AFRL-RI-RS-TR-2022-056

DATA DRIVEN DESIGN OF PROTEIN FUNCTION

UNIVERSITY OF WASHINGTON

MARCH 2022

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This report is available to the general public, including foreign nations. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TR-2022-056 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /
WILMAR W. SIFRE
Work Unit Manager

/ S /
GREGORY J. HADYNSKI
Assistant Technical Advisor
Computer & Communications Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
MARCH 2022	FINAL TECHNICAL REPORT	AUGUST 2017	AUGUST 2021

4. TITLE AND SUBTITLE

DATA DRIVEN DESIGN OF PROTEIN FUNCTION

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
FA8750-17-C-0219	N/A	61101E
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**
		R2CQ

6. AUTHOR(S)

Hugh Haddox, Ian Haydon, Kristina Herrera, Lance Stewart,and David Baker

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Washington
Box 351655
Seattle WA 98195-9472

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
Air Force Research Laboratory/RITA 525 Brooks Road Rome NY 13441-4505	RI	AFRL-RI-RS-TR-2022-056

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Our research first focused on a design task: designing mini-proteins that are stable, without requiring them to be functional. This work involved designing, testing, and learning from hundreds of thousands of mini-proteins, and resulted in machine learning (ML) models that are predictive of stability, as well as improvements to Rosetta's energy function. In the last two years, we have advanced to substantially more difficult tasks: designing proteins that are both stable and bind to a protein or DNA target. This work involved testing hundreds of thousands of binders, has resulted in dramatic increases to our success rates, and has provided several high-throughput datasets for future learning.

15. SUBJECT TERMS

De novo protein design; machine learning; Rosetta; mini-binding proteins

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
U	U	U	SAR	20

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
WILMAR W. SIFRE	**N/A**

PREVIOUS EDITION IS OBSOLETE.

Table of Contents

List of Figures

1.0 SUMMARY

Protein design applications range from the development of therapeutics to nanomaterials and biosensors, but the field still has large obstacles to overcome. First, the success rate of design is typically low. Second, iteration between designing, testing, and learning must be improved to efficiently optimize methods. Our work during the Synergistic Discovery and Design (SD2) program has helped address both of these obstacles.

At the start of SD2, our research focused on a relatively simple design task: designing mini-proteins (<70 amino acids) that are stable but non-functional. This work involved designing, testing, and learning from hundreds of thousands of mini-proteins, and resulted in machine learning (ML) models that are predictive of stability, as well as improvements to Rosetta's energy function. In the last two years, we have pursued much more difficult tasks: designing proteins that are both stable and bind to a protein or DNA target. This involved testing hundreds of thousands of binders, has yielded dramatic increases to our success rates, and has provided several high-throughput datasets for future learning. Several of the designed binders have therapeutic relevance, most notably mini-proteins that bind to the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) spike protein, which have been shown to reduce severity of infection in animals and are now being trialed for safety and efficacy in humans. Collectively, this work is a significant advance in our success rate in binder design and in our ability to rapidly design, test, and learn, both of which will serve as springboards for tackling important scientific challenges in the future.

2.0 INTRODUCTION

The Baker lab at the University of Washington seeks to understand the fundamental principles underlying protein structure and function, to encode these principles in the Rosetta computer program, and to use them to create a new world of *de novo* designed proteins to address 21st-century challenges in health and technology. In all cases, we start entirely from first principles; we do not re-engineer native proteins. By iterating between computation and experiment, we continually improve our design methodology.

Proteins carry out the critical functions of life: among other things, they catalyze chemical reactions, provide scaffolding to organize cells and compartments, regulate gene expression, convey signals, and transport molecules within the organism. The unique sequence of amino acids of a given protein drives the protein to adopt a very specific three-dimensional structure, or fold which in turn uniquely determines its function. If the rules by which amino acid sequences determine three dimensional structures and functions could be encoded in software, it would become possible to design molecular sensors, catalysts, and devices on the nanoscale with far greater potency than anything available today. Furthermore, unlike nanoscale devices currently in use, these designed nanomachines could be produced at the Avogadro's number (~6×10^{23}) scale using a universal production pipeline: they can be encoded in DNA, introduced into bacteria, and manufactured by standard protein expression and production procedures followed by spontaneous self-assembly (protein folding).

The leading software for predicting protein structure and function from sequence and designing new protein structures and functions is the Rosetta Molecular Modeling suite developed in our group[1–5]. In the past several years, we have improved Rosetta to the point where an expert can design a small number of new folds with new functions[6]. In this program, we aimed to dramatically improve capabilities for designing new proteins with new functions by transforming protein design into a data driven science.

2.1 Designing Stable Mini-Proteins.

Protein stability is often a prerequisite for a function. At the start of the SD2 program, one of our first goals was to learn basic principles of how to design stable proteins. Before SD2, the Baker lab had developed a high-throughput method for computationally designing and experimentally testing the stability of tens of thousands of mini-proteins [7]. Specifically, the high-throughput method quantified a protein's stability as a function of its resistance to protease degradation. There was no requirement for the protein to be functional in this assay -- only for it to stably fold. This initial study involved four cycles of design-test-learn and resulted in dramatically improved success rates in design. In total, the data consisted of ~16,000 miniproteins from four topologies. Building on this work, a central goal of the first few years of the SD2 program was to conduct additional design-test-learn cycles, and do so with an expanded repertoire of protein topologies to test the generalizability of our models. A second goal was to build SD2 infrastructure for use in addressing more difficult design challenges later in the program. This involved establishing a collaboration with the UW Biofabrication Center (BioFab), using Texas Advanced Computing Center (TACC) to design proteins, and building a GitLab repository to share data with machine learning (ML) collaborators.

2.2 Improving Rosetta's Energy Function.

Another one of our earliest goals in SD2 was to use high-throughput data on protein stability to improve Rosetta's energy function, which models the physics of proteins. An accurate energy function is crucial to design success, and improving its accuracy would presumably benefit the entire field of protein design, not just in designing proteins that are stable, but also ones that are functional. Before SD2, the energy function was trained almost exclusively on proteins from nature. While the energy function does a reasonable job of modeling these native proteins, many de novo protein designs still fail. Thus, we reasoned that it would be highly informative to retrain the energy function using data on do novo designs, allowing the energy function to learn from its past successes and failures, thereby resolving physical inaccuracies that lead to failure.

2.3 Designing Protein Binders.

After primarily focusing mainly on designing stable proteins for the first few years of SD2, we then advanced to a more difficult challenge: designing proteins that not only fold, but bind a protein target. Such binders have many applications, such as therapeutics against diseases like viruses or cancer. At the start of SD2, our success rate at designing protein binders was very low. For instance, it was not uncommon to get ~0-10 successful binders from a library of ~100,000

designs. And, before SD2, we had just started improving methods for rapidly designing and testing binders, but there was still much room for improvement.

2.4 Designing SARS-CoV-2 binders.

In early 2020, our team designed over two million candidate antiviral minibinder (MB) proteins targeting SARS-CoV-2 (Fig. 2A). Through computational modeling with Rosetta, and the reported nature of epitopes recognized by potent coronavirus neutralizing antibodies elicited upon infection, we identified several vulnerable sites on Sars-CoV-2 Spike (Spike or S protein), including the angiotensin converting enzyme 2 (ACE-2) cellular receptor recognition interface and the fusion peptide region.

2.5 Designing DNA Binders.

By the last year of the SD2 program, we had started to see a dramatic increase in the success rate of designing protein binders. We sought to leverage this success to address an even more difficult challenge: designing DNA binders. To our knowledge, before SD2, no one had ever designed a *de novo* DNA binder. If successful, such binders could be applied to engineering cellular circuits like the ones being used in other parts of the SD2 program, helping to fill the void left by the limited number of orthogonal native DNA binders that are currently available for use in the field.

3.0 METHODS, ASSUMPTIONS AND PROCEDURES

3.1 Designing Stable Mini-Proteins.

We designed proteins with Rosetta using computational resources from the University of Washington Baker lab, TACC, and Rosetta@Home. We collaborated with the BioFab to teach them how to perform the high-throughput stability assay from Rocklin et al.[7]. Briefly, this assay involves ordering tens of thousands of DNA oligonucleotides encoding computational protein designs. Next, it involves cloning these oligonucleotides into a plasmid for yeast display and transforming these plasmids into yeast, generating a library of cells, each displaying many copies of a single design. Designs are linked to a fluorescent marker for detection via fluorescence-activated cell sorting (FACS). Next, the library of yeast is treated with protease for a short time. Unfolded proteins tend to be cleaved much faster than folded proteins, and upon cleavage proteins will be detached from cells, along with their fluorescent tag, causing the cell to lose this label. FACS can then be used to separate populations of labeled cells with protease-resistant ("stable") designs from unlabeled cells with cleaved ("unstable") designs. Finally, deep sequencing is used to quantify the extent that each cell is enriched or depleted upon selection, allowing us to quantitatively estimate the stability of each protein.The BioFab went on to run this assay on several libraries. Note: this assay assumes that the stability scores from the protease assay are correlated with stability, which may not always be the case due to artifacts from yeast-display and since protease susceptibility is not a direct measurement of folding thermodynamics. We used a GitLab repo on TACC to share data with ML collaborators.

3.2 Improving Rosetta's Energy Function.

Park et al. describes the standard set of benchmarks used to assess the energy function's physical accuracy, and an automated ML protocol for optimizing parameters in the energy function based on these benchmarks [8]. We used this protocol when training and evaluating the energy function during SD2-related work. Briefly, this protocol involves quantifying Rosetta's performance on a standard set of benchmarks, which quantify Rosetta's accuracy in modeling high-resolution crystal structures, as well as thermodynamic data of small molecules. Scores from all benchmarks are summed together to generate a single number. This number is the output of the loss function that is minimized during parameter optimization. In parameter optimization, the Nelder-Mead simplex optimization method is used to efficiently search over parameter space to improve the loss function. During SD2, we used this approach to retune parameters in a way that reduced Rosetta's bias to introduce steric clashes. The updated energy function also had an improved aggregate score across benchmarks.

3.3 Designing SARS-CoV-2, Protein, and DNA Binders.

As with the design of stable proteins, we designed binders with Rosetta using computational resources from the Baker lab, TACC, and Rosetta@Home. The BioFab conducted a large number of the high-throughput binding assays that we used to experimentally test binders. In brief, these assays involve display of binder designs on the surface of yeast, incubating the yeast with a fluorescently labeled protein target, flow cytometry of yeast to select for cells with functional binders that have bound the target, and deep sequencing of cells before and after selection to determine which binders passed the selection.

4.0 RESULTS AND DISCUSSION

4.1 Designing Stable Mini-Proteins.

During the first few years of SD2, we designed and tested hundreds of thousands of new mini-proteins that spanned several new topologies. This >10X increase in testing was made possible by a few key pieces of infrastructure from SD2. First, we worked closely with the BioFab to help them master the high-throughput assay. This allowed them to rapidly and reproducibly test several mini-protein libraries within only a few years. Second, we leveraged the computational resources from TACC to analyze experimental results and then share them with ML collaborators in a structured and versioned-controlled GitLab repository. These data have led to multiple important discoveries, including improvements to Rosetta's energy function (described below), and ML models that are predictive of mini-protein stability (currently described in Jed Singer's manuscript in the publication pipeline). As part of this work, we also experimentally validated an improvement to a core computational algorithm used in protein design[9], and developed a method to systematically design mini-proteins with pockets with programmable shapes and sizes that are amenable to binding small molecules[10].

Aside from these discoveries, the first few years of work on protein stability were very useful in setting the stage for tackling more challenging design problems in the last few years. For

instance, our initial collaboration with the BioFab to run the high-throughput stability assay made it very easy to transition into running high-throughput binding assays. And our initial familiarity with TACC made it easy to transition into using TACC to help computationally design hundreds of thousands of mini-protein binders. Thus, being part of this ecosystem early on allowed us to use it later on to make rapid progress in designing binders.

4.2 Improving Rosetta's Energy Function.

At first, we only had limited success in this area, even after testing multiple libraries specifically tailored to help achieve this goal. However, in the second year of SD2, we finally had a breakthrough. With ML collaborators, we identified a set of ~20 designs that were highly interesting: although our models predicted that these designs should be highly stable, they were actually unstable in experiments. To gain insight into these designs, we experimentally tested the effects of all possible single-amino-acid mutations to each design. Excitingly, for most designs, we identified mutations that rescued the design, making it stable. By analyzing these mutations, we found that many of them relieved energetically unfavorable steric clashes that Rosetta had designed in the protein core, suggesting a physical inaccuracy in Rosetta's energy function—specifically, that it is too tolerant of such clashes. By re-tuning parameters in the energy function, we were able to resolve this bias (Fig. 1). Moreover, these changes substantially improved Rosetta's performance in a standard set of benchmarks used to quantify Rosetta's physical accuracy. Thus, we expect our changes to be generally useful in any design or modeling task.

While investigating this first inaccuracy, we unexpectedly identified two other inaccuracies in the energy function, which we are currently working to resolve. Although these discoveries did not directly come from high-throughput experiments through SD2, it was the original high-throughput experiment that set this chain of discoveries in motion. The first discovery came from analyzing crystal structures of designed proteins. We were curious to see if steric clashes designed by Rosetta were actually present in reality and found that many were actually absent in the experimentally determined structure, supporting our initial hypothesis. However, we also noticed a second difference: many of the designed salt bridges were also absent in the experimental structure. We traced this difference to a physically unrealistic bias in the energy function. In reality, salt bridges on the surfaces of proteins are weak due to electrostatic shielding from water. However, Rosetta is not aware of whether salt bridges are on the surface of proteins or buried in the core, so it does not model this shielding effect. Earlier this year, we implemented a new electrostatics term to resolve this inaccuracy. In turn, this fix helped us identify a third inaccuracy related to how Rosetta models the interaction between protein atoms and water. We have since implemented a new solvation term to resolve this third inaccuracy. We are eager to test whether these two newest fixes improve Rosetta's performance, which we will test using standard benchmarks and by designing and testing new mini-protein binders.

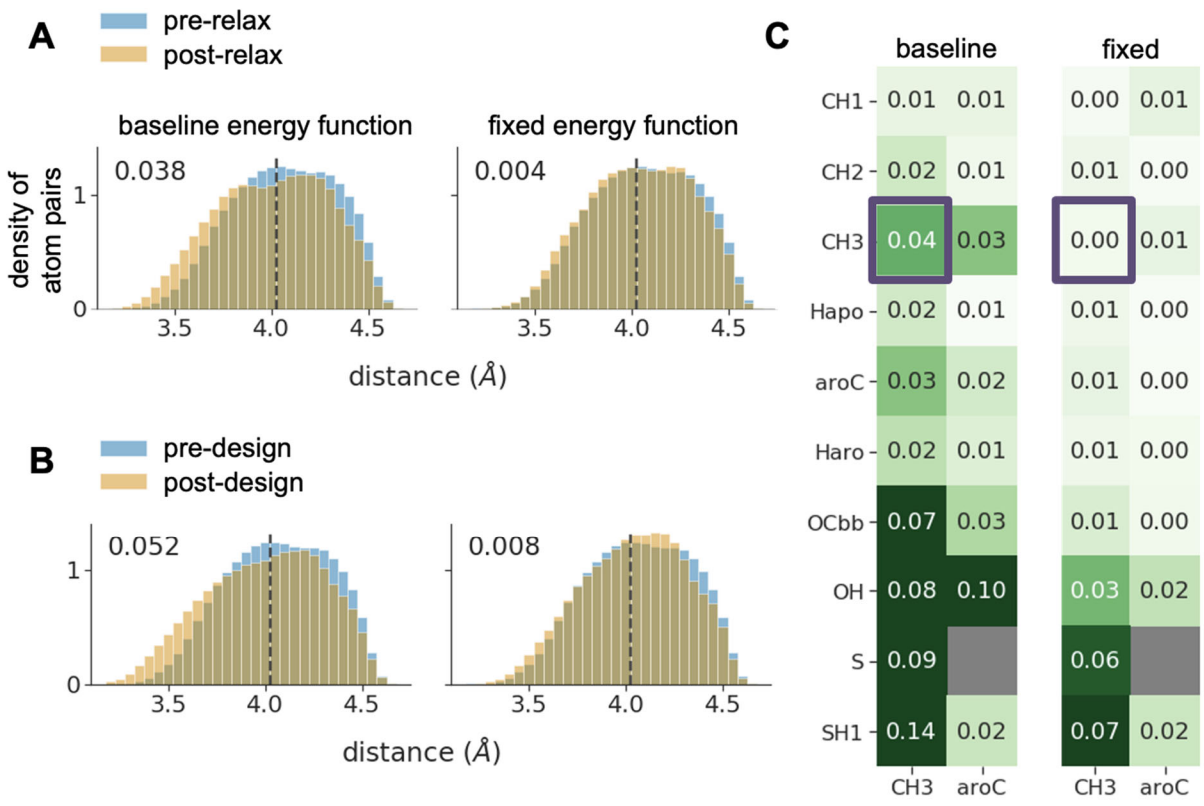

Figure 1: By retuning parameters in Rosetta's energy function, we eliminated a bias that led to energetically unfavorable steric clashing between atoms. Before SD2, Rosetta's energy function was too tolerant of steric clashing. A) For instance, we took ~80 high-resolution crystal structures of native proteins, relaxed their structures with Rosetta, and quantified distances between atoms. The left plot in this panel shows results from doing so with the standard Rosetta energy function, which was the baseline at the start of SD2. Specifically, it shows the distribution of all interatomic distances between pairs of methyl carbons, with the blue and orange distributions showing data before and after the relax, respectively. The vertical dashed line shows the sum of the van der Waals radii of these atoms -- atom pairs closer than this distance are considered to be "clashing". The orange distribution is shifted to the left of both the blue distribution and the vertical line, pointing to an unrealistic physical bias in Rosetta to introduce clashes. The right plot in this panel shows results of repeating this experiment with the new energy function that we retuned during SD2. This energy function does not show the same bias, as shown by the increased overlap between the orange and blue distributions. The numbers at the top left of each plot quantify this overlap as the Kullback–Leibler (KL) divergence, with numbers closer to zero indicating better overlap. B) This panel shows the same data as panel A, but with a protocol that allows Rosetta to completely redesign the protein sequence, as opposed to just relaxing the structure with a fixed sequence. As in (A), we observe a bias with the baseline energy function that is reduced upon retuning. C) Panels A and B show data for pairs of methyl carbons (CH3). This panel summarizes results across several pairs of atom types, for both the baseline energy function (left matrix) and fixed energy function (right matrix). Each box reports the KL divergence (number in top left of plots from panels A and B) for the indicated atom pair. Boxes in purple correspond to the data from panel A. The fixed energy function has improved KL divergence values across multiple atom pairs, showing general improvements upon retuning. (CH1, CH2, and CH3 are aliphatic carbons with different numbers of bonded hydrogens; aroC are aromatic carbons; Hapo and Haro are hydrogens from aliphatic and aromatic carbons, respectively; OCbb are oxygens from the protein backbone; OH are hydroxyl oxygens; S and SH1 are sulfur atoms without or with a bonded hydrogen).4.3 Designing Protein Binders, Including for SARS-CoV-2.

During SD2, TACC gave us resources for high-throughput design of millions of binders, and we used this resource to design libraries with ~100,000 binders for >20 different protein targets (Figure 2A). In turn, the BioFab helped us rapidly test each of these libraries in high-throughput binding experiments. Of note, the close collaboration between the UW and Biofab that we developed during work on protein stability made for a smooth transition to work on protein binding. Excitingly, our success rate dramatically improved over the course of the SD2 program. At the start of SD2, libraries of ~100,000 designs often yielded zero binders. During SD2, the success rate increased to 10s-100s of binders. Overall, this has led to a dramatic increase in the total number of successful binders identified via high-throughput experiments by the BioFab (Figure 2B). Following these experiments, the Baker lab has been validating and thoroughly characterizing individual binders using more precise low-throughput experiments. We have now validated multiple binders.

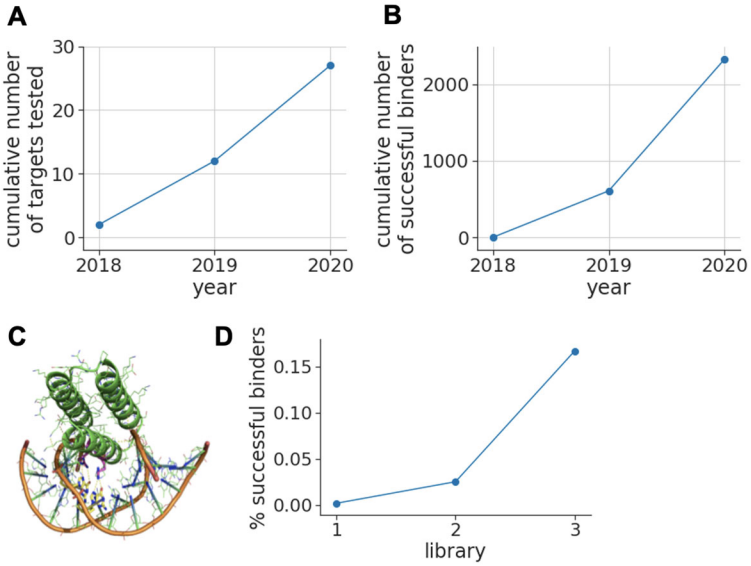

Figure 2. During SD2, we saw dramatic increases in the number of successful binders, for both protein binders and DNA binders. Panels A and B show data for protein binders, while C and D show data for DNA binders. A) Over a two year period, we dramatically increased the number of protein targets (e.g., SARS-Cov-2 RBD) that we put through our pipeline for binder design and testing. B) This increase, combined with improvements to design protocols, yielded a corresponding increase in the number of successful protein binders. C) In the last year of the SD2 program, we began designing and testing mini-proteins that bind to DNA, such as the one illustrated in this panel. D) During that year, we conducted three cycles of the design-test-learn pipeline, each time creating a new library with tens of thousands of DNA binders and experimentally testing them in the lab. Excitingly, our success rate has increased with each cycle.

Based on known binding modes of neutralizing mAbs, we focused our minibinder design pipeline on the ACE-2 receptor binding domain (RBD) at the distal end of CoV2 Spike. The best monomeric anti-CoV2 minibinders neutralize the virus with activities rivaling the most potent known antibodies[11], blocking infection of human cells with a half maximal inhibitory concentration (IC50) of ~15pM. When administered intranasally to K18-hACE2 transgenic mice[13]and Syrian hamsters[14] (Fig. 3B), these proteins also prevent infection as prophylaxis and prevent disease as a therapeutic at doses as low as 1 mg/kg[12].

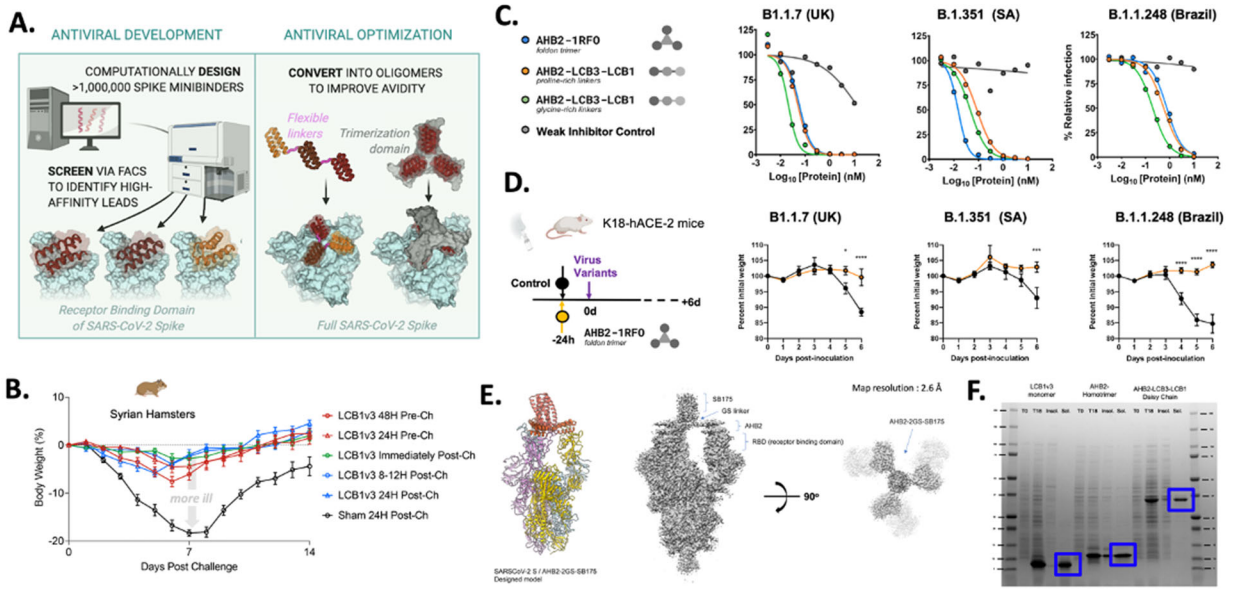

Figure 3. *De novo* Designed Anti-CoV2 RBD Minibinders. (A.) Rosetta design was used to compute the structures of high affinity MBs with slightly different binding modes predicted to bind the CoV2 viral Spike receptor binding domain and antagonize binding to cellular receptor ACE-2. The most potent antiviral MBs were then further designed as avid multi-headed daisy chained versions with flexible linkers, or geometric homotrimers to match the Spike trimer architecture. (B.) Syrian hamsters (N=9) received intranasal administration (i.n., 50 uL x 2 in both nostrils) of 1.5 mg/ml test articles (150 ug of LCB1v3, a monomeric anti-CoV2 MB with 15 pM affinity to the Wuhan RBD) or PBS control 16h prior to CoV2 challenge WA1/2020 (10^3 plaque forming units / mL @ 50 ul / nostril). Protein endotoxin levels are <10 E.U. / mg. Weight was measured daily *(Courtesy of Drs. Lisa Tostanoski and Daniel Barouch).* (C.) The multi-headed MB constructs are capable of neutralizing the most important CoV2 variants of concern in live virus neutralization assays with EC50s of ~0.02 nM conducted in the BSL-3 lab of Dr. Michael Diamond[13]. (D.) Single intranasal dosing of the AHB2 homotrimeric antiviral MB protects hACE-2 mice from lethal infection by the major CoV2 variants of concern as measured by weight loss over time following infection performed by the BSL-3 laboratory of Dr. Michael Diamond[13]. K18-hACE2-transgenic mice were dosed with 50ug C326-AHB2-1rfo by intranasal administration (i.n., 50 uL= 25 ul x 2 in both nostrils) 24h prior (T-24h) to infection with 10^3 plaque forming units of CoV2 Variants B.1.1.7, B1.351, B.1.1.24 intranasally at Day 0. Weight was measured daily thereafter and at 6 days post infection (6dpi) animals (n=6/timepoint). (E.) CryoEM structure determination of the AHB2 fused to a homotrimerizing domain (SB175) forms a highly avid bound structure to Spike trimer protein with RBDs in their open configuration, revealing a near atomic level accurate match with our design model (Structure determined in collaboration with Dr. Michael Jewett and David Veesler). (F.) The mininders can be readily produced in soluble form out of *E. coli* expression following auto induction (Time 0 vs Time 18 hours post induction). Following cell lysis by heating to 95°C and centrifugation to separate insoluble from soluble material, the LCB1v3 monomer MB, an AHB2 homotrimer, and three domain daisy chain AHB2-LCB3-LCB1 proteins are all >90% pure. Samples shown were analyzed by SDS-PAGE with reducing agent and Coomassie blue staining (molecular weight standards shown).

Based on proven strategies for generating multivalent minibinders to improve avidity for homotrimeric influenza hemagglutinin[15], the Baker, Veesler, Bloom, Diamond, Whelan, and Jewett labs collectively initiated a major campaign to design, select and characterize multivalent forms of MBs that resist escape mutagenesis, and are effective by single dose intranasal

administration in K18-hACE mice against all known mutant variants of concern25 that escape antibody responses[16–22] (Fig. 3C-D). The Veesler lab has solved cryo-EM structures confirming the binding modes to S trimers as designed (Fig. 3E).

4.4 Toward Clinical Development

Collectively, our data provides strong proof of concept for *de novo* designed anti-SARS-CoV-2 minibinders as viable agents for prevention or treatment of COVID-19. DARPA support together with philanthropic support (e.g. Schmidt Futures) has made this possible. The results have attracted the attention of the **Bill & Melinda Gates Foundation** (BMGF) and **International AIDS Vaccine Initiative** (IAVI) further amplifying your impact on our work. Both organizations will begin working with us to carefully define the target product profile for a multi-valent, escape resistant anti-SARS-CoV-2 minibinder as a preventative countermeasure for (i) preventing severe COVID-19 disease after infection, or (ii) protecting people from infection in high risk environments. We expect the final drug product to be a nebulized formulation that can be applied to the nasal cavity and/or inhaled into the lungs.

IAVI is a nonprofit research organization that develops vaccines and antibodies for human immunodeficiency virus (HIV), tuberculosis and emerging infectious diseases including COVID-19. We have also identified South Korea-based **SK Chemical** as a key partner for exploring manufacturing processes for our lead candidates and have completed material transfer agreements to initiate technology transfer in a multi-party collaboration with University of Washington (UW), BMGF, IAVI, and SK. Based on manufacturing and formulation development outcomes, we will be finalizing one primary candidate and one backup candidate for preclinical safety toxicology studies and eventual testing in people. Hence, we have now fully transitioned to conducting all the necessary preclinical research activities in manufacturing, formulation, and toxicology to submit an investigational new drug (IND) application to the Food and Drug Administration (FDA), enabling the initiation of human clinical trials in late 2022 or early 2023.

Additionally, the several high-throughput datasets generated during SD2 has provided an excellent opportunity for learning. And the Baker lab has recently used ML to identify multiple features that are predictive of success. One of the most predictive features is the size of the interface: larger interfaces tend to be more successful. Another feature is the hydrophobicity of the interface, where more hydrophobic interfaces tend to be more successful. However, this second feature is also nuanced. We discovered that large hydrophobic interfaces are more likely to fail if hydrophobic residues form large clusters. In contrast, they tend to be more successful if hydrophobic residues are interspersed with polar residues, which presumably help prevent designs from non-specifically aggregating via hydrophobic patches. A final predictive feature is the structural quality of the binder design in the absence of its target. Recently, the lab trained an ML algorithm to recognize crystal structures of native proteins. This model can be used to quantify how “native-like” designs are, which in part relates to whether the designed structure is physically realistic. We found that this quantity is also predictive of whether binders can successfully bind their target in an experiment. Overall, we have made considerable progress in designing protein binders in the last few years of SD2. With TACC, we have developed infrastructure to rapidly design binders. With the BioFab, we have developed infrastructure to rapidly experimentally test these binders. And we have learned features that are predictive of

binder success that can be applied to future design challenges. Together, these have all contributed to increasing our success rate in binder design from near zero successes per 100,000 designs to tens or hundreds of successes, which is a significant advance for the field. Manuscripts describing this work are either published or in the publication pipeline[23,24] (see Appendix I).

4.5 Designing DNA binders.

During the last year of SD2, we have tested three libraries of DNA binders, each with tens of thousands of designs. Our success rate has progressively increased with each attempt: a single hit with the first library, ~10 hits with the second library, and dozens of hits with the third library (Figure 2C and 2D). The increased success rate has likely come from multiple sources. First, we have made improvements to Rosetta's energy function for modeling protein-DNA interactions. Before SD2, this part of the energy function had not been updated for over a decade. During SD2, we established a set of biochemical benchmarks to quantify Rosetta's physical accuracy in modeling protein-DNA interactions. We then refit parameters in the energy function to optimize its performance on these benchmarks, which resulted in a large performance increase. Updated versions of the energy function were used to design the second and third libraries and may have contributed to increased experimental success. A second source of increased success may have come from improvements to our design protocol. As noted above, we identified multiple features that are predictive of success in designing protein-protein interactions. We also established an updated protocol for interface design that leads to higher-quality interfaces. We used this updated protocol in designing the second and third libraries, and used the predictive features to select which designs to order. These updates may have also contributed to increased experimental success.

With only ~100 successful binders, it has been difficult to use ML to learn principles that govern DNA binding. However, we are taking two approaches to improve our design pipeline. First, we are performing in-depth experimental characterization of a few of the successful binders. For example, for each of these binders, we have experimentally measured the effects of all single amino-acid mutations on binding. We plan to use these results to help validate that the design is binding at the expected location, and to test Rosetta's ability to predict the experimental results, which may reveal inaccuracies in Rosetta, as above. Next, we are comparing and contrasting our designed DNA binders with DNA binders from nature. We hypothesize that we can identify differences in binding mode that may reveal problems in our design protocol. Overall, this work has been a breakthrough in showing that it is possible to design DNA binders. We are eager to continue iterating on this challenge with design-test-learn cycles, with the eventual goal of using these binders to make synthetic cellular circuits.

5.0 CONCLUSIONS

A central goal of the start of the SD2 program was to use design-test-learn cycles to make rapid advances in design challenges. By applying this approach to protein design, we have made substantial advances in three design challenges of increasing difficulty: designing stable proteins, designing protein binders, and designing DNA binders. Not only have we increased our success rate in each of these challenges, but we have developed high-throughput methods that allow us to

rapidly design and test tens of thousands of proteins in each cycle. These methods leveraged computational resources from TACC, as well as the collaboration with the BioFab, which conducted a large number of high-throughput experiments. Learning from these data involved working with ML collaborators from SD2, as well as applying ML methods in the Baker lab, which included retraining the Rosetta energy function. All together, the above computational, experimental, and human resources through SD2 created an ecosystem that led to several exciting advances, from ML models that are predictive of stability, to improvements in Rosetta's energy function, to improvements in success rates of designing protein and DNA binders. The increase in success rates sets the stage for design to tackle important scientific challenges in the future. And the detailed framework that we developed for rapid and high-throughput design-test-learn cycles sets the stage for increasing these success rates even more.

6.0 REFERENCES

1. Davis IW, Baker D. RosettaLigand docking with full ligand and receptor flexibility. J Mol Biol. 2009 Jan 16;385(2):381–392. PMID: 19041878
2. DiMaio F, Leaver-Fay A, Bradley P, Baker D, André I. Modeling symmetric macromolecular structures in Rosetta3. PLoS One. 2011 Jun 22;6(6):e20450. PMCID: PMC3120754
3. Huang P-S, Ban Y-EA, Richter F, Andre I, Vernon R, Schief WR, Baker D. RosettaRemodel: a generalized framework for flexible backbone protein design. PLoS One. 2011 Aug 31;6(8):e24109. PMCID: PMC3166072
4. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, Kaufman K, Renfrew PD, Smith CA, Sheffler W, Davis IW, Cooper S, Treuille A, Mandell DJ, Richter F, Ban Y-EA, Fleishman SJ, Corn JE, Kim DE, Lyskov S, Berrondo M, Mentzer S, Popović Z, Havranek JJ, Karanicolas J, Das R, Meiler J, Kortemme T, Gray JJ, Kuhlman B, Baker D, Bradley P. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. 2011;487:545–574. PMCID: PMC4083816
5. Ovchinnikov S, Kinch L, Park H, Liao Y, Pei J, Kim DE, Kamisetty H, Grishin NV, Baker D. Large-scale determination of previously unsolved protein structures using evolutionary information. Elife. 2015 Sep 3;4:e09248. PMCID: PMC4602095
6. Huang P-S, Boyken SE, Baker D. The coming of age of de novo protein design. Nature. 2016 Sep 15;537(7620):320–327. PMID: 27629638
7. Rocklin GJ, Chidyausiku TM, Goreshnik I, Ford A, Houliston S, Lemak A, Carter L, Ravichandran R, Mulligan VK, Chevalier A, Arrowsmith CH, Baker D. Global analysis of protein folding using massively parallel design, synthesis, and testing. Science. 2017 Jul 14;357(6347):168–175. PMCID: PMC5568797
8. Park H, Bradley P, Greisen P Jr, Liu Y, Mulligan VK, Kim DE, Baker D, DiMaio F. Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. J Chem Theory Comput. 2016 Dec 13;12(12):6201–6212. PMCID: PMC5515585
9. Maguire JB, Haddox HK, Strickland D, Halabiya SF, Coventry B, Griffin JR, Pulavarti SVSRK, Cummins M, Thieker DF, Klavins E, Szyperski T, DiMaio F, Baker D, Kuhlman B. Perturbing the energy landscape for improved packing during computational protein design. Proteins. 2021 Apr;89(4):436–449. PMCID: PMC8299543
10. Basanta B, Bick MJ, Bera AK, Norn C, Chow CM, Carter LP, Goreshnik I, Dimaio F, Baker D. An enumerative algorithm for de novo design of proteins with diverse pocket

structures. Proc Natl Acad Sci U S A. 2020 Sep 8;117(36):22135–22145. PMCID: PMC7486743
11. Cao L, Goreshnik I, Coventry B, Case JB, Miller L, Kozodoy L, Chen RE, Carter L, Walls AC, Park Y-J, Strauch E-M, Stewart L, Diamond MS, Veesler D, Baker D. De novo design of picomolar SARS-CoV-2 miniprotein inhibitors. Science. 2020 Oct 23;370(6515):426–431. PMCID: PMC7857403
12. Case JB, Chen RE, Cao L, Ying B, Winkler ES, Johnson M, Goreshnik I, Pham MN, Shrihari S, Kafai NM, Bailey AL, Xie X, Shi P-Y, Ravichandran R, Carter L, Stewart L, Baker D, Diamond MS. Ultrapotent miniproteins targeting the SARS-CoV-2 receptor-binding domain protect against infection and disease. Cell Host Microbe. 2021 Jul 14;29(7):1151–1161.e5. PMCID: PMC8221914
13. Case JB, Rothlauf PW, Chen RE, Liu Z, Zhao H, Kim AS, Bloyet L-M, Zeng Q, Tahan S, Droit L, Ilagan MXG, Tartell MA, Amarasinghe G, Henderson JP, Miersch S, Ustav M, Sidhu S, Virgin HW, Wang D, Ding S, Corti D, Theel ES, Fremont DH, Diamond MS, Whelan SPJ. Neutralizing Antibody and Soluble ACE2 Inhibition of a Replication-Competent VSV-SARS-CoV-2 and a Clinical Isolate of SARS-CoV-2. Cell Host Microbe. 2020 Sep 9;28(3):475–485.e5. PMCID: PMC7332453
14. Muñoz-Fontela C, Dowling WE, Funnell SGP, Gsell P-S, Riveros-Balta AX, Albrecht RA, Andersen H, Baric RS, Carroll MW, Cavaleri M, Qin C, Crozier I, Dallmeier K, de Waal L, de Wit E, Delang L, Dohm E, Duprex WP, Falzarano D, Finch CL, Frieman MB, Graham BS, Gralinski LE, Guilfoyle K, Haagmans BL, Hamilton GA, Hartman AL, Herfst S, Kaptein SJF, Klimstra WB, Knezevic I, Krause PR, Kuhn JH, Le Grand R, Lewis MG, Liu W-C, Maisonnasse P, McElroy AK, Munster V, Oreshkova N, Rasmussen AL, Rocha-Pereira J, Rockx B, Rodríguez E, Rogers TF, Salguero FJ, Schotsaert M, Stittelaar KJ, Thibaut HJ, Tseng C-T, Vergara-Alert J, Beer M, Brasel T, Chan JFW, García-Sastre A, Neyts J, Perlman S, Reed DS, Richt JA, Roy CJ, Segalés J, Vasan SS, Henao-Restrepo AM, Barouch DH. Animal models for COVID-19. Nature. 2020 Oct;586(7830):509–515. PMCID: PMC8136862
15. Strauch E-M, Bernard SM, La D, Bohn AJ, Lee PS, Anderson CE, Nieusma T, Holstein CA, Garcia NK, Hooper KA, Ravichandran R, Nelson JW, Sheffler W, Bloom JD, Lee KK, Ward AB, Yager P, Fuller DH, Wilson IA, Baker D. Computational design of trimeric influenza-neutralizing proteins targeting the hemagglutinin receptor binding site. Nat Biotechnol. 2017 Jul;35(7):667–671. PMCID: PMC5512607
16. Wang Z, Schmidt F, Weisblum Y, Muecksch F, Barnes CO, Finkin S, Schaefer-Babajew D, Cipolla M, Gaebler C, Lieberman JA, Oliveira TY, Yang Z, Abernathy ME, Huey-Tubman KE, Hurley A, Turroja M, West KA, Gordon K, Millard KG, Ramos V, Da Silva J, Xu J, Colbert RA, Patel R, Dizon J, Unson-O'Brien C, Shimeliovich I, Gazumyan A, Caskey M, Bjorkman PJ, Casellas R, Hatziioannou T, Bieniasz PD, Nussenzweig MC. mRNA vaccine-elicited antibodies to SARS-CoV-2 and circulating variants. Nature. 2021 Apr;592(7855):616–622. PMCID: PMC8503938
17. Wibmer CK, Ayres F, Hermanus T, Madzivhandila M, Kgagudi P, Oosthuysen B, Lambson BE, de Oliveira T, Vermeulen M, van der Berg K, Rossouw T, Boswell M, Ueckermann V, Meiring S, von Gottberg A, Cohen C, Morris L, Bhiman JN, Moore PL. SARS-CoV-2 501Y.V2 escapes neutralization by South African COVID-19 donor plasma [Internet]. Available from: http://dx.doi.org/10.1101/2021.01.18.427166
18. Wu K, Werner AP, Moliva JI, Koch M, Choi A, Stewart-Jones GBE, Bennett H, Boyoglu-

Barnum S, Shi W, Graham BS, Carfi A, Corbett KS, Seder RA, Edwards DK. mRNA-1273 vaccine induces neutralizing antibodies against spike mutants from global SARS-CoV-2 variants [Internet]. Available from: http://dx.doi.org/10.1101/2021.01.25.427948
19. Starr TN, Greaney AJ, Hilton SK, Crawford KHD, Navarro MJ, Bowen JE, Alejandra Tortorici M, Walls AC, Veesler D, Bloom JD. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding [Internet]. Available from: http://dx.doi.org/10.1101/2020.06.17.157982
20. Greaney AJ, Loes AN, Crawford KHD, Starr TN, Malone KD, Chu HY, Bloom JD. Comprehensive mapping of mutations to the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human serum antibodies [Internet]. Available from: http://dx.doi.org/10.1101/2020.12.31.425021
21. Liu Z, VanBlargan LA, Bloyet L-M, Rothlauf PW, Chen RE, Stumpf S, Zhao H, Errico JM, Theel ES, Liebeskind MJ, Alford B, Buchser WJ, Ellebedy AH, Fremont DH, Diamond MS, Whelan SPJ. Landscape analysis of escape variants identifies SARS-CoV-2 spike mutations that attenuate monoclonal and serum antibody neutralization. bioRxiv [Internet]. 2021 Jan 11; Available from: http://dx.doi.org/10.1101/2020.11.06.372037 PMCID: PMC7805447
22. Andreano E, Piccini G, Licastro D, Casalino L, Johnson NV, Paciello I, Dal Monego S, Pantano E, Manganaro N, Manenti A, Manna R, Casa E, Hyseni I, Benincasa L, Montomoli E, Amaro RE, McLellan JS, Rappuoli R. SARS-CoV-2 escape from a highly neutralizing COVID-19 convalescent plasma. Proc Natl Acad Sci U S A [Internet]. 2021 Sep 7;118(36). Available from: http://dx.doi.org/10.1073/pnas.2103154118 PMCID: PMC8433494
23. Cao L, Coventry B, Goreshnik I, Huang B, Park JS, Jude KM, Marković I, Kadam RU, Verschueren KHG, Verstraete K, Walsh STR, Bennett N, Phal A, Yang A, Kozodoy L, DeWitt M, Picton L, Miller L, Strauch E-M, Halabiya S, Hammerson B, Yang W, Benard S, Stewart L, Wilson IA, Ruohola-Baker H, Schlessinger J, Lee S, Savvides SN, Christopher Garcia K, Baker D. Robust de novo design of protein binding proteins from target structural information alone [Internet]. bioRxiv. 2021 [cited 2021 Nov 10]. p. 2021.09.04.459002. Available from: https://www.biorxiv.org/content/10.1101/2021.09.04.459002v1.full
24. Singer JM, Novotney S, Strickland D, Haddox HK, Leiby N, Rocklin GJ, Chow CM, Roy A, Bera AK, Motta FC, Cao L, Strauch E-M, Chidyausiku TM, Ford A, Ho E, Mackenzie CO, Eramian H, DiMaio F, Grigoryan G, Vaughn M, Stewart LJ, Baker D, Klavins E. Large-scale design and refinement of stable proteins using sequence-only models [Internet]. bioRxiv. 2021 [cited 2021 Nov 18]. p. 2021.03.12.435185. Available from: https://www.biorxiv.org/content/10.1101/2021.03.12.435185v2

7.0 PUBLICATIONS RESULTING FROM PROGRAM

For additional information, please see the following publications resulting from the SD2 program.

Basanta B, Bick MJ, Bera AK, Norn C, Chow CM, Carter LP, Goreshnik I, Dimaio F, Baker D. An enumerative algorithm for de novo design of proteins with diverse pocket structures. Proc Natl Acad Sci U S A. 2020 Sep 8;117(36):22135–22145. PMCID: PMC7486743

Cao L, Coventry B, Goreshnik I, Huang B, Park JS, Jude KM, Marković I, Kadam RU, Verschueren KHG, Verstraete K, Walsh STR, Bennett N, Phal A, Yang A, Kozodoy L, DeWitt M, Picton L, Miller L, Strauch E-M, Halabiya S, Hammerson B, Yang W, Benard S, Stewart L, Wilson IA, Ruohola-Baker H, Schlessinger J, Lee S, Savvides SN, Christopher Garcia K, Baker D. Robust de novo design of protein binding proteins from target structural information alone [Internet]. bioRxiv. 2021 [cited 2021 Nov 10]. p. 2021.09.04.459002. Available from: https://www.biorxiv.org/content/10.1101/2021.09.04.459002v1.full

Cao L, Goreshnik I, Coventry B, Case JB, Miller L, Kozodoy L, Chen RE, Carter L, Walls AC, Park Y-J, Strauch E-M, Stewart L, Diamond MS, Veesler D, Baker D. De novo design of picomolar SARS-CoV-2 miniprotein inhibitors. Science. 2020 Oct 23;370(6515):426–431. PMCID: PMC7857403

Case JB, Chen RE, Cao L, Ying B, Winkler ES, Johnson M, Goreshnik I, Pham MN, Shrihari S, Kafai NM, Bailey AL, Xie X, Shi P-Y, Ravichandran R, Carter L, Stewart L, Baker D, Diamond MS. Ultrapotent miniproteins targeting the SARS-CoV-2 receptor-binding domain protect against infection and disease. Cell Host Microbe. 2021 Jul 14;29(7):1151–1161.e5. PMCID: PMC8221914

Maguire JB, Haddox HK, Strickland D, Halabiya SF, Coventry B, Griffin JR, Pulavarti SVSRK, Cummins M, Thieker DF, Klavins E, Szyperski T, DiMaio F, Baker D, Kuhlman B. Perturbing the energy landscape for improved packing during computational protein design. Proteins. 2021 Apr;89(4):436–449. PMCID: PMC8299543

Singer JM, Novotney S, Strickland D, Haddox HK, Leiby N, Rocklin GJ, Chow CM, Roy A, Bera AK, Motta FC, Cao L, Strauch E-M, Chidyausiku TM, Ford A, Ho E, Mackenzie CO, Eramian H, DiMaio F, Grigoryan G, Vaughn M, Stewart LJ, Baker D, Klavins E. Large-scale design and refinement of stable proteins using sequence-only models [Internet]. bioRxiv. 2021 [cited 2021 Nov 18]. p. 2021.03.12.435185. Available from: https://www.biorxiv.org/content/10.1101/2021.03.12.435185v2

8.0 SYMBOLS, ABBREVIATIONS, AND ACRONYMS

ACE-2	Angiotensin Converting Enzyme 2
BioFab	UW Biofabrication Center
BMGF	Bill & Melinda Gates Foundation
CH3	Methyl
DARPA	Defense Advanced Research Projects Agency
DNA	deoxyribonucleic acid
FACS	Fluorescence Activated Cell Sorting
HIV	Human Immunodeficiency Virus
IAVI	International AIDS Vaccine Initiative
IND	Investigational New Drug
IC50	Half maximal Inhibitory Concentration
KL	Kullback–Leibler
MB	Minibinder
ML	Machine Learning
RBD	Receptor Binding Domain
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
SD2	Synergistic Discovery and Design
TACC	Texas Advanced Computing Center
UW	University of Washington