map sequence-function relationships for CASP3 and CASP7. We generated CASP3 and CASP7 libraries using error-prone PCR. These libraries contained 2-4 amino acid substitutions per variant and approximately 25% of these variants were active caspases. We screened these CASP3 and CASP7 libraries for active caspases using our microfluidic platform. We screened each library in triplicate to evaluate the reproducibility of our methods and to ensure the robustness of our results. For each screening run, we analyzed over 1.5 million caspase variants on average and sorted $4 * 10^5 - 7 * 10^5$ active variants for downstream DNA sequencing analysis (Supp Table 1).

We verified the sorted caspase variants were active enzymes by retransforming the genes into *E. coli* and assaying individual clones in a plate-based format. The initial unsorted libraries were 20-25% functional, while the sorted libraries were 60-90% functional, indicating strong enrichment of functional sequences (**Fig 2.2a**). We then sequenced all six sorted samples and their corresponding initial unsorted libraries using Illumina sequencing. The data displayed excellent reproducibility across the three experimental replicates for CASP7 and two experimental replicates for CASP3 (**Supp Fig 2.2**). The third CASP3 replicate

displayed poor agreement with the first two replicates (**Supp Fig 2.2**), and was excluded from further analysis. The third CASP3 replicate had significant false positive sequences as indicated by its high fraction of inactive sequences (**Fig 2.2a**).

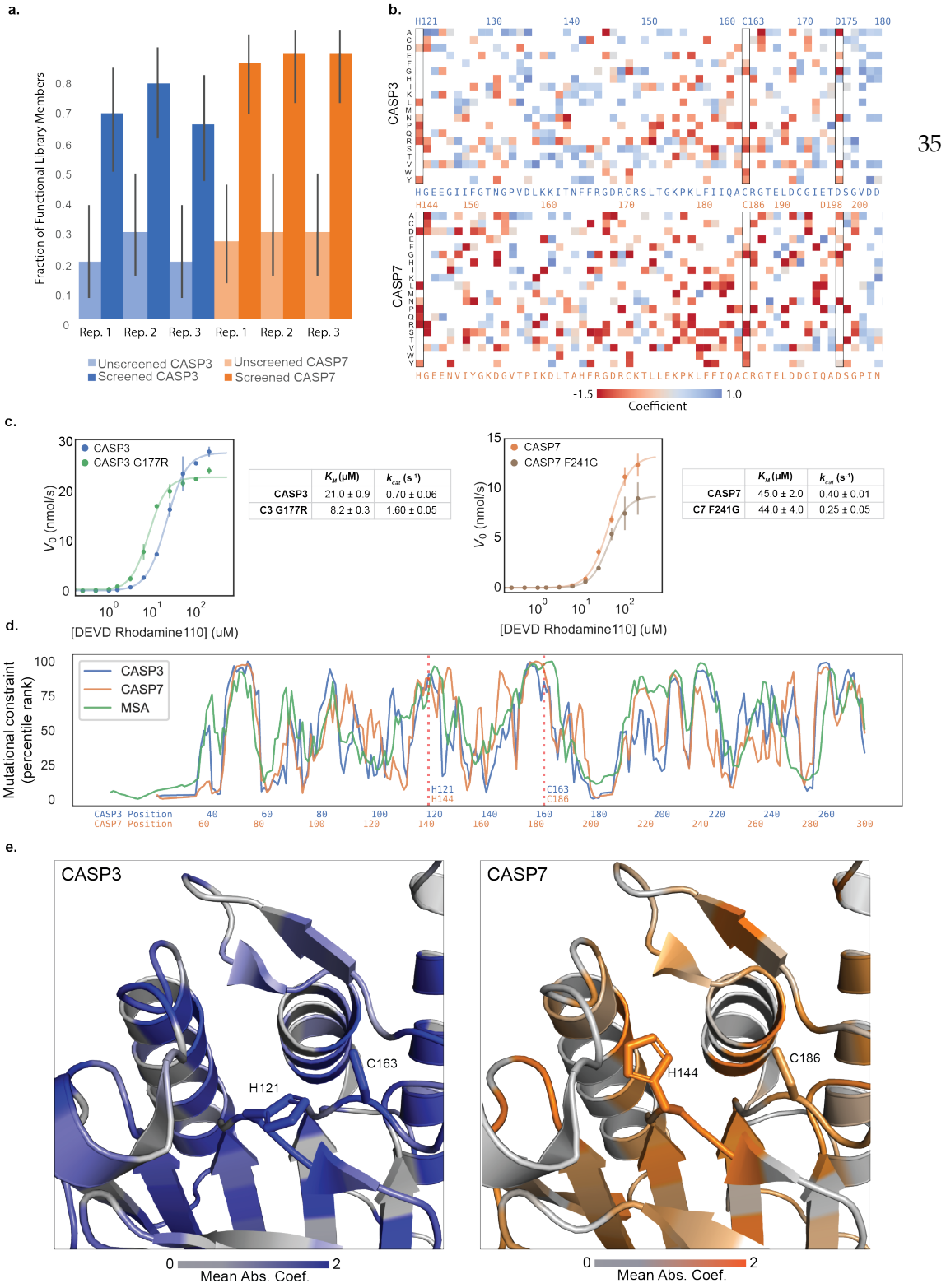
We used our deep mutational scanning data to build large-scale maps describing how individual mutations affect CASP3 and CASP7 activity (**Fig 2.2b**). These maps display expected mutational patterns for both caspases. Substitution of the active site cysteine and histidine residues is highly deleterious. Mutations to large aromatic residues in the hydrophobic core are not tolerated, whereas polar substitutions on the surface of the protein and chemically conservative mutations are generally more lenient. The internal processing sites D175 in CASP3 and D198 in CASP7, which are essential for maturation of zymogenic caspases to mature proteases, are also intolerant to mutation.

In addition to corroborating known and expected mutational patterns, our data also revealed new mutations that appear to enhance caspase activity. Our DMS analysis identified G177R as an activating mutation in CASP3. To validate this finding, we constructed, expressed, and measured CASP3 G177R's enzyme kinetic properties. CASP3 G177R's catalytic efficiency (k_{cat}/K_m) is over two-fold higher than wild-type CASP3. G177 is distant from CASP3's active site, but is located adjacent to the internal zymogen processing site D175. Mutations at position 177 could enhance enzyme activity through allosteric or conformational rearrangements related to CASP3's native activation mechanisms. CASP7 F241G was another putative activating mutation identified from our DMS analysis. Further kinetic characterization of this mutant found it was an active caspase, but actually had a

lower turnover number (k_{cat}) relative to wild-type CASP7. This discrepancy between the bulk assay and the droplet screen could be the result of enzyme expression level, since our droplet assay considers total activity that is not normalized to enzyme concentration All enzyme kinetic measurements are summarized in Supplemental Table 2.

We aggregated the individual mutational effects to obtain the mutational tolerance of each position in CASP3 and CASP7's primary sequence. This mutational tolerance is related to a site's importance for caspase function and allows us to analyze broader sequence and structural features. The site-wise mutational tolerance profiles of CASP3 and CASP7 are generally very similar, and also agree closely with profiles generated from a multiple sequence alignment (MSA) of natural caspases (**Fig 2.2c**). The beta-sheets that comprise the proteins' core are less mutable than the exterior helices, and the active site is evolutionarily conserved in the MSA and was also seen to be immutable in our deep mutational scan.

Contrasting mutational profiles reveals functional differences across executioner caspases Humans possess 12 separate caspases that all diverged from a common ancestor and share the same structurally conserved proteolytic domain. Despite their highly similar structure and biochemical activity, each caspase's regulation and cellular targets are unique and confer numerous non-redundant physiological roles. We explored our CASP3 and CASP7 sequence-function profiles to better understand functional differences between highly similar members of the caspase family.



	K_M (μM)	k_{cat} (s^{-1})
CASP3	21.0 ± 0.9	0.70 ± 0.06
C3 G177R	8.2 ± 0.3	1.60 ± 0.05

	K_M (μM)	k_{cat} (s^{-1})
CASP7	45.0 ± 2.0	0.40 ± 0.01
C7 F241G	44.0 ± 4.0	0.25 ± 0.05

Figure 2.2: Deep mutational scanning of CASP3 and CASP7.. a, Enrichment of active caspase variants in screened libraries. The fraction of active caspase variants was measured in a plate-based assay before and after screening. The error bars represent the 95% binomial proportion confidence intervals. All replicates showed significant enrichment. **b**, A heatmap of mutational coefficients surrounding the active site of CASP3 and CASP7. Mutations that are red have negative coefficients, corresponding to deleterious mutations. Mutations that are blue are positive and are either neutral or activating mutations. White boxes are mutations that did not appear on our DMS analysis. The outlined columns highlight the active site histidine and cysteine residues, as well as the internal aspartate where zymogen maturation occurs. **c**, Kinetic analysis of putatively activating mutations, CASP3 G177R and CASP7 F241G. Kinetic parameters were estimated from the Hill equation.**d**, The mutational tolerance of CASP3, CASP7, and the caspase family multiple sequence alignment (MSA) across sequence positions. The mutational tolerance at each position was calculated as the mean absolute value of all mutation coefficients at that position and plotted as a percentile rank. **e** The three-dimensional structures of the CASP3 and CASP7 active sites with their mutational tolerance scores mapped onto the structure. The active site residues are labeled and are strongly colored, indicating low tolerance to mutation.

We compared the mutational profiles of CASP3 and CASP7 to identify sites that display differing mutational tolerance and may have functionally diverged during caspase evolution and specialization (**Fig 2.3ab**). One notable sequence position was E173 in CASP3 and the equivalent residue Q196 in CASP7 (**Fig 2.3c**). CASP3 can tolerate any substitution at this position, whereas CASP7 can only accept substitution to glutamic acid. We verified this finding by performing enzyme kinetics measurements on CASP7 Q196A (**Fig 2.3d**). We found CASP7 Q196A had significantly diminished catalytic efficiency relative to wild-type CASP7. Intriguingly, Q196 is a known important regulatory site in CASP7 that is cleaved by Cathepsin G to activate procaspase-7.[17] While Cathepsin G is not present in our *E. coli*-based screen, it is possible that CASP7 can self-activate at this site and amino acid substitutions at this site reduce the pool of active enzyme.

Another differing sequence region were the adjacent sites H56/K57 in CASP3 and the corresponding D79/K80 in CASP7 (**Fig 2.3c**). Our analysis indicates that CASP3 residues H56 and K57 are completely insensitive to mutation, except for H56P. In CASP7, D79 and K80 show complete mutational intolerance, with all substitutions being deleterious. These residues are located in a solvent exposed loop that displays identical conformations between CASP3 and CASP7 and is located near the substrate binding site. Inspection of the crystal structures reveal that CASP7 has an extensive salt-bridge network in this region, while CASP3 does not. Presumably, mutations in CASP7 disrupt the salt-bridge network and lead to an altered conformation or destabilization in the protein structure.

A final site to note was I160 in CASP3 and I183 in CASP7 that are located within a

beta-sheet in the core of the enzyme (**Fig 2.3c**). I160T and I160S are well tolerated in CASP3, but CASP7 cannot tolerate any substitutions at I183. The packing environment of I160 and I183 are identical in the crystal structures, with the neighboring side-chains matching with sub-angstrom alignment. It is possible that substitutions at these sites are always destabilizing, but CASP3 has additional stabilization from elsewhere in the protein to permit these destabilizing core mutations.

Discussion

Caspases play a key role in numerous biological processes that are important for human health and disease. A deeper understanding of caspase structure, function, and regulation could open the door to novel therapeutic approaches.[4,7,18] In this work, we performed deep mutational scanning on CASP3 and CASP7 to reveal differences between highly similar members of the caspase family.

This work was enabled by our high-throughput droplet microfluidic screening platform that analyzes over 300,000 variants per hour in a highly controlled in vitro reaction environment.[11,19] Our device allowed us to have strict control over how long the proteolytic reactions were allowed to occur, which allowed us to more effectively differentiate caspase variants with altered activity. Cell-based assays that rely on proteolytic reporters and fluorescence-activated cell sorting (FACS) occur on much longer timescales and thus cannot distinguish WT-like activity from variants with severely diminished activity.[20] Even catalytically "dead" active site mutants such as CASP3 D175A display enough enzyme activity to

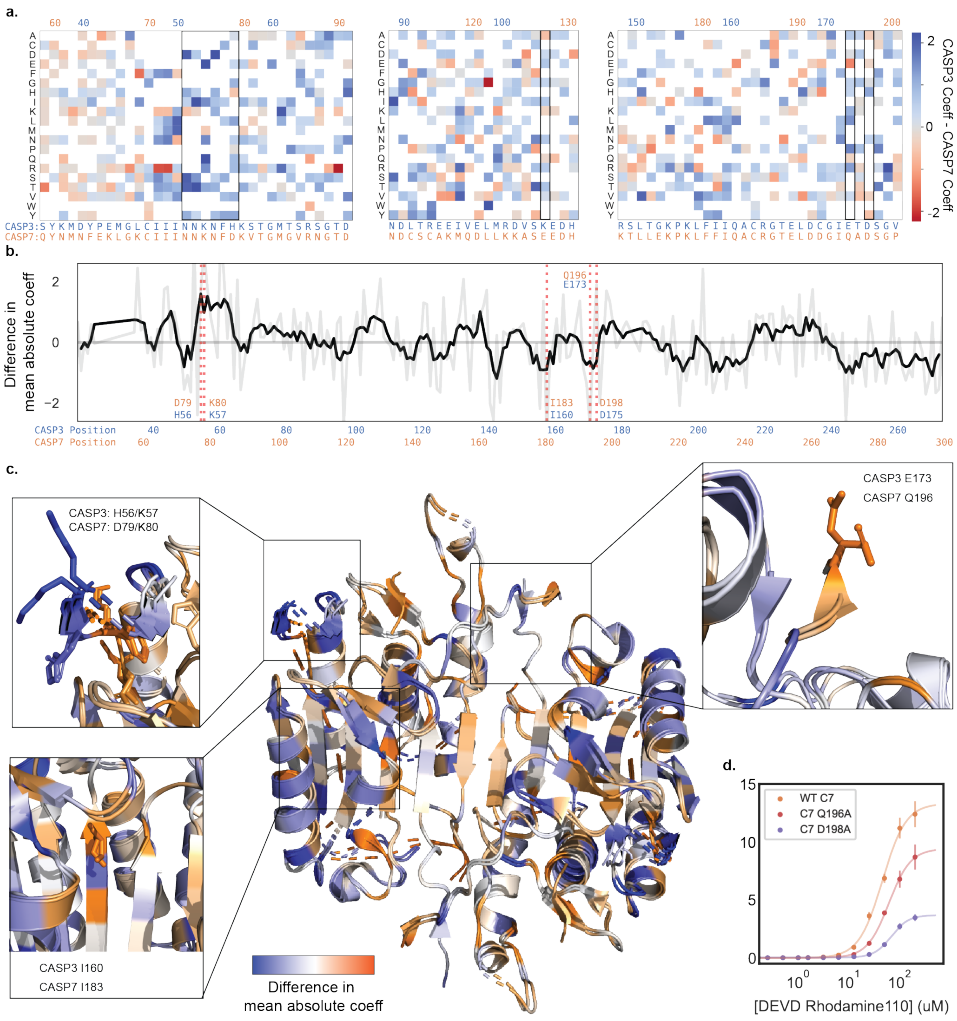


Figure 2.3: Divergence in the mutational landscapes of Caspase-3 and -7 . a, A heatmap of mutation coefficient differences between CASP3 and CASP7. Blue indicates a larger coefficient in CASP3 and orange indicates CASP7 has a larger coefficient. **b**,Differences in mutational tolerance between CASP3 and CASP7. The light grey line shows the difference in a the mean absolute coefficient of a site between CASP3 and CASP7, and black line is a moving average to highlight general differences. Positive values are positions where CASP3 has a larger mean absolute coefficient and thus mutations at that site have a larger effect. Negative values are where CASP7 has a larger effect. The red dotted lines indicate positions of interest. **c**,A mapping of the difference in mean absolute coefficient onto the aligned CASP3 and CASP7 structures (2H5J and 2QL5, respectively). The expanded boxes highlight the sequence regions shown in panels a and b. **d**, Kinetic analysis of CASP7 Q196A and CASP7 D198A.

completely hydrolyze the substrate within a few hours.[21,22] The ability to screen caspases based on fast reaction timescales is necessary to distinguish finer functional differences.[23,24] While our microfluidic screening platform provided these key advantages, it also presented technical challenges. Compared to standardized experimental platforms such as FACS, the design and optimization of our screen and surrounding workflow required dozens of rounds of optimization and engineering before we could reliably use it screen enzymes.[25] This upfront labor presents a non-trivial roadblock for other researchers looking to adapt the platform, especially if their lab is not already equipped to fabricate and use microfluidic devices.[25] We hope that as more demonstrations of our platform's utility follow, other researchers or private industries will begin to create easy-to-use prefabricated microfluidic chips and platforms decreasing the barrier to entry.[20,26]

We mapped the effects of 1644 amino acid substitutions in CASP3 and 1772 amino acid substitutions in CASP7—roughly one third of all possible single amino acid substitutions. Our results corroborated findings from previous research, such as mutational intolerance of the catalytic cysteine and histidine residues and other known allosteric and processing sites. We also observed mutational constraints in both enzymes that closely follow our understanding of protein stability from the structural perspective, such as the destabilizing effects of disrupting salt bridges or mutations to the hydrophobic cores.

We additionally characterized the two putatively activating mutations, CASP3 G177R and CASP7 F241G. CASP3 G177 is located two positions downstream of the D175 processing site and exists on an unstructured loop that is not visible in any crystal structure.

Mutation to arginine at position 177 significantly lowered the K_m of CASP3, but left the k_{cat} unchanged. The CASP7 F241G mutation displayed near wild-type activity, with no significant change in KM or expression, but with decreased k_{cat}. We hypothesize that disrupting core hydrophobic interactions may change the geometry and stability of the enzyme active site and make substrate proteolysis less efficient. These mutations would be difficult to identify without large-scale screening of random mutant libraries.[27]

We compared the mutational profiles between CASP3 and CASP7 to identify sites with differing mutational tolerance that may have diverged during caspase evolution and specialization.[28,29] As expected, a majority of the sites displayed similar mutational tolerance, but a small subset showed statistically significant differences. We identified several key sites that may hold potential for future drug design. CASP7 D79/K80 forms a structurally crucial salt bridge network that is not observed in CASP3. One may imagine designing a drug that could disrupt that network and selectively inhibit CASP7 while leaving CASP3 function relatively untouched.[4,6,9,18]

Our study had several key limitations. First, we chose to express caspases in *E. coli* due to the simplified molecular biology, high transformation efficiency, and the relative insensitivity of bacteria to caspase overexpression. The enzymes expressed in *E. coli* lack glycosylation and must operate in the absence their native regulatory partners such as other caspases and XIAP.[3,30,31] In addition, we assayed caspase activity on a single fluorogenic substrate that is likely not fully representative of their diverse cellular targets.[32,33] These factors could bias our results and reduce the relevance for caspase function in human cells.

Another limitation of deep mutational scanning (DMS) studies in general is the inability to dissect detailed molecular mechanisms.[34,35] Our DMS measurements describe how amino acid substitutions affect caspase activity, but they don't explain why. A mutation that decreases caspase activity could be the result of changes in protein expression, stability and folding, catalytic rate constants, substrate specificity, allosteric regulation, and more. Further biochemical characterization of individual mutants is necessary to obtain a complete picture of inner molecular workings of caspases.

Our results have highlighted several interesting future research directions. Residue Q196 appears to play an important role in CASP7 regulation, presumably because it serves as a secondary cleavage site for activation. Previous work found cleavage at the canonical D198 site or Q196 both activate CASP7, however the Q196 isoform is resistant to inhibition by BIR and XIAP.[17,36] We hypothesize that wild-type CASP7 exists as two different cleavage isoforms and alanine mutation at each of these two processing sites effectively shuts off formation of one of these isoforms. More specifically, the Q196A variant allows us to study the activity of the D198 cleavage isoform in isolation, and the D198A variant allows us to study the Q196 cleavage isoform. Our kinetic analysis of CASP7 Q196A and CASP7 D198A found both variants have diminished activity, but mutation at the canonical processing site D198 attenuates CASP7 activity more than at Q196. This result suggests the two different CASP7 isoforms may have mechanistic differences that account for their differences in kinetics. Considering previous research demonstrated the two isoforms have distinct interactions with native human inhibitory partners.[17], it may be that the endogenous

E. coli serine protease inhibitor Ecotin may also have distinct inhibitory modes with the recombinantly expressed CASP7 isoforms—since all our kinetic analyses were conducted in lysate, those effects could significantly alter the kinetics of the CASP7 mutants.

Further, exploring the possibility of leveraging sites like CASP7 D79/K80 to develop selective caspase inhibitors could be prudent to the field of drug design.[6,37] Demonstrating practical translational results from our screen could open possibilities for using deep mutational scanning for targeted and selective drug design for many other peptide targets.[4,38–40] Developing small molecule modulators that can selectively inhibit or activate members of the caspase family could open the door for novel therapeutics for a wide variety of human diseases.[6,41,42] Designing such molecules is incredibly challenging given the highly conserved structures and functions of caspases, and our limited understanding of protein dynamics and regulation.[42,43] DMS studies could narrow the space of potential target sites by directly and empirically correlating thousands of mutations to their functional effects and finding key protein features that functionally differ in closely related families of proteins.[12]

Materials and Methods

Caspase library generation Δpro-domain *casp3* and *casp7* genes were amplified using error-prone PCR to introduce random mutations. Error-prone PCR was performed following a protocol calling for 50 μM MnCl2 to decrease the fidelity of Taq polymerase.[44] We did 15 amplification cycles, introducing ~4.5 nucleotide mutations in the gene. We subsequently

purified the amplified product, digested it overnight with DpnI to remove remaining wildtype plasmid inserts, and cloned the insert back into pET22b using Circular Polymerase Extension Cloning (CPEC).[45,46]

The CPEC product was purified and used to transform electrocompetent *E. coli* C43(DE3) cells (Lucigen). Transformed cells were recovered for 45 minutes at 37° C then and diluted into 200 mL of sterile LB media with the added carbenicillin. Once the culture's optical density (OD) approached the lower detection limit of our spectrometer ($OD_{600} = 0.2$), the culture was concentrated, and freezer stocks of 25% glycerol were made and stored at -80° C. Each library had roughly 10^7 transformants. 10 transformants were picked from each library and their plasmids sequenced to find that each library had 2.5 amino acid substitutions per library member.

Plate reader-based caspase activity assay Individual clones from the mutagenized libraries were incubated in Magic Media (Invitrogen) for 18 hours at 30° C. Cells were pelleted and resuspended in solution 50 mM Tris, ph 7.4, 50 mM KCl after decanting the supernatant media to achieve a density of 1 OD_{600}/mL. 200 μL of resuspended culture was added to a black 96-well plate. 200 μL of assay reagent (0.3x BugBuster (Invitrogen), 20 uM DEVD-Rhodamine-110 (Bachem), 50 kU/mL Lysozyme, 50 mM Tris pH 7.4, 50 mM KCl, 100 uM EDTA) was added to the plate and the fluorescence (excitation at 480 nm, emission at 530 nm) over time measured on a plate reader. Sequences with >50% of the wildtype activity were considered to be functional.

Caspase kinetics experiments *E. coli* C43(DE3) cells expressing WT CASP3, WT CASP7, CASP3 D175A, CASP7 H144A, CASP3 G177R, CASP7 Q196A, or CASP7 D175A were grown for 18 hours at 30° C in Magic Media (Invitrogen). Cells were centrifuged and resuspended in 50 mM Tris, ph 7.4, 50 mM KCl to 1 OD mL. Enzyme concentration was determined using active site titration.[47] using the irreversible pan-Caspase inhibitor Z-VAD-FMK (Promega) and observing residual activity upon addition of assay reagents described above. Kinetic parameters were determined by observing proteolytic activity with a titrated range of the DEVD-Rhodamine-110 (Bachem) and fitting the observed initial velocity to the Hill equation.[47]

Microfluidic device fabrication An initial layer of photoresist resin, SU-8 3010, was coated onto a mirrored silicon wafer (University Wafers) and centrifuged at 1500 rpm to achieve 15 μm layer height. A photomask (**Supp Fig 2.4**) of the first layer of the microfluidic device was placed on the layer and 100 J/cm^2 of UV light is used to polymerize the features. The wafer was baked at 95° C for 10 minutes to catalyze the polymerization. A second 25 μm layer of SU8-3025 was coated onto the wafer by spinning at 4000 rpm, and similarly polymerized with the second photomask (Sup Fig 2b) to create the incubation line and baked again. Undeveloped photoresist is washed off with SU-8 developer (1-methoxy-2-propanol acetate, MicroChem).

The wafer was then used to create a relief in un-polymerized PDMS (Dow Corning Sylgard® 184, 11:1 polymer:cross-linker ratio), which was then polymerized by baking at

75° C. Inlet and outlet holes are punched with a 0.5 mm biopsy corer. The device was then thoroughly washed with isopropanol and double-deionized water and then plasma treated alongside a clean glass microscope slide, to which it was subsequently bonded. Prior to use, microfluidic channels were filled with Aquapel (Pittsburgh Glass Works) to ensure hydrophobicity, and then baked for 10 minutes at 100° C to vaporize any Aquapel left in the channels.

Microfluidic caspase screening 10 μL of either Caspase-3 or -7 library glycerol stocks was used to inoculate 5 mL of auto-induction media (Invitrogen Magic Media) and allowed to incubate and express for 18 hours at 30° C. The cultures were pelleted and resuspended in the assay buffer (50 mM Tris pH 7.4, 50 mM KCl, 100 μM EDTA) to a concentration of 0.075 OD_{600} to form the 2x cell suspension. A 2x assay reagent solution of 50 mM Tris pH 7.4, 50 mM KCl, 100 uM EDTA, 0.3x BugBuster (Invitrogen), 20 μM DEVD-Rhodamine-110 (Bachem), 50 kU/mL Lysozyme was also made. Both the 2x cell suspension and the 2x assay reagent were loaded into 1 mL luer lock syringes, which were purged of air and fitted with luer-to-PEEK tubing adapters. The cell syringe used PEEK tubing with 0.005" internal diameter, and all other syringes used 0.015" internal diameter PEEK tubing.

Droplets containing expressed Caspase library variants were generated at the co-flow drop maker junction. Both the 2x cell suspension and the 2x assay reagents flowed into the device at 15 μL/hr, and were pinched into droplets by fluorinated oil (HFE 7500) containing 1% (wt/wt) PEG–perfluoropolyether amphiphilic block copolymer surfactant flowing at

100 μL/hr.

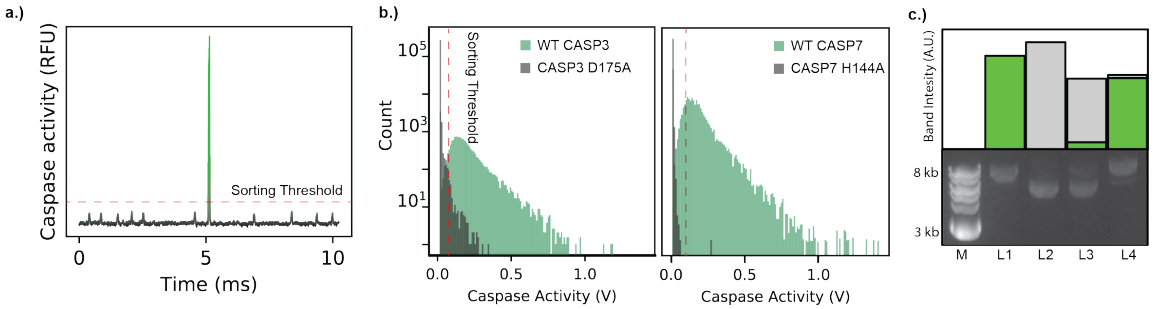
After incubating on-chip for ~3 minutes, droplets were sorted using electrocoalescense with an aqueous stream of 10 mM Tris, pH 8, 1 mM EDTA. A 473-nm laser was focused onto the channel just upstream of the sorting junction, each droplet was individually excited, and its fluorescence emission measured using a spectrally filtered PMT at 520 nm. A field-programmable gate array card controlled by custom LabVIEW code analyzed the droplet signal at 200 kHz, and if it detected sufficient fluorescence, a train of seven 180-V, 40-kHz pulses was applied by a high-voltage amplifier. This pulse destabilized the interface between the droplet and the adjacent aqueous stream, causing the droplet to merge with the stream via a thin-film instability, after which the droplet contents were injected into the collection stream via its surface tension. The contents of the sorted droplets were collected in a microcentrifuge tube for further processing. Droplets were processed at 800-1000 Hz. Because the cell occupancy of the droplets was 10%, we analyzed 80-100 cells per second. Caspase-3 and -7 libraries were sorted in triplicate over a total of 6 days. In total we analyzed.

DNA recovery and sequencing Recovered plasmid DNA was purified using Zymo spin columns and transformed into ultra-high efficiency 10G Supreme *E. coli* cells (Lucigen). Cells cultured in SOC media and recovered for 45 minutes at 37° C. The recovered culture was then used in totality to inoculate a larger 200 mL culture which was incubated overnight until its OD600 reached 0.5. The larger cultures were pelleted and resuspended in 20 mL

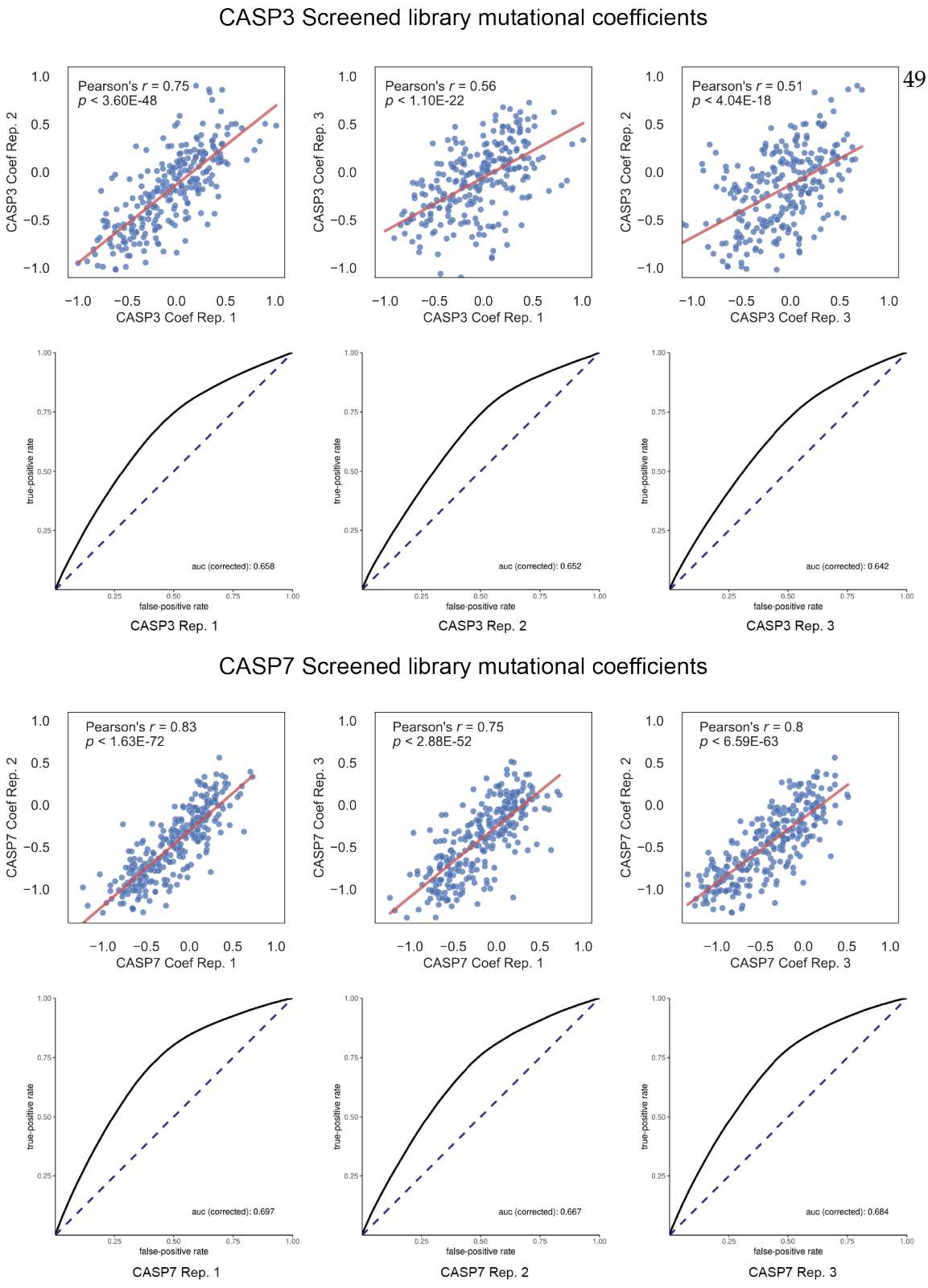
25% glycerol for storage at -80° C. Dilutions of the culture were plated prior to incubation to measure how many transformants were present. We generally observed 0.75–1x as many transformants as what we sorted. Plasmid purified from the larger culture was digested with the restriction enzyme DraIII and ScoI, gel extracted, tagmented using the Nextera XT Library Prepration Kit (Illumina) and sequenced using the Illumina MySeq 2x300. Read coverage for the sequencing runs are displayed in Supplemental Figure 3. Reads with a quality score less than 30 were discarded.

DMS data processing and analysis The reads from the Illumina FASTQ files were mapped to the caspase reference gene using Bowtie249, and translated to amino acid sequences. Mutations observed fewer than 10 times were discarded prior to continuing analysis. Fitness effects of each observed amino acid substitution was estimated using a positive-unlabeled learning framework that compares sequences from the presorted population with the sorted population40,50. Full sequencing databases can be found in the National Center for Bioinformatics Sequence Read Archive (NCBI SRA) under the following accession codes: SRX8049113, SRX8049114, SRX8049115, SRX8049116, SRX8049117, SRX8049118, SRX8049119, SRX8049120, SRX8049121, SRX8049122, SRX8049123, SRX8049124. Python and R scripts used to analyze data can be found at https://github.com/RomeroLab/pudms and https://github.com/RomeroLab/DMS-analysis.

Supplement



Supplementary Figure 2.1: Key controls demonstrate screening platform's ability to discriminate active caspases.. a, A time trace of droplets in the microfluidic device as they cross the fluorescence detector. The small peaks correspond to inactive or empty droplets, while the large peak in the center is a droplet containing active caspases. **b**, A histogram of fluorescence activity for WT CASP3, WT CASP7 and their respective inactive variants, CASP3 D175A and CASP7 H144A as they are observed in the microfluidic device. WT CASP3 and CASP7 display significantly higher fluorescence signal than CASP3 D175A or CASP7 H144A droplets. **c**, Quantification of recovered plasmid from a mock sorting experiment containing a 10:1 mixture of empty pET22 plasmid and CASP3-containing plasmid. Lanes M: NEB 1kb+ DNA standard; L1: pET 22 plasmid containing WT CASP3; L2: pET22 plasmid with no insert; L3: a 10:1 mixture of empty vector to CASP3 plasmid containing E. coli cells; L4: plasmid recovered after microfluidic screening the L3 input showing significant enrichment of CASP3 plasmid.



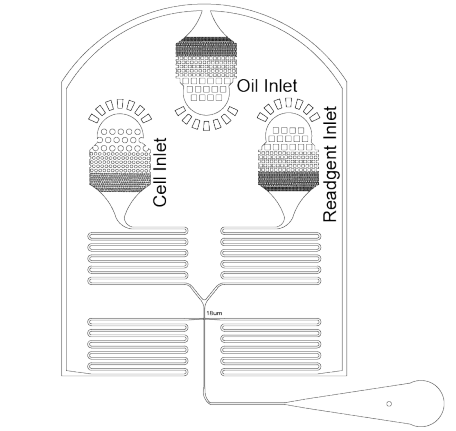
Supplementary Figure 2.2: Correlation of regression coefficients across experimental replicates.. For each experimental replicate, the mean regression coefficient at each site is plotted against each other. All three CASP7 experimental replicates correlate well with each other. CASP3 replicates 1 and 2 correlate well with each other, however replicate 3 correlates poorly with the others and was not used for further analysis. It's likely the microfluidic sorting in replicate 3 had sorting errors that resulted in false positive sequences. The Reciever-Operator Curve Area Under Curve (ROC AUC) for each replicate is > 0.5 suggesting that coefficient effects are non random predictors.

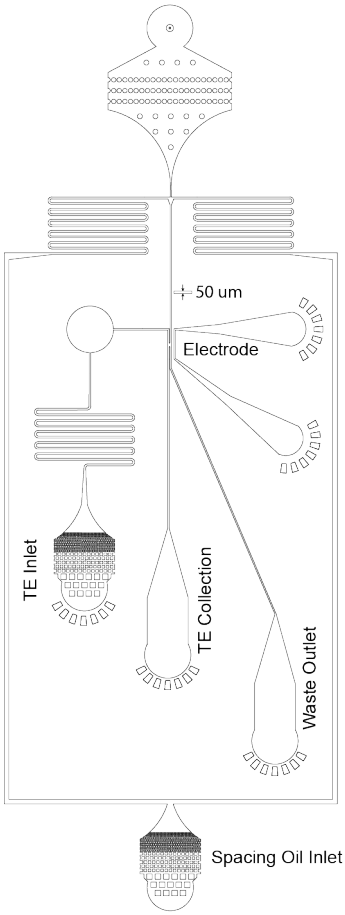
	Replicate	Drops analyzed	Positive drops sorted	Time (hrs)	Analysis frequency (Hz)	Sorting frequency (Hz)	Fraction of drops sorted	CFU recovered	Fraction functional after sorting
CASP3	1	15,639,854	693,057	7.3	592	26.2	4%	450,000	80%
	2	6,886,651	420,630	6.6	291	17.8	6%	200,000	70%
	3	15,061,487	499,529	7	598	19.8	3%	100,000	66%
CASP7	1	19,202,100	593,495	8	670	20.7	3%	280,000	85%
	2	22,474,041	602,890	8.3	755	20.3	3%	480,000	90%
	3	17,601,832	414,992	7	698	16.5	2%	120,000	90%

Table S1: Caspase Screening Statistics. CASP3 and CASP7 were sorted in triplicate. Each sort lasted for 6.5-8 hours and we were able to recover at least 10^5 variants per sort.

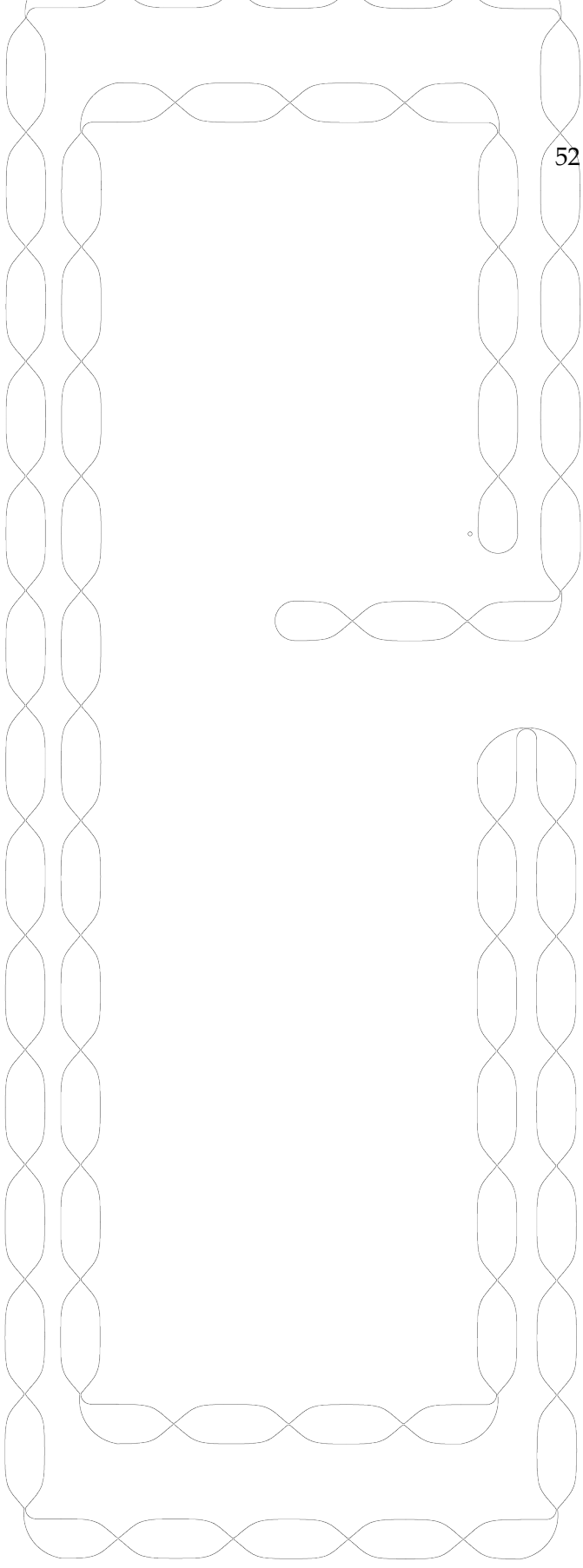
Enzyme	[Enzyme] (nM)	V_{max} (nMol/s)	K_m (uM)	K_{cat} (s^{-1})	K_{cat} / K_m ($s^{-1}uM^{-1}$)
WT CASP3	200 +/- 20	2.78 +/- 0.06	21.0 +/- 0.9	0.13 +/- 0.007	0.70 +/- 0.06
CASP3 D175A	190 +/- 100	n.m.	n.m.	n.m	n.m.
CASP3 G177R	170 +/- 20	2.29 +/- 0.04	8.2 +/- 0.3	0.28 +/- 0.01	1.60 +/- 0.05
WT CASP7	80 +/- 20	1.33 +/- 0.04	45 +/- 2	0.03 +/- 0.002	0.40 +/- 0.01
CASP7 H144A	180 +/- 70	n.m.	n.m.	n.m	n.m.
CASP7 Q196A	150 +/- 20	0.94 +/- 0.04	61 +/- 4	0.02 +/- 0.001	0.10 +/- 0.02
CASP7 D198A	100 +/- 20	0.92 +/- 0.05	68 +/- 3	0.0054 +/- 0.0003	0.05 +/- 0.02
CASP7 F241G	80 +/- 20	0.92 +/- 0.05	44 +/- 4	0.13 +/- 0.02	0.25 +/- .05

Table S2: Kinetic Paramters of CASP variants. Recombinant CASP mutants expressed in E. coli were assayed for proteolytic activity against the fluorescent substrate DEVD-Rhodamine-110 at a range of concentrations. Enzyme concentration was determined by active site titration using the irreversible pan-caspase inhibitor Z-VAD-FMK (Promega). Measurements were taken in triplicate and uncertainty calculated as the standard error.



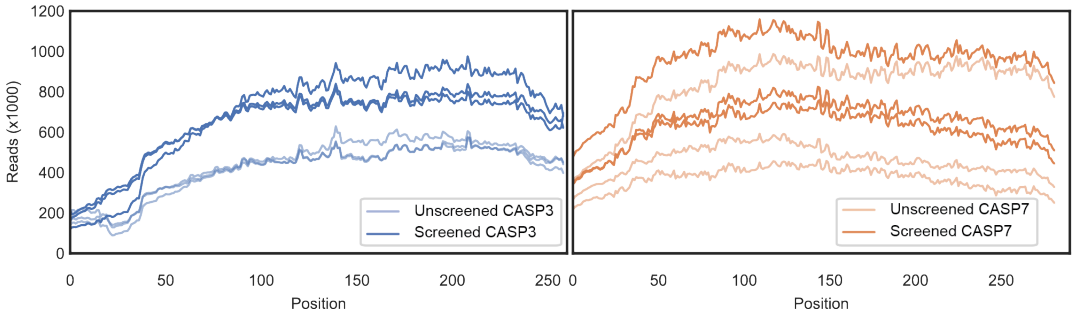


Layer 1: Drop Maker (top) and Sorter (bottom)



Layer 2: Incubation Line

Supplementary Figure 2.3: Scale schematic of the microfluidic design used in this study.. Layer 1 features the 15 μm-tall drop-maker and sorter. All inlets are labeled as such. Layer 2 is the 50 μm-tall incubation line that connects the drop maker and the sorter. The "sausage-like" repeating pattern serves to randomize the position of droplets transverse to the direction of flow to average out the effects of laminar flow that makes droplets in the center move faster than droplets near the edge of the channel.



Supplementary Figure 2.4: Read coverage of Illumina Sequencing runs for CASP3 and CASP7 datasets.. Each sequenced library (Unscreened CASP3, Screened CASP3, Unscreened CASP7, Screened CASP7) displays at least 200k reads at each position. Mutations that appeared fewer than 10 times at any given position were assumed to be noise and not considered in downstream analysis.

Author Contribution H.R. and P.A.R conceived the project. H.R. performed the experiments and analyzed the data. H.R. and P.A.R. wrote the manuscript.

Funding Statement This work was supported by the US National Institutes of Health (5R35GM119854). "The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health."

Competing Interest Statement The authors declare that they have no conflicts of interest with the contents of this article.

Acknowledgments We would like to acknowledge the UW-Madison Biotechnology Center DNA Sequencing Facility for assistance with Illumina library preparation and sequencing.

Data availability statement The datasets generated during and/or analysed during the current study are available in the National Center for Bioinformatics Sequence Read Archive (NCBI SRA) under the following accession codes: SRX8049113, SRX8049114, SRX8049115, SRX8049116, SRX8049117, SRX8049118, SRX8049119, SRX8049120, SRX8049121, SRX8049122, SRX8049123, SRX8049124.

References

[1] Shalini, S., Dorstyn, L., Dawar, S. & Kumar, S. Old, new and emerging functions of caspases (2015). URL `https://www.nature.com/cdd/journal/v22/n4/pdf/`

cdd2014216a.pdf.

[2] Graham, R. K. *et al.* Cleavage at the Caspase-6 Site Is Required for Neuronal Dysfunction and Degeneration Due to Mutant Huntingtin. *Cell* **125**, 1179–1191 (2006). URL https://www.sciencedirect.com/science/article/pii/S0092867406005587.

[3] Fuentes-Prior, P. & Salvesen, G. S. The protein structures that shape caspase activity, specificity, activation and inhibition. *Biochem. J* **384**, 201–232 (2004). URL http://www.biochemj.org/content/ppbiochemj/384/2/201.full.pdf.

[4] MacKenzie, S. H., Schipper, J. L. & Clark, A. C. The potential for caspases in drug discovery (2010). URL http://www.ncbi.nlm.nih.gov/pubmed/20812148http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3289102.

[5] McIlwain, D. R., Berger, T. & Mak, T. W. Caspase functions in cell death and disease. *Cold Spring Harbor perspectives in biology* **5**, a008656 (2013). URL http://www.ncbi.nlm.nih.gov/pubmed/23545416http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3683896.

[6] Häcker, H. G., Sisay, M. T. & Gütschow, M. Allosteric modulation of caspases (2011). URL http://www.sciencedirect.com/science/article/pii/S0163725811001604.

[7] Krishna Deepak, R. N., Abdullah, A., Talwar, P., Fan, H. & Ravanan, P. Identification of FDA-approved drugs as novel allosteric inhibitors of human exe-

cutioner caspases. *Proteins: Structure, Function and Bioinformatics* **86**, 1202–1210 (2018). URL https://onlinelibrary.wiley.com/doi/full/10.1002/prot.25601https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.25601https://onlinelibrary.wiley.com/doi/10.1002/prot.25601.

[8] Agniswamy, J., Fang, B. & Weber, I. T. Conformational similarity in the activation of caspase-3 and -7 revealed by the unliganded and inhibited structures of caspase-7. *Apoptosis* **14**, 1135–1144 (2009). URL http://www.pymol.orghttp://link.springer.com/10.1007/s10495-009-0388-9.

[9] Kudelova, J., Fleischmannova, J., Adamova, E. & Matalova, E. Pharmacological caspase inhibitors: research towards therapeutic perspectives. *Journal of physiology and pharmacology : an official journal of the Polish Physiological Society* **66**, 473–82 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/26348072.

[10] Hardy, J. A., Lam, J., Nguyen, J. T., O'Brien, T. & Wells, J. A. Discovery of an allosteric site in the caspases. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 12461–6 (2004). URL http://www.ncbi.nlm.nih.gov/pubmed/15314233http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC514654.

[11] Romero, P. A., Tran, T. M. & Abate, A. R. Dissecting enzyme function with microfluidic-based deep mutational scanning. *Proc. Nat. Adad. Sci.* **112**, 7159–7164 (2015). URL http://www.pnas.org/content/112/23/7159.full.pdf.

[12] Fowler, D. M., Stephany, J. J. & Fields, S. Measuring the activity of protein variants on a large scale using deep mutational scanning (2014). URL http://www.nature.com/doifinder/10.1038/nprot.2014.153.

[13] Fowler, D. M. & Fields, S. Deep mutational scanning: A new style of protein science (2014). URL http://www.nature.com/articles/nmeth.3027.

[14] Brentnall, M., Rodriguez-Menocal, L., De Guevara, R. L., Cepero, E. & Boise, L. H. Caspase-9, caspase-3 and caspase-7 have distinct roles during intrinsic apoptosis. *BMC Cell Biology* **14**, 32 (2013). URL http://bmccellbiol.biomedcentral.com/articles/10.1186/1471-2121-14-32.

[15] Slee, E. A., Adrain, C. & Martin, S. J. Executioner caspase-3, -6, and -7 perform distinct, non-redundant roles during the demolition phase of apoptosis. *The Journal of biological chemistry* **276**, 7320–6 (2001). URL http://www.ncbi.nlm.nih.gov/pubmed/11058599.

[16] Lakhani, S. A. *et al.* Caspases 3 and 7: Key mediators of mitochondrial events of apoptosis. *Science* **311**, 847–851 (2006). URL http://www.ncbi.nlm.nih.gov/pubmed/16469926http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3738210.

[17] Scott, F. L. *et al.* XIAP inhibits caspase-3 and -7 using two binding sites: evolutionarily conserved mechanism of IAPs. *The EMBO Journal* **24**, 645–655 (2005).

URL http://emboj.embopress.org/content/embojnl/24/3/645.full.pdfhttp://www.ncbi.nlm.nih.gov/pubmed/15650747http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC548652.

[18] O'Brien, T. & Lee, D. Prospects for caspase inhibitors. *Mini reviews in medicinal chemistry* **4**, 153–65 (2004). URL http://www.ncbi.nlm.nih.gov/pubmed/14965288.

[19] Frenz, L., Blank, K., Brouzes, E. & Griffiths, A. D. Reliable microfluidic on-chip incubation of droplets in delay-lines. *Lab on a Chip* **9**, 1344–1348 (2009). URL https://pubs.rsc.org/en/content/articlehtml/2009/lc/b816049jhttps://pubs.rsc.org/en/content/articlelanding/2009/lc/b816049j.

[20] Piyasena, M. E. & Graves, S. W. The intersection of flow cytometry with microfluidics and microfabrication. *Lab on a Chip* **14**, 1044 (2014). URL http://xlink.rsc.org/?DOI=c3lc51152a.

[21] MacKenzie, S. H. & Clark, A. C. Death by caspase dimerization. *Advances in Experimental Medicine and Biology* **747**, 55–73 (2012). URL http://www.ncbi.nlm.nih.gov/pubmed/22949111http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3877935.

[22] Denault, J.-B. & Salvesen, G. S. Human caspase-7 activity and regulation by its N-terminal peptide. *The Journal of biological chemistry* **278**, 34042–50 (2003). URL http://www.ncbi.nlm.nih.gov/pubmed/12824163.

[23] Ng, E. X., Miller, M. A., Jing, T. & Chen, C. H. Single cell multiplexed assay for proteolytic activity using droplet microfluidics. *Biosensors and Bioelectronics* **81**, 408–414 (2016). URL http://www.sciencedirect.com/science/article/pii/S0956566316301932?_rdoc=1&_fmt=high&_origin=gateway&_docanchor=&md5=b8429449ccfc9c30159a5f9aeaa92ffb&dgcid=raven_sd_recommender_email.

[24] Nicholls, S. B., Chu, J., Abbruzzese, G., Tremblay, K. D. & Hardy, J. A. Mechanism of a genetically encoded dark-to-bright reporter for caspase activity. *The Journal of biological chemistry* **286**, 24977–86 (2011). URL http://www.ncbi.nlm.nih.gov/pubmed/21558267http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3137071.

[25] Holtze, C. Large-scale droplet production in microfluidic devices - An industrial perspective. *Journal of Physics D: Applied Physics* **46**, 114008 (2013). URL https://iopscience.iop.org/article/10.1088/0022-3727/46/11/114008https://iopscience.iop.org/article/10.1088/0022-3727/46/11/114008/meta.

[26] Collins, D. J., Neild, A., DeMello, A., Liu, A.-Q. & Ai, Y. The Poisson distribution and beyond: methods for microfluidic droplet production and single cell encapsulation. *Lab on a chip* **15**, 3439–59 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/26226550.

[27] Hietpas, R. T., Jensen, J. D. & Bolon, D. N. Experimental illumination of a fitness landscape. *Proceedings of the National Academy of Sciences of the United States of America*

108, 7896–7901 (2011). URL http://www.ncbi.nlm.nih.gov/pubmed/21464309http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3093508.

[28] Reynolds, K. A., McLaughlin, R. N. & Ranganathan, R. Hot spots for allosteric regulation on protein surfaces. *Cell* **147**, 1564–1575 (2011). URL http://www.sciencedirect.com/science/article/pii/S0092867411013626.

[29] Lamkanfi, M., Festjens, N., Declercq, W., Berghe, T. V. & Vandenabeele, P. Caspases in cell survival, proliferation and differentiation. *Cell Death and Differentiation* **14**, 44–55 (2007).

[30] Hermel, E. *et al.* Specific caspase interactions and amplification are involved in selective neuronal vulnerability in Huntington's disease. *Cell Death and Differentiation* **11**, 424–438 (2004).

[31] Turowec, J. P. *et al.* An unbiased proteomic screen reveals caspase cleavage is positively and negatively regulated by substrate phosphorylation. *Molecular 'I&' cellular proteomics : MCP* **13**, 1184–97 (2014). URL http://www.ncbi.nlm.nih.gov/pubmed/24556848http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4014278.

[32] Julien, O. & Wells, J. A. Caspases and their substrates (2017).

[33] Pop, C. & Salvesen, G. S. Human caspases: Activation, specificity, and regula-

tion (2009). URL http://www.ncbi.nlm.nih.gov/pubmed/19473994http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2755903.

[34] Starr, T. N. & Thornton, J. W. Epistasis in protein evolution (2016). URL http://doi.wiley.com/10.1002/pro.2897.

[35] Firnberg, E., Labonte, J. W., Gray, J. J. & Ostermeier, M. A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape. *Molecular Biology and Evolution* **31**, 1581–1592 (2014). URL https://academic.oup.com/mbe/article/31/6/1581/2925654.

[36] Suzuki, Y., Nakabayashi, Y., Nakata, K., Reed, J. C. & Takahashi, R. X-linked inhibitor of apoptosis protein (XIAP) inhibits caspase-3 and -7 in distinct modes. *The Journal of biological chemistry* **276**, 27058–63 (2001). URL http://www.ncbi.nlm.nih.gov/pubmed/11359776.

[37] Feldman, T. *et al.* A Class of Allosteric Caspase Inhibitors Identified by High-Throughput Screening. *Molecular Cell* **47**, 585–595 (2012). URL https://www.sciencedirect.com/science/article/pii/S1097276512005023.

[38] Haddox, H. K., Dingens, A. S. & Bloom, J. D. Experimental Estimation of the Effects of All Amino-Acid Mutations to HIV's Envelope Protein on Viral Replication in Cell Culture. *PLoS Pathogens* **12** (2016). URL http://journals.plos.org/plospathogens/article/file?id=10.1371/journal.ppat.1006114&type=printable.

[39] Song, H., Bremer, B. J., Hinds, E. C., Raskutti, G. & Romero, P. A. Inferring Protein Sequence-Function Relationships with Large-Scale Positive-Unlabeled Learning. *Cell Systems* **12**, 92–101.e8 (2021).

[40] Pande, J., Szewczyk, M. M. & Grover, A. K. Phage display: Concept, innovations, applications and future (2010). URL `http://www.sciencedirect.com/science/article/pii/S0734975010000972`.

[41] Poreba, M. *et al.* Small Molecule Active Site Directed Tools for Studying Human Caspases (2015).

[42] Maciag, J. J. *et al.* Tunable allosteric library of caspase-3 identifies coupling between conserved water molecules and conformational selection. *Proceedings of the National Academy of Sciences of the United States of America* **113**, E6080–E6088 (2016). URL `http://www.ncbi.nlm.nih.gov/pubmed/27681633http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5068305`.

[43] Nussinov, R. & Tsai, C.-J. Allostery in Disease and in Drug Discovery. *Cell* **153**, 293–305 (2013). URL `http://dx.doi.org/10.1016/j.cell.2013.03.034`.

[44] Bloom, J. D. *et al.* Evolution favors protein mutational robustness in sufficiently large populations. *BMC Biology* **5**, 1–21 (2007). URL `https://link.springer.com/articles/10.1186/1741-7007-5-29https://link.springer.com/article/10.1186/1741-7007-5-29`.

[45] Quan, J. & Tian, J. Circular polymerase extension cloning of complex gene libraries and pathways. *PLoS ONE* **4**, 6441 (2009). URL `www.plosone.org`.

[46] Quan, J. & Tian, J. Circular polymerase extension cloning for high-throughput cloning of complex and combinatorial DNA libraries. *Nature Protocols* **6**, 242–251 (2011). URL `http://www.nature.com/doifinder/10.1038/nprot.2010.181`.

[47] Boucher, D., Duclos, C. & Denault, J. B. General in vitro caspase assay procedures. *Methods in Molecular Biology* **1133**, 3–39 (2014).

Chapter 3

DEEP MUTATIONAL SCANNING OF YAK RUMEN GLYCOSIDE HYDROLASE REVEALS MUTATIONS CONFERRING TOLERANCE TO INDUSTRIAL SOLVENT γ-VALEROLACTONE

HSR designed and conducted experiments, interpreted results, and analyzed data; JGL aided in assay development and optimization; HSR and PAR wrote the manuscript.

This chapter is being prepared for publication:

Roychowdhury R, RomeroPA. *Deep Mutational Scanning of Yak rumen glycoside hydrolase reveals mutations conferring tolerance to industrial solvent γ-valerolactone.* In preparation. **2022**.

Abstract

Engineering solvent resistance in glycoside hydrolases, and enzymes more broadly, has been uniquely challenging for researchers. Selection experiments to evolve such resilience are practically impossible to design, computational models poorly predict solvent tolerance, and most screening methods lack the throughput necessary to scan broad enough swaths of protein sequence space to find solvent resilient hits. In this study, we utilize a high-throughput droplet microfluidic screening platform to challenge the promiscuous glycoside hydrolase, CMX, with the industrial biofuel solvent, γ-valerolactone. Our screen identified several resilience-conferring mutations, from which we designed a CMX variant with significantly increased γ-valerolactone compared to the wild-type enzyme.

Introduction

Humanity's destructive impact on global biospheres and ecologies by way of fossil fuel consumption cannot be understated. Lignocellulose biomass, a major byproduct of industrial agriculture, may be utilized as an alternative biofuel and as a precursor to industrially relevant chemicals.[1] Itself a biomass-derived product, the solvent γ-valerolactone (GVL) can be used to pretreat lignocellulosic biomass for further saccharification, wherein biomass is solvated in an 80% GVL and 20% water.[2–4] The pretreatment strips over 80% of the lignin away, maintains near complete cellulose retention, and nearly all the GVL can be recovered and reused. After fractionation, downstream enzymatic hydrolysis turning polysaccharides

into sugars is seen to be the logical next step in biofuel production.[2,5]

RuCelA, isolated from the Yak gut microbiome, a promiscuous glycoside hydrolase with cellulase, mannase, and xylanase activity (earning the enzyme the moniker "CMX," which we use in this study) has been considered a strong candidate for enzymatically processing GVL treated products.[6,7] Predictably, having been evolved and optimized for lignocellulytic activity in the guts of yaks, we have found the enzyme's desired activity in the industrial conditions specified above to be deficient. Particularly, we have found that residual GVL leftover from fractionation halves CMX activity at concentrations as low as 3% volume/volume.

The goal in this study was to take the first steps in engineering CMX to better tolerate non-native conditions. We used a fully-integrated microfluidic lab-on-a-chip to screen a library of CMX mutants in high-throughput while subject to a 3% GVL challenge. We have identified an initial CMX variant with twice the activity of the wild type in the presence of 3% GVL. The microfluidic screening platform allowed us to screen millions of in vitro isotemporal enzymatic reactions against the fluorogenic substrate resorufin cellobioside, providing us quanitative information on individual mutations contribution to the enzyme sequence-function landscape.[8,9] We hope that with additional iterative rounds of evolution, we will soon evolve an enzyme well suited to the industrial decomposition of lignocellulose biomass.

Results

Droplet microfluidic screening platform enriches CMX Directed evolution is a key strategy in designing enzymes with characteristics conducive to pharmaceutical, industrial, and bioscience applications.[10,11] Typically, protein mutant libraries are subject to high-throughput screens or selections to introduce gain-of-function mutations. Designing an assay capable of selecting for resistance-conferring mutations in CMX is nearly impossible. One would first need to engineer an expression platform relies solely on CMX library variants for saccharide metabolism—no trivial feat—before even considering the addition of the solvent GVL, which disrupts cell membranes, to introduce evolutionary pressure.[12] For similar reasons, a fluorescent screen that could utilize fluorescence-activated cell sorting (FACS) is also not a feasible strategy unless one has already engineered a GVL resistant expression platform.[12–15]

Finding activating mutations that confer resistance to GVL requires screening a library of CMX variants in vitro. In lieu of laborious and reagent intensive plate-reader based assays, we used an ultra-highthroughput droplet microfluidic platform (**Fig 3.1**) capable of screening hundreds of CMX variants per second.[8,16] Individual CMX variants, generated through random mutagenesis and expressed in *E. coli* BL21 cells, were encapsulated in 10 pL droplets containing lysis reagents, the target fluorescent substrate, resorufin cellobioside, and a 3% GVL challenge. One in ten droplets contained an individual *E. coli* cell, physically separate from all the others, which is lysed and the CMX variant allowed to interact with

the substrate on a continuous flow reactor (**Fig 3.1a**). Following incubation, each droplet's fluorescence is measured using a laser fluorimeter. The incubation time, 1 hour, was calibrated such that the wild type CMX displays significantly lower fluorescence than the putatively GVL resistant CMX variants (**Fig 3.1c**). Highly fluorescent droplets', presumed to contain active CMX variants, are collected and their associated plasmid collected for Illumina sequencing.

We screened our library, where each CMX variant contained 2-5 amino acid substitutions, in triplicate. We were able to screen over 300,000 CMX variants per hour, and in each replicate collected ~100,000 GVL resistant CMX variants (Supp Fig). We confirmed our platform's ability to distinguish GVL-resistant CMX variants from wild type by assaying random individual screened variants against random individual variants from pre-screened library. We observed significant enrichment for near-wild-type activity in the screened population in the presence of 3% GVL challenge (**Fig 3.1d**).

Deep mutational Scanning of CMX By sequencing the CMX library prior to and after screening, we quantified the contributions of individual mutations to CMX function under the GVL challenge (**Fig3.2a**). Using a positive-unlabeled learning framework, we calculated a coefficient for each observed mutation, as well as the significance of that coefficient as a p-value .[9] (**Fig3.2c**). Mutations with coefficients less than zero were considered deleterious to enzyme activity, and conversely coefficients greater than zero were considered to be activating. Two of the three replicates were considered for analysis, the third replicate having

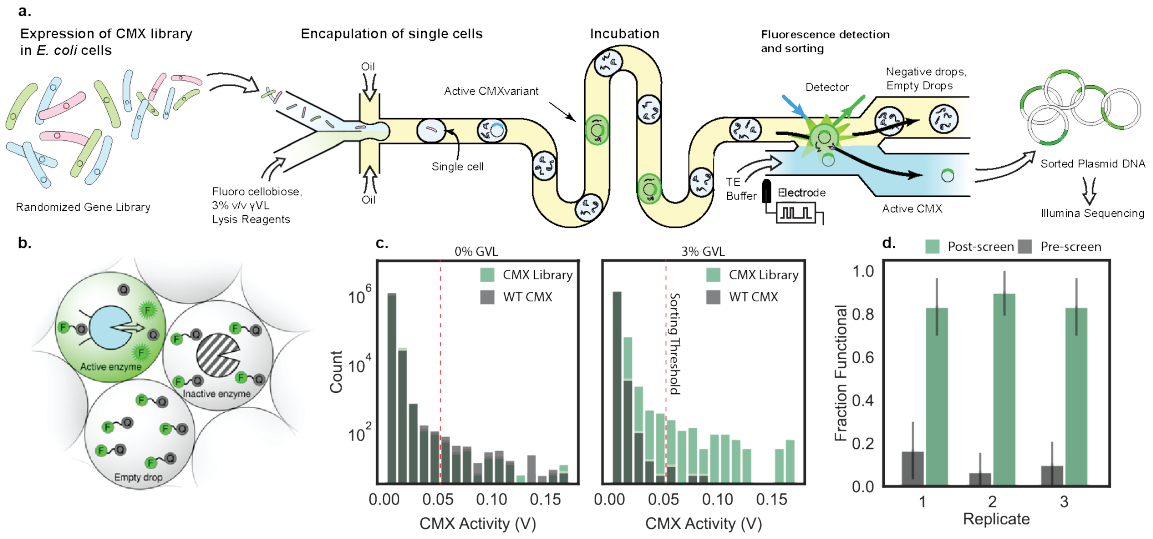


Figure 3.1: Microfluidic screening platform for screening CMX with 3% GVL challenge. a, A schematic of our microfluidic screening system. A dilute suspension of *E. coli* expressing CMX variants are injected into a microfluidic device and individual cells are encapsulated into microdroplets containing lysis reagents and a fluorogenic glycoside hydrolase substrate. The cells are lysed, the enzyme reaction is incubated on-chip, and the fluorescence of each droplet is analyzed using a laser. The fluorescent droplets are then sorted by electrocoalescence with an aqueous stream that collects the sorted plasmids for downstream analysis. **b**, Droplets containing active caspase variants will fluoresce, whereas empty droplets and droplets containing inactive CMXs will not.**c**, On-chip incubation time was calibrated to achieve maximum separation of the library's fluorescence from the WT CMX's under GVL conditions. In the absence of GVL, the distribution of droplets' fluorescent signal in both the library and WT population are similar; In the presence of 3% GVL, very few WT droplets display signal, allowing us to easily separate active CMX variants. **d**, Enrichment of active CMX variants in the presence of 3% GVL. All three technical replicates display significant enrichment for active CMX variants.

not been successfully sequenced with sufficient quality to extract meaningful information. The two replicates we analyzed correlated well with each other (**Supp Fig3.1**).

As expected, most mutations were deleterious to enzyme activity, especially those to the active site residues E177 and E288, labeled in red in **Fig3.2a**. Notably, glutamine mutations to active site residues were well tolerated, corroborating previous findings by Glasgow, et al, wherein Glu to Gln active site mutations made significant alterations to the enzyme's substrate specificity[7]. As observed in previous deep mutational scanning studies .[8,16,17], substitutions to and by aromatic residues in the protein core were ill tolerated, whereas the solvent exposed surface of the protein generally accepted mutations to polar residues (**Fig3.2ab**).

The five most putatively activating mutations ($p < 0.01$) observed, labeled and marked in green in **Fig3.2ac**, were chosen for further analysis. We designated the CMX variant containing the mutations H6P, S17T, D20H, D26H, S272W CMX-Top5. 4 of the 5 mutations occur on the N-terminus of CMX—and while the 5th, S272W occurs on the C-terminus, it is proximal to the other 4 (**Fig3.2b**). We chose to assess this variant's resistance to the 3% CMX challenge we screened for.

Activating mutations confer resistance to γ-valerolactone inhibition CMX is well optimized to function in the yak rumen where the survival of microbes (and by extension, the yak) depends on a suite of enzymes capable of converting the complex and indigestible lignocellulose comprising their diet into simpler sugars for energy. Performing the same

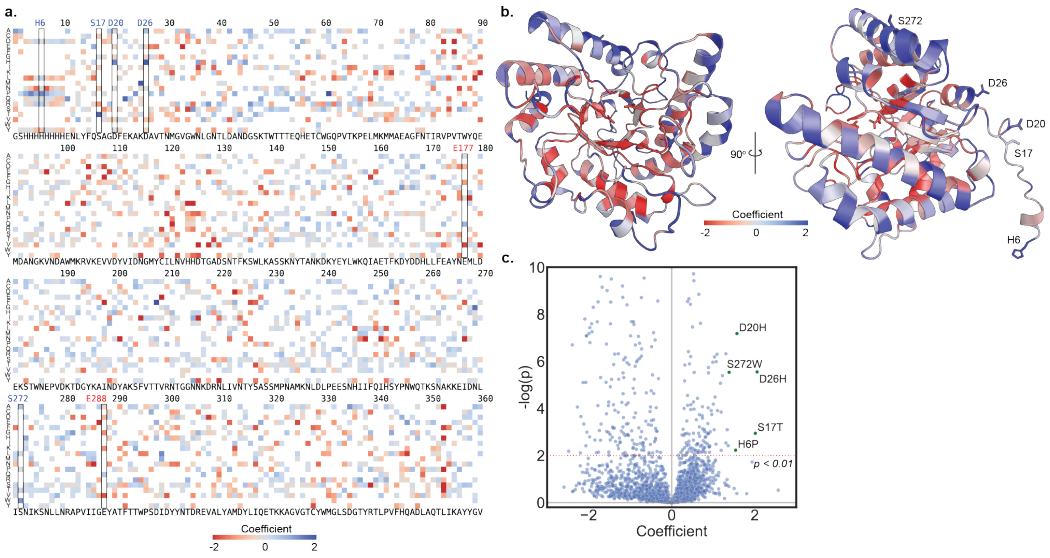
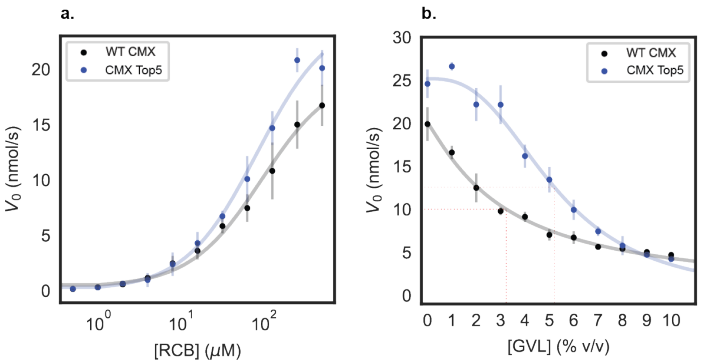


Figure 3.2: Mutational effects identified from deep mutational scan. a, A heatmap displaying the effects of all observed mutations to CMX in the presence of GVL. More negative (red) mutational coefficients indicate deleterious mutations, while more positive (blue) mutations indicate putatively activating mutations conferring resistance to GVL. The wild-type sequence is listed on the bottom of the heatmap. **b**, Average mutational effects at each position mapped onto the modeled enzyme structure. Sites colored red are largely intolerant of mutations, and sites colored blue are tolerant to mutations, similar to panel **a** The sites of the most activating mutations are shown—all but one are in the first 30 amino acids of the primary sequence and distal to the enzyme active site. **c**, A volcano plot depicting the effect and significance of all mutations to CMX. The five mutations with the highest coefficient where $p < 0.01$ are highlighted in green.

chemistry on the scale necessary to manufacture biofuels requires that we adapt CMX to a significantly different environment while maintaining (or hopefully improving) its glycosidase activity. With a 3% GVL challenge, we identified a series of activating mutations that we expect to confer CMX resilience to the conditions expected in biomass fuel conversion processes.

We expressed and purified the WT CMX and the putatively GVL resistant variant CMX Top5 to characterize their respective kinetics and tolerance to GVL. We titrated WT CMX and CMX Top5 with from 0 - 500 μM of the fluorescent substrate Resorufin Cellobioside, observing the initial rate of the reaction, V_0 (**Fig3.3a**). We estimated relevant kinetic parameters by modeling the resultant data with the Hill equation. We did not observe any significant change in CMX Top5's K_m or its catalytic specificity, k_{cat}/k_m compared to the wild type, but did record a significant~20% increase in its maximum velocity (V_{max}), from 20 nmol/s to 24.4 nmol/s (**Fig3.3c**).

To evaluate the primary goal of evolving GVL resiliency, we performed a dose-response assay on WT CMX and CMX Top5. We observed their respective V_{max} as GVL was titrated from 0-10% v/v (**Fig3.3b**). Validating our efforts, CMX Top5 displayed a 60% increase in its IC_{50}, increasing from 3% to 5%, and exhibiting WT-like activity under the screening conditions of 3% GVL (**Fig3.3bc**).



c.

Enzyme	V_{max} (nmol/s)	k_{cat}/k_m ($s^{-1}nmol^{-1}$)	IC_{50} (% v/v)
WT CMX	20 ± 1	0.26 ± 0.05	3.2 ± 0.2
CMX Top5	24.4 ± 1.3	0.24 ± 0.04	5.2 ± 0.2

Figure 3.3: Kinetics and dose response of CMX Top5 compared to WT. a, Kinetics of WT CMX and CMX Top5 variant. CMX Top5 retains the k_m and k_{cat} of the WT but displays a ~20% increase in V_{max}. **b**, CMX Top5 shows significantly increased tolerance to GVL, displaying WT-like activity under the 3% GVL challenge conditions, and its IC_{50} increases to ~5%. **c**, The kinetic parameters for WT CMX and CMX Top5 estimated by fitting kinetic data to the Hill equation.

Discussion

Biofuels represent a potentially crucial transitionary step in reducing our global dependance on fossil fuels for energy, particularly in the transportation sector[18]. Otherwise unusable organic biomass, such as agricultural byproducts from harvesting corn, is carbon rich and holds massive potential to replace fossil fuels as a source of both energy and precursors to plastics and other small molecules[19,20]. Such unused feedstock is woody, composed of complex oligosaccharide structures like lignocellulose and hemicellulose that are difficult to decompose into simpler and more useful sugars[21,22]—except for yaks and other ruminants with a cadre of highly effective grass-decomposing microbes with full-time residence in your gut[6,23,24].

A crucial step in our own efforts to decompose such biomatter into its constituent simple sugars is the cooption of such enzymes used by those ruminant gut flora[6,25,26] . CMX is a promising candidate for industrial use[6,7], considering its broad activity toward many different glycosidic bonds, but loses half its activity under the conditions presented in the industrial process—especially the use of γ-valerolactone (GVL) as a solvent. In this study, we aimed to engineer a variant of CMX more resilient to the 3% GVL present in the biofuel manufacturing process. We utilized contemporary microfluidic based platforms, capable of measuring enzyme veloicty in in vitro, to screen a library of CMX variants in ultra-high-throughput[16]. Under the relevant 3% GVL challenge conditions, we were able to isolate mutations that conferred resilience to the solvent conditions.

We expressed, purified, and characterized a variant of CMX with the 5 most activating mutations, designated CMX Top5, and found that it retained wild-type activity toward the screening substrate, cellobioside, in the 3% GVL challenge condition. CMX Top5 displayed a ~20% increase in V_{max} compared to the wild type. More importantly for our purposes, CMX Top5 displayed significantly increased resilience to the expected conditions on biofuel production. Where WT CMX lost half its activity at the expected 3% GVL, CMX Top5 maintained all its activity at that concentration and had a significantly increased IC_{50}—5% compared to 3%. This increase in GVL resistance is completely consistent with our screening parameters, which were sufficiently discriminatory to exclude the attenuated WT-like activity under the challenge conditions.

While we are satisfied with our initial foray into engineering a GVL resistant CMX variant, there are some key improvements that we have identified for future rounds of engineering. Firstly, having used error-prone PCR to generate a library of random mutants, we were only able to observe roughly one third of all possible amino acid substitutions and likely missed numerous potentially activating mutations. A more comprehensive library may have revealed a different, more effective set of resistance-conferring mutations. Conversely, a more targeted library, informed by contemporary machine learning models, could have added depth to our screen by increasing the overall functionality of our library.[27–30]

Secondly, our assay substrate was not necessarily representative of the expected substrates in biofuel processing reactors—cellobioside represents only one among many of the glycosidic bonds ($\beta 1 \rightarrow 4$) found in woody biomatter.[19,23,31] By screening using one

substrate over many, we may have inadvertently altered the enzyme's substrate specificity profile.

Thirdly, it is difficult to infer the molecular mechanisms by which the observed mutations confer GVL resistance.[28] The DMS measurements describe how amino acid substitutions affect enzyme activity, but fail to explain why.[17,32] The analysis we used also fails to account for epistatic interactions between co-occuring mutations, treating each individual mutations effects as independent. By treating mutational effects as independent in our analysis, we may have missed key pairings that have significant effects on enzyme function.[17,33,34] As with all studies of this nature, significant followup analysis is needed to gain a complete understanding of the inner molecular workings that confer GVL resistance[35,36].

We believe this study has legitimized a new approach to engineering enzymes in a way that offers more versatility that other FACS or plate-reader based assays[37,38]. Being able to strictly control in-vitro enzymatic reactions, especially how long they are allowed to react, allowed us to discriminate highly active enzyme variants from the basal activity of the wild type in our challenge conditions. By generating new libraries based off top mutants and performing additional rounds of screening, we believe we can further optimize this enzyme for biofuel process applications.

Materials and Methods

CMX library generation *RuCelA* genes were amplified using error-prone PCR to introduce random mutations. Error-prone PCR was performed following a protocol calling for 50

μM $MnCl_2$ to decrease the fidelity of Taq polymerase.[39] We did 15 amplification cycles, introducing ~4.5 nucleotide mutations in the gene. We subsequently purified the amplified product, digested it overnight with DpnI to remove remaining wildtype plasmid inserts, and cloned the insert back into pET22b using Circular Polymerase Extension Cloning (CPEC).[40,41]

The CPEC product was purified and used to transform electrocompetent *E. coli* BL21(DE3) cells (Lucigen). Transformed cells were recovered for 45 minutes at 37° C then and diluted into 200 mL of sterile LB media with the added carbenicillin. Once the culture's optical density (OD) approached the lower detection limit of our spectrometer (OD_{600} = 0.2), the culture was concentrated, and freezer stocks of 25% glycerol were made and stored at -80° C. Each library had roughly 10^7 transformants. 10 transformants were picked from each library and their plasmids sequenced to find that each library had 2.5 amino acid substitutions per library member.

Plate reader-based CMX activity assay Individual clones from the mutagenized libraries were incubated in Super Optimal Broth (SOB) for 4 hours at 37° C, or until their turbidity (OD_{600} reached 0.4-0.8. CMX expression was then induced with the addition of 100 μM Isopropyl β-D-1-thiogalactopyranoside (IPTG) and incubated for an additional 12 hours at 30° C.

Cells were pelleted and resuspended in solution 50 mM phosphate buffered saline (PBS), pH 6.2, after decanting the supernatant media to achieve a density of 1 OD_{600}/mL. 200 μL

of resuspended culture was added to a black 96-well plate. 200 μL of assay reagent (0.3x BugBuster (Invitrogen), 100 μM Resorufin Cellobioside (Bachem), 50 kU/mL Lysozyme, 50 mM PBS, pH 6.2) was added to the plate and the fluorescence (excitation at 570 nm, emission at 590 nm) over time measured on a plate reader. Sequences with >50% of the wildtype activity were considered to be functional.

CMX expression, purification, and kinetics *E. coli* BL21(DE3) cells expressing WT CMX and CMX-Top5 were grown in Super Optimal Broth (SOB)for 4 hours at 37° C, or until their turbidity (OD_{600} reached 0.4-0.8. CMX expression was then induced with the addition of 100 μM IPTG and incubated for an additional 12 hours at 30° C. Cells were centrifuged and resuspended in 50 mM PBS, pH 6.2, and lysed by sonication. Cell debris was clear by centrifugation, and subsequently nickel affinity chromatography was used to further purify CMX from the lysate.[7] The purified CMX was concentrated using a non-cellulose based 30 kDa centrifuge filtration unit, and the enzyme concentration measured using the Bradford Assay.

CMX characterization Kinetic parameters were determined by observing 5 mM CMX hydrolytic activity with a titrated range of the Resorufin Cellobioside (Bachem) and fitting the observed initial velocity to the Hill equation. Similarly, GVL inhibition was quantified by observing the initial velocity of CMX hydrolyzing 500 uM Resorufin Cellobioside with varying concentrations of GVL. The data were fit to the dose-response formulation of the

Hill equation to determine the GVL IC_{50}.

Microfluidic device fabrication An initial layer of photoresist resin, SU-8 3010, was coated onto a mirrored silicon wafer (University Wafers) and centrifuged at 1500 rpm to achieve 15 μm layer height. A photomask (**Supp Fig 2.4**) of the first layer of the microfluidic device was placed on the layer and 100 J/cm^2 of UV light is used to polymerize the features. The wafer was baked at 95° C for 10 minutes to catalyze the polymerization. A second 25 μm layer of SU8-3025 was coated onto the wafer by spinning at 4000 rpm, and similarly polymerized with the second photomask (Sup Fig 2b) to create the incubation line and baked again. Undeveloped photoresist is washed off with SU-8 developer (1-methoxy-2-propanol acetate, MicroChem).

The wafer was then used to create a relief in un-polymerized PDMS (Dow Corning Sylgard® 184, 11:1 polymer:cross-linker ratio), which was then polymerized by baking at 75° C. Inlet and outlet holes are punched with a 0.5 mm biopsy corer. The device was then thoroughly washed with isopropanol and double-deionized water and then plasma treated alongside a clean glass microscope slide, to which it was subsequently bonded. Prior to use, microfluidic channels were filled with Aquapel (Pittsburgh Glass Works) to ensure hydrophobicity, and then baked for 10 minutes at 100° C to vaporize any Aquapel left in the channels.

Microfluidic CMX screening 10 μL of CMX library glycerol stocks was used to inoculate 5 mL of SOB for 4 hours at 37° C, or until their turbidity (OD_{600} reached 0.4-0.8. CMX expression was then induced with the addition of 100 μM IPTG and incubated for an additional 12 hours at 30° C. The cultures were pelleted and resuspended in the assay buffer (50 mM PBS, pH 6.2) to a concentration of 0.075 OD_{600} to form the 2x cell suspension. A 2x assay reagent solution of 50 mM PBS, pH 6.2, 0.3x BugBuster (Invitrogen), 20 μM Fluorescein Cellobioside, 50 kU/mL Lysozyme was also made. Both the 2x cell suspension and the 2x assay reagent were loaded into 1 mL luer lock syringes, which were purged of air and fitted with luer-to-PEEK tubing adapters. The cell syringe used PEEK tubing with 0.005" internal diameter, and all other syringes used 0.015" internal diameter PEEK tubing.

Droplets containing expressed CMX library variants were generated at the co-flow drop maker junction. Both the 2x cell suspension and the 2x assay reagents flowed into the device at 15 μL/hr, and were pinched into droplets by fluorinated oil (HFE 7500) containing 1% (wt/wt) PEG–perfluoropolyether amphiphilic block copolymer surfactant flowing at 100 μL/hr.

After incubating on-chip for ~3 minutes, droplets were sorted using electrocoalescense with an aqueous stream of 10 mM Tris, pH 8, 1 mM EDTA. A 473-nm laser was focused onto the channel just upstream of the sorting junction, each droplet was individually excited, and its fluorescence emission measured using a spectrally filtered PMT at 520 nm. A field-programmable gate array card controlled by custom LabVIEW code analyzed the droplet signal at 200 kHz, and if it detected sufficient fluorescence, a train of seven 180-V,

40-kHz pulses was applied by a high-voltage amplifier. This pulse destabilized the interface between the droplet and the adjacent aqueous stream, causing the droplet to merge with the stream via a thin-film instability, after which the droplet contents were injected into the collection stream via its surface tension. The contents of the sorted droplets were collected in a microcentrifuge tube for further processing. Droplets were processed at 800-1000 Hz. Because the cell occupancy of the droplets was 10%, we analyzed 80-100 cells per second. Caspase-3 and -7 libraries were sorted in triplicate over a total of 6 days. In total we analyzed.

DNA recovery and sequencing Recovered plasmid DNA was purified using Zymo spin columns and transformed into ultra-high efficiency 10G Supreme *E. coli* cells (Lucigen). Cells cultured in SOC media and recovered for 45 minutes at 37° C. The recovered culture was then used in totality to inoculate a larger 200 mL culture which was incubated overnight until its OD600 reached 0.5. The larger cultures were pelleted and resuspended in 20 mL 25% glycerol for storage at -80° C. Dilutions of the culture were plated prior to incubation to measure how many transformants were present. We generally observed 0.75–1x as many transformants as what we sorted. Plasmid purified from the larger culture was digested with the restriction enzyme DraIII and ScoI, gel extracted, tagmented using the Nextera XT Library Prepration Kit (Illumina) and sequenced using the Illumina MySeq 2x300. Read coverage for the sequencing runs are displayed in Supplemental Figure 3. Reads with a quality score less than 30 were discarded.

DMS data processing and analysis The reads from the Illumina FASTQ files were mapped to the CMX reference gene using Bowtie2, and translated to amino acid sequences. Mutations observed fewer than 10 times were discarded prior to continuing analysis. Fitness effects of each observed amino acid substitution was estimated using a positive-unlabeled learning framework that compares sequences from the presorted population with the sorted population40,50. Full sequencing databases can be found in the National Center for Bioinformatics Sequence Read Archive (NCBI SRA) under the following accession codes: **CODES** Python and R scripts used to analyze data can be found at https://github.com/RomeroLab/pudms and https://github.com/RomeroLab/DMS-analysis.

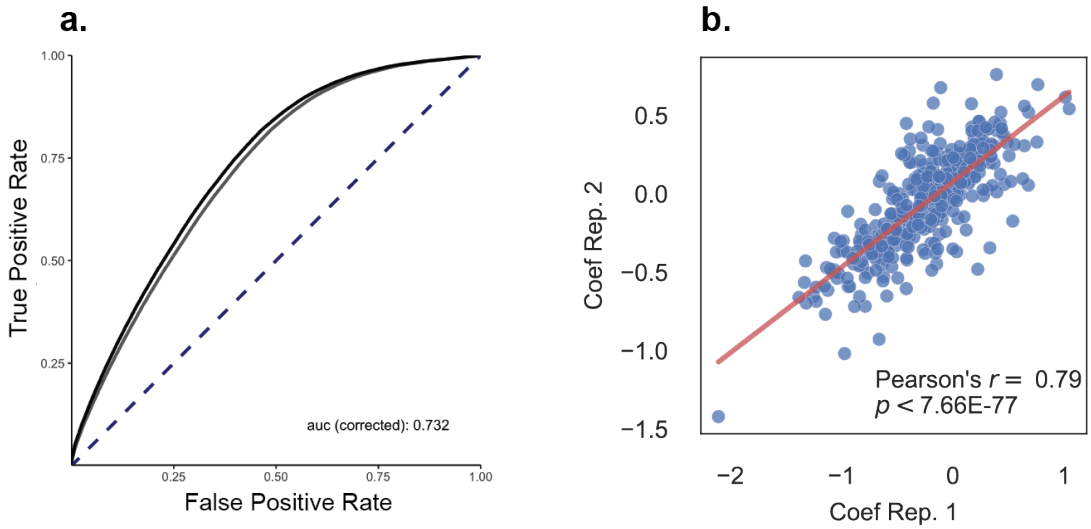
Author Contribution H.R. and P.A.R conceived the project. H.R. performed the experiments and analyzed the data. H.R. and P.A.R. wrote the manuscript.

Funding Statement This work was supported by the Great Lakes Bioenergy Research Consortium (GLBRC). "The content is solely the responsibility of the authors and does not necessarily represent the official views of the Great Lakes Bioenergy Research Consortium."

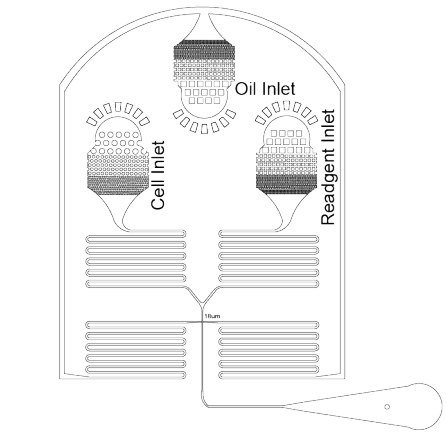
Competing Interest Statement The authors declare that they have no conflicts of interest with the contents of this article.

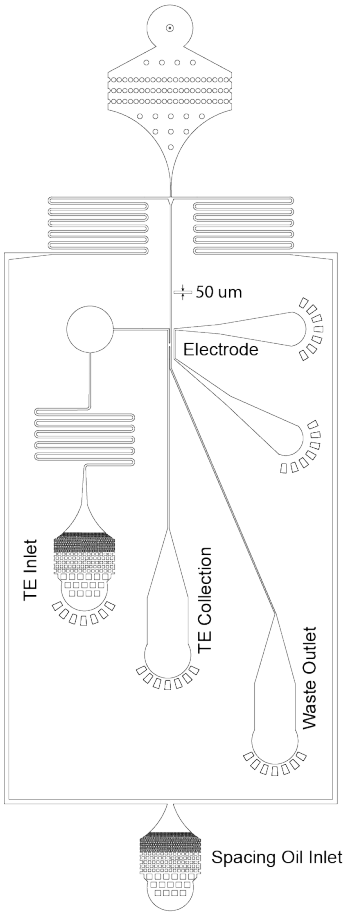
Acknowledgments We would like to acknowledge the UW-Madison Biotechnology Center DNA Sequencing Facility for assistance with Illumina library preparation and sequencing.

Supplement

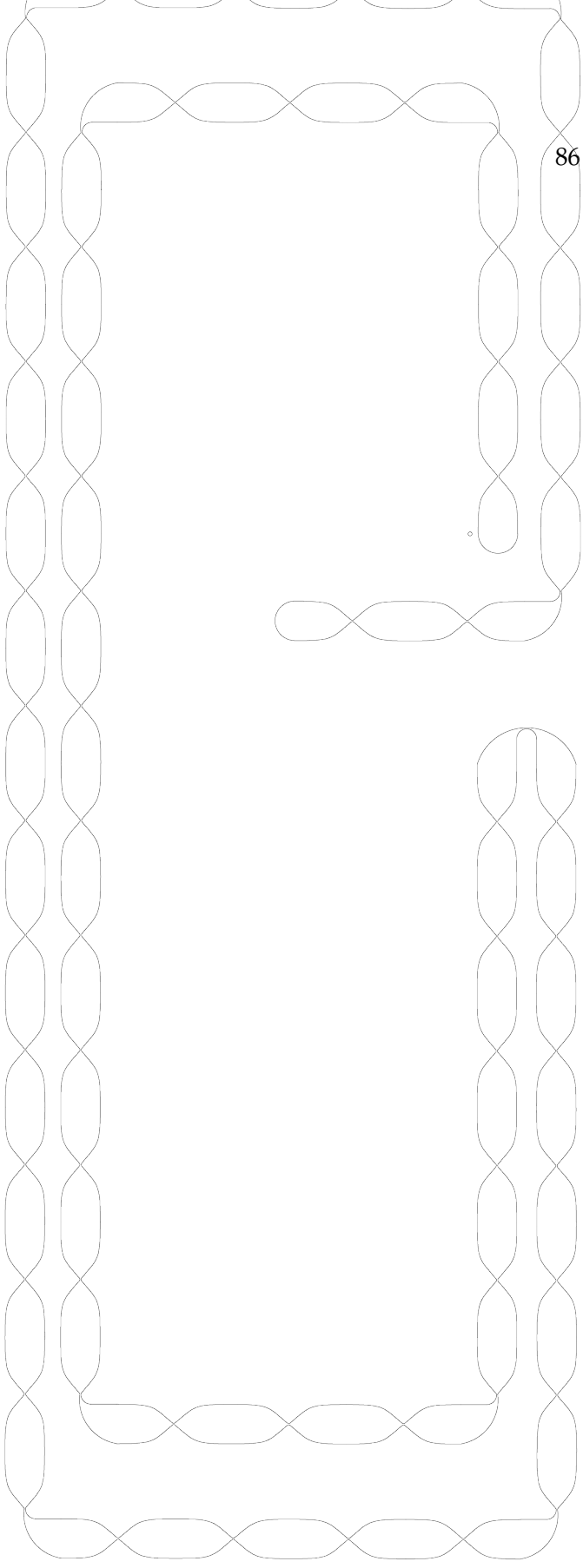


Supplementary Figure 3.1: Correlation of PU Learning coefficients across CMX experimental replicates.. **a**, The Receiver-Operator Curve Area Under Curve (ROC AUC) for each replicate is > 0.5 suggesting that coefficient effects are non random predictors. **b**, A scatter plot mutational coefficients from the two replicates used in our analysis





Layer 1: Drop Maker (top) and Sorter (bottom)



Layer 2: Incubation Line

Supplementary Figure 3.2: Scale schematic of the microfluidic design used in this study.. Layer 1 features the 15 μm-tall drop-maker and sorter. All inlets are labeled as such. Layer 2 is the 50 μm-tall incubation line that connects the drop maker and the sorter. The "sausage-like" repeating pattern serves to randomize the position of droplets transverse to the direction of flow to average out the effects of laminar flow that makes droplets in the center move faster than droplets near the edge of the channel.

References

[1] Serrano-Ruiz, J. C., Luque, R. & Sepúlveda-Escribano, A. Transformations of biomass-derived platform molecules: from high added-value chemicals to fuelsvia aqueous-phase processing. *Chem. Soc. Rev.* **40**, 5266–5281 (2011). URL http://dx.doi.org/10.1039/C1CS15131B.

[2] Shuai, L., Questell-Santiago, Y. M. & Luterbacher, J. S. A mild biomass pretreatment using γ-valerolactone for concentrated sugar production. *Green Chemistry* **18**, 937–943 (2016). URL https://pubs.rsc.org/en/content/articlehtml/2016/gc/c5gc02489ghttps://pubs.rsc.org/en/content/articlelanding/2016/gc/c5gc02489g.

[3] Mellmer, M. A., Martin Alonso, D., Luterbacher, J. S., Gallo, J. M. R. & Dumesic, J. A. Effects of γ-valerolactone in hydrolysis of lignocellulosic biomass to monosaccharides. *Green Chemistry* **16**, 4659–4662 (2014).

[4] Luterbacher, J. S. *et al.* Nonenzymatic sugar production from biomass using biomass-derived γ-valerolactone. *Science* **343**, 277–280 (2014).

[5] Chaturvedi, V. & Verma, P. An overview of key pretreatment processes employed for bioconversion of lignocellulosic biomass into biofuels and value added products. *3 Biotech* **3**, 415–431 (2013). URL https://doi.org/10.1007/s13205-013-0167-8.

[6] Chang, L. *et al.* Characterization of a bifunctional xylanase/endoglucanase from yak rumen microorganisms. *Applied Microbiology and Biotechnology* **90**, 1933–1942 (2011). URL `https://link.springer.com/article/10.1007/s00253-011-3182-x`.

[7] Glasgow, E., Vander Meulen, K., Kuch, N. & Fox, B. G. Multifunctional cellulases are potent, versatile tools for a renewable bioeconomy. *Current Opinion in Biotechnology* **67**, 141–148 (2021).

[8] Romero, P. A., Tran, T. M. & Abate, A. R. Dissecting enzyme function with microfluidic-based deep mutational scanning. *Proc. Nat. Adad. Sci.* **112**, 7159–7164 (2015). URL `http://www.pnas.org/content/112/23/7159.full.pdf`.

[9] Song, H., Bremer, B. J., Hinds, E. C., Raskutti, G. & Romero, P. A. Inferring Protein Sequence-Function Relationships with Large-Scale Positive-Unlabeled Learning. *Cell Systems* **12**, 92–101.e8 (2021).

[10] Francis, J. & Hansche, P. Directed evolution of metabolic pathways in microbial populations. i. modification of the acid phosphatase ph optimum in s. cerevisiae. *Genetics* **70**, 59–73 (1972).

[11] Chowdhury, R. & Maranas, C. D. From directed evolution to computational enzyme engineering—A review. *AIChE Journal* **66**, e16847 (2020). URL `https://onlinelibrary.wiley.com/doi/full/10.1002/aic.16847https:`

//onlinelibrary.wiley.com/doi/abs/10.1002/aic.16847https://aiche.onlinelibrary.wiley.com/doi/10.1002/aic.16847.

[12] Bottoms, S. *et al.* Chemical genomic guided engineering of gamma-valerolactone tolerant yeast. *Microbial cell factories* **17**, 1–12 (2018).

[13] Griffiths, A. D. & Tawfik, D. S. Miniaturising the laboratory in emulsion droplets. *Trends in biotechnology* **24**, 395–402 (2006).

[14] Mair, P., Gielen, F. & Hollfelder, F. Exploring sequence space in search of functional enzymes using microfluidic droplets. *Current Opinion in Chemical Biology* **37**, 137–144 (2017).

[15] Basova, E. Y. & Foret, F. Droplet microfluidics in (bio) chemical analysis. *Analyst* **140**, 22–38 (2015).

[16] Roychowdury, H. & Romero, P. A. Microfluidic deep mutational scanning of the human executioner caspases reveals differences in structure and regulation. *Cell Death Discovery* **8**, 1–8 (2022).

[17] Starr, T. N. & Thornton, J. W. Epistasis in protein evolution (2016). URL http://doi.wiley.com/10.1002/pro.2897.

[18] Sims, R. *et al.* Transport climate change 2014: Mitigation of climate change. contribution of working group iii to the fifth assessment report of the intergovernmental panel on climate change ed o edenhofer et al. *Cambridge and New York: Cambridge University Press.*

Available at: http://www. ipcc. ch/pdf/assessment-report/ar5/wg3/ipcc_wg3_ar5_chapter8. pdf (2014).

[19] Chen, H. Biotechnology of lignocellulose. *Theory and Practice. China: Chemical Industry Press and Springer* (2014).

[20] Houfani, A. A., Anders, N., Spiess, A. C., Baldrian, P. & Benallaoua, S. Insights from enzymatic degradation of cellulose and hemicellulose to fermentable sugars–a review. *Biomass and Bioenergy* **134**, 105481 (2020).

[21] Yoo, C. G., Zhang, S. & Pan, X. Effective conversion of biomass into bromomethylfurfural, furfural, and depolymerized lignin in lithium bromide molten salt hydrate of a biphasic system. *RSC advances* **7**, 300–308 (2017).

[22] Zhang, L., Xi, G., Zhang, J., Yu, H. & Wang, X. Efficient catalytic system for the direct transformation of lignocellulosic biomass to furfural and 5-hydroxymethylfurfural. *Bioresource technology* **224**, 656–661 (2017).

[23] Huang, C. *et al.* Lignin-enzyme interaction: A roadblock for efficient enzymatic hydrolysis of lignocellulosics. *Renewable and Sustainable Energy Reviews* **154** (2022).

[24] Jang, Y.-S., Malaviya, A., Cho, C., Lee, J. & Lee, S. Y. Butanol production from renewable biomass by clostridia. *Bioresource technology* **123**, 653–663 (2012).

[25] Ling, H. Batch submerged fermentation in shake flask culture and bioreactor: influence of different agricultural residuals as the substrate on the optimization of xylanase

production by bacillus subtilis and aspergillus brasiliensis. *J Appl Biotechnol Bioeng* **1**, 96–104 (2016).

[26] Sweeney, M. D. & Xu, F. Biomass converting enzymes as industrial biocatalysts for fuels and chemicals: recent developments. *Catalysts* **2**, 244–263 (2012).

[27] Yang, K. K., Wu, Z., Bedbrook, C. N. & Arnold, F. H. Learned protein embeddings for machine learning. *Bioinformatics* **34**, 2642–2648 (2018).

[28] Siedhoff, N. E., Schwaneberg, U. & Davari, M. D. Machine learning-assisted enzyme engineering. *Methods in Enzymology* **643**, 281–315 (2020).

[29] Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118** (2021).

[30] Hie, B. L. & Yang, K. K. Adaptive machine learning for protein engineering. *Current opinion in structural biology* **72**, 145–152 (2022).

[31] Houfani, A. A., Anders, N., Spiess, A. C., Baldrian, P. & Benallaoua, S. Insights from enzymatic degradation of cellulose and hemicellulose to fermentable sugars– a review (2020).

[32] Firnberg, E., Labonte, J. W., Gray, J. J. & Ostermeier, M. A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape. *Molecular Biology and Evolution* **31**, 1581–1592 (2014). URL `https://academic.oup.com/mbe/article/31/6/1581/2925654`.

[33] Haddox, H. K., Dingens, A. S. & Bloom, J. D. Experimental Estimation of the Effects of All Amino-Acid Mutations to HIV's Envelope Protein on Viral Replication in Cell Culture. *PLoS Pathogens* **12** (2016). URL `http://journals.plos.org/plospathogens/article/file?id=10.1371/journal.ppat.1006114&type=printable`.

[34] Hietpas, R. T., Jensen, J. D. & Bolon, D. N. Experimental illumination of a fitness landscape. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 7896–7901 (2011). URL `http://www.ncbi.nlm.nih.gov/pubmed/21464309http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3093508`.

[35] Fowler, D. M., Stephany, J. J. & Fields, S. Measuring the activity of protein variants on a large scale using deep mutational scanning (2014). URL `http://www.nature.com/doifinder/10.1038/nprot.2014.153`.

[36] Fowler, D. M. & Fields, S. Deep mutational scanning: A new style of protein science (2014). URL `http://www.nature.com/articles/nmeth.3027`.

[37] Stavrakis, S., Holzner, G., Choo, J. & DeMello, A. High-throughput microfluidic imaging flow cytometry. *Current opinion in biotechnology* **55**, 36–43 (2019).

[38] Qin, Y. *et al.* A fluorescence-activated single-droplet dispenser for high accuracy single-droplet and single-cell sorting and dispensing. *Analytical chemistry* **91**, 6815–6819 (2019).

[39] Bloom, J. D. *et al.* Evolution favors protein mutational robustness in sufficiently large populations. *BMC Biology* **5**, 1–21 (2007). URL https://link.springer.com/articles/10.1186/1741-7007-5-29https://link.springer.com/article/10.1186/1741-7007-5-29.

[40] Quan, J. & Tian, J. Circular polymerase extension cloning of complex gene libraries and pathways. *PLoS ONE* **4**, 6441 (2009). URL www.plosone.org.

[41] Quan, J. & Tian, J. Circular polymerase extension cloning for high-throughput cloning of complex and combinatorial DNA libraries. *Nature Protocols* **6**, 242–251 (2011). URL http://www.nature.com/doifinder/10.1038/nprot.2010.181.

Chapter 4

CONCLUSION

HSR wrote this chapter.

A brief summary

It seems unlikely that enzymes will lose their prominence as crucial scientific tools, as therapeutic targets, and in their burgeoning role as green industrial catalysts.[1–3] As society's specific enzymatic needs evolve (so to speak), we must continue to develop increasingly sophisticated, efficient, and effective means to engineer proteins to suit them.[4–7] Our ability to propose beneficial mutations to and design enzymes de novo continues to grow as we gain ever deeper fundamental understanding of proteins and the intramolecular mechanisms that govern their behavior.[8,9] Nascent machine learning methods also continue to provide increasingly accurate predictions of new protein sequences.[10,11]

Given the inherent physical complexity and heterogeneity of proteins and enzymes, however, I suspect that we will never be able to rationally predict all possible improved variants that may exist in the vastness of sequence space.[4,12–15] Directed evolution will therefore maintain its relevance as a fast and effective way to scan sequence space for better enzymes. By coupling massive, genetically diverse libraries of enzyme variants with ultra-high-throughput screening techniques, we gain an agnostic understanding of the enzyme's local sequence-function landscape capable of making predictions otherwise lost in rational design strategies. Effective protein engineering campaigns exploit both rational design and directed evolution, and indeed, it is rare to see contemporary approaches that do not leverage both.[16–19]

This dissertation presented a novel platform that enables ultra-high-throughput en-

zyme screening. In contrast to its nearest cousin, FACS, which limits enzymatic analysis to intracellular and membrane-display based methods,[20,21] droplet microfluidic devices allow us to encapsulate and lyse single cells and fluorogenically assay the resultant in vitro enzymatic reaction after a controlled incubation period.[4,15,18] With strict control over enzyme reaction time, we could screen enzyme libraries on the basis of their velocity, rather than total substrate turnover.[18]

We demonstrated our platform's utility by generating useful and comparative mutational mappings of the nearly identical apoptotic proteases, Caspase-3 and Caspase-7. Our screen pinpointed key functional differences between CASP3 and CASP7, including a previously reported secondary internal cleavage site, CASP7 Q196 that is not present in CASP3. Our results will open avenues for inquiry in caspase function and regulation that could potentially inform the development of future caspase-specific therapeutics.[18]

We also demonstrated our platforms utility in the particularly difficult enzyme engineering sub-genre of evolving solvent tolerance. Selection experiments to evolve such resilience are practically impossible to design, computational models poorly predict solvent tolerance, and most screening methods lack the throughput necessary to scan broad enough sections of protein sequence space to find solvent resilient hits. Using our platform we engineered the promiscuous glycoside hydrolase, CMX, for significantly increased tolerance to the green industrial solvent γ-valerolactone.

We believe this platform serves as a useful new addition to the protein engineer's toolbox, especially for enzyme engineering.

Considerations and Limitations

Our platform has some key limitations. Firstly, the effectiveness of our HTS platform is beholden to our choices in assays.[22–24] Our hardware limited our assays to fluorescent and luminescent measurements. While other spectroscopic methods, like fourier-transformed infrared (FTIR) and Raman spectroscopy, have been demonstrated to work on microfluidic screening platforms, they are significantly more difficult to implement and requires further specialized instrumentation, limiting the platform's generalizability.[25,26] Such methods also have comparatively limited throughput, making it more difficult to screen large libraries effectively.[27]

Secondly, it is practically impossible to control the intra-droplet enzyme concentration.[4] There is often massive cell-to-cell variance in enzyme production. One must consider that highly expressed but comparatively inactive enzyme variants may produce comparable signal to worse expressed but much more active enzyme variants—this may increase the noise in downstream analysis resulting in false positive hits.[18,28] Generalizing further, screens of this nature cannot be used to glean mechanistic insights about enzyme variants—further downstream biochemical analysis and validation are always required.[13,20]

Lastly, designing, fabricating, and utilizing microfluidic devices required not insignificant expertise and training. Most microfluidic devices and workflows in academic laboratories are designed and fabricated in-house and lack the standardization and intercompatibility of similar commercial options, like FACS.[27,29] Coupled with the myriad

quirks and particularities of each new design for each new experiment, learned through often extensive trial, error, and optimization, there is often a massive entry cost for researchers; it is difficult to learn on ones own.

Future Directions

There are several avenues of research continuing from the work in this dissertation, both from the particular use-cases described in chapters 2 and 3, but also in continuing further engineering of the screening platform itself.

Comparing Caspases, and perhaps other enzymes too. Deep mutational analysis of Caspase-3 and -7 revealed several putative divergences in their sequence function landscapes. While the study superficially explored a few of them by characterizing the kinetics of key mutants, it would be prudent to A) launch a significantly more in depth study on their potential physiological effects, and B) exploit those sites to design targeted small-molecule modulators. Selecting a few highly divergent mutants of Caspase-3 and Caspase-7 for mass spectrometry based proteomic analysis to probe changes in their native physiological interacting partners.[30]—this may lead to insights regarding their regulation in apoptosis and pyroptosis.

Researchers could leverage those latent biophysical and regulatory insights from further biochemical analysis and Rosetta modeling to screen small molecule libraries in silico, potentially leading to novel therapeutics capable of hyper-specifically targeting individual

caspases.[31,32] One may even consider generalizing such an approach to other potentially therapeutic protein targets. New approaches to understanding differences in proteins that circumvent costly and difficult structural analyses could greatly accelerate drug design by offering new target exosites on proteins to exploit.[33–35]

Engineering glycoside hydrolases, and other industrially relevant enzymes, for solvent tolerance Glycoside hydrolases like CMX are promising biocatalysts for industrial biofuel production, but like many enzymes, poorly tolerate industrial solvent conditions. Though it took just one iteration of screening to identify a variant of CMX that was resilient to the 3% GVL solvent condition we challenged it with, there are still key improvements that can be made. Solvent resilience is a major difficulty in engineering industrial biocatalysts, and this platform may extend the toolkit researchers use to approach that problem.[36–38] This design approach would be greatly aided by using ML and biophysical modeling to create better targeted libraries for screening. Only ~5% of our error-prone PCR library showed activity in the presence of GVL, meaning ~95% of our measurements wasted resources and gained us no new information. By generating more DMS data related to protein solvent resilience, perhaps we could train ML models to predict more functional libraries and increase the chances of successfully engineering better enzymes. An iterative and active learning approach to exploring sequence space as it relates to organic solvents could lead to greater biophysical insights about how solvents interact with proteins—further accelerating the design process.

Improvements and additions to the microfluidic screening platform. Finally, though our microfluidic platform has already gone through numerous iterations and optimizations since the start of this dissertation work, many improvements remain to be made. Most importantly, we should improve its accessibility and ease of use for future researchers, in our own lab and in others. When first embarking on this dissertation research with the earliest iterations of our platform, carrying through a successful screen had just a ~5% success rate. Even toward the end of these research projects, it was exceedingly rare to perform two consecutively successful sorting experiments.

The points of failure were many and diverse. For instance, the poly-dimethyl-sulfoxane polymer (PDMS) comprising the device would delaminate from the glass slide it was bonded to, making it impossible to create and direct emulsions, let alone assay and screen them. It turned out that the plasma-bonding process we used to bond PDMS devices to glass slides was very humidity dependent, and so in the summer months we would see few successful experiments compared to the winter months. Also for instance, despite how well-filtered our injected solutions were, the devices' fine, nanometer scale channels would regularly clog from whatever few particulates remained in suspension. For a more comprehensive list of foibles, Appendix A has them documented for troubleshooting. Finding better and less ad hoc solutions to these myriad problems will certainly spare the sanity of future graduate students wrangling with the platform.

In addition to quality-of-life improvements, one can also imagine expanding the capabilities of the platform. In the most ambitions cases, others have coupled their droplet

microfluidic devices to mass spectrometers.[39] Given aforementioned concerns, it may be more prudent to first start with simpler expansions like FTIR or Raman spectroscopy, both of which could greatly expand the possible assays we could screen based off.

Conclusions We have demonstrated that our platform has broad utility in the realms of enzyme engineering and related high-throughput screening enterprises. The ability to screen enzymes in high-throughput based on their velocity, which was previously limited to plate-based assays, has helped us engineer solvent resistances and generate comparative sequence-function mappings of enzymes that can be leveraged to find unique features for such endeavors as drug design. We hope that researchers going forward will find this platform and these demonstrations useful in their own ongoing efforts.

References

[1] Madhavan, A. *et al.* Design of novel enzyme biocatalysts for industrial bioprocess: Harnessing the power of protein engineering, high throughput screening and synthetic biology. *Bioresource Technology* **325**, 124617 (2021).

[2] Luetz, S., Giver, L. & Lalonde, J. Engineered enzymes for chemical production. *Biotechnology and bioengineering* **101**, 647–653 (2008).

[3] Ali, M., Ishqi, H. M. & Husain, Q. Enzyme engineering: Reshaping the biocatalytic functions (2020). URL `https://onlinelibrary.wiley.com/doi/full/10.1002/bit.27329https://onlinelibrary.wiley.com/doi/abs/10.1002/bit.27329https://onlinelibrary.wiley.com/doi/10.1002/bit.27329`.

[4] Romero, P. A., Tran, T. M. & Abate, A. R. Dissecting enzyme function with microfluidic-based deep mutational scanning. *Proc. Nat. Adad. Sci.* **112**, 7159–7164 (2015). URL `http://www.pnas.org/content/112/23/7159.full.pdf`.

[5] Basova, E. Y. & Foret, F. Droplet microfluidics in (bio) chemical analysis. *Analyst* **140**, 22–38 (2015).

[6] Chowdhury, R. & Maranas, C. D. From directed evolution to computational enzyme engineering—A review. *AIChE Journal* **66**, e16847 (2020). URL `https://onlinelibrary.wiley.com/doi/full/10.1002/aic.16847https:`

//onlinelibrary.wiley.com/doi/abs/10.1002/aic.16847https://aiche.onlinelibrary.wiley.com/doi/10.1002/aic.16847.

[7] Orencia, M. C., Yoon, J. S., Ness, J. E., Stemmer, W. P. & Stevens, R. C. Predicting the emergence of antibiotic resistance by directed evolution and structural analysis. *Nature structural biology* **8**, 238–242 (2001).

[8] Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118** (2021).

[9] Chen, C.-Y., Georgiev, I., Anderson, A. C. & Donald, B. R. Computational structure-based redesign of enzyme activity. *Proceedings of the National Academy of Sciences* **106**, 3764–3769 (2009).

[10] Mazurenko, S., Prokop, Z. & Damborsky, J. Machine learning in enzyme engineering. *ACS Catalysis* **10**, 1210–1223 (2019).

[11] Siedhoff, N. E., Schwaneberg, U. & Davari, M. D. Machine learning-assisted enzyme engineering. *Methods in Enzymology* **643**, 281–315 (2020).

[12] Turowec, J. P. *et al.* An unbiased proteomic screen reveals caspase cleavage is positively and negatively regulated by substrate phosphorylation. *Molecular 'I&' cellular proteomics : MCP* **13**, 1184–97 (2014). URL

http://www.ncbi.nlm.nih.gov/pubmed/24556848http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4014278.

[13] Starr, T. N. & Thornton, J. W. Epistasis in protein evolution (2016). URL http://doi.wiley.com/10.1002/pro.2897.

[14] Hietpas, R. T., Jensen, J. D. & Bolon, D. N. Experimental illumination of a fitness landscape. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 7896–7901 (2011). URL http://www.ncbi.nlm.nih.gov/pubmed/21464309http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3093508.

[15] Romero, P. A. & Arnold, F. H. Exploring protein fitness landscapes by directed evolution. *Nature reviews Molecular cell biology* **10**, 866–876 (2009).

[16] Wittmann, B. J., Johnston, K. E., Wu, Z. & Arnold, F. H. Advances in machine learning for directed evolution. *Current opinion in structural biology* **69**, 11–18 (2021).

[17] Bloom, J. D. *et al.* Evolution favors protein mutational robustness in sufficiently large populations. *BMC Biology* **5**, 1–21 (2007). URL https://link.springer.com/articles/10.1186/1741-7007-5-29https://link.springer.com/article/10.1186/1741-7007-5-29.

[18] Roychowdury, H. & Romero, P. A. Microfluidic deep mutational scanning of the human executioner caspases reveals differences in structure and regulation. *Cell Death Discovery* **8**, 1–8 (2022).

[19] Song, H., Bremer, B. J., Hinds, E. C., Raskutti, G. & Romero, P. A. Inferring Protein Sequence-Function Relationships with Large-Scale Positive-Unlabeled Learning. *Cell Systems* **12**, 92–101.e8 (2021).

[20] Firnberg, E., Labonte, J. W., Gray, J. J. & Ostermeier, M. A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape. *Molecular Biology and Evolution* **31**, 1581–1592 (2014). URL `https://academic.oup.com/mbe/article/31/6/1581/2925654`.

[21] Ali, M., Ishqi, H. M. & Husain, Q. Enzyme engineering: Reshaping the biocatalytic functions. *Biotechnology and bioengineering* **117**, 1877–1894 (2020).

[22] Mair, P., Gielen, F. & Hollfelder, F. Exploring sequence space in search of functional enzymes using microfluidic droplets. *Current Opinion in Chemical Biology* **37**, 137–144 (2017).

[23] Ng, E. X., Miller, M. A., Jing, T. & Chen, C. H. Single cell multiplexed assay for proteolytic activity using droplet microfluidics. *Biosensors and Bioelectronics* **81**, 408–414 (2016). URL `http://www.sciencedirect.com/science/article/pii/S0956566316301932?_rdoc=1&_fmt=high&_origin=gateway&_docanchor=&md5=b8429449ccfc9c30159a5f9aeaa92ffb&dgcid=raven_sd_recommender_email`.

[24] Romero, P. A., Tran, T. M. & Abate, A. R. Dissecting enzyme function with microfluidic-based deep mutational scanning. *Proceedings of the National Academy of Sciences* **112**, 7159–7164 (2015).

[25] Chrimes, A. F., Khoshmanesh, K., Stoddart, P. R., Mitchell, A. & Kalantar-Zadeh, K. Microfluidics and raman microscopy: current applications and future challenges. *Chemical Society Reviews* **42**, 5880–5906 (2013).

[26] Joly, M. *et al.* Scanning aperture approach for spatially selective atr-ftir spectroscopy: Application to microfluidics. *Analytical Chemistry* **93**, 14076–14087 (2021).

[27] Piyasena, M. E. & Graves, S. W. The intersection of flow cytometry with microfluidics and microfabrication. *Lab on a Chip* **14**, 1044 (2014). URL `http://xlink.rsc.org/?DOI=c3lc51152a`.

[28] Frenz, L., Blank, K., Brouzes, E. & Griffiths, A. D. Reliable microfluidic on-chip incubation of droplets in delay-lines. *Lab on a Chip* **9**, 1344–1348 (2009). URL `https://pubs.rsc.org/en/content/articlehtml/2009/lc/b816049jhttps://pubs.rsc.org/en/content/articlelanding/2009/lc/b816049j`.

[29] Holtze, C. Large-scale droplet production in microfluidic devices - An industrial perspective. *Journal of Physics D: Applied Physics* **46**, 114008 (2013). URL `https://iopscience.iop.org/article/10.1088/0022-3727/46/11/114008https://iopscience.iop.org/article/10.1088/0022-3727/46/11/114008/meta`.

[30] Cho, K. F. *et al.* Proximity labeling in mammalian cells with turboid and split-turboid. *Nature Protocols* **15**, 3971–3999 (2020).

[31] Agniswamy, J., Fang, B. & Weber, I. T. Conformational similarity in the activation of caspase-3 and -7 revealed by the unliganded and inhibited structures of caspase-7. *Apoptosis* **14**, 1135–1144 (2009). URL http://www.pymol.orghttp://link.springer.com/10.1007/s10495-009-0388-9.

[32] Häcker, H. G., Sisay, M. T. & Gütschow, M. Allosteric modulation of caspases (2011). URL http://www.sciencedirect.com/science/article/pii/S0163725811001604.

[33] Krishna Deepak, R. N., Abdullah, A., Talwar, P., Fan, H. & Ravanan, P. Identification of FDA-approved drugs as novel allosteric inhibitors of human executioner caspases. *Proteins: Structure, Function and Bioinformatics* **86**, 1202–1210 (2018). URL https://onlinelibrary.wiley.com/doi/full/10.1002/prot.25601https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.25601https://onlinelibrary.wiley.com/doi/10.1002/prot.25601.

[34] Kudelova, J., Fleischmannova, J., Adamova, E. & Matalova, E. Pharmacological caspase inhibitors: research towards therapeutic perspectives. *Journal of physiology and pharmacology : an official journal of the Polish Physiological Society* **66**, 473–82 (2015). URL http://www.ncbi.nlm.nih.gov/pubmed/26348072.

[35] MacKenzie, S. H., Schipper, J. L. & Clark, A. C. The potential for caspases in drug discovery (2010). URL http://www.ncbi.nlm.nih.gov/pubmed/20812148http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3289102.

[36] Bilal, M., Asgher, M., Cheng, H., Yan, Y. & Iqbal, H. M. Multi-point enzyme immobilization, surface chemistry, and novel platforms: a paradigm shift in biocatalyst design. *Critical reviews in biotechnology* **39**, 202–219 (2019).

[37] Bilal, M. & Iqbal, H. Tailoring multipurpose biocatalysts via protein engineering approaches: a review. *Catalysis Letters* **149**, 2204–2217 (2019).

[38] Ismail, A. R., Kashtoh, H. & Baek, K.-H. Temperature-resistant and solvent-tolerant lipases as industrial biocatalysts: Biotechnological approaches and applications. *International Journal of Biological Macromolecules* **187**, 127–142 (2021).

[39] Diefenbach, X. W. *et al.* Enabling biocatalysis by high-throughput protein engineering using droplet microfluidics coupled to mass spectrometry. *ACS omega* **3**, 1498–1508 (2018).

COLOPHON

This document was typeset with LaTeX 2$_\varepsilon$ using TeXShop. It is based on the University of Wisconsin dissertation template created by William C. Benton (available at `https://github.com/willb/wi-thesis-template`).