

DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY


Evolution and Computational Generation of Highly Functionalized Nucleic Acid Polymers

Citation

Chen, Jonathan Chris. 2021. Evolution and Computational Generation of Highly Functionalized Nucleic Acid Polymers. Doctoral dissertation, Harvard University Graduate School of Arts and Sciences.

Link

https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37368339

Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at
https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFlOWI2M2EwYTg

Accessibility

https://accessibility.huit.harvard.edu/digital-accessibility-policy

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story

HARVARD UNIVERSITY
Graduate School of Arts and Sciences



DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Department of

have examined a dissertation entitled

presented by

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature _____

Typed name: Prof.

Signature _____

Typed name: Prof.

Signature _____

Typed name: Prof.

Signature _____

Typed name: Prof.

Signature _____

Typed name: Prof.

Date:

Evolution and Computational Generation of Highly Functionalized Nucleic Acid Polymers

A dissertation presented

by

Jonathan C. Chen

to

The Graduate School of Arts and Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Chemical Biology

Harvard University

Cambridge, Massachusetts

April 2021

© 2021 Jonathan C. Chen

All rights reserved.

Dissertation Advisor: David R. Liu Jonathan C. Chen

Evolution and Computational Generation of Highly Functionalized Nucleic Acid Polymers

Abstract

In vitro selection enables the identification of polymers with fundamental activities such as target binding and reaction catalysis. While *in vitro* selections can begin with combinatorial libraries approaching ~10^{15} distinct molecules, these initial libraries often lack sufficient chemical functionality to support high on-target activity. Here we describe a ligase-mediated DNA-templated polymerization system with access to 32 building blocks containing eight chemically-diverse side-chains on a DNA backbone. This sequence-defined synthetic polymer system, highly-functionalized nucleic acid polymers (HFNAPs), supports iterated rounds of *in vitro* selection against protein and small molecule targets, yielding binders for PCSK9, IL-6, and daunomycin, all with nanomolar K_D's. Using high-throughput sequencing (HTS) to gain a deeper understanding of *in vitro* selection progression, we demonstrate the effectiveness of re-selection and evolution, while improving selection stringency by biasing for slow-off rates.

HTS also enables us to examine the role individual side-chains play in binding interactions and translation efficiency. Our neutral selections elucidate translation biases from side-chain inclusion and ligase sequence-preferences. The structure-activity relationship studies reveal that specific side-chains are necessary for binding, and our parallel selections with seven genetic codes demonstrate the critical importance of non-polar side-chains.

Despite these successful selection campaigns, *in vitro* selections sample only a miniscule fraction of the total HFNAP sequence space. Selection-induced sequence convergence and limited sequencing depth further constrain our ability to explore and understand *in vitro* selection fitness landscapes. Therefore, we trained a conditional variational autoencoder (CVAE) machine learning model on *in vitro* selection data, to learn the relationship

between sequence identity and binding affinity. Remarkably, the trained CVAE generated diverse and novel HFNAP sequences with binding affinities (K_D=13-15 nM) similar to those of the most active HFNAPs from *in vitro* selection, even though the CVAE-generated polymers lack sequence similarity with experimentally identified sequences. Select CVAE-generated polymers have secondary structure conservation with *in-vitro*-selection-derived polymers, demonstrating the model's ability to map the global fitness landscape for daunomycin binding. Coupling *in vitro* selection with a machine learning model that learned the fitness landscape of a binding task thus enabled direct generation of active variants, demonstrating a new approach to the discovery of highly active biopolymers.

Table of Contents

Title Page i
Copyright ii
Abstract iii
Table of Contents v
Acknowledgements vii
List of Figures x
Chapter 1 Introduction to in vitro selections and functionalized nucleic acid polymers 1
1.1 In vitro selections and directed evolution 2
1.2 Modified nucleic acids 5
1.2.1 Backbone and sugar-modified nucleic acids 5
1.2.2 Nucleobase functionalization expands chemical diversity 8
1.2.3 DNA-templated synthesis of highly functionalized nucleic acid polymers 11
1.3 Selection stringency and the size of search spaces 12
1.4 Conclusion and outlook 14
1.5 References 16
Chapter 2 Evolution of Sequence-Defined Highly Functionalized Nucleic Acid Polymers 29
2.1 Introduction 30
2.2 Improved HFNAP building blocks and translation system 32
2.3 In vitro selections against PCSK9 34
2.4 Diversification and evolution of HFNAPs for PCSK9 binding 38
2.5 In vitro selections against IL-6 43
2.5 Discussion 45
2.6 Methods 47
2.7 Acknowledgments 55
References 55
Chapter 3 Side-chain determinants of biopolymer function during selection and replication 61
3.1 Introduction 62
3.2 Parallel selections with seven different genetic codes 65
3.3 Head-to-head binding competition with seven different genetic codes 72
3.4 Examining translation bias in the HFNAP system 76
3.5 IL-6 selection outcomes 79
3.6 Discussion 82
3.7 Methods 84
3.8 Acknowledgments 95
References 95
Chapter 4 In vitro selection for HFNAPs with nanomolar affinity to a small molecule target 100
4.1 Introduction 101
4.2 Initial selection for HFNAPs that bind daunomycin from naïve libraries 102
4.3 Re-selection with increased stringency 106
4.4 Characterization of high-stringency selection outcomes 109
4.5 Discussion 114
4.6 Methods 116
References 122
Chapter 5 Machine learning generation of HFNAPs with nanomolar affinity to daunomycin 126
5.1 Introduction 127
5.2 Conceptual overview of generative machine learning model methodology 129

5.3 Conditional variational autoencoder (CVAE) architecture and model training 134
5.4 Characterization of CVAE-generated HFNAPs 140
5.5 Sequence diversity of CVAE-generated HFNAPs 145
5.6 CVAE-generated sequences conserve selection-2 secondary structures 149
5.7 Discussion 153
5.8 Methods 155
References 165
Chapter 6 Appendix 170
Chapter 2 Supplementary information 171
Chapter 3 Supplementary Information 183
Chapter 4 Supplementary Information 207
Chapter 5 Supplementary Information 216

Acknowledgements

To David, my mentor, from whom I've learned so much from during my time in Boston. I am grateful that you took a chance on someone with such diverse research interests, whose scientific path has always been less than straightforward. When I joined the DTS subgroup with the intention to learn about DNA and chemistry, you were nothing but supportive, even though I had never done organic synthesis before. You gave me the freedom to explore my research interests in a wide variety of areas, and have built an amazing group of scientists from whom I've learned so much. In particular, I am grateful for your mentorship when it comes to writing. Writing isn't and has never been a strong suit of mine, but through this iterative process, and your unending patience, I can begin to understand effective scientific communication.

I also don't think that this thesis project, generating HFNAPs, would have been possible anywhere else. The lab is often in a unique position of being able to attempt experiments because of the resources that you've assembled. While I'm appreciative of these advantages, it is your willingness to allow students to explore that I am most grateful. I always prefer to learn something myself, even if it takes more time than handing it off to someone else. Despite the fast pace nature of much of the science in the lab, I have always felt that you supported my scientific journey, as I set out to learn in a wide variety of areas. Without your support, I wouldn't have had the opportunity to learn both chemistry and machine learning. I still remember going to the ChemBio recruitment dinner where you asked me commit to Harvard on the spot. I wouldn't be the scientist I am today, had I gone anywhere else.

To my committee, thank you for your constant guidance and helpful suggestions. I've learned so much in our discussions together, and thank you for taking the time and effort to teach me about being a scientist. Your suggestions have impacted how I conduct science, from planning,

to execution, to analysis, and I am so grateful to have had the opportunity to learn from your combined set of expertise.

To CBP, thank you Dan, Suzanne, and Jason, for supporting me in my journey. Thank you for hosting such great recruiting events, that convinced me to come here. Thank you also to Ralph, for being so down to earth. I'm inspired by the strengths of your science and your character.

To the Liu Group, thank you for all the happy memories and the constant willingness to teach me new things. When I joined the lab, I knew so little chemistry, but thanks to you, JP, I have learned so much. Thank you for being so patient with my constant questions, and thank you for all the time you spent teaching me chemistry. Thank you to all of the DTS subgroup members, past and present, from whom I've learned so much about this area. Thank you to Phillip for mentoring me, and Zhen for looking out for me. Thank you to Dmitry for teaching me about macrocycle libraries, and Alix for paving the way ahead of me. Thank you to Alex and Beverly, as we sailed this tiny DTS boat together. Thank you to Alvin for putting up with me during one of the weirdest years for everyone, I hope we continue to work together in the future in this niche area. Thank you also to the rest of the Liu group members, past and present, from whom I've learned so much. Thank you to Ariel, for your unending patience. I've learned a lot about how to think about science from you. Thank you for always looking out for me. Thank you to Kevin for endless and inspiring conversations about science, technology, and life. Thank you to Jaron, an honorary DTS member, for always teaching me new things about chemistry. I hope I can help you somehow in the future. Thank you to Ahmed, who has constantly taught me new things and given me someone to look up to. You always made time for me, and for that I'm thankful. Thank you to Tina, from whom I've learned how to think about science. Thank you to Max, for teaching me about machine learning, and bouncing ideas off of me so I can learn more about this area.

It's guaranteed that I've forgotten people to thank individually, but please know that every single Liu lab member has made a significant difference in my scientific career.

To Minwoo, thank you for your friendship. My graduate school experience wouldn't have been the same without you. I'm so glad that you only work two T stops away now, as we can hopefully continue to catch EPL games together in the future. Thank you for all the work you put into the catalyst project.

To Michael, what started off as a random thing we wanted to try one summer, has evolved into the culmination of my thesis project. Thank you for being such a great person to work with as we learned, implemented, and tested. I couldn't have done it without you.

To my undergraduate mentors, Josh, I wouldn't be here today if I hadn't joined your lab in summer of freshman year. Thank you for providing such a great learning environment. I learned so much from Yi-Fan, about how to think about biology from a chemical point of view. I learned so much about mass spec from Wenyun. I am also so grateful to have been friends with Jun and Chel, who prepared me for graduate school and made the lab a great place to work.

To my family, the bedrock of my existence. Thank you for always supporting me.

To my childhood friends, Jon and Indraneel, without whom I wouldn't be who I am today.

To Anya, thank you for making me a better person.

List of Figures

Figure 1.1 2013 HFNAP building block and genetic code. 12

Figure 2.1 Design and construction of the sequence-defined polymer library. 31
Figure 2.2 Selection of PCSK9-binding polymers from a random HFNAP library 35
Figure 2.3 Sequence homology and characterization of selection-enriched PCSK9-binding HFNAPs. 37
Figure 2.4 Structures of bases in side-chain mutants characterized during structure-activity relationship studies of PCSK9-A5 and PCSK9-Evo5. 38
Figure 2.5 Evolution of an improved PCSK9-binding polymer. 39
Figure 2.6 Effects of truncating HFNAPs on their binding affinity to PCSK9 protein. 41
Figure 2.7 Characterization of IL-6-binding HFNAPs selected from a random library. 44

Figure 3.1 HFNAP translation, selection, and library design. 64
Figure 3.2 PCSK9 selection conditions and progress. 67
Figure 3.3 Protein target binding activity, sequence convergence, and sequence distribution of HFNAPs selected for PCSK9 or IL-6 affinity. 69
Figure 3.4 Target-binding activity of individual HFNAPs by SPR. 72
Figure 3.5 Analysis of mixed samples in polymer pool competition. 73
Figure 3.6 Head-to-head competition of seven pools following iterated selections for binding to PCSK9 or IL-6. 75
Figure 3.7 Individual polymer enrichment in PCSK9 competition experiment. 76
Figure 3.8 Changes in codon distribution. 78
Figure 3.9 IL-6 selection conditions and progress. 80

Figure 4.1 HFNAP building blocks, translation, and selection. 103
Figure 4.2 Selections against daunomycin yield HFNAPs with binding affinity. 104
Figure 4.3 Quantifying selection progress by comparing non-binding flow-throughs to the eluent. 105
Figure 4.4 MST characterization of the binding affinity for the top 4 most enriched sequences from selection 1. 106
Figure 4.5 The evolutionary trajectory of the top ten most enriched sequences in selection 1. 107
Figure 4.6 Sequence convergence for selection 1 and selection 2. 108
Figure 4.7 High-stringency selection against daunomycin. 110
Figure 4.8 The evolutionary trajectory of the top ten most enriched sequences in selection 2. 110
Figure 4.9 Changes in side-chain distribution for selection 2. 111
Figure 4.10 Changes in side-chain distribution for selection 1. 112
Figure 4.11 MST characterization of the binding affinity for the top ten most enriched sequences from selection 2. 113
Figure 4.12 MST characterization of the daunomycin-binding affinity of Dm-HS-1 mutants. 114

Figure 5.1 Overview of VAE structure, training and sequence generation. 131
Figure 5.2 CVAE architecture and fitness values from the in vitro selection. 132

Figure 5.3 Distribution of fitness values for training and general test set sequences. 133
Figure 5.4 CVAE performance of an early model. 135
Figure 5.5 Training performance of final model architecture. 136
Figure 5.6 tSNE plot of general fitness test set sequences and high fitness test set sequences. 138
Figure 5.7 Percent of generated sequences with at least 3-6 alcohol or amine side-chains. 139
Figure 5.8 Pairwise Levenshtein distances for 10,000 CVAE-generated HFNAPs.. 141
Figure 5.9 MST characterization of the binding affinity of CVAE-generated daunomycin-binding HFNAPs. 142
Figure 5.10 High-stringency selection of mixture of 10,000 CVAE-generated HFNAPs and 2,000 most highly enriched sequences from round 4 to round 8b of selection 2 (competition set). 143
Figure 5.11 High-stringency selection of mixture of 10,000 reference-based random HFNAPs and 2,000 most highly enriched sequences from round 4 to round 8b of selection 2 (competition set). 145
Figure 5.12 UMAP projection (a non-linear method of dimensionality reduction) of reference set, selection-2, and CVAE-generated daunomycin-binding HFNAP sequences. 147
Figure 5.13 Comparison of the Levenshtein distances from the 3,072-member reference set to the top seven sequences from each of the four groups. 148
Figure 5.14 Conserved secondary structures for Dm-HS-1. 150
Figure 5.15 Base pair distance between Dm-HS-1 conserved secondary structures and CVAE-generated sequences or reference-based random sequences from respective competition selections. 152
Figure 5.16 Secondary structures with the two closest distances to conserved Dm-HS-1 structures.. 153

Chapter 1 Introduction to *in vitro* selections and functionalized nucleic acid polymers

Contributions:

I wrote this, with feedback from Professor David Liu.

1.1 *In vitro* selections and directed evolution

Natural selection is the competitive process through which Darwinian evolution shapes biological life on a macro and molecular scale. Humans have implemented artificial selection since the domestication of plants and animals, by using selective breeding practices to improve food production output. Thousands of years later, Darwin used artificial selection in selective breeding practices to help explain the concept of natural selection in 1859[1]. It would take another 100+ years before researchers began to further probe and harness the power of artificial selection.

One of the earliest examples of researchers exploring artificial selection can be found in Sol Spiegelman's series of experiments in 1967, where he subjected bacteriophage RNA-dependent RNA polymerases (RNA replicases) to increasingly short replication windows[2-4]. This *in vitro* experiment removed the biological constraints imposed on the RNA genome for essential coding regions such as the coat protein, and resulted in large truncations to the RNA genome. Regions not required for genome replication could be readily discarded and had a selective advantage in the competitive environment. After a number of passages, the remaining RNA sequence replicated ~15-fold more quickly, but retained only 550 of the 3,600 nucleotides from the original sequence. As a result, these sequences were no longer biologically competent, and had lost their infectivity as early as the fourth passage. This experiment highlights the importance of carefully considering all possible methods for passing a selection during selection development. By selecting only for replication speed, the RNA sequences in Spiegelman's experiments were rewarded for producing more copies, eventually at the cost of minimizing the genome.

As scientists developed artificial selection methods further, two main techniques emerged: directed evolution and *in vitro* selections. Importantly, both of these techniques harnessed the power of artificial selections to search increasingly large libraries of biopolymer

variants. Directed evolution typically starts from naturally occurring activity (active proteins) and uses artificial selections to identify evolving macromolecules with altered activity[5-9]. For example, researchers may look to improve the activity, change substrates, resist degradation, or work in unfavorable conditions. *In vitro* selections on the other hand, typically start from large, random combinatorial libraries, and seek to identify active sequences *de novo* from naïve libraries[10-12].

Directed evolution was enabled in part by Greg Smith's pioneering work in 1985. Smith published a technique called phage display[13] enabling expression of fusion proteins on the surface of bacteriophage. By using antibodies against a protein of interest, phage that displayed the protein could be selectively enriched over 1,000-fold with the use of an antibody. This strategy was later used in directed evolution experiments to improve protein function and stability. Early phage display libraries utilized NNK libraries and expressed fairly short peptides (6-mer to 15-mer peptides)[14-17], based on strategies developed by Geysen and coworkers for epitope analysis using ELISA screening[18,19]. Later work included the expression of protein fragments[20-22], enzymes[23-25], hormones[26,27], and inhibitors[28,29]. Together, these works set the stage for protein-based directed evolution, a method that has become the mainstay of protein engineering efforts today.

Early efforts by Wells, Knowles, Arnold, and coworkers in the directed evolution of proteins highlight some of the major differences between *in vitro* selections and directed evolution[5-7]. Starting from active enzymes naturally selected over millions of years of evolution, Frances Arnold's early work on para-nitrobenzyl esterase[30] and subtilisin E[8,9] demonstrated the use of random mutagenesis and screening to evolve proteins with increased activity in organic solvents. While screening limits the number of variants that can be reasonably tested, directed evolution experiments can also utilize

selections, where variants are competed simultaneously[24,27,29]. However, because directed evolution relies on a host organism to produce the evolved protein, directed evolution library sizes are typically capped by the transformation efficiencies of *E. coli*[31] or phage[32,33], limiting the library sizes to ~10^9 and ~10^{11}, respectively. To access larger libraries and search greater mutational space, continuous evolution methods were developed.

Continuous evolution systems operate by subjecting viral or bacterial systems to growth rate challenges in unfavorable continuous flow conditions. Using chemostats, which have constant inflow and outflow to maintain a constant growth rate, Molineux and coworkers subjected phage to heat stress in early experiments, demonstrating an 18,000-fold improvement in growth rate at 43 ºC[34,35]. Scientists later incorporated a chemical mutagen MNNG (*N*-methyl-*N*'-nitro-*N*'-nitrosoguanidine), further increasing the size of search spaces that could be explored using this type of method[36]. Continuous evolution systems have enabled the understanding of antibiotic resistance[37,38], improved biofuel production[39,40], altered protease specificity[41-43], and improved genome editing agent activity[44,45]. Nevertheless, directed evolution requires that the protein has some level of activity to begin with, setting a prerequisite that can make developing new activity difficult. If there are no natural enzymes that catalyze a specific chemical transformation, for example, then directed evolution is limited in its ability to address this given challenge.

In contrast, *in vitro* selections typically begin from random, combinatorial libraries and are therefore not subject to the same starting constraints as directed evolution. Developed simultaneously in 1990 by Szostak, Gold, and Joyce, the initial experiments enabled identification of RNA sequences capable of small molecule binding[10,46,47], protein binding[11], and catalysis[12]. RNA, with its 2' hydroxyl, can catalyze additional reactions including DNA cleavage[12], tRNA charging with activated amino acids[48], amide bond formation[49], Diels-Alder cycloaddition[50], Michael addition[51], and aldol condensation[52]. Despite the lack of chemical

functionality, compared to amino acids, these examples demonstrated that nucleotides and the 2' hydroxyl are often sufficient chemical functionalities for specific binding interactions and catalytic transformations. Indeed, even DNA-based selections have been successful against both protein and small molecule targets, yielding binders for streptavidin,[53] thrombin,[54,55] adenosine,[56] ATP,[56] Ochratoxin A,[57] and many more.[58,59] In turn, DNA-based catalysts have also demonstrated an ability to catalyze phosphorylation[60,61], DNA ligation and cleavage[62,63], N-glycosylase activity[64], amide bond cleavage[65], porphyrin metalation[66], and Diels-Alder cycloaddition[67]. While these examples demonstrate a wide variety of successful oligonucleotide-based selections for binding and catalytic activity, lack of chemical functionality continues to limit the ability of oligonucleotides to serve as potent binders and catalysts. To address the lack of chemical functionality, a number of groups have developed nucleic acid systems with modified oligonucleotides.

1.2 Modified nucleic acids

1.2.1 Backbone and sugar-modified nucleic acids

Modified nucleic acids can increase the likelihood of conducting a successful *in vitro* selection campaign by increasing the chemical functionality available or altering the types of possible secondary structures. The modifications can occur on either the backbone, the sugar, or the nucleobase, which are the three major components of oligonucleotides.

Backbone modifications have been typically explored because of their effects on oligonucleotide stability. Because the phosphodiester bond is subject to attack by nucleases, phosphorothioate backbones are amongst the most common backbone modification, and have been demonstrated to exhibit increased nuclease resistance[68-70].

Similarly, methylphosphate linkages have demonstrated protective effects when used as capping linkages at the 3' or 5' ends[71,72]. Although researchers have attempted to change the chemical functionality of the backbone using boranophosphate linkages, these were not shown to yield improved binding compared to standard phosphodiester-derived oligonucleotides[73].

Sugar modifications can take many forms, the most common of which are 2'-position modifications. These modifications result in altered secondary structures, which can contribute both to activity and stability. Incorporation of 2'-amino-2'-deoxypyrimidines into *in vitro* selections yielded more stable human neutrophil elastase (HNE)[74], vascular endothelial growth factor (VEGF)[75], and basic fibroblast growth factor (bFGF) binders[76], where stability was increased upwards of 1000-fold in the case of the bFGF binding sequence. Janjić and coworkers also demonstrated that the sub-nanomolar potency against bFGF observed was ~4-fold weaker when the unmodified sequence was tested for affinity, highlighting the potential contribution of the 2'-amino modification in overall binding affinity.

Dang and coworkers also described the role stability can play in binding affinity, as the increased stability of intramolecular helices from 2'-fluoro-2'-deoxypyrimidines in comparison to 2'-amino-2'-deoxypyrimidines has actual effects for binding affinity. Using two parallel selections with either 2'-amino or 2'-fluoro modified oligonucleotides, they demonstrated how the selection utilizing 2'-fluoro oligonucleotides resulted in inhibitors of keratinocyte growth factor with almost 100-fold increased affinity (0.3-3 pM vs. 400 pM K_d) and substantially improved bioactivity (34 pM vs. 10 nM K_i)[77]. It is important to note, however, that the stabilities of these modifications are in part because of their nuclease resistance, but also because of the altered chemistry, resulting in more stable secondary structures[78-80].

Additionally, the other common 2'-position modification is the 2'-*O*-methyl modification. 2'-*O*-methyl modifications have been demonstrated to increase nuclease resistance[81,82] and are

also naturally occurring[83]. Keefe and coworkers conducted *in vitro* selections using a 2'-*O*-methyl modified oligonucleotide library, resulting in a VEGF binder with K_d = 2 nM and 96-hour plasma stability at 37 ºC with no signs of degradation[84,85]. These high affinities and stabilities may be partially attributed to the increased hydrophobicity that results from this modification[86].

A number of groups have experimented with replacing the standard ribose or deoxyribose sugar altogether. Collectively called xeno nucleic acids (XNAs), these modified nucleic acids contain five or six-member sugar rings including 1,5-anhydrohexitol nucleic acids (HNAs)[87], cyclohexenyl nucleic acids (CeNAs)[88], 2'-*O*,4'-*C*-methylene-β-D-ribonucleic acids (locked nucleic acids (LNAs))[89], arabinonucleic acids (ANAs)[90], 2'-fluoro-arabino-nucleic acids (FANAs)[91], and threose nucleic acids (TNAs)[92]. Chaput (TNAs) and Holliger (XNAs) have made substantial progress in developing the enzymatic machinery required to conduct *in vitro* selections on these XNAs, resulting in the discovery of engineered polymerases and reverse transcriptases (RTs) capable of translating DNA sequences to TNAs and XNAs, and back[93-95]. These enzymes then enable *in vitro* selections with these XNAs, yielding protein binders[93,94], and catalysts for RNA cleavage and ligation[96].

A related concept to XNAs is new base pairs with potentially non-canonical base pairing modes. Benner and coworkers have developed a number of additional hydrogen-bonding base pairs, expanding the standard genetic alphabet to include 6 or even 8 total bases[97-99]. In contrast, Hirao, Romesberg, and coworkers have developed hydrophobic base pairs[100-103]. Both work by Benner and Hirao demonstrated an ability to conduct *in vitro* selections with these expanded genetic systems, yielding sequences where the additional base pair was necessary for activity. Similar to how unnatural sugars can change the secondary structure of resulting polymers, these expanded genetic

alphabets may also benefit from increased chemical diversity and access to alternate structures not available to standard oligonucleotide-based polymers.

The majority of the modifications discussed thus far have largely resulted in improved stability with minimal changes to chemical functionality other than the hydrophobic base pairs. The final type of modification available for nucleic acids is nucleobase modification, many of which result in large changes in chemical functionality.

1.2.2 Nucleobase functionalization expands chemical diversity

Nucleobase functionalization largely takes place at the C-5 position of pyrimidines, with early examples consisting mostly of UTP and dUTP analogs. In 1981, Ward and coworkers synthesized dUTP and UTP analogs containing biotinylation at the C-5 position, and tested whether standard molecular biology polymerase tools could incorporate this modified analog[104]. Biotin, separated from the nucleobase by an allylamine linker, was successfully incorporated by *E. coli* DNA polymerase I, *E. coli* RNA polymerase, T4 DNA polymerase, T7 RNA polymerase, mouse DNA polymerase α and β, and human DNA polymerase α and β. Of note are the changes in incorporation efficiency, with dUTP-biotin incorporation less efficient than dTTP or UTP incorporation in their native contexts, and UTP-biotin underperforming dUTP-biotin for RNA polymerases and DNA polymerases, respectively. Additionally, UTP-biotin incorporation by eukaryotic RNA polymerases was limited, and dUTP-biotin also failed to serve as a successful reverse transcriptase substrate. Despite these minor drawbacks, incorporation was validated by antibiotin antibody capture. While hybridization studies indicate that biotin-labeled oligonucleotides form duplexes with lower melting temperatures, the sequences still demonstrate fairly robust duplex formation, indicating that the biotin modification at the 5'-position does not fully disrupt oligonucleotide duplex formation. In fact, DNA sequences with a 5.5% replacement rate for TTP by dUTP-biotin demonstrate identical melt curves to

unsubstituted DNA[104]. These experiments were important in establishing the feasibility of C-5 modifications, with initial data points indicating the non-disruptive nature of modifications built off this position.

Indeed, the non-disruptive nature of C-5 modifications was later confirmed crystallographically by Marx and coworkers. In a series of crystal structures with various C-5 modified analogs, they identified that C-5 analogs are well tolerated by the polymerase complex because the modification points into the major groove of the DNA helix, although some stabilizing interactions particularly against Arg660 are disrupted. As a result, C-5 modifications and linkers capable of hydrogen bonding are likely to be incorporated at higher efficiencies because of their stabilizing interactions with the polymerase complex[105,106].

As a result, there have been a wide variety of modifications incorporated at the C-5 position, many of which were later used in *in vitro* selections to achieve high affinity against targets of interest. Some key modifications include fluorescent labels[107,108], solvatochromic and pH-dependent fluorescent labels[109,110], and reactive groups for bioconjugation[111,112] (Click[113,114], Diels-Alder[115,116], Staudinger ligation[117]). While these instances are useful chemical biology tools, the incorporation of C-5 modifications for the purposes of expanding the chemical functionality of oligonucleotide-based *in vitro* selections began in the mid 1990s with work from Toole[118] and coworkers. Researchers isolated sequences with affinity towards their protein target of interest, thrombin, that were dependent on the modified analog. While authors pointed out that the resulting candidate sequences were unrelated to sequences from their previous *in vitro* selections against thrombin using unmodified ssDNA[54], these sequence differences could result from changes to selection library composition. For example, the primer differences could

result in altered secondary structures, leading to different preferred motifs and sequences within the random region.

Eventually Perrin and coworkers successfully attempted a dual modified nucleobase *in vitro* selection, laying the groundwork for the field as a whole[119,120]. Using allylamine-modified dUTP and imidazole-modified dATP, researchers selected for cleavage of a single RNA base within a DNA context. Although the cleavage substrate is positioned over the randomized active site region by base pairing, researchers attempted the selection without divalent cations, demonstrating how increased chemical functionality in oligonucleotides by means of nucleobase modification can overcome significant chemical functionality deficiencies.

The types of modified analogs as well as their usage in *in vitro* selections was then advanced by a number of research groups. In particular, Williams and coworkers explored linker lengths and mechanistic studies on polymerase incorporation efficiencies[121,122], while Hollenstein, Perrin, Famulok, Sugimoto, and others expanded the types of modifications[123-129]. In their 2017 publication, Gold and coworkers expanded the number and types of C-5 analogs beyond what had been commonly used in the field at the time[130,131]. The resulting libraries were then subjected to multiple parallel selections with different pairs of C-5 modified NTPs to determine if certain pairs of modifications were more likely to yield more potent binders. This work, along with work from our group demonstrated the important role of specific hydrophobic side-chains in developing protein binders[131-133].

The number of C-5 modified oligonucleotides has rapidly expanded because the C-5 position minimizes disruption to secondary structure and enzymatic transcription. The C-5 position is also readily modified via synthetic chemistry, which likely also played an important role in making many different substitutions readily available. The two general schemes involve triphosphorylation of functionalized nucleosides, or Sonogashira coupling with Iodo-dUTP, which is a readily available, commercial analog[125,134,135].

1.2.3 DNA-templated synthesis of highly functionalized nucleic acid polymers

So far, all examples have relied on polymerase-based translation and transcription systems, where one to three NTPs are used, resulting in 25-75% functionalization of the resulting oligonucleotide biopolymer. The goal of increasing the chemical functionality of oligonucleotides is, in part, to mimic nature and proteins and to access the chemical functionality afforded by the 20 amino acids. However, it's clear that even with a 75% functionalization and three separate NTPs, the level of chemical functionality available in such a system is fairly limited compared with what proteins have access to with their building block system. As a result, one of the major advantages of the ligase-based translation system developed in our group is the ability to incorporate up to 32 different nucleobase modifications into a sequence defined biopolymer, resulting in chemical diversity that truly rivals what is available to proteins.

This ligase-based system was first published by our group in 2013 and had the ability to incorporate up to eight different modified oligonucleotides in a single sequence (**Figure 1.1**)[136]. This highly functionalized nucleic acid polymer (HFNAP) translation system utilizes a system of trinucleotide monomer building blocks. Each building block consists of a trinucleotide sequence where the 5' most oligonucleotide is modified at the C-5 position. The biotinylated ssDNA template contains eight anti-codons that are complementary to the building block codon system. In this way, synthesis of sequence-defined polymers is possible, as base pairing results in building block hybridization, followed T4 DNA ligase catalyzed ligation of the phosphodiester backbones. After ligation, biotin pulldown followed by alkaline denaturation selectively elutes the single stranded HFNAP. The resulting HFNAP sequence has some clear advantages over polymerase-based translation systems, but the ligase-based system has disadvantages as well.

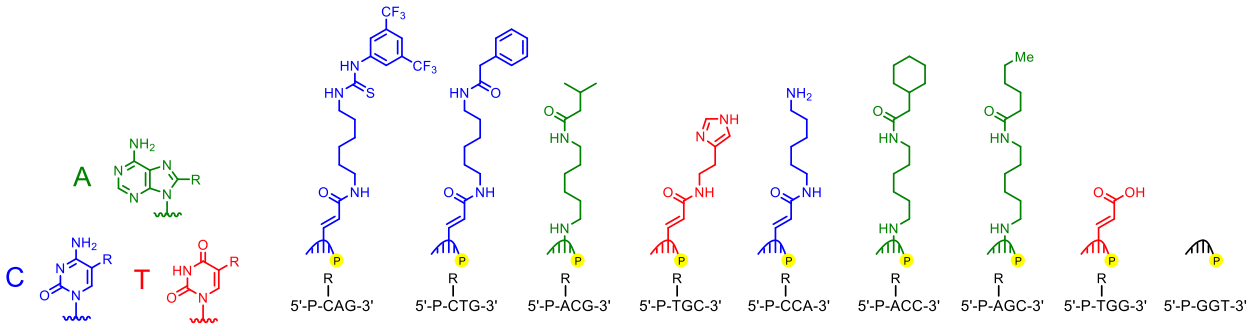


Figure 1.1 | 2013 HFNAP building block and genetic code. Building blocks have single codons and are built on either A, C or T modified nucleobases.

One of the major strengths of this translation system is the ability to incorporate up to 48 different nucleobase modifications. However, in its current iteration a number of issues are apparent. The length of the linkers could present a potential issue, as positioning of the side-chains by the HFNAP is less important if linkers are too long and flexible. Similarly, modified adenine nucleobases have long, non-hydrogen bonding linkers, which may lower translation efficiency. Furthermore, the current system constrains the modified nucleobases to a single trinucleotide codon, which may place secondary structure constraints on successful *in vitro* selection sequences. As a result, our group subsequently made dramatic changes to this system, enabling the discovery of sequence-defined polymers with high affinity for protein and small molecule targets of interest.

1.3 Selection stringency and the size of search spaces

One of the greatest advantages of *in vitro* selections is the lack of a host system, which enables synthesis of large starting libraries with $>10^{13}$ variants. By incorporating modified nucleobases into these libraries, researchers should gain access to chemically diverse libraries with high likelihood of discovering active variants. However, the size of these search spaces necessitates strong selection stringencies, in order to efficiently enrich those sequences. There have been three main strategies developed for increasing selection stringency.

The first strategy for increasing selection stringency is to decrease the target concentration, thereby limiting the number of variants that can continue to later rounds of the selection and increasing competition amongst remaining variants. However, decreasing the target concentration sufficiently compared with the selection variants can be difficult, as selections often occur on the order of picomoles. Therefore, the absolute amount of target available must be less than picomoles, which can be difficult to work with. Using microfluidics, Soh and coworkers have demonstrated how single bead manipulations can rapidly yield high affinity binders from large libraries[137-140].

The second strategy for increasing selection stringency is to vary the wash strength, which should remove weaker variants. Increasing the stringency of washes is a staple of modern *in vitro* selections, however the efficiency of this technique is reportedly low[97,141]. Soh and coworkers demonstrated how a microfluidic system enables washing using a dilution factor of 1,000, resulting in rapid enrichment of high affinity sequences[139]. However, they note that even specialized washing on its own still has limitations, as self-magnetization between beads can limit the amount of bead surface exposed to washing. As a result, their use of the volume dilution challenge, a technique borrowed from Gold and coworkers[130], is thought to further improve wash efficiency. Volume dilution entails conducting the target-binding portion of a selection in increasingly large volumes, which limits crowding effects and favors dissociation of weak binders.

The third strategy for increasing selection stringency is to include competitors[142]. Competitors can take the form of non-specific competitors like polyanionic compounds that compete against DNA's polyanionic backbone[130,131], or specific competitors or inhibitors that occupy binding sites[143]. Combined, these three strategies encompass the most common methods to modulate selection stringency.

Finally, the size of sequence search spaces remains an outstanding challenge for oligonucleotides and proteins. Even when starting *in vitro* selections with libraries of ~10^{13} molecules, the sequence space is often subsampled (as long as the random region is longer than 15 or 12 residues, for oligonucleotides and proteins, respectively). For typical *in vitro* selections that start with ~10^{13} molecules and a 40 N random region, it is unlikely that two selections against the same target would yield similar results. The initial starting libraries will have no overlap to begin with. Developing strategies to better analyze the ~10^{13} sequences within a single selection, will give insights as to how selections proceed. Yet current sequencing limitations preclude this possibility.

In summary, *in vitro* selection is a powerful technique that enables researchers to search large, combinatorial sequence spaces for activities of interest. These explorations have yielded a wide variety of active sequences from binding to catalysis—chemical interactions that would be surprising for a biomolecule originally considered just a store of information. It's clear that increased chemical functionality significantly augments the ability of oligonucleotides to accomplish new tasks. Further advances for *in vitro* selections may enable us to explore more of the vast sequence space that is available.

1.4 Conclusion and outlook

The strengths of *in vitro* selections, large, combinatorial libraries and efficient selection methods, are also sometimes outstanding challenges that prevent the technique from realizing its full potential. *In vitro* selections promise the ability to search large, combinatorial libraries of chemically diverse polymers for binding or catalytic activity. While a number of steps have been undertaken to increase the overall chemical diversity present within these libraries, researchers still do not have equivalent access to protein-like side-chains. Recent developments from our

group have demonstrated that broad chemical diversity is achievable in sequence-defined polymers, and the chemical diversity is integral to identifying polymers with high activity.

While *in vitro* selections are a largely established technique, selections remain labor intensive, limiting its widespread adoption. As such, it is critical to establish easily accessible high-stringency selection strategies. As outlined in the prior section, tuning selection stringency can be difficult without technical expertise in areas like microfluidics. A better understanding of *in vitro* selections is necessary in order to develop improved high-stringency selection strategies. The recent, rapid decrease in sequencing costs has made it possible to gather increasingly large amounts of data from *in vitro* selections. However, sequencing the entire *in vitro* selection remains impossible as most *in vitro* selections contain more unique molecules than the largest state of the art sequencing technologies can handle. Nevertheless, deep sequencing *in vitro* selections should yield a new understanding of evolutionary trajectories, helping us understand how *in vitro* selections proceed and how to improve the technique.

Recent observations about HFNAPs, and *in vitro* selections involving HFNAPs, may enable a deeper understanding of the selection process and the information contained within an individual *in vitro* selection. HFNAPs are uniquely poised to achieve protein-like activity because of their chemical diversity. This unique characteristic will likely increase the likelihood of identifying sequences with high activity. To date, the size of the sequence space and the limited space of possible solutions has made it challenging to directly generate unrelated, active variants. By producing more active sequences than is typical for standard selections, it may be possible to identify specific features that are contributing to desired activity. The ability to capture the specific features that are contributing to activity would enable direct generation of unrelated, active variants, a step forward in our ability to predict and generate active sequences computationally.

1.5 References

1. Darwin C. *On the Origin of Species by Means of Natural Selection, Or, The Preservation of Favoured Races in the Struggle for Life*. John Murray, 1859.

2. Mills D. R., Peterson R. L., Spiegelman S. An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. *Proceedings of the National Academy of Sciences of the United States of America* 1967, **58**(1)**:** 217--224.

3. Levisohn R., Spiegelman S. Further extracellular Darwinian experiments with replicating RNA molecules: diverse variants isolated under different selective conditions. *Proceedings of the National Academy of Sciences of the United States of America* 1969, **63**(3)**:** 805--811.

4. Kacian D. L., Mills D. R., Kramer F. R., Spiegelman S. A replicating RNA molecule suitable for a detailed analysis of extracellular evolution and replication. *Proceedings of the National Academy of Sciences of the United States of America* 1972, **69**(10)**:** 3038--3042.

5. Hermes J. D., Blacklow S. C., Knowles J. R. Searching sequence space by definably random mutagenesis: improving the catalytic potency of an enzyme. *Proceedings of the National Academy of Sciences* 1990, **87**(2)**:** 696.

6. Cunningham B. C., Wells J. A. Improvement in the alkaline stability of subtilisin using an efficient random mutagenesis and screening procedure. *Protein Engineering, Design and Selection* 1987, **1**(4)**:** 319-325.

7. Cunningham B. C., Henner D. J., Wells J. A. Engineering human prolactin to bind to the human growth hormone receptor. *Science* 1990, **247**(4949)**:** 1461.

8. Chen K., Arnold F. H. Enzyme Engineering for Nonaqueous Solvents: Random Mutagenesis to Enhance Activity of Subtilisin E in Polar Organic Media. *Nature biotechnology* 1991, **9**(11)**:** 1073--1077.

9. Chen K., Arnold F. H. Tuning the activity of an enzyme for unusual environments: sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide. *Proceedings of the National Academy of Sciences of the United States of America* 1993, **90**(12)**:** 5618--5622.

10. Ellington A. D., Szostak J. W. In vitro selection of RNA molecules that bind specific ligands. *Nature* 1990, **346**(6287)**:** 818--822.

11. Tuerk C., Gold L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science (New York, NY)* 1990, **249**(4968)**:** 505--510.

12. Robertson D. L., Joyce G. F. Selection in vitro of an RNA enzyme that specifically cleaves single-stranded DNA. *Nature* 1990, **344**(6265)**:** 467--468.

13. Smith G. P. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science (New York, NY)* 1985, **228**(4705)**:** 1315--1317.

14. Scott J. K., Smith G. P. Searching for peptide ligands with an epitope library. *Science (New York, NY)* 1990, **249**(4967)**:** 386--390.

15. Felici F., Castagnoli L., Musacchio A., Jappelli R., Cesareni G. Selection of antibody ligands from a large library of oligopeptides expressed on a multivalent exposition vector. *Journal of Molecular Biology* 1991, **222**(2)**:** 301--310.

16. Devlin J. J., Panganiban L. C., Devlin P. E. Random peptide libraries: a source of specific protein binding molecules. *Science (New York, NY)* 1990, **249**(4967)**:** 404--406.

17. Cwirla S. E., Peters E. A., Barrett R. W., Dower W. J. Peptides on phage: a vast library of peptides for identifying ligands. *Proceedings of the National Academy of Sciences of the United States of America* 1990, **87**(16)**:** 6378--6382.

18. Geysen H. M., Rodda S. J., Mason T. J. A priori delineation of a peptide which mimics a discontinuous antigenic determinant. *Molecular Immunology* 1986, **23**(7)**:** 709--715.

19. Geysen H. M., Rodda S. J., Mason T. J., Tribbick G., Schoofs P. G. Strategies for epitope analysis using peptide synthesis. *Journal of immunological methods* 1987, **102**(2)**:** 259--274.

20. Wang L. F., Du Plessis D. H., White J. R., Hyatt A. D., Eaton B. T. Use of a gene-targeted phage display random epitope library to map an antigenic determinant on the bluetongue virus outer capsid protein VP5. *Journal of immunological methods* 1995, **178**(1)**:** 1--12.

21. Petersen G., Song D., Hgle-Drr B., Oldenburg I., Bautz E. K. Mapping of linear epitopes recognized by monoclonal antibodies with gene-fragment phage display libraries. *Molecular \& General Genetics* 1995, **249**(4)**:** 425--431.

22. Jacobsson K., Frykberg L. Cloning of ligand-binding domains of bacterial receptors by phage display. *BioTechniques* 1995, **18**(5)**:** 878--885.

23. Corey D. R., Shiau A. K., Yang Q., Janowski B. A., Craik C. S. Trypsin display on the surface of bacteriophage. *Gene* 1993, **128**(1)**:** 129--134.

24. Marchand-Brynaert J., Bouchet M., Tetrahedron R. T. a. Design and synthesis of a bifunctional label for selection of β-lactamase displayed on filamentous bacteriophage by catalytic activity. *Tetrahedron* 1996, **52**(15)**:** 5591--5606.

25. Wang C. I., Yang Q., Craik C. S. Phage display of proteases and macromolecular inhibitors. *Methods in enzymology* 1996, **267:** 52--68.

26. McConnell S. J., Kendall M. L., Reilly T. M., Hoess R. H. Constrained peptide libraries as a tool for finding mimotopes. *Gene* 1994, **151**(1-2)**:** 115--118.

27. Lowman H. B., Bass S. H., Simpson N., Wells J. A. Selecting high-affinity binding proteins by monovalent phage display. 2002.

28. Markland W., Roberts B. L., Saxena M. J., Guterman S. K., Ladner R. C. Design, construction and function of a multicopy display vector using fusions to the major coat protein of bacteriophage M13. *Gene* 1991, **109**(1)**:** 13--19.

29. Roberts B. L.*, et al.* Protease inhibitor display M13 phage: selection of high-affinity neutrophil elastase inhibitors. *Gene* 1992, **121**(1)**:** 9--15.

30. Moore J. C., Arnold F. H. Directed evolution of a para-nitrobenzyl esterase for aqueous-organic solvents. *Nature biotechnology* 1996, **14**(4)**:** 458--467.

31. Liu J.*, et al.* An Improved Method of Preparing High Efficiency Transformation Escherichia coliwith Both Plasmids and Larger DNA Fragments. *Indian Journal of Microbiology* 2018, **58**(4)**:** 448--456.

32. Tiller T.*, et al.* A fully synthetic human Fab antibody library based on fixed VH/VL framework pairings with favorable biophysical properties. *mAbs* 2013, **5**(3)**:** 445--470.

33. Schwimmer L. J.*, et al.* Discovery of diverse and functional antibodies from large human repertoire antibody libraries. *Journal of immunological methods* 2013, **391**(1-2)**:** 60--71.

34. Bull J. J.*, et al.* Exceptional convergent evolution in a virus. *Genetics* 1997, **147**(4)**:** 1497--1507.

35. Wichman H. A., Badgett M. R., Scott L. A., Boulianne C. M., Bull J. J. Different trajectories of parallel evolution during viral adaptation. *Science (New York, NY)* 1999, **285**(5426)**:** 422--424.

36. Hillis D. M., Bull J. J., White M. E., Badgett M. R., Molineux I. J. Experimental phylogenetics: generation of a known phylogeny. *Science (New York, NY)* 1992, **255**(5044)**:** 589--592.

37. Toprak E.*, et al.* Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. *Nature Genetics* 2011, **44**(1)**:** 101--105.

38. Camps M., Naukkarinen J., Johnson B. P., Loeb L. A. Targeted gene evolution in Escherichia coli using a highly error-prone DNA polymerase I. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(17)**:** 9727--9732.

39. Luan G., Cai Z., Li Y., Ma Y. Genome replication engineering assisted continuous evolution (GREACE) to improve microbial tolerance for biofuels production. *Biotechnology for Biofuels* 2013, **6**(1)**:** 1--11.

40. Roth T. B., Woolston B. M., Stephanopoulos G., Liu D. R. Phage-Assisted Evolution of Bacillus methanolicus Methanol Dehydrogenase 2. *ACS Synthetic Biology* 2019, **8**(4)**:** 796--806.

41. 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* 2014, **5**(1)**:** 1--8.

42. Packer M. S., Rees H. A., Liu D. R. Phage-assisted continuous evolution of proteases with altered substrate specificity. *Nature communications* 2017, **8**(1)**:** 956.

43. Blum T. R.*, et al.* Phage-assisted evolution of botulinum neurotoxin proteases with reprogrammed specificity. *Science (New York, NY)* 2021, **371**(6531)**:** 803--810.

44. Richter M. F.*, et al.* Phage-assisted evolution of an adenine base editor with improved Cas domain compatibility and activity. *Nature biotechnology* 2020, **38**(7)**:** 883--891.

45. Thuronyi B. W.*, et al.* Continuous evolution of base editors with expanded target compatibility and improved activity. *Nature biotechnology* 2019, **37**(9)**:** 1070--1079.

46. Jiang F., Kumar R. A., Jones R. A., Patel D. J. Structural basis of RNA folding and recognition in an AMP-RNA aptamer complex. *Nature* 1996, **382**(6587)**:** 183--186.

47. Famulok M. Molecular recognition of amino-acids by RNA-aptamers-an L-Citrulline binding RNA motif and its evolution into an L-Arginine binder. *Journal of the American Chemical Society* 1994, **116**(5)**:** 1698--1706.

48. Saito H., Suga H. A Ribozyme Exclusively Aminoacylates the 3 \textquoteleft -Hydroxyl Group of the tRNA Terminal Adenosine. *Journal of the American Chemical Society* 2001, **123**(29)**:** 7178--7179.

49. Zhang B., Cech T. R. Peptide bond formation by in vitro selected ribozymes. *Nature* 1997.

50. Seelig B., Jschke A. A small catalytic RNA motif with Diels-Alderase activity. *Chemistry \& Biology* 1999, **6**(3)**:** 167--176.

51. Sengle G., Eisenfhr A., Arora P. S., Nowick J. S., Famulok M. Novel RNA catalysts for the Michael reaction. *Chemistry \& Biology* 2001, **8**(5)**:** 459--473.

52. Fusz S., Eisenfhr A., Srivatsan S. G., Heckel A., Famulok M. A Ribozyme for the Aldol Reaction. *Chemistry \& Biology* 2005, **12**(8)**:** 941--950.

53. Stoltenburg R., Reinemann C., Strehlitz B. FluMag-SELEX as an advantageous method for DNA aptamer selection. *Analytical and Bioanalytical Chemistry* 2005, **383**(1)**:** 83--91.

54. Bock L. C., Griffin L. C., Latham J. A., Vermaas E. H., Toole J. J. Selection of single-stranded DNA molecules that bind and inhibit human thrombin. *Nature* 1992, **355**(6360)**:** 564--566.

55. Hamaguchi N., Ellington A., Stanton M. Aptamer Beacons for the Direct Detection of Proteins. *Analytical Biochemistry* 2001, **294**(2)**:** 126--131.

56. Huizenga D. E., Szostak J. W. A DNA aptamer that binds adenosine and ATP. *Biochemistry* 1995.

57. Cruz-Aguado J. A., Penner G. Determination of Ochratoxin A with a DNA Aptamer. *Journal of Agricultural and Food Chemistry* 2008, **56**(22)**:** 10456--10461.

58. Breaker R. DNA aptamers and DNA enzymes. *Current Opinion in Chemical Biology* 1997, **1**(1)**:** 26--31.

59. McKeague M., DeRosa M. C. Challenges and Opportunities for Small Molecule Aptamer Development. *Journal of Nucleic Acids* 2012, **2012**.

60. Li Y., Breaker R. R. Phosphorylating DNA with DNA. *Proceedings of the National Academy of Sciences of the United States of America* 1999, **96**(6)**:** 2746--2751.

61. Walsh S. M., Sachdeva A., Silverman S. K. DNA Catalysts with Tyrosine Kinase Activity. *Journal of the American Chemical Society* 2013, **135**(40)**:** 14928--14931.

62. Sreedhara A., Li Y., Breaker R. R. Ligating DNA with DNA. *Journal of the American Chemical Society* 2004, **126**(11)**:** 3454--3460.

63. Chandra M., Sachdeva A., Silverman S. K. DNA-catalyzed sequence-specific hydrolysis of DNA. *Nature chemical biology* 2009, **5**(10)**:** 718--720.

64. Sheppard T. L., Ordoukhanian P., Joyce G. F. A DNA enzyme with N-glycosylase activity. *Proceedings of the National Academy of Sciences of the United States of America* 2000, **97**(14)**:** 7802--7807.

65. Zhou C.*, et al.* DNA-Catalyzed Amide Hydrolysis. *Journal of the American Chemical Society* 2016, **138**(7)**:** 2106--2109.

66. Li Y., Sen D. A catalytic DNA for porphyrin metallation. *Nature Structural \& Molecular Biology* 1996, **3**(9)**:** 743--747.

67. and M. C., Silverman S. K. *DNA and RNA Can Be Equally Efficient Catalysts for Carbon \textminus Carbon Bond Formation*. American Chemical Society, 2008.

68. Stein C. A., Subasinghe C., Shinozuka K., Cohen J. S. Physicochemical properties of phosphorothioate oligodeoxynucleotides. *Nucleic Acids Research* 1988, **16**(8)**:** 3209--3221.

69. Hoke G. D.*, et al.* Effects of phosphorothioate capping on antisense oligonucleotide stability, hybridization and antiviral efficacy versus herpes simplex virus infection. *Nucleic Acids Research* 1991, **19**(20)**:** 5743--5748.

70. Froehler B., Ng P., Matteucci M. Phosphoramidate analogues of DNA: synthesis and thermal stability of heteroduplexes. *Nucleic Acids Research* 1988, **16**(11)**:** 4831--4839.

71. Tidd D. M., Warenius H. M. Partial protection of oncogene, anti-sense oligodeoxynucleotides against serum nuclease degradation using terminal methylphosphonate groups. *British Journal of Cancer* 1989, **60**(3)**:** 343--350.

72. Miller P. S., Dreon N., Pulford S. M., McParland K. B. Oligothymidylate analogues having stereoregular, alternating methylphosphonate/phosphodiester backbones. Synthesis and physical studies. *The Journal of biological chemistry* 1980, **255**(20)**:** 9659--9665.

73. Lato S. M.*, et al.* Boron-containing aptamers to ATP. *Nucleic Acids Research* 2002, **30**(6)**:** 1401--1407.

74. Lin Y., Qiu Q., Gill S. C., Jayasena S. D. Modified RNA sequence pools for in vitro selection. *Nucleic Acids Research* 1994, **22**(24)**:** 5229--5234.

75. Green L. S.*, et al.* Nuclease-resistant nucleic acid ligands to vascular permeability factor/vascular endothelial growth factor. *Chemistry \& Biology* 1995, **2**(10)**:** 683--695.

76. Jellinek D.*, et al.* Potent 2'-amino-2'-deoxypyrimidine RNA inhibitors of basic fibroblast growth factor. *Biochemistry* 1995, **34**(36)**:** 11363--11372.

77. Pagratis N. C.*, et al.* Potent 2 \textasciiacute -amino-, and 2 \textasciiacute -fluoro-2 \textasciiacute - deoxyribonucleotide RNA inhibitors of keratinocyte growth factor. *Nature biotechnology* 1997, **15**(1)**:** 68--73.

78. Monia B. P.*, et al.* Evaluation of 2'-modified oligonucleotides containing 2'-deoxy gaps as antisense inhibitors of gene expression. *The Journal of biological chemistry* 1993, **268**(19)**:** 14514--14522.

79. Kawasaki A. M.*, et al.* Uniformly modified 2'-deoxy-2'-fluoro phosphorothioate oligonucleotides as nuclease-resistant antisense compounds with high affinity and specificity for RNA targets. *Journal of medicinal chemistry* 1993, **36**(7)**:** 831--841.

80. Aurup H., Tuschl T., Benseler F., Ludwig J., Eckstein F. Oligonucleotide duplexes containing 2'-amino-2'-deoxycytidines: thermal stability and chemical reactivity. *Nucleic Acids Research* 1994, **22**(1)**:** 20--24.

81. Geary R. S.*, et al.* Absolute bioavailability of 2'-O-(2-methoxyethyl)-modified antisense oligonucleotides following intraduodenal instillation in rats. *Journal of Pharmacology and Experimental Therapeutics* 2001, **296**(3)**:** 898--904.

82. Uhlmann E., Peyman A., Ryte A., Schmidt A., Buddecke E. Use of minimally modified antisense oligonucleotides for specific inhibition of gene expression. *Methods in enzymology* 2000, **313:** 268--284.

83. Smith C. M., Steitz J. A. Sno storm in the nucleolus: new roles for myriad small RNPs. *Cell* 1997, **89**(5)**:** 669--672.

84. Burmeister P. E.*, et al.* Direct in vitro selection of a 2'-O-methyl aptamer to VEGF. *Chemistry \& Biology* 2005, **12**(1)**:** 25--33.

85. Burmeister P. E.*, et al.* 2'-Deoxy purine, 2'-O-methyl pyrimidine (dRmY) aptamers as candidate therapeutics. *Oligonucleotides* 2006, **16**(4)**:** 337--351.

86. Friedman A. D., Kim D., Liu R. Highly stable aptamers selected from a 2'-fully modified fGmH RNA library for targeting biomaterials. *Biomaterials* 2015, **36:** 110--123.

87. Herdewijn P. Nucleic Acids with a Six-Membered \textquoteleft Carbohydrate \textquoteright Mimic in the Backbone. *Chemistry \& biodiversity* 2010, **7**(1)**:** 1--59.

88. Wang J.*, et al.* Cyclohexene nucleic acids (CeNA): Serum stable oligonucleotides that activate RNase H and increase duplex stability with complementary RNA. *Journal of the American Chemical Society* 2000, **122**(36)**:** 8595-8602.

89. Campbell M. A., Wengel J. Locked vs. unlocked nucleic acids (LNAvs.UNA): contrasting structures work towards common therapeutic goals. *Chemical Society Reviews* 2011, **40**(12)**:** 5680--5689.

90. Noronha A. M.*, et al.* Synthesis and Biophysical Properties of Arabinonucleic Acids (ANA):~ Circular Dichroic Spectra, Melting Temperatures, and Ribonuclease H Susceptibility of ANA \textperiodcentered RNA Hybrid Duplexes \dag. *Biochemistry* 2000, **39**(24)**:** 7050--7062.

91. Wilds C. J., Damha M. J. 2'-Deoxy-2'-fluoro-beta-D-arabinonucleosides and oligonucleotides (2'F-ANA): synthesis and physicochemical studies. *Nucleic Acids Research* 2000, **28**(18)**:** 3625--3635.

92. Schning K. U.*, et al.* Chemical Etiology of Nucleic Acid Structure: The α-Threofuranosyl-(3' \textrightarrow 2') Oligonucleotide System. *Science (New York, NY)* 2000, **290**(5495)**:** 1347--1351.

93. Pinheiro V. B.*, et al.* Synthetic Genetic Polymers Capable of Heredity and Evolution. *Science (New York, NY)* 2012, **336**(6079)**:** 341--344.

94. Yu H., Zhang S., Chaput J. C. Darwinian evolution of an alternative genetic system provides support for TNA as an RNA progenitor. *Nature chemistry* 2012, **4**(3)**:** 183--187.

95. Yu H., Zhang S., Dunn M. R., Chaput J. C. An Efficient and Faithful in Vitro Replication System for Threose Nucleic Acid. *Journal of the American Chemical Society* 2013, **135**(9)**:** 3583--3591.

96. Taylor A. I.*, et al.* Catalysts from synthetic genetic polymers. *Nature* 2015, **518**(7539)**:** 427--430.

97. Sefah K.*, et al.* In vitro selection with artificial expanded genetic information systems. *Proceedings of the National Academy of Sciences of the United States of America* 2014, **111**(4)**:** 1449--1454.

98. Zhang L.*, et al.* Evolution of Functional Six-Nucleotide DNA. *Journal of the American Chemical Society* 2015, **137**(21)**:** 6734--6737.

99. Hoshika S.*, et al.* Hachimoji DNA and RNA: A genetic system with eight building blocks. *Science (New York, NY)* 2019, **363**(6429)**:** 884--887.

100. Kimoto M., Yamashige R., Matsunaga K.-i., Yokoyama S., Hirao I. Generation of high-affinity DNA aptamers using an expanded genetic alphabet. *Nature biotechnology* 2013, **31**(5)**:** 453--457.

101. Matsunaga K.-i., Kimoto M., Hirao I. High-Affinity DNA Aptamer Generation Targeting von Willebrand Factor A1-Domain by Genetic Alphabet Expansion for Systematic Evolution of Ligands by Exponential Enrichment Using Two Types of Libraries Composed of Five Different Bases. 2016.

102. Zhang Y.*, et al.* A semi-synthetic organism that stores and retrieves increased genetic information. *Nature* 2017, **551**(7682)**:** 644--647.

103. Dhami K.*, et al.* Systematic exploration of a class of hydrophobic unnatural base pairs yields multiple new candidates for the expansion of the genetic alphabet. *Nucleic Acids Research* 2014, **42**(16)**:** 10235--10244.

104. Langer P. R., Waldrop A. A., Ward D. C. Enzymatic synthesis of biotin-labeled polynucleotides: novel nucleic acid affinity probes. *Proceedings of the National Academy of Sciences of the United States of America* 1981, **78**(11)**:** 6633--6637.

105. Obeid S., Baccaro A., Welte W., Diederichs K., Marx A. Structural basis for the synthesis of nucleobase modified DNA by Thermus aquaticus DNA polymerase. *Proceedings of the National Academy of Sciences* 2010, **107**(50)**:** 21327--21331.

106. Obeid S., Bukamp H., Welte W., Diederichs K., Marx A. Interactions of non-polar and \textquotedblleft Click-able \textquotedblright nucleotides in the confines of a DNA polymerase active site. *Chemical communications (Cambridge, England)* 2012, **48**(67)**:** 8320--8322.

107. Sinkeldam R. W., Greco N. J., Tor Y. Fluorescent Analogs of Biomolecular Building Blocks: Design, Properties, and Applications. *Chemical Reviews* 2010, **110**(5)**:** 2579-2619.

108. Tanpure A. A., Pawar M. G., Srivatsan S. G. Fluorescent Nucleoside Analogs: Probes for Investigating Nucleic Acid Structure and Function. *Israel Journal of Chemistry* 2013, **53**(6-7)**:** 366-378.

109. Riedl J., Pohl R., Rulíšek L., Hocek M. Synthesis and Photophysical Properties of Biaryl-Substituted Nucleos(t)ides. Polymerase Synthesis of DNA Probes Bearing Solvatochromic and pH-Sensitive Dual Fluorescent and 19F NMR Labels. *The Journal of Organic Chemistry* 2012, **77**(2)**:** 1026-1044.

110. Riedl J.*, et al.* Labelling of nucleosides and oligonucleotides by solvatochromic 4-aminophthalimide fluorophore for studying DNA–protein interactions. *Chemical Science* 2012, **3**(9)**:** 2797-2806.

111. Lnnberg H. Solid-Phase Synthesis of Oligonucleotide Conjugates Useful for Delivery and Targeting of Potential Nucleic Acid Therapeutics. *Bioconjugate Chemistry* 2009, **20**(6)**:** 1065--1094.

112. Singh Y., Murat P., Defrancq E. Recent developments in oligonucleotide conjugation. *Chemical Society Reviews* 2010, **39**(6)**:** 2054--2070.

113. Seela F., Sirivolu V. R. Nucleosides and Oligonucleotides with Diynyl Side Chains: Base Pairing and Functionalization of 2 \textasciiacute -Deoxyuridine Derivatives by the Copper(I)-Catalyzed Alkyne□Azide \textquoteleft Click \textquoteright Cycloaddition. *Helvetica Chimica Acta* 2007, **90**(3)**:** 535--552.

114. Sawant A. A., Mukherjee P. P., Jangid R. K., Galande S., Srivatsan S. G. A clickable UTP analog for the posttranscriptional chemical labeling and imaging of RNA. *Organic \& Biomolecular Chemistry* 2016, **14**(24)**:** 5832--5842.

115. Borsenberger V., Howorka S. Diene-modified nucleotides for the Diels-Alder-mediated functional tagging of DNA. *Nucleic Acids Research* 2009, **37**(5)**:** 1477--1485.

116. Schoch J., Wiessler M., Jschke A. Post-Synthetic Modification of DNA by Inverse-Electron-Demand Diels \textminus Alder Reaction. 2010.

117. Weisbrod S. H., Marx A. Novel strategies for the site-specific covalent labelling of nucleic acids. *Chemical communications (Cambridge, England)* 2008(44)**:** 5675--5685.

118. Latham J. A., Johnson R., Toole J. J. The application of a modified nucleotide in aptamer selection: novel thrombin aptamers containing 5-(1-pentynyl)-2'-deoxyuridine. *Nucleic Acids Research* 1994, **22**(14)**:** 2817--2822.

119. Perrin D. M., Garestier T., Hélène C. Bridging the Gap between Proteins and Nucleic Acids: A Metal-Independent RNAseA Mimic with Two Protein-Like Functionalities. 2001.

120. Lermer L., Roupioz Y., Ting R., Perrin D. M. *Toward an RNaseA Mimic: A DNAzyme with Imidazoles and Cationic Amines*. American Chemical Society, 2002.

121. Lee S. E.*, et al.* Enhancing the catalytic repertoire of nucleic acids: a systematic study of linker length and rigidity. *Nucleic Acids Research* 2001, **29**(7)**:** 1565-1573.

122. Gourlain T.*, et al.* Enhancing the catalytic repertoire of nucleic acids. II. Simultaneous incorporation of amino and imidazolyl functionalities by two modified triphosphates during PCR. *Nucleic Acids Research* 2001, **29**(9)**:** 1898--1905.

123. Hollenstein M., Hipolito C. J., Lam C. H., Perrin D. M. A DNAzyme with Three Protein-Like Functional Groups: Enhancing Catalytic Efficiency of M2+-Independent RNA

Cleavage. *Chembiochem : a European journal of chemical biology* 2009, **10**(12)**:** 1988--1992.

124. Hollenstein M., Hipolito C. J., Lam C. H., Perrin D. M. A self-cleaving DNA enzyme modified with amines, guanidines and imidazoles operates independently of divalent metal cations (M 2+). *Nucleic Acids Research* 2009, **37**(5)**:** 1638--1649.

125. Hocek M. Synthesis of Base-Modified 2 \textasciiacute -Deoxyribonucleoside Triphosphates and Their Use in Enzymatic Synthesis of Modified DNA for Applications in Bioanalysis and Chemical Biology. 2014.

126. Jger S.*, et al.* A versatile toolbox for variable DNA functionalization at high density. *Journal of the American Chemical Society* 2005, **127**(43)**:** 15071--15082.

127. Kuwahara M., Sugimoto N. Molecular Evolution of Functional Nucleic Acids with Chemical Modifications. *Molecules (Basel, Switzerland)* 2010, **15**(8)**:** 5423--5444.

128. Imaizumi Y.*, et al.* Efficacy of Base-Modification on Target Binding of Small Molecule DNA Aptamers. *Journal of the American Chemical Society* 2013, **135**(25)**:** 9412--9419.

129. Li M.*, et al.* Selecting Aptamers for a Glycoprotein through the Incorporation of the Boronic Acid Moiety. *Journal of the American Chemical Society* 2008, **130**(38)**:** 12636--12638.

130. Gold L.*, et al.* Aptamer-Based Multiplexed Proteomic Technology for Biomarker Discovery. *Plos One* 2010, **5**(12)**:** e15004.

131. Gawande B. N.*, et al.* Selection of DNA aptamers with two modified bases. *Proceedings of the National Academy of Sciences* 2017, **114**(11)**:** 2898--2903.

132. Chen Z., Lichtor P. A., Berliner A. P., Chen J. C., Liu D. R. Evolution of sequence-defined highly functionalized nucleic acid polymers. *Nature chemistry* 2018, **346:** 1.

133. Lichtor P. A., Chen Z., Elowe N. H., Chen J. C., Liu D. R. Side chain determinants of biopolymer function during selection and replication. *Nature chemical biology* 2019, **262**(4)**:** 1--426.

134. Hollenstein M. Nucleoside triphosphates--building blocks for the modification of nucleic acids. *Molecules (Basel, Switzerland)* 2012, **17**(11)**:** 13569--13591.

135. Hollenstein M. DNA Catalysis: The Chemical Repertoire of DNAzymes. *Molecules (Basel, Switzerland)* 2015, **20**(11)**:** 20777--20804.

136. Hili R., Niu J., Liu D. R. DNA ligase-mediated translation of DNA into densely functionalized nucleic acid polymers. *Journal of the American Chemical Society* 2013, **135**(1)**:** 98--101.

137. Lou X.*, et al.* Micromagnetic selection of aptamers in microfluidic channels. *Proceedings of the National Academy of Sciences* 2009, **106**(9)**:** 2989--2994.

138. Cho M.*, et al.* Quantitative selection of DNA aptamers through microfluidic selection and high-throughput sequencing. *Proceedings of the National Academy of Sciences of the United States of America* 2010, **107**(35)**:** 15373--15378.

139. Oh S. S.*, et al.* Improving Aptamer Selection Efficiency through Volume Dilution, Magnetic Concentration, and Continuous Washing in Microfluidic Channels. *Analytical chemistry* 2011, **83**(17)**:** 6883--6889.

140. Cho M.*, et al.* Quantitative selection and parallel characterization of aptamers. *Proceedings of the National Academy of Sciences* 2013, **110**(46)**:** 18460--18465.

141. Hanes J., Plückthun A. *In vitro* selection and evolution of functional proteins by using ribosome display. *Proceedings of the National Academy of Sciences* 1997, **94**(10)**:** 4937.

142. Marshall K. A., Ellington A. D. In vitro selection of RNA aptamers. *Methods in Enzymology*, vol. 318. Academic Press, 2000, pp 193-214.

143. Giver L., Bartel D. P., Zapp M. L., Green M. R., Ellington A. D. Selection and design of high-affinity rna ligands for hiv-1 rev. *Gene* 1993, **137**(1)**:** 19-24.

Chapter 2 Evolution of Sequence-Defined Highly Functionalized Nucleic Acid Polymers

This chapter has been adapted from:

Chen, Z., Lichtor, P. A., Berliner, A. P., Chen, J. C. & Liu, D. R. Evolution of sequence-defined highly functionalized nucleic acid polymers. *Nat Chem* **346,** 1 (2018).

Contributions:

Dr. Zhen Chen and Professor David Liu led the project. I contributed to characterization experiments. I thank Zhen and Phillip for teaching me about the world of DNA-templated synthesis and HFNAPs.

2.1 Introduction

The gene-encoded synthesis and Darwinian selection of sequence-defined biopolymers are fundamental features of all known forms of life. These processes have been harnessed in the laboratory to evolve RNA[1–3], DNA[4–6], and polypeptides[7–10] with a variety of binding and catalytic properties through iterated cycles of biopolymer translation, selection, replication, and mutation. The speed and effectiveness of the evolutionary process has inspired efforts to apply these principles to the much larger chemical space of synthetic polymers[11]. To date, however, the evolution of sequence-defined non-natural polymers in the laboratory has been limited to analogs of nucleic acids[12–15] and polypeptides[16] that can be synthesized by polymerases and ribosomes. Polymerases and ribosomes impose structural requirements on the building blocks that can be polymerized and thereby limit the diversity of synthetic polymers that are accessible to directed evolution. For example, polymerases use only mononucleotides as substrates, precluding the ability to encode a diverse set of codons and side chains. Known classes of polymerase-synthesized functional non-natural nucleic acid polymers, including those derived from non-natural sugar backbones[17–20], uniform installation of hydrophobic[21–28] or positively charged[29–35] side-chains on nucleobases, or introduction of novel nucleobases among the four possibilities[36–38], therefore have chemical diversities that are only modestly expanded beyond those of natural DNA and RNA, and fall short of the much more diverse chemical functionality present in proteins.

Previously we developed an *in vitro* system that uses DNA ligase to translate DNA sequences into sequence-defined highly functionalized nucleic acid polymers (HFNAPs) containing a wide range of side-chains chosen by the researcher[39]. In our original report, we showed that DNA sequences can be translated into HFNAPs using DNA ligase to catalyze the polymerization of up to 50 consecutive short, chemically functionalized oligonucleotide building blocks along a DNA template[39] (**Figure 2.1a**). We discovered that T4 DNA ligase accepts trinucleotide building blocks with a wide range of side-chains on the 5’ nucleobase, including

side-chains at the C5 position of pyrimidine nucleobases that are both synthetically accessible and unlikely to disrupt Watson-Crick base pairing[39].

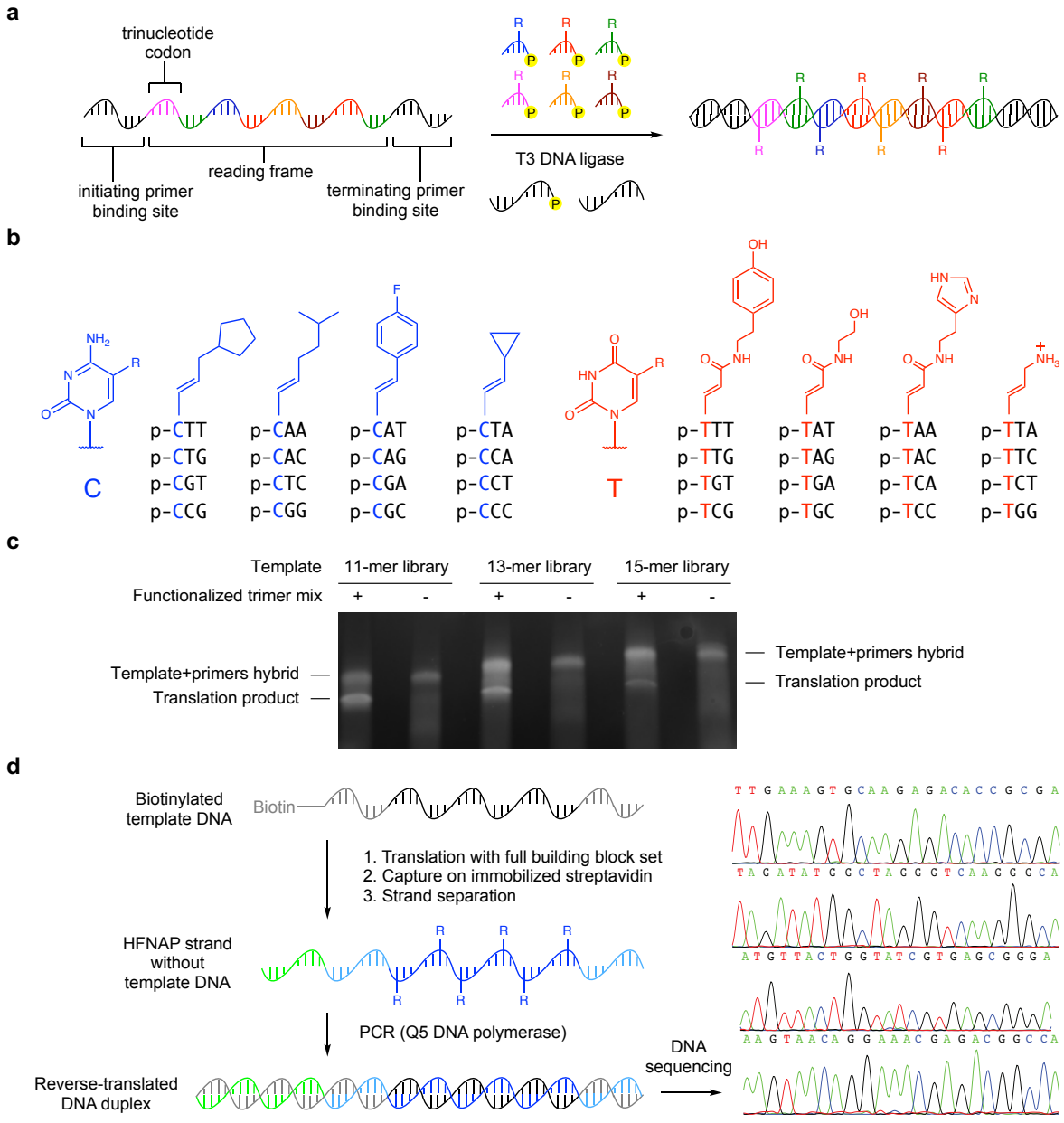


Figure 2.1 | Design and construction of the sequence-defined polymer library. a, Reaction scheme for DNA ligase-mediated translation of DNA templates into sequence-defined highly functionalized nucleic acid polymers (HFNAPs). **b**, Structures of 5'-phosphorylated trinucleotide building blocks for HFNAP library synthesis. **c**, Translation of libraries of randomized DNA templates into HFNAPs that incorporate up to 15 consecutive functionalized trinucleotide building blocks. The translation reactions, as well as control reactions from which the trinucleotide building blocks were omitted, were analyzed by polyacrylamide gel electrophoresis on a non-denaturing 10% TBE gel and imaged by SYBR gold staining. **d**, A complete cycle of

Figure 2.1 (Continued)
HFNAP translation, HFNAP strand isolation, and reverse translation back into DNA faithfully recovered sequence information from the original DNA templates. Left: experimental scheme; right: electrophoretic traces from Sanger sequencing. In control experiments in which the trinucleotide building blocks were omitted from the polymerization reactions, the PCR step did not generate any amplicons of the correct size.

This artificial translation system allows researchers, in principle, to mimic and even expand the chemical repertoire of protein building blocks in an evolvable synthetic polymer system. The broad chemical scope of HFNAPs gives them the potential to adopt unique folding and functional properties distinct from those of known natural or non-natural nucleic acid polymers. The original HFNAP system proved unable to support the evolution of functional polymers, however, likely because of the limited diversity provided by its eight-codon genetic code and the long, flexible linkers present in the building blocks. In this study, we designed a new HFNAP "genetic code", translation system, and *in vitro* selection system that overcomes these challenges, then applied the resulting HFNAP evolution system to generate sequence-defined synthetic polymers that binds two protein targets of biomedical interest.

2.2 Improved HFNAP building blocks and translation system

Our new genetic code was designed to offer a high degree of both codon and side-chain diversity to evolving polymers (**Figure 2.1b**). We included the maximum number of different trinucleotides containing a 5' pyrimidine (all 32 possible YNN combinations, where Y = C or T and N = A, C, G, or T) as codons. The corresponding 32 building blocks were each linked to one of eight side-chains (four codons per side-chain) that include hydrophobic, aliphatic, aromatic, halogenated, polar, and charged groups, several of which are not found among proteinogenic amino acid side-chains. The linkers between side-chains and nucleobases were redesigned compared with our original genetic code with limited conformational flexibility in order to increase the likelihood that the polymer backbone, nucleobases, and side-chains would cooperatively adopt defined folded structures. We assigned each side-chain to a set of four

codons that collectively contained the same balance of A/T versus C/G bases following the side-chain-functionalized 5' pyrimidine base.

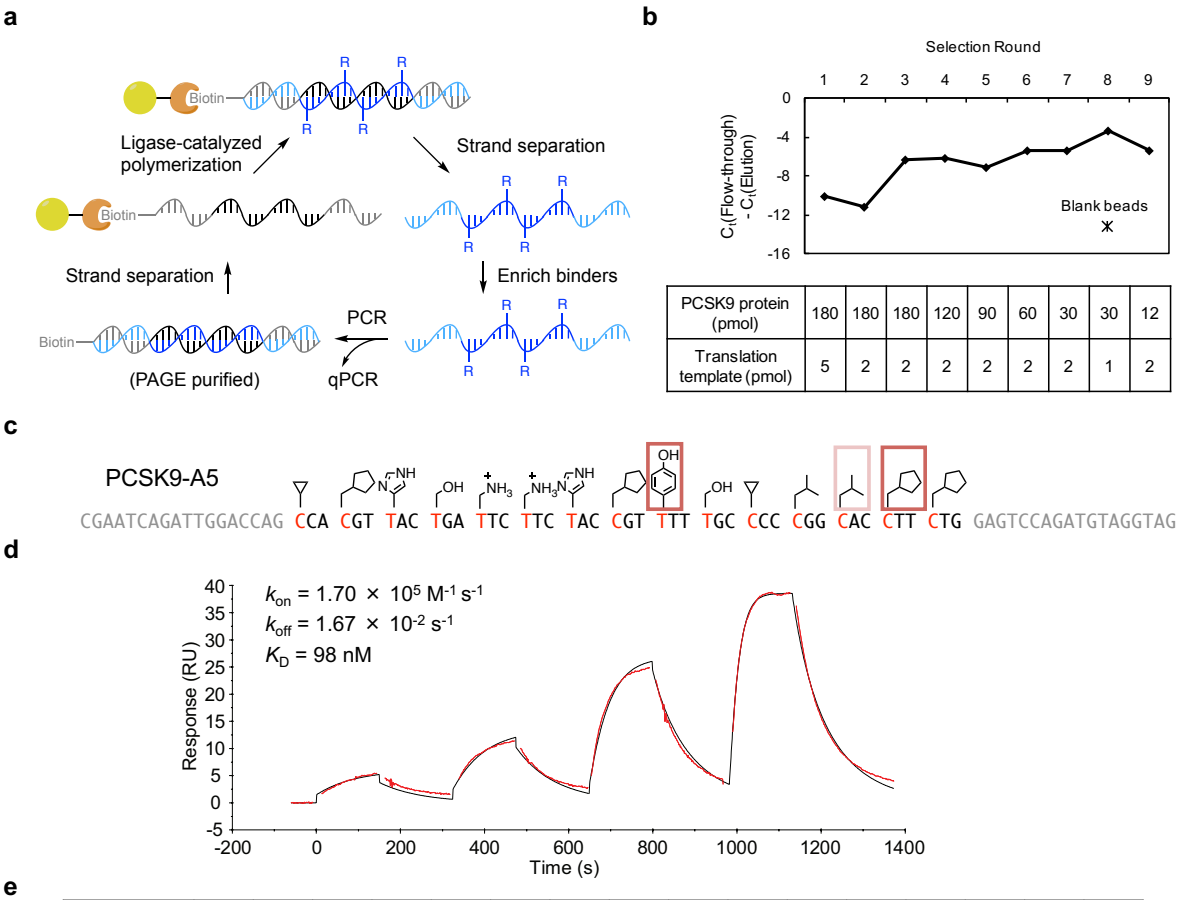
We improved DNA-templated polymerization (artificial "translation") reactions by screening ligase enzymes and adjusting polymerization conditions. We found that subjecting translation reactions to a slow (0.01 °C/s) temperature ramp to 4 °C before initiating ligation with T3 DNA ligase substantially improved yields of full-length HFNAP from libraries of DNA templates containing random coding regions of 45 nt, which encoded the incorporation of 15 consecutive side-chain-functionalized trinucleotide building blocks of mixed sequence (**Figure 2.1c**). To test the ability of translated polymers to be "reverse-translated" back into DNA, thereby enabling iterated cycles of translation, selection, reverse translation, and PCR amplification, we performed polymerization on four templates that each encoded the incorporation of eight different building blocks and collectively covered all 32 building blocks, and then subjected the resulting HFNAP products, separated from template DNA, to reverse translation in a PCR reaction using Q5 DNA polymerase[39] (**Figure 2.1d**). One of the PCR primers binds a 3'-overhang (**Figure 2.1d**, green) present in the HFNAP products but absent in template DNA, precluding the amplification of any contaminating template DNA. DNA sequencing of the resulting PCR products showed that the original sequence information in the templates was faithfully recovered (**Figure 2.1d**), indicating that both translation from DNA to HFNAPs and reverse translation from HFNAPs to their encoding DNA occur with high sequence fidelity using this set of building blocks.

These observations are qualitatively consistent with results from Hili and coworkers, who reported fidelities ranging from 95.1% to 98.4% per codon for ligase-mediated DNA-templated polymerization of functionalized pentanucleotides[40]. Perfect fidelity is not expected for a ligase-mediated polymerization, which lacks proofreading mechanisms, but we reasoned that the level of fidelity in our system may be sufficient to support iterated selection for functional polymers, consistent with our previous mock selection results[39]. Modest levels of mutations may also

confer a benefit to the selection, as reported by Benner and coworkers for selections of aptamers containing novel nucleobases[37].

2.3 *In vitro* selections against PCSK9

Encouraged by these developments, we generated a library of HFNAPs containing 15 consecutive building blocks drawn from the set of 32 (theoretical polymer library space = 3×10^{22}; average HFNAP molecular weight = 28 kDa) and subjected the resulting library (starting quantity = 3×10^{12} molecules) to iterated rounds of *in vitro* selection for binding to PCSK9 protein, a target implicated in low-density lipoprotein (LDL) metabolism and cardiovascular disease[41–43] (**Figure 2.2a**). The HFNAP library was incubated with PCSK9 protein immobilized on agarose beads. After washing the beads with buffer, the bound HFNAPs were eluted by boiling the beads in buffer containing detergent. The surviving HFNAPs were reverse translated in a PCR reaction using Q5 DNA polymerase, and the resulting double-stranded DNAs were purified by PAGE. The non-template strands were removed by alkaline denaturation, and the streptavidin-bound template strands were translated back into HFNAPs by ligase-catalyzed DNA-templated polymerization for the next round of PCSK9-binding selection. We note that the ability of HFNAPs to be directly reverse-translated by a DNA polymerase, though not an absolute requirement for iterated selection as demonstrated by the use of display methods to evolve nucleic acids[17,44], provides a practical and high-fidelity way to complete a selection cycle.



PCSK9 protein (pmol)	180	180	180	120	90	60	30	30	12
Translation template (pmol)	5	2	2	2	2	2	2	1	2

e

	A5	CCA ΔSide chain	CGT ΔSide chain	TAC ΔSide chain	TAC Linker only	TGA Linker only	TTC ΔSide chain	TTT ΔSide chain	TTT Linker only	TGC ΔSide chain	TGC Linker only	CCC ΔSide chain	CGG ΔSide chain	CAC ΔSide chain	CTT ΔSide chain	CTG ΔSide chain
k_{on} (10^5 M^{-1} s^{-1})	1.70	2.08	1.30	2.16	1.38	1.56	2.04	-	-	1.28	1.98	2.88	1.79	1.12	-	1.85
k_{off} (10^{-2} s^{-1})	1.67	1.76	1.08	1.79	1.10	1.25	1.77	-	-	1.06	1.76	2.13	1.50	2.29	-	2.18
K_D (nM)	98	85	83	83	80	80	87	>300	>300	83	89	74	84	204	>300	118

Figure 2.2 | Selection of PCSK9-binding polymers from a random HFNAP library. **a**, Overview of translation, selection, and reverse translation scheme. **b**, PCSK9 binding selection progress. The HFNAP pool's bulk affinity to PCSK9-coated beads was assessed by quantifying the amount of HFNAP in the flow-through versus the elution at each round of selection by quantitative PCR. Higher values in the graph indicate higher ratios of polymers that bound to immobilized PCSK9 and were eluted relative to polymers that flowed through the immobilized PCSK9. **c**, Sequence and side-chain structure of selected polymer PCSK9-A5. Side-chains essential for binding activity are boxed. **d**, SPR sensogram characterizing binding kinetics between surface-immobilized PCSK9-A5 polymer and the target PCSK9 protein. The concentrations of injected PCSK9 were 10, 30, 100, and 300 nM. The observed sensogram is shown in red and the fitted curve with the kinetic parameters listed is shown in black. **e**, Kinetic parameters for binding of PCSK9-A5 or its side-chain-deficient variants to PCSK9 protein, as measured by SPR. For the variants "TTT ΔSide chain", "TTT Linker only", and "CTT ΔSide chain", no SPR signal was observed at highest analyte concentration tested (300 nM PCSK9).

As the iterated rounds of selection progressed, the fraction of HFNAP that was retained on PCSK9-linked beads generally increased, consistent with enrichment of PCSK9-binding polymers, even though we steadily elevated selection stringency by decreasing the amount of PCSK9 protein (**Figure 2.2b**). At the eighth round of selection, the polymer population was retained by PCSK9 protein-conjugated agarose beads with an efficiency of approximately 10%. In contrast, less than 0.1% of the same polymer population was retained on agarose beads not conjugated to any protein (**Figure 2.2b**), suggesting that the ability of the selected polymers to bind PCSK9-linked beads arose from their ability to bind PCSK9, rather than agarose beads.

Results from high-throughput DNA sequencing after nine rounds of selection indicated that the HFNAP pool had strongly converged to just seven sequence families containing conserved sub-sequences suggestive of common binding motifs (**Figure 2.3a**). Sequences within the same family were likely descendants of a single parental polymer derived through mutations that accumulated through the selection process, as evidenced by high levels of homology. We individually synthesized the seven most highly enriched HFNAPs by ligase-mediated DNA-templated polymerization and tested their retention on immobilized PCSK9 or immobilized thrombin, an unrelated protein. Five of the seven tested polymers exhibited substantial apparent binding activity to immobilized PCSK9 beyond any apparent binding to immobilized thrombin. We also assayed unfunctionalized DNA sequences (lacking any side-chains) corresponding to these seven HFNAPs and observed no evident PCSK9-binding activity (**Figure 2.3b**). Together, these results suggest that the HFNAP populations emerging from nine rounds of *in vitro* selection converged on a small number of polymer families that bind immobilized PCSK9 in a manner dependent on the polymers' side-chains.

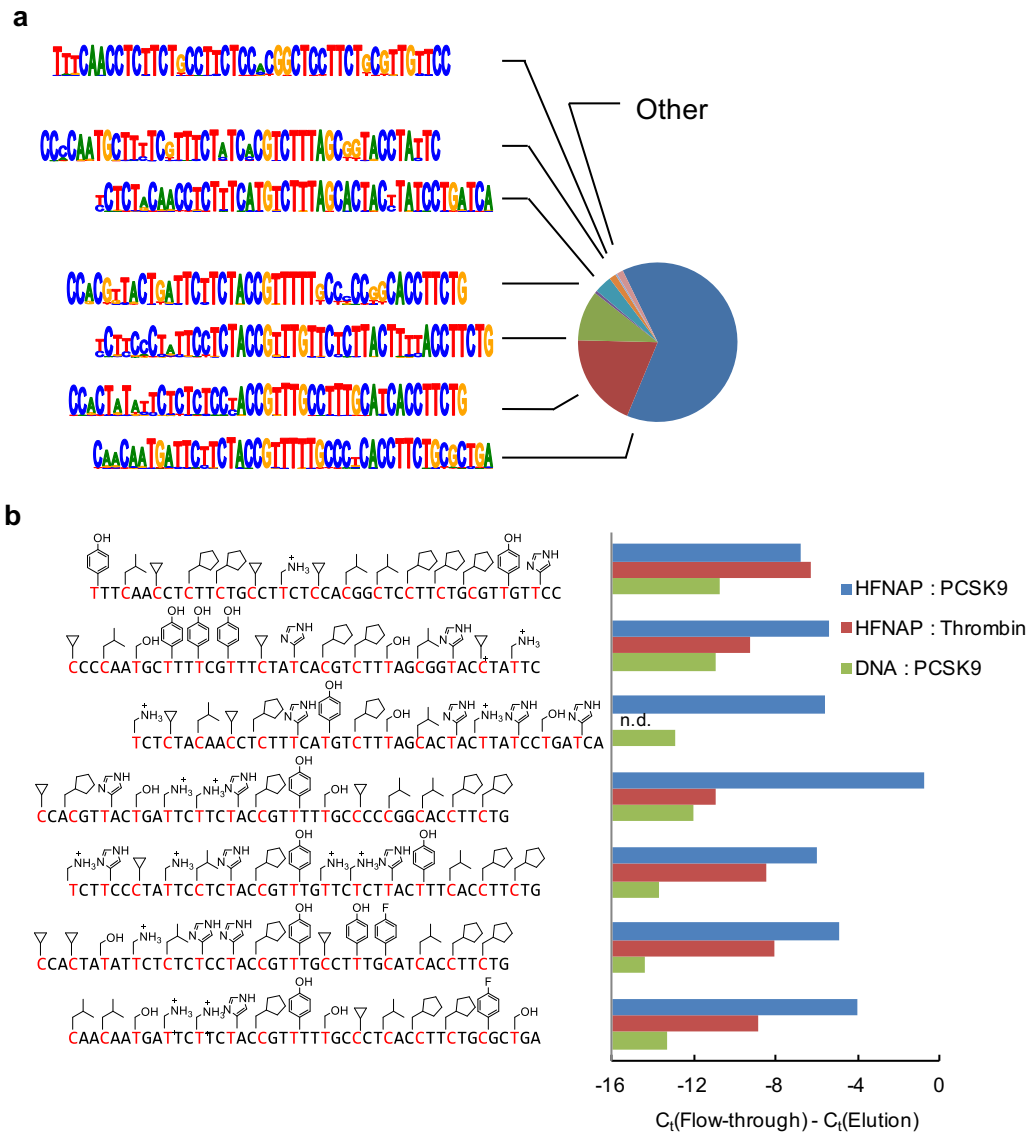


Figure 2.3 | Sequence homology and characterization of selection-enriched PCSK9-binding HFNAPs. a, Relative abundances and sequence logos of the dominant seven sequence clusters found in 3.1 million sequences from high-throughput DNA sequencing after nine rounds of PCSK9-binding selection from a naïve library. **b**, Retention of selection-enriched HFNAPs (only the variable, functionalized parts of the sequences are shown) on immobilized PCSK9 (target; blue bars) and immobilized thrombin (non-target; red bars) and of sequence-matched unfunctionalized DNA on immobilized PCSK9 (green bars), as quantified by qPCR. n.d. = not determined.

We characterized in depth PCSK9-A5, the polymer with the highest apparent PCSK9 binding activity (**Figure 2.2c**). We synthesized biotinylated PCSK9-A5 by templated translation and confirmed its binding affinity for PCSK9 (dissociation constant K_D = 98 nM) by surface plasmon resonance (SPR) (**Figure 2.2d**). To probe the role of the side-chain functional groups,

we synthesized side-chain mutants of PCSK9-A5 in which all instances of each building block were replaced by the corresponding trinucleotide either lacking any side-chain or containing a linker but missing the side-chain's functional group (**Figure 2.4**). The removal of a single phenol side-chain in codon 9 or a single cyclopentyl side-chain in codon 14 completely abolished the binding between PCSK9-A5 and its target (**Figure 2.2c** and **Figure 2.2e**, dark red). Furthermore, the removal of the isopentyl side-chain at codon 13 resulted in an approximately two-fold reduction in affinity (**Figure 2.2c** and **Figure 2.2e**, light red). The individual removal of other side-chains had less significant effects on binding affinity.

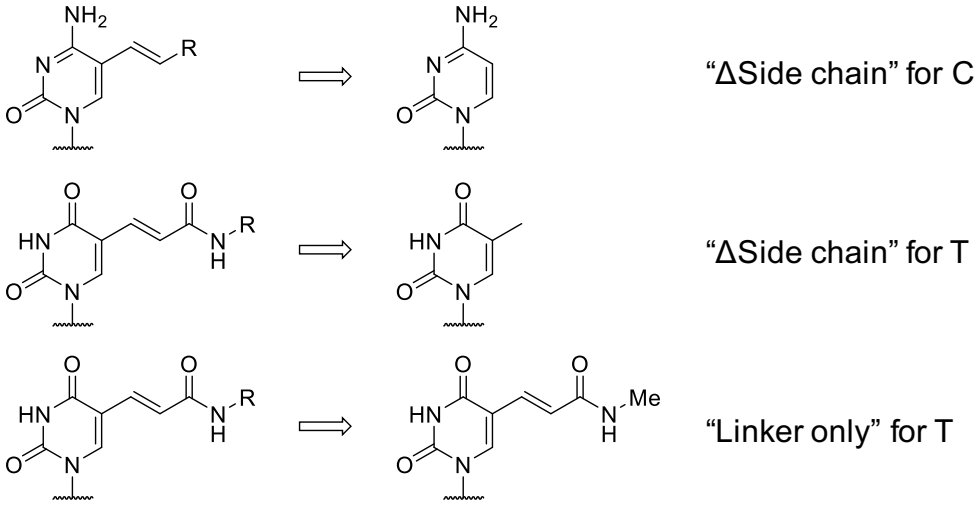
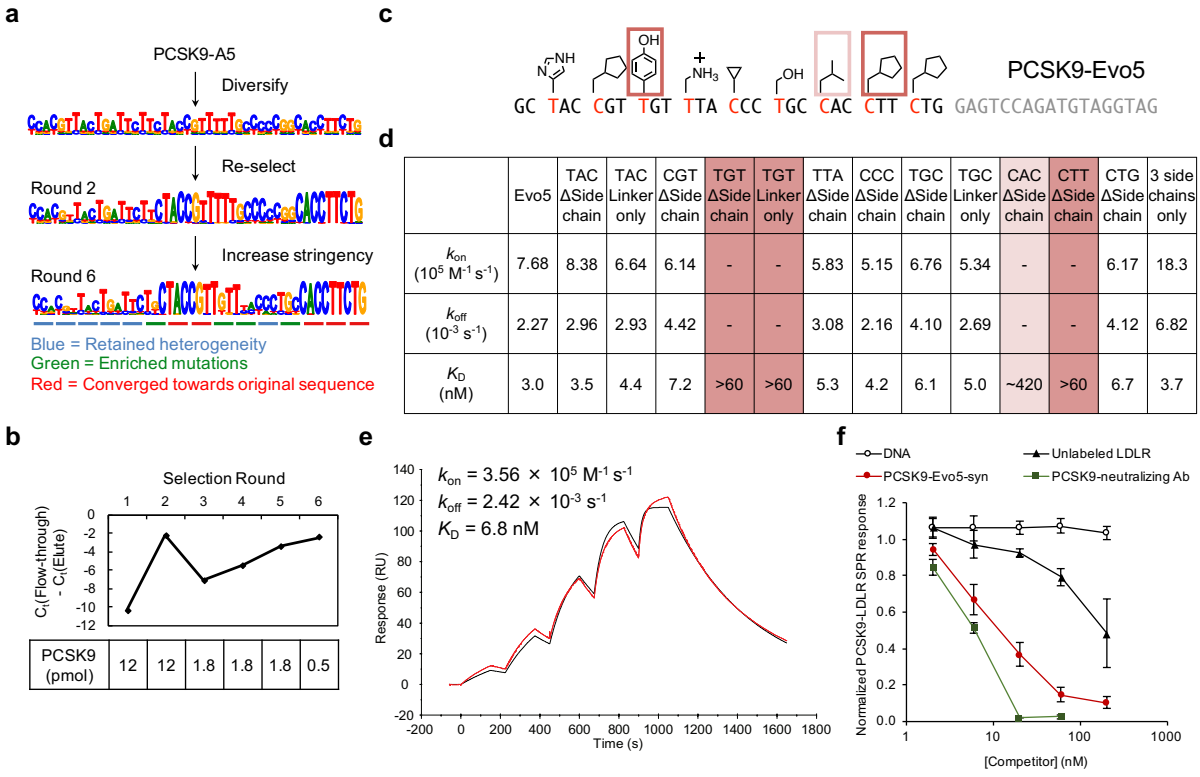


Figure 2.4 | Structures of bases in side-chain mutants characterized during structure-activity relationship studies of PCSK9-A5 and PCSK9-Evo5. Linker only version is shown for reference.

2.4 Diversification and evolution of HFNAPs for PCSK9 binding

Given the vast sequence space of the HFNAP library (3×10^{22} possible polymers), evolution would likely generate polymer variants with improved activity over those polymers present in the initial population of 3×10^{12} HFNAP molecules. To evolve the PCSK9-A5 polymer into variants with improved PCSK9 affinity, we synthesized a library of mutated PCSK9-A5 templates containing 79% identity and 21% diversity (79:21 at the pyrimidine-only first position and 79:7:7:7 at the second and third positions of each codon) for each nucleotide in the variable

region (**Figure 2.5a**). We subjected the resulting mutated PCSK9-A5 library to six additional iterated cycles of translation and selection for PCSK9 binding (**Figure 2.5b**). After just one round of enrichment at a stringency level comparable to that of the last round of our initial selection, the mutated PCSK9-A5 library exhibited bulk affinity for PCSK9-conjugated beads (**Figure 2.5b**). High-throughput DNA sequencing revealed that the mutant population after a second round of translation and selection began to converge toward the sequence of PCSK9-A5 at many positions (**Figure 2.5a**). In subsequent rounds, we reduced the amount of immobilized PCSK9 to further increase selection stringency. Four additional rounds of translation and selection resulted in steadily improved retention of the polymer population on immobilized PCSK9 (**Figure 2.5b**).



	Evo5	TAC ΔSide chain	TAC Linker only	CGT ΔSide chain	TGT ΔSide chain	TGT Linker only	TTA ΔSide chain	CCC ΔSide chain	TGC ΔSide chain	TGC Linker only	CAC ΔSide chain	CTT ΔSide chain	CTG ΔSide chain	3 side chains only
k_{on} (10^5 M^{-1} s^{-1})	7.68	8.38	6.64	6.14	-	-	5.83	5.15	6.76	5.34	-	-	6.17	18.3
k_{off} (10^{-3} s^{-1})	2.27	2.96	2.93	4.42	-	-	3.08	2.16	4.10	2.69	-	-	4.12	6.82
K_D (nM)	3.0	3.5	4.4	7.2	>60	>60	5.3	4.2	6.1	5.0	~420	>60	6.7	3.7

Figure 2.5 | Evolution of an improved PCSK9-binding polymer. a, Evolution scheme and DNA sequencing results of the diversification and iterated selection of PCSK9-A5 variants with increased PCSK9 binding activity. **b**, Affinity maturation of the diversified PCSK9-A5 pool. The evolving polymer pool's bulk affinity to immobilized PCSK9 was assessed by quantifying the amount of HFNAP in the flow-through and the elution at each round of selection by quantitative PCR. **c**, Sequence and side-chain structure of the resulting PCSK9-Evo5 polymer. Side chains

Figure 2.5 (Continued)
essential for binding activity are boxed. **d**, Kinetic parameters for binding of PCSK9-Evo5 or its side-chain-deficient variants to PCSK9 protein, as measured by SPR. For the variants "TGT ΔSide chain", "TGT Linker only", and "CTT ΔSide chain", no SPR signal was observed at the highest analyte concentration tested (60 nM PCSK9). For the variant "CAC ΔSide chain", the binding interaction fits a two-state reaction kinetic model with $K_D \approx 420$ nM. Representative sensograms are provided in **Supplementary Figure 2.1**. **e**, SPR sensogram characterizing binding kinetics between PCSK9-Evo5-syn and surface-immobilized PCSK9 protein. The concentrations of injected PCSK9-Evo5-syn were 1.8, 6, 18, 60, and 180 nM. The observed sensogram is shown in red and the fitted curve with the kinetic parameters listed is shown in black. **f**, SPR response on an LDLR-coated surface produced by flowing PCSK9 in the presence of either PCSK9-Evo5-syn, unfunctionalized DNA of identical sequence to PCSK9-Evo5-syn, unlabeled LDLR, or a known PCSK9-neutralizing monoclonal antibody. The SPR response shown is normalized to the response in experiments without any competitor (defined as an SPR response of 1). Error bars represent s.d. (n = 3).

High-throughput sequencing revealed new consensus codons at four out of 15 positions within the population of evolved polymers (**Figure 2.5a**). Among these four positions, codons 6, 10, and 12 evolved a different side-chain compared with that of PCSK9-A5, while codon 9 evolved a different codon (TGT instead of TTT) encoding the same phenol side-chain at this position. In addition, five other codons converged to the original sequence in PCSK9-A5, and six other codon positions, mostly near the 5' end, retained sequence heterogeneity introduced during mutation, suggesting that these positions do not strongly contribute to binding activity (**Figure 2.5a**). The three side-chain functional groups that were shown to be crucial to the PCSK9-binding activity of PCSK9-A5 were all maintained in the consensus sequence of the evolved polymer population.

We synthesized a biotinylated, truncated HFNAP (designated PCSK9-Evo5; **Figure 2.5c**) retaining only the evolved consensus sequence and the 3' constant region, and measured its affinity for PCSK9 to be $K_D = 3.0$ nM (**Figure 2.5d** and **Supplementary Figure 2.1**), representing a 33-fold increase over that of PCSK9-A5. In contrast, a similarly truncated version of PCSK9-A5 (designated PCSK9-A5Trunc) did not substantially improve affinity ($K_D = 68$ nM) over that of full-length PCSK9-A5 ($K_D = 98$ nM) (**Figure 2.6a**). Additional truncation of PCSK9-Evo5 from the 3'-end resulted in complete loss of affinity (**Figure 2.6b**). Similar to the structure-activity relationships observed for PCSK9-A5, the removal of a single phenol or cyclopentyl

side-chain from PCSK9-Evo5 abolished its affinity to PCSK9 protein (**Figure 2.5d**), and the removal of the isopentyl side-chain severely impaired target binding (the binding interaction approximately fits a two-state reaction kinetic model with $K_D \approx 420$ nM; **Figure 2.5d** and **Supplementary Figure 2.1**). The individual removal of other side-chains had less significant effects on binding affinity; indeed an HFNAP containing only three side chains (phenol at position 9, isopentyl at position 21, and cyclopentyl at position 24) and the rest of the Evo5 sequence as unfunctionalized DNA maintains strong binding to PCSK9 ($K_D = 3.7$ nM; **Figure 2.5d** and **Supplementary Figure 2.1**). Together, these results establish the evolution (iterated selection with intervening mutation and replication) of a sequence-defined synthetic polymer with improved target affinity.

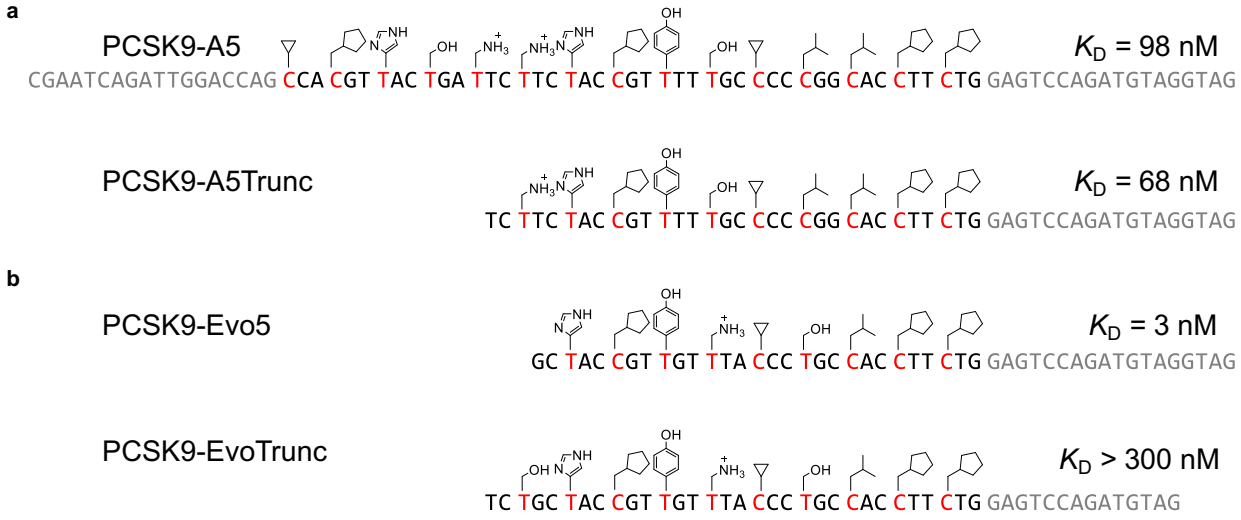


Figure 2.6 | Effects of truncating HFNAPs on their binding affinity to PCSK9 protein. a, Effect of truncating PCSK9-A5 from the 5' end. **b**, Effects of truncating PCSK9-Evo5 from the 3' end. No binding of up to 300 nM solution-phase PCSK9 to surface-immobilized PCSK9-EvoTrunc was observed by SPR.

To gain further insight into the secondary structures of PCSK9-binding HFNAPs and the affinity improvement of PCSK9-Evo5 over PCSK9-A5, we subjected the DNA sequences of PCSK9-Evo5 and PCSK9-A5Trunc to secondary structure prediction using mFold[45]. The 3' region, conserved between PCSK9-A5Trunc and PCSK9-Evo5, was predicted to form a stem-loop structure containing a bulge. The 5' region, which includes all the consensus mutations

leading to PCSK9-Evo5, contained significantly less (in the case of PCSK9-A5Trunc) or no (in the case of PCSK9-Evo5) predicted secondary structure based on Watson-Crick base-pairing (**Supplementary Figure 2.2**). Consistent with the observed side-chain structure-activity relationships shown in **Figure 2.5d**, these calculations further suggest crucial contributions of side chains, which are not captured in the secondary structure model, to HFNAP folding and target-binding activity.

To further characterize PCSK9-Evo5, we executed the multi-milligram-scale total synthesis of a variant of PCSK9-Evo5 (PCSK9-Evo5-syn), which has an additional 3' inverted dT base for exonuclease resistance[46], using standard phosphoramidite chemistry on solid support (**Supplementary Figure 2.3**). We planned to install all side-chain-functionalized building blocks entirely through corresponding side-chain-functionalized phosphoramidite reagents, with the nucleophilic functional groups protected with base-labile protecting groups that would be removed under standard oligonucleotide deprotection conditions. Although most of the side-chain-functionalized nucleoside phosphoramidites were readily synthesized, the imidazole-bearing thymidine reagent proved difficult to prepare. We therefore synthesized PCSK9-Evo5-syn on solid phase with an activated ester (NHS-carboxy-dT) in place of the imidazole-functionalized thymidine, and coupled histamine to that position in the bead-bound polymer chain to install the imidazole side-chain before global deprotection and cleavage from solid support with ammonium hydroxide (**Supplementary Figure 2.3**). The identity of PCSK9-Evo5-syn (approximately 2 mg from a one-micromole solid-phase synthesis, ~13% overall yield) was confirmed by high-resolution mass spectrometry (**Supplementary Figure 2.4**). High-affinity binding (K_D = 6.8 nM) between totally synthetic PCSK9-Evo5-syn and biotinylated Avi-tagged PCSK9 was confirmed by SPR (**Figure 2.5e**). In addition, binding of Alexa Fluor 647-labeled Evo5 at low nM concentrations to PCSK9 was confirmed in an electrophoretic mobility shift assay (EMSA) (**Supplementary Figure 2.5**).

PCSK9 regulates cholesterol metabolism by binding the LDL receptor (LDLR) and promoting the lysosomal degradation of LDLR [42]. We tested the ability of PCSK9-Evo5-syn to disrupt PCSK9-LDLR binding in an SPR assay. PCSK9-Evo5-syn dose-dependently reduced binding of PCSK9 to surface-immobilized LDLR (**Figure 2.5f** and **Supplementary Figure 2.6**). The potency of PCSK9-Evo5-syn inhibition of PCSK9-LDLR binding (IC_{50} = ~9 nM) was similar to that of a known PCSK9-neutralizing monoclonal antibody (**Figure 2.5f** and **Supplementary Figure 2.6**). In contrast, unfunctionalized DNA of the same sequence as PCSK9-Evo5-syn produced no inhibitory effect (**Figure 2.5f** and **Supplementary Figure 2.6**), consistent with the necessity of the side chains implicated in PCSK9 binding (**Figure 2.2c-e** and **Figure 2.5c-d**). We also tested the affinity of surface-immobilized PCSK9-Evo5 to different PCSK9 protein constructs and found that a truncated PCSK9 variant lacking the prodomain exhibited no apparent binding to PCSK9-Evo5 (**Supplementary Figure 2.7**), implicating the PCSK9 prodomain, known to be involved in a secondary binding interface between PCSK9 and LDLR[47], in mediating the interaction of PCSK9 with PCSK9-Evo5.

2.5 *In vitro* selections against IL-6

To test the generality of our polymer evolution system and to investigate the potential of this new class of polymers to evolve receptors to different proteins, we performed a separate selection for HFNAPs that bind a protein unrelated to PCSK9. We chose human interleukin-6 (IL-6), a key cytokine involved in inflammation and the target of many drugs and drug candidates[48], including modified DNA aptamers[24,25]. After seven iterated cycles of translation, selection for binding to immobilized IL-6 protein, reverse translation, and amplification, the most abundant sequence accounted for 3.6% of the population (**Figure 2.7a**). The top seven HFNAPs were individually synthesized and assayed for binding to immobilized IL-6. Based on its high apparent binding activity to immobilized IL-6, but not to immobilized PCSK9 (**Figure 2.7a**), the HFNAP IL6-A7 (**Figure 2.7b**) was chosen for further characterization. Binding of

biotinylated IL6-A7 to the target IL-6 protein was confirmed by SPR. Although the binding kinetics of IL6-A7 to IL-6 protein did not conform to a classical one-to-one binding model, a phenomenon often observed in aptamer-protein binding[18,24], fitting to a heterogeneous ligand model resulted in an apparent affinity of K_D = 12 nM for the major component and K_D = 22 nM for the minor component (**Figure 2.7c**).

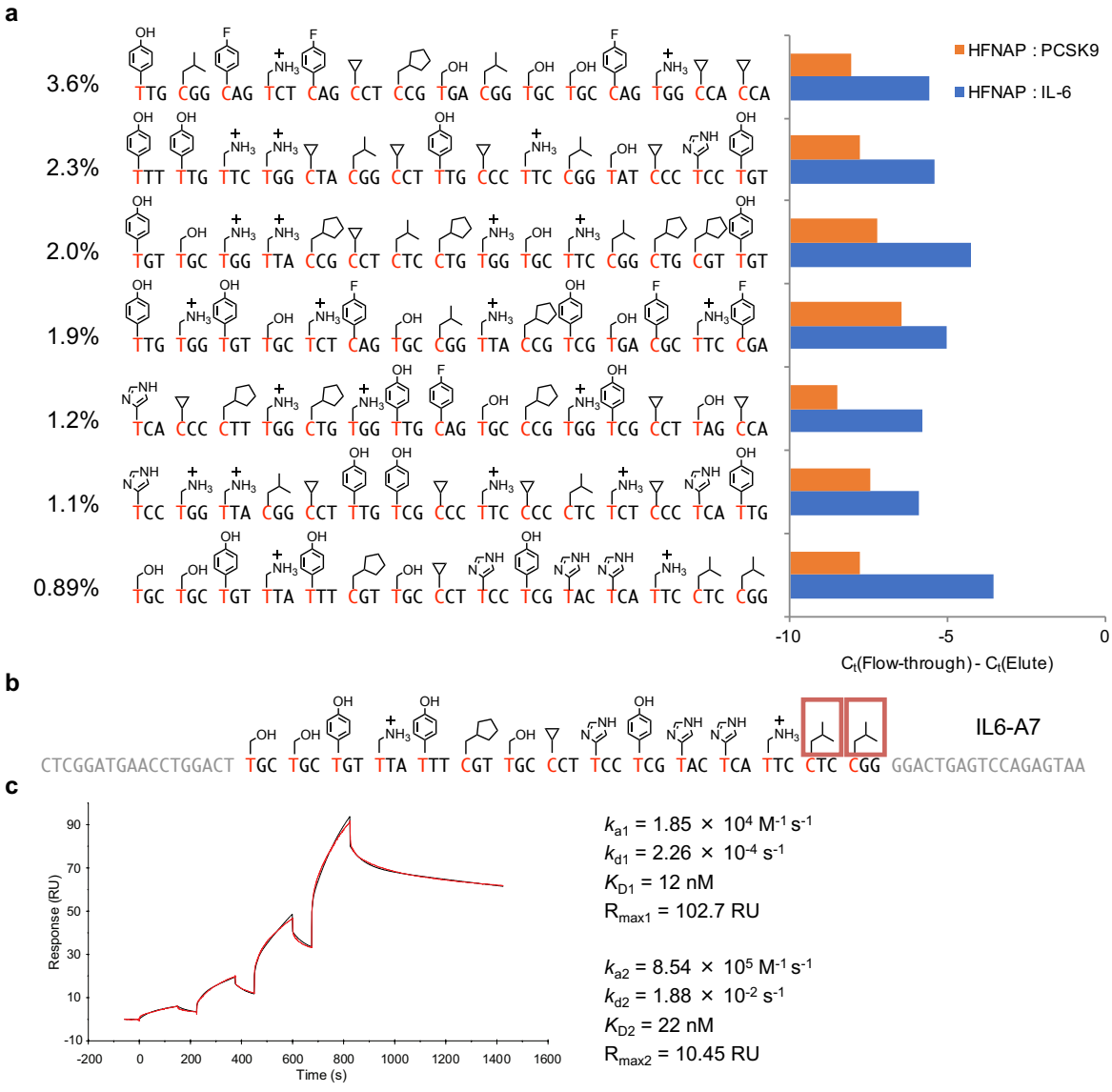


Figure 2.7 | Characterization of IL-6-binding HFNAPs selected from a random library. a, Retention of individual selection-enriched HFNAPs on immobilized IL-6 (target; blue bars) or immobilized PCSK9 (non-target; red bars). The percentages of each sequence in the pool after seven rounds of selection are listed to the left. The experiment was performed as a screen to identify hits for further characterization, and was not performed in replicates. **b**, Sequence and

Figure 2.7 (Continued)
side-chain structure of IL6-A7. Side-chains essential for binding activity are boxed. **c**, SPR sensogram characterizing binding kinetics between biotinylated IL6-A7 and its target IL-6 protein. The concentrations of injected IL-6 were 10, 30, 100, and 300 nM. The observed sensogram is shown in red and the fitted curve with the kinetic parameters listed is shown in black.

The binding kinetics and affinity of IL6-A7 to *E. coli*-expressed IL-6 (used in the selection) and to human HEK293 cell-expressed IL-6 protein were comparable (**Supplementary Figure 2.8**). Next we measured by SPR the IL-6 affinity of IL6-A7 side-chain mutants in which all instances of a trimer building block were replaced by the corresponding trinucleotide lacking any side-chain. The mutants missing isopentyl side-chain in either codon 14 or codon 15 exhibited severely impaired binding (**Supplementary Figure 2.8**), implicating these two side chains as key determinants of IL-6 binding activity in IL6-A7.

2.5 Discussion

We used a ligase-mediated DNA-templated polymerization system and *in vitro* selection to evolve HFNAPs, nucleic acid polymers that are densely functionalized with chemically diverse side-chains. HFNAPs that bind PCSK9 and IL-6 were selected from random polymer libraries. Through diversification and reselection, we evolved an improved PCSK9-binding HFNAP (Evo5) with K_D = 3 nM. We characterized structure-activity relationships within this polymer, revealing side chains at specific positions that are critical to target-binding activity. Evo5 potently inhibits binding between PCSK9 and the LDL receptor.

Collectively, these findings represent to our knowledge the first laboratory evolution of functional, genetically encoded sequence-defined synthetic polymers without the constraints imposed by polymerases or ribosomes. The DNA-templated, ligase-based translation system developed here supports many rounds of iterated selection of polymers with diverse side-chains, including side-chains that mimic—and extend beyond—the repertoire of amino acid side-chains found in proteins. Both the PCSK9-binding and IL-6-binding polymers generated in

this system exhibit position-dependent and side-chain dependent structure-activity relationships resembling those of proteins. Finally, we note that the PCSK9-binding polymers generated in this work are strongly dependent on the presence of multiple side-chains with different physical properties, consistent with the importance of chemical diversity to the functional potential of these polymers.

Recently, Gawande and coworkers performed selections for PCSK9 aptamers from modified DNA libraries in which all instances of one or both pyrimidines (C and/or T) were replaced by side-chain-functionalized variants[27]. High-affinity aptamers with dissociation constants similar to those of FDA-approved anti-PCSK9 monoclonal antibodies (evolocumab, K_D = 8.0 pM[49], and alirocumab, K_D = 0.58 nM[50]) were enriched from doubly modified libraries in which hydrophobic or phenolic side chains were present on 50% of the nucleobases on average. Aptamers enriched from singly modified libraries (25% hydrophobic side chains on average) were less potent ($K_D \geq 100$ pM), while libraries containing hydrophilic side chains or consisting of unmodified DNA did not produce aptamers with $K_D \leq 30$ nM. Consistent with their findings, the highest affinity binders from our HFNAP library, which contains a roughly equal mix of hydrophilic and hydrophobic side chains installed at 33% total frequency, has K_D = 3 nM to PCSK9. We note, though, that different modifications may be suitable for other applications, as demonstrated by DNA-based catalysts functionalized with nitrogen nucleophiles as side-chains[29–33,35]. Therefore, the diverse, balanced set of side-chains in HFNAPs, similar to the natural repertoire of proteins, may be more versatile in other settings.

The ligase-based polymerization method allows straightforward redesign of the genetic code of the polymer, as we exploited to expand the sequence and structural diversity of the polymers used in this work compared with those of our original system[39]. This feature also enables researchers to generate and select HFNAPs with side-chains tailored toward specific applications, as recently demonstrated by Hili and coworkers for scaffolding peptides on a DNA

template[51]. Moreover, the side-chain flexibility of this polymer evolution system raises the possibility of performing parallel evolution experiments with libraries of different side-chain compositions to shed light on the fundamental relationship between the structure of the building blocks in a genetic code and the evolutionary potential of the resulting polymers.

2.6 Methods

Additional experimental procedures and characterization data are provided in the Supplementary Information.

Synthesis of HFNAP by templated translation via DNA ligase-mediated polymerization. DNA template [up to 10 pmol, either in solution or immobilized on MyOne Streptavidin C1 magnetic beads (ThermoFisher Scientific)], polymerization initiation and termination primers (1.5 equivalents each relative to template), functionalized trinucleotide building blocks (10 equivalents relative to template for each occurrence of the corresponding codon) and 10x T4 RNA ligase reaction buffer (New England Biolabs; 1 µL) were mixed in a total volume of 8 µL in a PCR tube. The mixture was subjected to the following temperature program on a thermocycler: 95 °C for 10 sec; 65 °C for 4 min; a ramp from 65 °C to 4 °C at 0.1 °C per 10 s. To the PCR tube were added 1 µL of 10 mM ATP and 1 µL of T3 DNA ligase (New England Biolabs). The reaction was incubated at 4 °C for 12 h and then at 16 °C for 2 h.

Selections of HFNAP that bind protein targets. Selection bait was prepared by immobilizing recombinant protein onto AminoLink Plus aldehyde-functionalized agarose resin via reductive amination with a MicroLink Protein Coupling Kit (ThermoFisher Scientific). Loading was 1 mg PCSK9 protein (ACROBiosystems) per mL resin for the initial PCSK9 binder selection and the first two rounds of PCSK9 binder re-selection; 150 µg PCSK9 per mL resin for rounds 3-5 of the

re-selection; 40 µg protein per mL resin for round 6 of the re-selection; and 250 µg IL-6 protein (PeproTech) per mL resin throughout the IL-6 binder selection.

To initiate the selection, primer extension was performed with a biotinylated primer on 5 pmol of the sense strand randomized DNA library (Integrated DNA Technologies or TriLink BioTechnologies) with Klenow (exo-) polymerase (New England Biolabs). Biotinylated species was captured on streptavidin magnetic beads, which were then washed three times with 20 mM NaOH and then twice with 1x T4 RNA ligase reaction buffer. The bead-immobilized template strand library was then translated in a ligase-mediated polymerization to produce HFNAPs. The beads were suspended in 20mM NaOH to denature the HFNAP-template hybrids. HFNAP strands in the supernatant were cleaned up with a MinElute column (Qiagen).

The HFNAP library was added to DPBS (with calcium and magnesium; Lonza) supplemented with BSA (0.1 mg/ml final) and Tween-20 (0.01% final), and then incubated with PCSK9 resin in a micro-spin filtration column (Pierce) at room temperature for 1 h on a rotor. (The amounts of resin-bound protein used in each round of the PCSK9 selection are indicated in **Figure 2.2b**. Throughout the IL-6 selection, 240 pmol of immobilized IL-6 protein was used in each round.) The flow-through was collected by centrifugation at 1000 g into an Eppendorf tube. The beads were washed three times with 50 µL each of DPBS. The column was cut open, and the beads were collected by centrifugation into an Eppendorf tube and then incubated in 50 µL of lithium dodecyl sulfate (LDS) loading buffer (Life Technologies) at 95 °C for 15 min. After cooling, HFNAP strands were isolated from the mixture by cleaning up with a QiaQuick column (Qiagen), eluting the HFNAP into 50 µL of water.

Samples of 1 µL each from the flow-through, the three washes, and the elution were quantified by qPCR (20 µL reaction volume) using the iTaq Supermix (Bio-rad). The number of cycles for the qPCR curve of the elution sample to reach the end of exponential growth was used as the number of cycles for the preparative PCR (400 µL reaction volume split into 8 $\times$ 50 µL) of the selection elution pool (20 µL) with Q5 Hot Start High-Fidelity 2X Master Mix (New

England Biolabs), using a biotinylated primer for the strand that will serve as translation template. The PCR product was cleaned up with a MinElute column and PAGE purified on a non-denaturing 10% TBE gel. A portion (indicated in **Figure 2.2b**) of the dsDNA product was captured on streptavidin magnetic beads to initiate the next round of selection.

Surface plasmon resonance (SPR) assays. All SPR assays were performed at 25 °C on a Biacore X100 or Biacore T200 (GE Healthcare Life Sciences). Binding kinetics between enzymatically synthesized biotinylated HFNAPs and unlabeled recombinant proteins were measured using single-cycle kinetics with the Biotin CAPture kit (GE Life Sciences) using 0.9x HBS-EP buffer (GE Life Sciences) at a flow rate of 30 μL/min. The injected PCSK9 concentration ranged from 10 to 300 nM for PCSK9-A5 and its variants, or from 2 to 60 nM for PCSK9-Evo5 and its variants. The injected IL-6 concentration ranged from 10 to 300 nM.

Binding kinetics between chemically synthesized PCSK9-Evo5-syn and biotinylated Avi-tagged PCSK9 (ACROBiosystems) were measured using single-cycle kinetics on a Series S SA chip (GE Life Sciences) using 0.9x HBS-EP buffer at a flow rate of 30 μL/min. The injected PCSK9-Evo5-syn concentration ranged from 1.8 to 180 nM.

Binding of PCSK9 on surface-immobilized LDLR in the presence of various competing agents was measured on a Series S SA chip using 10 mM HEPES, 150 mM NaCl, 0.1 mM $CaCl_2$, 0.005% Tween-20, pH 7.5 as bulk buffer at a flow rate of 10 μL/min. The injected solutions contained 20 nM PCSK9 and various competing agents ranging from 2 to 200 nM.

Isolation of single-stranded HFNAP following ligase-mediated polymerization. To synthesize and isolate an unbiotinylated HFNAP, unbiotinylated primers and a doubly biotinylated ssDNA template was used in a solution phase polymerization reaction. After the reaction incubation period, the reaction mixture was combined with MyOne Streptavidin C1

magnetic beads (ThermoFisher Scientific, 65002; 1 µL of the stock 1% suspension per 4 pmol of biotinylated template), and then an equal volume of 2x bind-and-wash buffer (2M NaCl, 2mM EDTA, 20 mM Tris-HCl, pH 7.5) was added. After 30 minutes of incubation on a rotor, the supernatant was removed by magnetic separation, and the beads were suspended 18 µL of 20 mM NaOH. The supernatant was combined with 12 µL of formamide denaturing mix (95% formamide, 1 mM EDTA) and run on a 10% TBE-urea PAGE gel. Desired product was visualized by UV shadowing at 265 nm against a TLC plate (with F254 indicator), excised from the gel, eluted in 200 µL of TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH 8.0) overnight, filtered, mixed with 2 mL of ssDNA column loading mix (40:60:0.5 v/v/v of saturated aqueous guanidinium chloride/isopropanol/3M sodium acetate, pH 5.2) and cleaned up with a Qiagen QiaQuick column. Typical isolated yield of HFNAP from 200 pmol of template was between 5 and 15 pmol as determined by Nanodrop or quantitative PCR. For the validation of sequence specificity of translation and amplification (**Figure 2.1d**), a small sample (~1 fmol) was amplified by PCR using Q5 DNA polymerase and the primers T7-out-PCR2 and pp2-library. The amplicon was purified by PAGE on a nondenaturing 10% TBE gel and subjected to Sanger sequencing.

Synthesis and isolation of a biotinylated HFNAP followed the same procedure as above, except that an unbiotinylated ssDNA template was used, and one of polymerization primers was doubly biotinylated. After polymerization reaction, streptavidin bead capture, and alkaline denaturation, the bead-bound biotinylated molecules were eluted by heating in 95% formamide, 1 mM EDTA at 95 °C for 30 min. After magnetic separation, desired product was isolated from the supernatant by PAGE on a denaturing TBE-urea gel. Typical isolated yield of HFNAP from 200 pmol of template was between 5 and 15 pmol. Synthesis and isolation of a biotinylated, truncated HFNAP (such as PCSK9-Evo5) followed the same procedure, except that the primer contained a 2'-deoxy-U nucleotide, and the ligation reaction mixture was treated with USER enzyme (New England Biolabs) at 37 °C for 2 h before proceeding to streptavidin bead capture.

During the selections, polymerization reactions were performed with templates immobilized on streptavidin beads and processed as detailed in the main text Methods section.

High-throughput DNA sequencing and data analysis. Small samples from the elution pool of selection rounds were amplified by PCR using Q5 Hot Start High-Fidelity 2X Master Mix to sub-saturation number of cycles (determined during the selection by qPCR) with primers that install flanking sequences. The amplicons were PAGE-purified and amplified by PCR with Illumina adapter primers. The amplicons were again PAGE-purified and subjected to high-throughput sequencing on an Illumina MiSeq.

The FASTQ files from high-throughput sequencing were first processed with CutAdapt[52] for the following operations: a quality-based trim (with a threshold Phred score of 30), removal of constant regions (with a one-base error tolerance in each region; sequences were discarded if either constant region was not found), and filtering for the correct length (45) in the remaining sequence. Sequences that could not be completely parsed into trimer codons (whose first nucleobase should always be C or T) were discarded.

For the initial PCSK9 selection and the IL-6 selection, the copy numbers of all unique sequences were tallied and the unique sequences above 5 reads per million were clustered using FASTAptamer[53]. Sequence logos for PCSK9 selection-enriched sequences were then generated for individual clusters with WebLogo 3[54]. For the PCSK9-A5 evolution, sequence logos were generated directly from sequencing data with WebLogo 3.

Affinity characterization by bead retention assay. Candidate PCSK9-binding HFNAPs (from ligase-catalyzed polymerization reaction) or sequence-matched unfunctionalized DNA (0.5 pmol) in 50 µL of DPBS (with calcium and magnesium) containing 0.1 mg/ml BSA and 0.01 % Tween-20 was incubated with 1 µL of either PCSK9 beads or thrombin beads (1 mg protein per mL resin) in a micro-spin filtration column at room temperature for 1 h on a rotor. Following the

same procedure described for the selection, flow-through was collected, the beads were washed three times and the retained HFNAP or DNA was eluted by heating, and the amount of amplifiable HFNAP or DNA in the flow-through, wash, and elution samples were quantified by qPCR.

Candidate IL-6-binding HFNAPs and sequence-matched DNA were similarly assayed on PCSK9 beads (prepared as above, but serving as negative control) or on IL-6 beads (0.25 mg protein per mL resin).

Detailed procedures for surface plasmon resonance (SPR) assays. All SPR assays were performed at 25 °C on a Biacore X100 or Biacore T200 (GE Healthcare Life Sciences).

Binding kinetics between enzymatically synthesized biotinylated HFNAPs and unlabeled PCSK9 (ACROBiosystems, PC9-H5223) were measured using single-cycle kinetics with the Biotin CAPture kit (GE Life Sciences, 28920233 or 28920234). HBS-EP buffer (GE Life Sciences, BR100188), diluted by MilliQ water to 0.9x, was used as the bulk running buffer. Each experiment consisted of three start-up cycles followed by multiple data collection and blank cycles. In each data collection cycle, the CAP reagent was injected onto both active and control flow cells of the CAP chip to generate streptavidin-coated surfaces, and then a doubly biotinylated HFNAP was injected onto the active flow cell as the immobilized ligand. Afterwards, four ascending concentrations of PCSK9 protein [10, 30, 100, 300 nM protein supplemented with 1 mg/ml salmon sperm DNA (Invitrogen, 15632-011) as nonspecific binding reducer for PCSK9-A5 and its variants; 2, 6, 20, 60 nM protein without salmon sperm DNA for PCSK9-Evo5 and its variants] in 0.9x HBS-EP were injected onto both flow cells in series at 30 μL/min for 150 seconds each, followed by 240 seconds of dissociation. Both flow cells were then regenerated with a 3:1 mixture of 8 M guanidinium chloride and 1 M NaOH following manufacturer's instructions. Blank cycles were run similarly except that 0.9x HBS-EP (containing 1 mg/ml salmon sperm DNA when PCSK9-A5 and its variants were assayed) without PCSK9 protein

was injected. As signals from blank cycles were similar regardless of the immobilized HFNAP, one blank cycle was run for every two data collection cycles. Kinetic parameters were fitted to double-blank-subtracted sensograms using BIAEvaluation software under a 1:1 binding model, unless stated otherwise. Binding between biotinylated Evo5 and truncated PCSK9 protein missing the prodomain ("human mature PCSK9", ACROBiosystems, PC9-H5226) was also assayed using this protocol.

Binding kinetics between enzymatically synthesized biotinylated HFNAPs and unlabeled IL-6 protein, expressed in either *E. coli* (PeproTech, 200-06) or human HEK293 cells (ACROBiosystems, IL6-H4218), were assayed similarly with the following differences. Four ascending concentrations (10, 30, 100, 300 nM) of IL-6 without additional nonspecific binding reducer were injected in the single-cycle kinetic runs. As the binding kinetics did not fit a classical 1:1 binding model, a heterogeneous ligand model was used to fit the double-blank-subtracted sensograms.

Binding kinetics between chemically synthesized Evo5-syn and biotinylated Avi-tagged PCSK9 (ACROBiosystems, PC9-H82E7) were measured using single-cycle kinetics with on a Series S SA chip (GE Life Sciences, BR100531) using 0.9x HBS as the bulk running buffer. Both active and control flow cells were conditioned with three consecutive one-minute injections of 1 M NaCl in 50 mM NaOH. Biotinylated PCSK9 in 0.9x HBS buffer was immobilized onto the active flow cell to ~1000 RU. Either five portions of buffer or five ascending concentrations of Evo5-syn (1.8, 6, 18, 60, 180 nM) were injected onto both flow cells in series at 30 μL/min for 150 seconds each, followed by 600 seconds of dissociation. Kinetic parameters were fitted to double-blank-subtracted sensograms under a 1:1 binding model using BIAEvaluation software.

Binding of PCSK9 on surface-immobilized LDLR in the presence of various competing agents was measured on a Series S SA chip. The bulk running buffer was 10 mM HEPES, 150 mM NaCl, 0.1 mM $CaCl_2$, 0.005% Tween-20, pH 7.5. Both active and control flow cells were conditioned with three consecutive one-minute injections of 1 M NaCl in 50 mM NaOH, and then

biotinylated Avi-tagged LDLR (BPS Bioscience, 71206) was immobilized onto the active flow cell to ~2000 RU. In each data collection cycle, a solution consisting of PCSK9 (20 nM final), a carboxymethyl dextran-based non-specific binding reducer (GE Healthcare, BR-1006-91, 1 mg/ml final), and varying concentrations (0, 2, 6, 20, 60, or 200 nM final) of PCSK9-Evo5-syn, Evo5DNA-InvdT, unlabeled LDLR (AcroBiosystems, LDR-H5224), or a known PCSK9-neutralizing antibody (BPS Bioscience, 71207) in bulk running buffer was injected at 10 μL/min for 420 s, followed by 15 s of dissociation. The surface was regenerated using two consecutive one-minute injections of 50 mM HCl. Blank cycles were run similarly except that running buffer containing 1 mg/ml non-specific binding reducer without protein was injected. Response levels at the end of the injection periods from double-blank-subtracted sensograms were recorded.

Electrophoretic mobility shift assay (EMSA). A 7.5% Tris-Glycine polyacrylamide gel (Bio-rad, 5671024) was pre-run at 150 V for 1 hour at 4 °C in a cold room. Mixtures (12 μl each) of PCSK9-Evo5-Fluor or a sequence-matched DNA (1 nM final), PCSK9 protein (ACROBiosystems, PC9-H5223, between 0.3 and 300 nM final), and salmon sperm DNA (Invitrogen, 15632-011, 30 μg/ml final) in 0.5x HBS-EP buffer (diluted from HBS-EP buffer, GE Life Sciences, BR100188) containing 3% v/v glycerol was incubated at 25 °C for 30 minutes, and then at 4 °C for 5 minutes. The mixtures were loaded onto the pre-run gel and run at 150 V for 15 minutes at 4 °C. The gel was imaged with a Typhoon imager using the Cy5 channel.

DNA secondary structure prediction. DNA secondary structure prediction was performed on the mfold Web server[55] using parameters that reflect conditions used in SPR assays: the folding temperature was set to 25 °C, the sodium concentration to 0.15 M, and the magnesium concentration to zero.

Data availability

The principal data supporting the findings of this work are available within the figures and the Supplementary Information. Additional data that support the findings of this study are available from the authors on request.

2.7 Acknowledgments

This work was supported by the DARPA Fold Fx program (N66001-14-2-4053), the NIH R01 EB022376 (formerly R01 GM065400) and R35 GM118062, and the Howard Hughes Medical Institute. Z.C. was partially supported by the Y. Kishi Graduate Prize in Chemistry and Chemical Biology sponsored by the Eisai Corporation. The authors wish to thank Dr. Jia Niu, Dr. Ryan Hili, Dr. Juan Pablo Maianti, Dr. Dmitry L. Usanov, and Alix Chan for helpful discussions. We thank the Center for Macromolecular Interactions in the Department of Biological Chemistry and Molecular Pharmacology at Harvard Medical School for access to Biacore T200 SPR instrument, and thank Dr. Matt Blome, Dr. Brian Lang, and Dr. Kelly Arnett for technical assistance on SPR experiments. We thank the Harvard FAS Small Molecule Mass Spectrometry facility for access to ESI-MS instruments and thank Dr. Sunia A. Trauger for technical assistance.

References

1. Ellington, A. D. & Szostak, J. W. In vitro selection of RNA molecules that bind specific ligands. *Nature* **346,** 818–822 (1990).

2. Tuerk, C. & Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249,** 505–510 (1990).

3. Robertson, D. L. & Joyce, G. F. Selection in vitro of an RNA enzyme that specifically cleaves single-stranded DNA. *Nature* **344,** 467–468 (1990).

4. Bock, L. C., Griffin, L. C., Latham, J. A., Vermaas, E. H. & Toole, J. J. Selection of single-stranded DNA molecules that bind and inhibit human thrombin. *Nature* **355,** 564–566 (1992).

5. Ellington, A. D. & Szostak, J. W. Selection in vitro of single-stranded DNA molecules that fold into specific ligand-binding structures. *Nature* **355,** 850–852 (1992).

6. Breaker, R. R. & Joyce, G. F. A DNA enzyme that cleaves RNA. *Chem. Biol.* **1,** 223–229 (1994).

7. Mattheakis, L. C., Bhatt, R. R. & Dower, W. J. An in vitro polysome display system for identifying ligands from very large peptide libraries. *Proc. Natl. Acad. Sci. U. S. A.* **91,** 9022–9026 (1994).

8. Roberts, R. W. & Szostak, J. W. RNA-peptide fusions for the in vitro selection of peptides and proteins. *Proc. Natl. Acad. Sci. U. S. A.* **94,** 12297–12302 (1997).

9. Nemoto, N., Miyamoto-Sato, E., Husimi, Y. & Yanagawa, H. In vitro virus: Bonding of mRNA bearing puromycin at the 3'-terminal end to the C-terminal end of its encoded protein on the ribosome in vitro. *FEBS Lett.* **414,** 405–408 (1997).

10. Yamaguchi, J. *et al.* cDNA display: a novel screening method for functional disulfide-rich peptides by solid-phase synthesis and stabilization of mRNA-protein fusions. *Nucleic Acids Res.* **37,** e108 (2009).

11. Brudno, Y. & Liu, D. R. Recent Progress Toward the Templated Synthesis and Directed Evolution of Sequence-Defined Synthetic Polymers. *Chem. Biol.* **16,** 265–276 (2009).

12. Chaput, J. C., Yu, H. & Zhang, S. The Emerging World of Synthetic Genetics. *Chem. Biol.* **19,** 1360–1371 (2012).

13. Pinheiro, V. B. & Holliger, P. The XNA world: progress towards replication and evolution of synthetic genetic polymers. *Curr. Opin. Chem. Biol.* **16,** 245–252 (2012).

14. Pinheiro, V. B. & Holliger, P. Towards XNA nanotechnology: new materials from synthetic genetic polymers. *Trends Biotechnol.* **32,** 321–328 (2014).

15. Hollenstein, M. Nucleoside Triphosphates — Building Blocks for the Modification of Nucleic Acids. *Molecules* **17,** 13569–13591 (2012).

16. Rogers, J. M. & Suga, H. Discovering functional, non-proteinogenic amino acid containing, peptides using genetic code reprogramming. *Org. Biomol. Chem.* **13,** 9353–9363 (2015).

17. Yu, H., Zhang, S. & Chaput, J. C. Darwinian evolution of an alternative genetic system provides support for TNA as an RNA progenitor. *Nat. Chem.* **4,** 183–187 (2012).

18. Pinheiro, V. B. *et al.* Synthetic Genetic Polymers Capable of Heredity and Evolution. *Science* **336,** 341–344 (2012).

19. Taylor, A. I. *et al.* Catalysts from synthetic genetic polymers. *Nature* **518,** 427–430 (2014).

20. Dunn, M. R. & Chaput, J. C. Reverse Transcription of Threose Nucleic Acid by a Naturally Occurring DNA Polymerase. *ChemBioChem* **17,** 1804–1808 (2016).

21. Vaught, J. D. *et al.* Expanding the chemistry of DNA for in vitro selection. *J. Am. Chem. Soc.* **132,** 4141–4151 (2010).

22. Gold, L. *et al.* Aptamer-Based Multiplexed Proteomic Technology for Biomarker Discovery. *PLoS One* **5,** e15004 (2010).

23. Davies, D. R. *et al.* Unique motifs and hydrophobic interactions shape the binding of modified DNA ligands to protein targets. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 19971–19976 (2012).

24. Gupta, S. *et al.* Chemically-Modified DNA Aptamers Bind Interleukin-6 with High Affinity and Inhibit Signaling by Blocking its Interaction with Interleukin-6 Receptor. *J. Biol. Chem.* **289,** 8706–8719 (2014).

25. Gelinas, A. D. *et al.* Crystal Structure of Interleukin-6 in Complex with a Modified Nucleic Acid Ligand. *J. Biol. Chem.* **289,** 8720–8734 (2014).

26. Imaizumi, Y. *et al.* Efficacy of Base-Modification on Target Binding of Small Molecule DNA Aptamers. *J. Am. Chem. Soc.* **135,** 9412–9419 (2013).

27. Gawande, B. N. *et al.* Selection of DNA aptamers with two modified bases. *Proc. Natl. Acad. Sci. U. S. A.* **114,** 2898–2903 (2017).

28. Tolle, F., Brändle, G. M., Matzner, D. & Mayer, G. A Versatile Approach Towards Nucleobase-Modified Aptamers. *Angew. Chemie Int. Ed.* **54,** 10971–10974 (2015).

29. Santoro, S. W., Joyce, G. F., Sakthivel, K., Gramatikova, S. & Barbas, C. F. RNA cleavage by a DNA enzyme with extended chemical functionality. *J. Am. Chem. Soc.* **122,** 2433–2439 (2000).

30. Perrin, D. M., Garestier, T. & Hélène, C. Expanding the catalytic repertoire of nucleic acid catalysts: simultaneous incorporation of two modified deoxyribonucleoside triphosphates bearing ammonium and imidazolyl functionalities. *Nucleosides Nucleotides* **18,** 377–391 (1999).

31. Perrin, D. M., Garestier, T. & Hélène, C. Bridging the Gap between Proteins and Nucleic Acids: A Metal-Independent RNAseA Mimic with Two Protein-Like Functionalities. *J. Am. Chem. Soc.* **123,** 1556–1563 (2001).

32. Lermer, L., Roupioz, Y., Ting, R. & Perrin, D. M. Toward an RNaseA mimic: A DNAzyme with imidazoles and cationic amines. *J. Am. Chem. Soc.* **124,** 9960–9961 (2002).

33. Hollenstein, M., Hipolito, C. J., Lam, C. H. & Perrin, D. M. A self-cleaving DNA enzyme modified with amines, guanidines and imidazoles operates independently of divalent metal cations (M2+). *Nucleic Acids Res.* **37,** 1638–1649 (2009).

34. Shoji, A., Kuwahara, M., Ozaki, H. & Sawai, H. Modified DNA Aptamer That Binds the (R)-Isomer of a Thalidomide Derivative with High Enantioselectivity. *J. Am. Chem. Soc.* **129,** 1456–1464 (2007).

35. Sidorov, A. V., Grasby, J. A. & Williams, D. M. Sequence-specific cleavage of RNA in the absence of divalent metal ions by a DNAzyme incorporating imidazolyl and amino functionalities. *Nucleic Acids Res.* **32,** 1591–1601 (2004).

36. Sefah, K. *et al.* In vitro selection with artificial expanded genetic information systems. *Proc. Natl. Acad. Sci.* **111,** 1449–1454 (2014).

37. Zhang, L. *et al.* Evolution of functional six-nucleotide DNA. *J. Am. Chem. Soc.* **137,** 6734–6737 (2015).

38. Kimoto, M., Yamashige, R., Matsunaga, K., Yokoyama, S. & Hirao, I. Generation of high-affinity DNA aptamers using an expanded genetic alphabet. *Nat. Biotechnol.* **31,** 453–457 (2013).

39. Hili, R., Niu, J. & Liu, D. R. DNA ligase-mediated translation of DNA into densely functionalized nucleic acid polymers. *J. Am. Chem. Soc.* **135,** 98–101 (2013).

40. Lei, Y., Kong, D. & Hili, R. A High-Fidelity Codon Set for the T4 DNA Ligase-Catalyzed Polymerization of Modified Oligonucleotides. *ACS Comb. Sci.* **17,** 716–721 (2015).

41. Cohen, J. C., Boerwinkle, E., Mosley, Thomas, J. & Hobbs, H. H. Sequence Variations in PCSK9, Low LDL, and Protection against Coronary Heart Disease. *New Engl. J. Med.* **354,** 1264–1272 (2006).

42. Lambert, G., Charlton, F., Rye, K.-A. & Piper, D. E. Molecular basis of PCSK9 function. *Atherosclerosis* **203,** 1–7 (2009).

43. Stein, E. A. *et al.* Effect of a Monoclonal Antibody to PCSK9 on LDL Cholesterol. *N. Engl. J. Med.* **366,** 1108–1118 (2012).

44. Brudno, Y., Birnbaum, M. E., Kleiner, R. E. & Liu, D. R. An in vitro translation, selection and amplification system for peptide nucleic acids. *Nat. Chem. Biol.* **6,** 148–155 (2010).

45. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31,** 3406–3415 (2003).

46. Ortigão, J. F. R. *et al.* Antisense Effect of Oligodeoxynucleotides with Inverted Terminal Internucleotidic Linkages: A Minimal Modification Protecting against Nucleolytic Degradation. *Antisense Res. Dev.* **2,** 129–146 (1992).

47. Lo Surdo, P. *et al.* Mechanistic implications for LDL receptor degradation from the PCSK9/LDLR structure at neutral pH. *EMBO Rep.* **12,** 1300–1305 (2011).

48. Hunter, C. A. & Jones, S. A. IL-6 as a keystone cytokine in health and disease. *Nat. Immunol.* **16,** 448–457 (2015).

49. Gibbs, J. P. *et al.* Impact of Target-Mediated Elimination on the Dose and Regimen of Evolocumab, a Human Monoclonal Antibody Against Proprotein Convertase Subtilisin/Kexin Type 9 (PCSK9). *J. Clin. Pharmacol.* **57,** 616–626 (2017).

50. Kühnast, S. *et al.* Alirocumab inhibits atherosclerosis, improves the plaque morphology, and enhances the effects of a statin. *J. Lipid Res.* **55,** 2103–2112 (2014).

51. Guo, C., Watkins, C. P. & Hili, R. Sequence-Defined Scaffolding of Peptides on Nucleic Acid Polymers. *J. Am. Chem. Soc.* **137,** 11191–11196 (2015).

52. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10–12 (2011).

53. Alam, K. K., Chang, J. L. & Burke, D. H. FASTAptamer: A Bioinformatic Toolkit for High-throughput Sequence Analysis of Combinatorial Selections. *Mol. Ther. - Nucleic Acids* **4,** e230 (2015).

54. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14,** 1188–1190 (2004).

55. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31,** 3406–3415 (2003).

Chapter 3 Side-chain determinants of biopolymer function during selection and replication

This chapter has been adapted from:

Lichtor P. A., Chen Z., Elowe N. H., Chen J. C., Liu D. R. Side chain determinants of biopolymer function during selection and replication. *Nature Chemical Biology* 2019, **262**(4)**:** 1--426.

Contributions:

Dr. Phillip Lichtor and Professor David Liu led the project. I contributed to characterization experiments and data analysis. I thank Phillip for his guidance and patience, teaching me about the HFNAP system, SPR, HTS, and data analysis.

3.1 Introduction

The chemical structures of nucleic acid and protein building blocks determine the properties of biopolymers and thus many of the functions of living systems. What chemical functionalities are most likely to enable the evolution of polymers with activities essential to living systems, such as the ability to bind other molecules, or to catalyze chemical reactions? Insights into this fundamental question would enhance our understanding of life's molecular requirements, and inform our ability to navigate chemical space to discover non-natural polymers with useful properties. Many studies that have probed these questions have searched for function using limited sets of building blocks from naturally derived monomers of proteins[1-5] and nucleic acids[6-9], such as using just three DNA bases instead of four, or using a subset of amino acids out of the canonical 20. While these studies have shown the adequacy or limitations of smaller subsets of natural building blocks, the fundamental relationship between the availability of diverse combinations of many different side-chains in a biopolymer's genetic code and the evolutionary potential of the resulting polymers has not been experimentally illuminated.

Nucleic acids are well-suited biopolymers for exploring these questions because functional members can be rapidly enriched from large, diverse libraries of sequences using *in vitro* selections[10-12], iterated cycles of Darwinian selection and replication. Using *in vitro* selections and related processes, researchers have previously discovered and isolated many DNA-based receptors[13-15] and catalysts[16-21]. Researchers have also performed *in vitro* selections using side-chain-modified DNA, and have demonstrated that the replacement of one or two nucleotides with side-chain-appended nucleotides can enable target binding[22-24] or catalysis[25-28] that may be more difficult without the additional side-chain or pair of side-chains[29-31]. One hypothesis emerging from these studies is that replacing all cytosines (Cs) or thymines (Ts) with nucleotides containing aromatic side-chains may lead to polymer libraries with better affinity to protein targets[31,32]. To illuminate the relationship between side-chain access and

biopolymer functional potential more comprehensively, however, requires generating and comparing polymer libraries with access to larger sets of multiple, diverse side-chains, beyond one or two types of chemical modification in an otherwise unmodified nucleic acid.

We and others previously described an artificial translation system that uses DNA ligase to mediate the DNA-templated synthesis of long, sequence-defined highly functionalized nucleic acid polymers (HFNAPs)[33-36]5. In our system, each HFNAP building block contains a YNN trinucleotide backbone (where Y = T or C, and N = A, T, C, or G) that is covalently linked to one of up to 32 chemically diverse side-chains on the 5' pyrimidine base. In recent applications of the HFNAP system[34], eight different side-chains are encoded by 32 trinucleotide codons, such that four trinucleotide sequences provide access to each side-chain group. HFNAP libraries in which each polymer contains 15 consecutive building blocks can be generated in this system from corresponding libraries of DNA templates (**Figure 3.1**)[33,34]. The resulting libraries contain sequence-defined 45-mers with a theoretical library diversity of ~4 x 10^{22} different polymers with a variable sequence region of ~15 kD. These libraries can be selected in vitro for target binding, and surviving polymers can be "reverse translated" back into DNA using Q5 polymerase and PCR, resulting in DNA templates suitable for a subsequent round of translation and selection[34]. We recently used this system to discover HFNAPs that bind two disparate proteins of biomedical interest, PCSK9 and IL-6, with high affinity (K_d ~1-10 nM)[34].

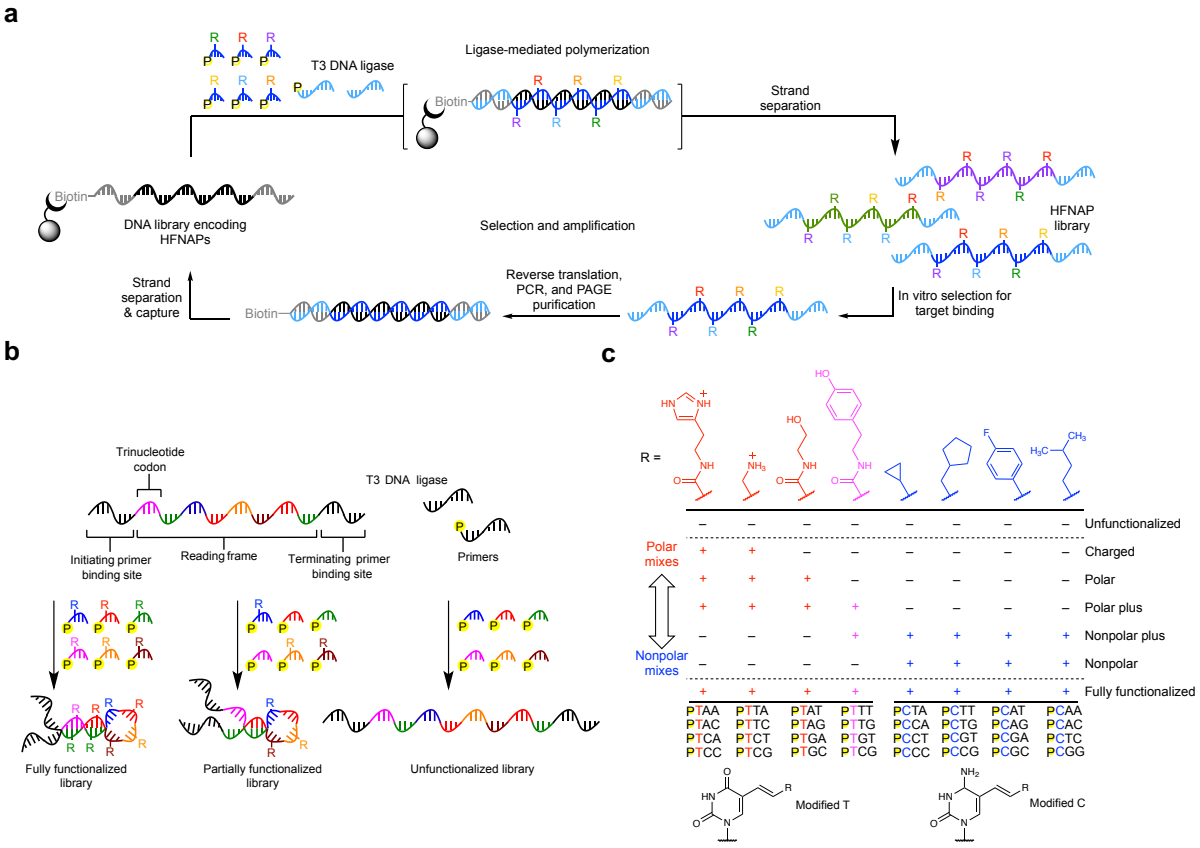


Figure 3.1 | HFNAP translation, selection, and library design. a, Overview of a complete cycle of HFNAP translation from DNA templates, selection for target affinity, reverse translation back to DNA templates, and template replication by PCR. **b**, Pool design showing how combinations of monomers are combined to make fully side-chain functionalized, partially functionalized, or unfunctionalized libraries. **c**, The seven sets of side-chain availability in each "genetic code" (top) and the 32 possible trinucleotide codons for each HFNAP.

Since researchers can choose the chemical structure of the side-chain in every building block, the HFNAP system provides an ideal opportunity to experimentally test how the side-chain repertoire in a sequence-defined polymer library containing many different side-chains influences the functional potential of the resulting polymers. Here we performed in parallel iterated rounds of translation, selection, and replication of HFNAPs for binding to PCSK9 or IL-6 using seven different "genetic codes", each of which allow polymers to access a different subset of side-chain functionalities. Using several metrics to evaluate both bulk populations of HFNAPs and individual polymers post-selection, we observed that access to hydrophobic side-chains strongly promoted the emergence of polymers that bind with high affinity to their protein targets

for both selections. Our findings illuminate the relationship between side-chain composition and polymer binding activities, and experimentally demonstrate how the availability of nonpolar side-chains could provide biopolymers such as proteins with evolutionary advantages over biopolymers such as natural DNA and RNA that lack such groups.

3.2 Parallel selections with seven different genetic codes

The HFNAP translation system (**Figure 3.1a**) uses a DNA template containing a 45-base coding region of 15 consecutive codons flanked by two primer binding sites. Trinucleotide building blocks hybridize to complementary codons on the template[33,34], and T3 DNA ligase catalyzes the covalent coupling of adjacent building blocks on a template in sequence-defined order[37]. From the resulting heteroduplex HFNAP:DNA double-stranded product, the desired polymer strands are isolated through alkaline denaturation (**Figure 3.1a**).

To study how side-chain composition in a polymer library influences selection outcomes, we prepared seven HFNAP libraries of random sequence from aliquots of the same DNA template library containing 15 consecutive NNR codons (where N = A, T, C, or G; and R = A or G). Each of the seven libraries differed in the presence or absence of different side-chains during translation (**Figure 3.1b**). The library templates were drawn from the same random sequence mixture, although statistically each library shares a negligible fraction of identical template sequences with the other six libraries. For libraries designated to omit specific side-chains, we replaced the building blocks that would normally contain that side-chain with sequence-matched unfunctionalized DNA trinucleotide building blocks lacking any side-chain.

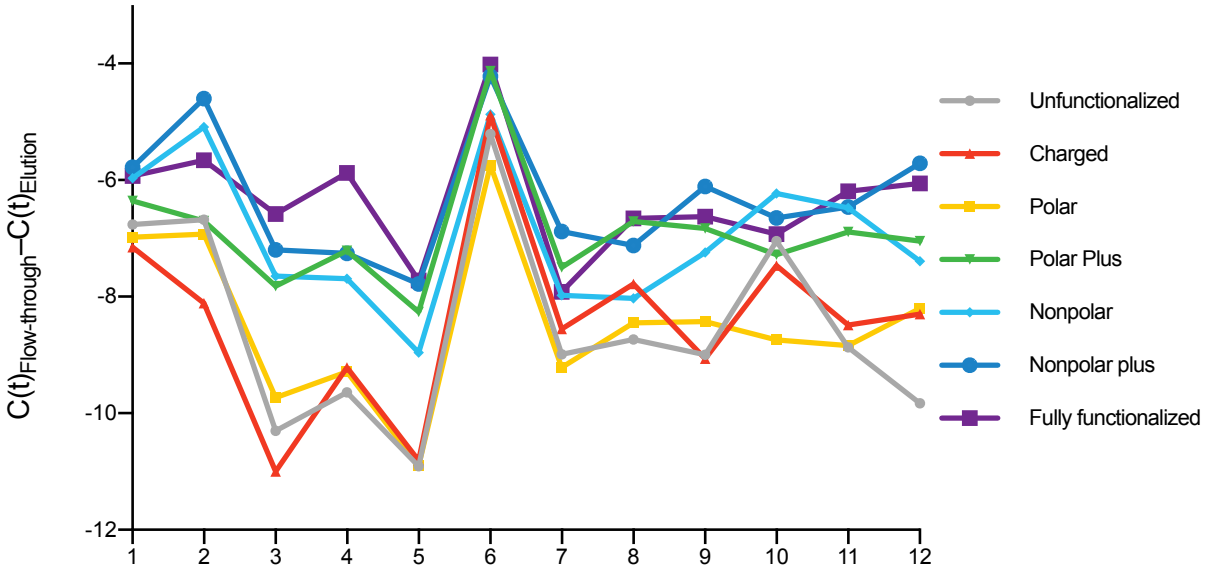
The seven different “genetic codes” included a fully functionalized code in which all eight side-chains including amine, alcohol, aliphatic, fluorinated, and aryl groups were introduced, and a completely unfunctionalized code, in which all building blocks were unfunctionalized DNA trinucleotides lacking side-chains (**Figure 3.1c**). In addition, we used five genetic codes in which different subsets of side-chains were included based on their polarity. The “charged” code

included building blocks with primary amine or imidazole side-chains that can be protonated in equilibrium with their neutral, non-protonated forms. The positively charged side-chains complement the negatively charged backbone of DNA. The “polar” code added an alcohol side-chain along with the charged side-chains. The “nonpolar” code contained only hydrocarbon and fluorocarbon side-chains, including cyclopropyl, cyclopentyl, p-fluorophenyl, and isopentyl side-chains. Since the phenolic side-chain possesses both polar and nonpolar characteristics[38] and is frequently involved in protein-protein interactions[39,40], we investigated how the addition of this side-chain influenced selection outcomes using two additional codes in which the phenolic side-chain was added to either the polar code (resulting in the “polar plus” code), or to the nonpolar code (resulting in the “nonpolar plus” code).

Following DNA-templated polymerization (translation) of each of the seven polymer libraries, we performed on each library iterated cycles of target binding selection, “reverse translation” from HFNAP back to DNA, PCR replication of surviving DNA templates, and re-translation into HFNAPs to enrich the population of HFNAP binders[34]. Briefly, translated polymers were incubated with the agarose bead-immobilized protein target. The non-binding flow-through was removed and the beads were washed three times with buffer. Bound polymers were eluted by heating in buffered detergent, then reverse translated and amplified by PCR with Q5 DNA polymerase. Full-length amplicons were purified by PAGE, and the template DNA strands were isolated by streptavidin capture and alkaline denaturation to enter the next round of HFNAP library translation. The sequence composition of each of the seven parallel populations was analyzed after each PCR step by high-throughput DNA sequencing (HTS).

We chose PCSK9 and interleukin 6 (IL-6), two disparate proteins of biomedical interest[41-43] for which we[34] and others[31,44,45] have previously isolated nucleic acid receptors, as the in vitro selection targets. In the first selection campaign, we examined how the seven libraries matured when iteratively challenged to bind PCSK9. We translated 7.5×10^{13} DNA template molecules (125 pmol) for each of the seven genetic codes and challenged the resulting polymer libraries to

bind PCSK9 using the selection method described above. After each round of selection, the quantity of the HFNAP pool retained by the PCSK9-linked beads was assessed by qPCR (**Figure 3.2**). The stringency of selections was increased gradually as the rounds progressed by decreasing the amount of target PCSK9 protein provided, or by decreasing the incubation time. All seven HFNAP libraries were subjected to the same selection conditions throughout all rounds of selection (**Figure 3.2**). Starting in round 7, we added a counter-selection at the end of each round to remove polymers that bound beads lacking PCSK9.



Round	1	2	3	4	5	6	7	8	9	10	11	12
Template (pmol)	125	1	1	1	1	1	1	1	1	1	1	1
PCSK9 (pmol)	125	13	3.7	3.7	3.7	1	1	1	1	1	1	1
Neg Blank beads (µL)	-	-	-	-	-	-	1.4	4	6	6	6	8
Time (h)	1	1	1	1	1	1	1	1	1	0.75	0.5	0.5

Figure 3.2 | PCSK9 selection conditions and progress. Population-level retention of each library on PCSK9-linked beads over seven rounds of selection, as measured by qPCR. Higher values indicate a greater proportion of amplifiable material eluted from the PCSK9-bound beads relative to the unbound material in the flow-through. The table indicates the selection conditions used at each round.

Over 12 rounds of iterated selection, reverse translation, template replication, and translation, qPCR analysis revealed that the polymer pools with access to the nonpolar or phenolic side-chains became enriched in PCSK9-binding polymers earlier in the iterated selection process, and to a greater extent, than the pools with access only to the unfunctionalized or polar side-chains (**Figure 3.2**). Using the templates from the final round of selection, we prepared 3'-biotinylated polymer pools and assessed the relative performance of these HFNAP populations by surface plasmon resonance (SPR), a direct binding assay that is distinct from bead retention. We examined the SPR response to PCSK9 using surfaces coated with each polymer pool and normalized the observed response by the maximum response expected from a 1:1 binding interaction (R_{max}) calculated from each polymer pool's immobilization level. Consistent with the qPCR data, the enriched pools with access to the full side-chain set and those containing nonpolar or the phenolic side-chains demonstrated the highest levels of polymer binding (%R_{max}) to PCSK9 (**Figure 3.3a** and **Supplementary Figure 3.1**). Naïve libraries, prepared and assayed similarly, demonstrated much lower binding responses (**Figure 3.3a**), suggesting that stronger binding to PCSK9 was a property enriched during the selection process. Together, these results suggest that the polymer libraries with highest side-chain diversity or access to nonpolar side-chains more strongly support the enrichment of PCSK9-binding HFNAPs.

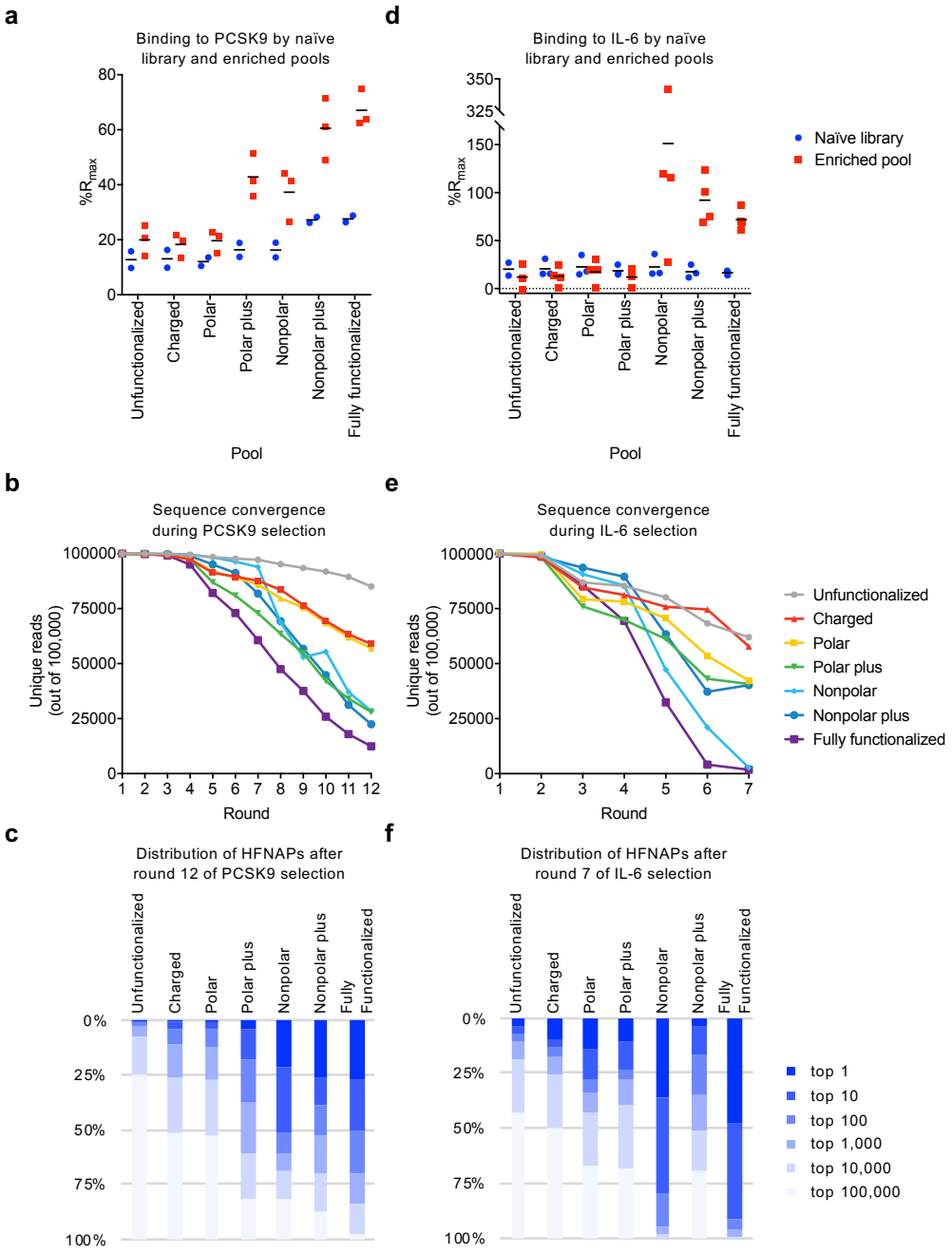


Figure 3.3 | Protein target binding activity, sequence convergence, and sequence distribution of HFNAPs selected for PCSK9 or IL-6 affinity. a, Normalized SPR response ($\%R_{max}$) observed when immobilizing biotinylated naïve libraries or enriched pools on an SPR chip and flowing PCSK9. **b**, Side-chain-dependent convergence of libraries in PCSK9 selections as determined by high-throughput sequencing (HTS). **c**, Polymer density in the final round of PCSK9 selection represented as the fraction of 100,000 HTS reads captured by the top 1, 10, 100, etc. HFNAPs. **d**, **e**, and **f**, analogous data to **a**, **b**, and **c** for the IL-6 selections. In a and d, horizontal bars denote mean values of three or more replicates; the value from each replicate is shown.

To further characterize the progress of each HFNAP population as it matured through the iterated translation, selection, and replication process, we subjected DNA templates emerging from each round of selection to HTS. The data revealed that the two populations containing mostly nonpolar side-chains (the nonpolar and nonpolar plus pools), the polar plus pool, as well the population with access to the full set of side-chains, converged much more quickly to a smaller number of unique HFNAPs compared to the populations having access to mostly polar side-chains (**Figure 3.3b**). Moreover, the most abundant HFNAPs from the pools with access to the nonpolar side-chains represented a higher percentage of the total pool than in the pools lacking nonpolar side-chains, suggesting that these strongly enriched polymers were indeed of high fitness in these populations (**Figure 3.3c**). For example, by the end of round 12, the ten most abundant HFNAPs from the fully functionalized and nonpolar plus pools, represented 50% and 39%, respectively, of all HFNAPs in the pool. In contrast, in the unfunctionalized DNA pool and the pool accessing only polar side-chains, the most abundant 10 HFNAPs represented only 0.72% and 3.7% of the total population, respectively (**Figure 3.3c**).

To illuminate the relationship between side-chain abundance in a polymer and enrichment during protein-binding selection, we examined the prevalence of each type of side-chain in the post-selection sequences. The most enriched sequences from the nonpolar pool were not predominantly composed of building blocks containing side-chains (**Supplementary Table 3.1**), even though the starting library contained 50% side-chain-containing building blocks, further suggesting that the benefit conferred by the availability of nonpolar side-chains is not simply to maximize the abundance of groups that participate in non-specific hydrophobic interactions with the target protein.

Next, we sought to validate and compare the binding of individual polymers identified during the selection. We synthesized biotinylated versions of the most abundant polymers at the end of the selection and also those HFNAPs undergoing the greatest enrichment during any stage of the selection. We subjected the resulting polymers to PCSK9-binding analysis by SPR

using a streptavidin-linked chip and a buffer similar to that used in selection (**Figure 3.4a**, **Supplementary Figure 3.2**, and **Supplementary Table 3.2**). Most of the individual polymers tested exhibited detectable binding to PCSK9, consistent with apparent bulk library affinity (**Figure 3.3a**), with K_d values as low as 11±14 nM (**Figure 3.4b**), consistent with our previous studies[34]. Some polymers did not demonstrate simple 1:1 binding kinetics to PCSK9, complicating the determination of K_d values and polymer comparison. We therefore injected a fixed concentration of PCSK9 onto surfaces coated with each polymer and normalized the observed response to R_{max}, calculated from each polymer's immobilization level. Individual HFNAPs from the polar plus, nonpolar plus, and fully functionalized pools generally exhibited higher normalized SPR response (%R_{max}) than the less functionalized and more polar pools (**Figure 3.4a**). Collectively these data on individual polymer abundance, enrichment, and binding characteristics suggest that individual polymers from the best-performing libraries indeed exhibit greater overall PCSK9 binding activity.

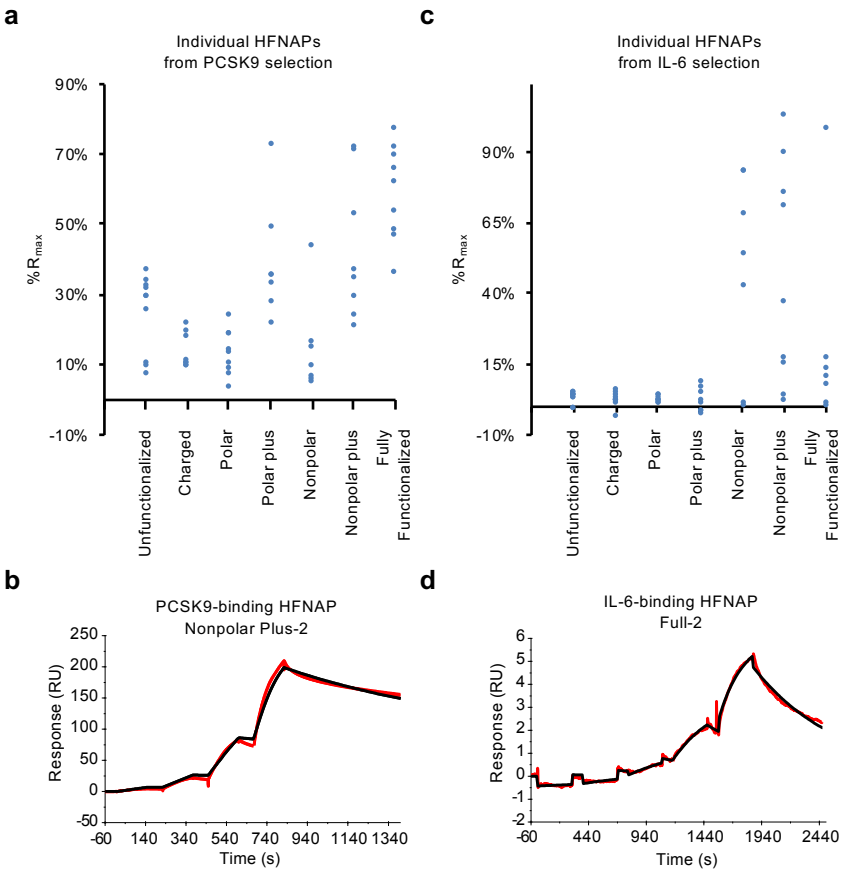


Figure 3.4 | Target-binding activity of individual HFNAPs by SPR. a, Observed $\%R_{max}$ for individual polymers emerging from iterated cycles of translation and selection for binding to PCSK9, calculated from SPR response following a single injection of 60 nM PCSK9. **b**, Background-subtracted sensorgram (red) of PCSK9-binding HFNAP Nonpolar Plus-2 binding to 10, 30, 100, and 300 nM PCSK9 and curve fitting with a 1:1 binding model (black) results in an apparent K_d of 28 nM (shown); the average among three replicates is K_d = 11±14 nM (see Methods). **c**, Observed $\%R_{max}$ for individual polymers emerging from iterated cycles of translation and selection for binding to IL-6, calculated from SPR response following the third of four injections of IL-6 in a single-cycle kinetics assay (8.9, 27, 80, and 240 nM IL-6; see Methods for details). **d**, Background-subtracted sensorgram (red) of IL-6-binding polymer Full-2 binding to 0.3, 1, 3, 10, and 30 nM IL-6 and fitting with a 1:1 binding model (black) results in an apparent K_d of 12 nM (shown); the average among three replicates is K_d = 5.3±5.9 nM, although we note that this polymer exhibits complex binding behavior at concentrations above 30 nM. K_d values represent mean ± standard deviation (n = 3).

3.3 Head-to-head binding competition with seven different genetic codes

To gain additional insight into the functional potential of the different pools that emerged after 12 rounds of selection, we performed a single-solution, head-to-head affinity competition. Such a competition offers a direct comparison of polymer library performance as a function of

side-chain availability. Polymers emerging from each of the seven pools' final round of selection for PCSK9 binding were mixed together in equimolar ratio (~160 fmol of each post-round 12 polymer library). The resulting mixture was subjected to an additional selection for binding to PCSK9 (**Figure 3.5a**). PCSK9-bound polymers were eluted, reverse-translated, sequenced by HTS, and assigned back to one of the seven pools based on alignment to sequences from previous rounds. The vast majority (94%) of the polymers surviving the head-to-head selection could be unambiguously assigned to their original pools (**Figure 3.5b**). The changes in relative abundance of each pool's polymers when combined and reselected for PCSK9 binding reflected the relative fitness of each population's polymers in the competition.

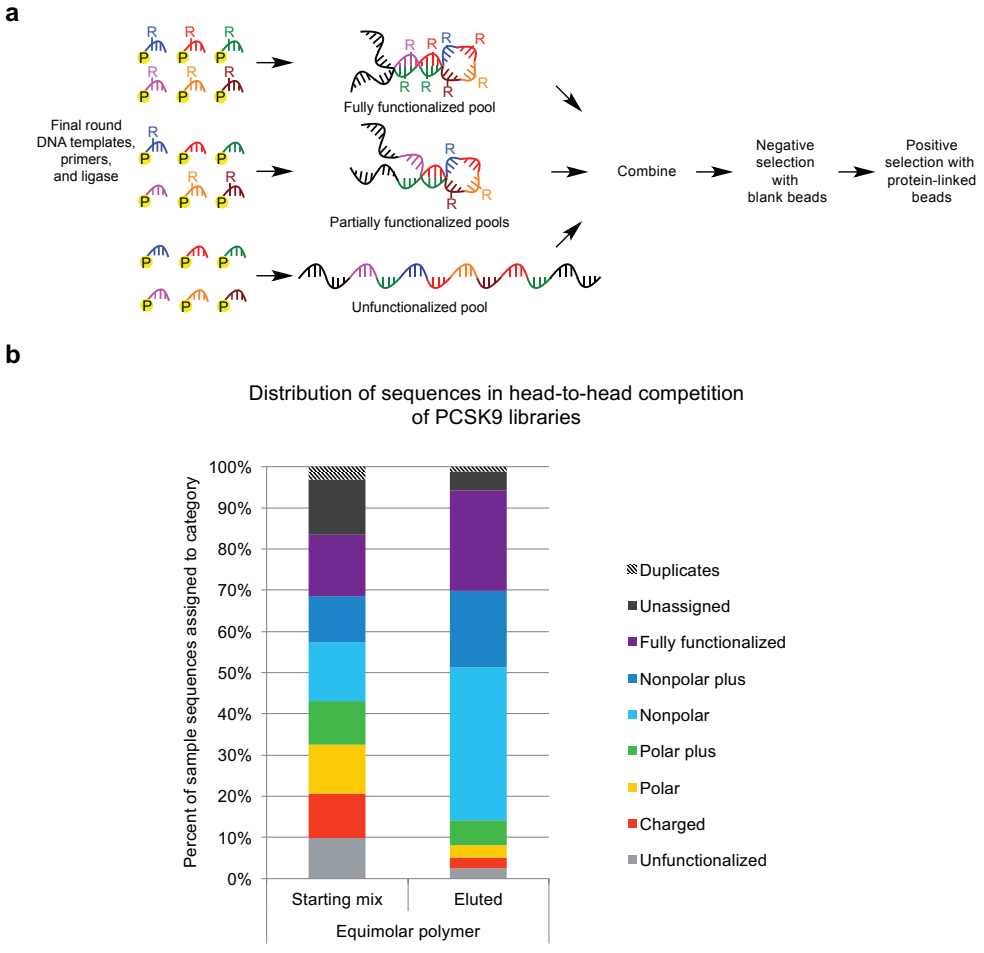


Figure 3.5 | Analysis of mixed samples in polymer pool competition. **a**, Schematic representation of head-to-head competition. **b**, Stacked bar plot showing the distribution of sequences from the starting and eluted material from the head-to-head competition experiment in which a round of selection against bead-bound PCSK9 was performed. Sequences are

Figure 3.5 (Continued)
assigned to a pool of origin based upon the occurrence of the same sequence in any of the 12 previous selection rounds. Sequences that were not previously found from HTS analysis of previous rounds are labeled as unassigned. Sequences appearing in multiple pools where assignment was ambiguous were removed, but labelled as duplicates, so bars add to 100%.

Consistent with the above findings, the ability of each polymer pool to survive the head-to-head PCSK9 binding competition correlated with the availability of nonpolar side-chains. The HFNAPs from the fully functionalized, nonpolar plus, and nonpolar populations enriched at the expense of polymers from the unfunctionalized, charged, polar, and polar plus populations (**Figure 3.6a**, **b**). This conclusion was not dependent on how the 5.7% of unassigned sequences were categorized, as the unfunctionalized, charged, and polar pools were depleted to a much larger extent than could be accounted for by those unassigned sequences (**Figure 3.5b**). On an individual polymer level, we also observed clear advantages for polymers that accessed nonpolar side-chains. For example, all of the 100 assignable polymers undergoing the greatest enrichment from the head-to-head competition were from the three pools (nonpolar, nonpolar plus, and fully functionalized) containing mostly nonpolar side-chains (**Figure 3.6b** and **Figure 3.7**). Importantly, the simple number of side-chains present in each polymer was not sufficient to explain these performance differences, as those 100 most strongly enriched polymers averaged only 7.4 out of 15 possible side-chains per polymer. Together, the results of this head-to-head competition experiment further support the finding that the populations containing nonpolar side-chains yielded polymers that outperform populations lacking nonpolar side-chains in a target protein-binding task.

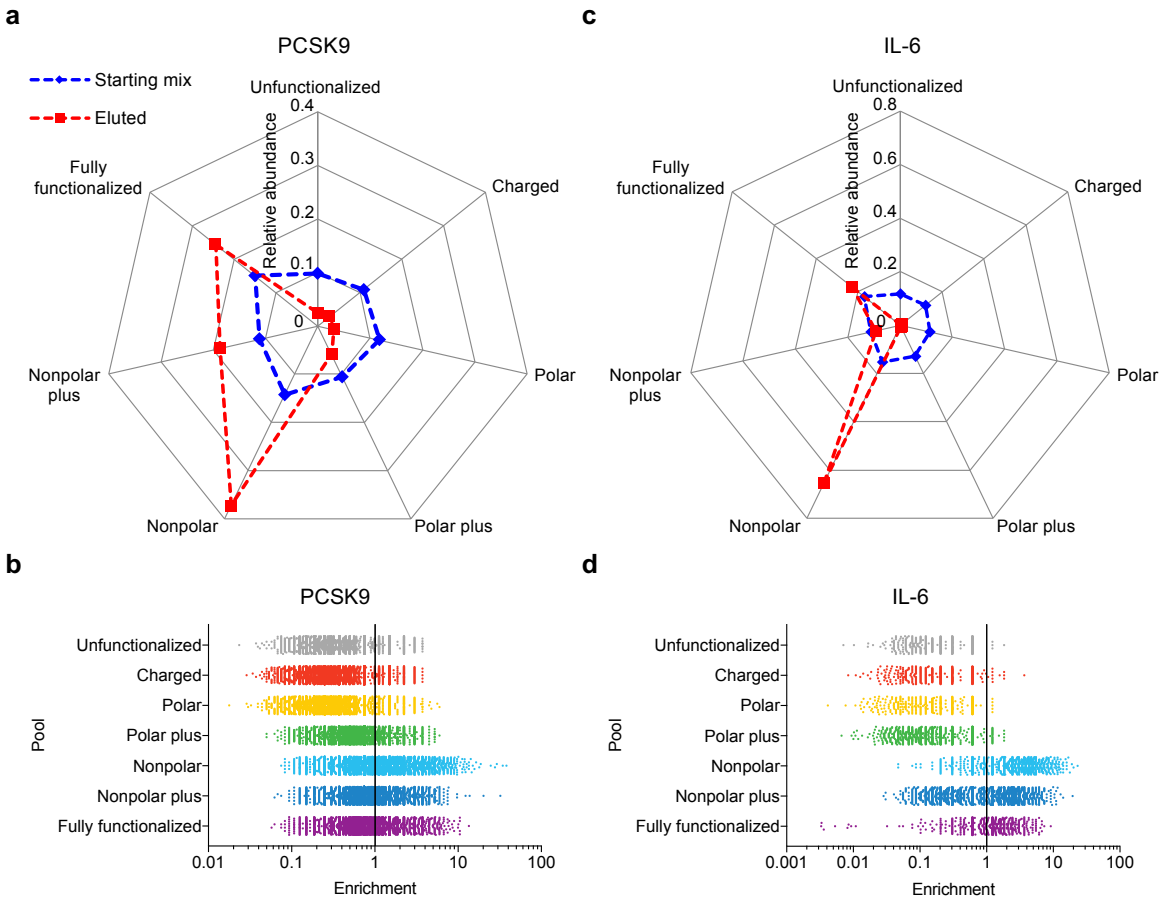


Figure 3.6 | Head-to-head competition of seven pools following iterated selections for binding to PCSK9 or IL-6. a, Plot showing the distribution of assigned polymers before head-to-head competition selection (blue) and after head-to-head selection (red) for PCSK9 binding. The values represent relative abundances of HFNAPs from each pool (total abundance = 1.0, including unassigned; see **Figure 3.5** for a detailed breakdown). **b**, The distribution of individual polymer enrichment values is shown for each pool in the PCSK9 competition experiment. Enrichment for each polymer was determined from the ratio of its abundance in the eluted material divided by its abundance in the initial combined pool. **c**, Distribution of assigned polymers before head-to-head competition selection (blue) and after head-to-head selection (red) for IL-6 binding. **d**, Distribution of individual polymer enrichment values for each pool in the IL-6 competition experiment.

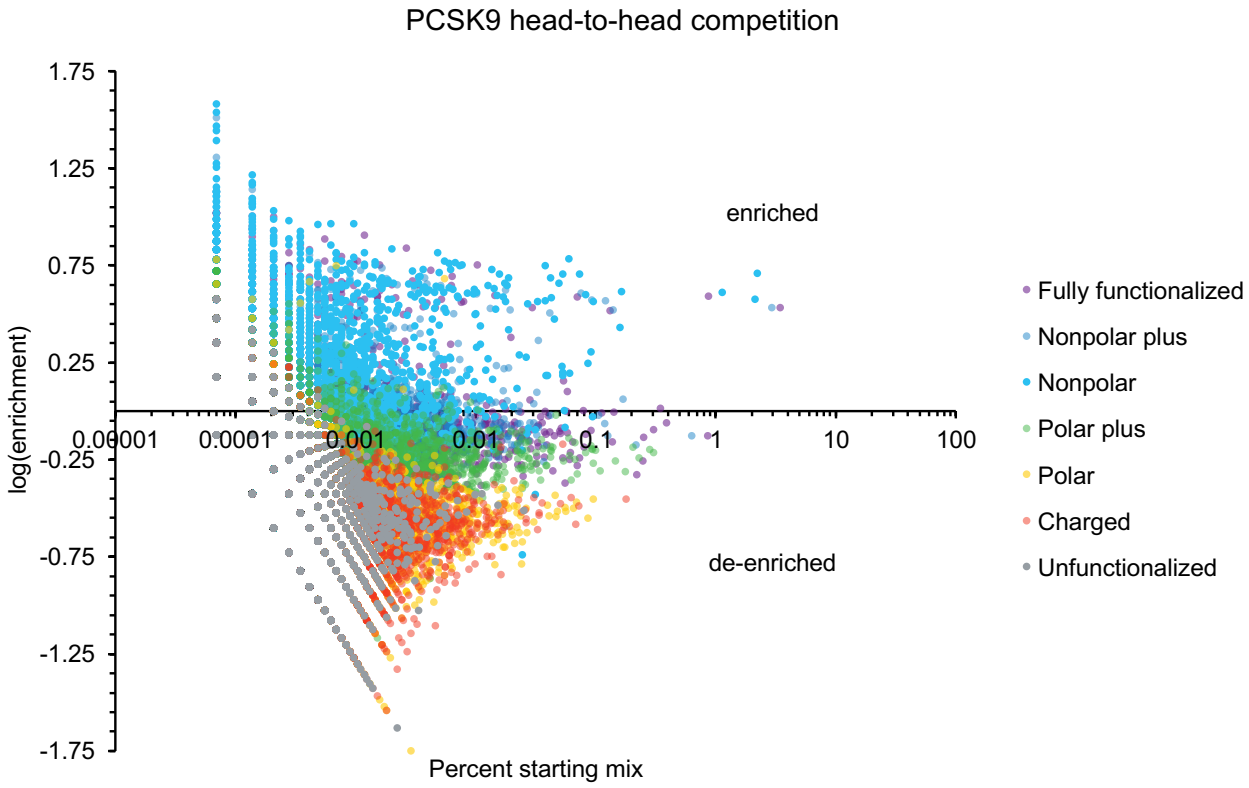


Figure 3.7 | Individual polymer enrichment in PCSK9 competition experiment. Each point represents a sequence that could be assigned to a pool and was identified in both the starting and eluted mixes.

3.4 Examining translation bias in the HFNAP system

To examine if the trends emerging from the enriched pools following the PCSK9 selections were due to target binding and not artifacts of other steps in the selection cycle, we also examined potential bias in the HFNAP translation, reverse translation and amplification steps of the selection process starting with aliquots of naïve template libraries used in the PCSK9 selection. We found that the overall translation and amplification efficiency of libraries containing side-chain functionalized building blocks is lower, rather than higher, compared to the translation and amplification efficiency of unfunctionalized DNA, up to a factor of approximately six for the fully functionalized library (**Supplementary Figure 3.3**). Given that the fully functionalized library was among the best-performing polymer libraries, this observation

suggests that the advantage of functionalization with side-chains overcomes any overall bias against incorporating side-chains in the polymer translation and amplification process.

To examine the potential bias of the system toward individual building blocks, we performed a neutral selection[46] in which we collected HTS data after two cycles of translation, dilution (without selection for target binding), reverse translation, and PCR for each genetic code set. Analysis of the distribution of all 32 codons revealed that over the course of two neutral selection rounds, none of the building blocks enriched more than approximately 2-fold or de-enriched more than approximately 4-fold (**Figure 3.8**). Importantly, the most prevalent building blocks in the most enriched pools from the PCSK9 selection were different from the most prevalent building blocks in the neutral selection (**Figure 3.8**). Moreover, none of the building blocks were completely eliminated during the course of any of the selections described above. Taken together, this analysis suggests that the changes to the HFNAP populations during target-binding selections are not primarily due to codon or side-chain biases separate from those that contribute to binding activity.

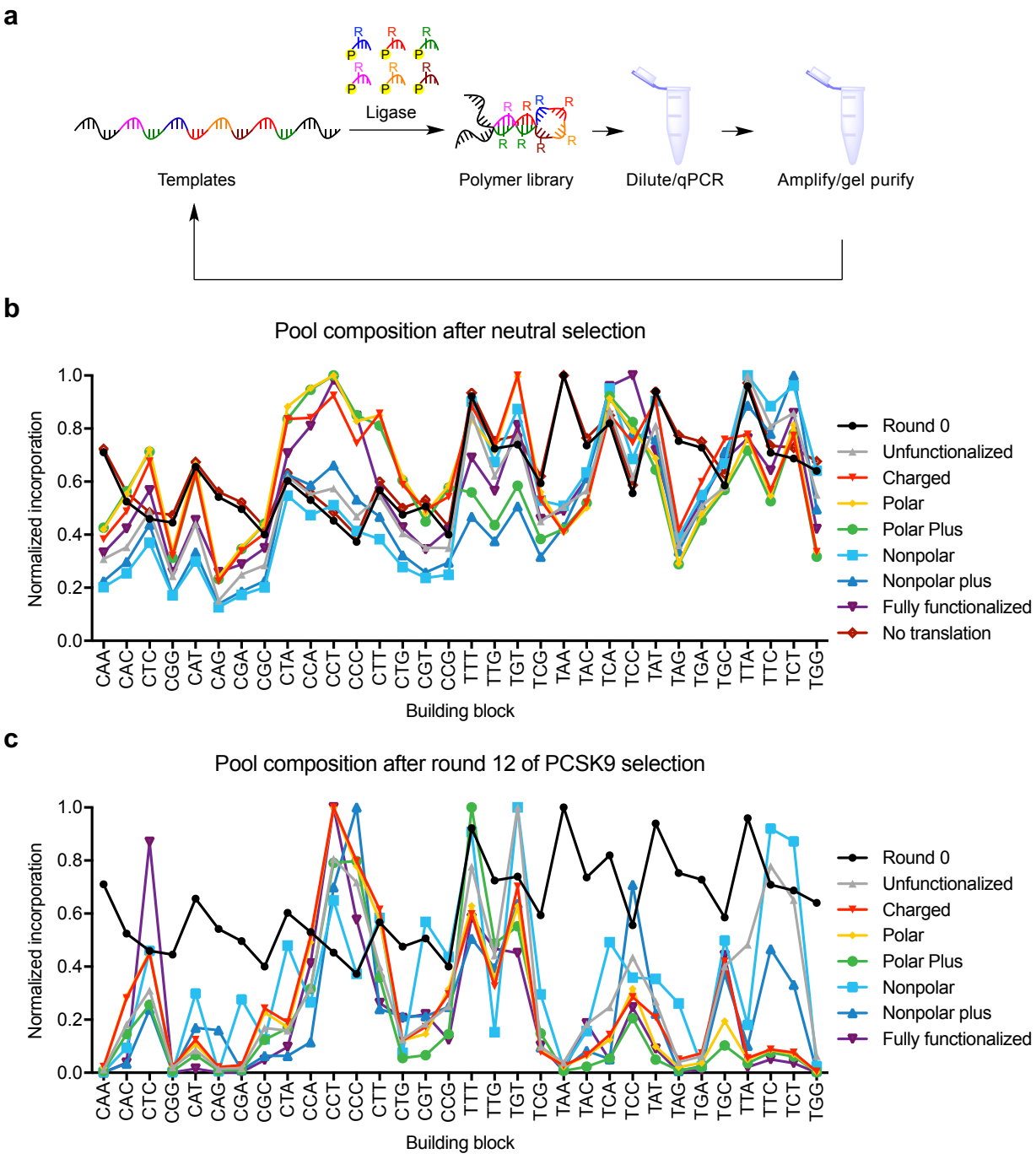
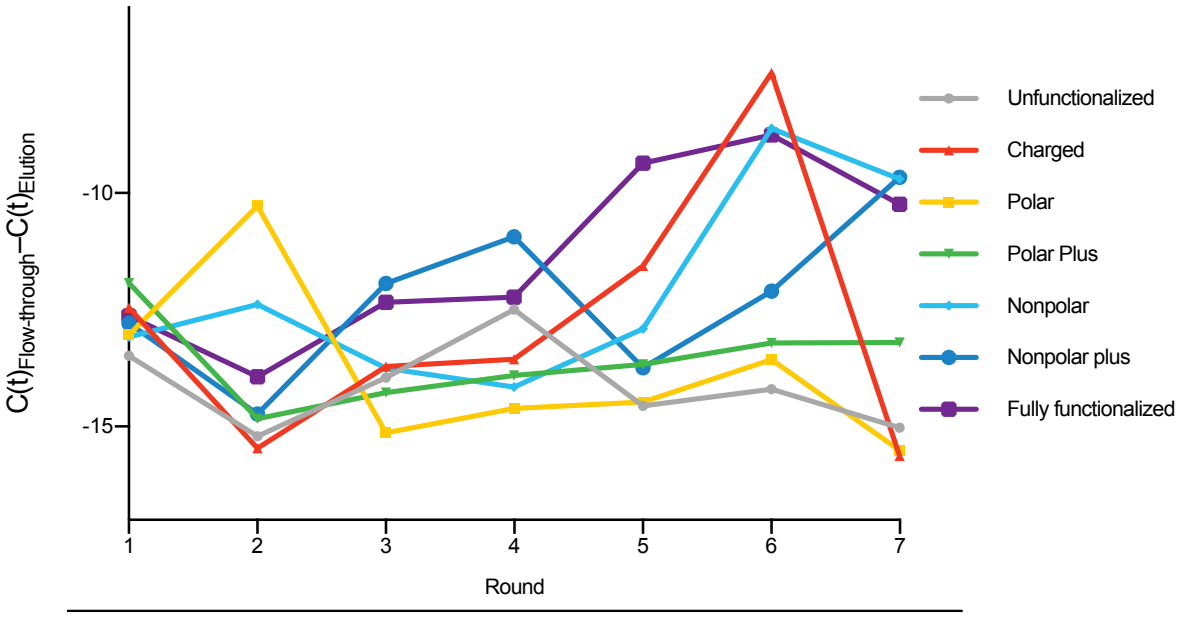


Figure 3.8 | Changes in codon distribution. a, Schematic representation of a "neutral selection" with a cycle of translation, reverse translation, and amplification. The seven different pools underwent two rounds of neutral selection and were interrogated by HTS. **b**, Plot of the distribution of the codons following blank selection with each of the different building-block sets (pools). Each pool's data is normalized to the most abundant codon in each library. The "no translation" sample shows data from two rounds of just dilution and amplification. **c**, Plot of the distribution of the codons following round 12 of the PCSK9-targeted selection.

3.5 IL-6 selection outcomes

To test if the above findings also apply to protein targets other than PCSK9, we repeated the HFNAP library translation and iterated selections, as well as all other experiments with enriched polymers described above, with IL-6, an unrelated protein involved in inflammation[47]. Over seven rounds of translation, selection, and replication, qPCR analysis once again suggested that the HFNAP libraries with access to the nonpolar side-chains generally exhibited superior binding to immobilized IL-6 than libraries lacking access to nonpolar side-chains (**Figure 3.9**). With substantial differences already evident among the enriched polymer pools, we stopped IL-6 selections after round 7. Following the seventh selection round, we immobilized the naïve polymer libraries and the enriched polymer pools to an SPR chip as described above to study their binding to IL-6 in a distinct assay modality. Similar to our findings for PCSK9 binding, all of the enriched polymer pools with access to the nonpolar set of side-chains (nonpolar, nonpolar plus, and the fully functionalized set) exhibited higher SPR responses to IL-6, suggesting greater binding activity (**Figure 3.3d**). These experiments further support that the bulk binding affinity of the enriched polymer libraries with nonpolar side-chains to IL-6 is greater than those without non-polar side-chains, consistent with the results from PCSK9-binding selections.



Template (pmol)	125	5	1	2	2	2	2
IL-6 (pmol)	125	1	1	2	2	2	2
Neg Blank beads (µL)	-	-	2	4	4	4	4
Time (h)	1	1	1	1	1	0.83	0.66

Figure 3.9 | IL-6 selection conditions and progress. Population-level retention of each library on IL-6-linked beads over seven rounds of selection, as measured by qPCR. Higher values indicate a greater proportion of amplifiable material eluted from the IL-6-bound beads relative to the unbound material in the flow-through. The table indicates the selection conditions used at each round.

HTS data indicated that all of the libraries in the IL-6 selection campaign converged more quickly than those in the PCSK9 selection campaign (**Figure 3.3e**). Nonetheless, during IL-6 binding selections the nonpolar and fully functionalized populations converged more quickly to a smaller number of HFNAPs than the other libraries, again suggesting the benefit of nonpolar side-chains to HFNAP enrichment in IL-6 binding. Moreover, the 10 most abundant HFNAPs from the fully functionalized and nonpolar pools following the seventh selection round represented 91% and 80%, respectively, of all HFNAPs in the pool (**Figure 3.3f**). The other side-chain-containing libraries (charged, polar, polar plus, and nonpolar plus) enriched much

less strongly, with the ten most abundant HFNAPs representing only 13-28% of the total population of each pool.

We performed a head-to-head IL-6-binding competition of an equimolar mixture of the IL-6 post-selection polymer pools. The nonpolar pool outperformed all others, while the fully functionalized and nonpolar plus pools approximately maintained their relative abundance before and after competitive selection (**Figure 3.6c**, **d** and **Supplementary Figures 3.4-3.5**). In contrast, the pools without side-chains or containing only polar side-chains were depleted in the IL-6 binding competition. As 99.4% of the sequences were assignable based on their similarity to sequences observed during selections on isolated pools, these results did not depend on the 0.6% unassigned sequences. In addition, all of the top 100 most enriched polymers from the head-to-head competition originated from pools with access to nonpolar side-chains. Furthermore, among this top set of polymers, the average number of side-chains present was 6.2 out of 15 possible, suggesting that the total number of nonpolar side-chains is not a primary determinant of IL-6 binding. Collectively, these results are consistent with the findings from the PCSK9 selections.

We prepared individual biotinylated polymers that exhibited the greatest enrichment during the IL-6 selection process for analysis, and immobilized them on the surface of a streptavidin-linked SPR chip to assess their binding to varying concentrations of IL-6. Individual polymers derived from the nonpolar, nonpolar plus, and fully functionalized pools exhibited substantially higher $\%R_{max}$ values than polymers from the unfunctionalized and exclusively polar side-chain sets (**Figure 3.4c**, **Supplementary Figure 3.7**, and **Supplementary Table 3.3**). Individual K_d values were as low as 5.3±5.9 nM (**Figure 3.4d**). These data confirm that individual polymers identified from the selections indeed bind IL-6 with high affinity.

3.6 Discussion

This work demonstrates that in both separate and competitive Darwinian selections for binding to two unrelated proteins, HFNAP libraries with access to nonpolar side-chains outperformed those without access to nonpolar side-chains. Despite some differences between the outcomes of PCSK9 and IL-6 selections, which may reflect differences in the rarity of polymers that bind to the target proteins, in selection stringency or other parameters during selection, or other factors that influence HFNAP fitness, the overall trends in polymer pool performance between the two selection campaigns were remarkably similar. In both selections, the libraries with access to nonpolar side-chains generally converged more quickly, yielded polymers with higher target protein binding activity, and competed more effectively compared with populations with access to only polar side-chains, or to those with no side-chains. Furthermore, the enriched pools and individual polymers containing nonpolar side-chains demonstrated higher apparent binding activity to their protein targets in an orthogonal SPR assay when compared to those from other pools.

These findings mirror observations during the development of some bioactive small-molecule probes and therapeutics[48], in which the addition of lipophilic groups is known to contribute to target-binding activity. While the addition of too many nonpolar groups in small-molecule drugs is known to lead to promiscuous protein binding[49], the data from our study suggest that nonpolar side-chains among the HFNAPs characterized in this work do not simply serve to increase non-specific protein affinity. We previously found that the specific structure, location, and sequence context of key nonpolar side-chains among PCSK9-binding and IL-6-binding HFNAPs, and not simply their abundance, were critical determinants of binding activity[34]. Indeed, our analysis of the top sequences from the head-to-head competition in this study indicate that the number of non-polar side-chains among top-performing sequences did not approach the maximum number allowed, further suggesting that the position of the nonpolar side-chains in the context of the surrounding building clocks, rather than the absolute number of

nonpolar side-chains, contributes strongly to target-binding activity. Consistent with these conclusions, the nonpolar side-chains in these DNA-based polymers represent only a minor fraction of groups among the extensive and fairly polar architecture of HFNAPs .

This study experimentally illuminates how nonpolar side-chains offer benefits when multiple side-chains are available, complementing previous findings that nonpolar groups in singly- or doubly-modified DNA-based polymers are beneficial[31] compared to unmodified DNA for generating protein-binding aptamers. It is tempting to speculate about the applicability of these findings to other protein targets and to other types of biopolymers. While the systems and experiments reported by others working with functionalized DNA are distinct from those in this work, our findings together with previous observations that DNA modified with nonpolar side chains has a higher propensity to provide functional receptors to a variety of protein targets[22,30] suggest that these principles may apply more broadly than to HFNAPs, PCSK9, and IL-6.

In addition to directly demonstrating how functional competition between classes of evolving polymers could have favored polymers with access to nonpolar side-chains similar to those found in proteins but absent from natural nucleic acids, these findings also inform future synthetic polymer evolution efforts. Of all known biopolymers, proteins arguably have access to the most diverse repertoire of chemical functionality—one that is further augmented by post-translational modifications. By providing exquisite control over the structure and position of diverse chemical groups including those unavailable to native biopolymers, the HFNAP translation system may eventually provide polymers with binding, catalytic, and other activities that complement—and perhaps even exceed—those in the impressive repertoire of natural proteins.

3.7 Methods

General Methods. Oligonucleotides were purchased from IDT, and are listed in the Supplementary Information. Monomer building blocks were synthesized as described in ref 34. T3 DNA ligase (3,000 U/μL), ATP, T4 RNA ligase buffer were purchased from New England Biolabs. PCSK9 (PC9-H5223) and IL-6 (IL6-H4218) were purchased from AcroBiosystems and reconstituted according to the manufacturers recommendations. Aliquots were flash frozen in liquid N_2 and stored at -80 ºC until ready to use. The proteins were prepared for SPR by dialyzing into the indicated buffer and concentration was determined by absorbance at 280 nm. Streptavidin-coated magnetic beads (Dynabeads MyOne C1) were purchased from Life Technologies. AminoLink Plus Micro Immobilization Kits (ThermoFisher Scientific) were used for protein immobilization and for blank beads. HBS-P+ was purchased from GE Healthcare Life Sciences.

Preparation of initial DNA libraries prior to selection. Starting template libraries were prepared through a primer extension of the appropriate naïve library (purchased as a hand mixed library from IDT; e.g., naïve library AZ) using a biotinylated primer (e.g., BtBt_EPrimerA, see Supporting Information). Binding the resulting dsDNA to magnetic beads, washing, and strand separation with 20 mM NaOH furnished the biotinylated complement after gel purification and cleanup with Omega Bio-tek HiBind DNA Midi columns.

DNA template libraries that were used to synthesize polymers in the start of the selections were from different aliquots from the same starting template library mix. Due to the high diversity of naïve libraries, a negligible fraction of template sequences in each of the pools are identical to each other.

General comments on library comparison studies. A 2',3'-dideoxy 3'-primer was used in the translation of the libraries to limit propagation of cheaters arising from incorporation of a primer-binding sequence in the polymer-coding region, as we previously described (ref. 34).

First round of HFNAP library synthesis. DNA template (125 pmol), initiation primer (188 pmol), termination primer (188 pmol), monomer mix containing 1.25 nmol each monomer, 12.5 µL 10x T4 RNA ligase buffer, and enough water to bring the volume to approximately 112 µL were added to tubes of a PCR strip, then mixed. The strips were placed in a thermal cycler and run with the following program: heat to 95°C for 10 seconds, cool to 65°C, hold at 65 ºC for 4 min, ramp down at 0.1 °C per 10 seconds to 4 °C. While still at 4 ºC, 6.25 µL of 10 mM ATP and 6.25 µL T3 DNA ligase (3,000 U/µL) were added to each reaction without removing the PCR strip from thermal cycler. The reactions were kept at 4 ºC for 12 h, then warmed to 16 ºC until ready for work up.

Each 125 µL reaction was mixed with a suspension of streptavidin-coated beads (250 µL from bottle, 2.5 mg beads; which had been prewashed with B&W buffer) that were suspended in 125 µL 2x B&W buffer (20 mM Tris•HCl, 2 M NaCl, 2 mM EDTA, pH 7.5). The bead suspensions were mixed on a rotary mixer for 1 h before removing the supernatant on a magnetic tube rack. The non-biotinylated DNA strands from the translations were separated from the bead-bound DNA through two ~10 min washes with freshly prepared 20 mM NaOH, which were combined. The combined supernatants from each reaction were mixed with a solution of 2.5 mL (5x volume) of 2:3 sat. aq. guanidine•HCl:isopropanol with 1% v/v 3 M NaOAc and 2.5 µL pH indicator. The resulting material was cleaned up with a Qiaprep 2.0 column and eluted into 52 µL buffer EB (Qiagen). A 1-µL aliquot was removed from each sample for analysis.

Protocol for second-round and later translations for HFNAP pools. An aliquot of double-stranded template in an amount based upon the table in **Figure 3.2** or **Figure 3.9** was added to a 10-µL aliquot of streptavidin-coated magnetic beads in 1x B&W buffer and mixed for 1 h. The non-biotinylated strand was removed from the template strand by washing twice (10 min each) with 20 mM NaOH (aq.) and discarding the flow through. The beads were washed twice with 1x B&W buffer and once with 1x T4 RNA ligase buffer. The supernatant was removed.

The beads were then mixed with a solution containing initiation primer (1.5 equiv), termination primer (1.5 equiv), prepared monomer mix containing 10 equiv each of 32 monomers, 1 µL 10x T4 RNA ligase buffer, and enough water to bring the volume to approximately 8 µL. The strips were placed in a thermal cycler and run with the following program: heat to 95°C for 10 seconds, cool to 65°C, hold at 65 ºC for 4 min, ramp down at 0.1 °C per 10 seconds to 4 °C. While still at 4 ºC, 1 µL of 10 mM ATP and 1 µL T3 DNA ligase (3,000 U/µL) were added to each reaction. The reactions were kept at 4 ºC for 12 h, then warmed to 16 ºC until ready for work up.

Synthesis of individual HFNAPs. A mixture of template (1 µL, 100 µM, 100 pmol), primers (e.g., pp1A and BtBt-pp1Z, 1.5 equiv, 150 pmol), 10x RNA ligase buffer (10 µL), 10 equiv of each of 15 monomers (either functionalized or unfunctionalized) found in HFNAP sequence of interest, and enough water to bring the volume to 80 µL was prepared. The mixture was distributed in 16-µL aliquots to five tubes of a strip, placed in a thermal cycler, and run with the following program: heat to 95°C for 10 seconds, cool to 65°C, hold at 65 ºC for 4 min, then ramp down at 0.1 °C per 10 seconds to 4 °C. While still at 4 ºC, 10 µL of 10 mM ATP and 10 µL T3 DNA ligase (3,000 U/µL) were distributed equally across the tubes. The reactions were kept at 4 ºC for at least 12 h.

The materials were pooled and mixed with a suspension of streptavidin-coated beads (100 µL from bottle, 1.0 mg beads; which had been prewashed with B&W buffer) that were

suspended in 100 µL 2x B&W buffer. The bead suspensions were mixed on a rotary mixer for 1 h before removing the supernatant on a magnetic tube rack. The translations were washed first with 100 µL 1x B&W buffer and then the non-biotinylated DNA strands from the translations were separated from the bead-bound DNA through two ~10-min washes with freshly prepared 20 mM NaOH. The beads were washed twice with 100 µL B&W buffer. The beads were mixed with a solution of 95% formamide/5% EDTA (500 mM, aqueous). The beads were heated at 95 ºC for 30 min, then cooled to room temperature. The supernatants were removed, mixed with ~1 µL Gel Pilot loading buffer, and added to a 10% TBE-Urea gel, which was run for 1 h at 200 V.

The gel was visualized by UV shadowing, identifying bands that appear when shining UV light over a gel placed over a TLC plate with F254 indicator. The gel bands containing the fully-translated HFNAPs were excised from the gel and the HFNAPs were extracted from the gel pieces into 300 µL TE buffer (10 mM Tris, 0.1 mM EDTA, pH 8) by shaking overnight at 37 ºC. The resulting material was mixed with 1.5 mL “Dmitry Buffer” (2:3 guanidine•HCl:IPA with 1% (v/v) aq. NaOAc (3 M)) and concentrated with a MinElute column.

Preparation of protein beads for *in vitro* selection. Protein aliquots were prepared for coupling to agarose beads in the PBS by first diluting in water and then concentrating using Amicon Ultra 0.5 mL 10k centrifugal filters. The concentrated material was dissolved in PBS spun down to concentrate. This dilution-concentration protocol was repeated again before removing the protein solution from the filter and diluting into PBS buffer. The resulting proteins were bound to agarose beads using an AminoLink Plus Micro Immobilization Kit (ThermoFisher Scientific), using 100 µg per 100 µL beads initially, and reducing the loading in later rounds, down to 20 µg PCSK9 or 10 µg IL-6 per 100 µL beads for the PCSK9 and IL-6 selections, respectively.

To make the blank beads for the negative selection, the same agarose beads were prepared without the addition of protein.

In vitro selections for protein binding. As a general note, the selection protocol used in each one of the rounds was similar, but the selection stringency applied at the different rounds varied. The same conditions were applied equally to all of the different pools, but these conditions varied between selections and rounds.

Protein-bound beads were transferred as a suspension, mixing thoroughly until homogeneous prior to transfer, to a spin column with filter. The supernatant was removed by spinning down and the protein beads were washed with selection buffer (0.1 mg/mL BSA, 0.05% Tween-20, in 0.5x DPBS with Ca and Mg ions (Lonza)). The translated materials were mixed with a 2x selection buffer and added to column containing protein-bound beads, then mixed for specified time.

For rounds in which a negative selection was implemented, the translated polymer solution in selection buffer was first mixed with blank beads for an hour and the supernatant was collected. A 1-µL aliquot was removed for analysis and the bulk of the material was mixed with protein-bound beads in the positive selection, mixing for the specified time.

The flow-through from the positive selection was collected and the beads from each of the selection pools were washed thrice with selection buffer. The beads were collected into a tube, mixed with LDS buffer, and incubated at 95 ºC for 15 min. The resulting material was purified by spin column (Qiagen MinElute) and eluted with buffer EB. Generally, the eluted material was mixed with additional buffer, making this material of equal volume to the washes (in the first round, this material was brought to only half the volume of the washes).

A 1-µL aliquot of all of the samples from each selection were analyzed by qPCR using Q5 Hot Start High Fidelity 2x Master Mix (New England Biolabs) and 0.5x SYBR Green I (Life Technologies). The data from the qPCR was used to compare the progress of each of the pools at each round and to determine the number of PCR cycles necessary to amplify the eluted material to late-exponential growth phase.

The preparative PCR was also performed with the same general conditions as the qPCR, except that a biotinylated version of one of the primers was introduced such that the template strand would carry a biotin. The 600-µL reactions were prepared and distributed among a PCR strip in aliquots of 100 µL. The number of PCR cycles from each selection pool was adjusted based upon the qPCR data such that each pool may be amplified with a differing number of cycles.

In the first round, the eluted material PCR-amplified with Q5 Hot Start High-Fidelity 2x Master Mix and purified in two portions such that all of the eluted material was carried forward. In subsequent rounds only about one-third of the material was amplified in future selection rounds.

Following amplification, the materials were cleaned up and concentrated by addition to 3 mL Dmitry Buffer (defined above) and purification using MinElute columns (Qiagen). The resulting material was further purified by PAGE gel using a non-denaturing 10% TBE gel (Bio-Rad). The gel was visualized by UV shadowing, identifying bands that appear when shining UV light over a gel placed over a TLC plate with F254 indicator. The gel bands were excised from the gel with special care to reduce potential contamination between the pools (e.g., a designated gel box was used for these studies, and gels were wrapped in plastic). The resulting double-stranded templates were extracted from the gel pieces into 300 µL TE buffer (10 mM Tris, 0.1 mM EDTA, pH 8) by shaking overnight at 37 ºC. The resulting material was mixed with 1.5 mL Dmitry Buffer and concentrated using a MinElute column. The resulting double-stranded DNA was quantified by Nanodrop.

HTS analysis of libraries. Aliquots collected over the course of the selection were amplified by PCR using Q5 Hot Start High-Fidelity 2x Master Mix with primers installing flanking sequences (see Oligonucleotide sequences: Sequences used for Illumina MiSeq analysis). To ensure adequate sample diversity, a portion of this PCR reaction was analyzed by qPCR to determine

the number of cycles to amplify the bulk of the material to a sub-saturation number of cycles. The resulting amplicons were PAGE-purified and amplified again with Illumina barcoding primers, again to sub-saturation level determined by separate qPCR analysis. The amplicons were PAGE purified and interrogated using an Illumina MiSeq.

When performing the HTS experiments, several barcoded samples were pooled together. Since all of the libraries from each selection campaign had identical primer binding sites, four different sets of tiled adapter primers were used to ensure sequence diversity. The adapter primers were designed to all yield amplicons of the same size. Furthermore, all of the samples from the same set (e.g., all samples from the round 5 eluted libraries) were amplified with the same set of adapters.

The FASTQ files were parsed to first furnish a list of sequences that contain both primer-binding sites with a coding region of correct length. Next, the quality of the resulting sequences was evaluated such that only reads with all base calls of Q32 or higher in the coding region were kept.

The resulting filtered list of reads was processed to computationally translate into the appropriate nucleic acid and/or functionalized polymer sequences. A list of the unique sequences was compiled and the number of instances of each of those unique sequences was recorded. For some of the analyses, the first 100,000 of the filtered reads were analyzed (e.g., for convergence data). In other analyses, like the overall ranking of sequences (which was used to select sequences to synthesize), the entire list of reads was processed.

Analysis of head-to-head competition experiments. Mixed samples prior to the head-to-head competition and those following the elution from the competition were prepared for HTS analysis as described above. Lists of reads were also processed as above, removing improperly sized and low-quality reads. We compiled a list of all of the unique sequences and associated read counts for each sample.

Each of the unique sequences from the head-to-head sample were compared against each of the seven sequence lists resulting from HTS analysis of the enriched pools in the final round of selection (e.g., round 12 for PCSK9). If a sequence from the head-to-head sample only appeared in a single pool, then it was assigned as originating from that pool. If a sequence appeared in multiple pools, then we looked at the ranking of the sequence in the originating pools. Some of the most abundant sequences in the final rounds of selection appeared in multiple pools, likely due to contamination facilitated by the higher copy number of template or polymers. If the sequence found in multiple parent pools was present in the top 1,000 most abundant sequences in one of the pools, that sequence was assigned to the pool in which it was most abundant. If the sequence did not make the top 1,000 threshold, it was considered an unassignable sequence and flagged as a duplicate (see **Figure 3.5** and **Supplementary Figure 3.4**).

Some of the sequences from the head-to-head competition samples were not matched to a parent pool after the initial assignment. In these cases, the same matching protocol was applied using larger lists of sequences that contained all of the unique sequences identified at any round. For instance, the larger list from the unfunctionalized pool contained all of the unique sequences found in rounds 1 through 11. Sequences that failed to match after this more extensive search were considered unassignable.

General protocols for SPR Analysis. SPR analysis was conducted using a Biacore T200 or S200 (GE Healthcare Life Sciences). Analyses were conducted at 25 ºC with a flow rate of 30 µL/min in a buffer of HBS-P+ containing 130 mg/L $CaCl_2$•H_2O, 100 mg/L $MgCl_2$•H_2O, and 200 mg/L KCl (unless otherwise specified). During the SPR experiments, reagents were stored at 4-10 ºC in the reagent compartment.

The choice of which report point to use in single-cycle kinetics runs to evaluate binding differences between different polymers or polymer mixtures was made such that reference binding of the analyte solution was minimal.

Unless otherwise specified, biotinylated HFNAPs or HFNAP libraries were immobilized to an SPR chip surface using a Biotin CAPture kit (GE Life Sciences 28-9202-34). Prior to use, the chip was conditioned three times for 1 min at 30 µL/min using a 3:1 solution of 8 M guanadine•HCl:1 M NaOH. Next, three startup cycles with buffer were performed.

Prior to immobilization on the surface, the active and reference surfaces of the chip were treated with the Biotin CAPture reagent according to the manufacturers recommendations and the ligands were captured onto the active cell. In a single cycle kinetics run, four or five concentrations of protein analyte were flowed over the active and reference surfaces, each for 150 s at 30 µL/min, followed by a dissociation phase of 600 s (unless otherwise specified). The chip surface was regenerated with a 30-s treatment with a 3:1 mixture of 8 M guanadine•HCl and 1 M NaOH at 30 µL/min.

The analysis of each sample ligand in binding to protein was accompanied by a blank run without protein. Sensorgrams were double reference subtracted. Specified response points were used with the following equation to calculate $\%R_{max}$, where *s*, the stoichiometry, is taken as 1:

$$R_{max,\ theoretical} = s \times RU_{ligand\_captured} \times mass_{protein} / mass_{polymer} \quad \text{(eq. 1)}$$

$$\%R_{max} = (RU_{observed} / R_{max,\ theoretical}) \times 100 \quad \text{(eq. 2)}$$

Assessment of pool binding by SPR. To prepare naïve libraries or enriched polymer pools for SPR analysis, the naïve library or the biotinylated templates following the final selection rounds were each amplified by PCR so as to furnish non-biotinylated template strands. PCR reactions were monitored by qPCR to ensure that the amplification was stopped prior to reaching

saturation. The non-biotinylated templates of the enriched libraries were translated with their respective monomer sets to furnish un-/partially-/fully-functionalized nucleic acid polymer pools.

The relative binding of the enriched polymer pools was studied by SPR by using a CAP chip, functionalizing with the CAPture reagent, and treating the chip surfaces with enriched polymer pools for an 1,800 s contact time at 1 µL/min or 900 s at 2 µL/min, followed by a 30 s stabilization period.

The naïve and enriched polymer pools from the PCKS9 selections were immobilized to achieve loading of 56-161 RU for all polymer pools. Concentrations of 7.3, 22, 67, and then 200 nM PCSK9, each with 0.05 mg/mL BSA, were flowed over the surface in the single-cycle kinetics mode. Example reference subtracted sensorgrams from the enriched library samples are shown in **Supplementary Figure 3.1**. The binding report points from the end of the 200 nM injections were used to calculate the $\%R_{max}$ values shown in **Figure 3.3a**, according to eq. 1 and 2. Enriched polymer pool data in **Figure 3.3a** represents $\%R_{max}$ determinations from different Biacore machines, CAP chips, batches of enriched polymer pools, and PCSK9 batches.

The libraries derived from the IL-6 selection were immobilized to the surface with a loading between 7 and 132 RU. Concentrations of 8.9, 27, 80, and 240 nM IL-6 were flowed over the surface in the single-cycle kinetics mode. Example reference subtracted sensorgrams from the enriched library samples are shown in **Supplementary Figure 3.7**. The values from **Figure 3.3d** were derived from calculation of the $\%R_{max}$ achieved following the 80 nM IL-6 injection in the single cycle kinetics run according to eq. 1 and 2. Enriched polymer pool data in **Figure 3.3d** represents $\%R_{max}$ determinations from different Biacore instruments, CAP chips, batches of enriched polymer pools, and IL-6 batches.

Since the identity and composition of the libraries determines the total mass, the mass used in the calculation of each of the polymer libraries was approximated based upon the

average of the calculated masses for each of the most abundant individual polymers synthesized.

Assessment of individual HFNAP binding by SPR. The $\%R_{max}$ for each polymer was determined using the instrument response from the double reference subtracted sensorgram. For the polymers resulting from the PCSK9 selections, the $\%R_{max}$ was calculated using the SPR response at the binding point after a single injection of 60 nM PCSK9 at 30 µL/min for 600 s. Measurements were taken using a Biacore S200 with the loading of polymers on the surface ranged from 9 to 119 RU. Polymers exhibited a range of behavior, with some representative background subtracted sensorgrams appearing in **Supplementary Figure 3.2**. Some values are calculated as the average of multiple measurements.

For assaying affinity of HFNAPs derived from the IL-6 selections, the buffer used was HBS-P+ with added $CaCl_2$, $MgCl_2$, and KCl. Injections of 8.8, 27, 80, and then 240 nM IL-6 were performed in single-cycle kinetics experiments. SPR response was measured at the binding point at the end of the final, 240 nM IL-6 injection and used to calculate the $\%R_{max}$. The amount captured on the surface varied between 46 and 176 RU. Example sensorgrams are shown in **Supplementary Figure 3.6**. Some values are calculated as the average of multiple measurements.

K_d values of individual polymers were determined from single-cycle kinetics analysis immobilizing biotinylated polymers. The affinity of HFNAP Nonpolar Plus-2 to PCSK9 was determined by immobilizing biotinylated HFNAP on the surface of an SA chip. Data were collected from runs with 10, 30, 100, and 300 nM PCSK9 and 0.3, 1, 3, 10, and 30 nM PCSK9 concentrations. Binding of HFNAP Full-2 from the IL-6 selection to IL-6 was studied by immobilizing biotinylated Full-2 to a CAP chip. Data were collected from runs with 0.3, 1, 3, 10, and 30 nM IL-6. All assessments were made using HBS-P+ buffer.

Code availability. The scripts used to analyze data from the head-to-head competition experiments are described and included in Appendix, Supplementary Notes 3.2-3.4.

Data availability. High-throughput sequencing data will be available from the NCBI Sequence Read Archive under accession code SRP153119.

3.8 Acknowledgments

This work was supported by DARPA N66001-14-2-4053, NIH R35 GM118062, and the Howard Hughes Medical Institute. Z.C. was partially supported by the Y. Kishi Graduate Prize in Chemistry and Chemical Biology sponsored by the Eisai Corporation. The authors thank Dr. Alix Chan, Dr. Juan Pablo Maianti, Dr. Michael Packer, Dr. Dmitry L. Usanov, Dr. Ben Thuronyi, and Dr. Chris Wilson for helpful comments and discussions. We also thank Dr. Kelly Arnett and Robyn Stoller for guidance with SPR experiments and analysis.

References

1. Kamtekar, S., Schiffer, J. M., Xiong, H., Babik, J. M. & Hecht, M. H. Protein design by binary patterning of polar and nonpolar amino acids. *Science* **262,** 1680–1685 (1993).

2. Davidson, A. R. & Sauer, R. T. Folded proteins occur frequently in libraries of random amino acid sequences. *Proc. Natl. Acad. Sci. U.S.A.* **91,** 2146–2150 (1994).

3. Biologically active LIL proteins built with minimal chemical diversity. **112,** E4717–E4725 (2015).

4. Taylor, S. V., Walter, K. U., Kast, P. & Hilvert, D. Searching sequence space for protein catalysts. *Proc. Natl. Acad. Sci. U.S.A.* **98,** 10596–10601 (2001).

5. Protein Design: The Choice of de Novo Sequences. **272,** 2031–2034 (1997).

6. A ribozyme composed of only two different nucleotides. **420,** 841–844 (2002).

7. Ruff, K. M., Snyder, T. M. & Liu, D. R. Enhanced Functional Potential of Nucleic Acid Aptamer Libraries Patterned to Increase Secondary Structure. *J. Am. Chem. Soc.* **132,** 9453–9464 (2010).

8. A ribozyme that lacks cytidine. **402,** 323–325 (1999).

9. Schlosser, K. & Li, Y. DNAzyme-mediated catalysis with only guanosine and cytidine nucleotides. *Nucleic Acids Research* **37,** 413–420 (2009).

10. Tuerk, C. & Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249,** 505–510 (1990).

11. Ellington, A. D. & Szostak, J. W. In vitro selection of RNA molecules that bind specific ligands. *Nature* **346,** 818–822 (1990).

12. Robertson, D. L. & Joyce, G. F. Selection in vitro of an RNA enzyme that specifically cleaves single-stranded DNA. *Nature* **344,** 467–468 (1990).

13. Bock, L. C., Griffin, L. C., Latham, J. A., Vermaas, E. H. & Toole, J. J. Selection of single-stranded DNA molecules that bind and inhibit human thrombin. *Nature* **355,** 564–566 (1992).

14. Selection in vitro of single-stranded DNA molecules that fold into specific ligand-binding structures. **355,** 850–852 (1992).

15. Keefe, A. D., Pai, S. & Ellington, A. Aptamers as therapeutics. *Nat Rev Drug Discov* **9,** 537–550 (2010).

16. A DNA enzyme that cleaves RNA. **1,** 223–229 (1994).

17. Capping DNA with DNA. **39,** 3106–3114 (2000).

18. In vitro selection, characterization, and application of deoxyribozymes that cleave RNA. **33,** 6151–6163 (2005).

19. Schlosser, K. & Li, Y. Biologically Inspired Synthetic Enzymes Made from DNA. *Chemistry & Biology* **16,** 311–322 (2009).

20. Walsh, S. M., Sachdeva, A. & Silverman, S. K. DNA Catalysts with Tyrosine Kinase Activity. *J. Am. Chem. Soc.* **135,** 14928–14931 (2013).

21. Silverman, S. K. Catalytic DNA: Scope, Applications, and Biochemistry of Deoxyribozymes. *Trends Biochem. Sci.* **41,** 595–609 (2016).

22. Gold, L. *et al.* Aptamer-Based Multiplexed Proteomic Technology for Biomarker Discovery. *PLoS ONE* **5,** e15004 (2010).

23. Ren, X., Gelinas, A. D., Carlowitz, I., Janjic, N. & Pyle, A. M. Structural basis for IL-1α recognition by a modified DNA aptamer that specifically inhibits IL-1α signaling. *Nature communications* **8,** 810 (2017).

24. Detection of Clostridium difficile toxins A, B and binary toxin with slow off-rate modified aptamers. **76,** 278–285 (2013).

25. Zhou, C. *et al.* DNA-Catalyzed Amide Hydrolysis. *J. Am. Chem. Soc.* **138,** 2106–2109 (2016).

26. Hollenstein, M., Hipolito, C. J., Lam, C. H. & Perrin, D. M. A DNAzyme with Three Protein-Like Functional Groups: Enhancing Catalytic Efficiency of M2+-Independent RNA Cleavage. *Chembiochem* **10,** 1988–1992 (2009).

27. Sequence-specific cleavage of RNA in the absence of divalent metal ions by a DNAzyme incorporating imidazolyl and amino functionalities. **32,** 1591–1601 (2004).

28. RNA cleavage by a DNA enzyme with extended chemical functionality. **122,** 2433–2439 (2000).

29. Thomas, J. M., Yoon, J.-K. & Perrin, D. M. Investigation of the Catalytic Mechanism of a Synthetic DNAzyme with Protein-like Functionality: An RNaseA Mimic? *J. Am. Chem. Soc.* **131,** 5648–5658 (2009).

30. Rohloff, J. C. *et al.* Nucleic Acid Ligands With Protein-like Side Chains: Modified Aptamers and Their Use as Diagnostic and Therapeutic Agents. *Molecular Therapy - Nucleic Acids* **3,** e201 (2014).

31. Gawande, B. N. *et al.* Selection of DNA aptamers with two modified bases. *Proceedings of the National Academy of Sciences* **114,** 2898–2903 (2017).

32. Davies, D. R. *et al.* Unique motifs and hydrophobic interactions shape the binding of modified DNA ligands to protein targets. *Proc. Natl. Acad. Sci. U.S.A.* **109,** 19971–19976 (2012).

33. Hili, R., Niu, J. & Liu, D. R. DNA ligase-mediated translation of DNA into densely functionalized nucleic acid polymers. *J. Am. Chem. Soc.* **135,** 98–101 (2013).

34. Chen, Z., Lichtor, P. A., Berliner, A. P., Chen, J. C. & Liu, D. R. Evolution of sequence-defined highly functionalized nucleic acid polymers. *Nat Chem* **346,** 1 (2018).

35. Kong, D., Lei, Y., Yeung, W. & Hili, R. Enzymatic Synthesis of Sequence-Defined Synthetic Nucleic Acid Polymers with Diverse Functional Groups. *Angew. Chem. Int. Ed.* **55,** 13164–13168 (2016).

36. Kong, D., Yeung, W. & Hili, R. In Vitro Selection of Diversely-Functionalized Aptamers. *J. Am. Chem. Soc.* jacs.7b07241 (2017). doi:10.1021/jacs.7b07241

37. Lei, Y., Washington, J. & Hili, R. Efficiency and fidelity of T3 DNA ligase in ligase-catalysed oligonucleotide polymerisations. *Org. Biomol. Chem.* **17,** 1962–1965 (2019).

38. Hydrophobicity of amino acid residues in globular proteins. **229,** 834–838 (1985).

39. The importance of being tyrosine: lessons in molecular recognition from minimalist synthetic binding proteins. **4,** 325–334 (2009).

40. Antigen-antibody interface properties: composition, residue interactions, and features of 53 non-redundant structures. **1824,** 520–532 (2012).

41. IL-6 as a keystone cytokine in health and disease. **16,** 448–457 (2015).

42. Cholesterol-Lowering PCSK9 Inhibitors Near Market Entry. **93,** 10–14 (2015).

43. Evolocumab and Clinical Outcomes in Patients with Cardiovascular Disease. **376,** 1713–1722 (2017).

44. Chemically Modified DNA Aptamers Bind Interleukin-6 with High Affinity and Inhibit Signaling by Blocking Its Interaction with Interleukin-6 Receptor. **289,** 8706–8719 (2014).

45. Crystal structure of interleukin-6 in complex with a modified nucleic acid ligand. **289,** 8720–8734 (2014).

46. Monitoring genomic sequences during SELEX using high-throughput sequencing: neutral SELEX. **5,** e9169 (2010).

47. IL-6 as a keystone cytokine in health and disease. **16,** 448–457 (2015).

48. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. **23,** 3–25 (1997).

49. Physiochemical drug properties associated with in vivo toxicological outcomes. **18,** 4872–4875 (2008).

Chapter 4 *In vitro* selection for HFNAPs with nanomolar affinity to a small molecule target

This chapter has been adapted from:

Chen, J. C., Chen, J. P., Shen, M. W., Wornow, M., Bae, M., Yeh, W. H., Hsu, A., & Liu, D. R. Machine learning-generated target-binding polymers from *in vitro* selection data. (*submitted*)

Contributions:

Professor David Liu and I initiated the project. I designed and conducted the major *in vitro* experiments including the *in vitro* selection, re-selection, and characterization experiments. I was assisted in these *in vitro* experiments by Minwoo Bae, Wei-Hsi Yeh, and Alvin Hsu. I designed and conducted the computational experiments, in conjunction with Jonathan P. Chen, Michael Wornow, and Max Shen.

4.1 Introduction

In vitro selection—iterative cycles of selection, amplification, and occasional mutagenesis on large combinatorial libraries—is a well-established technique[1,2] that enables the isolation of sequence-defined polymers with high levels of binding[3,4] or catalytic activity[5-9]. Because researchers do not have *a priori* knowledge about the location of active polymer variants in a vast sequence space, *in vitro* selections typically start with random or semi-random combinatorial libraries[10,11]. A selection campaign must satisfy two criteria to successfully isolate a variant with high activity. First, the starting library must contain a sufficient number of potential hits, or variants that are closely related to potential hits. Second, the selection methodology must efficiently enrich rare, active sequences. While larger (higher complexity) libraries may contain a greater absolute number of hits, the larger sequence space makes these hits more difficult to find. The challenge of satisfying both criteria frequently limits the ability of *in vitro* selections to identify desirable candidates.

To address this challenge, we combine selection improvements with a chemically diverse starting library of sequence-defined synthetic polymers previously developed by our group[12] known as highly functionalized nucleic acid polymers (HFNAPs)[12,13] to isolate variants with high activity. HFNAPs consist of oligonucleotides with side-chains on every third nucleotide that are translated from DNA templates using a ligation-based translation system (**Figure 4.1a**). Translation is mediated by T3 ligase and uses a set of customizable trinucleotide building blocks containing a 5' modified nucleobase with side-chains chosen by the researcher (**Figure 4.1b**)[12,13]. In contrast with other non-natural polymer systems compatible with Darwinian selection that replace individual nucleotide triphosphates with functionalized variants[14], or that use non-natural base pairs[15,16], the HFNAP system uses 32 codons, which collectively can encode up to 32 distinct side-chains. The current iteration of the HFNAP translation system consists of 15 consecutive trinucleotide building blocks, each of which contains one of eight different side-chains encoded by 32 codons. The resulting library's side-chain diversity (8^{15} =

~$3.5x10^{13}$) is orders of magnitude larger than the side-chain diversity achievable with polymerase-synthesized functionalized oligonucleotides of the same sequence length. The enhanced chemical diversity of HFNAPs compared to conventional oligonucleotides improves the likelihood of isolating variants with desirable properties such as target protein binding[12,17].

In this study, our judiciously chosen selection conditions and a chemically diverse starting library of HFNAPs enabled the discovery of highly active sequences that bind daunomycin, a small-molecule chemotherapeutic[18-20]. Enhanced selection stringency improved the efficiency with which active HFNAPs were enriched, resulting in the identification of rare sequences with increased affinity. The selection stringency also successfully selected against weaker binders. We observe altered side-chain preferences, in contrast to previously discovered preferences for protein binding[17]. These findings demonstrate that HFNAPs can be evolved to potently bind a target small molecule and provide an integrated experimental approach for designing selections with strong selection stringency.

4.2 Initial selection for HFNAPs that bind daunomycin from naïve libraries

The HFNAP translation system (**Figure 4.1a**) uses a DNA template with a 45-bp coding region containing 15 consecutive codons to recruit building blocks drawn from a set of 32 side-chain-functionalized trinucleotides (**Figure 4.1b**). As the building blocks hybridize to their complementary codons[12,13], T3 DNA ligase ligates their phosphodiester backbones into one continuous strand[21]. HFNAPs are then selectively eluted through biotin capture of the heteroduplex HFNAP–DNA double-stranded product followed by alkaline denaturation to remove the non-biotinylated HFNAP strand (**Figure 4.1a**). The total sequence space in this library is $32^{15} = 3.8x10^{22}$ sequences.

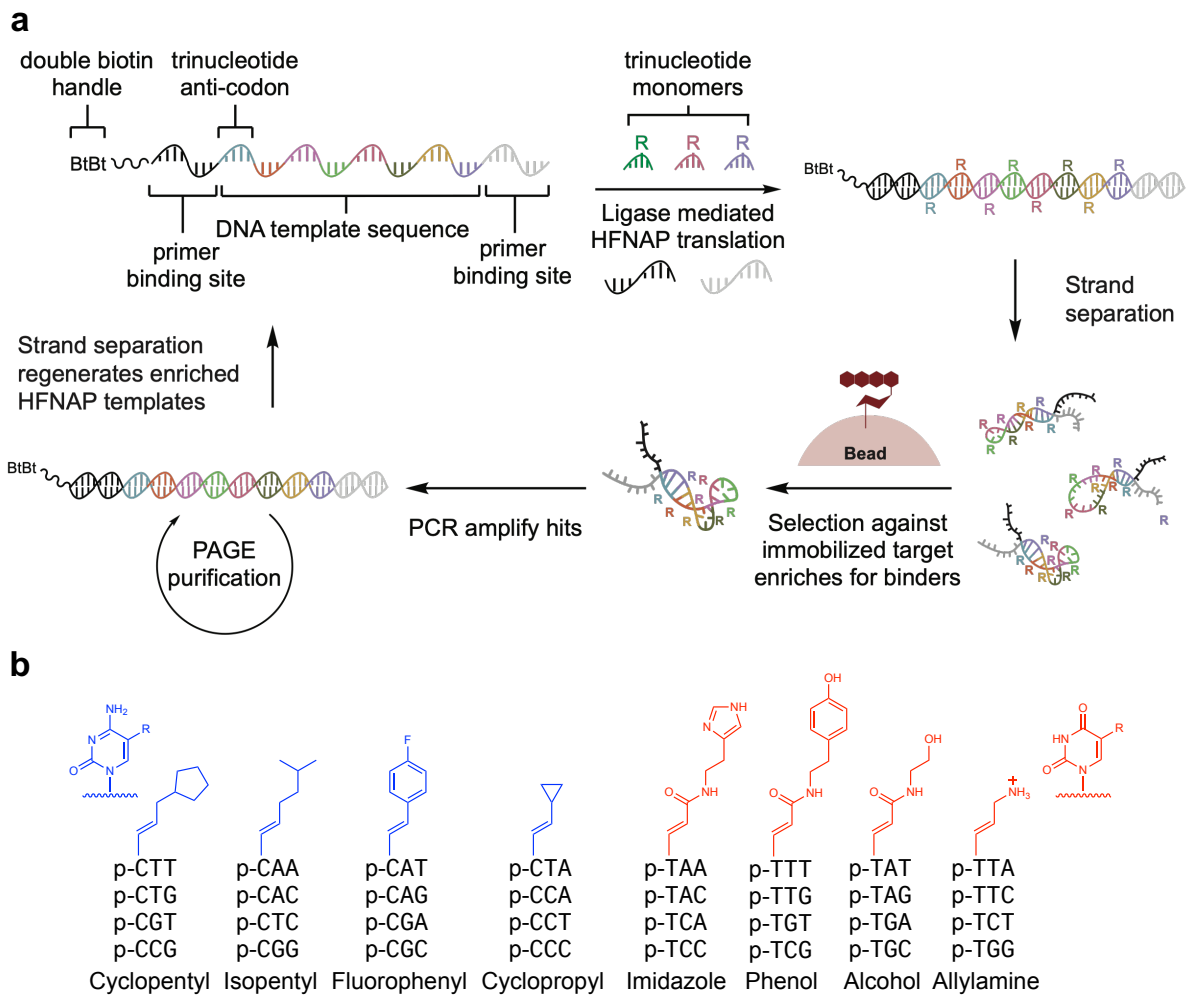


Figure 4.1 | HFNAP building blocks, translation, and selection. a, Overview of the HFNAP translation, selection against daunomycin, and reverse translation to regenerate DNA templates. **b**, Structures of the 5'-phosphorylated trinucleotide monomer building blocks used in this selection, along with their associated, four-fold redundant DNA codons.

Following DNA-templated translation of a 125 pmol starting template library (7.5×10^{13} molecules), we performed iterated rounds of *in vitro* selection for binding to immobilized daunomycin. Biotinylated daunomycin was synthesized by coupling a biotin linker to daunomycin's primary amine and immobilized by incubating the product with streptavidin-linked magnetic beads (see Methods). To prevent selection of HFNAPs that bind the biotin linker or the streptavidin-linked beads, a negative selection step was introduced after round 4. Translated HFNAPs were incubated with a 1:1 mixture of blank streptavidin-linked beads and streptavidin-linked beads with immobilized negative selection linker (see Methods). The flow-through from

this negative selection step was then subjected to positive selection for binding to immobilized daunomycin.

After removal of the positive selection flow-through, beads were washed three times with selection buffer (see Methods). Bound HFNAPs were eluted by reduction of the disulfide bond connecting daunomycin and the biotin linker (rounds 1-3) or by target elution with 1 mM daunomycin for 30 min, followed by two washes with selection buffer (rounds 4 and 5a-9a) (**Figure 4.2**). The eluted HFNAPs were "reverse translated" in a PCR reaction using Q5 DNA polymerase, which we previously found to efficiently copy HFNAPs back into complementary DNA[13]. Full-length amplicons were purified by native polyacrylamide gel electrophoresis (PAGE). Template DNA strands were isolated by biotin capture and alkaline denaturation for the next round of HFNAP library translation and selection.

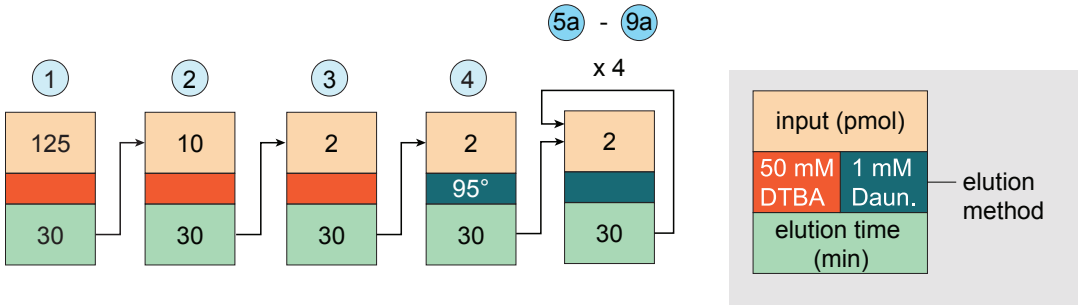


Figure 4.2 | Selections against daunomycin yield HFNAPs with binding affinity. Selection scheme and parameters are shown. Early rounds utilize reducing conditions to elute binders, to broadly capture HFNAPs. Subsequent rounds rely on target elution with 1 mM daunomycin. Selection 1 (rounds 1-4, 5a-9a) is denoted by the upper path, selection 2 (rounds 4, 5b-8b) by the lower path. **b**, MST characterization of the binding affinity for the top 4 most enriched sequences from selection 1. Sequences were found to bind with Kd = 1-2 μM. Values and error bars reflect the mean and SEM of n = 3 independent replicates. Error bars for some values are too small to extend beyond the data point. **c**, The evolutionary trajectory of the top ten most enriched sequences in selection 1.

We performed nine iterated rounds of *in vitro* selection for daunomycin binding using the conditions summarized in **Figure 4.2** (selection 1, rounds 1-4 and 5a-9a). As the selection progressed, the ratio of eluted HFNAPs to HFNAPs in the flow-through increased, consistent with enrichment of daunomycin-binding HFNAPs (**Figure 4.3**). The sequence composition of the

eluted HFNAPs at the end of each round of selection was analyzed by high-throughput DNA sequencing (HTS). We ranked sequences by overall frequency in round 9a, and the top four HFNAPs were synthesized by ligase-mediated DNA-templated translation of individual templates. We characterized the binding affinity of these individual polymers for free daunomycin by microscale thermophoresis (MST) and found that these HFNAPs bound with modest affinity (K_d = ~1-2 μM) (**Figure 4.4**). These results revealed that modest affinity was sufficient to pass the initial selection campaign and suggested that selection stringency could be increased to yield more active polymers.

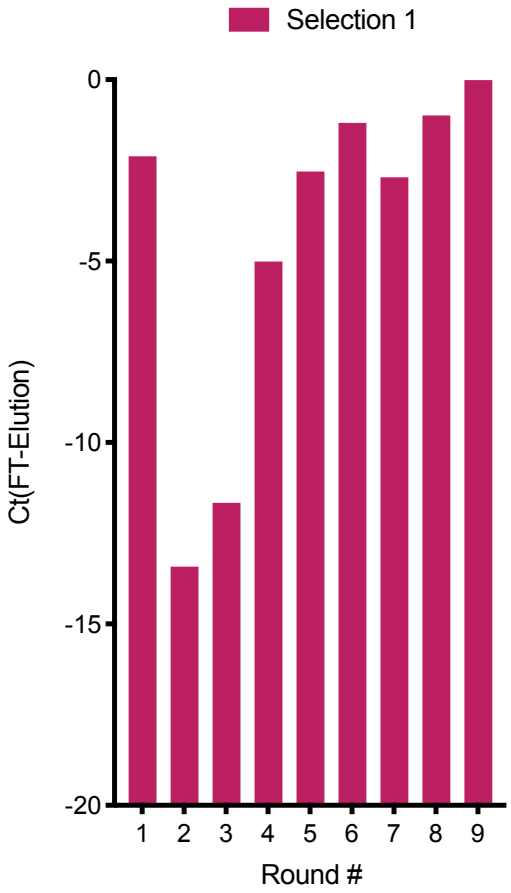
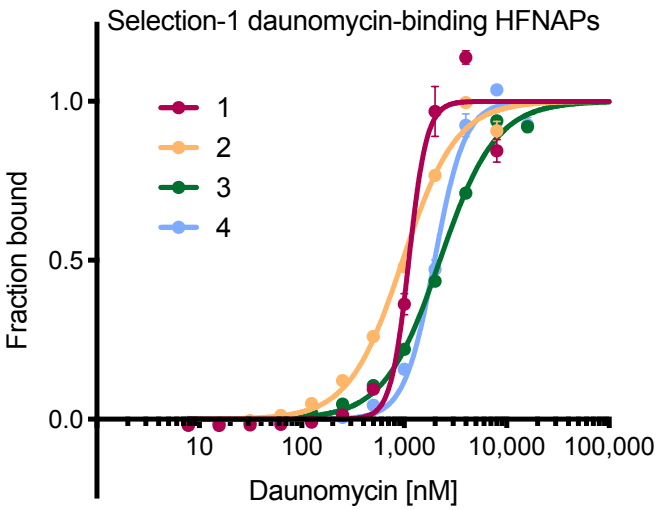


Figure 4.3 | Quantifying selection progress by comparing non-binding flow-throughs to the eluent. Selection-1 progress was followed by qPCR Ct value. Higher bars indicate more HFNAPs were found in the elution compared to the flow-through.



Sequence	K_d
Dm-LS-1	1.11 ± 0.06 μM
Dm-LS-2	0.97 ± 0.06 μM
Dm-LS-3	2.23 ± 0.10 μM
Dm-LS-4	1.98 ± 0.07 μM

Figure 4.4 | MST characterization of the binding affinity for the top 4 most enriched sequences from selection 1. Sequences were found to bind with K_d = 1-2 μM. Values and error bars reflect the mean and SEM of n = 3 independent replicates. Error bars for some values are too small to extend beyond the data point.

4.3 Re-selection with increased stringency

The optimal selection stringency for each round of selection is a delicate balance between preserving sequence diversity and enabling efficient enrichment of active variants. High-stringency conditions applied too early can result in selection failure (no legitimate survivors) or excessive loss of sequence diversity before rare, high-activity sequences can be accessed[22]. Carefully increasing selection stringency minimizes the probability that the stochastic, error-prone, and biased nature of selection steps results in irreversible loss of sequences that may give rise to the top-performing variants at later rounds of selection.

An analysis of the evolutionary trajectory of selection 1 informed our decision to restart the daunomycin-binding selection at round 4 with higher selection stringency. The top ten most enriched sequences from the start of the selection were already highly ranked as early as round 3 and subsequent rounds resulted in minimal changes to the overall ranking and distribution of top sequences (**Figure 4.5**). However, substantial enrichment occurred over these intermediate rounds, with the top ten most enriched sequences comprising 0.6%, 2.9%, and 10% of rounds

3, 4, and 5a, respectively. The most enriched sequence progressed similarly, representing 0.08%, 0.4%, and 1.4% of rounds 3, 4, and 5a, respectively. Overall, the intermediate rounds 3 to 5a exhibited a 5-fold sequence convergence that slowed thereafter (**Figure 4.6**). The overrepresentation of the top ten sequences in round 5, as well as the slowing of sequence convergence in subsequent rounds, eliminated the possibility of restarting the selection at round 5. We chose to restart the selection from the round 4 elution because of its favorable sequence diversity and also because the round 4 selection was the first in which we implemented elution with free daunomycin, which in principle should bias for polymers with binding affinity for free daunomycin and remove those that require the linker or bead.

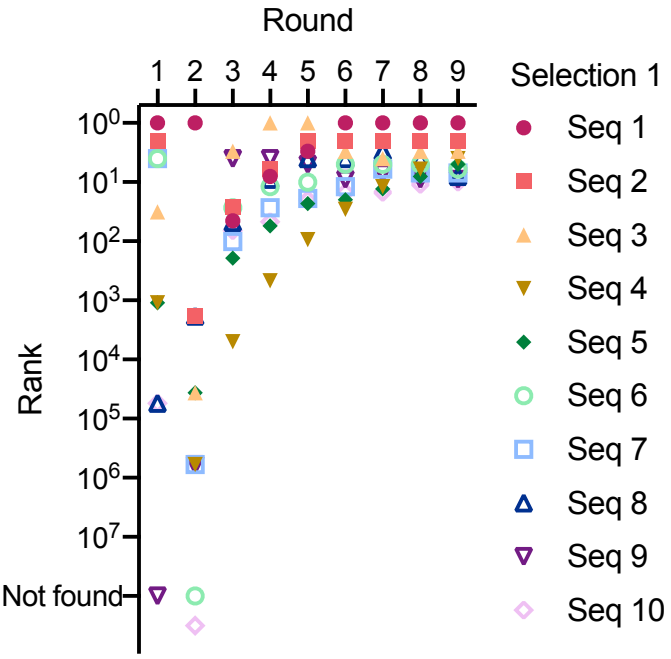


Figure 4.5 | The evolutionary trajectory of the top ten most enriched sequences in selection 1. The individual round ranks of the ten most frequent sequences in round 9 were identified in the previous rounds of selections.

To increase requirements for target binding (k_{on}) and dissociation (k_{off}), we began by raising the selection temperature to 37 °C, as increased temperature generally increases k_{off} since dissociation is an entropically favored, $T\Delta S>0$ process[23]. We also anticipated that elevated temperatures could denature marginally stable HFNAPs and thus enrich sequences

that adopt more stable conformations. We further increased stringency by introducing multiple elution steps to our protocol. The original selection used a single target elution step of 30 minutes, eluting ≥95% of HFNAPs with a target-bound half-life ≤7 minutes and only partially eluting HFNAPs with longer half-lives. The inverse correlation between binding half-life and k_{off} suggests that this approach may result in selection survivors dominated by modest-affinity binders with fast off-rates rather than with strong binders with slow off-rates. To improve selection outcomes by selectively enriching slow off-rate HFNAPs, we implemented an iterated target elution step in which initial 1-minute and 10-minute elutions with 1 mM daunomycin allow weak binders with fast off-rates to elute and be discarded. The final 60-minute elution captures slower off-rate binders to advance into subsequent rounds of selection. We anticipated that the combined changes to increase selection stringency would increase the probability of identifying high-affinity HFNAPs and may even select against modest-affinity binders.

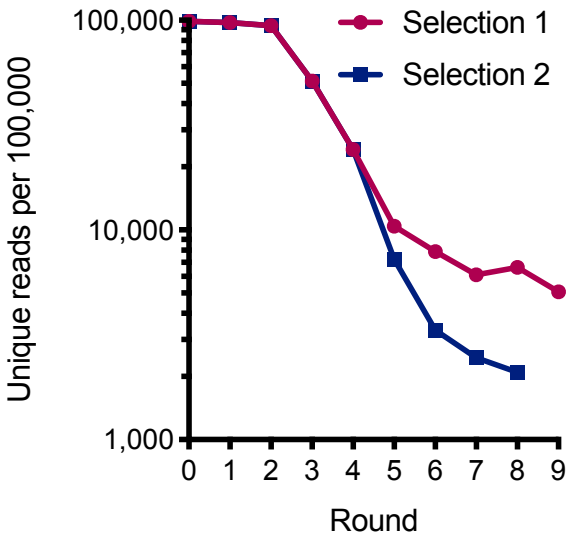


Figure 4.6 | Sequence convergence for selection 1 and selection 2. Sequencing reads for each round of selection were analyzed to determine the degree of sequence diversity, as quantified by unique reads per 100,000.

4.4 Characterization of high-stringency selection outcomes

After four additional rounds of high-stringency selection (**Figure 4.7**, selection 2, rounds 4 and 5b-8b), we observed accelerated sequence convergence, consistent with increased selection stringency (**Figure 4.6**). The increased selection stringency resulted in selection-2 enrichment values between round 4 and round 8b being greater than selection-1 enrichment values between round 4 and round 9a. The top ten most enriched sequences in selection 1 represented 2% of round 4 and 57% of round 9a, while the top ten most enriched sequences in selection 2 represented 0.3% of round 4 and 78% of round 8b, a 9-fold increase in the pooled enrichment (**Supplementary Figure 4.1**). A comparison of the evolutionary trajectories of the top ten sequences in the two selection conditions revealed faster and larger ranking changes for the top sequences in selection 2 (**Figure 4.8**) compared to selection 1 (**Figure 4.5**). Increased selection stringencies also resulted in strong side-chain preferences among surviving sequences that favored the polar allylamine and alcohol side-chains (**Figure 4.9**), which were not observed in selection 1 (**Figure 4.10**). Likewise, we observed decreases in the side-chain frequency of the hydrophobic isopentyl, cyclopentyl, and fluorophenyl side-chains over the course of the selection. These trends contrast sharply with the outcomes of our previous HFNAP selections for binding target proteins, which strongly enriched for hydrophobic side-chains[12,17]. Finally, 676 of the top 1,000 most enriched sequences at the end of selection 1 were found to de-enrich in selection 2 (**Supplementary Figure 4.2**), consistent with the modest daunomycin affinity observed for the selection-1 enriched sequences and demonstrating the effectiveness of high-stringency selection conditions in removing weak binders. Collectively, these results suggest that the more stringent selection conditions in selection 2 increased enrichment efficiency, enabling identification of rare sequences.

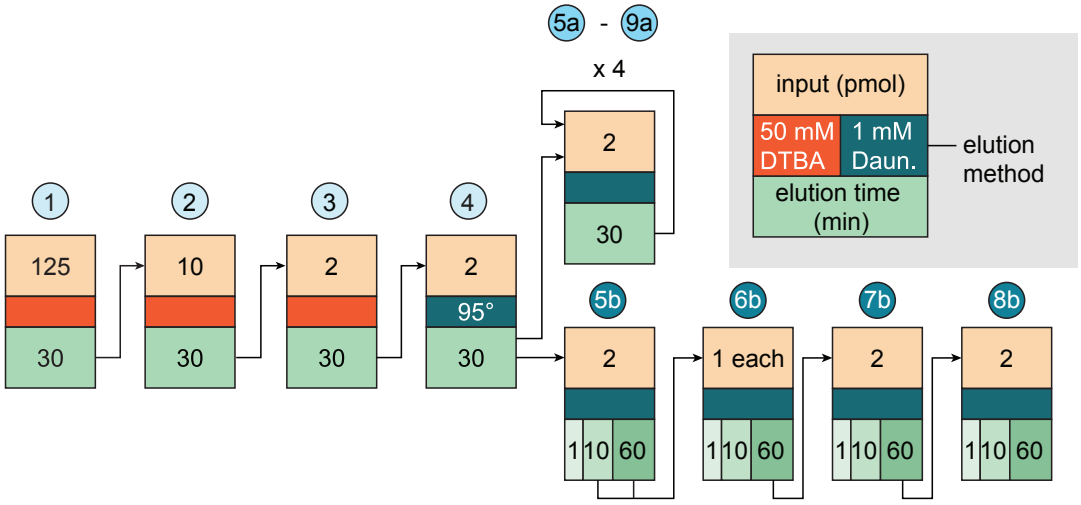


Figure 4.7 | High-stringency selection against daunomycin. Selection scheme and parameters are shown. Early rounds utilize reducing conditions to elute binders, to broadly capture HFNAPs. Subsequent rounds rely on target elution with 1 mM daunomycin. Selection 1 (rounds 1-4, 5a-9a) is denoted by the upper path, selection 2 (rounds 4, 5b-8b, high-stringency) by the lower path.

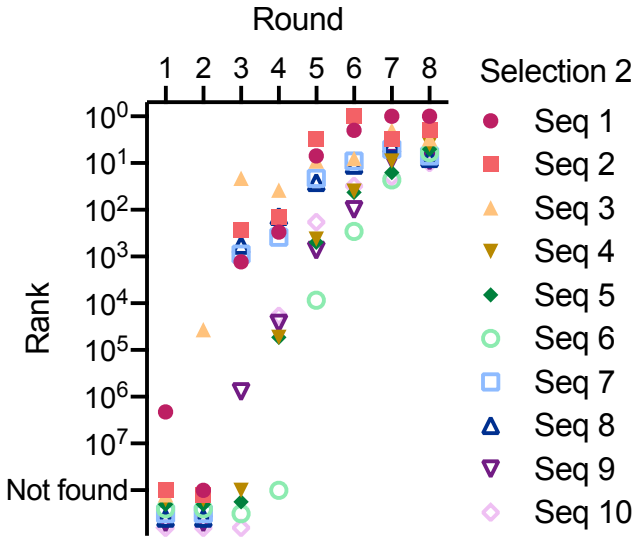


Figure 4.8 | The evolutionary trajectory of the top ten most enriched sequences in selection 2. The individual round ranks of the ten most frequent sequences in selection 2, round 8 were identified in the previous rounds of selections.

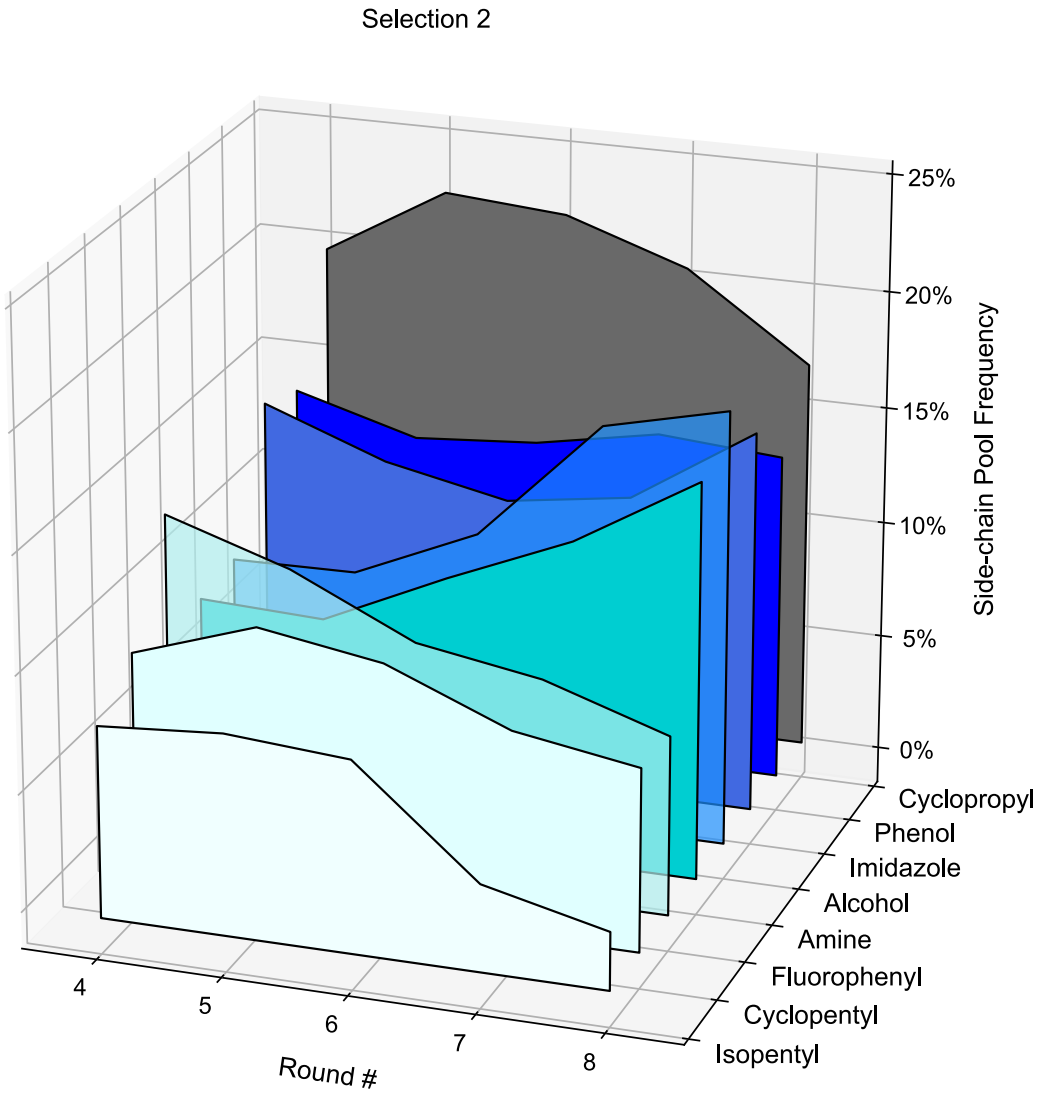


Figure 4.9 | Changes in side-chain distribution for selection 2. Overall side-chain frequencies are plotted over the course of the selection. Decreases in the frequency of isopentyl, cyclopropyl and fluorophenyl side-chains are accompanied by increases in the frequency of amine and alcohol side-chains.

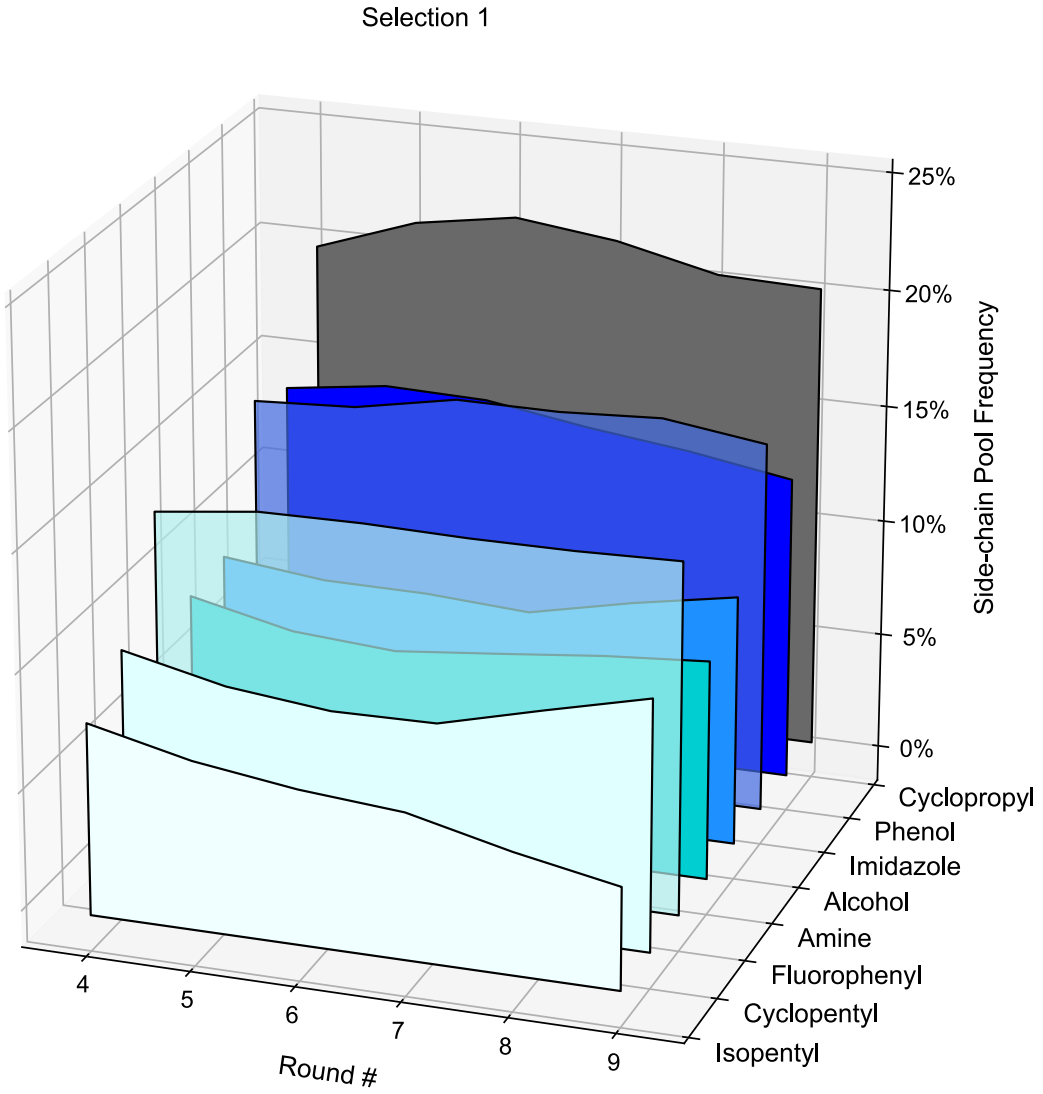


Figure 4.10 | Changes in side-chain distribution for selection 1. Overall side-chain frequencies are plotted over the course of the selection for daunomycin binding. Minimal changes to the side-chain frequencies are observed.

To directly test whether increased selection stringency resulted in the isolation of HFNAPs with improved binding affinity, we used MST to characterize the binding affinities of the top ten most enriched sequences at the end of selection 2. These HFNAPs demonstrated binding affinities between K_d=5-65 nM, representing a ~15-200-fold improvement over HFNAPs emerging from selection 1 (**Figure 4.11** and **Supplementary Figure 4.3**). An independent binding assay using gel filtration of radiolabeled daunomycin corroborated this greatly improved binding affinity compared with the most highly enriched selection 1 HFNAPs (**Supplementary**

Figure 4.4). The same HFNAP sequences without the side-chains were also assayed for binding affinity, and showed ~10-1,000-fold lower daunomycin affinity (**Figure 4.11**), indicating that the HFNAP side-chains play important roles in daunomycin binding.

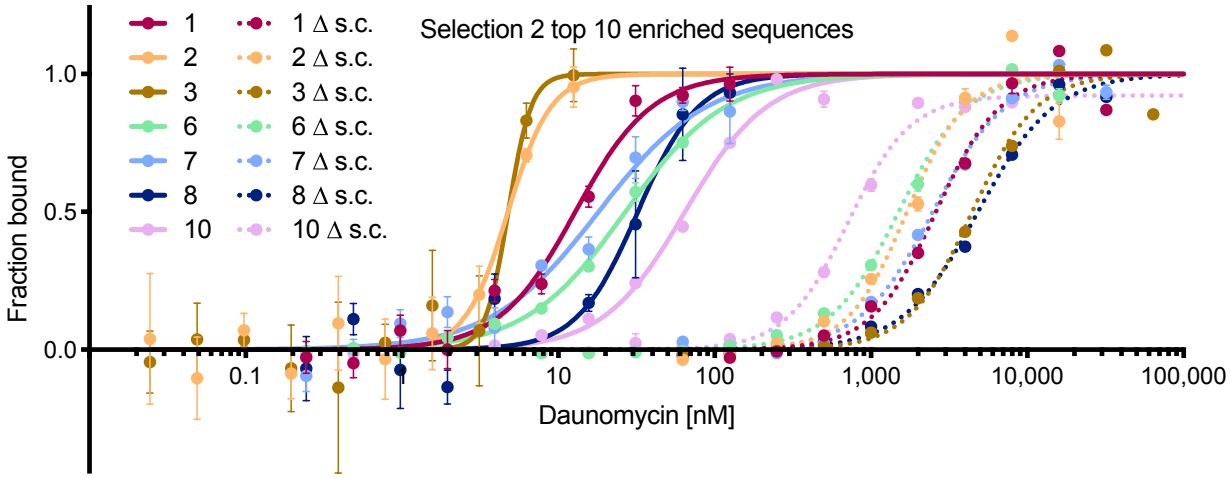
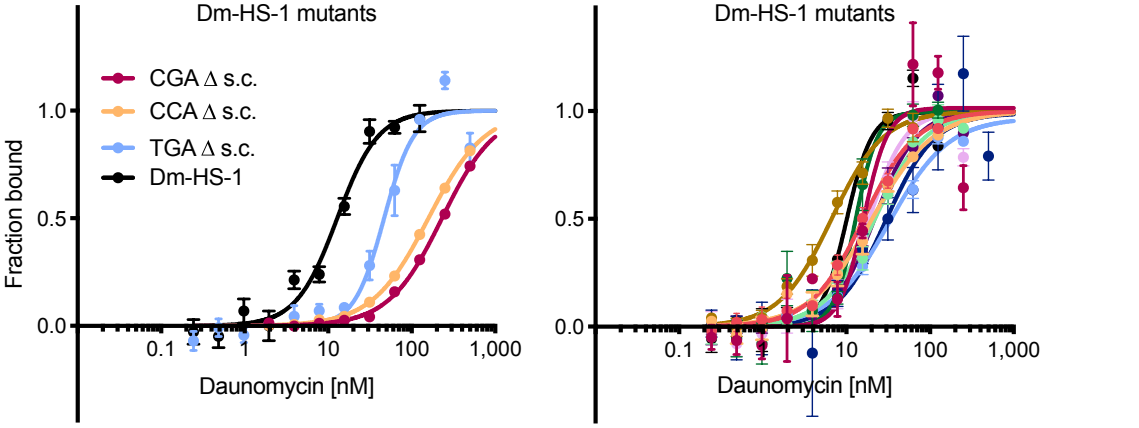


Figure 4.11 | MST characterization of the binding affinity for the top ten most enriched sequences from selection 2. Sequences 4, 5, and 9 are omitted for clarity, as sequences 1, 4, 5 and sequences 3, 9 are each related by 1 nucleotide. 1 Δ s.c. denotes the sequence (Dm-HS-1) without side-chains. Sequences were found to bind with K_d = 5-65 nM. Values and error bars reflect the mean and SEM for n = 3 independent replicates.

To further characterize specific structure-activity relationships between the HFNAPs surviving selection and their side-chains, we synthesized 14 variants of Dm-HS-1, the most highly enriched sequence at the end of selection 2, round 8b, in which we replaced each building block with the corresponding DNA trinucleotide lacking any side-chain (**Figure 4.12**) (one building block, TGC, was present twice in Dm-HS-1). We found that two side-chains play a critical role in binding affinity. Removal of the fluorophenyl side-chain from the CGA building block at position 1 or removal of the cyclopropyl side-chain from the CCA building block at position 9 resulted in >10-fold loss in binding affinity. Additionally, loss of the alcohol side-chain from the TGA building block at position 6 decreased binding affinity by 3-fold. Removal of each of the remaining side chains individually did not substantially affect binding affinity (**Figure 4.12** and **Supplementary Table 4.1**). Together, the results establish that increased selection

stringency enriched rare HFNAPs with improved daunomycin binding affinities that are strongly dependent on the presence of specific side-chains.



Codon	Side-chain	K_d	K_d change
CGA	Fluorophenyl	231.1 ± 23.6 nM	18x
CCA	Cyclopropyl	158.8 ± 14.2 nM	12x
TGA	Alcohol	46.8 ± 4.8 nM	3.6x

Figure 4.12 | MST characterization of the daunomycin-binding affinity of Dm-HS-1 mutants. 14 Dm-HS-1 mutants were translated where a single building block is replaced with the corresponding building block lacking any side-chain. MST characterization of the binding affinity for these Dm-HS-1 mutants is shown. CGA Δ s.c. denotes Dm-HS-1 without a side-chain for the CGA building block. The table shows individual side-chain removals that result in the largest increase in K_d. The activity of the original Dm-HS-1 sequence is shown in black for comparison. Values and error bars reflect the mean and SEM for n = 3 independent replicates.

4.5 Discussion

Current sequencing throughputs and costs do not support analyzing selection progress in real time, leaving researchers to estimate selection progress using various analytical techniques. The resulting selections may arrive at a local optimum, with premature sequence convergence that makes enrichment and identification of rare, higher activity sequences difficult. In the present study, we implemented a high-stringency selection strategy at an optimal, intermediate selection round, based on a deep-sequencing-enabled understanding of selection-1 progress. The strategies we used to increase selection stringency are straightforward and

easily accessible to all researchers, and do not require specialized expertise or technologies to implement. The de-enrichment of top-performing selection-1 HFNAPs indicates that these positive selection strategies for slow off-rates, can have negative selection effects on weaker binders. In the current study, our resulting modifications to selection 2 result in the identification of several HFNAPs with affinity as high as K_d=5-65 nM, representing a ~15-200-fold improvement in affinity over HFNAPs identified from selection 1.

The promise of oligonucleotide-based technologies lies within their large, combinatorial search spaces. Without efficient search methods, *in vitro* selections are confined to search broadly rather than deeply, resulting in a failure to identify rare sequences with high activity. Improving selection methodology is an active area of research, with attempts to increase stringency by narrowing the search space[24] or automating the process[25]. Our work shifts the retention probability of sequences in a selection round based on off-rates, achieving increased selection stringency and higher affinity binders.

Increased selection stringency comes at the cost of sequence diversity. Comparing selection 2 to selection 1, the former demonstrated stronger sequence convergence and preference for hydrogen-bonding side-chains. Loss of sequence diversity has implications for therapeutics, as drug discovery has highlighted the importance of having multiple, diverse candidates. Because chemical modifications to either the sugar[26], phosphate backbone[27-29], or nucleobase[30] can result in loss of activity with increasing modification[31,32], sequence diversity is critical to increasing our chances of identifying candidate oligonucleotides for therapeutic applications. At present, we are limited to conducting labor intensive, parallel *in vitro* selections to achieve multiple, unrelated sequences with high on-target activity.

4.6 Methods

General methods. Oligonucleotides were purchased from IDT and are listed below. Building blocks were synthesized as described previously[12,13]. Building blocks were synthesized either directly from phosphoramidites (Glen Research) or via coupling of an amine to the NHS ester containing phosphoramidite as described previously[12,13]. T3 DNA ligase, ATP, and T4 RNA ligase buffer were purchased from New England BioLabs. MyOne Streptavidin C1 magnetic beads (Life Technologies) were used in immobilization. MST samples were prepared using HBS-P+ buffer (GE Healthcare Life Sciences).

Synthesis of HFNAP libraries by DNA-templated translation. DNA template libraries (125 pmol), initiation primer (188 pmol), termination primer (188 pmol), monomer mix (1.25 nmole/monomer), and 13 µL of water were mixed and added to a PCR strip. Strips were then incubated at the following temperatures using a thermal cycler: 95 °C for 10 s, 65 °C for 4 min, followed by a 0.1 °C per 10 s ramp to 4 °C. 6.25 µL of ATP and 6.25 µL of T3 DNA ligase were then added at 4 °C, and then incubated for 16 h at 4 °C.

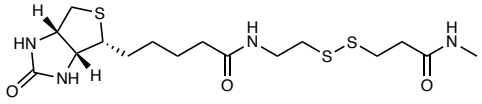
250 µL of streptavidin beads were then prepared by washing in B&W buffer according to manufacturer protocol. The beads were then changed into 125 µL of 2× B&W (20 mM Tris-HCl, 2 M NaCl, 2 mM EDTA, pH 7.5). The translation reaction was then combined with the streptavidin beads and mixed for 1 h on a rotary mixer. The supernatant was discarded, and beads were washed thrice to remove nonspecific binders. The HFNAP strands were then selectively eluted from the biotinylated template using two 5-min washes with 250 µL of freshly prepared 20 mM NaOH. The combined elution was mixed with 5 volumes of column-binding buffer (2:3 sat. aq. guanidine•HCl:isopropanol, 1% v/v 3 M NaOAc, 0.1% pH indicator). The combined sample was then column purified using a QiaPrep 2.0 column and eluted with 52 µL of EB (Qiagen). A 1-µL aliquot was saved as the pre-selection library.

Subsequent rounds of translation require streptavidin bead immobilization of double stranded DNA templates and 20 mM NaOH washes to obtain ssDNA templates for translation. Beads are then washed twice with T4 RNA ligase buffer to neutralize any remaining base before proceeding with synthesis and translation as usual. These bead immobilized templates are then used in lieu of in solution templates.

Synthesis of individual HFNAPs by DNA-templated translation. Synthesis of individual HFNAPs follows a similar protocol as library synthesis. The major difference is use of 10 equivalents of the building blocks for each HFNAP as opposed to a master mix. Following translation, HFNAP isolation and elution, and column cleanup, the sample is then PAGE purified on a denaturing 10% TBE-Urea gel (Criterion, Bio-Rad). The band is visualized by UV shadowing, where DNA bands were identified by the shadow cast by DNA onto a TLC plate with F254 indicator when the gel was illuminated by a UV lamp. The bands were excised from the gel and extruded through stacked microcentrifuge tubes (0.2 mL, 0.5 mL, 1.5 mL), with 17 gauge and 27 gauge holes made in the first two tubes respectively. The resulting material was then resuspended in 400 µL of Crush Soak (1× TE, 200 mM NaCl), and then shaken overnight at 37 ºC.

The gel pieces were then removed using 0.22 µm PVDF centrifugal filters (Millipore). The remaining liquid was then combined with 5× equivalents of column binding buffer, DNA column purified, and eluted with water.

Synthesis of positive and negative selection targets. Biotinylated daunomycin was synthesized by reacting EZ-Link Sulfo-NHS-SS-Biotin (Thermo Scientific) with daunomycin (Selleckchem)[20]. The product was HPLC purified, and concentration determined by A260.



The negative selection target, Me-SS-Biotin, was prepared according to the following procedure. The reaction was conducted under N_2 atmosphere. EZ-Link™ Sulfo-NHS-SS-Biotin (ThermoFisher Scientific) (31.0 mg, 0.0511 mmol) was weighed directly into an oven-dried vial sealed with a Teflon septum and containing a magnetic stir bar. The vial was cooled to 0 °C, and 10 mL of 2 M methylamine in methanol was added to the vial dropwise. The reaction mixture was allowed to warm to room temperature slowly and stirred for 24 h. After this time, the solvent and excess methylamine were removed *in vacuo*. The crude reaction mixture was dissolved in 30/70 acetonitrile:water and purified by reverse phase HPLC (0/100 to 30/70 of acetonitrile/water with 0.1% trifluoroacetic acid). The product was lyophilized overnight before further use.

^{1}H NMR (400 MHz, CD_3OD) δ 4.521(ddd, *J* = 8, 4.8, 1.2 Hz, 1H), 4.336 (dd, *J* = 8, 4.4 Hz, 1H), 3.509 (t, *J* = 6.8 Hz, 2H), 3.239 (ddd, *J* = 10.4, 4.4, 4 Hz, 1H), 2.979 (t, *J* = 8 Hz, 2H), 2.960 (dd, *J* = 14, 4.8 Hz, 1H), 2.856 (t, J = 5.2 Hz, 2H), 2.755 (s, 3H), 2.738 (d, *J* = 14 Hz, 1H), 2.615 (t, *J* = 7.2 Hz, 2H), 2.250 (t, *J* = 7.6 Hz, 2H), 1.695 (m, 4H), 1.479 (p, *J* = 7.6 Hz, 2H).

^{13}C{^{1}H} NMR (151 MHz, CD_3OD) δ 177.09, 175.15, 166.99, 64.24, 62.50, 57.85, 41.90, 40.37, 39.41, 37.59, 37.43, 36.05, 30.60, 30.33, 27.67, 27.24.

MS (ESI): *m/z* for $C_{16}H_{28}N_4O_3S_3$ $[M + TFA - H]^{-1}$ calcd.: 533.12, found: 533.14 (see **Supplementary Note 4.2**).

Daunomycin affinity selection. Daunomycin-linked beads and negative selection beads were prepared prior to the start of each selection round. Streptavidin-linked beads were washed with B&W according to manufacturer protocol, then incubated with biotinylated daunomycin or the negative selection linker (see Methods) for 1 h. The beads were then washed in selection buffer (20 mM Tris-HCl (pH 7.4), 140 mM NaCl, 5 mM KCl, 1 mM $MgCl_2$, 1 mM $CaCl_2$, 0.05% [v/v] Tween 20) thrice, and then resuspended in 2× selection buffer.

The selection began with incubation of column purified HFNAP libraries with immobilized daunomycin on streptavidin-linked beads. The selection was placed on a rotary mixer for 1 h. The flow-through was then collected, the beads washed 3× using selection buffer, and finally putative binders were eluted according to the scheme in **Figure 4.7**. When eluting with 1 mM daunomycin, the beads were washed twice with selection buffer after the initial incubation with daunomycin. The combined material constitutes the eluted material. The eluted material was then mixed with 5× equivalents of column binding buffer, DNA column purified and eluted with water.

Selections started with a negative selection after round 4. The column-purified HFNAP libraries were mixed with negative selection beads on a rotary mixer for 30 min. The flow-through was then collected, and mixed with positive selection, daunomycin-linked beads for 1 h.

A 1-μL aliquot of each sample was analyzed by qPCR using Q5 Hot Start High Fidelity 2× Master Mix (New England BioLabs) and 0.5× SYBR Green I (Life Technologies). The qPCR cycle number was used to determine the number of cycles of amplification for the eluted material. The eluted material was amplified by PCR using Q5 Hot Start High Fidelity 2× Master Mix (New England BioLabs), and a primer set in which one of the primers was biotinylated. The resulting PCR reaction was then DNA column-purified using 5× equivalents of column binding buffer, and eluted with water.

The eluted material was then gel purified in a 10% TBE native PAGE gel (Criterion, Bio-Rad). The gel was visualized by UV shadowing, where DNA bands were identified by the shadow cast by DNA onto a TLC plate with F254 in which when the gel was illuminated by a UV lamp. The bands were excised from the gel and extruded through stacked microcentrifuge tubes (0.2 mL, 0.5 mL, 1.5 mL), with 17-gauge and 27-gauge holes made in the first two tubes respectively. The resulting material was then resuspended in 400 μL of Crush Soak (1× TE, 200 mM NaCl), and then shaken overnight at 37 °C.

The gel pieces were then removed using 0.22 μm PVDF centrifugal filters (Millipore). The remaining liquid was then combined with 5× equivalents of column binding buffer, DNA column purified, and eluted with EB. The absorbance was then measured to determine the concentration of the eluent, for subsequent rounds of translation.

High stringency selection scheme: The two main differences for the high stringency selection scheme were the selection temperature and the elution conditions. The positive and negative selections were both performed at 37 °C, with an initial 2-min incubation in a thermocycler to rapidly raise the temperature of the selection to 37 °C prior to incubation on a rotary mixer in a 37 °C warm room. The elution condition was then modified to contain multiple washes with 1 mM daunomycin, followed by 2 washes with selection buffer. The incubation times and the elution that was carried forward to the next round of selection are shown in **Figure 4.7**.

HTS analysis of libraries. Samples to be analyzed were amplified by PCR using Q5 Hot Start High-Fidelity 2× Master Mix to install Illumina sequencing adapters, with amplification cycle number determined by qPCR. Samples were then DNA column purified, followed by an additional PCR round to install Illumina barcodes. The resulting samples were then combined, DNA column purified, and the concentration was determined by Kapa Library Quantification Kit (Kapa Biosystems). Fastq files were then parsed using AptaSuite, setting a strict cutoff for full-length sequences[33]. The parsed sequences were exported through AptaSuite in Fastq format, and subsequently converted into csv's for downstream analysis.

Microscale Thermophoresis (MST). MST analysis was conducted using a Monolith NT.Automated (Nanotemper). Analyses were conducted at 25 ºC in HBS-P+ buffer (Cytiva/GE Healthcare Life Sciences) with additional 5 mM KCl, 1 mM $MgCl_2$, and 1 mM $CaCl_2$. Sequences were translated using a Cy5 tag and PEG18 spacer, for use in tracking the HFNAP. Generally,

the laser excitation energy was set to 10 % to minimize photobleaching, and the IR laser power was set to high for all readings.

Dilution series for each replicate was made separately, using serial dilutions from the highest concentration to the lowest concentration. The amount of labeled HFNAP that was used was subsaturating, as determined by fluorescence intensity measured against a known standard. Measurements were made in triplicate. Data was processed and fitted using standard Nanotemper software (MO.Affinity) and exported to Prism for plotting.

Gel filtration. Gel filtration was conducted using [^{3}H]-daunomycin, from Perkin-Elmer or American Radiolabeled Chemicals[34]. Analyses were conducted at 25 ºC in selection buffer (20 mM Tris-HCl (pH 7.4), 140 mM NaCl, 5 mM KCl, 1 mM $MgCl_2$, 1 mM $CaCl_2$, 0.05% [v/v] Tween 20). 25 µL of a 40 nM solution of the HFNAP in selection buffer was mixed 1:1 with 25 µL of serial dilutions of radiolabeled daunomycin. The initial daunomycin mixture was made by mixing [^{3}H]-daunomycin with non-radiolabeled daunomycin, followed by 1:1 serial dilution with selection buffer. Samples were incubated in the dark at 25 ºC for at least 30 min. During incubation, Centri-Sep gel filtration column (Princeton Separations catalog number CS- 901) were prepared according to manufacturer instructions by reconstituting with selection buffer. The 50 µL reaction was loaded onto the prepared Centri-Sep columns, followed by centrifugation according to manufacturer instructions. The resulting flow-through was pipetted into 5 mL of Ultima Gold (Perkin-Elmer) in a scintillation vial, vortexed, and counted using a Tri-Carb 2910 TR liquid scintillation counter. The process was conducted in parallel with an unrelated sequence from a different selection against a different target (gel filtration control sequence). Fraction bound by Dm-HS-1 was calculated by taking the difference between the amount of [^{3}H]-daunomcyin bound by Dm-HS-1 and the amount bound by the control sequence.

4.7 Acknowledgments

This work was supported by NIH R35 GM118062, DARPA N66001-14-2-4053, and the Howard Hughes Medical Institute. We thank J.P. Maianti, P. Lichtor, Z. Chen, and A. Berliner for helpful comments. We also thank K. Arnett for help with MST experiments and analysis. We thank A. Vieira for assistance editing the manuscript. We thank Mary O'Reilly and Bang Wong for help with figure design and illustrations.

References

1. Ellington A. D., Szostak J. W. In vitro selection of RNA molecules that bind specific ligands. *Nature* 1990, **346**(6287)**:** 818--822.

2. Tuerk C., Gold L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science (New York, NY)* 1990, **249**(4968)**:** 505--510.

3. Bock L. C., Griffin L. C., Latham J. A., Vermaas E. H., Toole J. J. Selection of single-stranded DNA molecules that bind and inhibit human thrombin. *Nature* 1992, **355**(6360)**:** 564--566.

4. Huizenga D. E., Szostak J. W. A DNA aptamer that binds adenosine and ATP. *Biochemistry* 1995.

5. Robertson D. L., Joyce G. F. Selection in vitro of an RNA enzyme that specifically cleaves single-stranded DNA. *Nature* 1990, **344**(6265)**:** 467--468.

6. Zhang B., Cech T. R. Peptide bond formation by in vitro selected ribozymes. *Nature* 1997.

7. Seelig B., Jschke A. A small catalytic RNA motif with Diels-Alderase activity. *Chemistry \& Biology* 1999, **6**(3)**:** 167--176.

8. Sheppard T. L., Ordoukhanian P., Joyce G. F. A DNA enzyme with N-glycosylase activity. *Proceedings of the National Academy of Sciences of the United States of America* 2000, **97**(14)**:** 7802--7807.

9. Zhou C.*, et al.* DNA-Catalyzed Amide Hydrolysis. *Journal of the American Chemical Society* 2016, **138**(7)**:** 2106--2109.

10. Ruff K. M., Snyder T. M., Liu D. R. Enhanced Functional Potential of Nucleic Acid Aptamer Libraries Patterned to Increase Secondary Structure. *Journal of the American Chemical Society* 2010, **132**(27)**:** 9453--9464.

11. Davis J. H., Szostak J. W. Isolation of high-affinity GTP aptamers from partially structured RNA libraries. *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99**(18)**:** 11616--11621.

12. Chen Z., Lichtor P. A., Berliner A. P., Chen J. C., Liu D. R. Evolution of sequence-defined highly functionalized nucleic acid polymers. *Nature chemistry* 2018, **346:** 1.

13. Hili R., Niu J., Liu D. R. DNA ligase-mediated translation of DNA into densely functionalized nucleic acid polymers. *Journal of the American Chemical Society* 2013, **135**(1)**:** 98--101.

14. Gawande B. N.*, et al.* Selection of DNA aptamers with two modified bases. *Proceedings of the National Academy of Sciences* 2017, **114**(11)**:** 2898--2903.

15. Kimoto M., Yamashige R., Matsunaga K.-i., Yokoyama S., Hirao I. Generation of high-affinity DNA aptamers using an expanded genetic alphabet. *Nature biotechnology* 2013, **31**(5)**:** 453--457.

16. Sefah K.*, et al.* In vitro selection with artificial expanded genetic information systems. *Proceedings of the National Academy of Sciences of the United States of America* 2014, **111**(4)**:** 1449--1454.

17. Lichtor P. A., Chen Z., Elowe N. H., Chen J. C., Liu D. R. Side chain determinants of biopolymer function during selection and replication. *Nature chemical biology* 2019, **262**(4)**:** 1--426.

18. Gui Fang C., Jie Z., Yong Hua T., Pin Gang H., Yu Zhi F. Study on the Interaction between Antitumor Drug Daunomycin and DNA. *Chinese Journal of Chemistry* 2005, **23**(5)**:** 576--580.

19. Blum R. H., Carter S. K. AdriamycinA New Anticancer Drug with Significant Clinical Activity. *Annals of Internal Medicine* 1974, **80**(2)**:** 249--259.

20. Wochner A.*, et al.* A DNA aptamer with high affinity and specificity for therapeutic anthracyclines. *Analytical Biochemistry* 2008, **373**(1)**:** 34--42.

21. Lei Y., Washington J., Hili R. Efficiency and fidelity of T3 DNA ligase in ligase-catalysed oligonucleotide polymerisations. *Organic \& Biomolecular Chemistry* 2019, **17**(7)**:** 1962--1965.

22. B L. Stochastic modeling and optimization of phage display. *Journal of Molecular Biology* 1998, **277**(4)**:** 893--916.

23. Johnstone R. W., Andrew S. M., Hogarth M. P., Pietersz G. A., McKenzie I. F. C. The effect of temperature on the binding kinetics and equilibrium constants of monoclonal antibodies to cell surface antigens. *Molecular Immunology* 1990, **27**(4)**:** 327--333.

24. Wang J.*, et al.* Particle Display: A Quantitative Screening Method for Generating High-Affinity Aptamers. *Angewandte Chemie International Edition in English* 2014, **53**(19)**:** 4796--4801.

25. Glkler J., Schtze T., Konthur Z. Automation in the High-throughput Selection of Random Combinatorial Libraries \textemdash Different Approaches for Select Applications. *Molecules (Basel, Switzerland)* 2010, **15**(4)**:** 2478--2490.

26. Eaton B. E., Pieken W. A. Ribonucleosides and RNA. *Annual review of biochemistry* 1995, **64**(1)**:** 837--863.

27. Vosberg H. P., Eckstein F. Effect of deoxynucleoside phosphorothioates incorporated in DNA on cleavage by restriction enzymes. *The Journal of biological chemistry* 1982, **257**(11)**:** 6595--6599.

28. Wickstrom E. Oligodeoxynucleotide stability in subcellular extracts and culture media. *Journal of Biochemical and Biophysical Methods* 1986, **13**(2)**:** 97--102.

29. Purcell J., Hengge A. C. The Thermodynamics of Phosphate versus Phosphorothioate Ester Hydrolysis. *The Journal of organic chemistry* 2005, **70**(21)**:** 8437--8442.

30. Vaught J. D.*, et al.* Expanding the Chemistry of DNA for in Vitro Selection. *Journal of the American Chemical Society* 2010, **132**(12)**:** 4141--4151.

31. Ni S.*, et al.* Chemical Modifications of Nucleic Acid Aptamers for Therapeutic Purposes. *International Journal of Molecular Sciences* 2017, **18**(8)**:** 1683.

32. Khvorova A., Watts J. K. The chemical evolution of oligonucleotide therapies of clinical utility. *Nature biotechnology* 2017, **35**(3)**:** 238--248.

33. Hoinka J., Backofen R., Przytycka T. M. AptaSUITE: A Full-Featured Bioinformatics Framework for the Comprehensive Analysis of Aptamers from HT-SELEX Experiments. *Molecular Therapy - Nucleic Acids* 2018, **11:** 515--517.

34. Curtis E. A., Liu D. R. Discovery of widespread GTP-binding motifs in genomic DNA and RNA. *Chemistry \& Biology* 2013, **20**(4)**:** 521--532.

Chapter 5 Machine learning generation of HFNAPs with nanomolar affinity to daunomycin

This chapter has been adapted from:

Chen, J. C., Chen, J. P., Shen, M. W., Wornow, M., Bae, M., Yeh, W. H., Hsu, A., & Liu, D. R. Machine learning-generated target-binding polymers from *in vitro* selection data. (*submitted*)

Contributions:

Professor David Liu and I initiated the project. I designed and conducted the major *in vitro* experiments including the *in vitro* selection, re-selection, and characterization experiments. I was assisted in these *in vitro* experiments by Minwoo Bae, Wei-Hsi Yeh, and Alvin Hsu. I designed and conducted the computational experiments, in conjunction with Jonathan P. Chen, Michael Wornow, and Max Shen.

5.1 Introduction

In vitro selection—iterative cycles of selection, amplification, and occasional mutagenesis on large combinatorial libraries—is a well-established technique[1,2] that enables the isolation of sequence-defined polymers with high levels of binding[3,4] or catalytic activity[5-9]. When *in vitro* selection was first developed, limited sequencing capabilities restricted the identification of selection outcomes to only a modest number of active sequences. Subsequent developments in high-throughput sequencing (HTS) and lowered sequencing costs have made it possible to reconstruct broader fitness landscapes, revealing broader relationships between sequence and activity. Even state-of-the-art sequencing technologies, however, are still limited to ~10^{10} reads, which precludes full sequencing of the typical 10-1,000 pmol ($6x10^{12}$ to $6x10^{14}$ molecules) starting libraries used for *in vitro* selections. As a result, attempts to map sequences onto the global fitness landscape are typically limited to regions of local space from late-stage selection rounds, where the high convergence and low sequence diversity makes comprehensive identification of active sequences possible[10-12].

Because most biopolymer fitness landscapes are rugged[12,13], with multiple local optima, characterization of the global fitness landscape requires identification of multiple, distinct, active variants[14]. To address this challenge, improvements to *in vitro* selections have increased the probability of finding active sequences in a random, combinatorial library[15-18] and made enrichment more efficient[19,20]. However, these approaches are still subject to limitations imposed by *in vitro* selection itself, sequence convergence, and limited sequencing depth. In the current study, we present a novel approach for discovering diverse, highly active sequences unrelated to any previously known variants: direct generation using a machine learning model trained on *in vitro* selection data. High-activity sequences generated by this machine learning model have no apparent sequence similarity to experimental sequences used to train the model, suggesting that the model has learned the fitness landscape sufficiently well to generate novel yet highly active variants.

The use of machine learning to enable functional synthetic polymer discovery presents new challenges. Many machine learning-aided protein evolution efforts have relied extensively on evolved, active protein sequences that are often incorporated directly into the resulting sequence predictions[21,22]. When starting from protein variants with low but non-zero levels of activity, machine learning can incorporate previously solved protein structures[23], or related protein sequences optimized over millions of years of natural evolution[24]. In contrast, *in vitro* selections start from naïve libraries[18,25], resulting in sparse and noisy datasets in which active examples are vastly outnumbered by inactive sequences. Sequence convergence during selection also limits the sequence diversity of active variants. Moreover, low-throughput hit characterization often limits the number of sequences for which activity can been measured. In the absence of comprehensive ground-truth activity measurements from labor-intensive biochemical assays, enrichment values are used as a proxy, introducing substantial measurement noise as the correlation between enrichment values and desired activity can be modest, influenced by the stochasticity of selections and activity-independent biases that occur during translation, selection, or PCR amplification[26,27]. Finally, limited sequencing depth restricts the ability to sample and observe the evolutionary trajectory of large numbers of input sequences, resulting in a noisy view of already noisy data.

Despite the lack of optimal training data, we envisioned a model that could learn aspects of a fitness landscape sufficient to generate active sequences unrelated to the limited diversity of sequences that can be identified from an *in vitro* selection. Generating active sequences in this manner is difficult because the space of possible outputs is very large, while the space of "correct" outputs is comparatively small. In this study, we address these challenges experimentally using judiciously chosen selection conditions and a chemically diverse starting library of HFNAPs to acquire data of sufficient quality to train a machine learning model. These selection improvements increased enrichment efficiency, enabling the discovery of highly active sequences that bind daunomycin, a small-molecule chemotherapeutic[28-30].

We then applied a generative machine learning approach to learn the joint probability distribution between HFNAP sequence space and daunomycin binding affinity. Our trained model successfully used data from a single *in vitro* selection to generate multiple diverse, highly active polymer sequences with no apparent sequence similarity to one another or to the sequences from the *in vitro* selection, in contrast with the tendency of individual selection campaigns to produce highly sequence-related hits. While the sequences may lack similarity, secondary structure conservation was identified between top-performing CVAE-generated sequences and selection-2 sequences, consistent with the machine learning model having achieved an understanding of the daunomycin-binding fitness landscape. These findings establish a novel integrated experimental and machine-learning approach to generate many sequence-defined polymer sequences with strong target binding activity, including those unrelated to any previously identified experimental variants, with broad implications for the discovery and application of non-natural biopolymers.

5.2 Conceptual overview of generative machine learning model methodology

Identifying active and inactive biopolymers beyond those closely related to current variants is a critical step towards mapping the global fitness landscape, and is a primary challenge of biopolymer evolution efforts[16]. For example, more comprehensive sequence space search can be achieved by increased experimental sampling. While multiple, parallel selections against the same target are likely to yield a more diverse set of active variants, each of these selections is still limited by the same sequencing constraints and sequence convergence that results from selection progress. Recently, generative machine learning models have risen to prominence for their ability to generate entirely new text[31], music[32], or faces[33] based on training examples. A generative machine learning model trained on HFNAP *in vitro* selection data could, in principle, learn the HFNAP fitness landscape for daunomycin binding. The resulting model could then directly generate many high-activity HFNAPs, in theory including those unrelated to

any sequences experimentally known to bind daunomycin, obviating the need for increased sampling or search depth.

To generate HFNAP sequences with daunomycin affinity, we used a variational autoencoder (VAE), a type of generative model commonly used to make predictions with biological sequence data[34,35]. VAEs attempt to model the observed data, HFNAP sequences, as a function of a latent or "hidden" variable (using Bayes' Theorem). The VAE model architecture begins with an encoder neural network that compresses the input sequence into a lower-dimensional space called the latent space. This dimensionality reduction from the input space to the latent space is critical because it forces the network to embed the latent space with only the information required for successful sequence reconstruction. The decoder network must then reconstruct the same input sequence by taking independent samples from this lower-dimensional representation, the latent probability distribution. During model training, the parameter weights for the two networks and the latent distribution are jointly optimized such that sequences can be accurately embedded into and reconstructed from the latent probability distribution. A visual example is shown in **Figure 5.1**, in which correct sequence reconstruction results in minimal changes to the model, while incorrect sequence reconstruction results in larger modification of the model to minimize the magnitude of the error. The trained model can then generate new HFNAP sequences by sampling from the latent probability distribution and using the generative decoder to produce new HFNAP sequences (**Figure 5.1**).

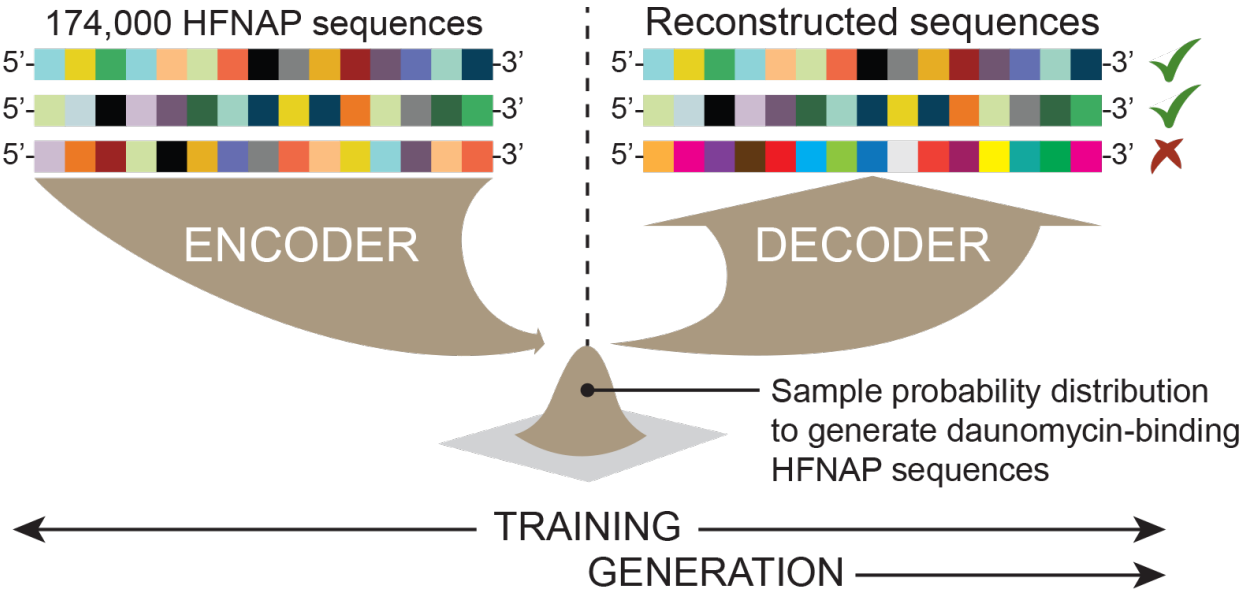


Figure 5.1 | Overview of VAE structure, training and sequence generation. Sequence inputs are encoded into a probability distribution. A sample of the probability distribution is given to the decoder, which must reconstruct the original input sequence. Failure to accurately reconstruct the sequence results in penalties, which are used to optimize the parameters of the encoder, decoder, and probability distribution to improve reconstruction accuracy. After training, the probability distribution is then sampled from and given to the decoder to generate new HFNAP sequences.

A trained VAE model could therefore generate HFNAP sequences with a broad range of affinities for daunomycin. In order to direct the model to generate only the most active sequences, however, the model requires data labels that indicate whether sequences are active or inactive. We thus moved from an unsupervised model (without labels), the VAE, to a semi-supervised model (with labels), known as a conditional variational autoencoder (CVAE)[36] (**Figure 5.2a**). To label the *in vitro* selection data, we defined a fitness value between 0 and 1 to represent an HFNAP's anticipated binding affinity for daunomycin based on the sequence's observed enrichment between round 4 and round 8b of selection 2 (see Methods, **Figure 5.2b**, **Figure 5.3**). The majority of the finesses are near 0, with ~2,000 sequences with fitness >0.25. The latent distribution and fitness value are provided as inputs to the decoder, which first uses the information to reconstruct the sequence (**Figure 5.2a**). The reconstructed sequence is then used as the input to an additional network, which predicts the fitness value given the sequence information. Because fitness prediction and sequence reconstruction are jointly optimized as

parts of the decoder, sequence information becomes conditioned on fitness values, which can associate fitness values with regions of sequence and latent space. Successful conditioning will enable the CVAE to generate novel HFNAP sequences with high predicted daunomycin binding affinity.

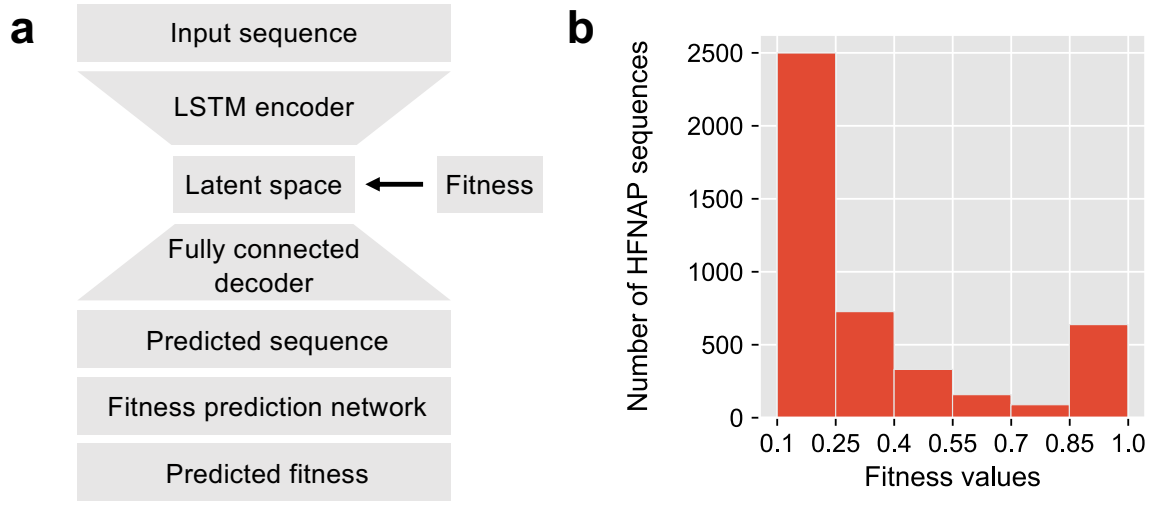


Figure 5.2 | CVAE architecture and fitness values from the *in vitro* selection. a, The CVAE architecture used in this study. Conditional VAEs introduce an additional, conditional variable, which enables the user to specify a specific condition or state. Here, we introduce sequence fitness, a value between 0 and 1 that represents the anticipated binding affinity for daunomycin based on the sequence's observed enrichment between round 4 and round 8b of selection 2, as the conditional variable (see Methods). During training, the fitness condition becomes associated with sequence identity and the latent distribution. Subsequently, the fitness condition enables the user to specify that the trained CVAE generates HFNAPs with desirable binding affinities. **b**, Distribution of fitness values assigned to training sequences that were greater than 0.1.

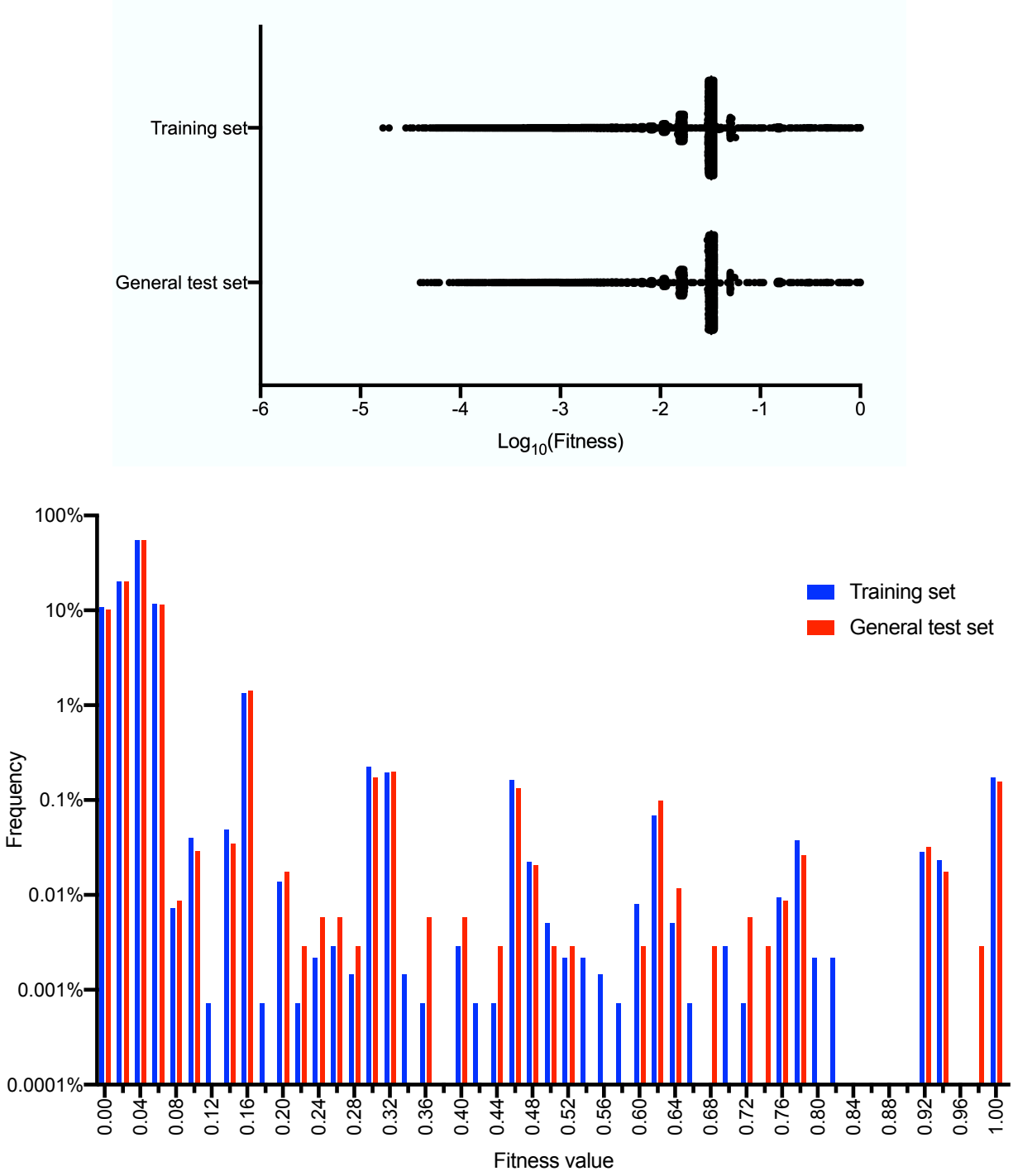


Figure 5.3 | Distribution of fitness values for training and general test set sequences. Selection-2 sequences were randomly assigned to the training set or general test set using an 80:20 ratio. Training set and general test set sequences are plotted against $\log_{10}$(fitness) and demonstrate comparable fitness distributions. The histogram of fitness values is then plotted with the bin center labeled on the X-axis. For example, the fitness value = 0.50 bin includes all sequences in the range: 0.49 ≥ fitness value > 0.51. Fitness values were assigned based on selection-2 enrichment values (see Methods), resulting in discontinuity in assigned fitnesses.

5.3 Conditional variational autoencoder (CVAE) architecture and model training

We performed extensive architecture search to arrive at a CVAE capable of successful sequence reconstruction and fitness prediction. The CVAEs were trained on the 172,545 HFNAP sequences that were identified between round 4 and round 8b of the selection 2. We used standard 80:20 train:test splits, in which we randomly withheld 20% of the data from training (**Figure 5.3**), then used the withheld sequences to test the performance of the model on unseen data. Initial progress on model training seemed promising, with >90% DNA reconstruction accuracy on the test set, with reconstruction accuracy defined as the percentage of accurate DNA backbone bases when comparing the correct HFNAP sequence to the HFNAP sequence predicted by the model (**Figure 5.4**). Since only 1% of the sequences were likely to be active, represented by the ~2,000 sequences with fitness >0.25 (**Figure 5.2b**), we were concerned the model may not be accurate for high-affinity sequences and might overemphasize inactive sequences, resulting in a failure to accurately approximate the HFNAP daunomycin binding landscape.

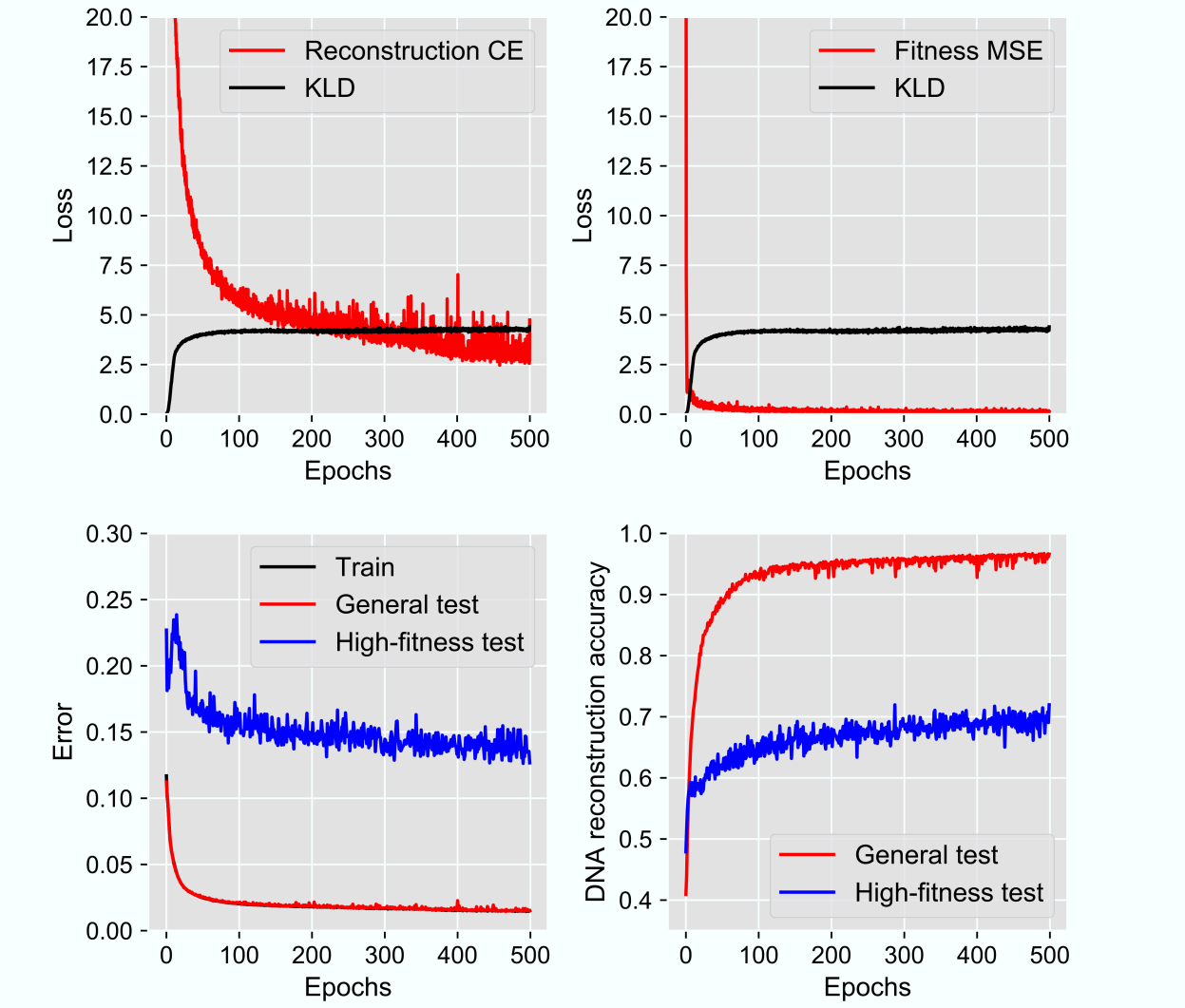


Figure 5.4 | CVAE performance of an early model. Total sequence reconstruction loss as calculated by cross entropy and fitness reconstruction loss as calculated by mean-squared error are shown. Model DNA reconstruction accuracy on 20% held out test set was high. Further analysis of the model reveals low DNA reconstruction accuracy for high-fitness sequences.

To mitigate this possibility, we modified our train:test splits and added an additional held-out high-fitness test set consisting of sequences with fitness >0.85. While this additional test set enabled us to specifically monitor model performance on 256 HFNAP sequences that we were confident had a high probability of daunomycin binding, it decreased the number of sequences with fitness values >0.85 available for model training by nearly half. These sequences were chosen based on having high round 4 to round 8b enrichment values in selection 2 (see Methods). Together, the general fitness test set, comprised mostly of low fitness values, and the

high-fitness test set, enabled us to more accurately observe changes to the model's sequence reconstruction and fitness prediction performance in response to CVAE architecture modifications. After extensive architecture search, we achieved >95% DNA reconstruction accuracy on the high-fitness test set and, importantly, 89% DNA reconstruction accuracy on the general test set when comparing the input DNA backbone sequences and reconstructed DNA backbone sequences (**Figure 5.5**).

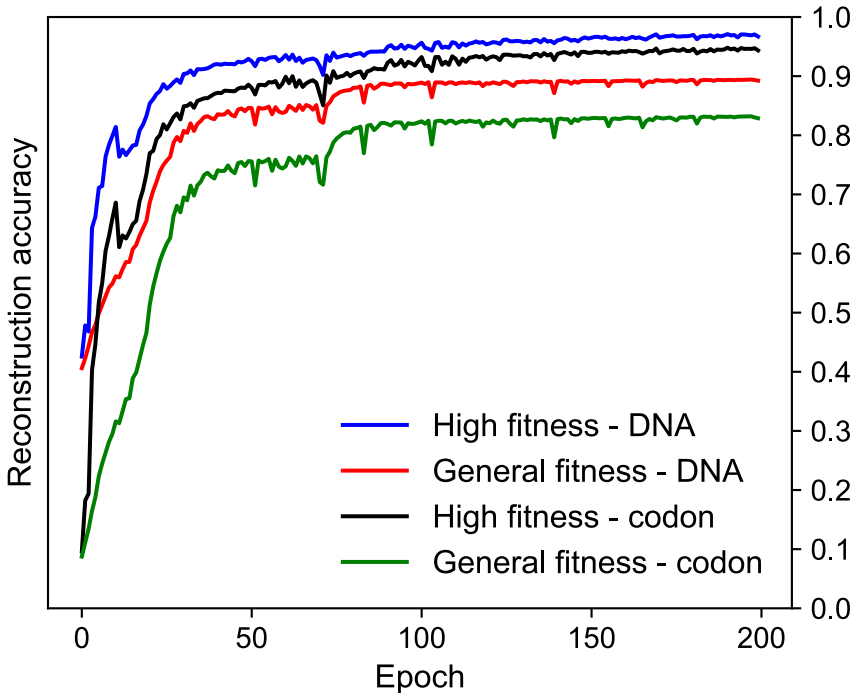


Figure 5.5 | Training performance of final model architecture. Reconstruction accuracy was used to assess training progress. The graph shows the reconstruction accuracies of the final, trained CVAE. Accuracy was calculated either by calculating the percentage of DNA bases that were correct, or the number of codons that were correct. e, Percent of generated sequences with at least 3-6 alcohol or amine side-chains. Each point along the x-axis represents 10,000 CVAE-generated sequences that were generated with that given fitness value as the condition.

Model performance on both test sets is important. **Figure 5.6** shows the sequence-converged nature of the high-fitness test sequences compared to the general-fitness test sequences. If the model had only achieved high accuracy on the sequence-converged, 256-member high-fitness test set and low accuracy on the general-fitness test set, the trained CVAE might only generate sequences similar to the high-fitness test set. Conversely, high accuracy on

the general-fitness test set and low accuracy on the high-fitness test set would likely translate to a failure to generate high-affinity sequences. The model's high reconstruction accuracy for both test sets suggests that it may have learned the HFNAP fitness landscape for daunomycin binding, which is further corroborated by the fitness prediction accuracy (Pearson correlation = 0.9999). The average mean squared error (MSE) between the assigned fitness and the predicted fitness for general test set sequences was 3.36×10^{-4}, which suggests high accuracy. A low value is unsurprising given that low fitness values make up a majority of the data, where 97% of the sequences have fitness values ≤0.05 (**Figure 5.3**). The model also performed well on the 318 sequences in the general test set with fitness >0.25, and on the 849 sequences with fitness >0.05, with the average MSE to be 8.26×10^{-3} and 1.00×10^{-2}, respectively. Fitness prediction on the high-fitness test set was also good, with an average MSE of 7.86×10^{-3}. The encouraging performance of the trained model on these test sets enabled us to move forward with further characterization of the model.

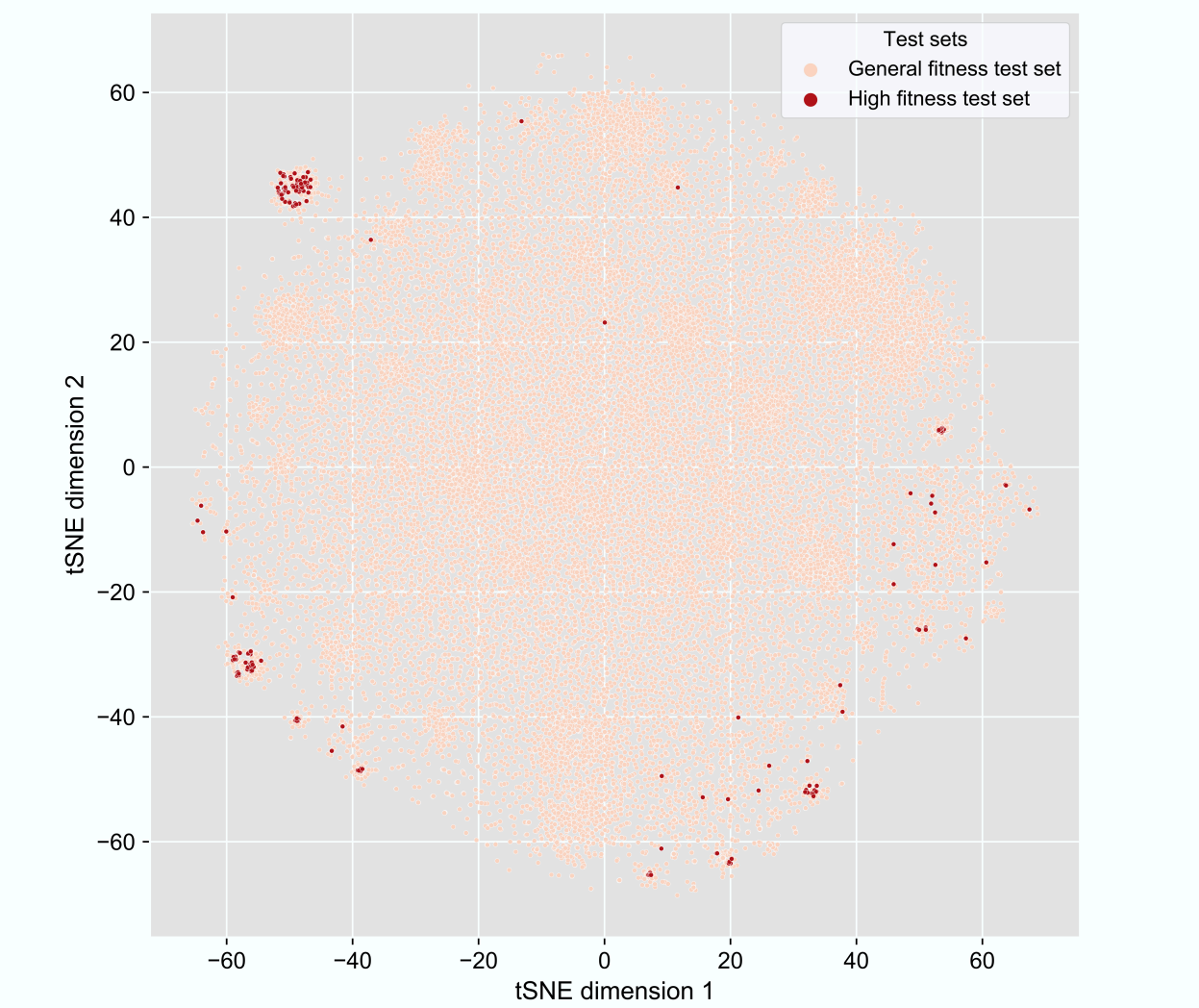


Figure 5.6 | tSNE plot of general fitness test set sequences and high fitness test set sequences. Members of the high fitness test set are converged, as evidenced by local clustering of sequences.

Our earlier analysis of the selection-2 results had indicated a strong preference for amine and alcohol side-chains among enriched sequences (**Figure 4.9**). Because HFNAP fitness is a function of the sequence, we expect that an accurate understanding of the fitness landscape should capture relationships such as this one. This preference for amine and alcohol side-chains can be further quantified in the selection data by looking at the proportion of sequences with at least 3-6 amine or alcohol sidechains as a function of our assigned fitness values (**Supplementary Figure 5.1**). To observe whether our CVAE captures a similar

relationship, we generated HFNAPs by sampling the latent distribution using the “reference set,” the top 3,072 HFNAP sequences from the training set with the highest fitness values. This data pre-processing step allowed us to discard poor-performing HFNAP sequences and focus on regions of latent space from the best-performing sequences. Sampling from this modified latent distribution and using the full range of fitness values to generate ten million new HFNAP sequences, we found that the model captured both the general frequencies and the overall rising then plateauing relationship between fitness and the number of amine or alcohol side-chains that was found in the training data (**Figure 5.7** and **Supplementary Figure 5.1**). These analyses indicated that the CVAE had learned a relationship between sequence and fitness, and suggested that the CVAE could be used to generate novel sequences with activity. Thus, we proceeded with experimental characterization, to determine if the CVAE had developed a sufficient understanding of the HFNAP fitness landscape for daunomycin binding to generate high-affinity sequences.

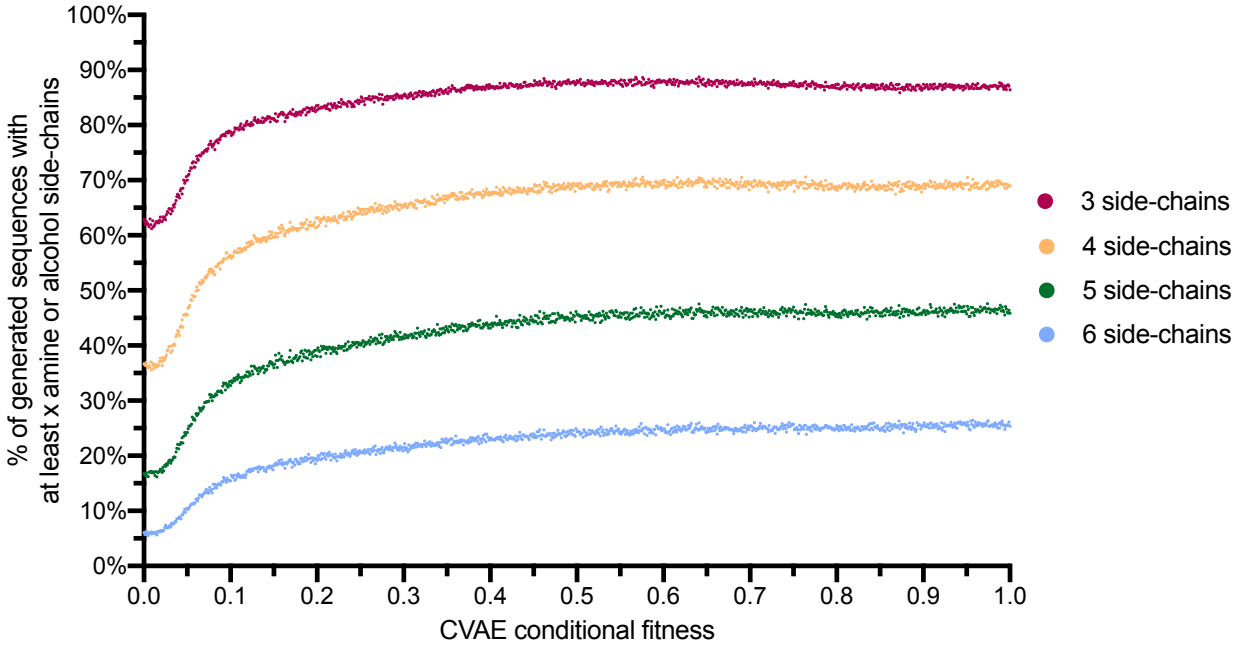


Figure 5.7 | Percent of generated sequences with at least 3-6 alcohol or amine side-chains. Each point along the x-axis represents 10,000 CVAE-generated sequences that were generated with that given fitness value as the condition.

5.4 Characterization of CVAE-generated HFNAPs

Satisfied with the CVAE's sequence reconstruction and fitness prediction performance on computational tasks, we next investigated whether the model could generate HFNAP sequences with experimentally measured binding affinity for daunomycin. To generate 10,000 HFNAP sequences, we specified the desired fitness values in increments of 0.05 ranging from 0.5 to 0.95 and sampled from the reference set latent distribution. The result was 1,000 sequences at each fitness value: 0.5, 0.55, … 0.95, constituting 10,000 CVAE-generated HFNAP sequences for experimental validation.

To choose sequences for low-throughput binding affinity characterization of individual HFNAPs, we identified groups of generated sequences with higher levels of sequence similarity. Although we sampled randomly from the latent distribution, our previous experiment demonstrates that CVAE-generated sequences exhibit a strong preference for amine and alcohol side-chain encoding codons (**Figure 5.7**). Any sequence similarities found in the generated sequences could arise from this codon preference, or represent regions of increased model confidence that reflect important factors relevant for daunomycin binding. To determine sequence similarity among the CVAE-generated sequences, we calculated the pairwise Levenshtein distance (LD)[37,38] between all 10,000 CVAE-generated HFNAP sequences (**Figure 5.8**). Levenshtein distance is the edit distance between two sequences using insertions, deletions, or substitutions; low Levenshtein distances indicate high sequence similarity.

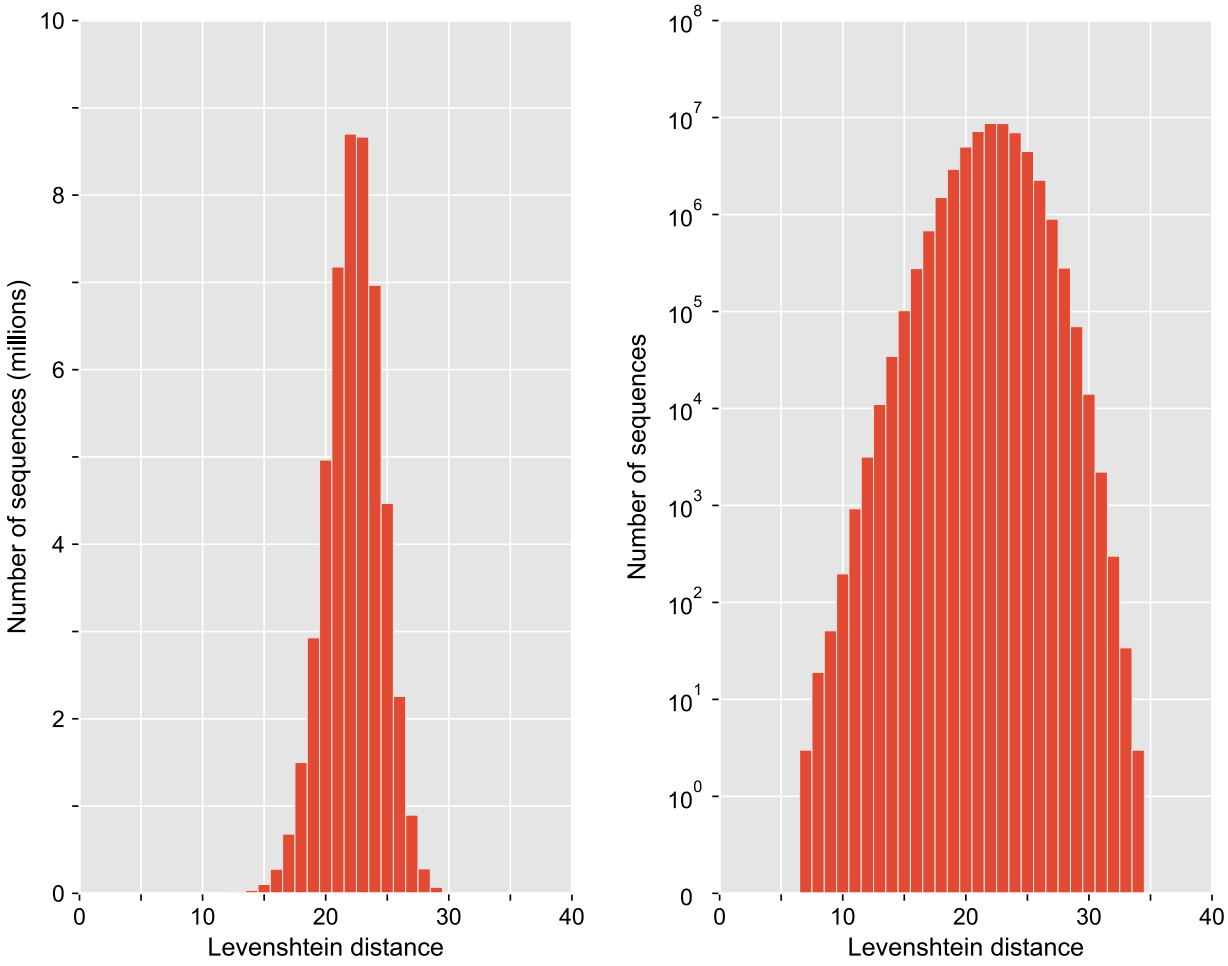
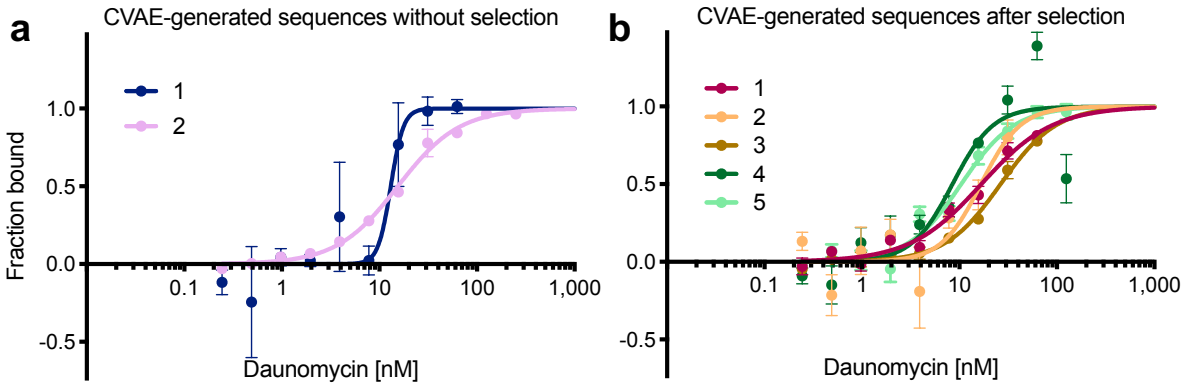


Figure 5.8 | Pairwise Levenshtein distances for 10,000 CVAE-generated HFNAPs. Pairwise Levenshtein distances were calculated to identify areas with increased sequence similarity.

Although most of the CVAE-generated sequences were distinct and characterized by distances similar to those of randomly sampled HFNAP sequences (**Supplementary Figure 5.2**), 136 sequences exhibited pairwise LD ≤9. It is tempting to speculate that these clusters of similar sequences reflect possible areas of model confidence. This subset of sequences characterized by pairwise LD ≤9 constituted about 1% of the generated HFNAPs, and can be further segmented into 128 single-pairs, six double-pairs, and two triple-pairs. A triple-pair is a generated HFNAP sequence that is related to three other generated HFNAPs by LD ≤9. Because triple-pairs may reflect local areas of model confidence, we chose the two triple-pairs

(eight sequences total) for further characterization. MST characterization indicated that two of these sequences bound with K_d=13-15 nM affinity (**Figure 5.9a**).



CVAE-generated sequences without selection	K_d
Dm-CVAE-without-selection-1	13.2 ± 3.2 nM
Dm-CVAE-without-selection-2	15.5 ± 1.3 nM

CVAE-generated sequences after selection	K_d
Dm-CVAE-after-selection-1	16.7 ± 6.0 nM
Dm-CVAE-after-selection-2	16.6 ± 14.0 nM
Dm-CVAE-after-selection-3	26.4 ± 6.9 nM
Dm-CVAE-after-selection-4	8.5 ± 2.3 nM
Dm-CVAE-after-selection-5	10.4 ± 1.5 nM

Figure 5.9 | MST characterization of the binding affinity of CVAE-generated daunomycin-binding HFNAPs. a, Two sequences identified directly without additional rounds of selection bind with K_d = 13-15 nM. Values and error bars reflect mean and SEM for n = 3 independent replicates. **b**, Five sequences identified after a single round of high-stringency selection were found to bind with K_d = 9-26 nM. Values and error bars reflect mean and SEM for n = 3 independent replicates.

To compare the binding affinity of model-generated HFNAPs to the top experimentally derived sequences from selection 2, we performed a competition selection. DNA templates for the 10,000 CVAE-generated HFNAPs and the top 2,000 HFNAPs from the training set with the highest fitness values (competition set) were translated as a mixture into HFNAPs, and subjected to a single round of high-stringency selection against daunomycin. Analysis of the HTS results indicated that of the 2,433 sequences that had enriched during the selection, 48.6% were from the CVAE-generated sequences and 51.4% were from the competition set. The 1,183 CVAE-generated sequences performed as well as, or better than members of the competition set in the single round of competitive selection (**Figure 5.10**). We individually

translated the five CVAE-generated HFNAP sequences with the highest enrichments for MST characterization and found them to bind daunomycin with 9-26 nM affinity (**Figure 5.9b**). These model-generated, high-affinity binders exceed or match the daunomycin-binding affinity of experimentally isolated sequences from selection 1 and selection 2 that were subject to competition with ≥10^{13} variants. These results show further evidence that our trained CVAE is capable of generating multiple validated HFNAPs with affinity for daunomycin comparable to that of HFNAPs that had survived multiple rounds of high-stringency selection.

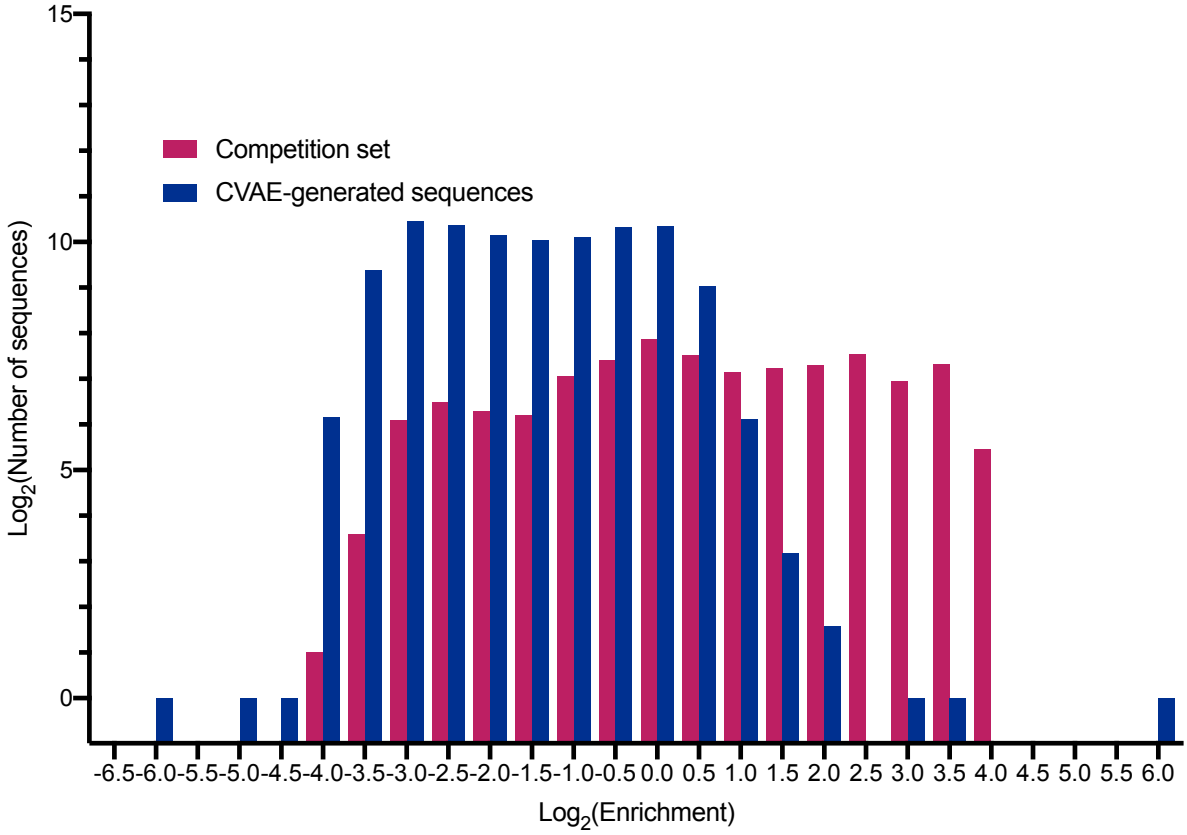


Figure 5.10 | High-stringency selection of mixture of 10,000 CVAE-generated HFNAPs and 2,000 most highly enriched sequences from round 4 to round 8b of selection 2 (competition set). Enrichment values and the number of sequences are shown on $\log_2$ scale, with enrichment calculated by dividing the frequency for that sequence at the 60 min elution sample by the pre-selection frequency. For competition set HFNAPs, 1,250 of the 2,000 sequences were identified with $\log_2$(enrichments)>0. For CVAE-generated HFNAPs, 1,183 of the 10,000 sequences were identified with $\log_2$(enrichments)>0. Two CVAE-generated sequences were identified in the elution but not in the pre-selection sequencing reads, and so approximate enrichment values were assigned using the rarest identified sequence frequency in the pre-selection sequencing reads. 19 CVAE-generated sequences that were found in the pre-

Figure 5.10 (Continued)
selection reads but not the elution reads, and 34 CVAE-generated sequences that were not found in either the pre-selection or elution reads, are omitted. A single competition selection sequence was not identified in either the pre-selection or elution reads and also omitted. X-axis bin values denote bin centers. For example, the $\log_2$(enrichment) = 0 bin includes all sequences in the range: $-0.25 \geq \log_2(\text{enrichment}) > 0.25$.

To ensure that the CVAE-generated HFNAPs were dependent on the model and not simply the observed amine and alcohol side-chain preference, we randomly sampled from the same building block frequency distribution as the reference set, the 3,072 top performing sequences from selection-2, to create the reference-based random set for a competition selection. DNA templates for the 10,000-member reference-based random set and the 2,000-member competition set were translated as a mixture into HFNAPs, and subjected to a single round of high-stringency selection for daunomycin binding. HTS results revealed that no sequence in the reference-based random set enriched, with the highest enrichment value being 0.69, indicating widespread de-enrichment (**Figure 5.11**). We individually translated five sequences with the highest enrichment values for MST characterization and found none with notable daunomycin binding activity (all had $K_d \geq 10$ μM) (**Supplementary Figure 5.3**). This result indicates that the building block distribution alone is not sufficient to confer daunomycin binding activity, and that the affinity of the CVAE-generated sequences is indeed the result of the model having learned a useful representation of the daunomycin fitness landscape.

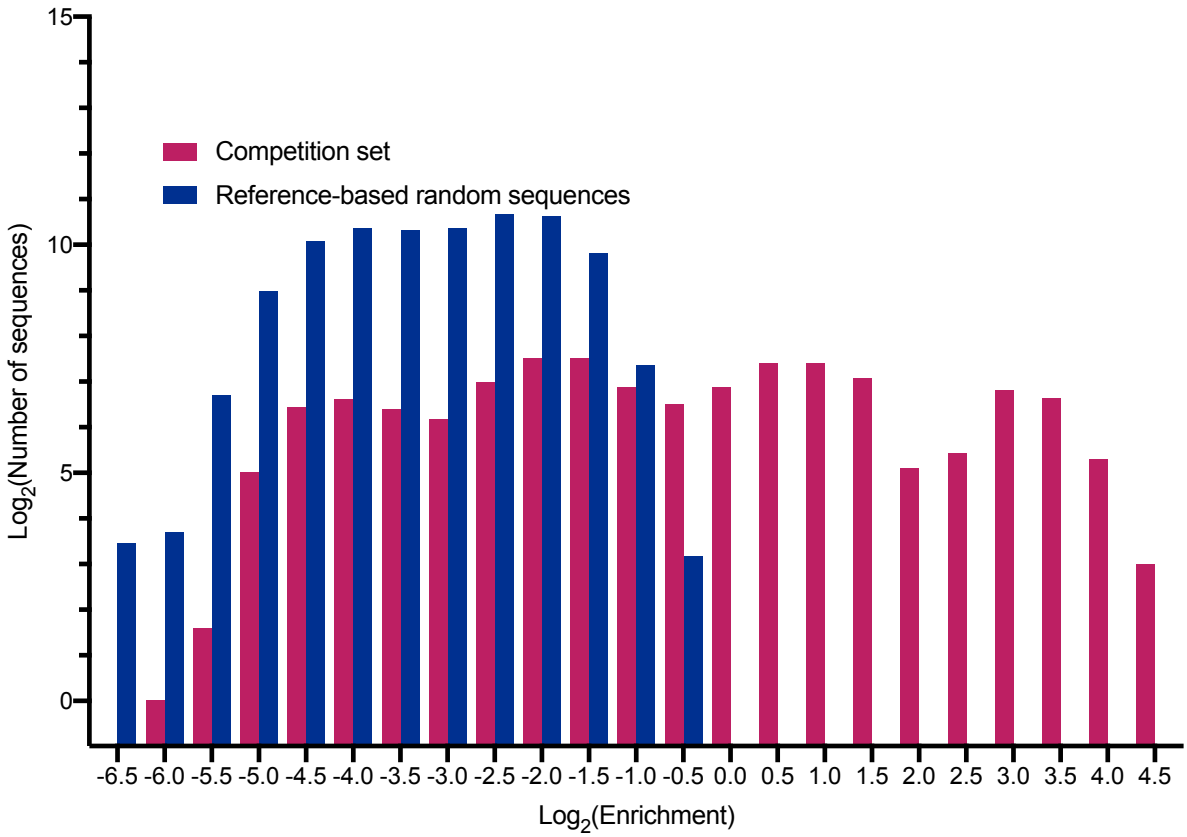


Figure 5.11 | High-stringency selection of mixture of 10,000 reference-based random HFNAPs and 2,000 most highly enriched sequences from round 4 to round 8b of selection 2 (competition set). Enrichment values and the number of sequences are shown on $\log_2$ scale, with enrichment calculated by dividing the frequency for that sequence at the 60 min elution sample by the pre-selection frequency. No reference-based random sequence enriched, as denoted by the lack of reference-based random sequences with $\log_2$(enrichment) ≥ 0. The highest observed enrichment value was 0.69 (linear scale value). X-axis bin values denote bin centers. For example, the $\log_2$(enrichment) = 0 bin includes all sequences in the range: -0.25 ≥ $\log_2$(enrichment) > 0.25.

5.5 Sequence diversity of CVAE-generated HFNAPs

Understanding the global fitness landscape is critical to achieving multi-objective optimization for properties such as stability, binding affinity, and selectivity. To populate unknown regions of the global fitness landscape, model-generated sequences should be unrelated to known sequences. Establishing the degree of dissimilarity between model-generated active sequences and subsets of the training data also demonstrates whether the model has learned a fitness landscape that is distal to the local fitness landscape identified from

selection 2. Here, we use Levenshtein distance and dimensionality reduction to determine if CVAE-generated HFNAP sequences are dissimilar to the reference set, establishing that the CVAE had learned sequence determinants of daunomycin activity beyond sequence similarity to reference set sequences, and also dissimilar to each other, establishing that the CVAE was capable of identifying diverse and distinct regions of active sequence space.

To determine whether the CVAE-generated daunomycin-binding HFNAPs are related to the reference set, we used the Levenshtein distance as our metric of sequence similarity and visualized sequences in two dimensions using dimensionality reduction by UMAP[39]. The UMAP plot visually demonstrates that the seven validated CVAE-generated HFNAP sequences are dissimilar from the reference set, as they remain separated from the clusters (**Figure 5.12**). In contrast, the seven validated selection-2 sequences are found within large clusters, indicating sequence similarity resulting from selection induced sequence convergence. The distribution of Levenshtein distances to the reference set can be visualized directly, where we observe the seven CVAE-generated daunomycin-binding HFNAPs are separated from all 3,072 reference set sequences by a Levenshtein distance of 15 or more (**Figure 5.13**, mean $LD_{reference\ set}$ = 22.9 ± 1.21, n=7) and from all training sequences by a Levenshtein distance of 12 or more (**Supplementary Figure 5.4**, mean $LD_{training\ set}$ = 23.20 ± 2.05, n=7), indicating large dissimilarity. The distance distribution from CVAE-generated sequences to the reference set are similar to ones from seven reference-based random HFNAP sequences to the reference set (mean $LD_{reference\ set}$ = 24.1 ± 0.78, n=7) and from seven truly random HFNAP sequences to the reference set (1/32 building block distribution, mean $LD_{reference\ set}$ = 24.9 ± 0.64, n=7), with the closest sequence being 17 and 16 LD away, respectively (**Figure 5.13**). In contrast, 39% of the reference set is within a Levenshtein distance of only 1-2 to at least one of the seven validated selection-2 daunomycin-binding HFNAPs (**Figure 5.13**, mean $LD_{reference\ set}$ = 20.2 ± 1.90, n=7). These data demonstrate that the CVAE-generated daunomycin-binding HFNAPs are unrelated

to the reference set, and provides further support that the model has learned the daunomycin-binding fitness landscape distal to the local fitness landscape from selection 2.



Figure 5.12 | UMAP projection (a non-linear method of dimensionality reduction) of reference set, selection-2, and CVAE-generated daunomycin-binding HFNAP sequences. 3,072 reference-based random sequences and 3,072 truly random sequences were included in the distance matrix and embedding but omitted from the final UMAP plot for clarity. Reference set sequences form clusters, indicating sequence similarity. Selection-2 daunomycin-binding HFNAPs, which are sequence related to the reference set, are found within these clusters. The selection-2 daunomycin-binding HFNAPs are Dm-HS-1, 2, 3, 6, 7, 8, and 10. Dm-HS-4 and Dm-HS-5 were omitted for sequence similarity to Dm-HS-1, and Dm-HS-9 was omitted for sequence similarity to Dm-HS-3.

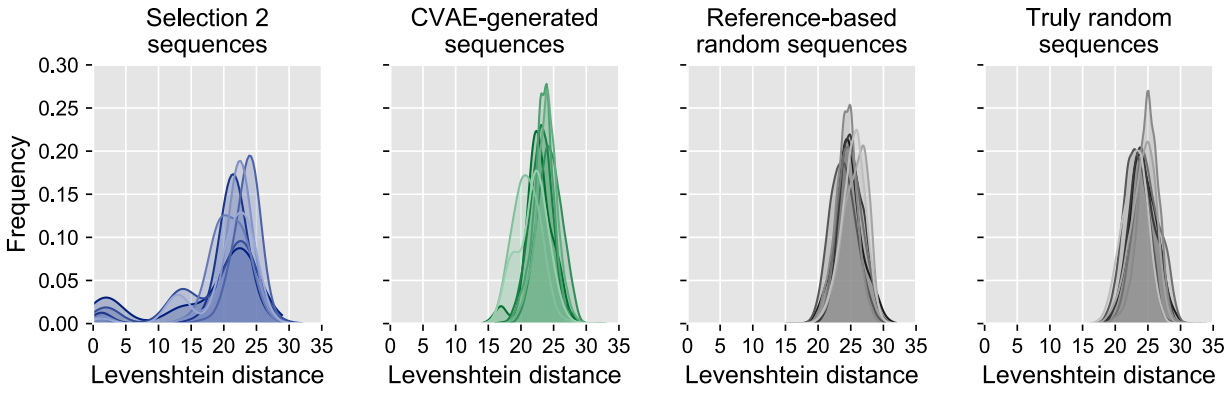


Figure 5.13 | Comparison of the Levenshtein distances from the 3,072-member reference set to the top seven sequences from each of the four groups. Reference-based random sequences are HFNAP sequences randomly sampled using the same building block frequency as the reference set. Truly random sequences are HFNAP sequences randomly sampled using a uniform building block frequency distribution of 1/32. Selection-2 daunomycin-binding HFNAPs include Dm-HS-1, 2, 3, 6, 7, 8, and 10. Dm-HS-4 and Dm-HS-5 were omitted for sequence similarity to Dm-HS-1, and Dm-HS-9 was omitted for sequence similarity to Dm-HS-3.

Similar relationships are maintained, even when comparing Levenshtein distances between all 10,000 CVAE-generated sequences and the reference set (**Supplementary Figure 5.5-5.6**). Using the Jensen-Shannon (JS) distance[40,41] between the resulting distributions, we found the distributions calculated from the CVAE-generated HFNAPs to be closer to the distribution from the reference-based random sequences (0.333 JS distance) and the distribution from the truly random sequences (0.569 JS distance) than the distribution from the top 10,000 training sequences by fitness (0.616 JS distance). Additionally, the closest Levenshtein distances for the 10,000 reference-based random sequences and the 10,000 truly random sequences to the reference set are 11 and 12, respectively (**Supplementary Figure 5.5**). We find that only 16 and 43 of the 10,000 CVAE-generated sequences exhibit Levenshtein distances to the reference set that are <11 or <12, respectively. The dissimilarity between CVAE-generated HFNAPs and the reference set indicates the model is likely generating HFNAPs based on an understanding of the broad fitness landscape, rather than relying on sequence similarity to the reference set.

To assess whether the validated CVAE-generated daunomycin binding HFNAPs are derived from a single local fitness landscape as opposed to different areas of the global fitness landscape, we compared the pairwise Levenshtein distances between experimentally-derived and computationally-derived active HFNAPs. The seven validated CVAE-generated daunomycin-binding HFNAPs have a median pairwise distance of 23 LD in contrast to the median pairwise distance of 21 LD for the selection-2 validated sequences used in **Figure 5.12** (**Supplementary Figure 5.7**). We can then use the Mann-Whitney U test to compare the resulting distance distributions to determine whether these distributions are dissimilar (indicated by low U statistic values and small P values). We find that the validated CVAE-generated HFNAPs' distance distribution is different from the validated selection-2 HFNAPs' distance distribution ($U=652$, $p_{two\text{-}tailed}=0.0378$, $n_1=n_2=42$). This result suggests that the model produces active HFNAPs with sequence diversity exceeding what we find from our experimental *in vitro* selection. While model-generated diversity does not match that of the two random controls, we find that the selection-2 HFNAPs' distance distribution is more dissimilar to these random controls than the CVAE-generated HFNAPs' distance distribution, as indicated by lower U statistic values and smaller P values for both cases (**Supplementary Figure 5.7**). Extending the pairwise distance calculations to all 10,000 CVAE-generated HFNAPs, we observe similar relationships qualitatively (**Supplementary Figure 5.8**). The pairwise distance distribution between the validated CVAE-generated daunomycin-binding HFNAPs exhibits sequence diversity exceeding that of experimentally derived sequences from *in vitro* selection, which further demonstrates that the CVAE is able to identify distinct areas of the fitness landscape.

5.6 CVAE-generated sequences conserve selection-2 secondary structures

To begin to understand the CVAE's ability to generate active, daunomycin-binding HFNAPs despite the lack of apparent sequence similarity with HFNAPs from *in vitro* selection, we analyzed the training data for conserved secondary structures that may be responsible for

activity. Because Dm-HS-1 dominates the training data, with 23.5% of the top 1,000 sequences by fitness within 5 LD, we reasoned that this sequence family was a good starting point for secondary structure analysis. Using the RNAstructure suite of programs[42], we ran multilign on the ten most abundant Dm-HS-1 related sequences (inclusive, abundance at end of selection 2, round 8). Multilign predicts secondary structure conservation in multiple oligonucleotide sequences[43]. The aligned structures were then compared with the individually folded structures, yielding a set of four secondary structures for Dm-HS-1 that constitute the possible conserved secondary structures (**Figure 5.14**). These structures include the minimum free energy structure (MFE) (**Figure 5.14**, Structure 1), and were subsequently compared with the secondary structures from CVAE-generated sequences.

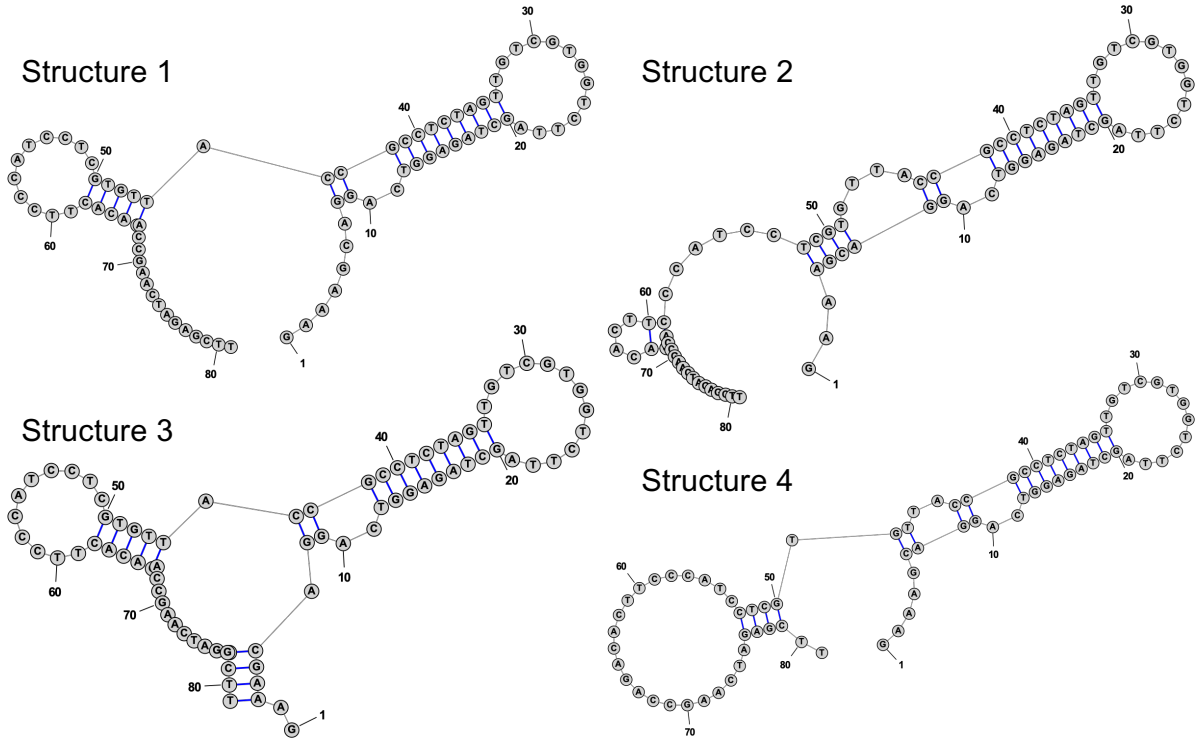


Figure 5.14 | Conserved secondary structures for Dm-HS-1. Dm-HS-1 was folded using RNAstructure with DNA parameters, and related sequences were folded and structures analyzed for conservation both manually and computationally via multilign. Structure 1 is the MFE.

The ten most enriched CVAE-generated sequences and ten most enriched reference-based random sequences from the respective competition selections were individually folded using RNAstructure's Fold program. Because folding often yields a variety of different structures, all structures generated from the standard settings were compared with the four Dm-HS-1 conserved secondary structures found in **Figure 5.14**. These secondary structures were then converted to dot-bracket notation, and base pairing distance was analyzed by RNAdistance, a package within ViennaRNA[44,45]. Base pair edit distance quantifies the number of edits required to go from one structure to the other. For each CVAE-generated sequence or reference-based random sequence, we identified the closest secondary structure to any of the four conserved Dm-HS-1 structures (**Figure 5.15a**), or averaged the closest distance to each of the four conserved Dm-HS-1 structures (**Figure 5.15b**). Data containing every distance comparison is shown in **Figure 5.15c**. These distance comparisons demonstrate that the secondary structures of CVAE-generated sequences are closer to Dm-HS-1 conserved secondary structures than the secondary structures of reference-based random sequences are.

The quantitative distances values are also supported by visual inspection of the structures. We examined the top two closest structures from each group. Dm-CVAE-after-selection-3 (**Figure 5.16**) is 24 base pair edit distance from Dm-HS-1 conserved structures 1 and 2 (**Figure 5.14**) and captures similar stem loops on both sides. Dm-CVAE-after-selection-5 (**Figure 5.16**) is 30 base pair edit distance away from Dm-HS-1 conserved structure 3 (**Figure 5.14**), maintaining all three base pairing regions in a similar fashion. However, when considering reference-based random sequences, Dm-reference-based-random-after-selection-2 structure 2 (**Figure 5.16**) is 32 base pair edit distance away from Dm-HS-1 conserved structure 2 (**Figure 5.14**) but lacks a 3' stem loop as the 3' end is unstructured. While Dm-reference-based-random-after-selection-3 structure 1 (**Figure 5.16**) is 38 base pair edit distance away from Dm-HS-1 conserved structure 3 (**Figure 5.14**), the 5' stem loop contains some structural differences. Despite the lack of sequence similarity, these CVAE-generated sequences fold into secondary

structures that resemble those of Dm-HS-1, the most dominant family of sequences in our high-stringency *in vitro* selection. Secondary structure similarity may explain the performance of CVAE-generated sequences in the competition selection, which suggests that the CVAE's understanding of the HFNAP fitness landscape for daunomycin binding may be derived from an understanding of secondary structure motifs present within high-fitness sequences.

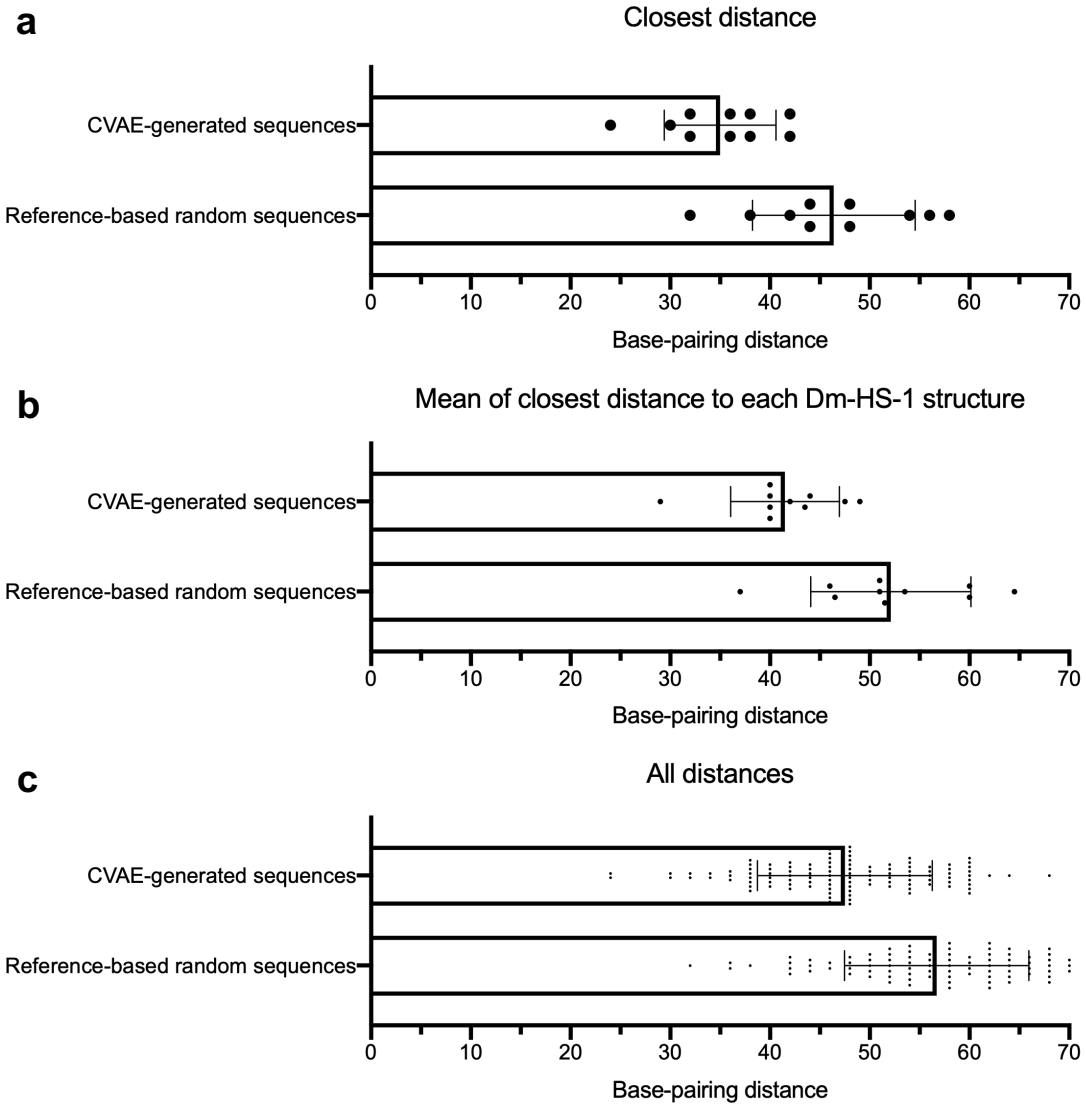
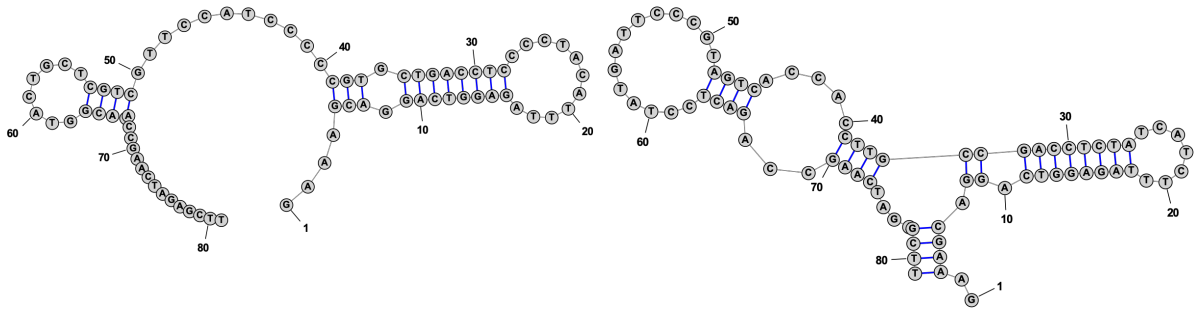
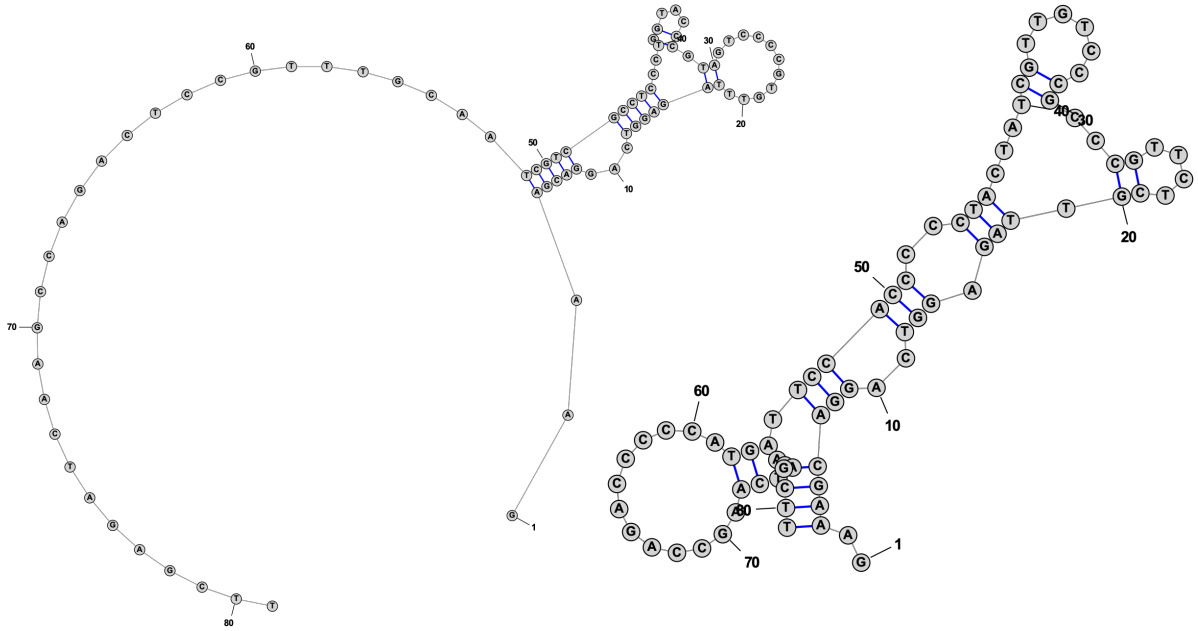


Figure 5.15 | Base pair distance between Dm-HS-1 conserved secondary structures and CVAE-generated sequences or reference-based random sequences from respective competition selections. a, The structure from each CVAE-generated sequence with the closest distance to any of the four conserved Dm-HS-1 structures is shown. **b**, The structures from each CVAE-generated sequence with the closest distance to each of the four conserved Dm-HS-1 structures are averaged and shown. **c**, All distances are plotted. Error bars denote standard deviation in all panels. Reference-based random sequences are analyzed similarly.



Dm-CVAE-after-selection-3

Dm-CVAE-after-selection-5



Dm-reference-based-random-after-selection-2

Dm-reference-based-random-after-selection-3

Figure 5.16 | Secondary structures with the two closest distances to conserved Dm-HS-1 structures. Dm-CVAE-after-selection-3 structure 2 is 24 base pair edit distance away from Dm-HS-1 conserved structures 1 and 2. Dm-CVAE-after-selection-5 structure 1 is 30 base pair edit distance away from Dm-HS-1 conserved structure 3. Dm-reference-based-random-after-selection-2 structure 2 is 32 base pair edit distance away from Dm-HS-1 conserved structure 2. Dm-reference-based-random-after-selection-3 structure 1 is 38 base pair edit distance away from Dm-HS-1 conserved structure 3.

5.7 Discussion

The global fitness landscapes of functional sequence-defined polymers such as proteins or nucleic acids are traditionally considered too large and rugged to be captured by an algorithm. In the present study, we trained a generative machine learning model on HFNAP *in*

vitro selection data to generate diverse HFNAPs with high binding affinity for the small molecule daunomycin. By leveraging the chemical functionality of HFNAPs and improving selection conditions, we produced a modest dataset of 10^5 experimental sequences to train our conditional variational autoencoder (CVAE), a type of generative machine learning model. CVAEs can generate useful, unique outputs because CVAE training results in an optimized posterior latent probability distribution. In the current study, our CVAE learned the joint probability distribution between sequence identity and binding affinity and then directly generated HFNAPs, unrelated to experimentally identified sequences, that bound daunomycin with affinity as high as K_d=13-15 nM.

Notably, the CVAE-generated sequences performed just as well as the sequences emerging from multiple rounds of high-stringency *in vitro* selection, despite the lack of sequence similarity between the CVAE-generated and experimental polymers. The diversity of CVAE-generated sequences counteracts the sequence convergence typical of traditional *in vitro* selections, thereby increasing the likelihood of identifying candidate sequences with diverse properties including those that may not have been explicitly selected, but which are desirable. For example, increasing levels of backbone and 2'-position modifications usually reduces on-target activity[46,47], but having diverse sequences may identify sequences more amenable to chemical modification. Because CVAE-generated sequences contain structural motifs similar to conserved secondary structures from selection 2, our method enables further expansion of sequence and structural diversity, while maintaining high on-target activity. Together, these results demonstrate that generative machine learning models can be used to directly generate active polymer sequences and to reduce the number of *in vitro* selections required to discover diverse and highly active variants.

Researchers continue to recognize the power of applying machine learning techniques to address a wide range of life sciences questions, to make predictions about biological data[48,49], and to improve protein function by efficiently exploring fitness landscapes[21,24]. Notably,

Biswas and colleagues have made impressive progress in successfully limiting the amount of data a model needs to be trained on to achieve predictive power[24]. Pre-training models on the ~24 million UniRef50 amino-acid sequences allows them to learn fundamental features necessary for protein function. The resulting models can then be fine-tuned using <100 functional variants of the desired protein, minimizing the amount of labeled data that needs to be individually characterized. In addition, Stokes and co-workers showed that message passing neural networks trained on small molecules can guide drug discovery by making predictions on new, unrelated compounds[48]. Collectively, these examples illustrate the importance of training on useful datasets to achieve predictive power. We expect that advances in model construction and training will only continue to improve researchers' abilities to generate useful predictions that can outperform training examples.

More broadly, this work demonstrates that integrating *in vitro* selection with machine learning can help researchers explore a much larger swath of a polymer's total sequence space than would be possible from iterated *in vitro* selections alone. The ability of this approach to directly generate many unrelated sequences with high on-target activity is an important advance that increases the likelihood of discovering a polymer with a desired combination of attributes suitable for biotechnological or therapeutic applications. By making better use of the abundance of information-rich HTS data generated by *in vitro* selections than traditional approaches—which use only a small fraction of this data to identify enriched clones and putative consensus sequences—we may be better poised to discover active sequences that survive difficult selection criteria with few solutions. Our integrated *in vitro* selection and machine learning approach may also facilitate the generation of synthetic polymer catalysts with activities not known among previously described enzymes and ribozymes.

5.8 Methods

General methods. Oligonucleotides were purchased from IDT and are listed below. Building blocks were synthesized as described previously[16,50]. Building blocks were synthesized either directly from phosphoramidites (Glen Research) or via coupling of an amine to the NHS ester containing phosphoramidite as described previously[16,50]. T3 DNA ligase, ATP, and T4 RNA ligase buffer were purchased from New England BioLabs. MyOne Streptavidin C1 magnetic beads (Life Technologies) were used in immobilization. MST samples were prepared using HBS-P+ buffer (GE Healthcare Life Sciences).

Synthesis of HFNAP libraries by DNA-templated translation. DNA template libraries (125 pmol), initiation primer (188 pmol), termination primer (188 pmol), monomer mix (1.25 nmole/monomer), and 13 µL of water were mixed and added to a PCR strip. Strips were then incubated at the following temperatures using a thermal cycler: 95 °C for 10 s, 65 °C for 4 min, followed by a 0.1 °C per 10 s ramp to 4 °C. 6.25 µL of ATP and 6.25 µL of T3 DNA ligase were then added at 4 °C, and then incubated for 16 h at 4 °C.

250 µL of streptavidin beads were then prepared by washing in B&W buffer according to manufacturer protocol. The beads were then changed into 125 µL of 2× B&W (20 mM Tris-HCl, 2 M NaCl, 2 mM EDTA, pH 7.5). The translation reaction was then combined with the streptavidin beads and mixed for 1 h on a rotary mixer. The supernatant was discarded, and beads were washed thrice to remove nonspecific binders. The HFNAP strands were then selectively eluted from the biotinylated template using two 5-min washes with 250 µL of freshly prepared 20 mM NaOH. The combined elution was mixed with 5 volumes of column-binding buffer (2:3 sat. aq. guanidine•HCl:isopropanol, 1% v/v 3 M NaOAc, 0.1% pH indicator). The combined sample was then column purified using a QiaPrep 2.0 column and eluted with 52 µL of EB (Qiagen). A 1-µL aliquot was saved as the pre-selection library.

Subsequent rounds of translation require streptavidin bead immobilization of double stranded DNA templates and 20 mM NaOH washes to obtain ssDNA templates for translation.

Beads are then washed twice with T4 RNA ligase buffer to neutralize any remaining base before proceeding with synthesis and translation as usual. These bead immobilized templates are then used in lieu of in solution templates.

Synthesis of individual HFNAPs by DNA-templated translation. Synthesis of individual HFNAPs follows a similar protocol as library synthesis. The major difference is use of 10 equivalents of the building blocks for each HFNAP as opposed to a master mix. Following translation, HFNAP isolation and elution, and column cleanup, the sample is then PAGE purified on a denaturing 10% TBE-Urea gel (Criterion, Bio-Rad). The band is visualized by UV shadowing, where DNA bands were identified by the shadow cast by DNA onto a TLC plate with F254 indicator when the gel was illuminated by a UV lamp. The bands were excised from the gel and extruded through stacked microcentrifuge tubes (0.2 mL, 0.5 mL, 1.5 mL), with 17 gauge and 27 gauge holes made in the first two tubes respectively. The resulting material was then resuspended in 400 µL of Crush Soak (1× TE, 200 mM NaCl), and then shaken overnight at 37 ºC.

The gel pieces were then removed using 0.22 µm PVDF centrifugal filters (Millipore). The remaining liquid was then combined with 5× equivalents of column binding buffer, DNA column purified, and eluted with water.

Daunomycin affinity selection. Daunomycin-linked beads and negative selection beads were prepared prior to the start of each selection round. Streptavidin-linked beads were washed with B&W according to manufacturer protocol, then incubated with biotinylated daunomycin or the negative selection linker (see Methods) for 1 h. The beads were then washed in selection buffer (20 mM Tris-HCl (pH 7.4), 140 mM NaCl, 5 mM KCl, 1 mM $MgCl_2$, 1 mM $CaCl_2$, 0.05% [v/v] Tween 20) thrice, and then resuspended in 2× selection buffer.

The selection began with incubation of column purified HFNAP libraries with immobilized daunomycin on streptavidin-linked beads. The selection was placed on a rotary mixer for 1 h. The flow-through was then collected, the beads washed 3× using selection buffer, and finally putative binders were eluted according to the scheme in **Figure 4.7**. When eluting with 1 mM daunomycin, the beads were washed twice with selection buffer after the initial incubation with daunomycin. The combined material constitutes the eluted material. The eluted material was then mixed with 5× equivalents of column binding buffer, DNA column purified and eluted with water.

Selections started with a negative selection after round 4. The column-purified HFNAP libraries were mixed with negative selection beads on a rotary mixer for 30 min. The flow-through was then collected, and mixed with positive selection, daunomycin-linked beads for 1 h.

A 1-μL aliquot of each sample was analyzed by qPCR using Q5 Hot Start High Fidelity 2× Master Mix (New England BioLabs) and 0.5× SYBR Green I (Life Technologies). The qPCR cycle number was used to determine the number of cycles of amplification for the eluted material. The eluted material was amplified by PCR using Q5 Hot Start High Fidelity 2× Master Mix (New England BioLabs), and a primer set in which one of the primers was biotinylated. The resulting PCR reaction was then DNA column-purified using 5× equivalents of column binding buffer, and eluted with water.

The eluted material was then gel purified in a 10% TBE native PAGE gel (Criterion, Bio-Rad). The gel was visualized by UV shadowing, where DNA bands were identified by the shadow cast by DNA onto a TLC plate with F254 in which when the gel was illuminated by a UV lamp. The bands were excised from the gel and extruded through stacked microcentrifuge tubes (0.2 mL, 0.5 mL, 1.5 mL), with 17-gauge and 27-gauge holes made in the first two tubes respectively. The resulting material was then resuspended in 400 μL of Crush Soak (1× TE, 200 mM NaCl), and then shaken overnight at 37 °C.

The gel pieces were then removed using 0.22 μm PVDF centrifugal filters (Millipore). The remaining liquid was then combined with 5× equivalents of column binding buffer, DNA column purified, and eluted with EB. The absorbance was then measured to determine the concentration of the eluent, for subsequent rounds of translation.

High stringency selection scheme: The two main differences for the high stringency selection scheme were the selection temperature and the elution conditions. The positive and negative selections were both performed at 37 °C, with an initial 2-min incubation in a thermocycler to rapidly raise the temperature of the selection to 37 °C prior to incubation on a rotary mixer in a 37 °C warm room. The elution condition was then modified to contain multiple washes with 1 mM daunomycin, followed by 2 washes with selection buffer. The incubation times and the elution that was carried forward to the next round of selection are shown in **Figure 4.7**.

HTS analysis of libraries. Samples to be analyzed were amplified by PCR using Q5 Hot Start High-Fidelity 2× Master Mix to install Illumina sequencing adapters, with amplification cycle number determined by qPCR. Samples were then DNA column purified, followed by an additional PCR round to install Illumina barcodes. The resulting samples were then combined, DNA column purified, and the concentration was determined by Kapa Library Quantification Kit (Kapa Biosystems). Fastq files were then parsed using AptaSuite, setting a strict cutoff for full-length sequences[51]. The parsed sequences were exported through AptaSuite in Fastq format, and subsequently converted into csv's for downstream analysis.

Microscale Thermophoresis (MST). MST analysis was conducted using a Monolith NT.Automated (Nanotemper). Analyses were conducted at 25 ºC in HBS-P+ buffer (Cytiva/GE Healthcare Life Sciences) with additional 5 mM KCl, 1 mM $MgCl_2$, and 1 mM $CaCl_2$. Sequences were translated using a Cy5 tag and PEG18 spacer, for use in tracking the HFNAP. Generally,

the laser excitation energy was set to 10 % to minimize photobleaching, and the IR laser power was set to high for all readings.

Dilution series for each replicate was made separately, using serial dilutions from the highest concentration to the lowest concentration. The amount of labeled HFNAP that was used was subsaturating, as determined by fluorescence intensity measured against a known standard. Measurements were made in triplicate. Data was processed and fitted using standard Nanotemper software (MO.Affinity) and exported to Prism for plotting.

Machine Learning Model

To gain further insight into the fitness landscape of daunomycin binding, we applied deep learning techniques to model the HFNAP sequences from selection 2.

Training dataset. The training dataset was constructed from HTS analysis of the high stringency selection, round 4 through round 8b. We bootstrapped the dataset to account for differences in read depth for each sample and to normalize the number of reads per sample. Data for each round was randomly sampled 1 million times with replacement, and this process was iterated 1,000 times to calculate bootstrapped sequence abundances. The rounds were then combined using a series of merges, and missing values were assigned an abundance of 1.

The fitness value was then calculated by dividing the round 8b abundance by the round 4 abundance, and then dividing the resulting enrichment value by 20. A ceiling was introduced such that enrichment values greater than 1 were then adjusted to 0.999. This gave us ~2000 sequences with fitness greater than 0.25, which seemed to align with what we might expect given the sequence similarities within the top 2000 sequences. The 256-member high-fitness test set was chosen by ranking the sequences by enrichment (round 4 to round 8b). Every other highest-ranking sequence was included in the high-fitness test set, which prevented withholding the best 256 sequences from the data.

Model construction and training. To learn the relationship between HFNAP sequence space and daunomycin binding, we begin by positing that the relationship is a function of a hidden (latent) variable z. To learn the posterior distribution $p(z|x)$ over the latent z, we can apply Bayes Theorem given our observed data x.

$$p_\theta(z|x) = \frac{p_\theta(x|z)p_\theta(z)}{p(x)} = \frac{p_\theta(x,z)}{p(x)}$$

However, computing $p_\theta(x)$ is intractable because it scales exponentially with the size of the latent space.

$$p_\theta(x) = \int p_\theta(x,z)dz = \int p_\theta(x|z)p_\theta(z)dz$$

Instead, we can use variational inference to approximate the posterior $p_\theta(z|x)$, resulting in a variational distribution q with parameters ϕ[34,52].

$$p_\theta(z|x) \approx q_\phi(z|x)$$

Because approximation $q_\phi(z|x)$ and $p_\theta(x,z)$ are tractable, we can find the marginal likelihood of $p_\theta(x)$.

$$\begin{aligned}
\log p_\theta(x) &= \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x)] \\
&= \mathbb{E}_{q_\phi(z|x)}\left[\log\left[\frac{p_\theta(x,z)}{p_\theta(z|x)}\right]\right] \\
&= \mathbb{E}_{q_\phi(z|x)}\left[\log\left[\frac{p_\theta(x,z)}{q_\phi(z|x)}\right]\right] + \mathbb{E}_{q_\phi(z|x)}\left[\log\left[\frac{q_\phi(z|x)}{p_\theta(z|x)}\right]\right] \\
&= \mathcal{L}_{\theta,\phi}(x) + D_{KL}\left(q_\phi(z|x)||p_\theta(z|x)\right)
\end{aligned}$$

The first term is the evidence lower bound (ELBO), and the second term is the Kullback-Leibler (KL) divergence. Because the KL divergence is always non-negative, the ELBO becomes the lower bound of $\log p_\theta(x)$.

$$\mathcal{L}_{\theta,\phi}(x) = \log p_\theta(x) - D_{KL}\left(q_\phi(z|x)||p_\theta(z|x)\right)$$

$$\mathcal{L}_{\theta,\phi}(x) \leq \log p_\theta(x) - D_{KL}\left(q_\phi(z|x)||p_\theta(z|x)\right)$$

Therefore, maximizing the ELBO will simultaneously maximize the log-likelihood of the model $p_\theta(x)$ and minimize the KL divergence between the approximation $q_\phi(z|x)$ and the posterior $p_\theta(z|x)$. Re-writing the ELBO can also capture a more intuitive understanding of the VAE objective function, where we see terms for the inference model $q_\phi(z|x)$ and the generative model $p_\theta(x|z)$.

$$\mathcal{L}_{\theta,\phi}(x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}\big(q_\phi(z|x)||p_\theta(z)\big)$$

Here we see that the objective is to maximize the log-likelihood of reconstructing sequences and minimize the KL divergence between the inference model $q_\phi(z|x)$ and the prior $p_\theta(z)$. To take gradients of the ELBO, we used the reparameterization trick[34,53]. In practice, we incorporated convex combination linear inverse autoregressive flows to obtain full rank covariance Gaussian posteriors[54].

Models were built using Python 3.6 and PyTorch 1.3.1. The encoder consisted of bidirectional stateful LSTMs (2 layers, 200 nodes) and the decoder used fully connected nodes, one network for sequence reconstruction and another network for fitness prediction. Fitness values are concatenated to the z's and fed to the decoder. Sequences were input to the model as 15mer vectors, where 45mer ACGT sequences were converted to 15mers of numbers ranging from 0-31 based on the 32 available building blocks for HFNAP translation. Reconstruction accuracy was determined by building block reconstruction accuracy, where the input 15mer was checked against the output 15mer. DNA reconstruction accuracy was used for qualitative analysis. Reconstruction loss is computed with a cross entropy loss, and fitness loss is computed by mean squared error. The loss is regularized using a KL divergence between the latent distribution and a multivariate Gaussian normal. The model was trained on the training set described previously, comprised of 172,545 selection-2 sequences, with corresponding fitness values. To select model architecture and hyperparameters, we empirically performed grid search across a range of values including the size and number of neural network layers, the

layer architecture, the KL annealing, and the weighting between sequence reconstruction loss, fitness MSE, and the KL divergence. The sequence reconstruction loss weighting was varied between 1 to 15 and the fitness MSE weighting was varied between 100 to 200. KL annealing occurred over 5000 batches using a sigmoid annealing schedule[55]. The Adam optimizer was used with a learning rate of 0.001 and β_1=0.9 and β_2=0.99[56].

Generating HFNAP sequences with putative binding affinity for daunomycin. After model parameters were sufficiently trained and reconstruction loss was minimized, sequences were generated for testing. The top 3,072 sequences as determined by fitness, was designated as the reference set. The reference set was fed into the trained model to calculate the z distribution to sample from. The corresponding z distribution was then randomly sampled, and the decoder used to generate 10,000 HFNAP sequences for testing. We varied the fitness in increments of 0.05 ranging from 0.5 to 0.95. This gave us 1,000 sequences at each fitness value: 0.5, 0.55, … 0.95.

Generated sequences selection competition. The 10,000 CVAE-generated HFNAP sequences and top 2,000 sequences by fitness from selection 2 were synthesized (Twist Bioscience). The templates were amplified using a biotinylated primer, native PAGE purified on a 10% TBE gel, gel extracted, and DNA column purified. A mastermix of the building block distribution of the 12,000 sequences was then made, and used in the HFNAP translation. The translated sequences were then subjected to the high stringency selection conditions, with 1 min, 10 min, and 60 min elutions. The 60 min elution sample was then DNA column purified, PCR amplified, and native PAGE purified on a 10% TBE gel. The gel extracted material was then DNA column purified and prepared for HTS as shown previously. The pre-selection library, a saved sample post translation and before selection, was prepared similarly. The two samples were subject to HTS on a NextSeq, and analyzed by AptaSuite[51]. The resulting data was used

to calculate the enrichments of all of the original 12,000 sequences. The templates for the top 5 most enriched sequences from the 10,000 generated HFNAPs were then ordered from IDT. The templates were then used for HFNAP translations to prepare samples for MST characterization.

Random sequences selection competition. 10,000 reference-based random HFNAPs were sampled using the building block distribution frequency of the reference set. These were then subject to the same selection that the 10,000 CVAE-generated HFNAPs underwent.

5.9 Acknowledgments

This work was supported by NIH R35 GM118062, DARPA N66001-14-2-4053, and the Howard Hughes Medical Institute. We thank J.P. Maianti, P. Lichtor, Z. Chen, and A. Berliner for helpful comments. We also thank K. Arnett for help with MST experiments and analysis. We thank A. Vieira for assistance editing the manuscript. We thank Mary O'Reilly and Bang Wong for help with figure design and illustrations.

References

1. Ellington A. D., Szostak J. W. In vitro selection of RNA molecules that bind specific ligands. *Nature* 1990, **346**(6287)**:** 818--822.

2. Tuerk C., Gold L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science (New York, NY)* 1990, **249**(4968)**:** 505--510.

3. Bock L. C., Griffin L. C., Latham J. A., Vermaas E. H., Toole J. J. Selection of single-stranded DNA molecules that bind and inhibit human thrombin. *Nature* 1992, **355**(6360)**:** 564--566.

4. Huizenga D. E., Szostak J. W. A DNA aptamer that binds adenosine and ATP. *Biochemistry* 1995.

5. Robertson D. L., Joyce G. F. Selection in vitro of an RNA enzyme that specifically cleaves single-stranded DNA. *Nature* 1990, **344**(6265)**:** 467--468.

6. Zhang B., Cech T. R. Peptide bond formation by in vitro selected ribozymes. *Nature* 1997.

7. Seelig B., Jschke A. A small catalytic RNA motif with Diels-Alderase activity. *Chemistry \& Biology* 1999, **6**(3)**:** 167--176.

8. Sheppard T. L., Ordoukhanian P., Joyce G. F. A DNA enzyme with N-glycosylase activity. *Proceedings of the National Academy of Sciences of the United States of America* 2000, **97**(14)**:** 7802--7807.

9. Zhou C.*, et al.* DNA-Catalyzed Amide Hydrolysis. *Journal of the American Chemical Society* 2016, **138**(7)**:** 2106--2109.

10. Jiménez J. I., Xulvi-Brunet R., Campbell G. W., Turk-MacLeod R., Chen I. A. Comprehensive experimental fitness landscape and evolutionary network for small RNA. *Proceedings of the National Academy of Sciences of the United States of America* 2013, **110**(37)**:** 14984--14989.

11. Ameta S., Winz M.-L., Previti C., Jschke A. Next-generation sequencing reveals how RNA catalysts evolve from random space. *Nucleic Acids Research* 2013, **42**(2)**:** 1303--1310.

12. de Visser J. A. G. M., Krug J. Empirical fitness landscapes and the predictability of evolution. *Nature Reviews Genetics* 2014, **15**(7)**:** 480--490.

13. Szendro I. G., Schenk M. F., Franke J., Krug J., de Visser J. A. G. M. Quantitative analyses of empirical fitness landscapes. *Journal of Statistical Mechanics: Theory and Experiment* 2013, **2013**(01)**:** P01005.

14. Pitt J. N., Ferré D. A. Rapid Construction of Empirical RNA Fitness Landscapes. *Science (New York, NY)* 2010, **330**(6002)**:** 376--379.

15. Gawande B. N.*, et al.* Selection of DNA aptamers with two modified bases. *Proceedings of the National Academy of Sciences* 2017, **114**(11)**:** 2898--2903.

16. Chen Z., Lichtor P. A., Berliner A. P., Chen J. C., Liu D. R. Evolution of sequence-defined highly functionalized nucleic acid polymers. *Nature chemistry* 2018, **346:** 1.

17. Kimoto M., Yamashige R., Matsunaga K.-i., Yokoyama S., Hirao I. Generation of high-affinity DNA aptamers using an expanded genetic alphabet. *Nature biotechnology* 2013, **31**(5)**:** 453--457.

18. Ruff K. M., Snyder T. M., Liu D. R. Enhanced Functional Potential of Nucleic Acid Aptamer Libraries Patterned to Increase Secondary Structure. *Journal of the American Chemical Society* 2010, **132**(27)**:** 9453--9464.

19. Lou X.*, et al.* Micromagnetic selection of aptamers in microfluidic channels. *Proceedings of the National Academy of Sciences* 2009, **106**(9)**:** 2989--2994.

20. Oh S. S.*, et al.* Improving Aptamer Selection Efficiency through Volume Dilution, Magnetic Concentration, and Continuous Washing in Microfluidic Channels. *Analytical chemistry* 2011, **83**(17)**:** 6883--6889.

21. Bedbrook C. N.*, et al.* Machine learning-guided channelrhodopsin engineering enables minimally invasive optogenetics. *Nature Methods* 2019, **16**(11)**:** 1176--1184.

22. Mason D. M.*, et al.* Deep learning enables therapeutic antibody optimization in mammalian cells by deciphering high-dimensional protein sequence space. *bioRxiv* 2019, **10:** 617860.

23. Wang J., Cao H., Zhang J. Z. H., Qi Y. Computational Protein Design with Deep Learning Neural Networks. *Scientific reports* 2018, **8**(1)**:** 1--9.

24. Biswas S., Khimulya G., Alley E. C., Esvelt K. M., Church G. M. Low-N protein engineering with data-efficient deep learning. *bioRxiv* 2020, **6:** 2020.2001.2023.917682.

25. Davis J. H., Szostak J. W. Isolation of high-affinity GTP aptamers from partially structured RNA libraries. *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99**(18)**:** 11616--11621.

26. Cho M.*, et al.* Quantitative selection of DNA aptamers through microfluidic selection and high-throughput sequencing. *Proceedings of the National Academy of Sciences of the United States of America* 2010, **107**(35)**:** 15373--15378.

27. Hoinka J., Berezhnoy A., Sauna Z. E., Gilboa E., Przytycka T. M. AptaCluster – A Method to Cluster HT-SELEX Aptamer Pools and Lessons from Its Application. 2014**:** 115--128.

28. Gui Fang C., Jie Z., Yong Hua T., Pin Gang H., Yu Zhi F. Study on the Interaction between Antitumor Drug Daunomycin and DNA. *Chinese Journal of Chemistry* 2005, **23**(5)**:** 576--580.

29. Blum R. H., Carter S. K. AdriamycinA New Anticancer Drug with Significant Clinical Activity. *Annals of Internal Medicine* 1974, **80**(2)**:** 249--259.

30. Wochner A.*, et al.* A DNA aptamer with high affinity and specificity for therapeutic anthracyclines. *Analytical Biochemistry* 2008, **373**(1)**:** 34--42.

31. Brown T. B.*, et al.* Language Models are Few-Shot Learners. 2020.

32. Roberts A., Engel J., Raffel C., Hawthorne C., Eck D. A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music. 2018.

33. Karras T., Laine S., Aila T. A Style-Based Generator Architecture for Generative Adversarial Networks. 2018.

34. Kingma D. P., Welling M. Auto-Encoding Variational Bayes. 2013.

35. Greener J. G., Moffat L., Jones D. T. Design of metalloproteins and novel protein folds using variational autoencoders. *Scientific reports* 2018, **8**(1)**:** 16189.

36. Sohn K., Lee H., Yan X. Learning Structured Output Representation using Deep Conditional Generative Models. 2015**:** 3483--3491.

37. Levenshtein V. I. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics - Doklady* 1966, **10**(8)**:** 707--710.

38. A V H., R D. Approximate String Matching. *ACM Computing Surveys (CSUR)* 1980, **12**(4)**:** 381--402.

39. McInnes L., Healy J., Saul N., Groberger L. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* 2018, **3**(29)**:** 861.

40. Fuglede B., Topsoe F. Jensen-Shannon divergence and Hilbert space embedding. 30--30.

41. Endres D. M., Schindelin J. E. A new metric for probability distributions. *IEEE Transactions on Information Theory* 2003, **49**(7)**:** 1858--1860.

42. Reuter J. S., Mathews D. H. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* 2010, **11**(1)**:** 129.

43. Xu Z., Mathews D. H. Multilign: an algorithm to predict secondary structures conserved in multiple RNA sequences. *Bioinformatics* 2011, **27**(5)**:** 626-632.

44. Hofacker I. L.*, et al.* Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie / Chemical Monthly* 1994, **125**(2)**:** 167-188.

45. Lorenz R.*, et al.* ViennaRNA Package 2.0. *Algorithms for Molecular Biology* 2011, **6**(1)**:** 26.

46. Ni S.*, et al.* Chemical Modifications of Nucleic Acid Aptamers for Therapeutic Purposes. *International Journal of Molecular Sciences* 2017, **18**(8)**:** 1683.

47. Khvorova A., Watts J. K. The chemical evolution of oligonucleotide therapies of clinical utility. *Nature biotechnology* 2017, **35**(3)**:** 238--248.

48. Stokes J. M.*, et al.* A Deep Learning Approach to Antibiotic Discovery. *Cell* 2020, **180**(4)**:** 688--702.e613.

49. Arbab M.*, et al.* Determinants of Base Editing Outcomes from Target Library Analysis and Machine Learning. *Cell* 2020, **182**(2)**:** 463--480.e430.

50. Hili R., Niu J., Liu D. R. DNA ligase-mediated translation of DNA into densely functionalized nucleic acid polymers. *Journal of the American Chemical Society* 2013, **135**(1)**:** 98--101.

51. Hoinka J., Backofen R., Przytycka T. M. AptaSUITE: A Full-Featured Bioinformatics Framework for the Comprehensive Analysis of Aptamers from HT-SELEX Experiments. *Molecular Therapy - Nucleic Acids* 2018, **11:** 515--517.

52. Kingma D. P., Welling M. An Introduction to Variational Autoencoders. 2019.

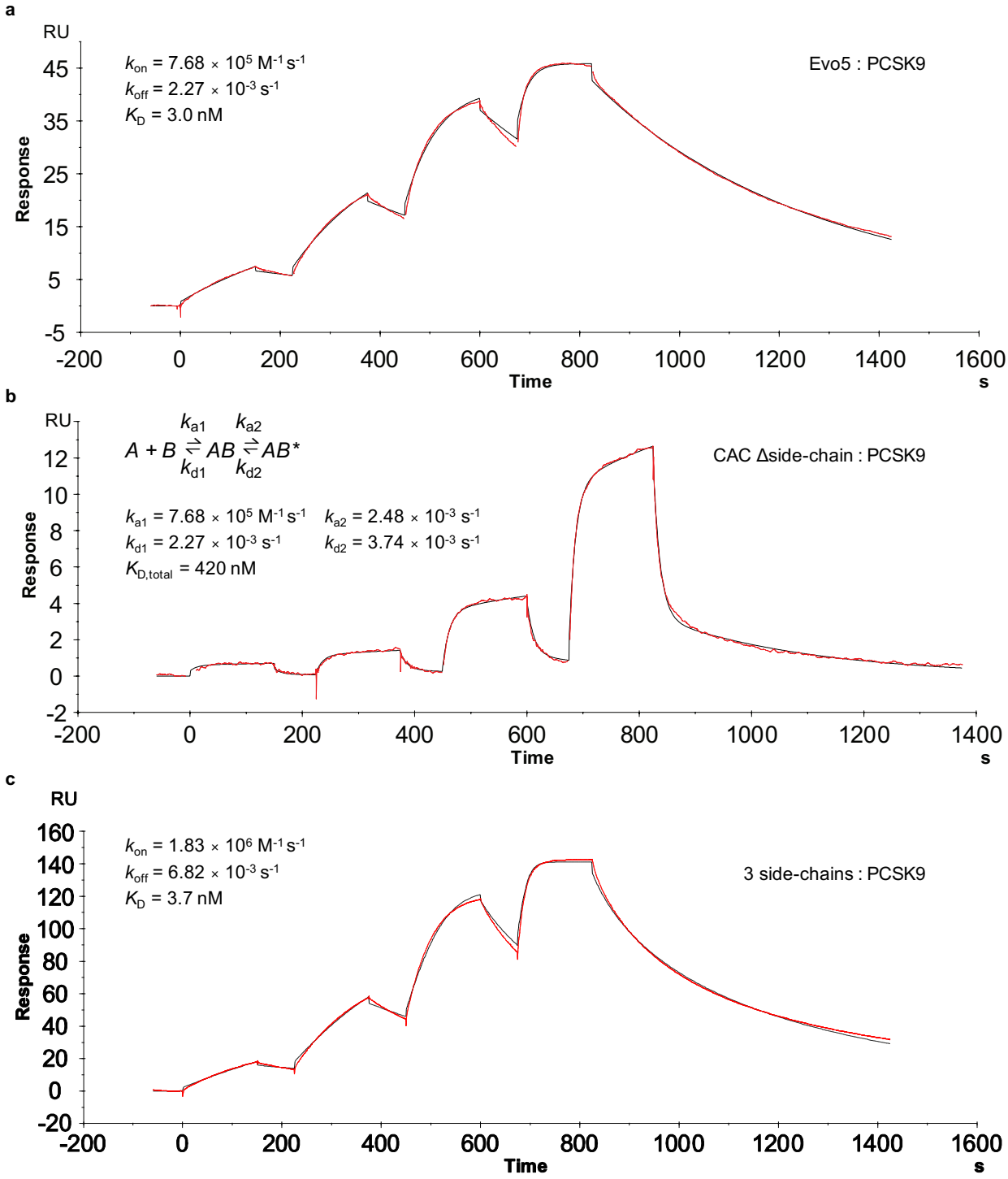
53. Rezende D. J., Mohamed S., Wierstra D. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. 2014.

54. Tomczak J. M., Welling M. Improving Variational Auto-Encoders using convex combination linear Inverse Autoregressive Flow. 2017.

55. Bowman S. R.*, et al.* Generating Sentences from a Continuous Space. 2015.

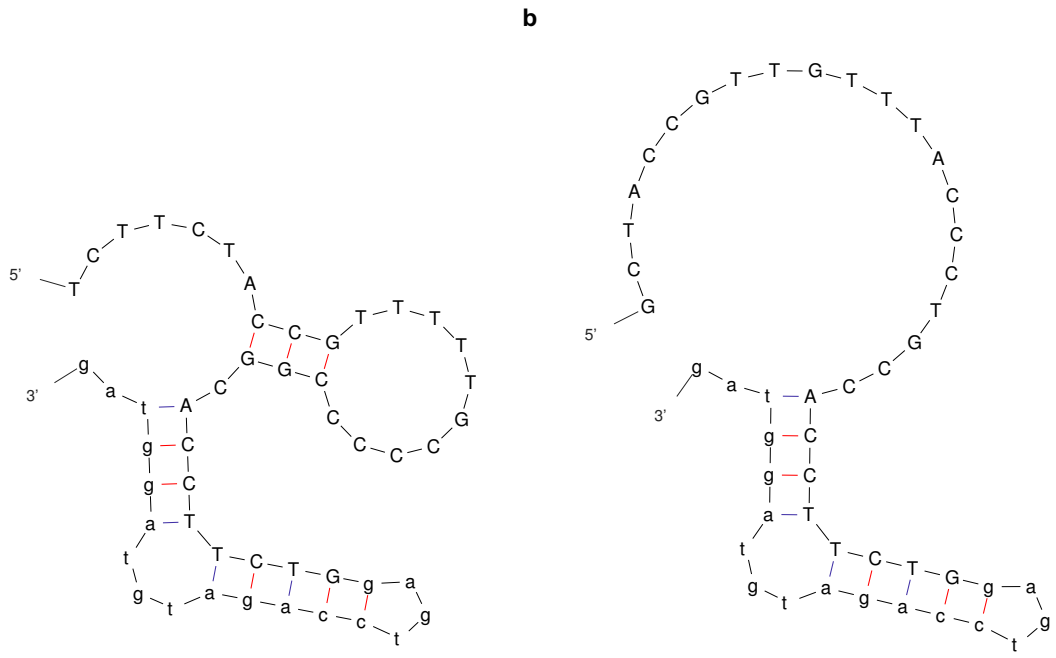
56. Kingma D. P., Ba J. Adam: A Method for Stochastic Optimization. 2014.

Chapter 6 Appendix



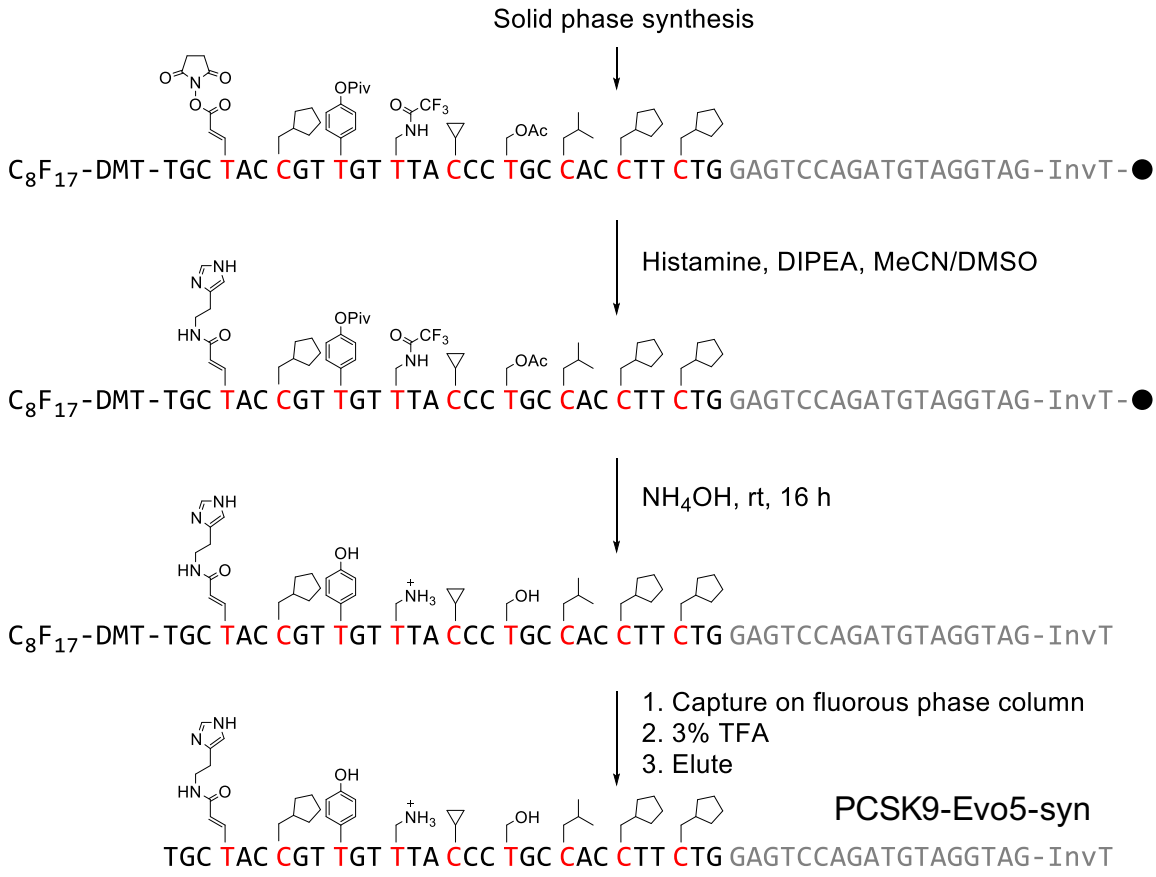
Supplementary Figure 2.1 | Binding kinetics characterization between PCSK9 protein and PCSK9-Evo5 or a side-chain mutant. a, SPR sensogram characterizing binding kinetics between PCSK9 protein and surface-immobilized biotinylated PCSK9-Evo5. **b**, SPR sensogram characterizing binding kinetics between PCSK9 protein and surface-immobilized biotinylated PCSK9-Evo5-"CACΔSide chain". The concentrations of injected PCSK9 were 2, 6, 20, and 60

nM. The raw sensograms are shown in red and the fitted curves with the kinetic parameters listed are shown in black.

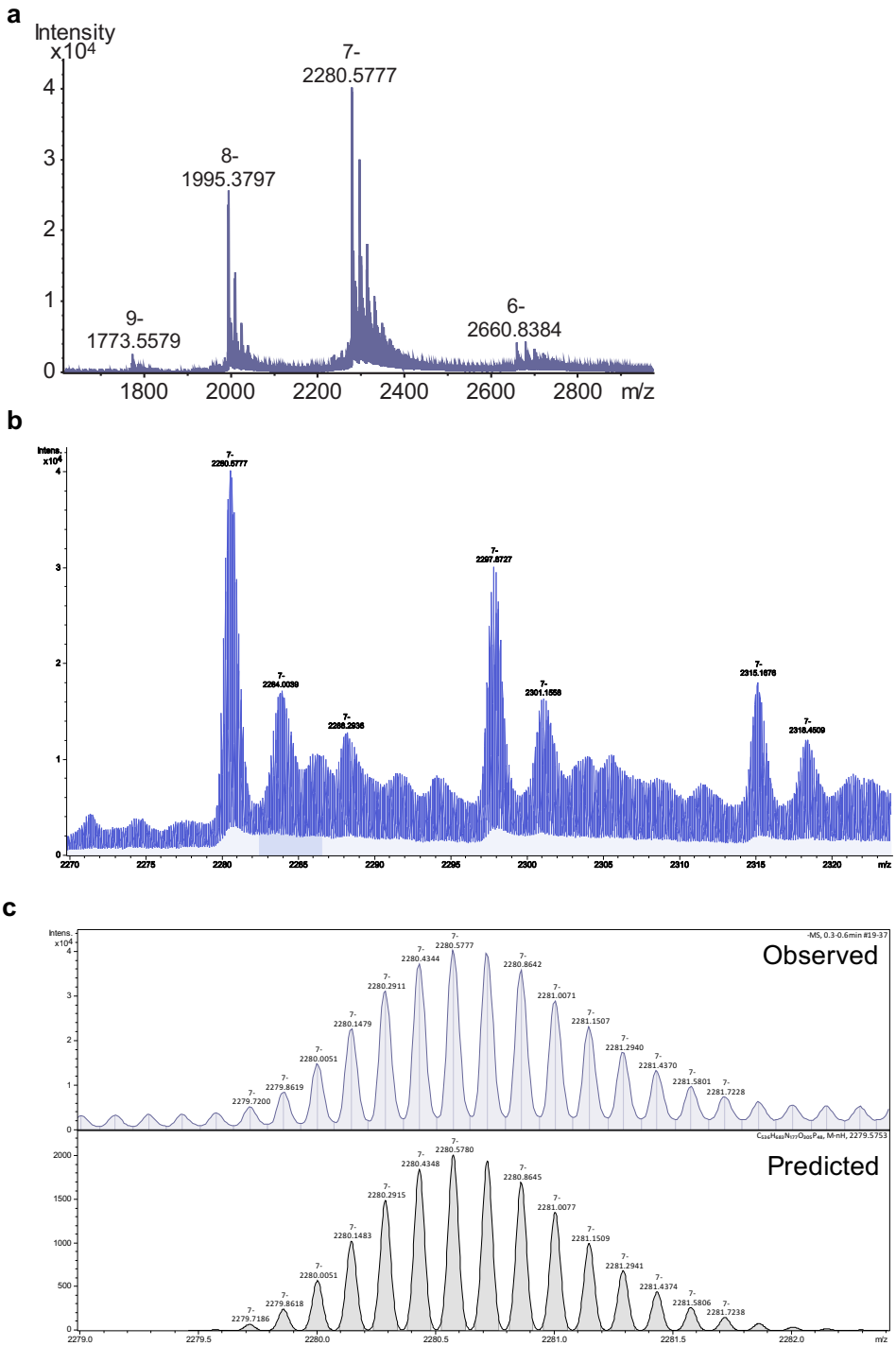


PCSK9-A5Trunc (DNA) PCSK9-Evo5 (DNA)

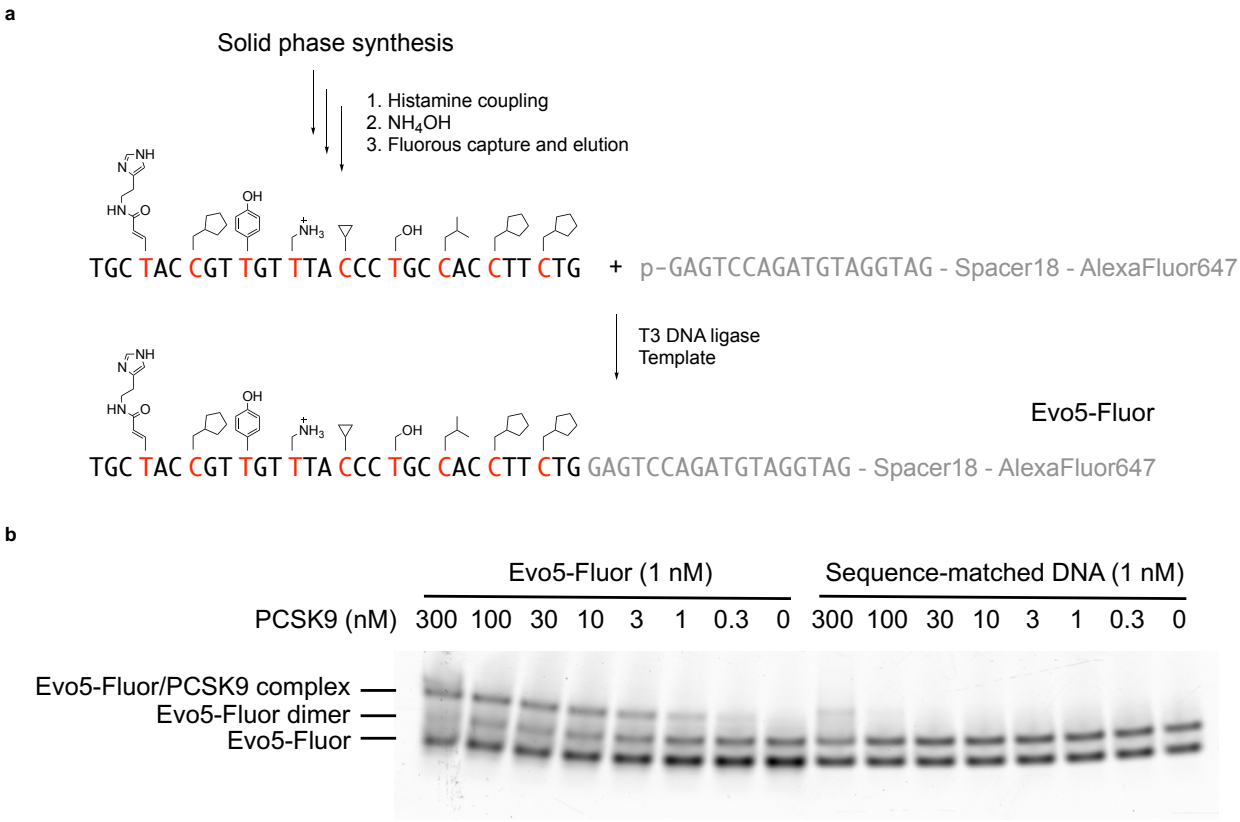
Supplementary Figure 2.2 | Secondary structure predictions of HFNAPs based on their DNA sequences. a, Predicted minimum free energy secondary structure based on the DNA sequence of PCSK9-A5Trunc. **b**, Predicted minimum free energy secondary structure based on the DNA sequence of PCSK9-Evo5. Nucleotides in the constant, unfunctionalized primer-binding region are shown in lowercase letters, and nucleotides in the variable, functionalized region are shown in uppercase letters. Note that predictions do not consider side chains, which are likely to be crucial in HFNAP folding.



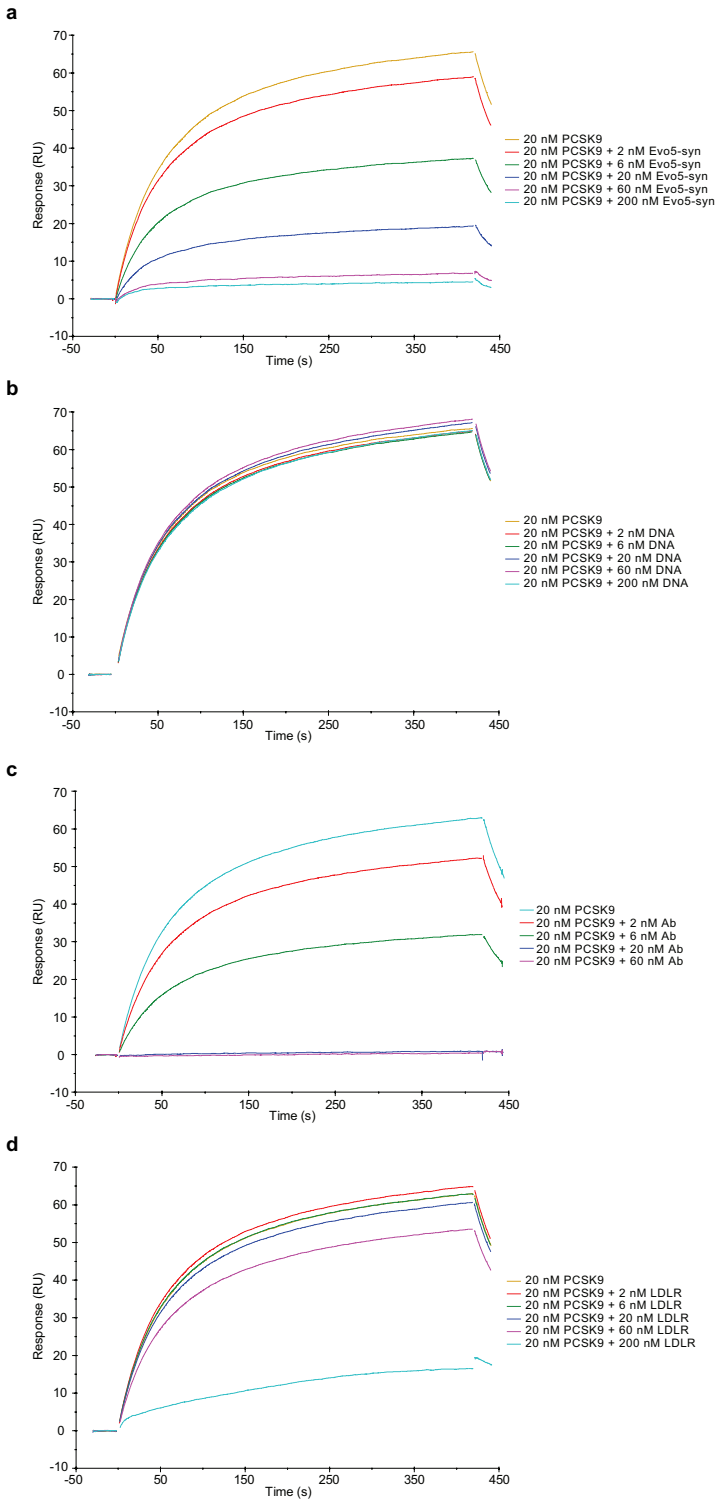
Supplementary Figure 2.3 | Multi-milligram-scale synthesis scheme for PCSK9-Evo5-syn.



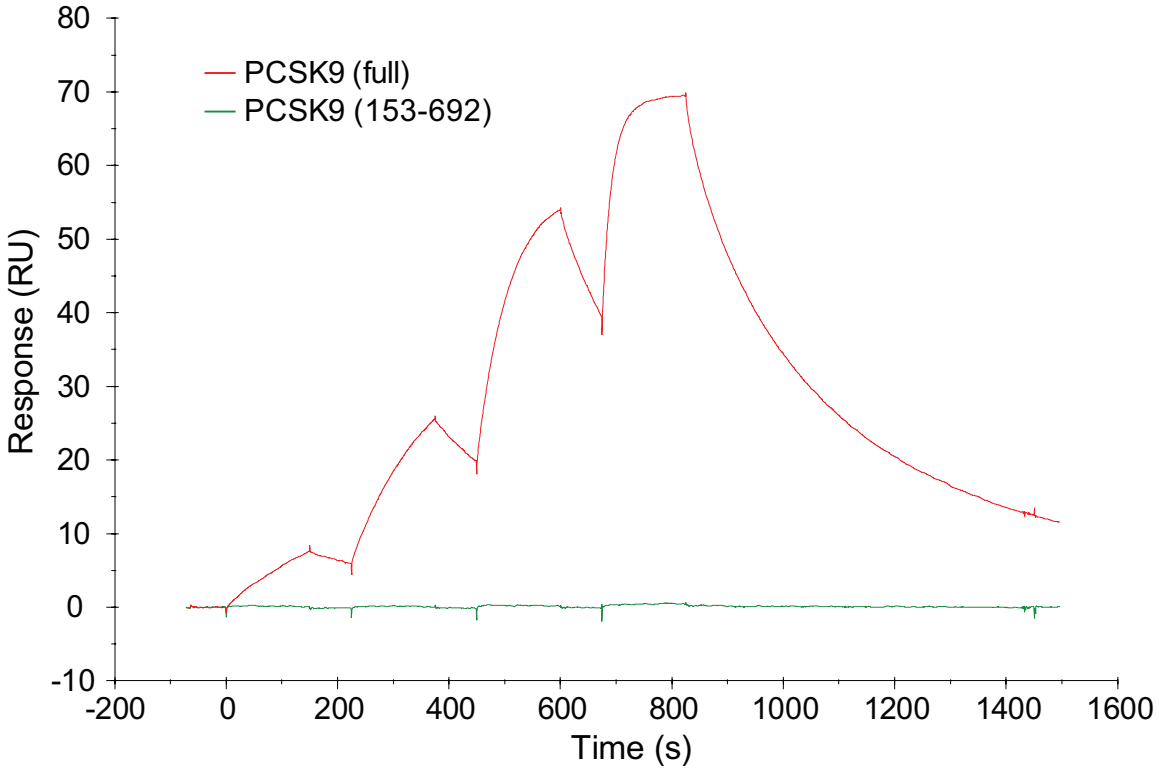
Supplementary Figure 2.4 | ESI-MS spectrum of PCSK9-Evo5-syn. Molecular formula of PCSK9-Evo5-syn is $C_{536}H_{683}N_{177}O_{305}P_{48}$. a, Full spectrum. **b**, Spectrum detailing the "7-" group of peaks, showing the peak clusters for $[M-7H]^{7-}$ (around m/z 2280.5777), as well as the sodium adduct (around m/z 2284.0039), the Tris adduct (around m/z 2297.8727), and other multiple adducts. **c**, Top: observed spectrum of the $[M-7H]^{7-}$ peak cluster. Bottom: predicted spectrum based on the molecular formula and isotopic abundances.



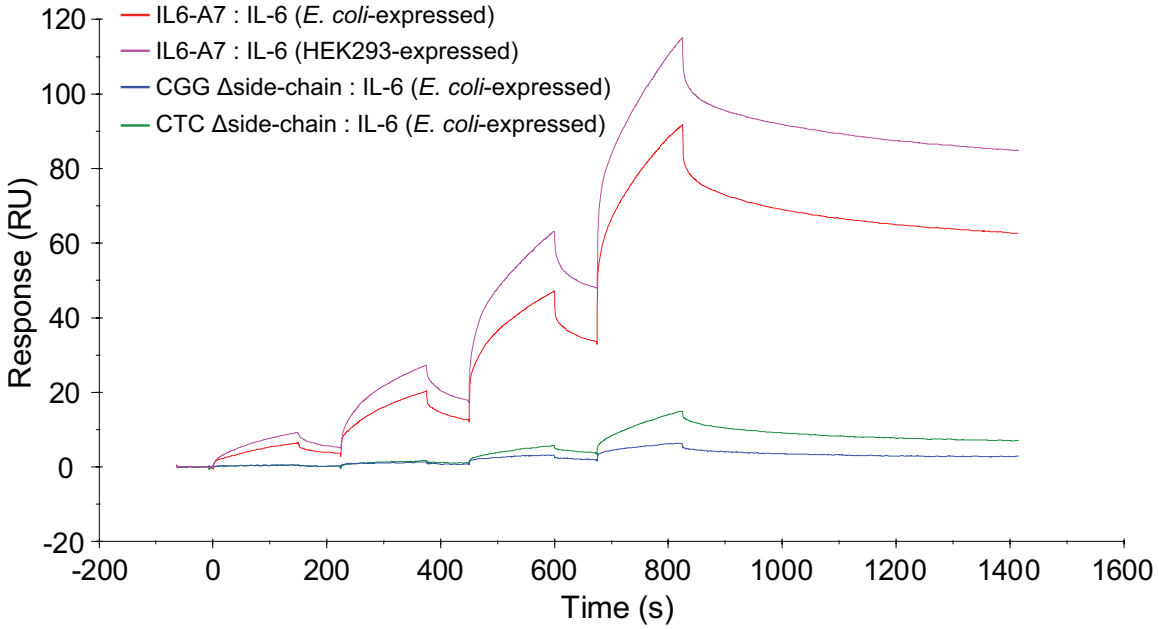
Supplementary Figure 2.5 | Electrophoretic mobility shift assay (EMSA) confirms binding of Evo5-Fluor to PCSK9 protein. a, Synthesis of the Alexa Fluor 647-labeled Evo5-Fluor. **b**, EMSA results testing Evo5-Fluor binding to PCSK9. Evo5-Fluor (1 nM) could form a complex with PCSK9 protein at concentrations down to 0.3 nM, while a sequence-matched DNA (1 nM) formed a small amount of complex only at 300 nM PCSK9. Both Evo5-Fluor and the sequence-matched DNA dimerize under the assay conditions (4 °C), resulting in the extra band. We note that the EMSA gel electrophoresis experiment takes place on a slower time scale (~15 min) than dissociation of the polymer:PCSK9 complex (k_{off} = 2.42 × 10^{-3} s^{-1} by SPR, which corresponds to $t_{½}$ ~5 min), and therefore underestimates binding affinity.



Supplementary Figure 2.6 | SPR characterization of PCSK9-LDLR binding in the presence of PCSK9-Evo5-syn or various controls. The sensograms show the interaction between surface-immobilized LDLR and solution-phase PCSK9 pre-incubated with varying concentrations of **a**, PCSK9-Evo5-syn, **b**, sequence-matched unfunctionalized DNA, **c**, a known PCSK9-neutralizing antibody, or **d**, unlabeled LDLR.



Supplementary Figure 2.7 | Effect of truncating PCSK9 on its binding affinity to PCSK9-Evo5-syn. SPR sensograms show binding kinetics between surface-immobilized biotinylated PCSK9-Evo5 and either full-length PCSK9 protein (red) or a truncated PCSK9 protein missing the prodomain (green). The concentrations of injected protein were 2, 6, 20, and 60 nM.



Supplementary Figure 2.8 | SPR characterization of binding between IL-6 and HFNAPs. The SPR sensograms show binding kinetics between biotinylated IL6-A7 (red) or two side-chain variants (blue and green) and *E. coli*-expressed IL-6 protein, and between biotinylated IL6-A7 and HEK293 cell-expressed IL-6 protein (purple). For all experiments, comparable amounts of

the biotinylated HFNAPs (all between 90 and 120 RU) were immobilized on the active flow cell. The concentrations of injected IL-6 were 10, 30, 100, and 300 nM.

Supplementary Note 2.1

All occurrences of U below are 2'-deoxy-U. Commercially available oligonucleotide modifiers are denoted by shorthand notations used by IDT.

Evaluation of translation yield on template libraries

Name	Sequence
Template-11codon	CGTACGGTCGACGCTAGCNNRNNRNNRNNRNNRNNRNNRNNRNNRNNRCACGTGGAGCTCGGATCC
Template-13codon	CGTACGGTCGACGCTAGCNNRNNRNNRNNRNNRNNRNNRNNRNNRNNRNNRNNRCACGTGGAGCTCGGATCC
Template-15codon	CGTACGGTCGACGCTAGCNNRNNRNNRNNRNNRNNRNNRNNRNNRNNRNNRNNRNNRNNRCACGTGGAGCTCGGATCC
pp1-library	/5Phos/GCTAGCGTCGACCGTACG
pp2-library	GGATCCGAGCTCCACGTG

Validation of sequence specificity of translation and amplification

Name	Sequence
Template-Bt-CATA	/52-Bio//iSp18/CGTACGGTCGACGCTAGCTTGAAAGTGCAAGAGACACCGCGACACGTGGAGCTCGGATCC
Template-Bt-CBTB	/52-Bio//iSp18/CGTACGGTCGACGCTAGCATGTTACTGGTATCGTGAGCGGGACACGTGGAGCTCGGATCC
Template-Bt-CCTC	/52-Bio//iSp18/CGTACGGTCGACGCTAGCTAGATATGGCTAGGGTCAAGGGCACACGTGGAGCTCGGATCC
Template-Bt-CDTD	/52-Bio//iSp18/CGTACGGTCGACGCTAGCAAGTAACAGGAAACGAGACGGCCACACGTGGAGCTCGGAUCC
pp1-library-T7	/5Phos/GCTAGCGTCGACCGTACGAGCGTCGCTACGCGTGAC
pp2-library	GGATCCGAGCTCCACGTG
T7-out-PCR2	TAATACGACTCACTATAGGGCTCGATTTAATTTCGCCGACGTGATGACATTCCAGGCAGTGTCACGCGTAGCGACGCT

PCSK9 binder selection and evolution

Name	Sequence
Naïve library AZ15	CGA ATC AGA TTG GAC CAG YNN YNN YNN YNN YNN YNN YNN YNN YNN YNN YNN YNN YNN YNN GAG TCC AGA TGT AGG TAG
BtBt-ExtA	/52-Bio//iSp18/CTA CCT ACA TCT GGA CTC
ExtA	CTA CCT ACA TCT GGA CTC
pp1A	/5Phos/GAG TCC AGA TGT AGG TAG
pp2Z	CGA ATC AGA TTG GAC CAG
MiSeqA	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT NNNN CTA CCT ACA TCT GGA CTC
MiSeqZ	TGG AGT TCA GAC GTG TGC TCT TCC GAT CT NNNN CGA ATC AGA TTG GAC CAG
IlluminaAdapterFwd	AATGATACGGCGACCACCGAGATCTACAC[8-base barcode]ACACTCTTTCCCTACACGAC
IlluminaAdapterRev	CAAGCAGAAGACGGCATACGAGAT[8-base barcode] GTGACTGGAGTTCAGACGTGTGCT

Rediv library AZ15	CGA ATC AGA TTG GAC CAG XZP XFO JPZ JFP JOZ JOZ JPZ XFO JOO JFZ XZZ XFF XPZ XOO XOF GAG TCC AGA TGT AGG TAG X = 79% dC, 21% T Z = 7% dA, 79% dC, 7% dG, 7% T P = 79% dA, 7% dC, 7% dG, 7% T F = 7% dA, 7% dC, 79% dG, 7% T O = 7% dA, 7% dC, 7% dG, 79% T J = 21% dC, 79% T
pp1A-3ddC	/5Phos/GAG TCC AGA TGT AGG TAG/3ddC/

Synthesis of putative PCSK9 binders for bead retention assay

Name	Sequence
pp1A	/5Phos/GAG TCC AGA TGT AGG TAG
pp2Z	CGA ATC AGA TTG GAC CAG
PCSK9A1-BtTempl	/52-Bio//iSp18/CTA CCT ACA TCT GGA CTC CAG AAG GTG ATG CAA AGG CAA ACG GTA GGA GAG AGA ATA TAG TGG CTG GTC CAA TCT GAT TCG
PCSK9A1-DNA	CGA ATC AGA TTG GAC CAG CCA CTA TAT TCT CTC TCC TAC CGT TTG CCT TTG CAT CAC CTT CTG GAG TCC AGA TGT AGG TAG
PCSK9A2-BtTempl	/52-Bio//iSp18/CTA CCT ACA TCT GGA CTC TCA GCG CAG AAG GTG AGG GCA AAA ACG GTA GAA GAA TCA TTG TTG CTG GTC CAA TCT GAT TCG
PCSK9A2-DNA	CGA ATC AGA TTG GAC CAG CAA CAA TGA TTC TTC TAC CGT TTT TGC CCT CAC CTT CTG CGC TGA GAG TCC AGA TGT AGG TAG
PCSK9A3-BtTempl	/52-Bio//iSp18/CTA CCT ACA TCT GGA CTC TGA TCA GGA TAA GTA GTG CTA AAG ACA TGA AAG AGG TTG TAG AGA CTG GTC CAA TCT GAT TCG
PCSK9A3-DNA	CGA ATC AGA TTG GAC CAG TCT CTA CAA CCT CTT TCA TGT CTT TAG CAC TAC TTA TCC TGA TCA GAG TCC AGA TGT AGG TAG
PCSK9A4-BtTempl	/52-Bio//iSp18/CTA CCT ACA TCT GGA CTC GAA TAG GTA CCG CTA AAG ACG TGA TAG AAA CGA AAA GCA TTG GGG CTG GTC CAA TCT GAT TCG
PCSK9A4-DNA	CGA ATC AGA TTG GAC CAG CCC CAA TGC TTT TCG TTT CTA TCA CGT CTT TAG CGG TAC CTA TTC GAG TCC AGA TGT AGG TAG
PCSK9A5-BtTempl	/52-Bio//iSp18/CTA CCT ACA TCT GGA CTC CAG AAG GTG CCG GGG GCA AAA ACG GTA GAA GAA TCA GTA ACG TGG CTG GTC CAA TCT GAT TCG
PCSK9A5-DNA	CGA ATC AGA TTG GAC CAG CCA CGT TAC TGA TTC TTC TAC CGT TTT TGC CCC CGG CAC CTT CTG GAG TCC AGA TGT AGG TAG
PCSK9A6-BtTempl	/52-Bio//iSp18/CTA CCT ACA TCT GGA CTC GGA ACA ACG CAG AAG GAG CCG TGG AGA AGG CAG AAG AGG TTG AAA CTG GTC CAA TCT GAT TCG
PCSK9A6-DNA	CGA ATC AGA TTG GAC CAG TTT CAA CCT CTT CTG CCT TCT CCA CGG CTC CTT CTG CGT TGT TCC GAG TCC AGA TGT AGG TAG

PCSK9A7-BtTempl	/52-Bio//iSp18/CTA CCT ACA TCT GGA CTC CAG AAG GTG AAA GTA AGA GAA CAA ACG GTA GAG GAA TAG GGA AGA CTG GTC CAA TCT GAT TCG
PCSK9A7-DNA	CGA ATC AGA TTG GAC CAG TCT TCC CTA TTC CTC TAC CGT TTG TTC TCT TAC TTT CAC CTT CTG GAG TCC AGA TGT AGG TAG

Synthesis of biotinylated PCSK9-A5 and variants for surface plasmon resonance assay

Name	Sequence
pp1A	/5Phos/GAG TCC AGA TGT AGG TAG
BtBt-pp2Z	/52-Bio//iSp18/CGA ATC AGA TTG GAC CAG
PCSK9A5-Templ	CTA CCT ACA TCT GGA CTC CAG AAG GTG CCG GGG GCA AAA ACG GTA GAA GAA TCA GTA ACG TGG CTG GTC CAA TCT GAT TCG

Synthesis of biotinylated PCSK9-Evo5 and variants for surface plasmon resonance assay

Name	Sequence
pp1A-3primeBtBt	/5Phos/GAG TCC AGA TGT AGG TAG/iSp18/iBiodT/3Bio/
pp1A-3primeBtBt-14nt	/5Phos/GAG TCC AGA TGT AG/iSp18/iBiodT/3Bio/
Evo5-pp2-dU	CCA CGT TAC TGA TTC UGC
PCSK9Evo5-Templ	CTA CCT ACA TCT GGA CTC CAG AAG GTG GCA GGG TAA ACA ACG GTA GCA GAA TCA GTA ACG TGG

Synthesis of PCSK9-Evo5-Fluor and negative control for EMSA assay

Name	Sequence
pp1A-Alexa647	/5phos/GAG TCC AGA TGT AGG TAG /iSp18//3AlexF647N/
Evo5_DNA-LeftHalf	TGCTACCGTTGTTTACCCTGCCACCTTCTG
BtBt_Evo5-Template	/52-Bio//iSp18/CTA CCT ACA TCT GGA CTC CAG AAG GTG GCA GGG TAA ACA ACG GTA GCA GAA TCA GTA ACG TGG CTG

Negative control for PCSK9-Evo5-syn in SPR assay

Name	Sequence
Evo5DNA-InvdT	TGC TAC CGT TGT TTA CCC TGC CAC CTT CTG GAG TCC AGA TGT AGG TAG /3InvdT/

IL-6 binder selection

Name	Sequence
Naïve library CW15	CTC GGA TGA ACC TGG ACT YNN YNN YNN YNN YNN YNN YNN YNN YNN YNN YNN YNN YNN YNN GGA CTG AGT CCA GAG TAA
BtBt-ExtC	/52-Bio//iSp18/TTA CTC TGG ACT CAG TCC
ExtC	TTA CTC TGG ACT CAG TCC
pp1C	/5Phos/GGA CTG AGT CCA GAG TAA
pp2W	CTC GGA TGA ACC TGG ACT
MiSeqC	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT NNNNNN TTA CTC TGG ACT CAG TCC
MiSeqW	TGG AGT TCA GAC GTG TGC TCT TCC GAT CT NNNN CTC GGA TGA ACC TGG ACT

IlluminaAdapterFwd	AATGATACGGCGACCACCGAGATCTACAC[8-base barcode] ACACTCTTTCCCTACACGAC
IlluminaAdapterRev	CAAGCAGAAGACGGCATACGAGAT[8-base barcode] GTGACTGGAGTTCAGACGTGTGCT

Synthesis of putative IL-6-binding HFNAPs for bead retention assay

Name	Sequence
pp1C	/5phos/GGA CTG AGT CCA GAG TAA
pp2W	CTC GGA TGA ACC TGG ACT
IL6-A1-BtTempl	/52-Bio//iSp18/TTA CTC TGG ACT CAG TCC TGG TGG CCA CTG GCA GCA CCG TCA CGG AGG CTG AGA CTG CCG CAA AGT CCA GGT TCA TCC GAG
IL6-A2-BtTempl	/52-Bio//iSp18/TTA CTC TGG ACT CAG TCC ACA GGA GGG ATA CCG GAA GGG CAA AGG CCG TAG CCA GAA CAA AAA AGT CCA GGT TCA TCC GAG
IL6-A3-BtTempl	/52-Bio//iSp18/TTA CTC TGG ACT CAG TCC ACA ACG CAG CCG GAA GCA CCA CAG GAG AGG CGG TAA CCA GCA ACA AGT CCA GGT TCA TCC GAG
IL6-A4-BtTempl	/52-Bio//iSp18/TTA CTC TGG ACT CAG TCC TCG GAA GCG TCA CGA CGG TAA CCG GCA CTG AGA GCA ACA CCA CAA AGT CCA GGT TCA TCC GAG
IL6-A5-BtTempl	/52-Bio//iSp18/TTA CTC TGG ACT CAG TCC TGG CTA AGG CGA CCA CGG GCA CTG CAA CCA CAG CCA AAG GGG TGA AGT CCA GGT TCA TCC GAG
IL6-A6-BtTempl	/52-Bio//iSp18/TTA CTC TGG ACT CAG TCC CAA TGA GGG AGA GAG GGG GAA GGG CGA CAA AGG CCG TAA CCA GGA AGT CCA GGT TCA TCC GAG
IL6-A7-BtTempl	/52-Bio//iSp18/TTA CTC TGG ACT CAG TCC CCG GAG GAA TGA GTA CGA GGA AGG GCA ACG AAA TAA ACA GCA GCA AGT CCA GGT TCA TCC GAG

Synthesis of biotinylated IL6-A7 and variants for surface plasmon resonance assay

Name	Sequence
pp1C-3ddC	/5Phos/GGACTGAGTCCAGAGTAA/3ddC/
BtBt-pp2W	/52-Bio//iSp18/CTCGGATGAACCTGGACT
IL6-A7-Templ	TTA CTC TGG ACT CAG TCC CCG GAG GAA TGA GTA CGA GGA AGG GCA ACG AAA TAA ACA GCA GCA AGT CCA GGT TCA TCC GAG

Chapter 3 Supplementary Information

Supplementary Table 3.1. Distribution of unfunctionalized and side-chain-functionalized building blocks in the top 100 most abundant polymers emerging from each pool in the PCSK9 selection. The theoretical distributions were calculated assuming an equal incorporation of all 32 building blocks based upon the number of building blocks used in each pool with side-chains or without. The actual distribution building blocks in the top 100 most abundant polymers was determined from HTS data.

	Unfunctionalized	Charged	Polar	Polar Plus	Nonpolar	Nonpolar Plus	Fully Functionalized
Theoretical							
Unfunctionalized	100%	75%	63%	50%	50%	38%	0%
Functionalized	0%	25%	38%	50%	50%	63%	100%
Found							
Unfunctionalized	100%	94%	88%	53%	58%	29%	0%
Functionalized	0%	6%	12%	47%	42%	71%	100%

Supplementary Table 3.2. Most abundant or enriched sequences from the PCSK9 binding selection. Sequences identified from HTS data that were most abundant after round 12 or exhibited the highest enrichment between any two selection rounds. Polymer sequences are shown without constant regions, which are CGAATCAGATTGGACCAG (pp2Z) and GAGTCCAGATGTAGGTAG (pp1A) on the 5'- and 3'-sides, respectively, of the polymer sequences shown. Functionalized side-chains are abbreviated as follows: histidine-containing, H; amine-containing, K; hydroxyl-containing, S; phenolyl-containing, Y; cyclopropyl-containing, 3; cyclopentyl-containing, 5; fluorophenyl-containing, F; isopentyl-containing, L.

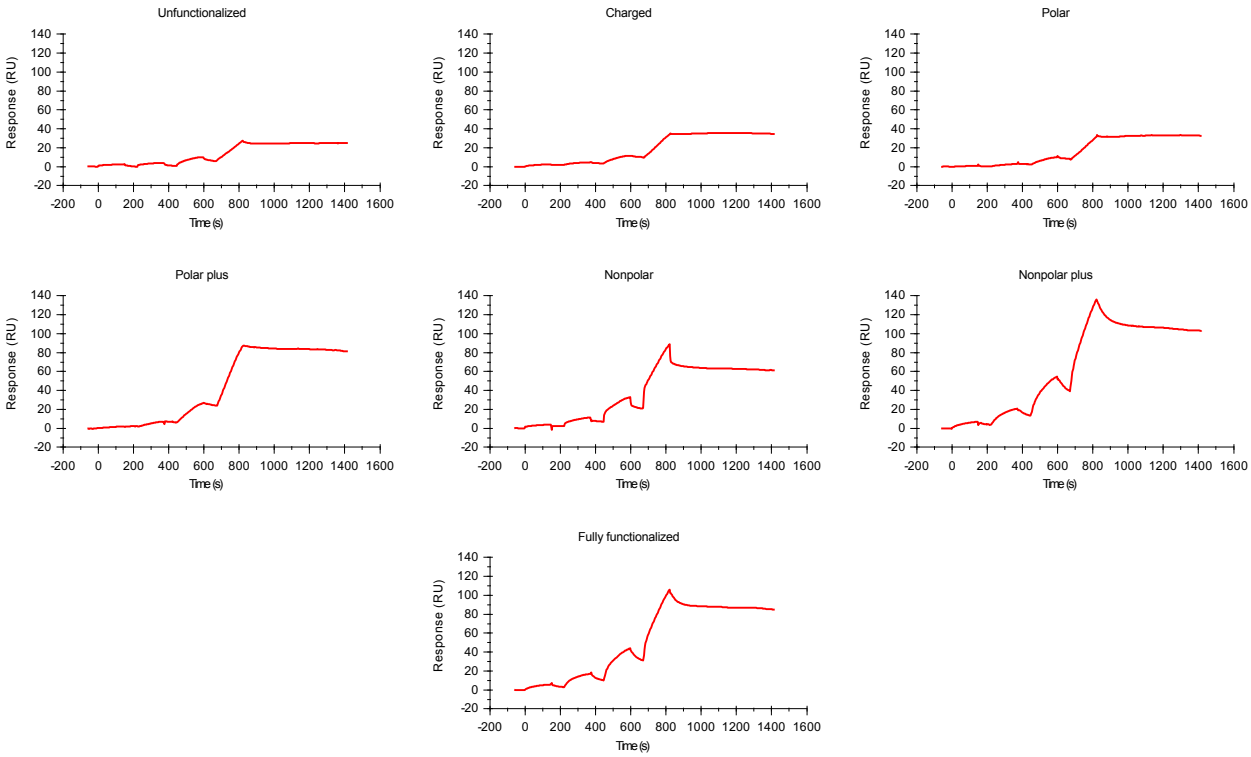
Pool	Abund. rank	Polymer sequence	% of total	Enrichment Rank	$\%R_{max}$
Unfunctionalized	1	TTACCGCCCTCTCCCTTTTGTCCCCTCCGTTGTTTCTCTTTCCCT	0.172	24100	30
	2	CCGCTTTTGCCTTCTTACCCTTTTTGTCATCCCTGTTGTTTCCCT	0.093	10629	32
	3	CACCCTCCGTCTCCGCTCCCCTCGTCTTGTTGTTTCCCTTTCCCT	0.077	19741	34
	4	TTTTATCTTTTGCCACCCTTGCCACCTTGTTGTTGTTGTTGTCCT	0.064	23796	9.4
	5	TTTCCGCTCTCGCCCCCTTGTTTACTCTCTTGGTTCCCCTTCCCT	0.050	152	10
	63	CGCTTACCATTGTCTTGTTTGCCTTCCCCTTTGTCCTGTTGTTTA	0.015	2	37
	1754	CACTTCTCATTATTGTTCTTATGACCTTGCCCGTTTTCACCCCCT	0.002	1	25
	2223	TGCTTTTTCCTCTCTTATCACTTCTTGTTGCTTTGTTGTTTCTCT	0.002	4	7.4
	2407	CTTTCTTGTCGCCGCCCTCCTTGTTGTTGATGTCCTTCCTTTCCT	0.002	4	30
	4416	CGCTTTCTCTTTTTCCCTTTTCCATGCCCATGTTTATCTTTCCGC	0.001	3	32
Charged	1	TTTTGTTGCCACCCTTGTTGTCCCCCTHCACCCCCACGTTGTCCT	1.219	13282	20
	2	TTTCATCCCTTTTTGCCACCCCTTTGTTGTCCACCCCCATGTCCT	0.570	1434	11
	3	TTTCTCCCTCTTCCACACCTTTGTCTTTGCCACCCTCCTTGTCCT	0.451	2703	22
	4	TGATTTTGCHAACGTCTTHCACTCHCCCAACCGCACCCTCCACCG	0.445	41	9.5
	5	TTTCTTTAGCCACCCTGCCCTTGTTAGCACCCTTGTTGTTGTCCT	0.427	58	n.d.
	47	CTTTTTCTTTGTKCTHCACCCCTTCCGTGCCCCTTTHCACCCCCT	0.071	3	10
	167	CGTTTTTTGCTTCTTCTATGCCCTTTTCTCTGCCCGTTTCCCCGC	0.033	2	9.8
	168	TAGTTGCCCCCTCGTTTTHCCCCTTGTCACCCTTGTTTTTGTCTC	0.033	3	18
	928	CACTAGCCTTTTTGCCCCTTTCTTTGTCGTCTTCCCTGTTTTCTC	0.008	1	10
	10736	CCGTTTHCCCCGCTTTATCCCCGGTTTCCATGTTTTTGTCCTCCT	0.001	5	n.d.
Polar	1	TTGTTGCCCTCGCCTCCACTCCCGCCTCCTHCCCCGTGTTGTCCT	0.6155	4356	24
	2	CCCTTTCTTCCTHCCCCCSGCCTCHCCCTATTGCCATGTTGTCCT	0.5872	55895	19
	3	TTTCCGCCACCACCGCCTCTTHCCCTCTTTCCTTGTTGTCCCCCT	0.3706	6601	15
	4	TTTTTTCCTTGTCAGCCTCTCTTGCCCCCGCCTCCTTGTTGTCCT	0.3559	64417	13
	5	CCGCCTCACCTTHCCCTTTGTKTCCCTTGTTGTCTTTGTKTCCCT	0.3463	4087	n.d.
	31	TTTHCCCCTTTTCTCCCACCGTTTSATTGTCTTTTTCCCTGTCCT	0.1183	2	7.4
	1239	CAGSATKCTHCACATTTTCCTTGTTTTCCCCCTTGTKCTCGTSAG	0.0068	4	19

	101845	SGACGGCACHACCCTCGTTTTHCCCCCTGTTTTCTTCCTCTCHAC	0.0002	5	3.4
	n.d.	CCTTTTHCCHCCCGGSATCTCCCGTTTHCCCCTTGTKCTCGTSAG	n.d.	1	8.7
	n.d.	CGGHACCCATTTCTTCTCSGAHCCTGTCTTTTTTGTKCTCGTSAG	n.d.	3	10
Polar plus	1	CCACCCYTTYTTCCTYTTYTGCCTYTTCTTCTTCCTCCCYTGCTA	4.100	70	73
	2	YCGCTAYTTCCTYTTCTCCCCYTTYGTCCCYGTCCCYTTCCACTA	2.536	94	49
	3	CCCYTTYGTYTTCTCSGCCCTCGTYTGCCTYTGCCCCGCCACCCT	2.277	343	35
	4	YGTYTTYTTCCCKTCYTGCCTYTTHCCCCGYTGCCACCCCTTCCT	1.798	632	n.d.
	5	YTTYTTCCTCTTYTGCCCCGTHCCCCTYGTCCAYTGCCACACCCT	1.717	10718	n.d.
	25	CGCYTTYGTYTTYTGCCCYTTCACCCTHCCYTGCCTCTTCCGCCA	0.338	2	22
	38	CCTCTGYTGYTTYTTHCCCACYTTCTTCCCYTTCACCCAYGTCCT	0.248	4	n.d.
	64	CTTCTAYTTYTTCCCYGTYTTYTGCCTYTTCCTYGTYTGCCCCGC	0.148	4	35
	66	YCGHACYTTYTTCCGCCTYTTSGCCCTYTTCTCCCCCTTYTTCCT	0.143	1	28
	445	HCACCTYCGYCGCGCCCCCCCYTTYGTCCTYTGYTTCTTCTCCCT	0.022	3	33
Nonpolar	1	TAGTTTFGATAT3TATCTTGCLTCTCGFAT5GTTGTTCA3CT5CG	21.861	16	17
	2	TTT5GTTTCTCA5TTTCTTTT3CATTCTGTTCCTTCTGTTCT3TA	17.723	70	44
	3	TAGTTTFGATAC3TATCTTGCLTCTCGFAT5GTTGTTCA3CT5CG	6.608	2	10
	4	TAGTTTFGCTAT3TATCTTGCLTCTCGFAT5GTTGTTCA3CT5CG	1.086	6	n.d.
	5	TAGTTAFGATAT3TATCTTGCLTCTCGFAT5GTTGTTCA3CT5CG	0.902	3	5.2
	15	TAGTTTTGATAT3TATCTTGCLTCTCGFAT5GTTGTTCA3CT5CG	0.320	7	n.d.
	20	TAGTTTFGATAT3TATCCTGCLTCTCGFAT5GTTGTTCA3CT5CG	0.292	1	5.7
	51	TTT5GTTTG5TTTCTTTT3CCTTCTGTTCTTTG3TATAC3CATCG	0.068	4	15
	61	TAGTTTFGATAT3TATCTTGCLTCTCGFAT5GTTGTTCC3CT5CG	0.057	5	6.7
Nonpolar plus	1	FAG5GTYTGFATTGC3CT3CC5TGTCCTTC3CCTAT3CC5CGTCC	26.0222	250	72
	2	3CCLTCTCTYGTYGTTCC3CCYTTTCCYGTTCC3CTYTGTTC3CT	4.1645	145	71
	3	TACLTCYGTYTGTCC3CCTTATCTTTC3CC5TGYGTTCCYGTTAA	1.6224	301	35
	4	YTG3CCYTTYGTYGTTTCLTCTCCYCG3CC3CTYGTTTC3CTTGA	1.5915	1471	n.d.
	5	FAG5GTYTGFATTGC3CT3CC5TGTCCTTC3CATAT3CC5CGTCC	1.2917	295	n.d.
	88	5CGFGA5TG3CCYTTTCC3CTTACFGC3CAFGC5TGYGTTATLTC	0.0799	2	21
	395	TGCYTT5TTYTT3CT3CC5CGYTTTTCTTCYGTTTA3CCTCA3TA	0.0187	4	37
	2478	YTG3CCYTTYGTYGTTTCLTCTCCYCG3CC3CTYGTTTC3CTTGC	0.0028	1	29
	14282	5TTTCTTCAYGTTTC3CTYTTYGT5TT3CC3CCTCTTCC3CCTTA	0.0003	5	n.d.
	21725	YTT5TGYTGYGTTTC3CTTCCYTGTCTTCCYGTTTATCC3CCTCC	0.0003	3	24
Fully functionalized	1	5GTYTGYTGSGC3CTLTCLTC3CTLTCSGC5TGLTCHAC3CT3CA	26.957	162	62
	2	5GTYTGYTGSGC3CTLTCLTC5CGLTCSGC5TGLTCHAC3CT3CA	6.400	187	78
	3	3CT5CGYTTYTT5TT3CCYGTYTT3CC3CA5GT3CC3CTSATYGT	5.745	3430	47
	4	YTTYGTSAT3CAHCC3CCYGTYTT3CCHCCYGTYGTLTC3CC3CT	2.494	1776	54
	5	YCGYTTYGTYTT3CTHCCYGTYTGLTCLTCLTCHCC3CTYGT3CT	2.213	43	n.d.
	9	YTTYTGYCGYTTSGC5TTHCC3CCYTG3CTLTC3CT3CC5TTLTC	1.103	2	70
	32	YTT5CG5TTYTT3CCSGC3CTYGTYTG3CC5TG3CC5TTSGA3CA	0.291	5	36
	74	YCG3CC3CTYTT5TTYGTYGTYTTHCCFATSGC3CTSATYTT3CA	0.109	3	72
	99	YCG5TTHCC3CT3CCYGTYTT5TT3CCYGT5TT3CC3CTYTGKTA	0.082	1	66
	391	5TTSGC5TT3CCYGTYTT3CTLTC3CTYGT3CTYGT3CC5TTHCA	0.015	4	48

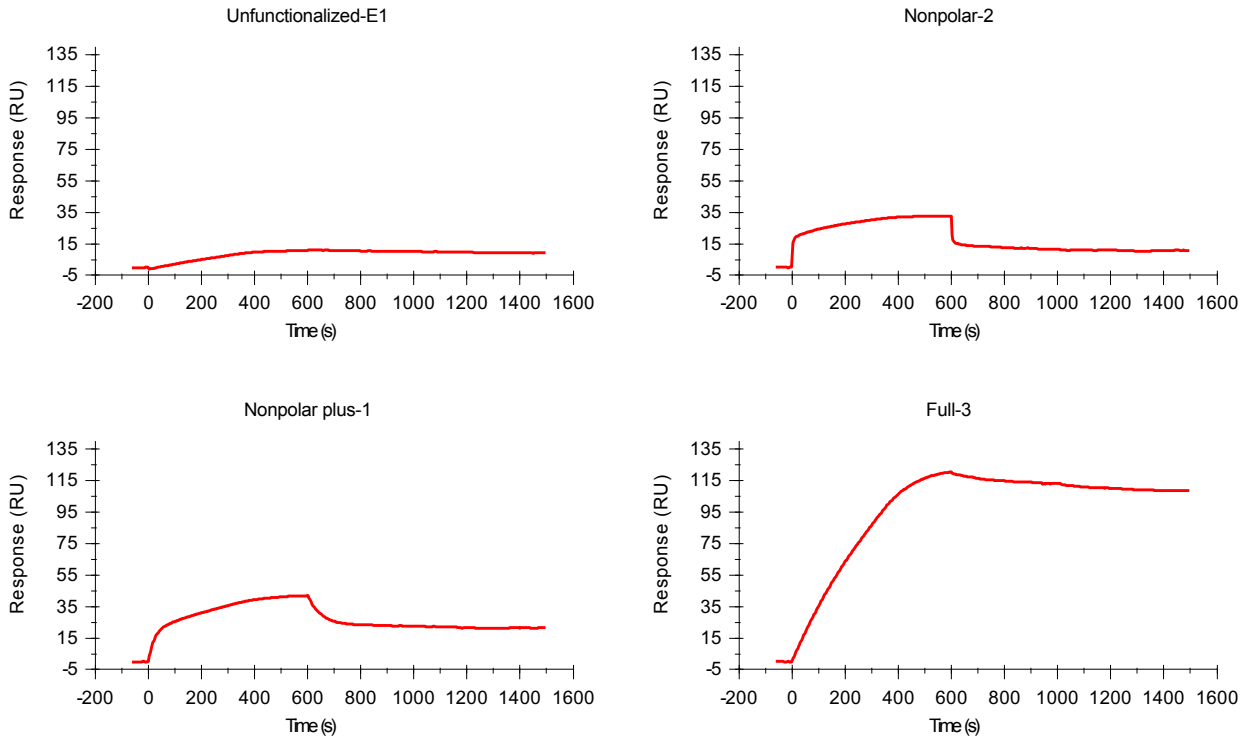
Supplementary Table 3.3. Most abundant or enriched sequences from the IL-6 binding selection. Sequences identified from HTS data that were most abundant after round 7 or exhibited the highest enrichment between any two selection rounds. Polymer sequences are shown without constant regions, which are CTCGGATGAACCTGGACT (pp2W) and GCATCGAAGCCAAGATTC (pp1B) on the 5'- and 3'-sides, respectively, of the polymer sequence shown. Functionalized side-chains are abbreviated as follows: histidine-containing, H; amine-containing, K; hydroxyl-containing, S; phenolyl-containing, Y; cyclopropyl-containing, 3; cyclopentyl-containing, 5; fluorophenyl-containing, F; isopentyl-containing, L. Not determined (n.d.).

Pool	Abund. rank	Polymer sequence	% of total	Enrichment Rank	$\%R_{max}$
Unfunctionalized	1	TTATTGCACTGGCTGTACTTTTCGTATCCATGGTCATCTTTTCTA	3.830	34	4
	2	TTACTGCGGCGCCAGCCGTGCCCATCCTTTCTTCATCCGTGCCCA	1.741	10	3
	3	CTGTTTTTACTACCGCGCTAGCCGCCTCATCCGTTTTAGCGCCAT	0.324	105	3
	4	TAATGATGCTCATGGTGCTCTTCATCACCATCTTCGTTCCATCCT	0.320	16851	4
	5	TCTTCATTGCGCTGGTGCCTGCCATCCTTGTTCCATCCCCTTTGC	0.236	9068	4
	9	TTATTGCACTGGCTGTACTTTTCGTATCCATGGTCATCTTTCCTA	0.118	1	-1
	46	TTATTGCACTGGCTGTACTTTTCGTATCCATGGTCATCTTTTCTG	0.034	5	4

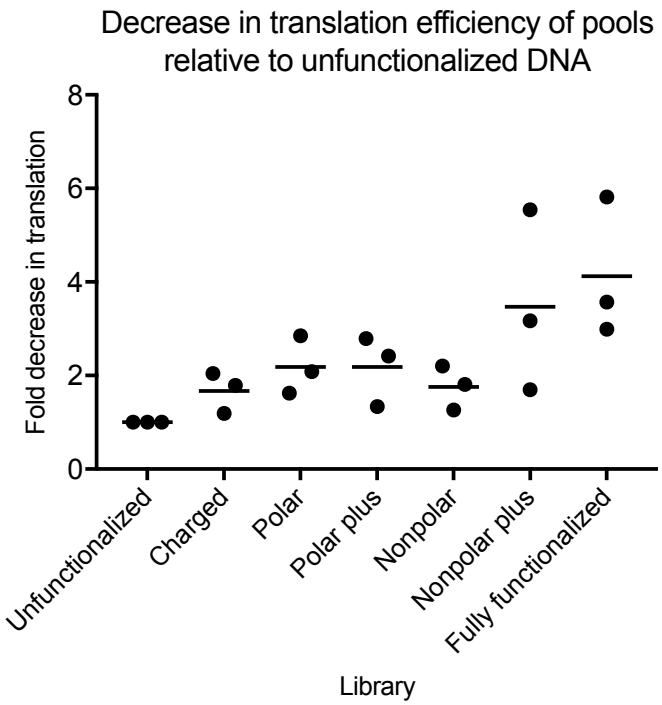
	55	TTATTGCACTGGCTGTACTTCTCGTATCCATGGTCATCTTTTCTA	0.030	4	5
	58	TTATTGCACTGGCTGTACTTTTCGTATCCATGGTCATCCTTTCTA	0.029	3	5
	232	TTACTGCGGCGCCAGCCGTGCCCATCCTTTCTCCATCCGTGCCCA	0.011	2	0
Charged	1	CTGCAATGCKTAKGGHCCKGGCGTTAGTGTTGTTGCKTCCACCGG	9.294	1	4
	2	TCGCGAKGGTCGTGCHCCCGATTTCTAKTACCGCGCTAGCGGCTC	1.812	20	5
	3	CTGCAATGCKTAKGGHCCKGGCGGTAGTGTTGTTGCKTCCACCGG	0.647	4	1
	4	TAGKGGCTTCTATGCTGCCTGCGCTTGHCACCCCGCCCTCTTCGT	0.406	255	-3
	181	TCGCGAKGGTCGHACHCCCGATTTCTAKTACCGCGCTAGCGGCTC	0.013	3	6
	294	TCGCGAKGGTCGKGGHCCCGATTTCTAKTACCGCGCTAGCGGCTC	0.010	2	2
	788	CTGCAATGAKTAKGGHCCKGGCGTTAGTGTTGTTGCKTCCACCGG	0.006	5	2
	985	TCGCGAKGGTCGTGCCCCCGATTTCTAKTACCGCGCTAGCGGCTC	0.005	6	3
	25	TCGCGAKGGTCGTGAHCCCGATTTCTAKTACCGCGCTAGCGGCTC	0.070	7	2
Polar	1	KGGCCCCCTHAATGTTGTHACCTCHCASGAKCTCGGSGCCCGCCC	14.160	3	4
	2	CGGCAATTTCTTTCGKTCKTACTGKTCCTTCCCTTGCCCCCAHCC	8.281	6	2
	3	HCATGTKGGCTCSGATGTCCTTCGKTCCTACTGKTCCATCCTCCG	3.552	11	4
	4	CTGCCTSGCKTACGGTCGSGATTGCACSATTGTCCAHCCCCTCTT	0.671	49	n.d.
	5	CTGSATKGGCTGSGACCCSGCTTGCCCCTCTTGCGTTGTCATCCC	0.655	13	n.d.
	7	HCATGTKGGCTCSGASGCCCTTCGKTCCTACTGKTCCATCCTCCG	0.294	7	2
	10	CGGCAATTTCTCTCGKTCKTACTGKTCCTTCCCTTGCCCCCAHCC	0.167	1	3
	15	HCATGTKGGCTCSGATGTCCTTCGKTCCTACTGKTCCACCCTCCG	0.119	2	2
	35	KCTTGTKGGCTCSGATGTCCTTCGKTCCTACTGKTCCATCCTCCG	0.059	5	2
	130	KGGCCCCCCHAATGTTGTHACCTCHCASGAKCTCGGSGCCCGCCC	0.021	4	2
Polar plus	1	YTTCGAYTGCCAYTGSGCYGTCATCCCCGTHCASGAYGTCGACTG	10.140	57	9
	2	KTCCCTHAACTGYTGCTGSATYTTSGCYTGSGAKCTCATCCCCTG	7.954	6	5
	3	CCAHCACGGCTTYTTYCGKTCKTACTGKTCCACCCTKTCCCCCGT	2.146	5	7
	4	CCAHAACTGYGTCGACTACGGHCCHCCCGACTGYGTKTCCATYGT	1.476	30	n.d.
	5	CAGSGAKCTSATCCAHCACCTYGTCATCGCYTGCTGYGTHCCKGG	0.516	74	n.d.
	23	YTTCGAYTGCCAYTGSGCYGTCATCCCCGCHCASGAYGTCGACTG	0.083	3	-2
	68	HCCCTGSATKGGCTTCAAKCTCCAKCTCTGCCTCACCCCHCACCC	0.036	7	1
	469	CGGYTGCTGCTTYTGYTTYCGKTCKTACTGKTCCTTCTCCGTCGC	0.011	4	2
	2262	CCAHAACTGYGTCGACTACGGHCCHCCCGACTGYGTKTCCACYGT	0.004	2	-2
	26647	CCTKTAHCAHAAHACKGGSGAKCTYTGCTGKGGCGAYTGCTCCGG	0.001	1	-2
Nonpolar	1	3CTTTATCATAATACTGGTGATCTTTG5TGTGGFGATTGLTCLGG	35.775	13	83
	2	TACTTATAC3CTTTG5GTTTATGTTGG5TGTGGFGATTGLTCLGG	10.110	16	66
	3	LTC3CATGGTTTTGGTCCTAT3TATGG5TGTGGFGATTGLTCLGG	9.371	9	83
	4	5TG3CATGGTTTTGGTCCTAT3TATGG5TGTGGFGATTGLTCLGG	7.403	53	n.d.
	5	3CTTTATCATAATGCTGGTGATCTTTG5TGTGGFGATTGLTCLGG	5.254	3	n.d.
	27	TTCTATTCT5TTTGCTGGFGATAT5TGTGCTTTTGTTTA3TA5TG	0.159	1	1
	28	3CTTTGTCATAATGCTGGTGATCTTTG5TGTGGFGATTGLTCLGG	0.139	6	43
	67	TTCTTGTAC3CTTTG5GTTTATGTTGG5TGTGGFGATTGLTCLGG	0.038	4	54
	87	TTTTAGTGTTCTTCCTTGTTC3CTTGGTTGTGGFGATTGLTCLGG	0.029	2	n.d.
	11	5TGTGGTATTCGTTGTCCLTCTCC5CGTGC3CCTCALTCTCCFGA	1.744	5	1
Nonpolar plus	1	TCA5CG3CALAALGGLTCTTCYCGFGCYTTTAGTGATGA3TATGG	3.646	4	17
	2	FGCFGA3CCYTGYTGYTTTTCTCTTCCTGCTTCTGAYTGLTCLGG	3.191	208	103
	3	TGGTGGYTGYCGFGC5CGFGCTGATAC3CT3CTTCATGATGCTTC	2.412	43	15
	4	3TATTCTATTGCTTATATTCTTCCYTGTGGTGATGAYTGLTCLGG	1.785	114	90
	5	3CT5CGTGAYGTYGTTCC3CTTTCYTGYGT5TTTGAYTGLTCLGG	1.185	163	n.d.
	100	3TATTCYGTTGCTTATATTCTTCCYTGTGGTGATGAYTGLTCLGG	0.074	6	76
	132	FGCFGC3CCYTGYTGYTTTTCTCTTCCTGCTTCTGAYTGLTCLGG	0.051	5	70
	135	5CG5CGYTTYTTYTGTCTTCATACTCATGCTTCTGAYTGLTCLGG	0.051	2	37
	230	TGGFGCTGAYTT5TTTGCTACYTG3TAYTGTGGYGT5GT3CATGG	0.031	1	4
	237	TGGTGGFGCYTTFAGTGG3CTLTC5TGTCATAT3TA3CAFGCTGC	0.030	3	2
Fully functionalized	1	HCALGG3CTKCTFGA5TGSGC3CCYCGFGC3CAYTGHAC3TASGC	47.467	2	17
	2	YCGYGTKCTYGTYTGKGGYTGSGCYTTYTG5TGYTG5GTLTCFGC	19.972	9	98
	3	HCAFGA3CTKCTFGA5TGSGC3CCYCGFGC3CAYTGHAC3TASGC	13.238	4	13
	4	YCGLGG3CTKCTFGA5TGSGC3CCYCGFGC3CAYTGHAC3TASGC	8.110	3	10
	13	YCG5GTYTTLGGKTASGCKTCLGGKTAHCC5CGLGGHCCKCTKGG	0.175	8	8
	30	HCALGG5CGKCTFGA5TGSGC3CCYCGFGC3CAYTGHAC3TASGC	0.074	1	1
	32	KCT3TAHCA3CTKCTKGG5GT5GTKTALTCFAT3CAYTTHAC3CA	0.072	5	n.d.
	44	KCT3TAKCT3CTKCTKGG5GT5GTKTALTCFAT3CAYTTHAC3CA	0.048	7	n.d.
	60	HCALGG3CTYCGFGA5TGSGC3CCYCGFGC3CAYTGHAC3TASGC	0.030	6	1



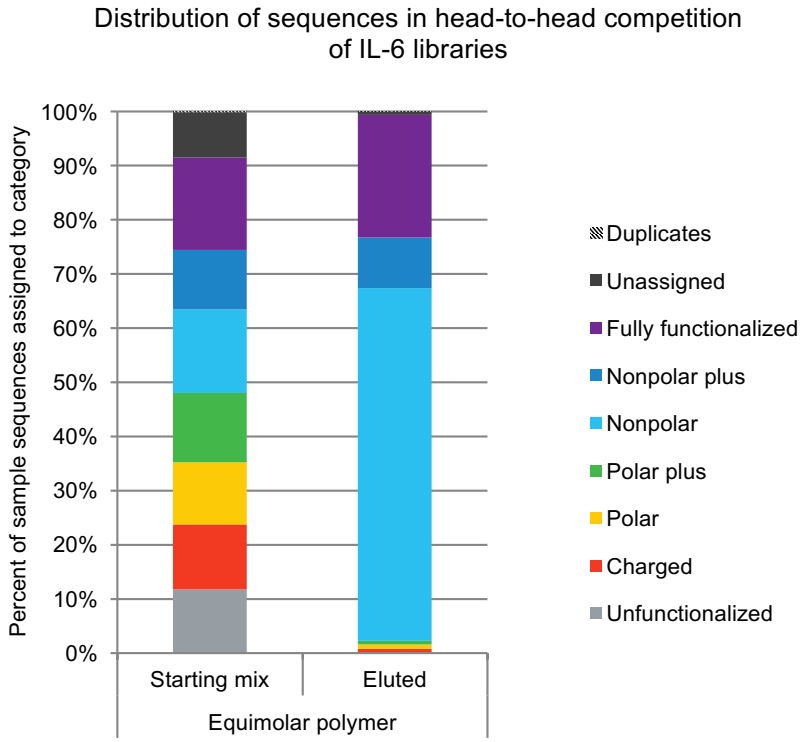
Supplementary Figure 3.1 | PCSK9 affinity of selection-enriched polymer pools. Representative SPR sensorgrams of single cycle kinetics run of different 3'-biotinylated enriched pools following selection round 12. The polymer pools were bound to the surface of a CAP chip and concentrations of 7.3, 22, 67, and 200 nM PCSK9, each with 0.05 mg/mL BSA, were flowed over the chip's surface at 30 µL/min in HBS-P+ buffer with $CaCl_2$, $MgCl_2$, and KCl. Data in **Figure 3.3a** were plotted based on the response at third binding point, at the end of the 200 nM injection of PCSK9 to minimize any influence of reference binding in the pool comparison. Sensorgrams shown represent the first of three replicates collected with the enriched polymer libraries and PCSK9 shown in **Figure 3.3a**.



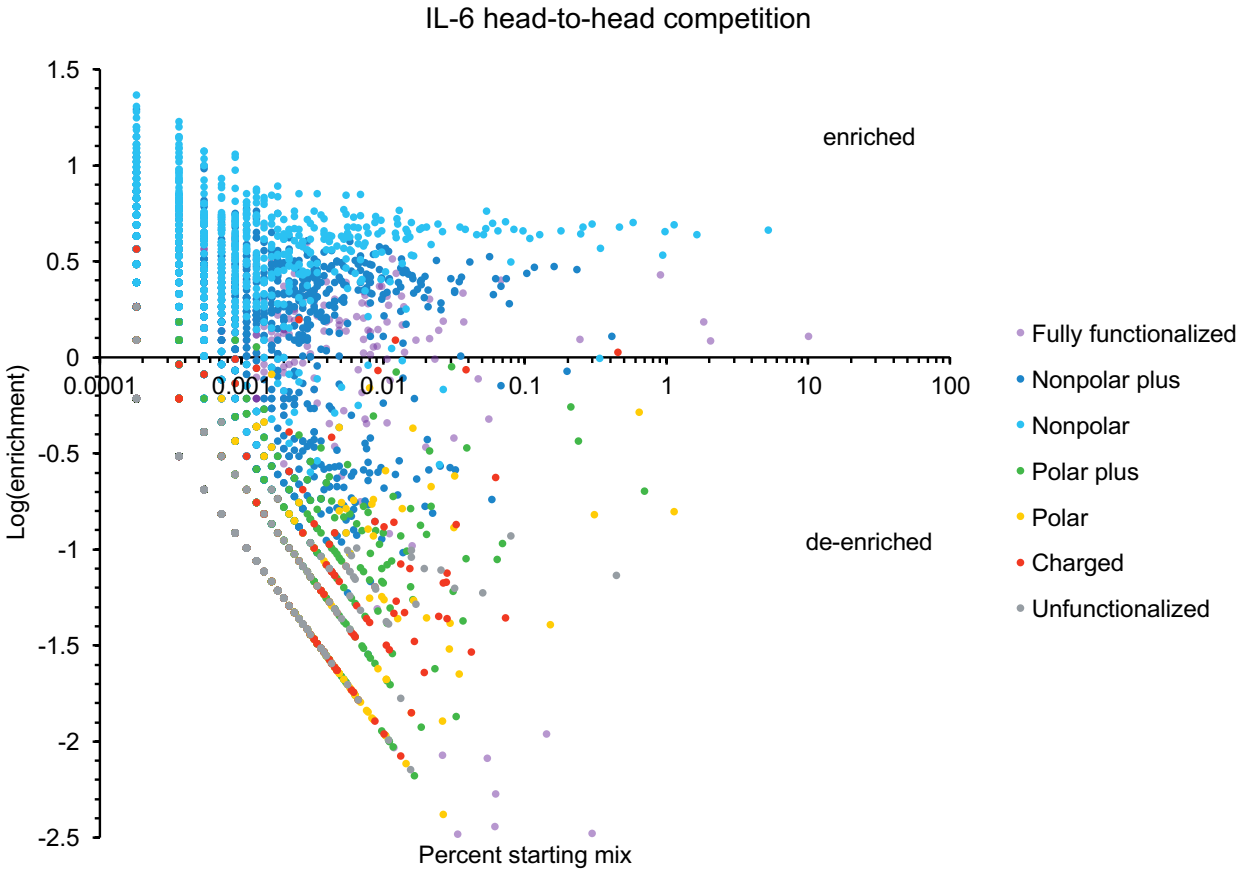
Supplementary Figure 3.2 | Individual polymer binding to PCSK9. Reference- and background-subtracted sample sensorgrams in response to a single 600-s injection of 60 nM PCSK9. Polymer Unfunctionalized-E1 is the polymer exhibiting the greatest enrichment from any round in the IL-6 selection from the unfunctionalized (DNA) library. Polymers Nonpolar-2, Nonpolar Plus-1, and Full-3 are the 2nd-, 1st-, and 3rd-most abundant sequences from the nonpolar, nonpolar plus, and fully functionalized pools following PCSK9 selection, respectively. Data presented are representative of sensorgrams used to derive data in **Figure 3.4a** and **Supplementary Table 3.2**. For this measurement, most polymers were analyzed once. For the above sensorgrams, Full-3 was analyzed twice, giving similar results.



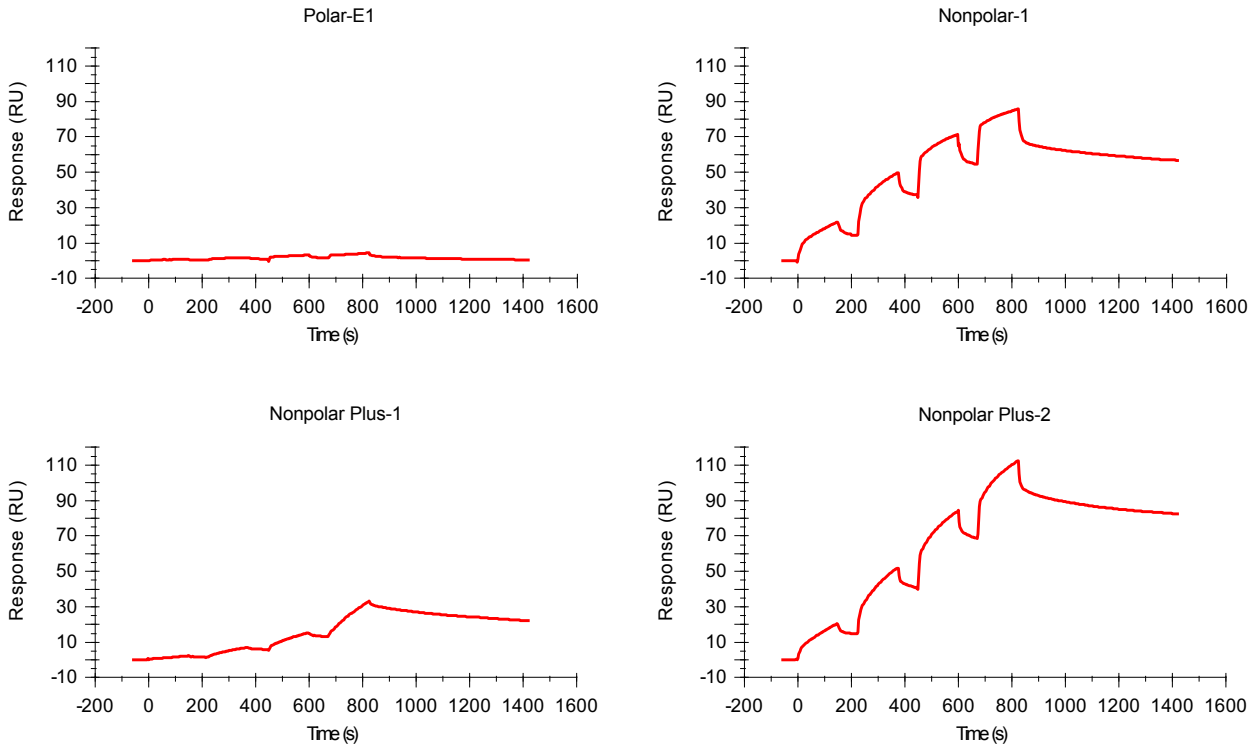
Supplementary Figure 3.3 | Translation efficiency of each pool. Calculated fold decrease in polymer populations as determined by qPCR following translation of the first-round libraries for the PCSK9, IL-6, and neutral selections, before any selection step has taken place. The horizontal bars show the mean values of three replicates; individual values for each replicate are shown. Assessment of translation efficiency assumes the PCR amplification rates of the polymer pool templates is equal.



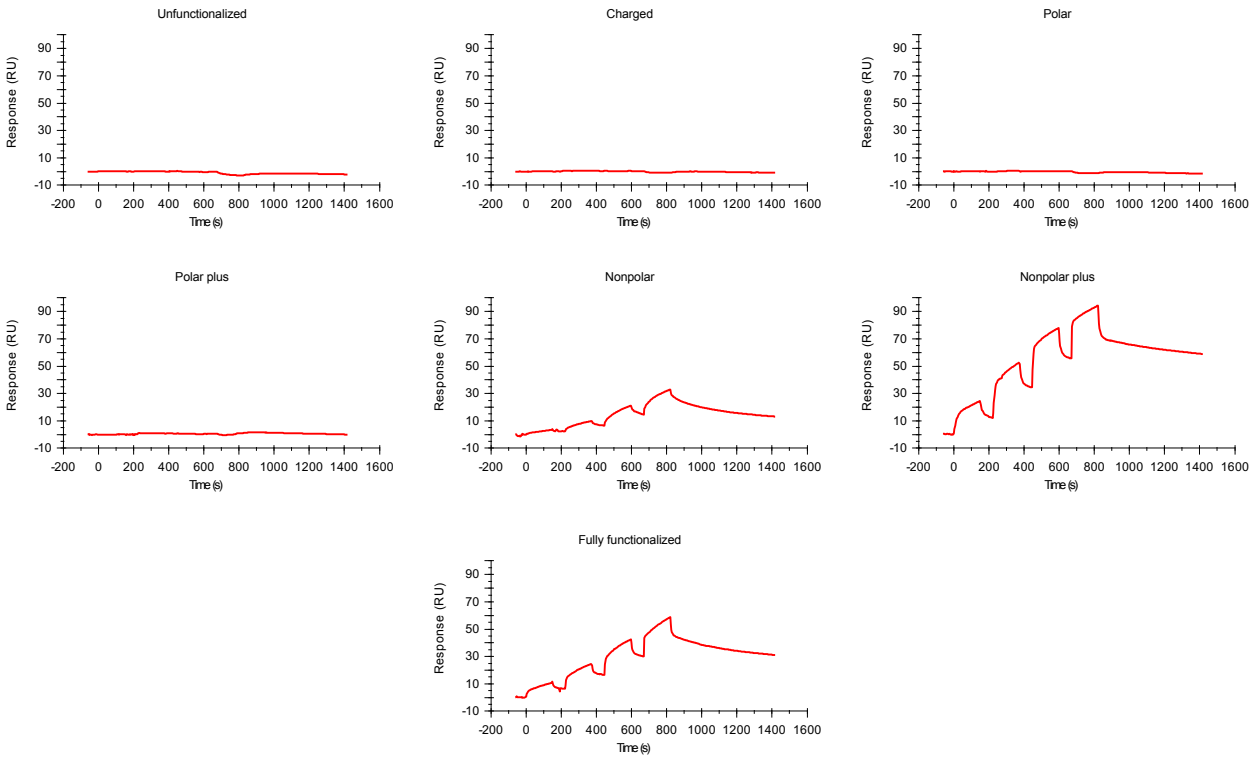
Supplementary Figure 3.4 | Analysis of mixed samples in polymer pool competition. Stacked bar plot showing the distribution of sequences from the starting and eluted material from the head-to-head competition experiments in which a round of selection against bead-bound IL-6 was performed. Sequences are assigned to a pool of origin based upon the occurrence of the same sequence in any of the 7 previous selection rounds. Sequences that were not previously found from HTS analysis of previous rounds are labeled as unidentified. Sequences appearing in multiple pools where assignment was ambiguous were removed, but labelled as duplicates, so bars add to 100%.



Supplementary Figure 3.5 | Individual polymer enrichment in IL-6 competition experiment. Each point represents a sequence that could be assigned to a pool and was identified in both the starting and eluted mixes.



Supplementary Figure 3.6 | Individual polymer binding to IL-6 example data. Sample SPR sensorgrams of individual polymers in response to 8.8, 27, 80, and then 240 nM injection of IL-6. Data plotted in **Figure 3.4b** were calculated based upon based on the response at the fourth binding point, at the end of the end the final injection of IL-6. Polymer Polar-E1 is the polymer exhibiting the greatest enrichment from any round in the IL-6 selection with the polar polymer library. Polymers Nonpolar-1, Nonpolar Plus-1, and Nonpolar Plus-2 are the 1st-, 1st-, and 2nd-most abundant sequences from their respective pools following IL-6 selection. Data presented are representative of sensorgrams used to derive data in **Figure 3.4c** and **Supplementary Table 3.3**. For this measurement, most polymers were analyzed once. For the above sensorgrams, Nonpolar-1 and Nonpolar Plus 2 were analyzed twice, and Nonpolar Plus-1 was analyzed three times, giving similar results for each measurement.



Supplementary Figure 3.7 | Affinity of selection-enriched polymer pools to IL-6. SPR sensorgrams for single cycle kinetics run of different 3'-biotinylated enriched pools following selection round 7. The polymer pools were bound to the surface of a CAP chip and concentrations of 8.9, 27, 80, and 240 nM IL-6 were flowed over the chip's surface at 30 µL/min. Data plotted in **Figure 3.3c** were plotted based on the response at the third binding point, at the end of the 80 nM injection of IL-6 to minimize any influence of reference binding in the pool comparison. Sensorgrams shown represent the first of 3-4 replicates collected with the enriched polymer libraries and IL-6 shown in **Figure 3.3c**.

Supplementary Note 3.1. Oligos used in selections, HFNAP preparation, and HTS.

Sequences used for selection of PCSK9-binding HFNAPs.

Name	Sequence
Naïve library AZ	CGA ATC AGA TTG GAC CAG YNN YNN YNN YNN YNN YNN YNN YNN YNN YNN YNN YNN YNN GAG TCC AGA TGT AGG TAG
ddC-pp1A	/5phos/GAG TCC AGA TGT AGG TAG/3ddC/
pp2Z	CGA ATC AGA TTG GAC CAG
ExtPrimerA	CTA CCT ACA TCT GGA CTC
BtBt_EPrimA	/52Bio//iSp18/CTA CCT ACA TCT GGA CTC

Sequences used for selection of IL-6-binding HFNAPs.

Name	Sequence
Naïve library BW	CTC GGA TGA ACC TGG ACT YNN YNN YNN YNN YNN YNN YNN YNN YNN YNN YNN YNN YNN GCA TCG AAG CCA AGA TTC
ddC-pp1B	/5Phos/GCATCGAAGCCAAGATT/3ddC/
pp2W	CTC GGA TGA ACC TGG ACT
ExtPrimerB	GAA TCT TGG CTT CGA TGC
BtBt_EPrimerB	/52-Bio//iSp18/GAA TCT TGG CTT CGA TGC

Adapter primers used for Illumina MiSeq analysis of PCSK9 selections.

Name	Sequence
Seq_Adapt_A1	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT N CTA CCT ACA TCT GGA CTC
Seq_Adapt_A2	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT NN CTA CCT ACA TCT GGA CTC
Seq_Adapt_A3	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT NNN CTA CCT ACA TCT GGA CTC
Seq_Adapt_A	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT NNNN CTA CCT ACA TCT GGA CTC
Seq_Adapt_Z1	TGG AGT TCA GAC GTG TGC TCT TCC GAT CT N CGA ATC AGA TTG GAC CAG
Seq_Adapt_Z2	TGG AGT TCA GAC GTG TGC TCT TCC GAT CT NN CGA ATC AGA TTG GAC CAG
Seq_Adapt_Z3	TGG AGT TCA GAC GTG TGC TCT TCC GAT CT NNN CGA ATC AGA TTG GAC CAG
Seq_Adapt_Z	TGG AGT TCA GAC GTG TGC TCT TCC GAT CT NNNN CGA ATC AGA TTG GAC CAG

Adapter primers used for Illumina MiSeq analysis of IL-6 selections.

Name	Sequence
Seq_Adapt_B1	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT N GAA TCT TGG CTT CGA TGC
Seq_Adapt_B2	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT NN GAA TCT TGG CTT CGA TGC
Seq_Adapt_B3	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT NNN GAA TCT TGG CTT CGA TGC
Seq_Adapt_B4	ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT NNNN GAA TCT TGG CTT CGA TGC
Seq_Adapt_W1	TGG AGT TCA GAC GTG TGC TCT TCC GAT CT N CTC GGA TGA ACC TGG ACT
Seq_Adapt_W2	TGG AGT TCA GAC GTG TGC TCT TCC GAT CT NN CTC GGA TGA ACC TGG ACT
Seq_Adapt_W3	TGG AGT TCA GAC GTG TGC TCT TCC GAT CT NNN CTC GGA TGA ACC TGG ACT
Seq_Adapt_W4	TGG AGT TCA GAC GTG TGC TCT TCC GAT CT NNNN CTC GGA TGA ACC TGG ACT

Barcoding adapters used in Illumina MiSeq analysis.

Name	Sequence
PE-REV1	CAAGCAGAAGACGGCATACGAGAT ATCGTGAT GTGACTGGAGTTCAGACGTGTGCTCTTC
PE-REV2	CAAGCAGAAGACGGCATACGAGAT ATACATCG GTGACTGGAGTTCAGACGTGTGCTCTTC
PE-REV3	CAAGCAGAAGACGGCATACGAGAT ATGCCTAA GTGACTGGAGTTCAGACGTGTGCTCTTC
PE-REV4	CAAGCAGAAGACGGCATACGAGAT ATTGGTCA GTGACTGGAGTTCAGACGTGTGCTCTTC
PE-REV5	CAAGCAGAAGACGGCATACGAGAT ATCACTGT GTGACTGGAGTTCAGACGTGTGCTCTTC
PE-REV6	CAAGCAGAAGACGGCATACGAGAT ATATTGGC GTGACTGGAGTTCAGACGTGTGCTCTTC
PE-REV7	CAAGCAGAAGACGGCATACGAGAT ATGATCTG GTGACTGGAGTTCAGACGTGTGCTCTTC
PE-REV8	CAAGCAGAAGACGGCATACGAGAT ATTCAAGT GTGACTGGAGTTCAGACGTGTGCTCTTC

PE-REV9	CAAGCAGAAGACGGCATACGAGAT ATCTGATC GTGACTGGAGTTCAGACGTGTGCTCTTC
PE-REV10	CAAGCAGAAGACGGCATACGAGAT ATAAGCTA GTGACTGGAGTTCAGACGTGTGCTCTTC
PE-REV11	CAAGCAGAAGACGGCATACGAGAT ATGTAGCC GTGACTGGAGTTCAGACGTGTGCTCTTC
PE-REV12	CAAGCAGAAGACGGCATACGAGAT ATTACAAG GTGACTGGAGTTCAGACGTGTGCTCTTC
PE-REV13	CAAGCAGAAGACGGCATACGAGAT TGTTGACT GTGACTGGAGTTCAGACGTGTGCTCTTC
PE-REV14	CAAGCAGAAGACGGCATACGAGAT ACGGAACT GTGACTGGAGTTCAGACGTGTGCTCTTC
PE-REV15	CAAGCAGAAGACGGCATACGAGAT TCTGACAT GTGACTGGAGTTCAGACGTGTGCTCTTC
PE-REV16	CAAGCAGAAGACGGCATACGAGAT CGGGACGG GTGACTGGAGTTCAGACGTGTGCTCTTC
PE-REV18	CAAGCAGAAGACGGCATACGAGAT GTGCGGAC GTGACTGGAGTTCAGACGTGTGCTCTTC
PE-REV19	CAAGCAGAAGACGGCATACGAGAT CGTTTCAC GTGACTGGAGTTCAGACGTGTGCTCTTC
PE-REV20	CAAGCAGAAGACGGCATACGAGAT AAGGCCAC GTGACTGGAGTTCAGACGTGTGCTCTTC
PE-REV21	CAAGCAGAAGACGGCATACGAGAT TCCGAAAC GTGACTGGAGTTCAGACGTGTGCTCTTC
PE-REV22	CAAGCAGAAGACGGCATACGAGAT TACGTACG GTGACTGGAGTTCAGACGTGTGCTCTTC
PE-REV23	CAAGCAGAAGACGGCATACGAGAT ATCCACTC GTGACTGGAGTTCAGACGTGTGCTCTTC
PE-REV25	CAAGCAGAAGACGGCATACGAGAT AAAGGAAT GTGACTGGAGTTCAGACGTGTGCTCTTC
PE-REV27	CAAGCAGAAGACGGCATACGAGAT ATATCAGT GTGACTGGAGTTCAGACGTGTGCTCTTC
PE_REV-18B	CAAGCAGAAGACGGCATACGAGAT GCGTACGT GTGACTGGAGTTCAGACGTGTGCT
PE_REV-19B	CAAGCAGAAGACGGCATACGAGAT TTTCACCG GTGACTGGAGTTCAGACGTGTGCT
PE_REV-20B	CAAGCAGAAGACGGCATACGAGAT GGCCACAA GTGACTGGAGTTCAGACGTGTGCT
PE_REV-21B	CAAGCAGAAGACGGCATACGAGAT CGAAACTC GTGACTGGAGTTCAGACGTGTGCT
PE_REV-22B	CAAGCAGAAGACGGCATACGAGAT CGTACGTA GTGACTGGAGTTCAGACGTGTGCT
PE_REV-23B	CAAGCAGAAGACGGCATACGAGAT CCACTCAT GTGACTGGAGTTCAGACGTGTGCT
PE_REV-25B	CAAGCAGAAGACGGCATACGAGAT AGGAATAA GTGACTGGAGTTCAGACGTGTGCT
PE_REV-27B	CAAGCAGAAGACGGCATACGAGAT ATCAGTAT GTGACTGGAGTTCAGACGTGTGCT
FWD-2a	AATGATACGGCGACCACCGAGATCTACAC ATTACTCG ACACTCTTTCCCTACACGAC
FWD-2b	AATGATACGGCGACCACCGAGATCTACAC TCCGGAGA ACACTCTTTCCCTACACGAC
FWD-2c	AATGATACGGCGACCACCGAGATCTACAC CGCTCATT ACACTCTTTCCCTACACGAC
FWD-2d	AATGATACGGCGACCACCGAGATCTACAC GAGATTCC ACACTCTTTCCCTACACGAC
FWD-2e	AATGATACGGCGACCACCGAGATCTACAC ATTCAGAA ACACTCTTTCCCTACACGAC
FWD-2f	AATGATACGGCGACCACCGAGATCTACAC GAATTCGT ACACTCTTTCCCTACACGAC
FWD-2g	AATGATACGGCGACCACCGAGATCTACAC CTGAAGCT ACACTCTTTCCCTACACGAC
FWD-2h	AATGATACGGCGACCACCGAGATCTACAC TAATGCGC ACACTCTTTCCCTACACGAC
FWD-2i	AATGATACGGCGACCACCGAGATCTACAC CGGCTATG ACACTCTTTCCCTACACGAC
FWD-2j	AATGATACGGCGACCACCGAGATCTACAC TCCGCGAA ACACTCTTTCCCTACACGAC
FWD-2k	AATGATACGGCGACCACCGAGATCTACAC TCTCGCGC ACACTCTTTCCCTACACGAC
FWD-2l	AATGATACGGCGACCACCGAGATCTACAC AGCGATAG ACACTCTTTCCCTACACGAC

Sequences for SPR analysis derived from the unfunctionalized pool of PCSK9 selection.

Template Name	Template
Rnd12-1_seq1	CGA ATC AGA TTG GAC CAG TTA CCG CCC TCT CCC TTT TGT CCC CTC CGT TGT TTC TCT TTC CCT GAG TCC AGA TGT AGG TAG /iSp18//3Bio/
Rnd12-1_seq2	CGA ATC AGA TTG GAC CAG CCG CTT TTG CCT TCT TAC CCT TTT TGT CAT CCC TGT TGT TTC CCT GAG TCC AGA TGT AGG TAG /iSp18//3Bio/
Rnd12-1_seq3	CGA ATC AGA TTG GAC CAG CAC CCT CCG TCT CCG CTC CCC TCG TCT TGT TGT TTC CCT TTC CCT GAG TCC AGA TGT AGG TAG /iSp18//3Bio/
PCSK9-MOB-1_seq1	CGA ATC AGA TTG GAC CAG CAC TTC TCA TTA TTG TTC TTA TGA CCT TGC CCG TTT TCA CCC CCT GAG TCC AGA TGT AGG TAG /iSp18//3Bio/
PCSK9-MOB-1_seq2	CGA ATC AGA TTG GAC CAG CGC TTA CCA TTG TCT TGT TTG CCT TCC CCT TTG TCC TGT TGT TTA GAG TCC AGA TGT AGG TAG /iSp18//3Bio/
PCSK9-MOB-1_seq3	CGA ATC AGA TTG GAC CAG CGC TTT CTC TTT TTC CCT TTT CCA TGC CCA TGT TTA TCT TTC CGC GAG TCC AGA TGT AGG TAG /iSp18//3Bio/
PCSK9-MOB-1_seq4	CGA ATC AGA TTG GAC CAG CTT TCT TGT CGC CGC CCT CCT TGT TGT TGA TGT CCT TCC TTT CCT GAG TCC AGA TGT AGG TAG /iSp18//3Bio/
PCSK9-MOB-1_seq5	CGA ATC AGA TTG GAC CAG TGC TTT TTC CTC TCT TAT CAC TTC TTG TTG CTT TGT TGT TTC TCT GAG TCC AGA TGT AGG TAG /iSp18//3Bio/

Templates used to synthesize HFNAPs from the charged pool derived from PCSK9 selection for SPR analysis.

Template Name	Template

Rnd12-2_seq1-temp	CTA CCT ACA TCT GGA CTC AGG ACA ACG TGG GGG TGA AGG GGG ACA ACA AGG GTG GCA ACA AAA CTG GTC CAA TCT GAT TCG
Rnd12-2_seq2	CGA ATC AGA TTG GAC CAG TTT CAT CCC TTT TTG CCA CCC CTT TGT TGT CCA CCC CCA TGT CCT GAG TCC AGA TGT AGG TAG /iSp18//3Bio/
Rnd12-2_seq3	CGA ATC AGA TTG GAC CAG TTT CTC CCT CTT CCA CAC CTT TGT CTT TGC CAC CCT CCT TGT CCT GAG TCC AGA TGT AGG TAG /iSp18//3Bio/
Rnd12-2_seq4-temp	CTA CCT ACA TCT GGA CTC CGG TGG AGG GTG CGG TTG GGA GAG TGA AAG ACG TTA GCA AAA TCA CTG GTC CAA TCT GAT TCG
PCSK9-MOB-2_seq1	CGA ATC AGA TTG GAC CAG CAC TAG CCT TTT TGC CCC TTT CTT TGT CGT CTT CCC TGT TTT CTC GAG TCC AGA TGT AGG TAG /iSp18//3Bio/
PCSK9-MOB-2_seq2	CGA ATC AGA TTG GAC CAG CGT TTT TTG CTT CTT CTA TGC CCT TTT CTC TGC CCG TTT CCC CGC GAG TCC AGA TGT AGG TAG /iSp18//3Bio/
PCSK9-MOB-2_temp3	CTA CCT ACA TCT GGA CTC GAG ACA AAA ACA AGG GTG ACA AGG GGA AAA ACG AGG GGG CAA CTA CTG GTC CAA TCT GAT TCG
PCSK9-MOB-2_temp4	CTA CCT ACA TCT GGA CTC AGG GGG TGA AAA GGG GCA CGG AAG GGG TGA AGA ACA AAG AAA AAG CTG GTC CAA TCT GAT TCG
PCSK9-MOB-2_temp5	CTA CCT ACA TCT GGA CTC AGG AGG ACA AAA ACA TGG AAA CCG GGG ATA AAG CGG GGA AAA CGG CTG GTC CAA TCT GAT TCG

Templates used to synthesize HFNAPs from the polar pool derived from PCSK9 selection for SPR analysis.

Template Name	Template
Rnd12-3_seq1-temp	CTA CCT ACA TCT GGA CTC AGG ACA ACA CGG GGA AGG AGG CGG GAG TGG AGG CGA GGG CAA CAA CTG GTC CAA TCT GAT TCG
Rnd12-3_seq2-temp	CTA CCT ACA TCT GGA CTC AGG ACA ACA TGG CAA TAG GGA GAG GCA GGG GGA AGG AAG AAA GGG CTG GTC CAA TCT GAT TCG
Rnd12-3_seq3-temp	CTA CCT ACA TCT GGA CTC AGG GGG ACA ACA AGG AAA GAG GGA AAG AGG CGG TGG TGG CGG AAA CTG GTC CAA TCT GAT TCG
Rnd12-3_seq4	CGA ATC AGA TTG GAC CAG TTT TTT CCT TGT CAG CCT CTC TTG CCC CCG CCT CCT TGT TGT CCT GAG TCC AGA TGT AGG TAG /iSp18//3Bio/
PCSK9-MOB-3_temp1	CTA CCT ACA TCT GGA CTC CTA ACG AGA ACA AGG GGA AAA CGG GAG ATA CCG GGA GGA AAA AGG CTG GTC CAA TCT GAT TCG
PCSK9-MOB-3_temp2	CTA CCT ACA TCT GGA CTC AGG ACA GGG AAA AAG ACA ATA AAA CGG TGG GAG AAA AGG GGA AAA CTG GTC CAA TCT GAT TCG
PCSK9-MOB-3_temp3	CTA CCT ACA TCT GGA CTC CTA ACG AGA ACA AAA AAG ACA GGA TCA GAG AAG AAA TGG GTA CCG CTG GTC CAA TCT GAT TCG
PCSK9-MOB-3_temp4	CTA CCT ACA TCT GGA CTC CTA ACG AGA ACA AGG GGG AAA ACA AGG AAA ATG TGA AGA ATA CTG CTG GTC CAA TCT GAT TCG
PCSK9-MOB-3_temp5	CTA CCT ACA TCT GGA CTC GTA GAG AGG AAG AAA ACA GGG GGA AAA ACG AGG GTA GTG CCG TCA CTG GTC CAA TCT GAT TCG

Templates used to synthesize HFNAPs from the polar plus pool derived from PCSK9 selection for SPR analysis.

Template Name	Template
Rnd12-4_seq1-temp	CTA CCT ACA TCT GGA CTC TAG CAA GGG AGG AAG AAG AAA AGG CAA AAA AGG AAA AAA GGG TGG CTG GTC CAA TCT GAT TCG
Rnd12-4_seq2-temp	CTA CCT ACA TCT GGA CTC TAG TGG AAA GGG ACA GGG ACA AAA GGG GAG AAA AGG AAA TAG CGA CTG GTC CAA TCT GAT TCG
Rnd12-4_seq3-temp	CTA CCT ACA TCT GGA CTC AGG GTG GCG GGG CAA AGG CAA ACG AGG GCA GAG AAA ACA AAA GGG CTG GTC CAA TCT GAT TCG
PCSK9-MOB-4_temp1	CTA CCT ACA TCT GGA CTC AGG AAA AAG GGG GAG AAA AGG GCA AAA AGG CGG AAA AAA GTA CGA CTG GTC CAA TCT GAT TCG
PCSK9-MOB-4_temp2	CTA CCT ACA TCT GGA CTC TGG CGG AAG AGG CAA GGA AGG GTG AAA GGG CAA AAA ACA AAA GCG CTG GTC CAA TCT GAT TCG
PCSK9-MOB-4_temp3	CTA CCT ACA TCT GGA CTC AGG GAG AAG AAA CAA AGG ACA AAA GGG GGG GCG CGA CGA AGG TGA CTG GTC CAA TCT GAT TCG
PCSK9-MOB-4_temp4	CTA CCT ACA TCT GGA CTC GCG GGG CAA ACA AGG AAA AGG CAA AAA ACA GGG AAA AAA TAG AAG CTG GTC CAA TCT GAT TCG
PCSK9-MOB-4_temp5	CTA CCT ACA TCT GGA CTC AGG ACA TGG GTG AAA GGG AAG AAA GTG GGA AAA AAA CAA CAG AGG CTG GTC CAA TCT GAT TCG

Templates used to synthesize HFNAPs from the nonpolar pool derived from PCSK9 selection for SPR analysis.

Template Name	Template
12-5_seq1-temp	CTA CCT ACA TCT GGA CTC CGG AGG TGA ACA ACG ATG CGA GAG GCA AGA TAG ATA TCG AAA CTA CTG GTC CAA TCT GAT TCG
Rnd12-5_seq2-temp	CTA CCT ACA TCT GGA CTC TAG AGA ACA GAA GGA ACA GAA TGG AAA AGA AAG TGA GAA ACG AAA CTG GTC CAA TCT GAT TCG
Rnd12-5_seq3-temp	CTA CCT ACA TCT GGA CTC CGG AGG TGA ACA ACG ATG CGA GAG GCA AGA TAG GTA TCG AAA CTA CTG GTC CAA TCT GAT TCG
PCSK9-MOB-5_temp1	CTA CCT ACA TCT GGA CTC CGG AGG TGA ACA ACG ATG CGA GAG GCA GGA TAG ATA TCG AAA CTA CTG GTC CAA TCT GAT TCG
PCSK9-MOB-5_temp3	CTA CCT ACA TCT GGA CTC CGG AGG TGA ACA ACG ATG CGA GAG GCA AGA TAG ATA TCG TAA CTA CTG GTC CAA TCT GAT TCG
PCSK9-MOB-5_temp4	CTA CCT ACA TCT GGA CTC CGA TGG GTA TAG CAA AGA ACA GAA GGG AAA AGA AAG CAA ACG AAA CTG GTC CAA TCT GAT TCG
PCSK9-MOB-5_temp5	CTA CCT ACA TCT GGA CTC CGG AGG GGA ACA ACG ATG CGA GAG GCA AGA TAG ATA TCG AAA CTA CTG GTC CAA TCT GAT TCG
PCSK9-MOB-5_temp7	CTA CCT ACA TCT GGA CTC CGG AGG TGA ACA ACG ATG CGA GAG GCA AGA TAG ATA TCA AAA CTA CTG GTC CAA TCT GAT TCG

Templates used to synthesize HFNAPs from the nonpolar plus pool derived from PCSK9 selection for SPR analysis.

Template Name	Template
Rnd12-6_seq1-temp	CTA CCT ACA TCT GGA CTC GGA CGG GGG ATA GGG GAA GGA CAG GGG AGG GCA ATG CAA ACG CTG CTG GTC CAA TCT GAT TCG
Rnd12-6_seq2-temp	CTA CCT ACA TCT GGA CTC AGG GAA CAA AGG GGA ACA GGA AAA GGG GGA ACA ACA AGA GAG GGG CTG GTC CAA TCT GAT TCG
Rnd12-6_seq3-temp	CTA CCT ACA TCT GGA CTC TTA ACA GGA ACA CAG GGG GAA AGA TAA GGG GGA CAA ACA GAG GTA CTG GTC CAA TCT GAT TCG
Rnd12-6_seq4-temp	CTA CCT ACA TCT GGA CTC TCA AGG GAA ACA AGG GGG CGA GGA GAG GAA ACA ACA AAA GGG CAA CTG GTC CAA TCT GAT TCG
PCSK9-MOB-6_temp1	CTA CCT ACA TCT GGA CTC GCA AGG GAA ACA AGG GGG CGA GGA GAG GAA ACA ACA AAA GGG CAA CTG GTC CAA TCT GAT TCG
PCSK9-MOB-6_temp2	CTA CCT ACA TCT GGA CTC GAG ATA ACA CAG GCG TGG GCG GTA AGG GGA AAA GGG CAG TCG CGG CTG GTC CAA TCT GAT TCG
PCSK9-MOB-6_temp3	CTA CCT ACA TCT GGA CTC GGA GGG GGA TAA ACA GGA AGA CAA GGA AGG GAA ACA CAA CAG AAA CTG GTC CAA TCT GAT TCG
PCSK9-MOB-6_temp4	CTA CCT ACA TCT GGA CTC TAG TGA GGG TAA ACA GAA GAA AAA CGG GGG AGG AAA AAG AAA GCA CTG GTC CAA TCT GAT TCG
PCSK9-MOB-6_temp5	CTA CCT ACA TCT GGA CTC TAA GGG GGA AGA GGG GGG AAG ACA AAA AGG GAA ACA TGA AGA AAG CTG GTC CAA TCT GAT TCG

Templates used to synthesize HFNAPs from the fully functionalized pool derived from PCSK9 selection for SPR analysis.

Template Name	Template
Rnd12-7_seq1	CTA CCT ACA TCT GGA CTC TGG AGG GTA GAG CAG GCA GAG AGG GAG GAG AGG GCA CAA CAA ACG CTG GTC CAA TCT GAT TCG
Rnd12-7-seq2	CTA CCT ACA TCT GGA CTC TGG AGG GTA GAG CAG GCA GAG CGG GAG GAG AGG GCA CAA CAA ACG CTG GTC CAA TCT GAT TCG
Rnd12-7-seq3	CTA CCT ACA TCT GGA CTC ACA ATA AGG GGG ACG TGG GGG AAA ACA GGG AAG AAA AAA CGG AGG CTG GTC CAA TCT GAT TCG
Rnd12-7-seq4	CTA CCT ACA TCT GGA CTC AGG GGG GAG ACA ACA GGA GGG AAA ACA GGG GGA TGG ATA ACA AAA CTG GTC CAA TCT GAT TCG
PCSK9-MOB-7_temp1	CTA CCT ACA TCT GGA CTC TAA CAA AGG GGG AAG ACA GGG AAG AAA ACA GGG AGG GGA AAG CGA CTG GTC CAA TCT GAT TCG
PCSK9-MOB-7_temp2	CTA CCT ACA TCT GGA CTC GAG AAG GGG AGG GAG AGG CAA GGG GGA AAG GCA AAA CGA CAA AAA CTG GTC CAA TCT GAT TCG
PCSK9-MOB-7_temp3	CTA CCT ACA TCT GGA CTC TGG AAA ATA AGG GCA ATG GGA AAA ACA ACA AAG AAA AGG GGG CGA CTG GTC CAA TCT GAT TCG
PCSK9-MOB-7_temp4	CTA CCT ACA TCT GGA CTC TGA AAG GGG ACA AGG ACA AGG GAG AGG AAA ACA GGG AAG GCA AAG CTG GTC CAA TCT GAT TCG

PCSK9-MOB-7_temp5	CTA CCT ACA TCT GGA CTC TGG TCA AAG GGG CAG GGG CAA ACA AGG GCA GGG AAA AAG CGG AAA CTG GTC CAA TCT GAT TCG

Sequences for SPR analysis derived from the unfunctionalized pool of IL-6 selection.

Template Name	Template
IL6-R7-1_seq1	CTC GGA TGA ACC TGG ACT TTA TTG CAC TGG CTG TAC TTT TCG TAT CCA TGG TCA TCT TTT CTA GCA TCG AAG CCA AGA TTC /iSp18//3Bio/
IL6-R7-1_seq2	CTC GGA TGA ACC TGG ACT TTA CTG CGG CGC CAG CCG TGC CCA TCC TTT CTT CAT CCG TGC CCA GCA TCG AAG CCA AGA TTC /iSp18//3Bio/
IL6-R7-1_seq3	CTC GGA TGA ACC TGG ACT CTG TTT TTA CTA CCG CGC TAG CCG CCT CAT CCG TTT TAG CGC CAT GCA TCG AAG CCA AGA TTC /iSp18//3Bio/
IL6-R7-1_seq4	CTC GGA TGA ACC TGG ACT TAA TGA TGC TCA TGG TGC TCT TCA TCA CCA TCT TCG TTC CAT CCT GCA TCG AAG CCA AGA TTC /iSp18//3Bio/
IL6-R7-1_seq5	CTC GGA TGA ACC TGG ACT TCT TCA TTG CGC TGG TGC CTG CCA TCC TTG TTC CAT CCC CTT TGC GCA TCG AAG CCA AGA TTC /iSp18//3Bio/
IL6-MOB-1_seq1	CTC GGA TGA ACC TGG ACT TTA TTG CAC TGG CTG TAC TTT TCG TAT CCA TGG TCA TCT TTC CTA GCA TCG AAG CCA AGA TTC /iSp18//3Bio/
IL6-MOB-1_seq2	CTC GGA TGA ACC TGG ACT TTA CTG CGG CGC CAG CCG TGC CCA TCC TTT CTC CAT CCG TGC CCA GCA TCG AAG CCA AGA TTC /iSp18//3Bio/
IL6-MOB-1_seq3	CTC GGA TGA ACC TGG ACT TTA TTG CAC TGG CTG TAC TTT TCG TAT CCA TGG TCA TCC TTT CTA GCA TCG AAG CCA AGA TTC /iSp18//3Bio/
IL6-MOB-1_seq4	CTC GGA TGA ACC TGG ACT TTA TTG CAC TGG CTG TAC TTC TCG TAT CCA TGG TCA TCT TTT CTA GCA TCG AAG CCA AGA TTC /iSp18//3Bio/
IL6-MOB-1_seq5	CTC GGA TGA ACC TGG ACT TTA TTG CAC TGG CTG TAC TTT TCG TAT CCA TGG TCA TCT TTT CTG GCA TCG AAG CCA AGA TTC /iSp18//3Bio/

Templates used to synthesize HFNAPs from the charged pool derived from IL-6 selection for SPR analysis.

Template Name	Template
IL6-R7-2_seq1-temp	GAA TCT TGG CTT CGA TGC CCG GTG GAA GCA ACA ACA CTA ACG CCA GGA CCA TAA GCA TTG CAG AGT CCA GGT TCA TCC GAG
IL6-R7-2_seq2-temp	GAA TCT TGG CTT CGA TGC GAG CCG CTA GCG CGG TAA TAG AAA TCG GGA GCA CGA CCA TCG CGA AGT CCA GGT TCA TCC GAG
IL6-R7-2_seq3-temp	GAA TCT TGG CTT CGA TGC CCG GTG GAA GCA ACA ACA CTA CCG CCA GGA CCA TAA GCA TTG CAG AGT CCA GGT TCA TCC GAG
IL6-R7-2_seq4-temp	GAA TCT TGG CTT CGA TGC ACG AAG AGG GCG GGG TGA CAA GCG CAG GCA GCA TAG AAG CCA CTA AGT CCA GGT TCA TCC GAG
IL6-MOB-2_temp2	GAA TCT TGG CTT CGA TGC GAG CCG CTA GCG CGG TAA TAG AAA TCG GGA CCA CGA CCA TCG CGA AGT CCA GGT TCA TCC GAG
IL6-MOB-2_temp3	GAA TCT TGG CTT CGA TGC GAG CCG CTA GCG CGG TAA TAG AAA TCG GGA GTA CGA CCA TCG CGA AGT CCA GGT TCA TCC GAG
IL6-MOB-2_temp5	GAA TCT TGG CTT CGA TGC CCG GTG GAA GCA ACA ACA CTA ACG CCA GGA CCA TAA TCA TTG CAG AGT CCA GGT TCA TCC GAG
IL6-MOB-2_temp6	GAA TCT TGG CTT CGA TGC GAG CCG CTA GCG CGG TAA TAG AAA TCG GGG GCA CGA CCA TCG CGA AGT CCA GGT TCA TCC GAG
IL6-MOB-2_temp7	GAA TCT TGG CTT CGA TGC GAG CCG CTA GCG CGG TAA TAG AAA TCG GGA TCA CGA CCA TCG CGA AGT CCA GGT TCA TCC GAG

Templates used to synthesize HFNAPs from the polar pool derived from IL-6 selection for SPR analysis.

Template Name	Template
IL6-R7-3_seq1-temp	GAA TCT TGG CTT CGA TGC GGG CGG GCA CCG AGA TCA TGA GAG GTA ACA ACA TTA AGG GGG CCA AGT CCA GGT TCA TCC GAG
IL6-R7-3_seq2-temp	GAA TCT TGG CTT CGA TGC GGA TGG GGG CAA GGG AAG GAA CAG TAA GAA CGA AAG AAA TTG CCG AGT CCA GGT TCA TCC GAG
IL6-R7-3_seq3-temp	GAA TCT TGG CTT CGA TGC CGG AGG ATG GAA CAG TAG GAA CGA AGG ACA TCA GAG CCA ACA TGA AGT CCA GGT TCA TCC GAG
IL6-MOB-3_temp1	GAA TCT TGG CTT CGA TGC GGA TGG GGG CAA GGG AAG GAA CAG TAA GAA CGA GAG AAA TTG CCG AGT CCA GGT TCA TCC GAG
IL6-MOB-3_temp2	GAA TCT TGG CTT CGA TGC CGG AGG GTG GAA CAG TAG GAA CGA AGG ACA TCA GAG CCA ACA TGA AGT CCA GGT TCA TCC GAG

IL6-MOB-3_temp4	GAA TCT TGG CTT CGA TGC GGG CGG GCA CCG AGA TCA TGA GAG GTA ACA ACA TTA GGG GGG CCA AGT CCA GGT TCA TCC GAG
IL6-MOB-3_temp5	GAA TCT TGG CTT CGA TGC CGG AGG ATG GAA CAG TAG GAA CGA AGG ACA TCA GAG CCA ACA AGA AGT CCA GGT TCA TCC GAG
IL6-MOB-3_temp7	GAA TCT TGG CTT CGA TGC CGG AGG ATG GAA CAG TAG GAA CGA AGG GCA TCA GAG CCA ACA TGA AGT CCA GGT TCA TCC GAG

Templates used to synthesize HFNAPs from the polar plus pool derived from IL-6 selection for SPR analysis.

Template Name	Template
IL6-R7-4_seq1-temp	GAA TCT TGG CTT CGA TGC CAG TCG ACA TCA TGA ACG GGG ATG ACA GCA CAA TGG CAA TCG AAA AGT CCA GGT TCA TCC GAG
IL6-R7-4_seq2-temp	GAA TCT TGG CTT CGA TGC CAG GGG ATG AGA TCA CAA GCA AAA ATA CAG CAA CAG TTA AGG GAA AGT CCA GGT TCA TCC GAG
IL6-R7-4_seq3-temp	GAA TCT TGG CTT CGA TGC ACG GGG GAA AGG GTG GAA CAG TAA GAA CGA AAA AAG CCG TGA TGG AGT CCA GGT TCA TCC GAG
IL6-MOB-4_temp1	GAA TCT TGG CTT CGA TGC CCG GAG CAA TCG CCA CAG CAA AGA TCA CCA GTA TTA TGA TAA AGG AGT CCA GGT TCA TCC GAG
IL6-MOB-4_temp2	GAA TCT TGG CTT CGA TGC ACA GTG GAA ACA CAG TCG GGA GGA CCG TAG TCG ACA CAG TTA TGG AGT CCA GGT TCA TCC GAG
IL6-MOB-4_temp3	GAA TCT TGG CTT CGA TGC CAG TCG ACA TCA TGA GCG GGG ATG ACA GCA CAA TGG CAA TCG AAA AGT CCA GGT TCA TCC GAG
IL6-MOB-4_temp4	GAA TCT TGG CTT CGA TGC GCG ACG GAG AAG GAA CAG TAA GAA CGA AAA CAA AAG CAG CAA CCG AGT CCA GGT TCA TCC GAG
IL6-MOB-4_temp7	GAA TCT TGG CTT CGA TGC GGG TGA GGG GTG AGG CAG AGA TGG AGA TTG AAG CCA ATA CAG GGA AGT CCA GGT TCA TCC GAG

Templates used to synthesize HFNAPs from the nonpolar pool derived from IL-6 selection for SPR analysis.

Template Name	Template
IL6-R7-5_seq1-temp	GAA TCT TGG CTT CGA TGC CCG GAG CAA TCG CCA CAG CAA AGA TCA CCA GTA TTA TGA TAA AGG AGT CCA GGT TCA TCC GAG
IL6-R7-5_seq2-temp	GAA TCT TGG CTT CGA TGC CCG GAG CAA TCG CCA CAG CCA ACA TAA ACG CAA AGG GTA TAA GTA AGT CCA GGT TCA TCC GAG
IL6-R7-5_seq3-temp	GAA TCT TGG CTT CGA TGC CCG GAG CAA TCG CCA CAG CCA TAG ATA GGA CCA AAA CCA TGG GAG AGT CCA GGT TCA TCC GAG
IL6-MOB-5_temp1	GAA TCT TGG CTT CGA TGC CAG TAG TAA ACA AAA GCA CAG ATA TCG CCA GCA AAG AGA ATA GAA AGT CCA GGT TCA TCC GAG
IL6-MOB-5_temp2*	GAA TCT TGG CTT CGA TGC CCG GAG CAA TCG CCA CAA CCA AGG GAA CAA GGA AGA ACA CTA AAA AGT CCA GGT TCA TCC GAG
IL6-MOB-5_temp4	GAA TCT TGG CTT CGA TGC CCG GAG CAA TCG CCA CAG CCA ACA TAA ACG CAA AGG GTA CAA GAA AGT CCA GGT TCA TCC GAG
IL6-MOB-5_temp5	GAA TCT TGG CTT CGA TGC TCG GGA GAG TGA GGG GCA CGG GGA GAG GGA CAA CGA ATA CCA CAG AGT CCA GGT TCA TCC GAG
IL6-MOB-5_temp6	GAA TCT TGG CTT CGA TGC CCG GAG CAA TCG CCA CAG CAA AGA TCA CCA GCA TTA TGA CAA AGG AGT CCA GGT TCA TCC GAG

* Translation did not yield sufficient full-length product

Templates used to synthesize HFNAPs from the nonpolar plus pool derived from IL-6 selection for SPR analysis.

Template Name	Template
IL6-R7-6_seq1-temp	GAA TCT TGG CTT CGA TGC CCA TAG TCA TCA CTA AAA GCG CGA GAA GAG CCG TTG TGG CGG TGA AGT CCA GGT TCA TCC GAG
IL6-R7-6_seq2-temp	GAA TCT TGG CTT CGA TGC CCG GAG CAA TCA GAA GCA GGA AGA GAA AAA CAA CAA GGG TCG GCG AGT CCA GGT TCA TCC GAG
IL6-R7-6_seq3-temp	GAA TCT TGG CTT CGA TGC GAA GCA TCA TGA AGG AGG GTA TCA GCG CGG GCG CGA CAA CCA CCA AGT CCA GGT TCA TCC GAG
IL6-R7-6_seq4-temp	GAA TCT TGG CTT CGA TGC CCG GAG CAA TCA TCA CCA CAA GGA AGA ATA TAA GCA ATA GAA TAG AGT CCA GGT TCA TCC GAG
IL6-MOB-6_temp1	GAA TCT TGG CTT CGA TGC CCA TGG ACG ACA CCA CAA TAG CAA GTA GCA AAG AAA TCA GCG CCA AGT CCA GGT TCA TCC GAG

IL6-MOB-6_temp2	GAA TCT TGG CTT CGA TGC CCG GAG CAA TCA GAA GCA TGA GTA TGA AGA CAA AAA AAA CGG CGG AGT CCA GGT TCA TCC GAG
IL6-MOB-6_temp3	GAA TCT TGG CTT CGA TGC GCA GCG TGG TAG ATA TGA CAG GAG AGG CCA CTG AAA GCG CCA CCA AGT CCA GGT TCA TCC GAG
IL6-MOB-6_temp5	GAA TCT TGG CTT CGA TGC CCG GAG CAA TCA GAA GCA GGA AGA GAA AAA CAA CAA GGG GCG GCG AGT CCA GGT TCA TCC GAG
IL6-MOB-6_temp6	GAA TCT TGG CTT CGA TGC CCG GAG CAA TCA TCA CCA CAA GGA AGA ATA TAA GCA ACA GAA TAG AGT CCA GGT TCA TCC GAG

Templates used to synthesize HFNAPs from the fully functionalized pool derived from IL-6 selection for SPR analysis.

Template Name	Template
IL6-R7-7_seq1-temp	GAA TCT TGG CTT CGA TGC GCA TAG GTA CAA TGG GCG CGA GGG GCA CAG TCG AGA AGG CCG TGA AGT CCA GGT TCA TCC GAG
IL6-R7-7_seq2-temp	GAA TCT TGG CTT CGA TGC GCG GAG ACG CAA CAG CAA AAA GCA CAA CCA CAA ACA AGA ACA CGA AGT CCA GGT TCA TCC GAG
IL6-R7-7_seq3-temp	GAA TCT TGG CTT CGA TGC GCA TAG GTA CAA TGG GCG CGA GGG GCA CAG TCG AGA AGG TCG TGA AGT CCA GGT TCA TCC GAG
IL6-R7-7_seq4-temp	GAA TCT TGG CTT CGA TGC GCA TAG GTA CAA TGG GCG CGA GGG GCA CAG TCG AGA AGG CCG CGA AGT CCA GGT TCA TCC GAG
IL6-MOB-7_temp1	GAA TCT TGG CTT CGA TGC GCA TAG GTA CAA TGG GCG CGA GGG GCA CAG TCG AGA CGG CCG TGA AGT CCA GGT TCA TCC GAG
IL6-MOB-7_temp5*	GAA TCT TGG CTT CGA TGC TGG GTA AAA TGG ATG GAG TAA ACG ACG CCA AGA AGG TGA TAG AGA AGT CCA GGT TCA TCC GAG
IL6-MOB-7_temp6	GAA TCT TGG CTT CGA TGC GCA TAG GTA CAA TGG GCG CGA GGG GCA CAG TCG CGA AGG CCG TGA AGT CCA GGT TCA TCC GAG
IL6-MOB-7_temp7*	GAA TCT TGG CTT CGA TGC TGG GTA AAA TGG ATG GAG TAA ACG ACG CCA AGA AGG AGA TAG AGA AGT CCA GGT TCA TCC GAG
IL6-MOB-7_temp8	GAA TCT TGG CTT CGA TGC CCA AGA GGA CCG CGG GGA TAA CCG GAA GCA TAA CCG AAA ACG CGA AGT CCA GGT TCA TCC GAG

* Translation did not yield full-length product

Supplementary Note 3.2.

Description of pool assignment protocol for HTS sequence data from head-to-head competition. Filtered sequences of the appropriate length were obtained from HTS of the mixed samples and the eluted material. These data were used as input files containing a list of the DNA sequences, ordered by relative abundance, along with the number of reads for each sequence in the sample. Each of the unique sequences from these lists were matched first against the HTS data gathered after each of the individual pools was subjected to the final round of parallel selection (e.g., round 12 for the PCSK9 selection). Another data file was produced from this analysis containing each of the unique sequences from the head-to-head competition sample, ranked by abundance, along with a number corresponding to the pool to which it was assigned, the number of reads in the sample, and a list of the pools from which the sequence appeared. Of note, some of the most abundant sequences also appeared at low frequency in other pools, presumably because the probability of contamination increases with higher abundance. At the end of output file is a list of the total number of reads assigned to each starting pool and another list containing the fraction of the total ascribed to each pool. Also included is the percentage of sequences discarded because they appeared in relatively low abundance in several pools and the percentage of sequences that were unmatched. These unmatched sequences were also output in a separate file.

For the bulk library comparison, to make additional assignments, the unmatched sequences from the head-to-head-competition samples were then matched against a much larger list of polymers from each pool containing the entire list of sequences that had been observed in any round of selection (i.e., sequences that were also found from rounds 1-11 in the PCSK9 selection).

All code was run using Python 2.7.10.

Supplementary Note 3.3. Code used for pool assignment of sequences from head-to-head competition data.

```
#!/usr/bin/env python
import sys, math, collections, time
from numpy import *
from collections import Counter
from decimal import *

# to run this program, type something like python tester.py List_of_sequences output_file

try:
    infilename = sys.argv[1]        # This will input the file containing the big list of sequences to be analyzed
    infilenametwo = sys.argv[2]     # This inputs the file where the list of sequences will be pulled, i.e. the file with sequences from an
individual pool
    outfilename = sys.argv[3]        # This inputs the file where the matches will go
    outfilelast = sys.argv[4]    # This is where the record will be created

except:
    print "Usage:",sys.argv[0], "1_sequence_to_analyze 2_Filelist_file 3_Output_of_matches 4_record_file"; sys.exit(1)

# Sample input:
# python assignall6_mob.py Eluted_data_seqinterp allpools_combined Eluted_data_MOBcomplete_NoDup
assignmentrecordedXXMix-NoDup

ifile = open(infilename, 'r')  # open sequence list file from combined pools
ifile2 = open(infilenametwo, 'r') #opens the list of sequence files to look at
recordfile = open(outfilelast, 'a') # open a file where info will be written for a number of sequences
matchfile = open(outfilename, 'w')

unmatchedfilelist = [outfilename, 'unmatched']
unmatchedfilename = '_'.join(unmatchedfilelist)
unmatchedfile = open(unmatchedfilename, 'w')

results = ifile.readlines(); ifile.close() #opens combined pool
files = ifile2.readlines(); ifile2.close()
timestr = time.strftime("%Y-%m-%d %H:%M:%S")

matchfile.write("> %s %s: %s %s %s %s %s %s\n" % (timestr, sys.argv[0], "In", len(results)-1, "sequences of ", infilename,
"searching for matches from", files))        # stamps the date and time on the output count file
matchfile.write("%s\n \n" % ("index, sequence, best rank, pools in which sequence appears, (rank in pool different pools)"))

matches = []

pool1 = []
pool2 = []
pool3 = []
pool4 = []
pool5 = []
pool6 = []
pool7 = []

print files
templist = {}
readlist = {} # dictionary that ranks which pool that sequence likely came from, updated over time
reads = {} # dictionary that will hold the number of the sequence reads
# Calculating the total reads in the file
totalcount = 0 # this will eventually hold the total number of sequence reads from the mix

duplicates = {} #opens a dictionary of duplicates

# results contains the sequencing reads in the following format:
# > 2016-07-07 15:12:33 processfull237.py PL237-PElute1
# 1 CGTTTGTTGTGCCCTCTCCTCCCTCTCTGCCTGCTCTACCCTCCA 226396
# 2 TTTCGTTTCTCACTTTCTTTTCCATTCTGTTCCTTCTGTTCTCTA 218770
# 3 CAGCGTTTGCATTGCCCTCCCCTGTCCTTCCCCTATCCCCCGTCC 192318
# 4 TAGTTTCGATATCTATCTTGCCTCTCGCATCGTTGTTCACCTCCG 152445
```

```
# 5 TAGTTTCGATACCTATCTTGCCTCTCGCATCGTTGTTCACCTCCG 88141

for row in results[1:]:
	row = row.split()
	sequence = row[1]  # depending on the input file, this may need to be column 2
	seq_count = int(row[2])
	totalcount = totalcount + seq_count
	reads[sequence] = seq_count
	templist[sequence] = 0
	readlist[sequence] = (0,0,0,0,0,0,0,0) # sequence rank in each pool, if found


print totalcount

poolsumlist = []

poolcount = 0
for pool_db in files:
	pool_db = pool_db.split()[0] # removes the /n from the line
	ifile3 = open(pool_db, 'r') # opens the pool hit file
	print pool_db
	pool_db = ifile3.readlines(); ifile3.close() # pool lines defined
	poolhits = []
	poolcount = poolcount + 1
	for line in pool_db[1:]: #CHANGE BACK FOR ALL
		entry = line.split() #should split into 0:rank, 1:interpreted sequence, 2:DNA sequence, 3: reads
		tup = (entry[0], entry[2], entry[3], pool_db) # enter data as rank, sequence, reads, pool
		poolhits.append(tup)

	checksum = 0

	for ele in poolhits:
		sequence = ele[1]
		poolrank = ele[0]
		y = 0
		tuplst = []
		if sequence in templist:
			if templist[sequence] == 0:
				templist[sequence] = poolcount
			else:
				temp = []
				temp.append(str(templist[sequence])); temp.append(str(poolcount))
				templist[sequence] = ','.join(temp[0:])
			tuplst = list(readlist[sequence])
			tuplst[poolcount - 1] = int(poolrank) + 1
			readlist[sequence] = tuple(tuplst)
		checksum = checksum + 1

i = 1
unmatched = 0
unmatchedreads = []
unmatchedsequences = []
poolreadtotal = {1:0, 2:0, 3:0, 4:0, 5:0, 6:0, 7:0, 8:0}

for row in results[1:]:
	row = row.split()
	sequence = row[1] # pulls sequence
	seq_reads = int(row[2])
	ranks = list(readlist[sequence])
	if sum(ranks) > 0:
		toprank = min(i for i in ranks if i > 0)
		if toprank >= 1000 and toprank != sum(ranks): # <-- Threshold for duplicate exclusion here
			assigned_pool = 8
		else:
			assigned_pool = ranks.index(toprank) + 1
		poolreadtotal[assigned_pool] = int(poolreadtotal[assigned_pool]) + seq_reads

#create new dictionary with key = number of reads for the pool, update the key with new sum
	else:
		assigned_pool = 0
```

```
                        unmatched = unmatched + 1
                        unmatchedreads.append(int(row[2]))      # adds to list containing the read number of unmatched sequences
                        unmatchedsequences.append(sequence) # adds unmatched sequences to a list
            matchfile.write("%s %s %s %s %s %s\n" % (i, sequence, assigned_pool, reads[sequence], templist[sequence],
readlist[sequence]))
            i = i+1

print "\n \n"
print poolreadtotal

poolpercentages = []
####  calculate the percentage of each of the read totals:

matchfile.write("###\n \n %s\n" % ("The summed reads are"))


for i in range(1, 8, 1):
            matchfile.write("%s\n" % (poolreadtotal[i]))

matchfile.write("\n %s\n" % ("The pool percentages are"))
for i in range(1, 8, 1):
            poolper = float(poolreadtotal[i])/float(totalcount)
            poolpercentages.append(poolper)
            matchfile.write("%s %s\n" % (i, poolper))
print poolpercentages

matchfile.write("\n %s\n" % ("The duplicate reads and fraction of whole are"))
matchfile.write("%s\n" % (poolreadtotal[8])) # prints duplicate reads
matchfile.write("%s\n\n" % (float(poolreadtotal[8])/float(totalcount))) # prints fraction total

matchfile.write("\n %s %s\n" % ("The percentage of total reads accounted for above is", sum(poolpercentages)))

totalunmatched = sum(unmatchedreads)
print totalunmatched
percentunmatched = 100*float(totalunmatched)/float(totalcount)

matchfile.write("\n \n %s %s %s %s %s %s %s\n \n" % ("There are", unmatched, "unidentified sequences, comprising",
totalunmatched, " reads, or", percentunmatched, "% of total"))
timestr = time.strftime("%Y-%m-%d %H:%M:%S")
matchfile.write("> %s %s" % (timestr, sys.argv[0])) # stamps the date and time on the output count file

unmatchedfile.write("> %s %s: %s %s %s %s %s %s\n" % (timestr, sys.argv[0], "In", len(results)-1, "sequences of ", infilename,
"searching for matches from", files))        # stamps the date and time on the output count file



for j in range(0, len(unmatchedsequences), 1):
            unmatchedfile.write("%s %s %s\n" % (j, unmatchedsequences[j], unmatchedreads[j]))          # stamps the date and time on
the output count file

recordfile.write("%s, %s %s %s %s %s\n" % (timestr, sys.argv[0], sys.argv[1], sys.argv[2], sys.argv[3], sys.argv[4]))
```

Supplementary Note 3.4. Code used to calculate individual sequence enrichments of assigned polymers in head-to-head competition. To assess the enrichment of individual sequences in the head-to-head competition that had been assigned to their pools of origin, the output file generated above was amended with the following code to also output a percentage of the total sample for each sequence.

```
#!/usr/bin/env python
import sys, math, collections, time
from numpy import *
from collections import Counter

try:
   infilename = sys.argv[1]        # This will input the file containing the list of sequences to be analyzed
   outfilenametwo = sys.argv[2]  # This will input a file where all of the values will be stored

except:
   print "Usage:",sys.argv[0], "1_sequence_list 2_record_file"; sys.exit(1)

outputfilename = '_'.join([infilename, 'amend'])

ifile = open(infilename, 'r')  # open sequence list file
outputfile = open(outputfilename, 'w') # open file where amended information will be kept
recordfile = open(outfilenametwo, 'a') # open a file where record will be kept

rank = []
DNAseqlist = []
HFNAPseqlist = []
readcount = []

print ifile.readline() # removes header line and prints it in the real-time display
print ifile.readline() # removes header line and prints it in the real-time display
print ifile.readline() # removes header line and prints it in the real-time display

for line in ifile:    # scans sequence list file line by line
   datalist = line.split() # split up the entries in the file
   if datalist[0] == '###':
            break
   rank.append(datalist[0]) #adds rank list
   DNAseqlist.append(datalist[1]) # adds to growing list of DNA sequences
   HFNAPseqlist.append(datalist[2]) # adds to the growing list of HFNAP sequences
   readcount.append(int(datalist[3])) #adds reads

readtotal= sum(readcount)

timestr = time.strftime("%Y-%m-%d %H:%M:%S")
print "> %s %s %s" % (timestr, sys.argv[0], infilename) #Prints out the date and time for logging of multiple runs
print infilename

print "There are", readtotal, "total reads"
print "and", len(rank), "unique DNA sequences"

recordfile.write("%s, %s\n" % (timestr, infilename))

outputfile.write("> %s %s %s\n" % (timestr, sys.argv[0], infilename))      # stamps the date and time on the output count file
outputfile.write("> %s\n" % (readtotal))    # stamps total number of sequences
for i in range(0, len(rank), 1):
   outputfile.write("%s %s %s %s %s\n" % (rank[i], DNAseqlist[i], HFNAPseqlist[i], readcount[i],
100*float(readcount[i])/float(readtotal)))
```

The amended files containing the list of assigned sequences from samples before and after the selection were then compared against each other. Sequences that were found in both samples were compiled into a list ordered by their relative enrichment in the head-to-head competition.

```
#!/usr/bin/env python
import sys, math, collections, time
```

```
from numpy import *
from collections import Counter

# The goal of this program is to take in the combined list of sequences and to track how well each of those sequences
# perform in the head-to-head competition experiment - basically want to compare enrichment of sequences in selection round

try:
    infilename = sys.argv[1]        # This will input the file containing the big list of sequences from each pool
    infilenametwo = sys.argv[2]     # This inputs the file where the list of sequences from individual round/pool will be pulled
    outfilename = sys.argv[3]       # This inputs the file where the matches will go
    outfilelast = sys.argv[4]    # This is where the record will be created

except:
    print "Usage:",sys.argv[0], "1_combined_sequence_list 2_file_containing_list_of_all_seqinterp_files_from_pool
3_Output_of_matches 4_record_file"; sys.exit(1)

ifile = open(infilename, 'r')  # open sequence list file from combined pools
ifile2 = open(infilenametwo, 'r') #opens the list of sequence files to look at
recordfile = open(outfilelast, 'a') # open a file where info will be written for a number of sequences

fulllist = ifile.readlines(); ifile.close() #opens combined list of all sequences

subfiles = ifile2.readlines(); ifile2.close() #reads the lines of the file containing all of the sequence files from the rounds

timestr = time.strftime("%Y-%m-%d %H:%M:%S")

# Format of the data in fulllist is index, DNA sequence, DNA sequence, total counts

percentpoollist = {}
readroundlist = {}
lastranklist = {}
pooloforigin = {}

for line in fulllist[2:]: # This will be the loop that extracts all of the sequneces that are to be examined
    sequencedata = line.split()
    sequence = sequencedata[1] #extracts sequence from the big list of sequences
    percentpoollist[sequence] = (0,0) # defines tuple that will include an entry for each round for each sequence's percentages
    readroundlist[sequence] = (0,0) # defines tuple that will include an entry for each round for each sequence's total reads
    lastranklist[sequence] = 'X'
    pooloforigin[sequence] = sequencedata[2]

roundcount = 0 # defines a value for indexing the round count
for subfil in subfiles:
    subfile = subfil.split()[0] # removes the /n from the line
    ifile3 = open(subfile, 'r') # opens the data file from the round
    print "subfil is", subfil # should print the round number?
    print "subfile is", subfile
    lines = ifile3.readlines(); ifile3.close() # reads the sequence data from each of the rounds

    roundcount = roundcount + 1
    print "roundcount is", roundcount

    poolhits = [] #creates list where each sequence will go
    readcount = lines[1].split()[1] #[1]picks out total number of reads in the second line of document
    print "readcount is", readcount
#    print lines
    for line in lines[2:]:  # Loop parsing data from the amended _seqinterp files
        entry = line.split()  # should split into 0:rank, 1:DNA sequence, 2:pool origin, 3: reads, 4: percentage of total
        #       print entry[2]
        if entry[1] in percentpoollist.keys():
                    tup = (entry[0], entry[1], entry[2], entry[3], entry [4], subfile)  # enter data as rank, sequence, pool origin, reads, pool
                    poolhits.append(tup)

        checksum = 0

    for i in range(0, len(poolhits), 1):
        sequence = poolhits[i][1]
        seqorigin = poolhits[i][2]
        seqreads = poolhits[i][3]
        seqpercent = poolhits[i][4]
```

```
        if roundcount == 2:
                lastrank = poolhits[i][0]
                lastranklist[sequence] = lastrank

        #for each sequence in the individual round, want to create tuple that has the reads for each round
        #also want to create a tuple that has the percentage for each read

        # want to output the tuple to a list and then modify the particular entry
        temppercentlist = list(percentpoollist[sequence])
        tempreadslist = list(readroundlist[sequence])
        temppercentlist[roundcount-1] = seqpercent
        tempreadslist[roundcount-1] = seqreads

        percentpoollist[sequence] = tuple(temppercentlist)
        readroundlist[sequence] = tuple(tempreadslist)
        pooloforigin[sequence] = seqorigin


countfilename = '_'.join([outfilename, 'count'])
countfile = open(countfilename, 'w')
countfile.write("> %s %s\n" % (timestr, sys.argv[0])) #Header to output file
counter = 0

mobilityfilename = '_'.join([outfilename, 'mobility'])
mobilityfile = open(mobilityfilename, 'w')
mobilityfile.write("> %s %s\n" % (timestr, sys.argv[0])) #Header to output file

percentagefilename = '_'.join([outfilename, 'percent'])
percentfile = open(percentagefilename, 'w')
percentfile.write("> %s %s\n" % (timestr, sys.argv[0])) #Header to output file
counter = 0

newdictionary1 = {}
newdictionary2 = {}

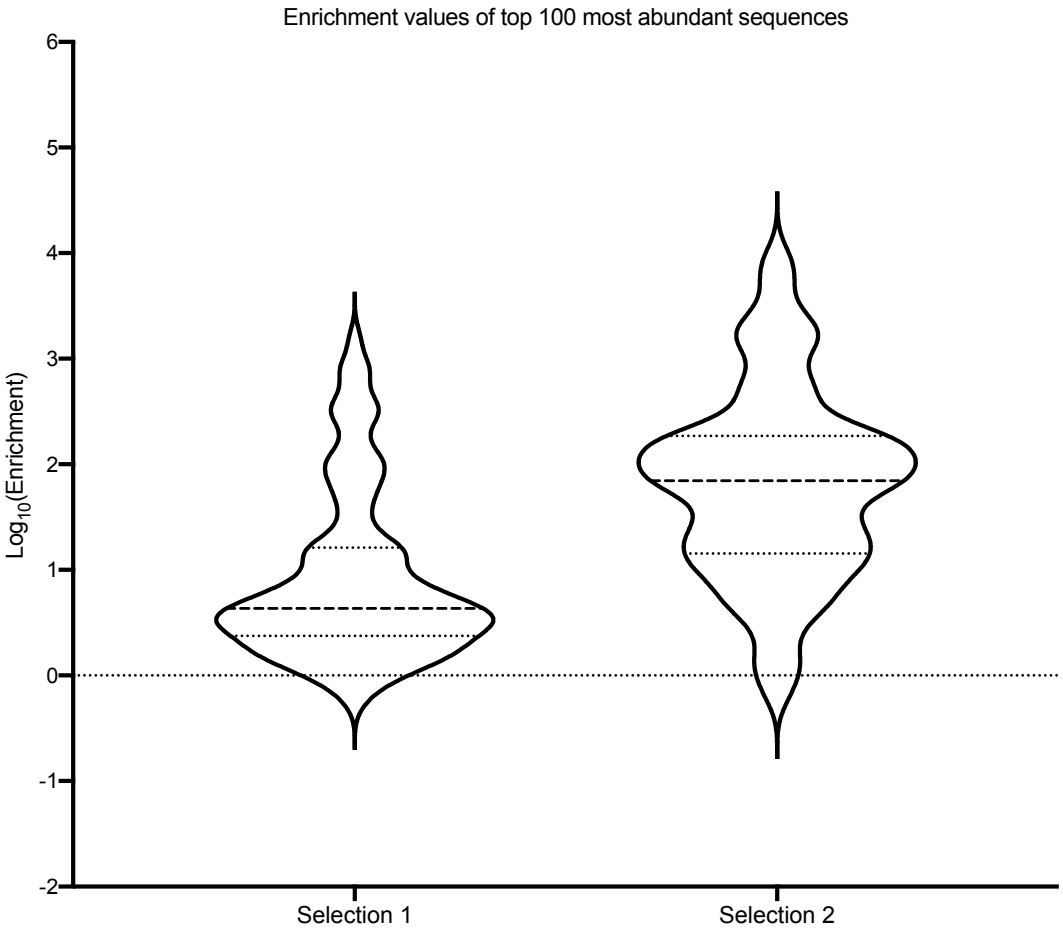
for w, z in percentpoollist.items():
        counter += 1
        percentfile.write("%s %s %s\n" % (counter, w, z))   #
        templist = []
        for i in range(1, len(z), 1):
                if z[i-1] > 0:
                        templist.append(round(float(z[i])/float(z[i-1]),2))
                elif z[i-1] and z[i] == 0:
                        templist.append(0)
                else:
                        templist.append(-0.01)
        total = sum(templist)
        newdictionary1[w] = max(templist) # adds total
        newdictionary2[w] = templist # adds templist

counter = 0
for k, v in sorted(newdictionary1.items(), key=lambda p:p[1], reverse = True):
        counter +=1
        if v > 1:
                mobilityfile.write("%s %s %s %s %s %s\n" % (counter, pooloforigin[k], lastranklist[k], k, v, newdictionary2[k])) #
prints index, sequence, total, tuple

recordfile.write("%s, %s %s %s %s %s\n" % (timestr, sys.argv[0], sys.argv[1], sys.argv[2], sys.argv[3], sys.argv[4]))
```

Codon	Sidechain	K_d	K_d change
CGA	Fluorophenyl	231.1 ± 23.6 nM	18x
TTC	Allylamine	16.2 ± 3.2 nM	1.25x
TGG	Allylamine	16.0 ± 2.2 nM	1.25x
TGC	Alcohol (2x)	20.3 ± 4.4 nM	1.6x
TGT	Phenol	6.7 ± 1.0 nM	0.52x
TGA	Alcohol	46.8 ± 4.8 nM	3.6x
TCT	Allylamine	12.9 ± 1.5 nM	1x
CCG	Cyclopentyl	21.9 ± 2.6 nM	1.7x
CCA	Cyclopropyl	158.8 ± 14.2 nM	12x
TTG	Phenol	32.6 ± 5.0 nM	2.5x
TCC	Imidazole	29.9 ± 10.2 nM	2.3x
TAC	Imidazole	21.7 ± 2.1 nM	1.7x
CCT	Cyclopropyl	18.8 ± 2.6 nM	1.5x
TCA	Imidazole	10.1 ± 1.6 nM	0.8x

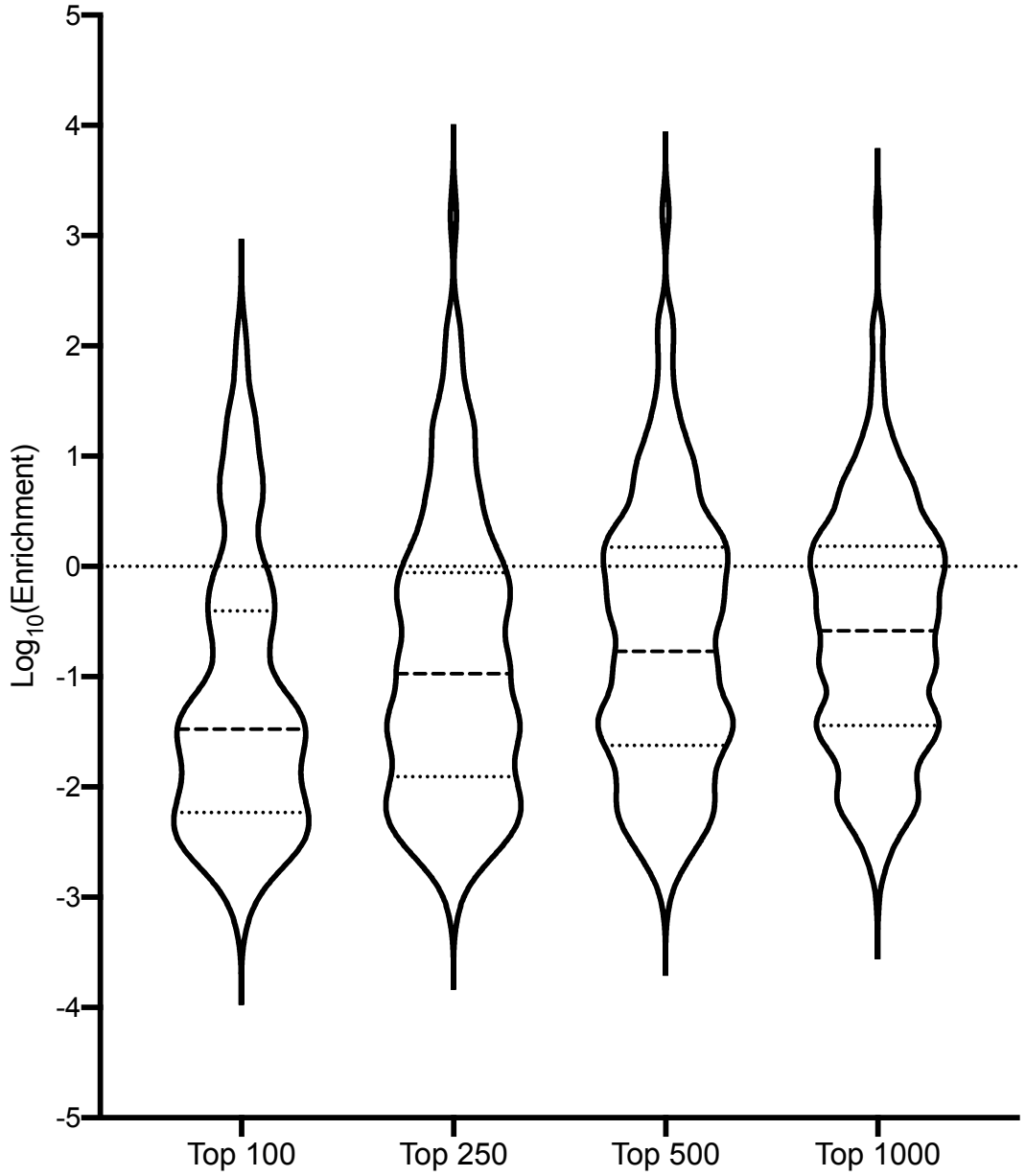
Supplementary Table 4.1 | MST characterization of daunomycin binding affinity for mutants of Dm-HS-1 lacking side-chains. Fourteen single mutants of Dm-HS-1 were translated without side-chains for the indicated codon position. MST characterization of the binding affinities of the resulting mutants are shown. The K_d change is relative to unmutated Dm-HS-1. The mean and SEM of n = 3 independent replicates are shown.



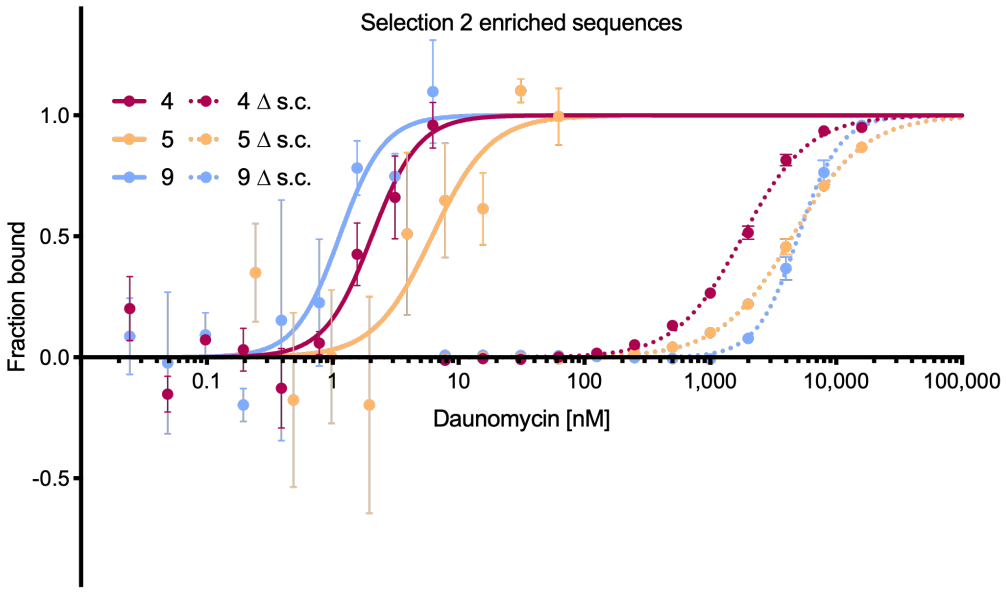
Median enrichment values			
Top	Selection 1	Selection 2	Fold-change
10	14.9	1635	109.5
20	10.1	142	14.0
25	8.0	141	17.5
30	7.2	119	16.6
40	5.4	90	16.8
50	5.4	81	15.2
100	4.3	70	16.2

Average enrichment values			
Top	Selection 1	Selection 2	Fold-change
10	36.3	2581	71.1
20	36.4	1719	47.2
25	30.7	1466	47.8
30	26.3	1293	49.3
40	20.5	1042	50.7
50	64.2	848	13.2
100	69.3	481	6.9

Supplementary Figure 4.1 | Differences in enrichment values between selection 1 and selection 2. Violin plot center line designates the median; dotted line used for upper and lower quartiles. Selection-1 enrichment values were calculated by dividing each HFNAP sequence's frequency in the last round of selection 1 (9a) by its frequency in round 4. Selection-2 enrichment values were calculated by dividing each HFNAP sequence's frequency in the last round of selection 2 (8b) by its frequency in round 4. If sequences were not found in round 4, the subsequent round frequency was used. A $\log_{10}$(enrichment) value of 0 denotes no change. Positive values indicate enrichment and negative values indicate de-enrichment.

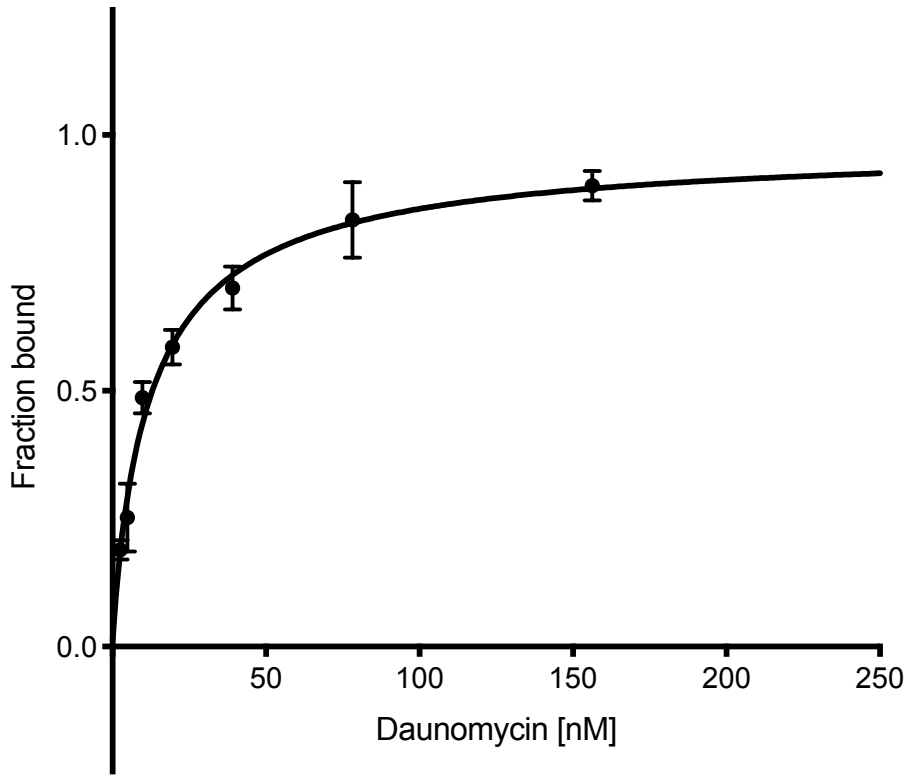


Supplementary Figure 4.2 | Performance of selection-1 sequences in selection 2. Violin plot center line designates the median; dotted line used for upper and lower quartiles. The selection-2 enrichment values of the top 1000 sequences from selection 1 (by relative abundance) were calculated. The 1000 selection-1 sequences were identified in selection 2, and their sequence frequencies at the end of selection 2 (round 8b) were divided by the sequence frequencies at the start of selection 2 (round 4). If sequences were not found in round 4, the subsequent round frequency was used. If sequences were not found in round 8, the previous round frequency was used. A $\log_{10}$(enrichment) value of 0 denotes no change. Positive values indicate enrichment and negative values indicate de-enrichment.



Sequence	K_d	K_d Δ side-chains	K_d change
Dm-HS-1	12.85 ± 1.6 nM	2.61 ± 0.19 μM	200x
Dm-HS-2	4.77 ± 1.0 nM	1.70 ± 0.17 μM	350x
Dm-HS-3	4.78 ± 0.95 nM	4.45 ± 0.28 μM	930x
Dm-HS-4	2.05 ± 0.65 nM	1.81 ± 0.057 μM	880x
Dm-HS-5	6.16 ± 3.4 nM	4.50 ± 0.21 μM	730x
Dm-HS-6	26.86 ± 4.2 nM	1.51 ± 0.065 μM	56x
Dm-HS-7	18.09 ± 5.2 nM	2.49 ± 0.12 μM	140x
Dm-HS-8	32.26 ± 8.9 nM	4.84 ± 0.26 μM	150x
Dm-HS-9	1.13 ± 0.45 nM	5.00 ± 0.14 μM	4000x
Dm-HS-10	64.68 ± 3.9 nM	0.713 ± 0.028 μM	11x

Supplementary Figure 4.3 | MST characterization of daunomycin affinity for the top 10 most enriched sequences from selection 2. Dm-HS-4 and Dm-HS-5 differ from Dm-HS-1 by one nucleotide, and Dm-HS-9 differs from Dm-HS-3 by one nucleotide. The mean and SEM of n = 3 independent replicates are shown. Δ s.c. denotes sequences without side-chains. Sequences were found to bind with K_d = 5-65 nM. The mean and SEM of n = 3 independent replicates are shown.



Supplementary Figure 4.4 | Gel filtration characterization of Dm-HS-1 daunomycin affinity. ^{3}H-daunomycin was used to characterize Dm-HS-1 daunomycin affinity in a gel filtration assay. The mean and SEM of n = 3 independent replicates are shown. K_d = 12.6 ± 3.5 nM which reflects mean and SEM of n = 3 independent replicates.

Supplementary Note 4.1 – Oligonucleotides used in selections, HFNAP preparation, and HTS.

Oligonucleotides for daunomcyin affinity selection

Name	Sequence
Biotinylated DX template library	/52-Bio/iSp18/AAGCTCTAGTTCGGTCTG NNY NNY NNY NNY NNY NNY NNY NNY NNY NNY NNY NNY NNY NNY ATCTCCAGTCCTGCTTTC
BtBt ExtPrimerD	/52-Bio/iSp18/AAGCTCTAGTTCGGTCTG
ExtPrimer D	AAGCTCTAGTTCGGTCTG
pp1D	/5Phos/CAGACCGAACTAGAGCTT
dc pp1D	/5Phos/CAGACCGAACTAGAGCTT/3ddC/
pp2X	GAAAGCAGGACTGGAGAT
Seq Adapt D1	ACACTCTTTCCCTACACGACGCTCTTCCGATCT N AAGCTCTAGTTCGGTCTG
Seq Adapt D2	ACACTCTTTCCCTACACGACGCTCTTCCGATCT NN AAGCTCTAGTTCGGTCTG
Seq Adapt D3	ACACTCTTTCCCTACACGACGCTCTTCCGATCT NNN AAGCTCTAGTTCGGTCTG
Seq Adapt D4	ACACTCTTTCCCTACACGACGCTCTTCCGATCT NNNN AAGCTCTAGTTCGGTCTG
Seq Adapt X1	TGGAGTTCAGACGTGTGCTCTTCCGATCT N GAAAGCAGGACTGGAGAT
Seq Adapt X2	TGGAGTTCAGACGTGTGCTCTTCCGATCT NN GAAAGCAGGACTGGAGAT
Seq Adapt X3	TGGAGTTCAGACGTGTGCTCTTCCGATCT NNN GAAAGCAGGACTGGAGAT
Seq Adapt X4	TGGAGTTCAGACGTGTGCTCTTCCGATCT NNNN GAAAGCAGGACTGGAGAT
IlluminaAdapterFwd	AATGATACGGCGACCACCGAGATCTACAC [8-base barcode] ACACTCTTTCCCTACACGAC
IlluminaAdapterRev	CAAGCAGAAGACGGCATACGAGAT [8-base barcode] GTGACTGGAGTTCAGACGTGTGCT

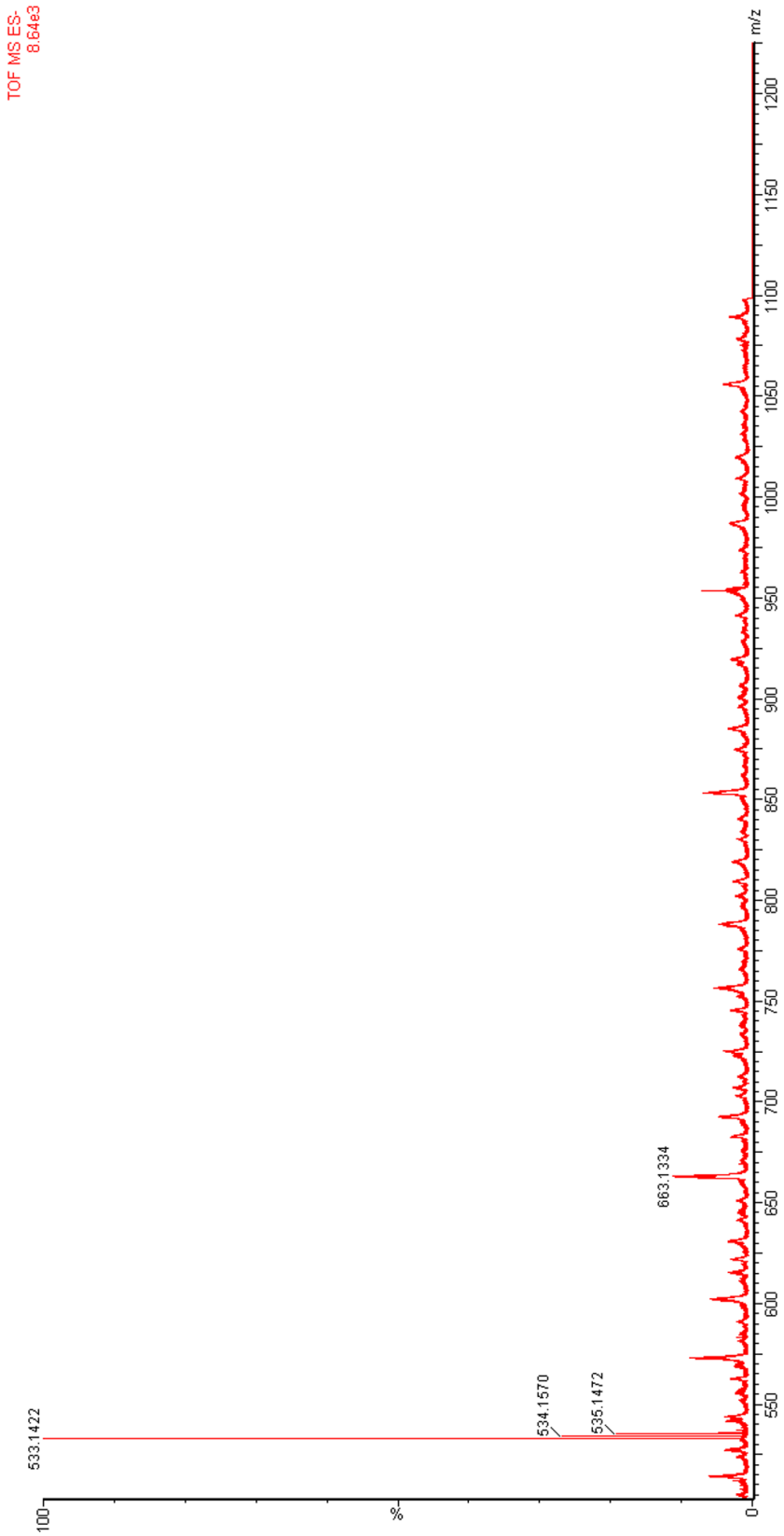
Oligonucleotides for synthesis of individual hits for characterization

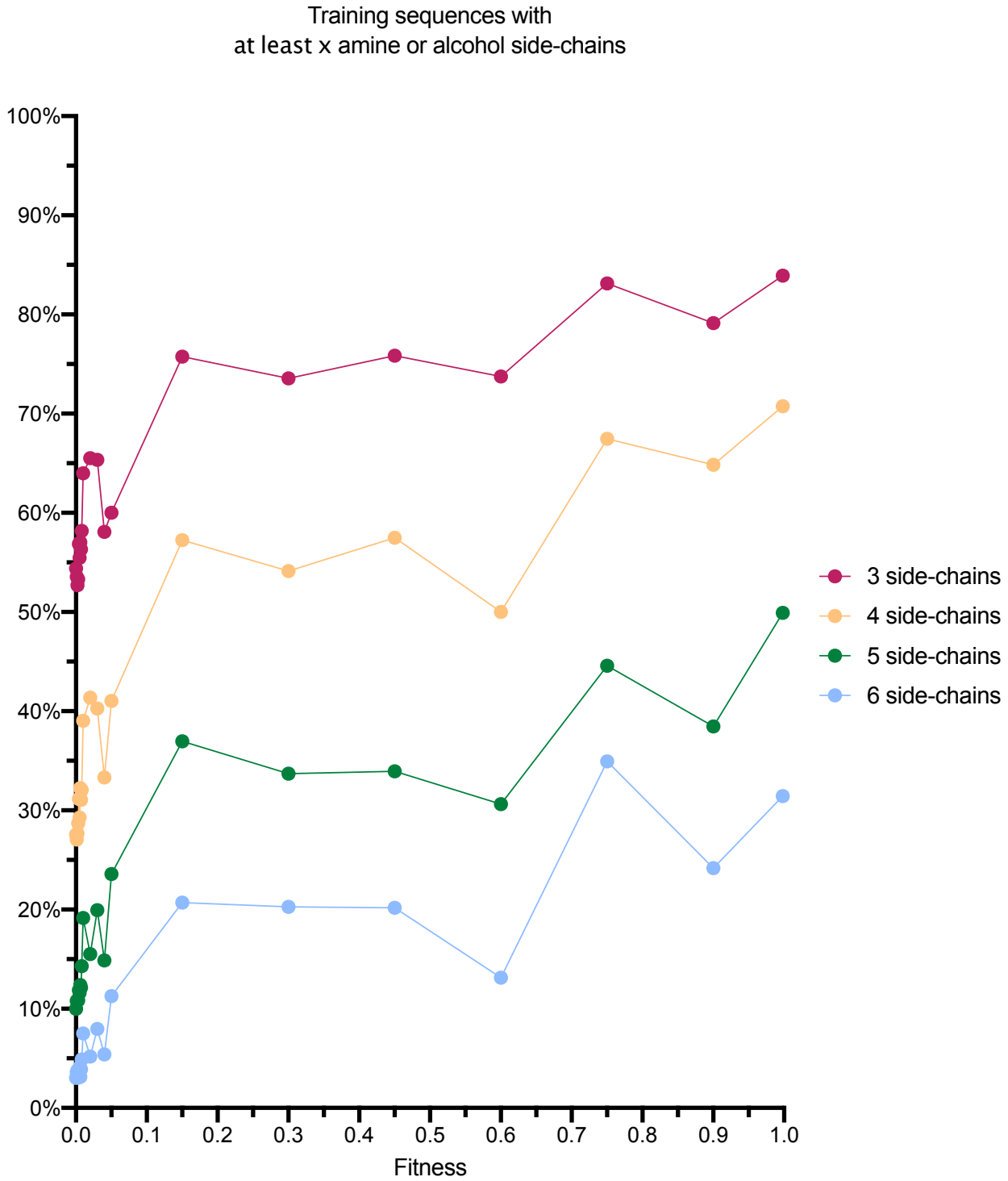
Name	Sequence
pp1D	/5Phos/CAGACCGAACTAGAGCTT
dc pp1D	/5Phos/CAGACCGAACTAGAGCTT/3ddC/

pp2X	GAAAGCAGGACTGGAGAT
pp2x Cy5	/5Cy5//iSp18/GA AAG CAG GAC TGG AGA T
Dm-LS-1 template	/52-Bio/iSp18/AAGCTCTAGTTCGGTCTG TAAGGGGAAGTAGGGTAAAAGGGACTAGGGATAAAAGCGGCGCGG ATCTCCAGTCCTGCTTTC
Dm-LS-2 template	/52-Bio/iSp18/AAGCTCTAGTTCGGTCTG CAAGGAGGGATACGGGTAGCGGGGACAGGGATAAAAGCGGCGCGG ATCTCCAGTCCTGCTTTC
Dm-LS-3 template	/52-Bio/iSp18/AAGCTCTAGTTCGGTCTG AGGTAGCAGTGGGCGGTGGGACGGCAAGGAGGAGTAGCGATGCGA ATCTCCAGTCCTGCTTTC
Dm-LS-4 template	/52-Bio/iSp18/AAGCTCTAGTTCGGTCTG TAAGGGGAAGTAGGGTAAGAGGGACTAGGGATAAAAGCGGCGCGG ATCTCCAGTCCTGCTTTC
Dm-HS-1 template	/52-Bio//iSp18/AAGCTCTAGTTCGGTCTG TGAAGGGTAGGAGCACAATGGCGGAGATCAACAGCACCAGAATCG ATCTCCAGTCCTGCTTTC
Dm-HS-2 template	/52-Bio//iSp18/AAGCTCTAGTTCGGTCTG TTGAGAGAGACAGGGGCAGGACAAGGGTCAACGATGCGACCATAG ATCTCCAGTCCTGCTTTC
Dm-HS-3 template	/52-Bio//iSp18/AAGCTCTAGTTCGGTCTG CGAGTAGGGAAGGGGATATGGAGGAGATCAAAATCACAATAATCG ATCTCCAGTCCTGCTTTC
Dm-HS-4 template	/52-Bio//iSp18/AAGCTCTAGTTCGGTCTG CGAAGGGTAGGAGCACAATGGCGGAGATCAACAGCACCAGAATCG ATCTCCAGTCCTGCTTTC
Dm-HS-5 template	/52-Bio//iSp18/AAGCTCTAGTTCGGTCTG TGAAGGGCAGGAGCACAATGGCGGAGATCAACAGCACCAGAATCG ATCTCCAGTCCTGCTTTC
Dm-HS-6 template	/52-Bio//iSp18/AAGCTCTAGTTCGGTCTG AGGTCAGCGTGAACGGCATGATTAGCAGGACAGCAGTTACTACCA ATCTCCAGTCCTGCTTTC
Dm-HS-7 template	/52-Bio//iSp18/AAGCTCTAGTTCGGTCTG AGGTGACGGAGGTGATCAAGGGGACTAGCAACGTAACTGCAATCG ATCTCCAGTCCTGCTTTC
Dm-HS-8 template	/52-Bio//iSp18/AAGCTCTAGTTCGGTCTG TTATGAATGGCAATAAGGTGGGCAGGGGAGACGTAACTGCAATCG ATCTCCAGTCCTGCTTTC

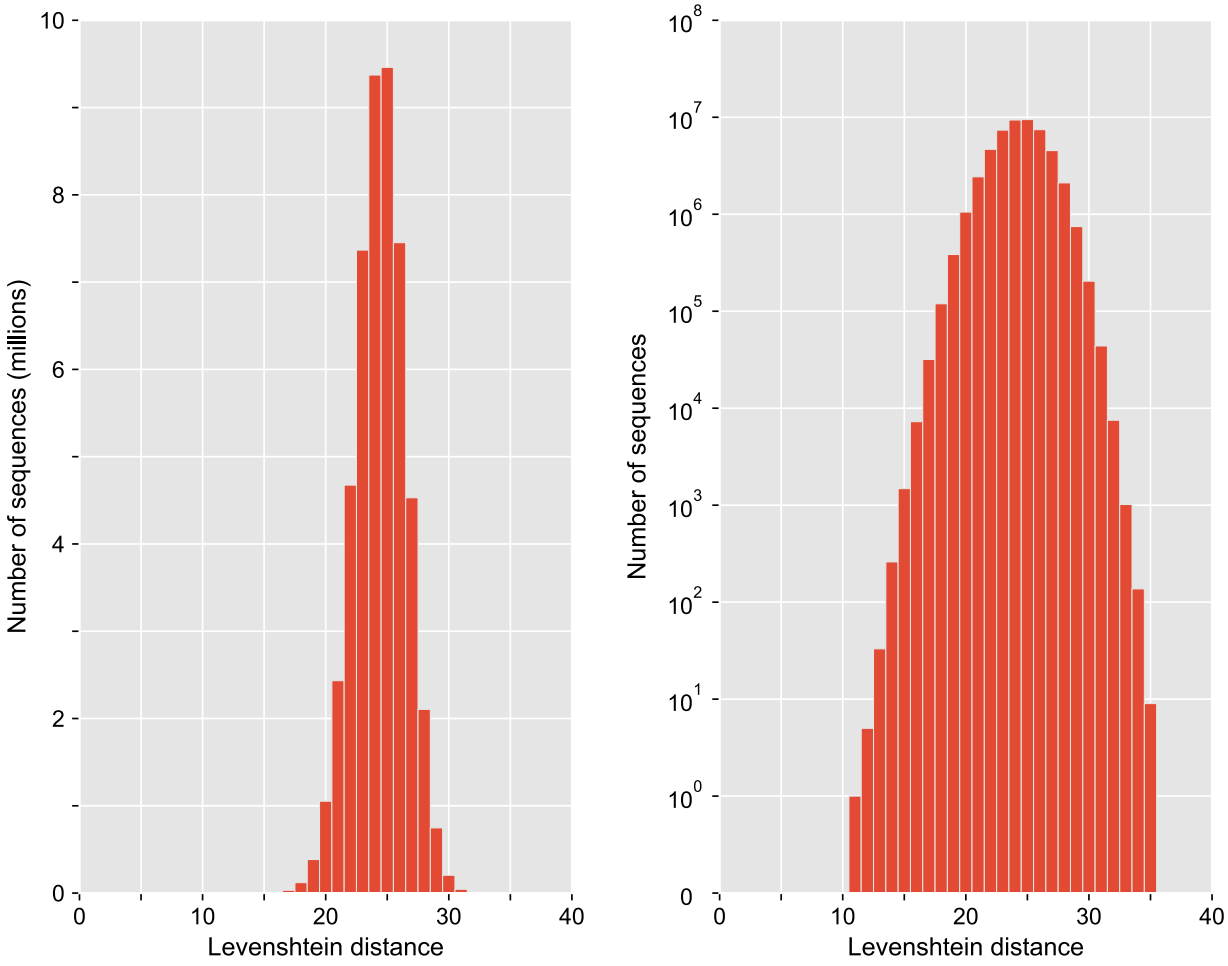
Dm-HS-9 template	/52-Bio//iSp18/AAGCTCTAGTTCGGTCTG CGAGTAGGGAAGGGGATATGGGGGAGATCAAAATCACAATAATCG ATCTCCAGTCCTGCTTTC
Dm-HS-10 template	/52-Bio//iSp18/AAGCTCTAGTTCGGTCTG CGGTCAAGGGGAACAGGGGTGACAGGGAGAAAGCCGATGCGAAGG ATCTCCAGTCCTGCTTTC

Supplementary Note 4.2 – Mass spectrum of negative selection target

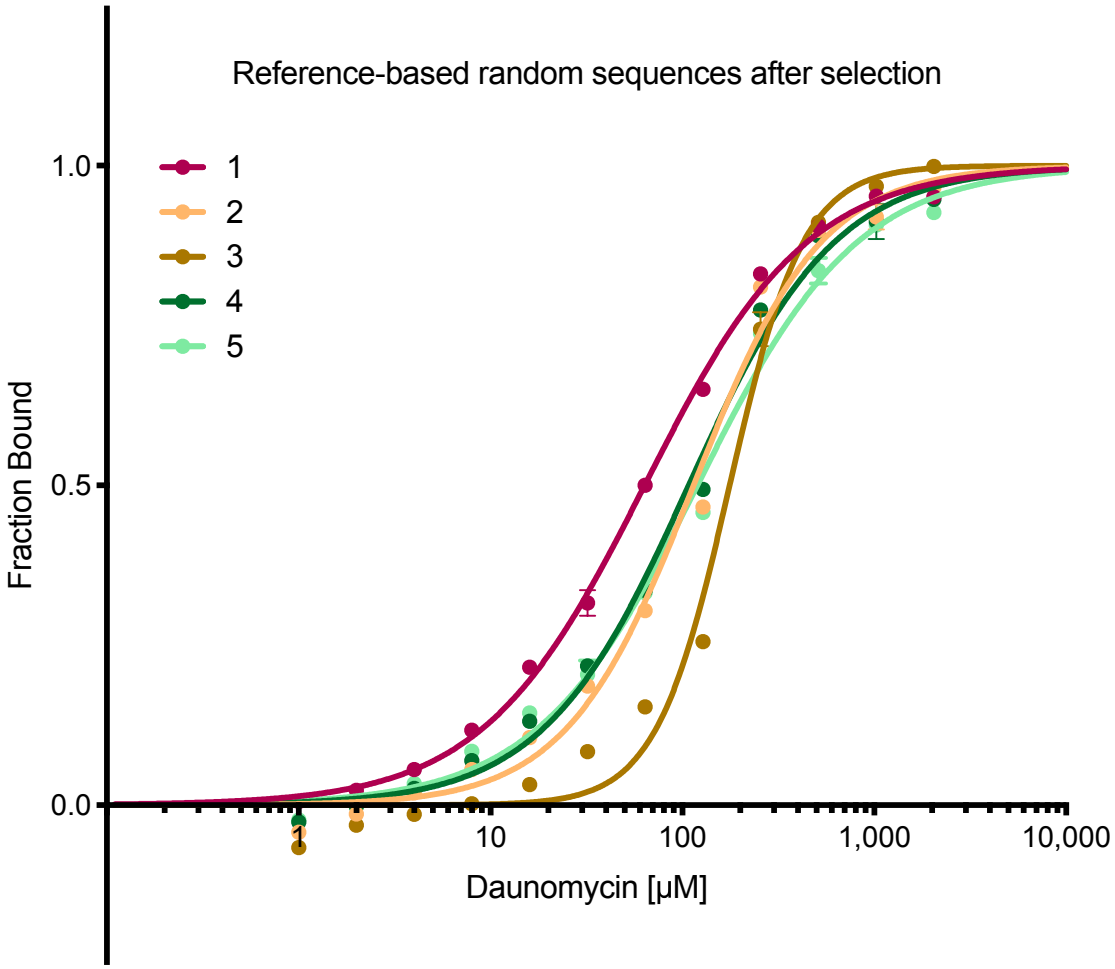




Supplementary Figure 5.1 | Percent of training sequences with at least 3-6 alcohol or amine side-chains. Selection sequences were grouped based on their fitness values. Each point along the X-axis represents the sequences with fitness greater than the x value and less than or equal to the next x value. For example, the 0.6 data point refer to all sequences with fitness >0.6 and ≤0.75.

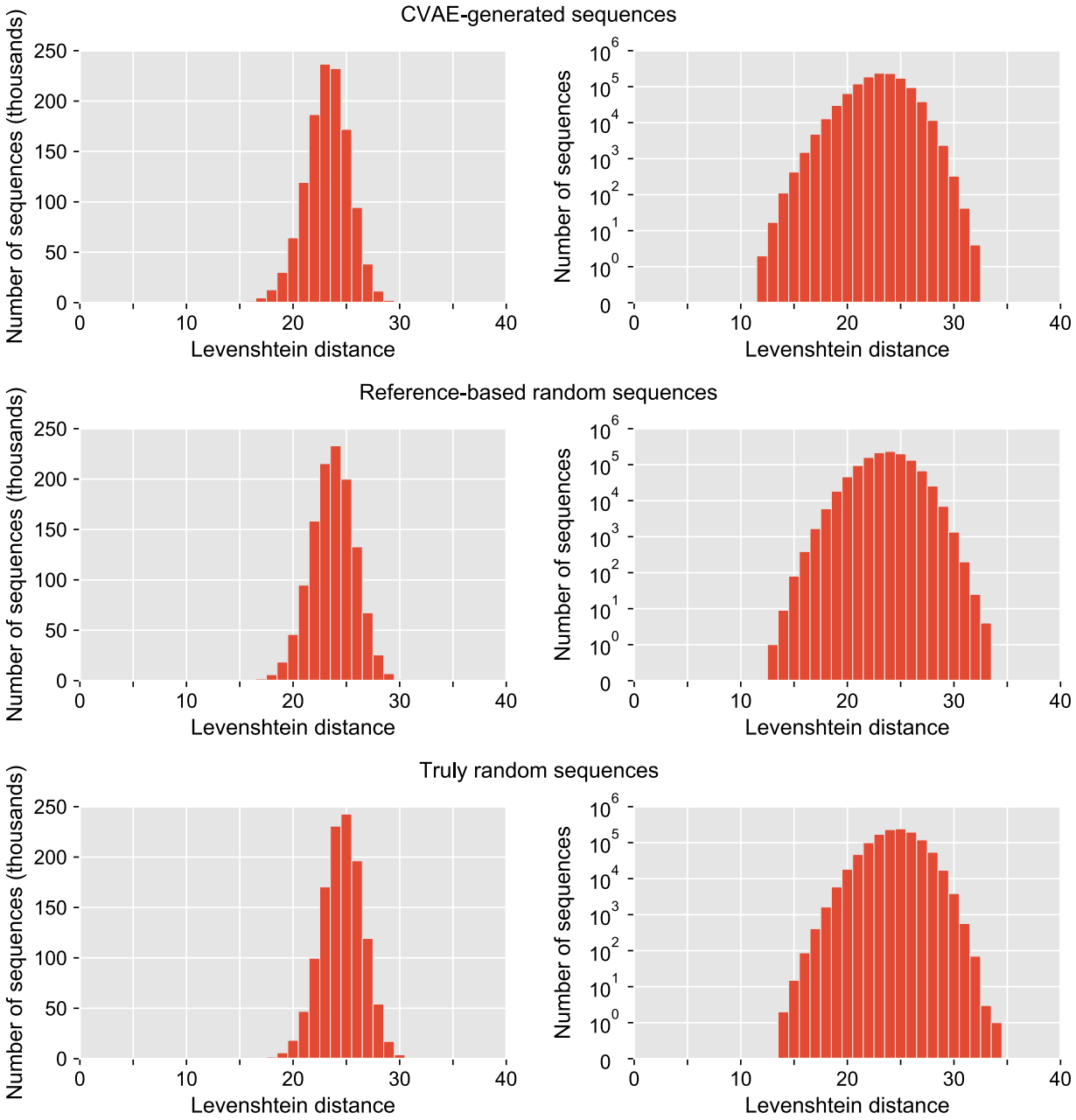


Supplementary Figure 5.2 | Pairwise Levenshtein distances for 10,000 truly random HFNAPs. Pairwise distance distributions for randomly sampled sequences using a 1/32 frequency for each codon are shown.

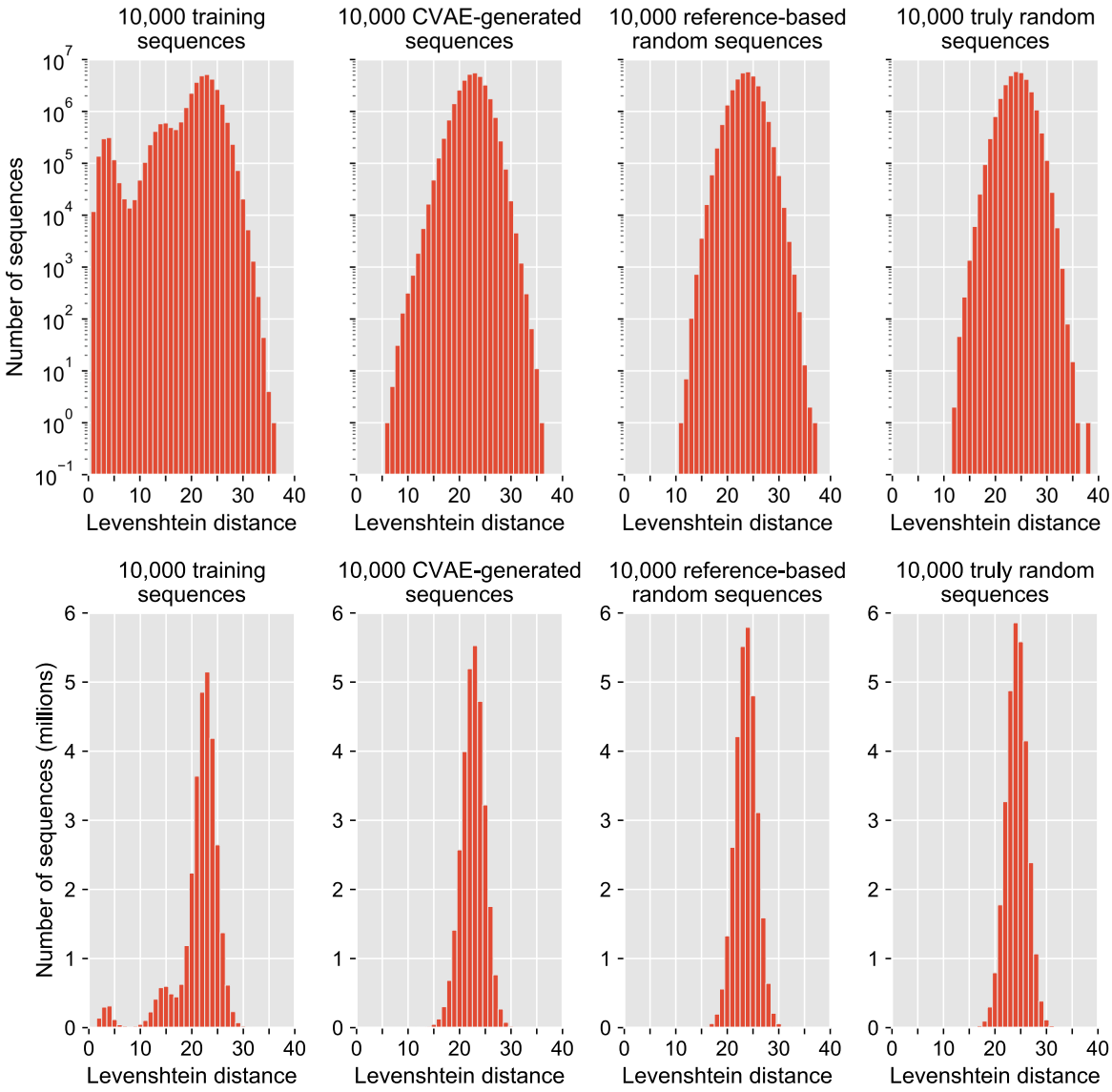


Sequence	K_d
Dm-reference-based-random-after-selection-1	63.6 ± 2.4 μM
Dm-reference-based-random-after-selection-2	114.4 ± 7.4 μM
Dm-reference-based-random-after-selection-3	176.5 ± 8.0 μM
Dm-reference-based-random-after-selection-4	107.9 ± 6.9 μM
Dm-reference-based-random-after-selection-5	119.7 ± 9.3 μM

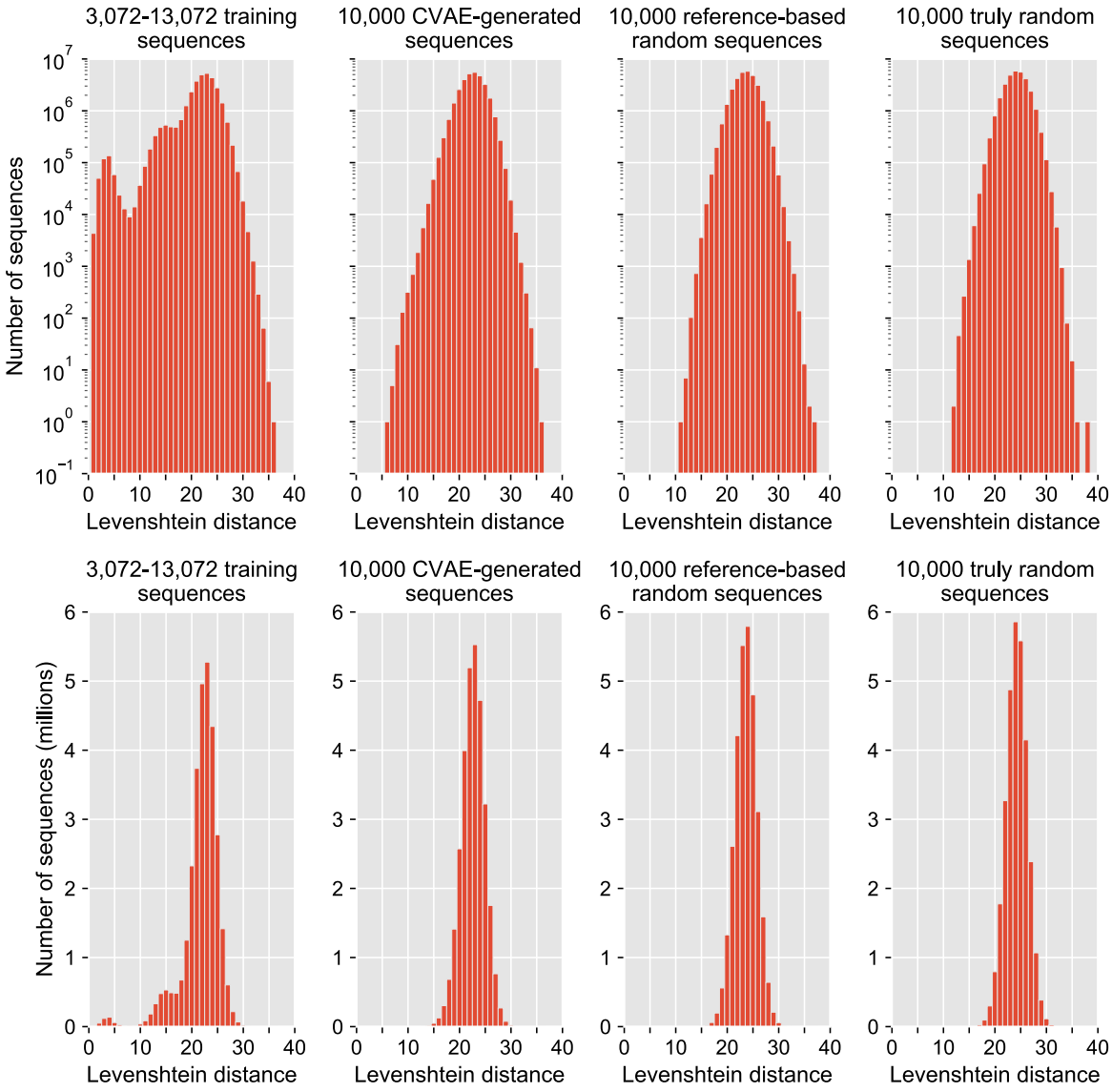
Supplementary Figure 5.3 | MST characterization of the five most enriched sequences from the reference-based random set after selection. 10,000 reference-based random sequences and the top 2,000 sequences by fitness from selection 2 (competition set) were subject to HFNAP translation and one round of high-stringency selection. The top five most enriched sequences from the reference-based random set were translated and characterized by MST. Sequences bound with K_d >10 μM. The mean and SEM of n = 3 independent replicates are shown. Error bars for some values are too small to extend beyond the data point.



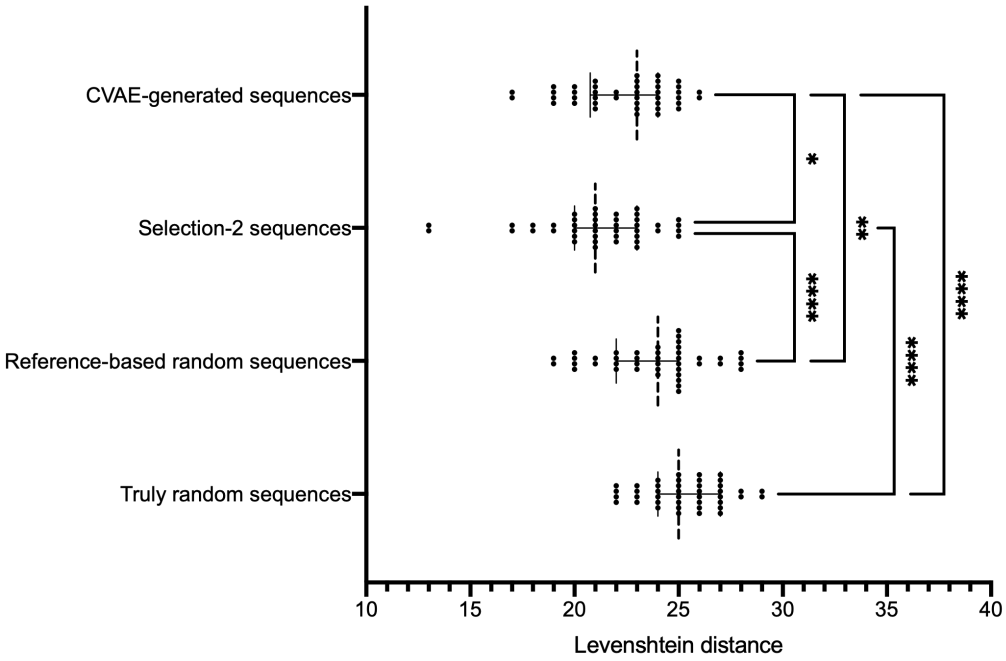
Supplementary Figure 5.4 | Pairwise Levenshtein distances between seven sequences and 172,545-member training set. Pairwise distances for the seven CVAE-generated daunomycin-binding HFNAPs, seven reference-based random sequences, and seven truly random sequences were calculated with respect to the training set, and are shown in aggregate.



Supplementary Figure 5.5 | Pairwise Levenshtein distances from reference set to 10,000-member libraries. Pairwise Levenshtein distances are calculated and plotted in aggregate.

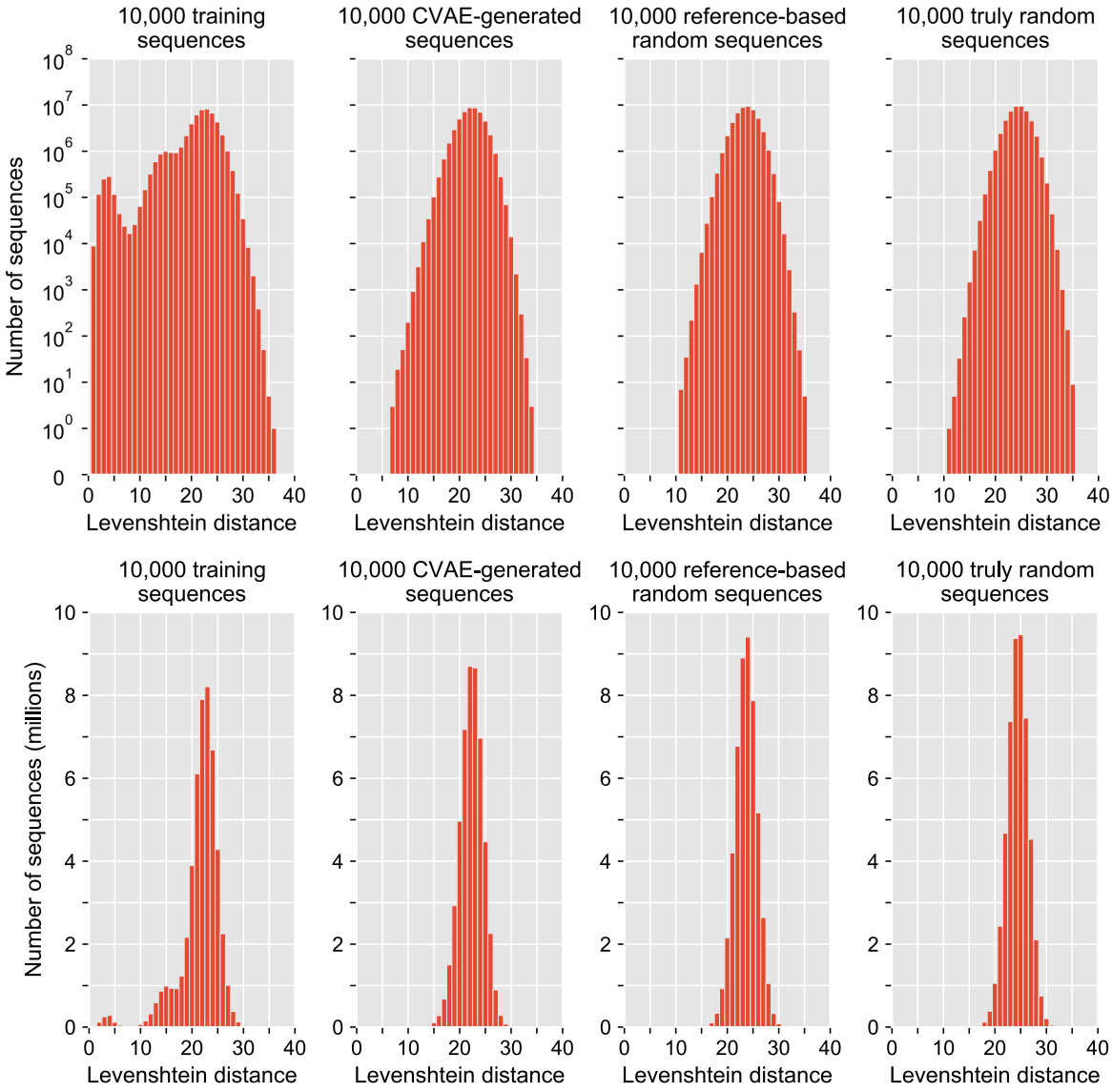


Supplementary Figure 5.6 | Pairwise Levenshtein distances from reference set to 10,000-member libraries. The reference set consists of the top 3,072 HFNAP sequences with the highest fitness in selection 2. Top 3,072 to 13,072 training set sequences by fitness are used to remove direct overlaps present in the reference set to the top training set sequences. Pairwise Levenshtein distances are calculated and plotted in aggregate.



A	B	U	p value
CVAE-generated sequences	Selection-2 sequences	652	0.0378
CVAE-generated sequences	Reference-based random sequences	584	0.0067
CVAE-generated sequences	Truly random sequences	300	3.1E-08
Selection-2 sequences	Reference-based random sequences	404	8.9E-06
Selection-2 sequences	Truly random sequences	154	4.5E-13

Supplementary Figure 5.7 | Distance distributions for pairwise Levenshtein distances for seven validated CVAE-generated HFNAP sequences, seven validated selection-2 HFNAP sequences, seven reference-based random HFNAP sequences, and seven truly random HFNAP sequences. Median and quartile distances are shown. Distance distributions are compared using the Mann-Whitney U test with two-tailed p values.



Supplementary Figure 5.8 | Pairwise Levenshtein distances amongst 10,000-member libraries. Pairwise Levenshtein distances are calculated between every member of the given library and shown in aggregate.

Supplementary Note 5.1 – Oligonucleotides used in selections, HFNAP preparation, and HTS.

Oligonucleotides for daunomycin affinity selection

Name	Sequence
Biotinylated DX template library	/52-Bio/iSp18/AAGCTCTAGTTCGGTCTG NNY NNY NNY NNY NNY NNY NNY NNY NNY NNY NNY NNY NNY NNY ATCTCCAGTCCTGCTTTC
BtBt ExtPrimerD	/52-Bio/iSp18/AAGCTCTAGTTCGGTCTG
ExtPrimer D	AAGCTCTAGTTCGGTCTG
pp1D	/5Phos/CAGACCGAACTAGAGCTT
dc pp1D	/5Phos/CAGACCGAACTAGAGCTT/3ddC/
pp2X	GAAAGCAGGACTGGAGAT
Seq Adapt D1	ACACTCTTTCCCTACACGACGCTCTTCCGATCT N AAGCTCTAGTTCGGTCTG
Seq Adapt D2	ACACTCTTTCCCTACACGACGCTCTTCCGATCT NN AAGCTCTAGTTCGGTCTG
Seq Adapt D3	ACACTCTTTCCCTACACGACGCTCTTCCGATCT NNN AAGCTCTAGTTCGGTCTG
Seq Adapt D4	ACACTCTTTCCCTACACGACGCTCTTCCGATCT NNNN AAGCTCTAGTTCGGTCTG
Seq Adapt X1	TGGAGTTCAGACGTGTGCTCTTCCGATCT N GAAAGCAGGACTGGAGAT
Seq Adapt X2	TGGAGTTCAGACGTGTGCTCTTCCGATCT NN GAAAGCAGGACTGGAGAT
Seq Adapt X3	TGGAGTTCAGACGTGTGCTCTTCCGATCT NNN GAAAGCAGGACTGGAGAT
Seq Adapt X4	TGGAGTTCAGACGTGTGCTCTTCCGATCT NNNN GAAAGCAGGACTGGAGAT
IlluminaAdapterFwd	AATGATACGGCGACCACCGAGATCTACAC [8-base barcode] ACACTCTTTCCCTACACGAC
IlluminaAdapterRev	CAAGCAGAAGACGGCATACGAGAT [8-base barcode] GTGACTGGAGTTCAGACGTGTGCT

Oligonucleotides for synthesis of individual hits for characterization

Name	Sequence
pp1D	/5Phos/CAGACCGAACTAGAGCTT
dc pp1D	/5Phos/CAGACCGAACTAGAGCTT/3ddC/
pp2X	GAAAGCAGGACTGGAGAT
pp2x Cy5	/5Cy5//iSp18/GA AAG CAG GAC TGG AGA T
Dm-CVAE-without-selection-1 template	/52-Bio//iSp18/AAGCTCTAGTTCGGTCTG GTAGGAGGAAAGGCAATATGGAAAAGATCAGCATAAGCGATGC GG ATCTCCAGTCCTGCTTTC
Dm-CVAE-without-selection-2 template	/52-Bio//iSp18/AAGCTCTAGTTCGGTCTG GGGGGAAAAGGAGCACAACTGAGAAGAGAGCAACAGCGGTAAC GA ATCTCCAGTCCTGCTTTC
Dm-CVAE-after-selection-1 template	/52-Bio//iSp18/AAGCTCTAGTTCGGTCTG ACAGCAGAGGCAGCAGCAGGGTTAGGGCTGGAGCTGCCATTAA GG ATCTCCAGTCCTGCTTTC

Dm-CVAE-after-selection-2 template	/52-Bio//iSp18/AAGCTCTAGTTCGGTCTG GGAGGACTGGAGGCAGGGCAACTGTCAAGGAGGTGGCCATAAA CG ATCTCCAGTCCTGCTTTC
Dm-CVAE-after-selection-3 template	/52-Bio//iSp18/AAGCTCTAGTTCGGTCTG CCATGACGAGCAGCAAGGTAGGGGGCACGACTGGAGGGGATG TAA ATCTCCAGTCCTGCTTTC
Dm-CVAE-after-selection-4 template	/52-Bio//iSp18/AAGCTCTAGTTCGGTCTG GCGCGAGAGGCAGCAGCAAGATCACAGTCGGGATCGCTGTCAG CG ATCTCCAGTCCTGCTTTC
Dm-CVAE-after-selection-5 template	/52-Bio//iSp18/AAGCTCTAGTTCGGTCTG AGGATACTAAGGGCATCAGTGGTGGAACGGCTGGAGATAGTAG AA ATCTCCAGTCCTGCTTTC
Dm-reference-based-random-after-selection-1 template	/52-Bio//iSp18/AAGCTCTAGTTCGGTCTG ACAGTACCGAAGCAAATAGTAGCAGTAGCGGTACGGTGGACAT GG ATCTCCAGTCCTGCTTTC
Dm-reference-based-random-after-selection-2 template	/52-Bio//iSp18/AAGCTCTAGTTCGGTCTG AGGCAAACGTTAGCAGCGGAGGGACCATGGGCATCAGGGGCAC AA ATCTCCAGTCCTGCTTTC
Dm-reference-based-random-after-selection-3 template	/52-Bio//iSp18/AAGCTCTAGTTCGGTCTG GGGGTACTAAGGTGGGGGATGATAGCAACAGGGCGGGCAAGA GCA ATCTCCAGTCCTGCTTTC
Dm-reference-based-random-after-selection-4 template	/52-Bio//iSp18/AAGCTCTAGTTCGGTCTG TGGGGGGGATGGCAACGGAAACGGCCGATGCTATAAGGAAGGA TG ATCTCCAGTCCTGCTTTC
Dm-reference-based-random-after-selection-5 template	/52-Bio//iSp18/AAGCTCTAGTTCGGTCTG GAAAAAGCAGTATAAAAAAGGAGATTATCGATAAGGGAGGAAAT G ATCTCCAGTCCTGCTTTC