Designing Next Generation Nucleic Acid Diagnostics Using Synthetic Biology and Artificial Intelligence

by

Nicolaas Manuel Angenent-Mari

B.S, University of Wisconsin-Madison (2016)

Submitted to the Department of Biological Engineering in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Biological Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2021

Author . . . .
Department of Biological Engineering
May 24, 2021

Certified by . . . .
James J. Collins
Professor of Biological Engineering
Thesis Supervisor

Accepted by . . . .
Feng Zhang
Professor of Biological Engineering

Designing Next Generation Nucleic Acid Diagnostics Using Synthetic Biology and Artificial Intelligence

by

Nicolaas Manuel Angenent-Mari

Submitted to the Department of Biological Engineering in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Biological Engineering

Abstract.

Nucleic Acid Testing (NAT) is an indispensable tool for effective disease diagnosis. Analyzing pathogen and host RNA or DNA often provides otherwise unobtainable information necessary for proper patient treatment. Unfortunately, a number of technical barriers prevent the expansion of NAT technology into novel application spheres, such as wearable, digital, and direct-to-consumer or point-of care diagnostic testing. These limitations include the cost of NAT, the equipment needed to perform it, and assay sensitivity. No available technologies have simultaneously achieved the combination of a consumer-tolerable cost, equipment-free passive operation, and gold standard sensitivity. The design of novel assays that overcoming all such limitations in concert would allow for the deployment of NAT in previously unprecedented environments, improving the range and accessibility of crucial disease monitoring.

In this thesis I outline four efforts to expand the capacity of NAT in this direction. First, I describe the design of a novel CRISPR-Cas13 activated riboswitch which demonstrates the potential of synthetic biology tools for nucleic acid detection. Second, I describe the prototyping of a platform for the deployment of freeze-dried synthetic biology-based diagnostic assays in wearable formats, including examples of NAT assays and also assays for small molecule analytes. Third, I describe the synthesis and subsequent analysis using deep learning of a toehold switch library, demonstrating the potential for high-throughput AI-guided design of diagnostic tools. Fourth, I describe the design of a novel isothermal nucleic acid amplification method that functions at low temperatures. I conclude by discussing the future direction of NAT technologies, and describe new opportunities for improved health outcomes that could arise from a new generation of diagnostic tools.

Thesis Supervisor: James J. Collins
Title: Professor of Biological Engineering

Acknowledgements

I have to start this section with my parents. Mama and Papa, you always cultivated curiosity in me and encouraged me to not be afraid of puzzles and interesting but difficult problems. It's not a coincidence that I dressed up as a carbon atom for Halloween when I was ten – I got it in my head that it would be wild and amazing to be an atom because of how you raised me and no other reason. The best part of it is that you never pressured me, or caused me to stress over school. You never had expectations for my grades, you just showed Vicky and I how fun math and science can be and left the rest up to us. I didn't grow up associating school with a fear of failure or disappointing you. So thank you both, I wouldn't have reached grad school without that incredible starting path you built for us. I count myself incredibly lucky.

There have of course been countless other people who have pushed me and pulled me upward along my journey. I want to thank Jim for inviting me into his group and teaching me the true challenges of practicing science. When I showed up at MIT I knew how to set up experiments and dream up design solutions, but I struggled to choose projects that balanced reward and chances for success, and I had yet to learn what it takes to finish something. These have been invaluable lessons that I will continue to apply to many aspects of life.

I want to thank my committee members Feng Zhang and Jacquin Niles for providing both professional and personal advice, particularly during the tougher parts of my PhD (the end, as is common). I feel much more equipped to explore the job market with their counsel.

To the Collins lab – wow, what a phenomenal bunch. If anything, I feel that I've taken for granted the genuine good nature of the group, and how effortless it was to work on teams with them. I want to thank Luis for his endless enthusiasm, Helena for helping me keep my Spanish in solid shape for four years, Angelo for sharing his poetry with me, Shimyn for being grumpy but

secretly having a heard of gold, Rose for having interesting opinions on cannibalism, Devora for teaching me about Jewish dating apps, Evan for primo lunchtime banter, Peter for many fascinating technical discussions, Nina for always being dependable, Ian for secretly being a master manipulator in Mafia-style games, Bernie for his calm insightfulness, Erica for keeping our grad student subgroup together, Ally for being a patient beta reader, Jackie for humoring my sporadic ML questions, Atti for briefly and gloriously bringing Civ5 to IMES, Raph for keeping the 100-years war alive, Miguel for being an empathetic man, Sarah for having excellent taste in SciFi books, Melissa for being my lab-mom while I rotated, Chad for being the Chaddiest Chad, and Nate for his conspiracy theories. There was a lot of scientific mentoring that also happened, but these will be the important things that I remember looking back.

Vicky, I've been a bit of an absent brother in the past and we haven't always talked as much I'd like, but I'm so grateful that we've turned that around. I really hope when we're Mama's age, that we can stay as close as she is with all her siblings.

Carissa, it's been indescribably weird not having you here for this last leg of the journey. I'm amazed at how much both of us have grown since separating, but I know that everything I've learned about myself this last year comes at the top of a mountain of lessons I learned from you. You were there to console me through an infinite number of tough times in lab starting when I was 18, and I'll never forget it. I hope you feel like this is partially your victory, too.

And lastly, I'd like to thank all the friends I've made these last five years, both in the BE department and elsewhere. You guys are the real reason I get up every day, definitely not science. You are too many to enumerate, but I hope to hold onto the connections I have with all of you as I transition onto the next stages of my life! Keeping long-term friends over the years has always been hard for me, but I will try my darndest!

Contents

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Bibliography

Chapter 1

1.1 Background and Motivation for Next-Generation Nucleic Acid Testing.

Nucleic Acid Testing (NAT) is an indispensable tool for effective disease diagnosis. Analyzing pathogenic RNA or DNA often provides otherwise unobtainable information necessary for proper patient treatment [1,2]. Compared with assays for other relevant analytes such as proteins and small molecules, NAT assays have the potential for orders of magnitude greater sensitivity [3,4], and frequently encode strain-type information with greater specificity [4]. Unfortunately, compared to other types of diagnostics like antigen-based tests, greater hurdles exist for deploying NAT in low-resource settings (Figure 1-1). These barriers include high per-test costs due to expensive materials, complex and expensive equipment needs, and variable robustness, among other limitations [1,2,3]. A "next-generation" NAT technology circumventing these limitations could dramatically expand the range of applications suitable for nucleic acid diagnostics, improving accessibility in medical clinics, bringing NAT to the consumer home, and allowing for deployment in a number of passive surveillance capacities such as wearable and wireless assays.

In the last decade, a concerted push to solve this problem has produced many technologies that attempt to address NAT's inherent drawbacks. In particular, there has been an emphasis on point-of-care (POC) technologies. Such "sample-in, results-out" testing platforms have focused on infectious diseases and try to satisfy the WHO's ASSURED criteria: affordable, sensitive, specific, user-friendly, rapid and robust, equipment-free, and deliverable to end-users [1,2,3]. A combination of simplified temperature control equipment, improved biochemistry, and microfluidics have allowed for significant progress. Their use precludes the need for transporting patient samples to faraway diagnostic labs, reducing diagnosis time and bringing NAT to clinics in communities

otherwise unable to perform NAT testing. However, despite progress with POC testing, many

hurdles remain for bringing NAT to emerging economies.

Figure 1-1. Complexity of protein and nucleic acid assays. Available diagnostic technologies are compared for nucleic acid based and antigen (protein) based assays. While tests for protein analytes are available at a range of complexities allowing for useful clinical information to be diagnosed at hospitals, smaller clinics, or at the home, tests for nucleic acid analytes tend to be more complex and expensive such that at-home tests have not properly been marketed and diagnostic availability is limited for smaller clinics. In particular this thesis aims to fill the missing gap of an at-home nucleic acid test.

1.2 Summary of Unresolved Barriers for Point-of-Care NAT Diagnostics

Unfortunately, point-of-care rarely equates with reduced per-test costs compared to laboratory-based testing, and equipment is still always required. Currently the per-test cost for a POC NAT is $10-50, and the equipment cost is $1,000-10,000 [4,5,6]. By contrast, lab-based testing at centralized facilities requires more expensive equipment, but once sample transport chains are established, it is cheaper for the patient on a per-test basis (material costs can be as low as $1). Therefore, local clinics often prefer to send samples to a centralized lab. This saves their patients

money, and also spares the local clinic the expense of having to purchase any equipment [2,3,5,6]. POC platforms hence reduce the incentive to expand diagnostics closer to the patient due to uncompetitive per-test and equipment costs, defeating their purpose.

Additionally, while $10-50 is cheaper than previous platforms, hundreds of millions still live below the global poverty line of $1.90/day [7]. For comparison, a test purchased at that income level equates to $160-800 per test in the US based on a $30/day poverty line. One can see significant room for price reduction, as many cannot afford even this reduced-cost testing.

Lastly and most importantly, cost aside, the needs for NAT diagnostics have moved beyond single-plex sequence detection (which is what most POC platforms have focused on). When combating both infectious and non-infectious illnesses, more in-depth sequence information on multiple targets is often necessary for proper treatment decisions and disease monitoring. Hence significant need exists for tests capable of higher-throughput screening of dozens of nucleic acid sequences and single-nucleotide polymorphisms (SNP's), while still reducing costs and maintaining point-of-care utility.

To highlight the magnitude of the challenges to be overcome, I will discuss in more detail current gold-standard POC NAT technologies using isothermal amplification, as well as the needs for more multiplexed and sequence-specific nucleic acid testing.

1.3 Equipment vs Price Tradeoffs for Isothermal Amplification Technologies

By far the closest competitors for an all-purpose "next generation" POC NAT technology are those that utilize isothermal amplification as their core detection mechanism. I will outline their benefits as well as how they remain limited as per the previously stated hurdles.

The main advance of isothermal amplification techniques (the most commonly used being Recombinase Polymerase Amplification or RPA, Loop-Mediated Isothermal Amplification or LAMP, Nucleic Acid Sequence Based Amplification or NASBA, and Strand Displacement Amplification or SDA) is to remove the requirement for temperature cycling inherent to the previous gold-standard technology qPCR [4,8,9,10,11]. These amplification detection schemes all function at either a single temperature in the case of NASBA and LAMP (41C and 65C, respectively), or at a narrow range of temperatures in the case of RPA and SDA (37C-48C), which significantly simplifies the technology needed to produce temperature control [4,10,11]. Unfortunately however, this does not eliminate the need for equipment, as even isothermal temperature control still requires electricity-driven devices, often with precise and robust control needed (especially with 65C incubation for LAMP and 41C for NASBA). A case has been made for equipment-free RPA amplification: since the assay functions at body temperature for example, it can sometimes be run simply by holding it close enough to a person [10]. In chapter two of this thesis I outline a wearable NAT assay we developed taking advantage of this. However, there are many situations where body heat is not available (ie, remotely placed digital and wireless diagnostics), and in addition not all humans produce enough body heat to reliably keep a skin-surface reaction at 37C and layers of clothing may separate the diagnostic from the skin of the user [12]. Because the function of RPA declines between 37C and 25C, this significantly limits the range of applications for which RPA can be applied without the use of external equipment, leaving further room for improvement.

Even when not considering equipment, isothermal amplification methods also have significant limitations in terms of price and sensitivity. The cost of isothermal amplification reagents is not significantly cheaper than that of previously available qPCR reagents, and the cartridge materials needed to carry out a POC assay using them can be more expensive than the

reagents used in qPCR assays [13]. In addition, isothermal amplification methods often suffer from reduced sensitivity compared to qPCR by up to several orders of magnitude [14,15,16].

Nonetheless despite these drawbacks, many POC nucleic acid diagnostic platforms currently exist, able to employ both conventional qPCR and isothermal amplification. These mostly include products that still require equipment in addition to a disposable assay cartridge, such as the Cepheid GeneXpert Omni, Lumora BART, Alere Q HIV, DRW SAMBA II, and Iquum Liat (Figure 1-2) [1,2,3,4,5,6]. Such products are portable, room-temperature storable, operate on battery power, and are relatively simple to use. However, the need to purchase equipment precludes at-home use and limits deployment for next-generation applications such as clothing-embedded sensors and remote digital sensors.

Cepheid Omni   **SAMBA II Semi-Q**

Figure 1-2. Example point-of-care NAT technologies. Two technologies for "sample-in, results-out" nucleic acid testing are shown, both of which are at present commercially available. The Cepheid Omni relies on qPCR assays and miniaturized thermocycling technology, while the SAMBA II uses isothermal amplification and a single-temperature heating step. Both have equipment costs on the order of ~$2-10k, and cartridge costs on the order of $20-50.

Other products where the equipment has been miniaturized sufficiently to be incorporated directly into an assay cartridge will soon reach the market, including at-home SARS-Cov2 tests from Sense and Lucira [25,26]. These do eliminate the equipment cost, but at the price of increasing the cartridge expense from $10-20 to $50-100, making it more of a luxury product than a broadly accessible health care device. This added cost arises from the continued need to heat the reaction, actuate liquid transfer, and purify the sample on-chip. Aside from price, these same limitations also preclude the deployment of such assays in clothing-embedded sensors and remote digital sensors.

In short, based on the previously outlined criteria, isothermal amplification methods have significant limitations in deployment as a "next-generation" NAT.

1.4 Screening for Drug Resistance and Viral Escape Mutations

One of the most challenging and largely unresolved issues in global health is the rise of drug resistant and viral escape pathogens.

Major diseases affected by drug resistance include HIV, HCV, TB, and malaria [17,18,19]. For many such diseases new treatments remain in the early stages of testing or are entirely absent, meaning that the need to contain their spread through careful monitoring is urgent. While most drug-resistance assays culture the pathogen to directly asses the organism's drug resistance, this is a costly process and can take weeks before diagnosis, limiting throughput and accessibility [21]. However, another way to asses a pathogen's drug resistance potential is through a mutation screen for single nucleotide polymorphisms (SNP's) known to confer drug resistance [20,21,22]. This assay type has the potential to be significantly faster in turnaround time.

Viral escape has been a similarly challenging health issue, most notably during the ongoing SARS-Cov2 pandemic, and the diagnosis of evolving viral strains potentially able to evade vaccine-conferred immunity can also require sequence-level nucleic acid information [23,24]. However even prior to the SARS-Cov2 pandemic, the monitoring of other viral diseases like Flavivirus infection was of significant concern (ie, Zika, Dengue, Chikungunya, West Nile, Yellow Fever). The ranges of these viruses often change rapidly and overlap, present with similar disease symptoms, and antibody-based assays can fail to distinguish them [23,24]. For example, antibodies against Zika, Dengue, and Chikungunya often give identical results. Additionally, unique strains of each virus can be difficult to distinguish by antibody, which complicates monitoring the spread of the pathogen as it evolves.

More recently, a number of SARS-CoV-2 strains of concern have emerged [27]. These new variants may affect transmissibility, treatment efficacy, and the degree of immunity that is generated by both natural infection and immunization [28]. Of particular concern are variants B.1.1.7 (originally discovered in the United Kingdom), B.1.351 (originally discovered in South Africa), and P.1 / B.1.1.28.1 (originally discovered in Brazil / Japan) [29,30,31]. The N501Y spike mutation (common to all three variants) has 4-10x increased affinity to the human ACE2 receptor for SARS-CoV-2 [32], which is hypothesized to contribute to the observed increase in transmissibility of B.1.1.7. The B.1.351 and P.1 variants have additional receptor binding domain (RBD) mutations such as E484K that show significantly reduced neutralization by antibodies generated by current vaccines and by prior natural infection presumably from non-variant SARS-CoV-2 strains [31,33,34,35,36]. These variants also have additional mutations in the spike N-terminal domain and appear to be resistant to several mAbs targeting that region [31].

These are all examples of highly relevant nucleotide-level information that a sequence-specific POC NAT could provide, where other POC assays such as a rapid antigen test might fail. There is therefore great potential to improve outcomes through increased multiplex and SNP-level sequence discrimination capability in POC NAT technologies.

However, very few POC tests exist for detecting drug-resistance SNP's and none for distinguishing viral strains (either for SARS-Cov2, or otherwise). Presently for HIV, HCV, and malaria there is no such commercially available POC drug resistance test, while for TB one test exists (GeneXpert MTB/RIF), but it can only detect resistance to a single drug out of the many for which resistance has been observed [20]. Research shows that for accurate drug resistance diagnoses in TB, screening of up to dozens of SNP's is necessary [18,19,21].

There is hence a need for broad POC screens of emerging and drug-resistant pathogens, able to distinguish viral species and strains, and able to detect drug-resistance mutations, in order to help monitor rapid outbreaks as they occur. Equally important, this can help determine proper treatment for patients.

1.5 Concluding Thoughts on the Limitations of NAT Testing

Significant progress has been made towards reducing the cost and equipment burden of NAT technologies without compromising their sensitivity. In many cases this has enabled their deployment further out from centralized hospitals to more local clinics and doctor's offices. However, the potential for further improvement and expansion into as-yet untapped markets is great, with at-home consumer NAT diagnostics and passive forms of NAT testing such as wearable and wireless/digital assays being key examples.

Chapter 2

2.1 Background and Motivation for Synthetic Biology Diagnostics

In response to the need for new classes of diagnostics, the Collins lab has been involved in the development of two new NAT technologies that leverage advances in the modularity, specificity, and sensitivity of technologies available in the synthetic biology toolkit: the toehold switch [38,39,40], and SHERLOCK [37,42,43,44]. These technologies are examples of cell-free synthetic biology. In the last two decades a plethora of genetic actuators, logic gates, and biosensors have been designed to modularly control the cellular environment of bacterial, fungal, mammalian, and other organisms.
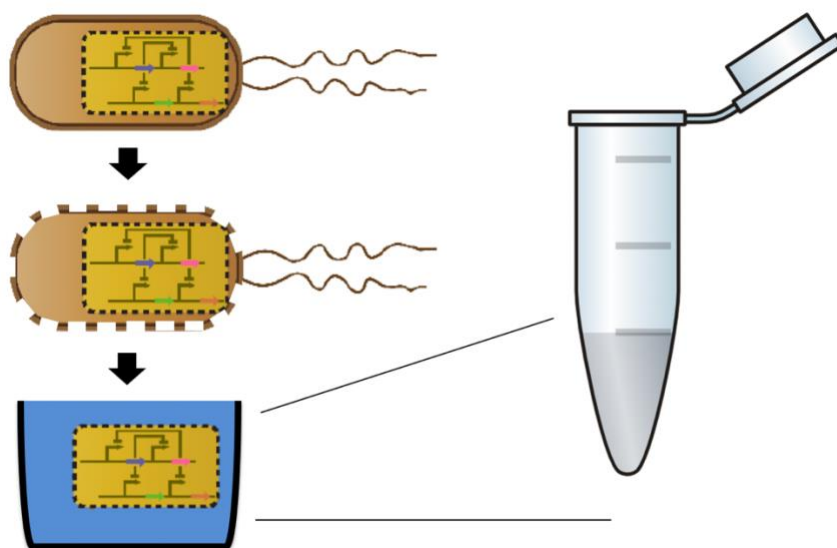
Figure 2-1. Adaptation of cellular synthetic biology to a cell-free environment. One method of reproducing synthetic biology tools that rely on both transcription and translation is to lyse the organism that originally hosted the synthetic biology tool, and to suspend the lysate in an energy buffer containing all the necessary small molecule components for desired reactions to proceed (transcription, translation, or other enzymatic reactions).

These efforts have been conceptually analogous to the control and actuation of electronic circuits. However, in order to expand their use, many of these synthetic biology components have been engineered for function outside of the cellular environment. Such cell-free reactions are self-contained, abiotic, aqueous chemical systems. They can be as simple as one or more purified enzymes interacting, or as complex as a cell-free protein synthesis system (CFPS) which contains all the biomolecular components required for efficient transcription and translation (Figure 2-1). Such systems can also be freeze-dried into shelf-stable formats in porous substrates, which allow for robust distribution, storage and use without specialized environmental requirements [38,40,41].

The toehold switch and SHERLOCK are both examples of cell-free synthetic biology reactions that are highly amenable to being freeze-dried [37,40]. As NAT diagnostics, they both build on existing isothermal amplification techniques, improving their sensitivity and specificity as a post-amplification amplicon-detection mechanism, whereby the amplified DNA or RNA is positively identified with a sequence-specific signal generation mechanism (to distinguish it from nonspecific amplification products). These approaches allow for more robust detection, but due to the reliance on isothermal pre-amplification, still suffer from said technology's limitations. I will describe them both, and then describe the design of a novel Cas13-activated riboswitch that combines elements of both in an attempt to circumvent the need for isothermal amplification.

2.2 The Toehold Switch: Riboregulator-Based RNA Detection

The first Collins lab diagnostic tool taking advantage of advances in synthetic biology that I will describe is the toehold switch (THS), a class of RNA responsive riboregulators that can be used in cell-free protein synthesis systems (CFPS) (Figure 2-2) [38,39,40].
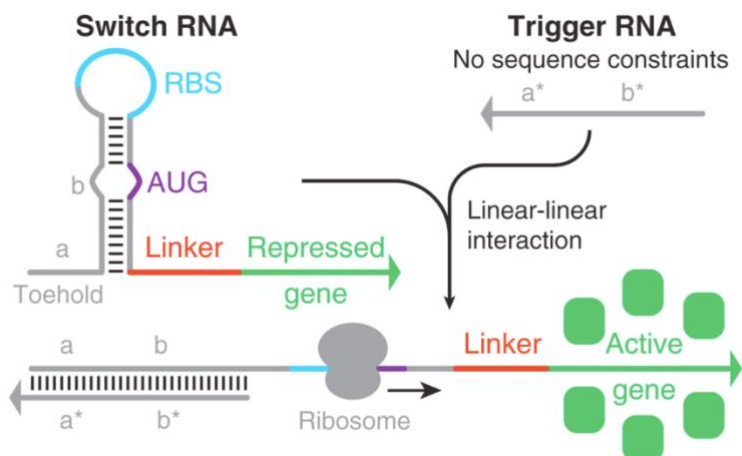
Figure 2-2. Illustrating toehold switch architecture. In its off state (top left), the closed switch RNA has its translation repressed by the presence of a hairpin stem, sterically hindering RBS access. Any trigger RNA (top right) to be detected can be specified by making the switch RNA complementary in regions a and b, to a* and b*. In the presence of trigger RNA, the switch RNA undergoes toehold-mediated strand displacement, making it competent for translation (bottom). Figure adapted from Green et al [39].

These mRNA switches repress translation by the e. coli ribosome in their off state by virtue of a hairpin secondary structure that reduces access to the Shine-Delgarno (SD) sequence of the ribosome binding site (RBS). However in the presence of a fully programmable activator RNA ('trigger RNA'), the hairpin stem is melted by toehold-mediated strand displacement, allowing for the SD sequence to be accessed and resulting in a more translationally active mRNA (Figure 2-2). The end effect is that specific RNA sequences of 30-40nt can be detected, with a programmable protein being produced as the sensor output. There are a number of advantages to the THS. The most notable advantage is its incredibly low cost: when produced at scale, the main reagent CFE can cost as little as a single cent per reaction, which allows the price of reagents to be essentially negligible [41]. Another essential advantage is the robustness to temperature variance, with the expression of many proteins varying little across a wide range of ambient temperatures 16C-37C

(based on our own unpublished data, and in particular for luciferases as luminescent output rather than fluorescent proteins). These properties are notably lacking for competing isothermal amplification technologies and allow the THS to operate without equipment and at low cost.

Unfortunately however, the THS suffers from one key and intractable limitation, which is its poor sensitivity. A typical LOD for Toehold Switches is in the low nanomolar range [38,39,40], which makes them entirely unsuitable for direct clinical use. Published and unpublished work in the Collins lab has attempted to ameliorate this, but the only successful approach has been to use an isothermal NASBA or RPA pre-amplification step [38]. Such a solution abolishes the THS's main advantages and replaces them with isothermal amplification's cost and temperature control limitations, relegating the THS to an amplicon detection role. Although the use of THS in concert with isothermal amplification can improve the sensitivity and specificity of the assay compared to isothermal amplification alone (by discriminating non-specific byproducts from true amplicon), it also adds an extra step to the diagnostic workflow because isothermal amplification methods tend to be incompatible with the buffer and enzyme mixes that most CFPS systems contain, requiring a two-pot assay.

2.3 SHERLOCK: CRISPR-Based RNA and DNA Detection

The second notable synthetic biology diagnostic tool that the Collins lab has been involved in developing in collaboration with the Zhang lab, is SHERLOCK (specific high-sensitivity enzymatic reporter unlocking) [37]. This platform leverages the versatile function of a class of Type-VI CRISPR-Cas effectors called Cas13 (also previously known as C2C2) [43,44]. Like many CRISPR-Cas effectors, these enzymes can be programmed with an RNA guide known as the crRNA. For Cas13, the crRNA targets the enzyme-guide complex to bind to a complementary

ssRNA target. Upon binding to the correct complementary sequence, the enzyme activates a nonspecific collateral RNAse activity that degrades any RNAs in the surrounding solution [37,42,43,44] (Figure 2-3). This function can be leveraged to detect specific ssRNA targets by including fluoresent reporters: these molecules consist of a fluorophore and quencher chemically synthesized at either end of an unstructured ssRNA molecule. When Cas13 encounters an RNA target previously specified by the complexed crRNA, the binding event can be detected by the cleavage of the fluorescent reporter, which produces a fluorescent signal when the quencher is physically separated from the fluorophore.
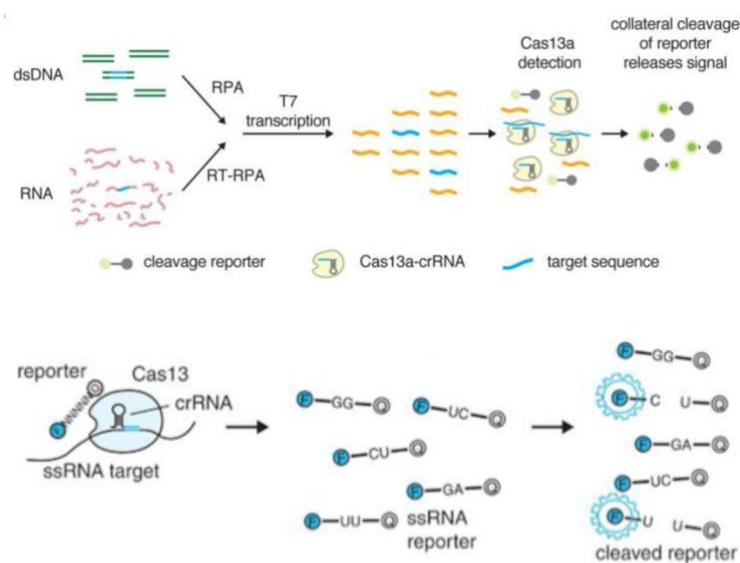
Figure 2-3. Workflow of a SHERLOCK assay. A) First an RNA or DNA target is amplified using RT-RPA or RPA, an isothermal amplification technique. The amplicon is then transcribed using a T7 polymerase, and finally the transcribed RNA is detected by Cas13 cleavage assay. The cleavage assay proceeds as follows: B) The Cas13a enzyme-crRNA complex identifies an ssRNA target complementary to the crRNA, and upon binding, a nonspecific collateral RNAse activity is generated, resulting in a quenched-fluorophore reporter being cleaved and generating signal. Figure adapted from Gootenberg et al. [42]

Alone, most CRISPR-Cas13 orthologs show only modest sensitivity for detecting target RNA's (at best reaching low-pM concentrations for the ortholog from *Leptotrichia wadei* bacterium LwaCas13a), and so the enzymes were combined with an RPA pre-amplification step [37,42]. The combination of an RPA amplification step, followed by T7 transcription of the resulting amplicon, and finally followed by Cas13 enzyme detection of the transcribed RNA, has been dubbed SHERLOCK (Figure 2-3). The platform has been expanded with up to four orthogonal Cas13 orthologs allowing for multiplexing (LwaCas13a, CcaCas13b, LbaCas13a, and PsmCas13b) [42]. In addition a similar and in some ways orthogonal system called DETECTR operates almost identically but leverages a different class of gRNA-guided CRISPR enzymes called Cas12, which bind to and recognize dsDNA and occasionally ssDNA, and produce ssDNAse cleavage activity in response to target activation [45].

The main benefit of SHERLOCK/DETECTR is unparalleled accuracy and sensitivity, matching and often surpassing the previous gold-standard qPCR [37,45]. These methods can achieve single-copy detection in volumes up to hundreds of microliters, and with single nucleotide discrimination capabilities [37,42,45]. In this fashion it is superior to most isothermal amplification schemes that operate on their own without an amplicon-detection step.

Unfortunately, because SHERLOCK uses RPA isothermal amplification, it still suffers from the same limitations that isothermal amplification methods do. SHERLOCK is as expensive if not more so than RPA by itself, and has the same temperature dependency issues as RPA. While by far one of the best candidates for a general purpose "next generation" NAT (in chapter three we use it in limited capacity for a wearable nucleic acid sensor), it still can only perform optimally and robustly with some form of temperature-control equipment. In summary the ideal confluence

of ambient-temperature robustness for equipment-free use, high sensitivity, and low cost are still not fully realized with SHERLOCK.

2.4 Leaderless Riboswitch: CRISPR-Cas13-Activated Translational Regulator

Interestingly, Cas13 alone (without using pre-amplification) nearly achieves the main aims of a "next generation" POC NAT diagnostic, its main limitation being sensitivity. LwaCas13a is a Cas13 ortholog that demonstrates some of the best on/off behavior compared to other known Cas13 orthologs [37,43,44], and it demonstrates other desirable properties such as a well-characterized nucleotide cutting preference [42]. Without pre-amplification with RPA, the enzyme demonstrates detection sensitivity in the low pM range when using a visual lateral-flow based output [37]. This is insufficient to be considered gold-standard, but is still significantly better than the sensitivity of toehold switches by at least 1000-fold. Its function is also relatively invariant across a range of ambient temperatures from 20C-37C (unpublished data), and the materials and manufacturing cost to produce the enzyme itself and its crRNA guide could be quite low at scale [41].

One path to a "next generation" diagnostic would hence be to simply improve the sensitivity of the LwaCas13a enzyme. The approach described in this chapter that I took to achieve this goal was to create a reporter for LwaCas13a that, rather than producing a fluorescent output based on a fluorophore-quencher system, would generate a protein expression signal analogous to the output of a toehold switch. Due to the high dynamic range that can be achieved by modifying ribosome translation initiation rates with riboregulators [46,51,52,53] compared with the dynamic range of fluorophore-quencher reporters, I hypothesized that such a reporter would improve the sensitivity of LwaCas13a detection. This could occur without any compromise in temperature-invariance, and could actually result in a reduction in cost. A cell-free riboswitch reporter would

also present a number of other advantages over fluorescence-based reporters, including improved robustness against contaminating RNAses in samples, and a longer shelf-life due to being storable in a DNA-encoded format.

A cell-free expression reporter system made to interact with LwaCas13a's RNAse activity would need to generate a significant change in translation initiation rate based on a single cleavage event. Based on this criterion, designs similar to the THS were discarded as they relied mainly on changes in secondary structure, not primary structure [39,51,52,53]. Instead a novel class of mRNA's was investigated, known as Leaderless mRNA's (LL-mRNA). These mRNA's use a unique and relatively understudied pathway for translation initiation that does not rely on a Shine-Delgarno or other ribosomal recognition sequence. Instead these mRNA's have only an AUG start codon at their 5' end, with no upstream sequence attached [47,48,49,50]. After screening many LL-mRNA leader sequences (Figure 2-4), a particularly active leaderless sequence from the naturally leaderless cI gene of lambda bacteriophage was used [50], with the first 20-36nt of the gene placed in front of an open reading frame (ORF) to generate leaderless translation in e. coli cell-free protein expression systems (with or without an amino acid linker as needed).

Expression from LL-mRNA is strongly inhibited by both secondary structure, as well as additional sequence added upstream of the start codon [48,49,50]. As a result, an exceptionally tight Riboswitch can be designed that relies on the removal of a sequence 5' to the start codon which blocks the start codon with secondary structure (Figure 2-5). This data represents to my knowledge the first instance of the use of LL-mRNA for bioengineering purposes. The resulting difference in expression of the repressed and unrepressed LL-mRNA has an exceptionally wide dynamic range of up to 10^4, likely due to the lack of an SD sequence, which is known to assist the ribosome in the unwinding of mRNA secondary structure [46,51,52,53]. The difference in expression between the

on and off states of the LL-Riboswitch approaches the biological limit for the e. coli ribosome [46], and to my knowledge, this potential dynamic range is greater than that of most published riboregulators (Figure 2-5).
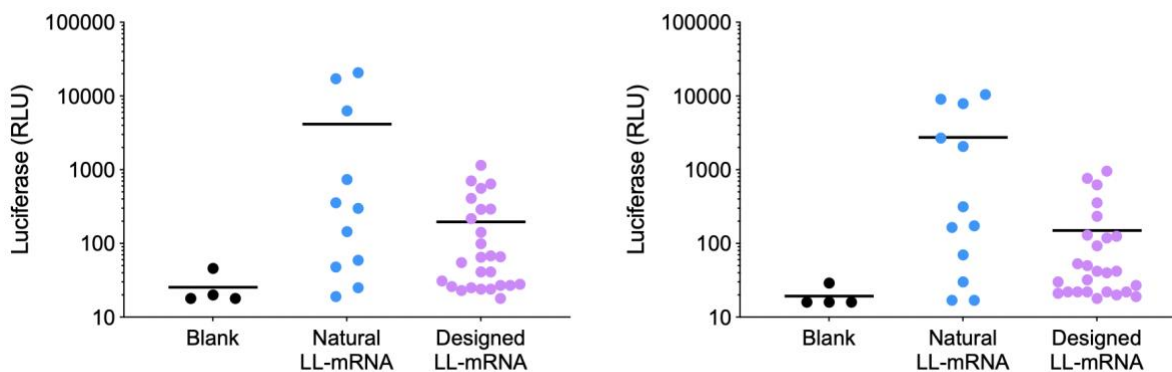

Figure 2-4. Panel of Leaderless mRNA Candidates. A series of 48 leader sequences were cloned upstream of a NanoLuc gene and allowed to express at 23C (left) or 37C (right) in a cell-free protein expression system. Of these sequences, 12 were naturally occurring LL-mRNA's, and 36 were artificially designed with low secondary structure. Mean of two replicates (n=2).
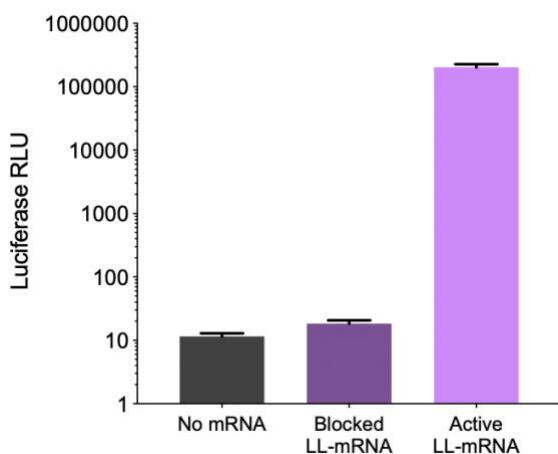

Figure 2-5. Maximum repression of LL-mRNA. A LL-mRNA consisting of a cI leader upstream of a Nano-Luciferase gene was expressed at room temperature in a cell-free protein expression system. Blocked LL-mRNA had an inhibitory secondary structure upstream of the cI leader. Mean and SD of two replicates (n=2).
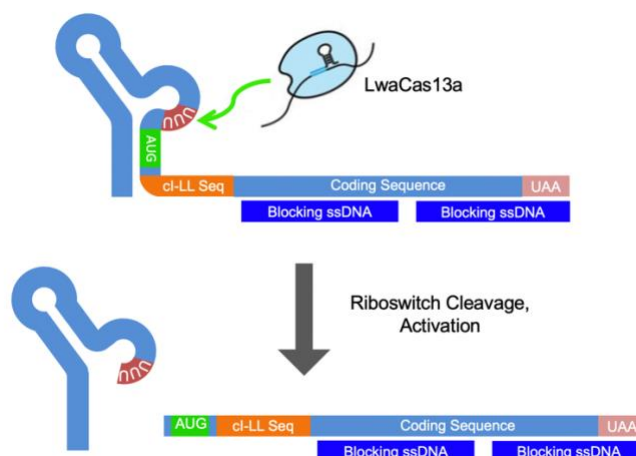
Figure 2-6. Leaderless Riboswitch mechanism of activation. The final LL-Riboswitch design consists of a cI leader sequence upstream of a coding sequence (unless specified, the P86 Nano-BiT split-luciferase peptide), with inhibitory secondary structures blocking ribosome access to the start codon. RNAse cleavage activity of LwaCas13a is directed to a location immediately 5' to the start codon by the presence of a series of U's, and blocked elsewhere by tiled ssDNA oligos.

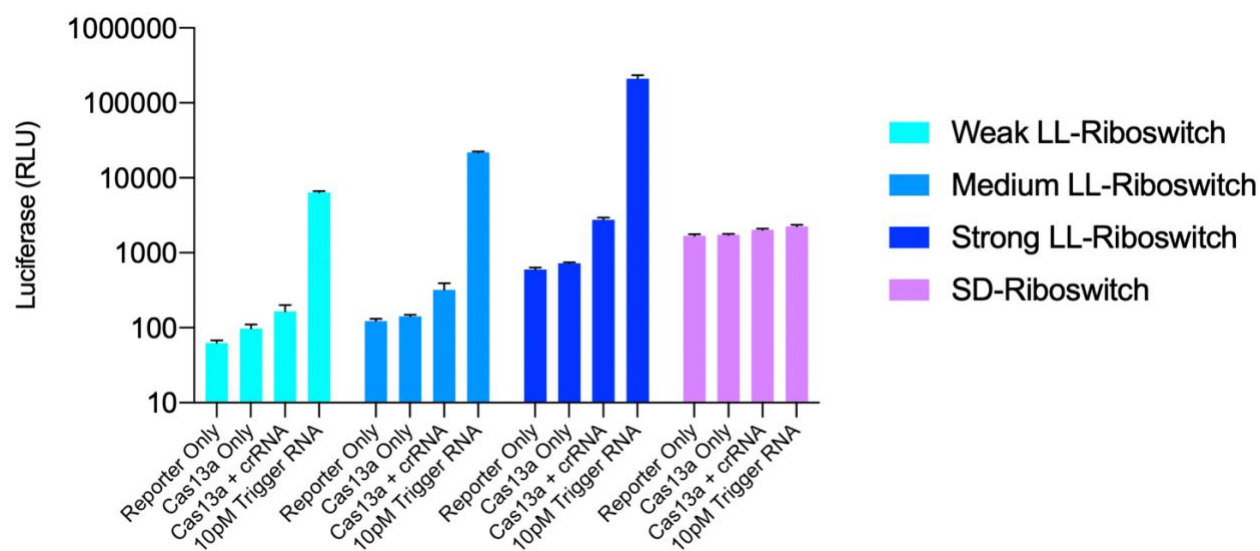**Figure 2-7. Maximum repression of LL-mRNA.** Three Leaderless Riboswitches with tunable levels of protein expression (weak, medium, and strong) were compared with a riboswitch that utilized a Shine-Delgarno leader (SD-Riboswitch). Their expression at 23C in a CFPS was measured alone, with Cas13a, with Cas13a and a crRNA guide, or with Cas13a-crRNA and 10pM of complementary RNA trigger. Mean and SD of two replicates (n=2).

In order to direct LwaCas13a to activate this LL-Riboswitch, the two main constraints on its RNAse activity were leveraged: 1) LwaCas13a can only cleave at U residues, and 2) LwaCas13a can only cleave ssRNA and not dsRNA. All regions on the repressed LL-Riboswitch were hence protected by secondary structure in the form of tiled ssDNA oligomers or hairpin structures, save for a target loop immediately 5' to the AUG start codon and the cI leaderless sequence. To protect the single-stranded cI leader sequence, all U's were removed from it, while a range of U residues were placed in the target loop to facilitate cleavage at that location (Figure 2-6). The resulting LL-mRNA is readily activated by LwaCas13a's RNAse activity, with tunable levels of leakage and signal based on the strength of repressive secondary structures (Figure 2-7). The length and position of the repressive hairpin, and number of U residues upstream of the start codon were all optimized for maximal fold-change, with 5xU being optimal (data not shown).

Interestingly, when the same style of Riboswitch is attempted with an SD-driven mRNA (with both the SD and AUG codon repressed by the hairpin secondary structure), only very little induction can be achieved (Figure 2-7). This is likely due to the higher leakage in the inactivated state that occurs when the ribosome can bind an SD sequence, and the larger secondary structure that the ribosome must overcome in the activated state when both the start codon and SD are blocked, as opposed to only the start codon.

As a reporter, the LL-Riboswitch signal allows for between 10x and 100x improved sensitivity over fluorescent reporters depending on the concentration of background RNA (Figure 2-8). The final limit of detection using this system is in the mid-fM range when expressing a short split-luciferase peptide as the output gene (from the NanoBiT system, peptide P86 complementing with pre-expressed 11S NanoLuc)[54]. While a plate reader instrument was used for these experiments, a low-cost colorimetric output solution could be to capture the luciferase's

chemiluminescent light on instant film similar to that used in kodak cameras. This passive readout is possible since no excitation light is required, as in the case of a fluorescence reporter. Importantly, the assay was carried out at room temperature.
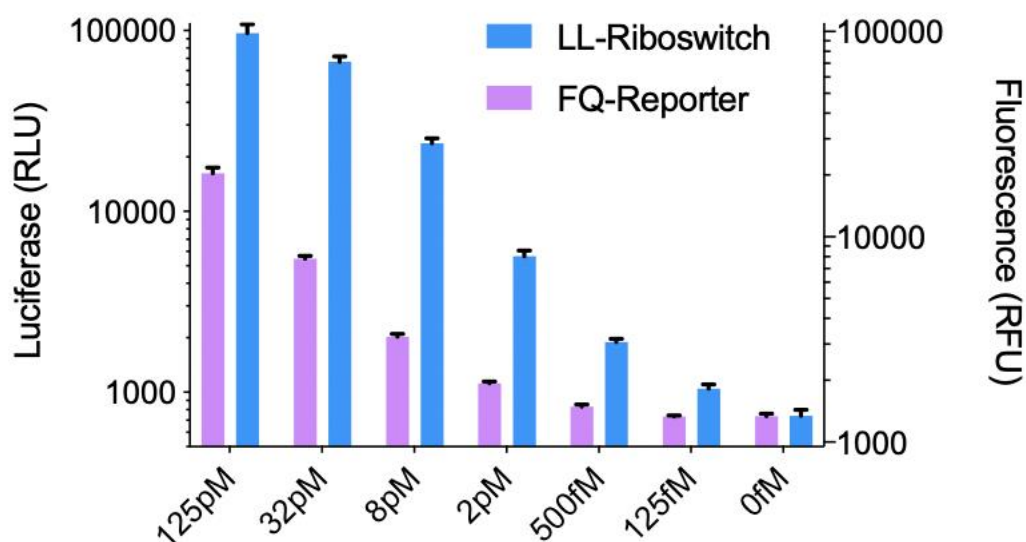

Figure 2-8. Cas13a cleavage signal comparison for LL-Riboswitch and fluorescence-quencher reporter. The collateral cleavage activity for LwaCas13a was measured when detecting varying concentrations of trigger RNA (125fM-125pM) using either an optimized LL-Riboswitch or a fluorescence-quencher (FQ) reporter. Cleavage was carried out for one hour. Mean and SD of three replicates (n=3).

2.5 Flexible Protein Outputs from a LL-Riboswitch Diagnostic

In addition to improving the sensitivity of RNA detection by Cas13, the use of LL-Riboswitches allows for increased flexibility in the output of the detection reaction. Whereas previously only behaviors induced by the physical separation of molecules joined by RNA could be induced (mostly fluorescence signal), with a LL-Riboswitch the bioactive enzymatic or structural behavior of an arbitrary protein can be leveraged as a diagnostic output. However, for

large ORFs, it became intractable to tile their entire length with ssDNA oligomers, as the efficiency

of translation was impeded and the effective protection from cleavage by Cas13 was diminished.

In my hands, only ORFs shorter than what can be covered by about three ssDNA oligomers were

inducible by Cas13a – if more than that many oligomers were required, then the RNAse activity

of Cas13a would destroy the ORF more rapidly than the ribosome could generate protein

expression from it (data not shown). Hence an intermediary solution was necessary if larger

proteins were to be expressed from LL-Riboswitches. To facilitate this process, a more orthogonal

system was created where rather than directly express a protein that would need to be tiled by

ssDNA oligomers, a smaller peptide was designed as an output that could induce transcription and

expression of the intended full-length output protein. This was achieved using a split-intein system,

with an output peptide engineered to drive the transcription of arbitrary ORFs by activating T7

RNA polymerase (RNAP). First, the C-terminal five amino acids of T7 RNAP were removed,

leaving a catalytically dead protein with fully abolished transcriptional activity. Second, a peptide

consisting of those missing five amino acids of T7 RNAP and the CfaGEP intein [55] was designed

that upon being expressed in the presence of the catalytically dead T7 RNAP, could covalently re-

attach the missing amino acids to T7 RNAP, and thereby fully restore transcriptional activity.

When used as the output of a LL-Riboswitch in the presence of the split/deficient T7 RNAP, this

short intein peptide could induce downstream transcription of a larger full-length gene such as

GFP or NanoLuciferase [73] (Figures 2-9, 2-10). Because unprotected mRNAs produced by T7

RNAP were being continually transcribed and replenished upon induction, protein expression from

this two-step system was much more robust to the RNAse activity of Cas13 than a one-step system

that directly expressed protein from the finite quantitiy of LL-Riboswitch.
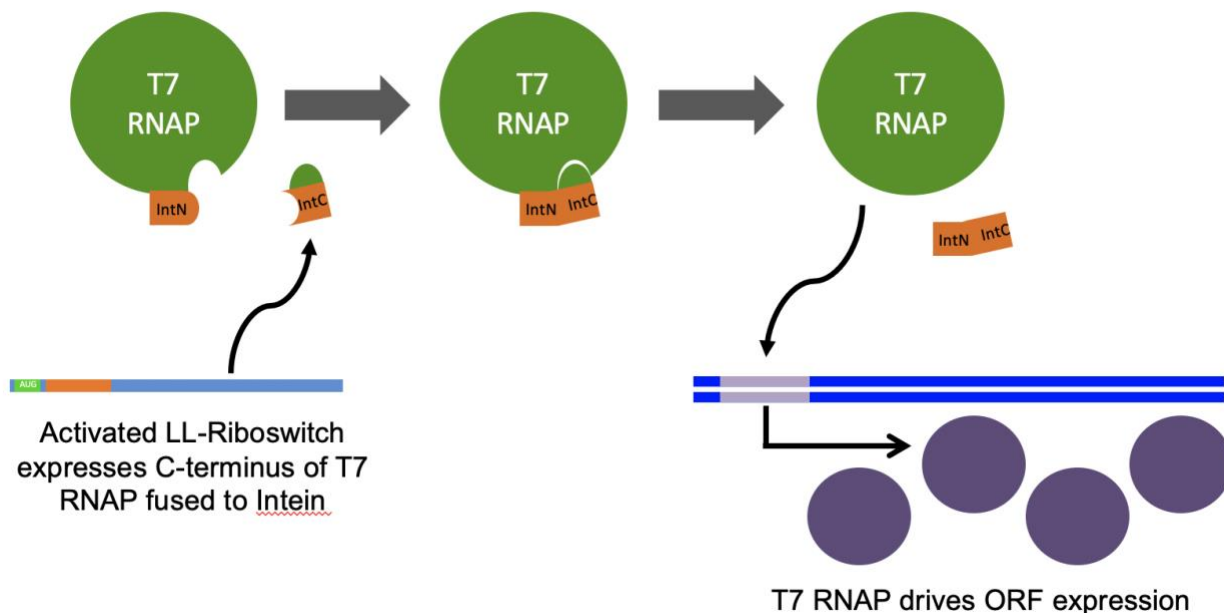
Figure 2-9. A split-intein T7 RNAP induction system for more general LL-Riboswitch expression. An engineered output peptide produced from a LL-Riboswitch induces transcription of an arbitrary ORF by 1) binding to a C-terminally deficient T7 RNAP, then 2) using a split-intein to splice five missing amino acid residues to its C-terminus, and 3) induce transcription of the desired ORF from a dsDNA template with an upstream T7 promoter.
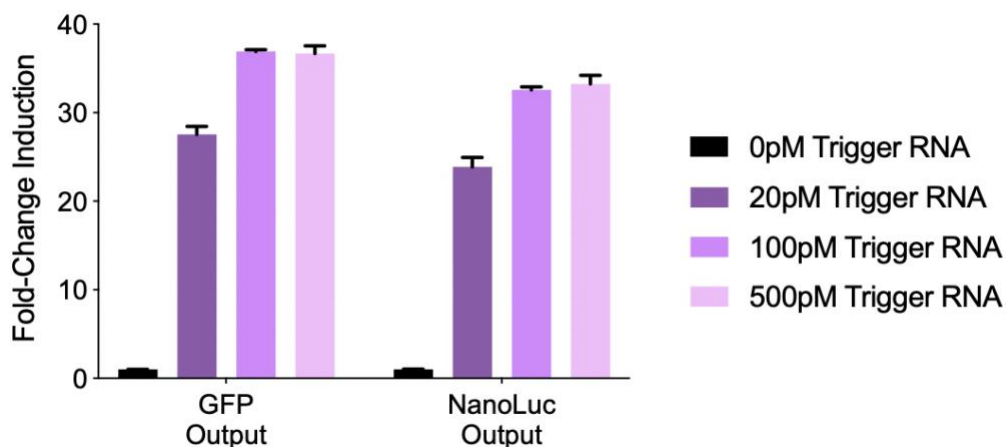

Figure 2-10. Cas13a cleavage signal comparison for LL-Riboswitch and fluorescence-quencher reporter. The collateral cleavage activity of LwaCas13a was used to drive expression of two full-length ORFs, superfolder-GFP and NanoLuciferase, using the split-intein LL-Riboswitch reporter system outlined in Figure 2-9. Mean and SD of two replicates (n=2).

2.6 Improved Limit of Detection for LL-Riboswitches with Gated-crRNA

While the limit of detection of the Cas13 LL-Riboswitch diagnostic thus far has been improved to the mid-fM range, in order to truly achieve useful detection performance the sensitivity will need to be further improved by at least an order of magnitude, if not more.

The greatest current limitation to the LL-Riboswitch diagnostic is the off-state untriggered RNAse activity generated by the LwaCas13a-crRNA complex itself. In the presence of crRNA guide but in the absence of any trigger RNA, LwaCas13a displays background RNAse activity that is equivalent to roughly 10-100x the background signal of the LL-Reporter alone (Figure 2-11). Hence any attempt to improve the sensitivity of the LL-Riboswitch diagnostic platform will need to first reduce the background produced by this leakage RNAse activity, until it is of the same level of signal as the LL-Riboswitch reporter on its own. This could be accomplished in a variety of ways, but the design path I chose attempted to maintain separation between the crRNA and the LwaCas13a enzyme except in the presence of the desired activator RNA. By preventing the LwaCas13a:crRNA complex from forming, the leakage produced by it should be reduced, but without affecting the on state signal as the presence of an activator RNA will enable the guide to complex with the LwaCas13a enzyme (Figure 2-12). It leverages a similar approach to strand-separation seen in molecular beacons.
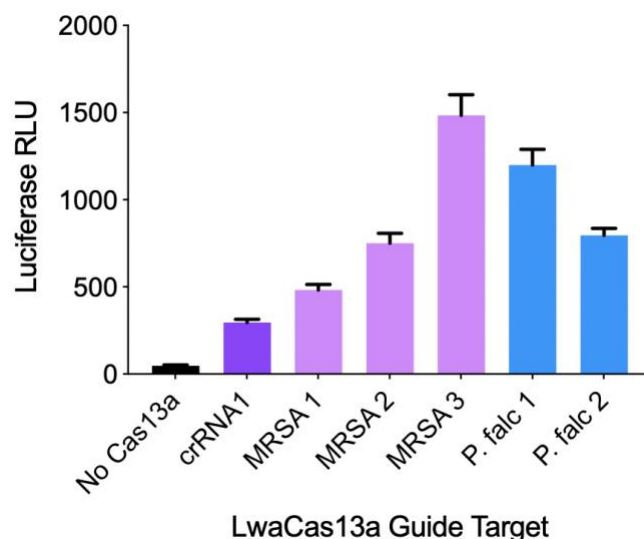
Figure 2-11. Background leakage of LwaCas13a complexed with different crRNA guides. The signal produced from cleavage of a LL-Riboswitch reporter by LwaCas13a in the absence of trigger RNA for one hour was measured at 23C. The different crRNAs tested targeted either a synthetic sequence (crRNA1, dark purple), a sequence from the mecA gene of Methicillin-resistant *Staphylococcus aureus* (MRSA, pink), or a sequence from *Plasmodium falciparum* (P. falc, blue). Mean and SD of three replicates (n=3).
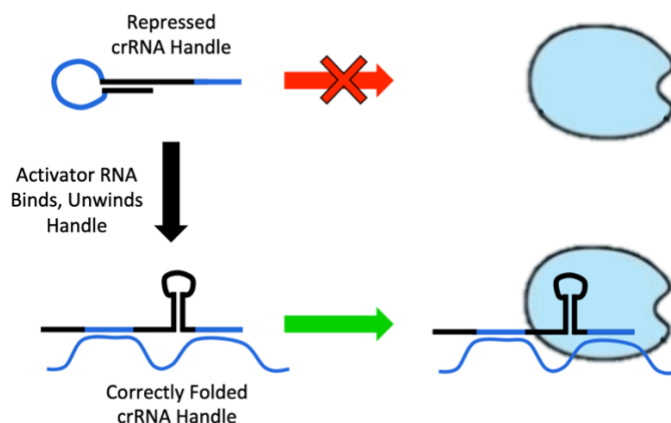
Figure 2-12. Architecture of a Gated-crRNA for Cas13a. In the repressed inactive state, a 5'-appended sequence complementary to the handle of the crRNA forms an inhibitory hairpin structure that prevents the crRNA from being recognized by Cas13a. However, in the presence of an activator (trigger) RNA with sequences complementary to both the crRNA spacer and the gate region, the crRNA handle is freed and the crRNA-trigger complex can bind to Cas13a.
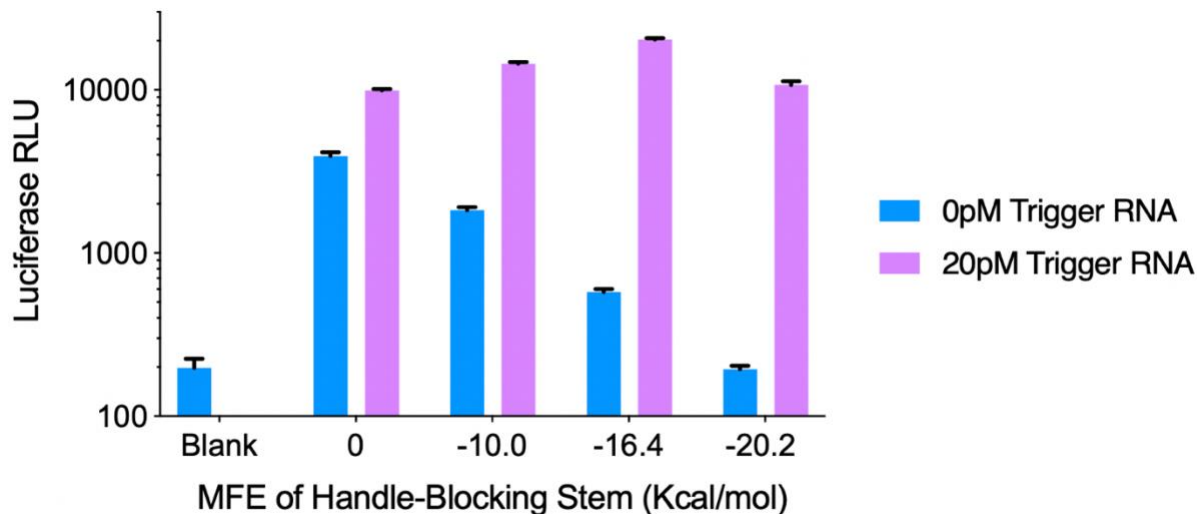
Figure 2-13. Background Cas13a RNAse activity is repressed by a gated handle. Signal from a Cas13a-induced LL-Riboswitch in the presence of 0pM or 20pM of a Trigger RNA is shown using four gated crRNA with repressive hairpins of varying strength (MFE values calculated by NUPACK). The trigger RNA sequence was designed to bind to the crRNA spacer and sterically unwind the gate region. Mean and SD of two replicates (n=2).

A number of preliminary attempts to prevent the LwaCas13a:crRNA complex from forming relied on designing secondary structures in the crRNA guide itself, including blocking the guide handle with complementary ssDNA oligos or extensions of the guide from the 5' and 3' directions. These could then hypothetically be melted after binding of the trigger RNA to the spacer region of the crRNA, by binding of a second region on the trigger RNA to a "gate" region, allowing for trigger-induced on-state RNAse signal to be unaffected while the off-state RNAse signal is eliminated.

Of the attempted designs, only one succeeded and is shown here. The most effective "gating" process functioned conceptually like a molecular beacon, where the blocking hairpin stem is unwound by steric hindrance rather than toehold-mediated strand displacement. While not as

efficient a method of changing RNA secondary structure conformation, it allows for the activator

RNA sequence to be fully unconstrained.

This modified crRNA handle significantly repressed background leakage, and the presence

of ssRNA and/or dsRNA at the 5' end of the crRNA handle did not effect on-state RNAse activity

(Figure 2-13). Furthermore, the on-state activity with activator present was reduced by the need to

de-repress the crRNA guide to a lesser extent than the degree of background reduction, improving

the overall on/off behavior. This resulted in a subsequent improvement in limit of detection of five

to ten-fold over the standard crRNA design for the same spacer sequence (Figure 2-8 and Figure
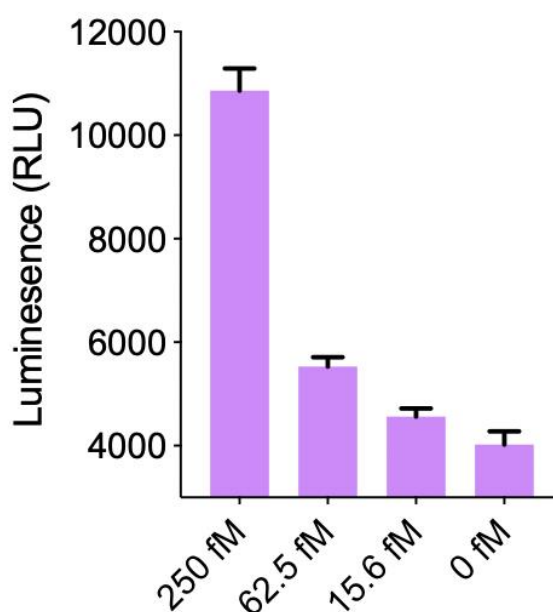
2-14).

Figure 2-14. Limit of detection for a crRNA design with a gated handle. An LOD curve for an optimized gated-crRNA architecture for LwaCas13a is shown against a trigger RNA that both unwinds the gate region and binds the spacer of the crRNA. An optimized LL-Riboswitch was used to generate signal. Mean and SD of three replicates (n=3).

Unfortunately, while the design of a two-trigger "gated" crRNA was largely successful and can serve to provide greater specificity for larger RNA targets, the improvement in sensitivity was not sufficient for the LL-Riboswitch to reach a limit of detection that would make it suitable as a replacement for isothermal amplification based techniques like SHERLOCK.

2.7 Concluding thoughts on Cell-Free Expression Reporters for CRISPR-Cas13

Although the creation of a CRISPR-Cas13 actuatable riboswitch in conjunction with a gated Cas13 crRNA design improved the sensitivity of direct RNA detection using Cas13, the optimized limit of detection did not surpass the femtomolar range. For a gold-standard NAT diagnostic to have clinical utility, it should be able to robustly generate a high fold-change signal when detecting analytes well into the attomolar range [1,2,3,4,5,6]. Indeed, since the completion of this work, another group has published on efforts to directly detect RNA samples including SARS-Cov2 using the collateral activity of Cas13 (notably improving the sensitivity by using LbuCas13 instead of LwaCas13) [56]. However, that work did not robustly achieve attomolar limits of detection and relied on relatively small changes in fluorescent signal over time using a form of fluorescence measurement equipment. In addition, they required a temperature incubation at 37C to achieve optimal Cas13 cleavage activity. The necessity to track kinetic behavior of a fluorescence signal, and the requirement of a 37C incubation step, obviates the advantage of using Cas13 alone in comparison to approaches melding Cas13 with an isothermal amplification tool such as SHERLOCK.

In the end, a different approach to improving the limitations inherent in isothermal amplification had to be taken, other than to try to circumvent it entirely. Rather than improve the sensitivity of a synthetic biology tool that operated well at low temperatures but was insufficiently

sensitive, in Chapter Five I describe an effort to improve the low-temperature robustness of an existing isothermal amplification method that was already sufficiently sensitive at higher temperatures. This proved to be a more amenable design challenge and brought us much closer to an ideal POC NAT technology suitable for next-generation applications. The old maxim held true that, "if you can't beat them, join them".

Chapter 3

3.1 Deploying freeze-dried synthetic biology diagnostics in wearable format

While the diversity of modular biosensors, genetic logic gates, and output effectors in the synthetic biology design toolkit have been increasingly validated in benchtop assays for the purposes of NAT and other diagnostic assays[59,60], recent developments in wireless technology, wearable electronics, smart materials, and functional fibers with novel mechanical, electrical and optical properties have marked the dawn of next-generation biosensing systems[61]. To date, practical examples of devices that enable the wearable deployment of these tools are limited. For the most part only a few demonstrations of hygroscopically actuated vents and response signals to induction molecules have been achieved using living engineered bacteria encapsulated in flexible substrates and hydrogels (notably PDMS) in a wearable format (Figure 3-1) [58,62,63,65].
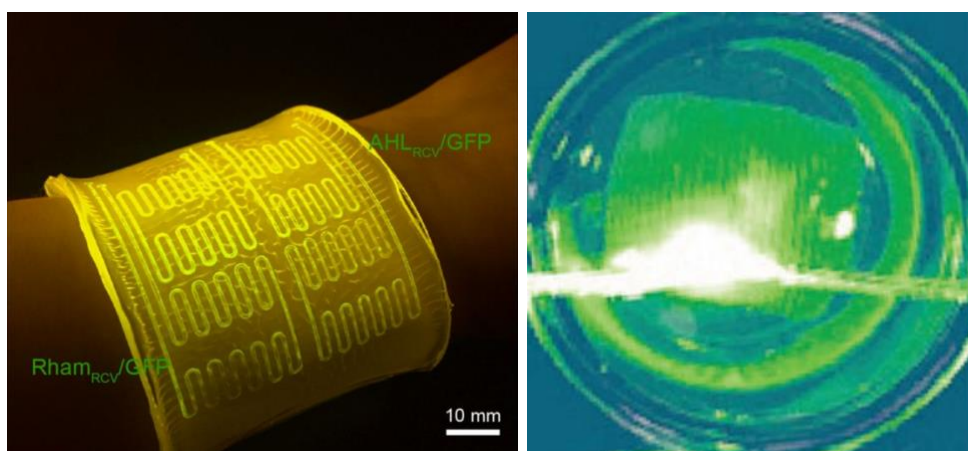
Figure 3-1. Examples of wearable synthetic biology using living cells. Two examples from the literature are shown for wearable deployment of living synthetic biology systems. On the left is a PDMS patch containing live *E. coli* cells able to produce GFP in response to small molecule inducers that have been smeared onto the skin [57]. On the right is a fabric that allows for live engineered *E. coli* to adhere in a biofilm only upon exposure to blue light, and are in this case also constitutively expressing GFP for visualization purposes [58].

While seminal, such examples encounter several limitations. In particular, sustaining living organisms within these devices for extended periods of time is an as-yet unsolved and seemingly intractable obstacle, as retaining the viability and function of wearable sensing systems based on living cells requires nutrient delivery, waste extraction, as well as temperature and gas regulation, all of which involve numerous technological apparatus. Genetically engineered cells can also pose biocontainment or biohazard concerns, particularly if integrated into consumer-level garments, leading to stringent regulatory pathways in many critical applications. Moreover, such genetically engineered cell populations will continually evolve and suffer mutational pressures over time, potentially resulting in loss of the genetic phenotype and function [63,64]. Thus, a new approach in synthetic biology is needed to resolve the mismatch between practical requirements of wearable synthetic biology use and the operational limitations of available biomolecular circuits for sensing and response, allowing their utilization ideally in a wide range of wearable substrates (e.g., functional fibers or fabrics) to assess molecular targets in a manner that is passive and fully field-deployable for consumers or professionals [65].
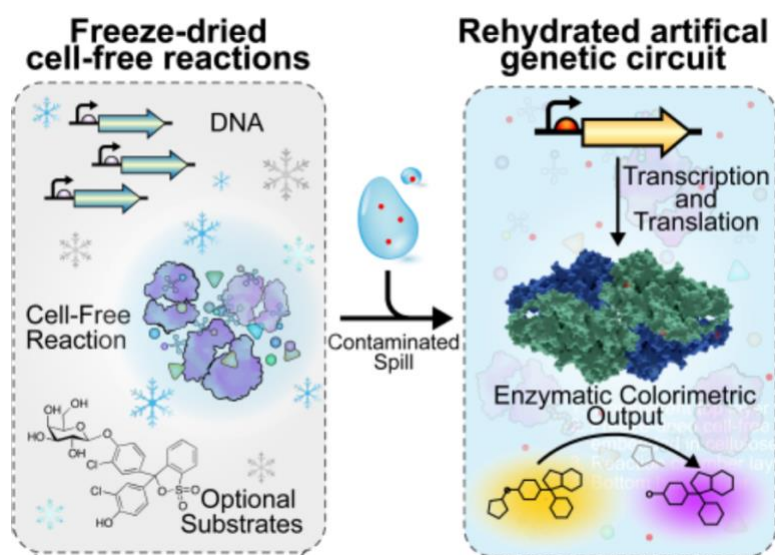

Figure 3-2. Freeze-dried cell-free synthetic biology. Schematic shows cell-free reaction components inertly stored at ambient temperature until rehydration with an analyte.

As mentioned at the start of chapter two, cell-free synthetic biology reactions are self-contained abiotic chemical systems that can carry out the function of a synthetic biology tool in the absence of a living cell, and can be complex enough to contain all the biomolecular components required for efficient transcription and translation (Figure 3-2). Genetically engineered circuits, encoded in DNA or RNA, can be added to such freeze-dryable, cell-free (FDCF) reactions for activation by simple rehydration. Robust FDCF systems have already been developed for inexpensive paper-based nucleic acid diagnostics using the toehold switch [39] and highly sensitive programmable CRISPR-based nucleic acid sensors using Cas13/Cas12 [37,42,45,66] (as described in chapter two), in addition to other functions such as on-demand production of antimicrobials, antibodies, and enzymes [67], and low-cost educational kits for teaching [68,69,70]. The ability for such systems to be freeze-dried into shelf-stable formats allows for robust distribution and storage in a way that circumvents the limitations of cellular synthetic biology systems [67].

Interestingly, we found that cell-free protein expression reactions (among other cell-free reaction types) were amenable to many different textile and paper substrates (Figure 3-3). Hence, we chose to use freeze-dried, cell-free genetic circuits in combination with specifically designed flexible and textile substrates as a new direction towards practical wearable biosensors. In this chapter I will describe the design and validation of a number of wearable freeze-dried, cell-free synthetic biology (wFDCF) sensors not just for NAT diagnostics, but also for small molecule and toxin detection. These sensors are integrated into flexible multi-material substrates (e.g., silicone elastomers and textiles) and use genetically engineered components including toehold switches, transcriptional factors, riboswitches, fluorescent aptamers, and SHERLOCK/DETECTR (CRISPR-Cas13a/Cas12a).
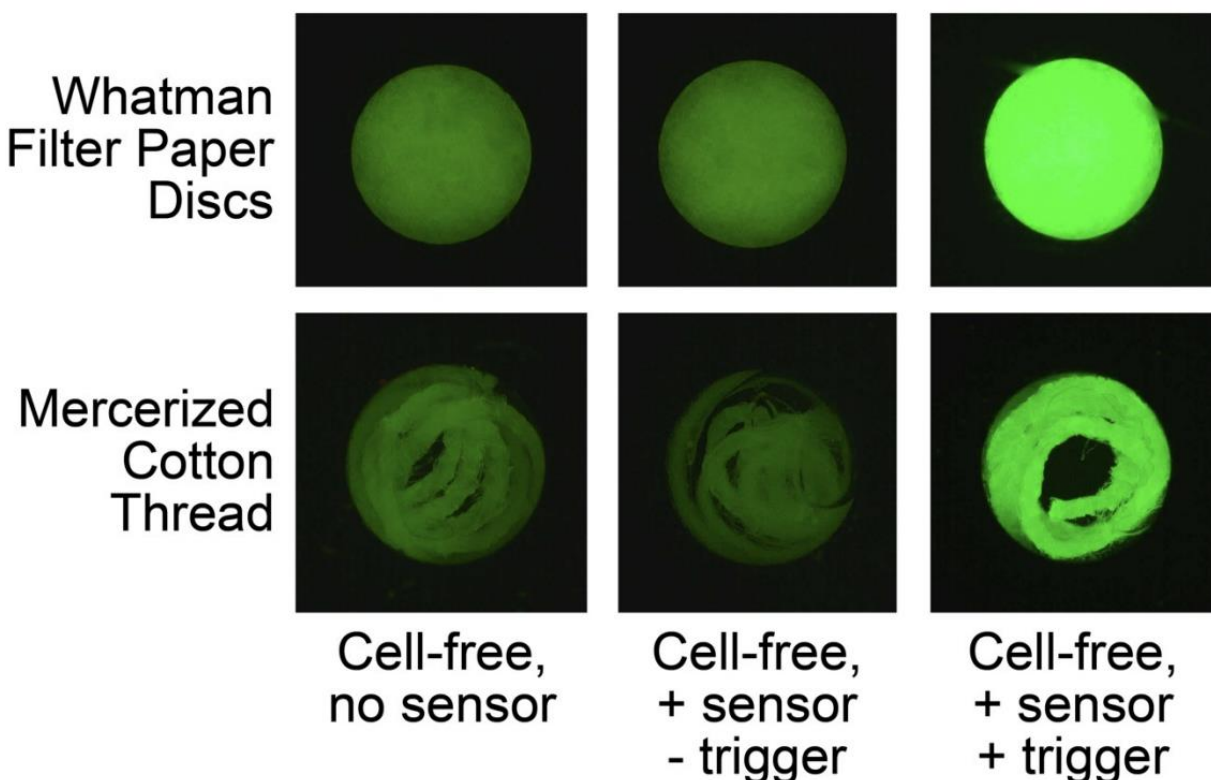
Figure 3-3. Textile and paper embedding of freeze dried cell-free protein synthesis. Figure shows fluorescence microscopy images of cell-free protein synthesis reactions expressing GFP from a toehold switch sensor, when wicked either into a filter paper disc or a cotton thread and then freeze-dried. In sensor-added conditions 33nM of Zika-targeting toehold DNA were added, and in trigger-added conditions, 12uM of Zika RNA was added.

3.2 Wearable Diagnostics using a Colorimetric Synthetic Biology Platform

For our first wFDCF demonstration, we embedded colorimetric genetic circuits into cellulose substrates surrounded by a fluid wicking and containment assembly made of flexible elastomers (Figure 3-4, top left). These prototypes were assembled layer-by-layer to form reaction chambers fluidically connected to top sample portals. The devices are flexible, elastic, and can rapidly wick in splashed fluids through capillary action (Figure 3-4, top right). Pinning geometries throughout the device direct sample fluids towards enclosed hydrophilic paper networks allowing

for reaction rehydration. Using a lacZ β-galactosidase operon as the circuit output to hydrolyze chlorophenol red-β-Dgalactopyranoside (CPRG), a yellow to purple color change develops upon exposure to a target (Figure 3-4, bottom).
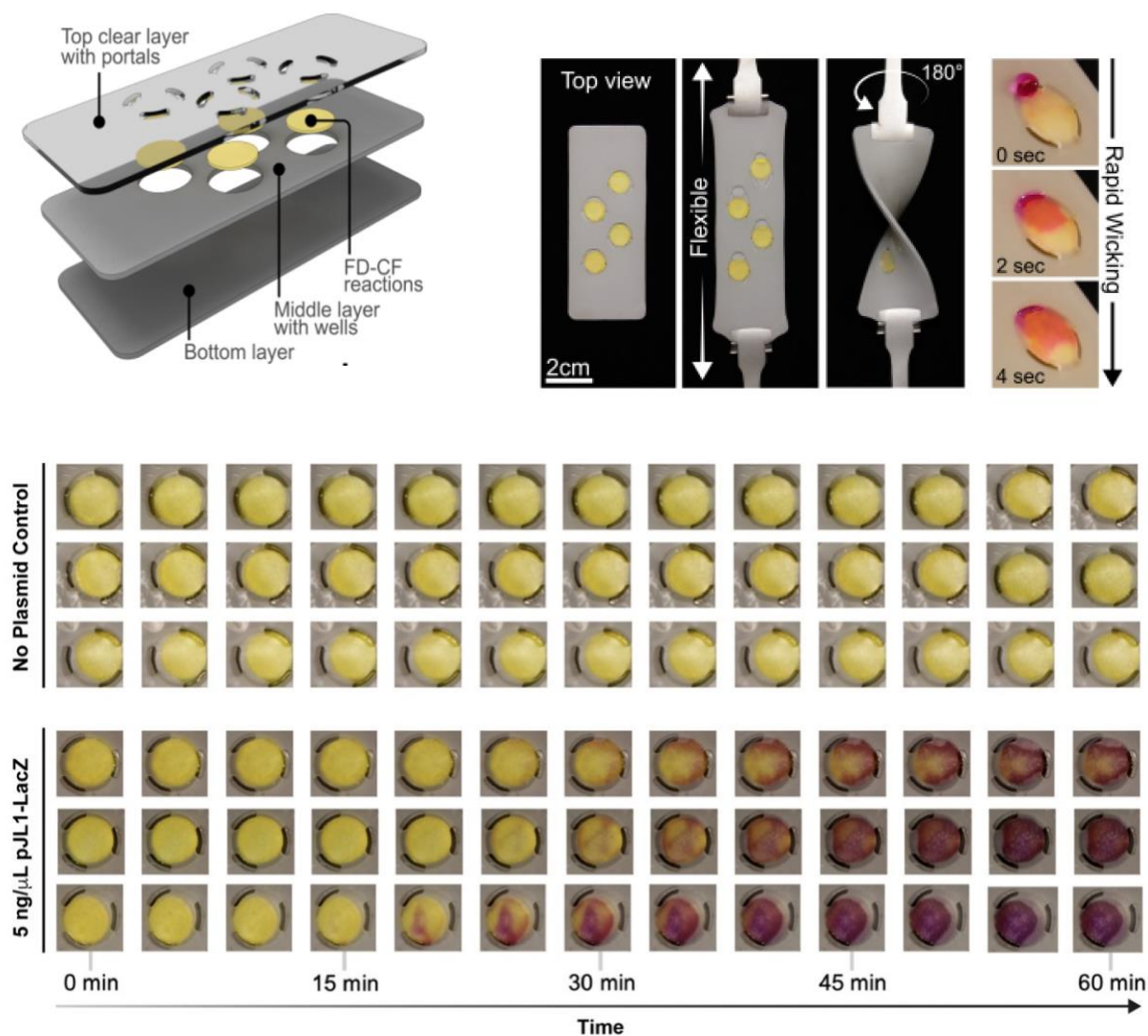
Figure 3-4. Wearable colorimetric reaction patch design. Top left: overview schematic of three-layered reaction patches with two opaque and one transparent silicone layer bonded to surround paper disks containing FDCF reactions. Top right: wFDCF reaction patches can resist axial and torsional strain, and allow for rapid passive wicking from environmental splashes through entry ports. Bottom: images were taken every five minutes of CFPS reactions in triplicate constitutively expressing LacZ at 30C. Signal was visible within 20-30 minutes.

Key environmental factors were considered for the design of these prototypes. For instance, sample exposure in the field could occur with variable splash volumes (we assumed a reasonable range to be between 50-100uL), relative humidity (RH, 20-40%), and temperature (20-37C). Thus, we optimized our design to reduce inhibition of genetic circuit operation caused either by evaporative water loss to the environment (resulting in reaction concentration), or by excessive water gain from large environmental sample splashes (resulting in reaction dilution). In particular, our devices use impermeable chambers exhibiting low evaporation rates (<20% volume/hour at room temperature), and which also constrain the rehydration volume to ~50 uL per sensor. In addition, the wFDCF reactions were optimized to generate a higher concentrated reaction upon rehydration. We found that cell-free reactions (in this case using the PURE system, obtained from NEB) when concentrated 1.5x by rehydration at a lower volume after freeze-drying, displayed increased reaction kinetics, enabling signal output at least 10 min faster and ensuring that the desired circuit is completed before eventual evaporation in the device terminates the reaction (Figure 3-5). The resulting stand-alone colorimetric system is modular and can be used in garments such as patches or bracelets.
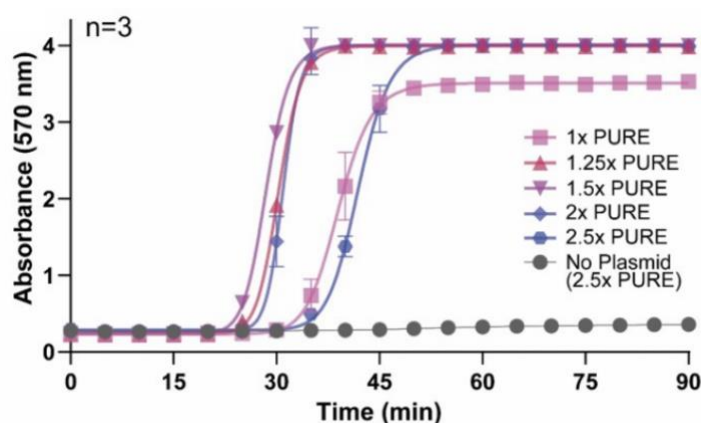

Figure 3-5. Cell-free protein synthesis reactions at varying rehydration volumes. Freeze-dried CFPS reactions containing constitutively expressed LacZ were rehydrated at various concentrations. Mean and SD of three replicates (n=3).

Functional testing of this colorimetric wearable platform was performed utilizing four different synthetic biology biosensors with lacZ as the output (Figure 3-6). These various demonstrations include a constitutive lacZ expression reaction (Figure 3-6, top left), a transcription factor-regulated circuit using the tetracycline repressor (TetR) (Figure 3-6, top right), a toehold switch for Ebola virus RNA detection (Figure 3-6, bottom left), and a theophylline riboswitch for small-molecule sensing (Figure 3-6, bottom right). Genetic circuits using transcriptional regulators are among some of the most common elements used in synthetic biology. The wFDCF TetR sensor demonstrates the capacity of the colorimetric platform for facile integration of well-established genetic modules into a wearable format. Furthermore, a functional theophylline riboswitch wFDCF circuit is functionally validated in our platform for the environmental detection of small molecules via engineered cis-regulated RNA circuits. This specific riboswitch was selected as a model test case, although a plethora of similar riboswitches for various targets have been reported and could be used in a modular fashion. All of the colorimetric wFDCF sensors reported here exhibited visible changes within ~40-60 min after exposure to the respective trigger molecules or inducer, and were performed at ambient conditions of 30-40% RH and 30ºC to simulate the average skin surface temperature[71].
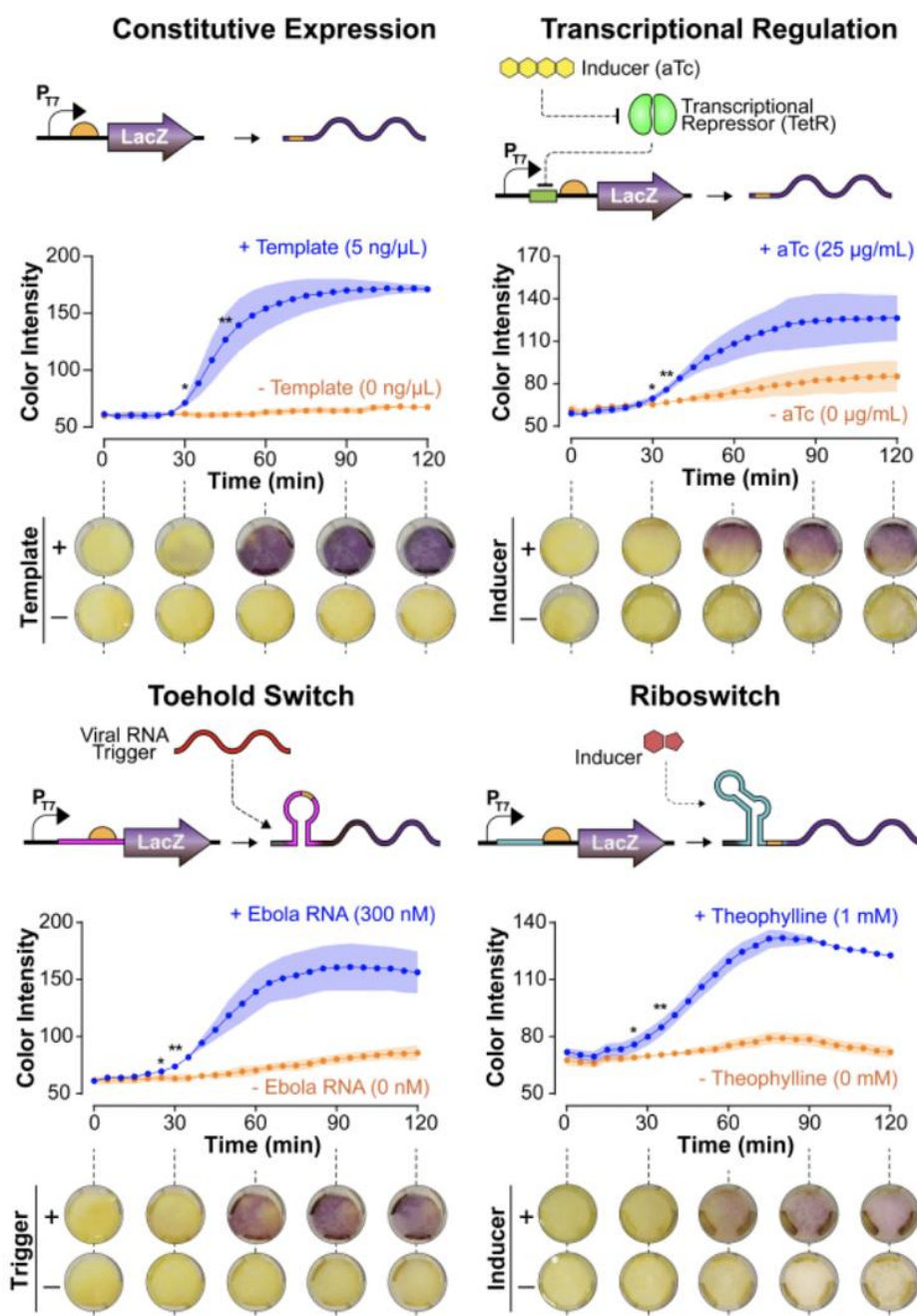
Figure 3-6. Colorimetric cell-free synthetic biology implementations in wearable patches. Various synthetic biology circuits were freeze-dried in wearable patches, including constitutively expressed outputs (top left), transcription factor-regulated circuits for small molecule detection (top right), toehold switches for nucleic acid-sensing (bottom left), and riboswitches to detect small molecules (bottom right). Each graph shows color deconvoluted values, n=3. Statistical significance is indicated for specific time points (*P ≤ 0.05 and **P ≤ 0.01). Bottom images are representative images.

Importantly for the purposes of developing this technology towards a wearable platform for NAT diagnostics, we show that a wFDCF Ebola virus RNA toehold switch sensor is capable of rapid and sensitive detection of an RNA sequence originating from Ebola virus. Though the sensitivity of the toehold switch makes it an unsuitable tool for detection of RNA analytes in true clinical samples (as previously explained), this example paves the way for the development of similar viral or bacterial wearable nucleic acid sensors which will be discussed further in this chapter. Similarly for these demonstrations a mild 30C temperature incubation from close contact with human skin was assumed to be present, and allowed for enhanced detection. Though non-ideal for reasons previously discussed in chapter one, in further iterations of our design the need for temperature incubation was reduced and will be covered further in this chapter.

3.3 Wearable Diagnostics using a Fluorescent Synthetic Biology Platform

Though our demonstrations with colorimetric wFDCF reaction patches showed great potential, our prototypes continued to have limited sensitivity for NAT diagnostic use, as well as a reliance on low-level temperature incubation at 30C. In order to expand on the attractiveness and versatility of textiles as ubiquitous wearable substrates, we attempted to broaden the signal types detectable by our prototypes to include those reliant on light emission (as opposed to light absorbance for colorimetric systems). Figure 3-7 shows a highly sensitive, textile based system capable of containing and monitoring the activation of wFDCF reactions with fluorescent or luminescent outputs. To achieve this, we fabricated a second wearable platform that integrates: (a) hydrophilic threads (85% polyester / 15% polyamide) for cell-free reagent immobilization, (b) patterns of skin-safe hydrophobic silicone elastomers for reaction containment, and (c) inter-weaved polymeric optic fibers (POFs) for signal interrogation (Figure 3-8). This fabric was chosen

as our main immobilization substrate after conducting a compatibility screening of over 100 textiles (e.g., silks, cotton, rayon, linen, hemp bamboo, wool, polyester, polyamide, nylon, and combination materials) using a lyophilized constitutive lacZ cell-free reaction, though we found that a variety of substrates were amenable to the cell-free reaction. The analysis of sensor outputs was done using a custom-built wearable POF spectrometer that allowed for monitoring with a mobile phone application (Figure 3-9). This is achieved by illuminating the wFDCF textile reaction with blue light (447 nm) via etched excitation POFs (Figure 3-8). The light emitted from the activated system is then collected by the second set of emission POFs (Figure 3-8), which exit the fabric weave and bundle into a connection to the optical sensor (Figure 3-7) of our wearable spectrometer (Figure 3-9). Signals coming from each of the devices are filtered and processed to generate temporally and spatially resolved fluorescence images of the POF bundle-ends (510 nm) and averaged pixel intensity traces per channel for quantitative analysis (Figure 3-9). In the case of luminescence demonstrations, all POFs bundles are treated as signal inputs, without the need for sample illumination. All reported wFDCF fluorescence and luminescence sensor replicates (n=3) exhibited visible fluorescence or luminescence within 5-20 min after exposure to relevant trigger conditions, at 30-40% RH and 30ºC (or lower if indicated).
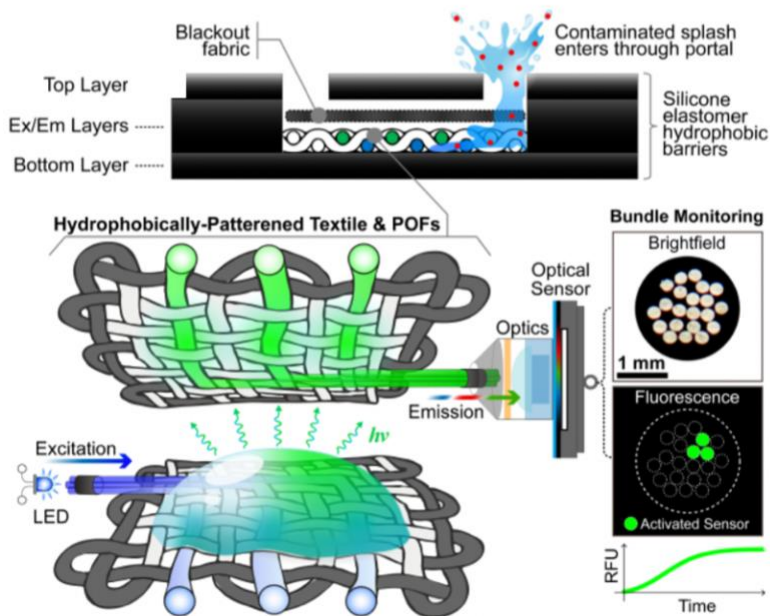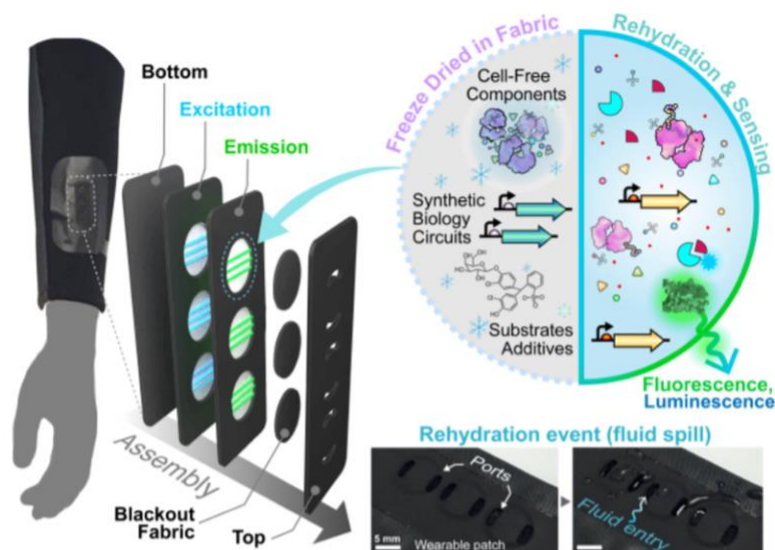
Figure 3-7. Fluorescent wearable patch design and operation. Top: Details of fiber-optic based wFDCF module for fluorescence/luminescence outputs. Fiber-optic embedded textiles allow excitation of the samples and detection by sensing emission light. A layers of blackout cover made of polyester fabric prevent the entry of environmental light. Rehydration ports over the device allow aqueous sample to be wicked into internal reaction wells. Bottom: A cross-sectional view of the interior of the device, where two layers of hydrophobically patterned fabric inter-woven with polymeric optic fibers allow for rehydration of freeze-dried cell-free reaction components as well as to provide light input/output for excitation and emission signals.
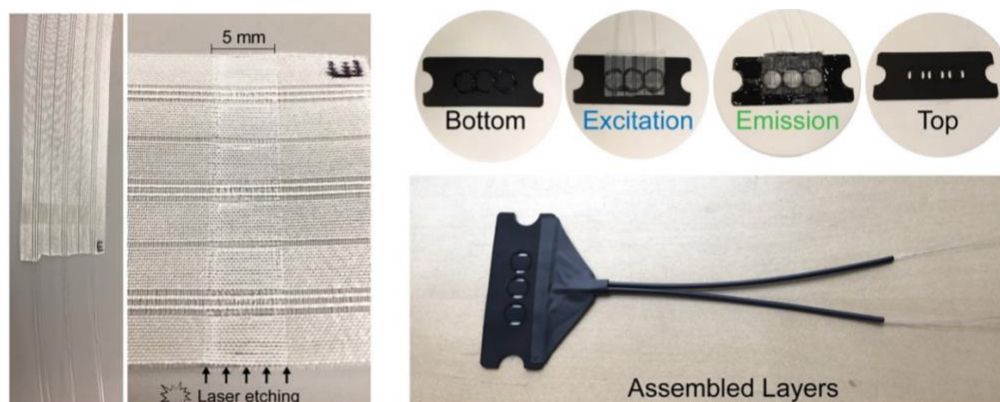
Figure 3-8. Fluorescent wearable patch materials and fabrication. Left: A cut strip of hydrophilic POF fabric was laser-etched (5 mm) to disrupt the POF outer cladding in the POFs sections closest to the reaction zone. Right: Examples of prepared wFDCF fabric-elastomer layers and final assembly 5 into a three-well sensor for garment integration. POFs in these devices were covered with black heat shrink tubing (6 mm). Top elastomer cover features two 5.19 x 1.85 mm curved sample ports instead of three as in the colorimetric prototypes to reduce direct light leakage on top of the POFs that may cause background light detection.
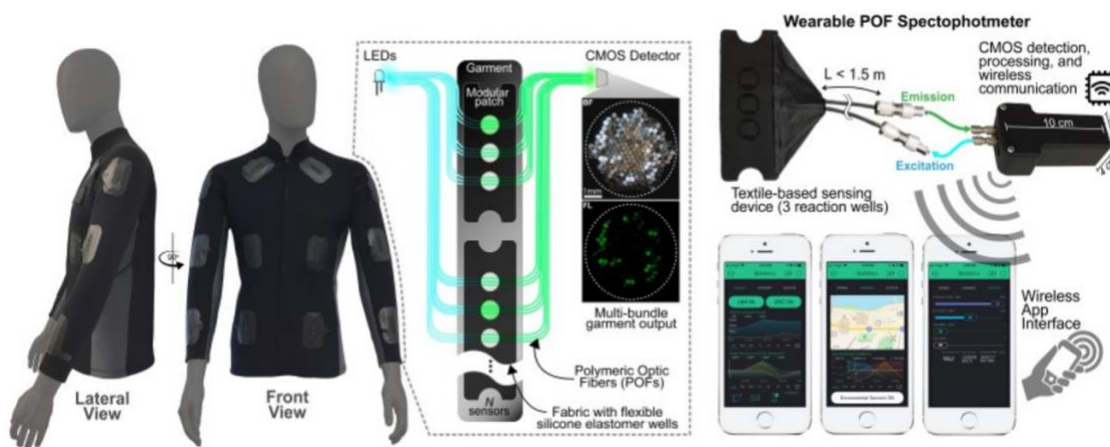
Figure 3-9. Integration of fluorescent wearable patches into wireless digital sensing platform. Left: Distributed continuous sensing of garment sensors can be achieved through multi-bundle imaging. Right: Connection of wearable POF spectrometer to digital wireless module. The spectrometer electronics consist of a Raspberry Pi Zero with a camera module as well as LED illumination, environmental sensing, and custom-fabricated shields for battery power. Smartphone application for visualization and alarm of wFDCF sensor activation was based on the blynk.io platform. This application allows for wireless recording of experiments, control of device parameters, as well as environmental and geolocation information.

Using this integrated platform, we performed distributed on-body sensing of various target exposures. A sample activation through fluid splashing can be seen in Figure 3-7 (top panel), where the sample wicks through the entry ports with blackout fabrics to rehydrate the freeze-dried, cell-free synthetic biology reactions immobilized within the hydrophilic textile fibers. These fibers are located within the excitation and emission layers of the device as shown in Figure 3-7 (bottom panel) and 3-8. Trigger presence in the splash fluid leads to activation of the sensor circuits, which produce fluorescent or luminescent reporters.

The versatility of this textile platform in fluorescence mode was first verified using two independent synthetic biology modules upstream of a superfolder green fluorescent protein (sfGFP) operon. These demonstrations included the activation of constitutive sfGFP expression and sensing of theophylline using an inducible riboswitch (Figure 3-10, top and bottom left). Although the detection of theophylline on its own is not of direct clinical relevance, as stated previously it is an important model test case for the use of a plethora of other small-molecule sensing riboswitches. A third fluorescence demonstration was done via activation of a 49-nucleotide Broccoli aptamer (Figure 3-10, top right) with substrate-specificity to (Z)-4-(3,5-difluoro-4-hydroxybenzylidene)-1,2-dimethyl-1Himidazol-5(4H)-one (DFHBI-1T), evincing functionality of this emerging class of fluorescent sensors in synthetic biology [72]. Furthermore, demonstrations utilizing luminescence outputs were conducted using a nanoLuciferase [73] operon downstream of an HIV RNA toehold switch (Figure 3-10, bottom right), as well as a B. burgdorferi RNA toehold switch for the wearable detection of Lyme disease (not shown). As previously stated the toehold switch is on its own not capable of sufficiently sensitive detection as to be deployed for clinical utility, but it nonetheless serves as an encouraging test case.
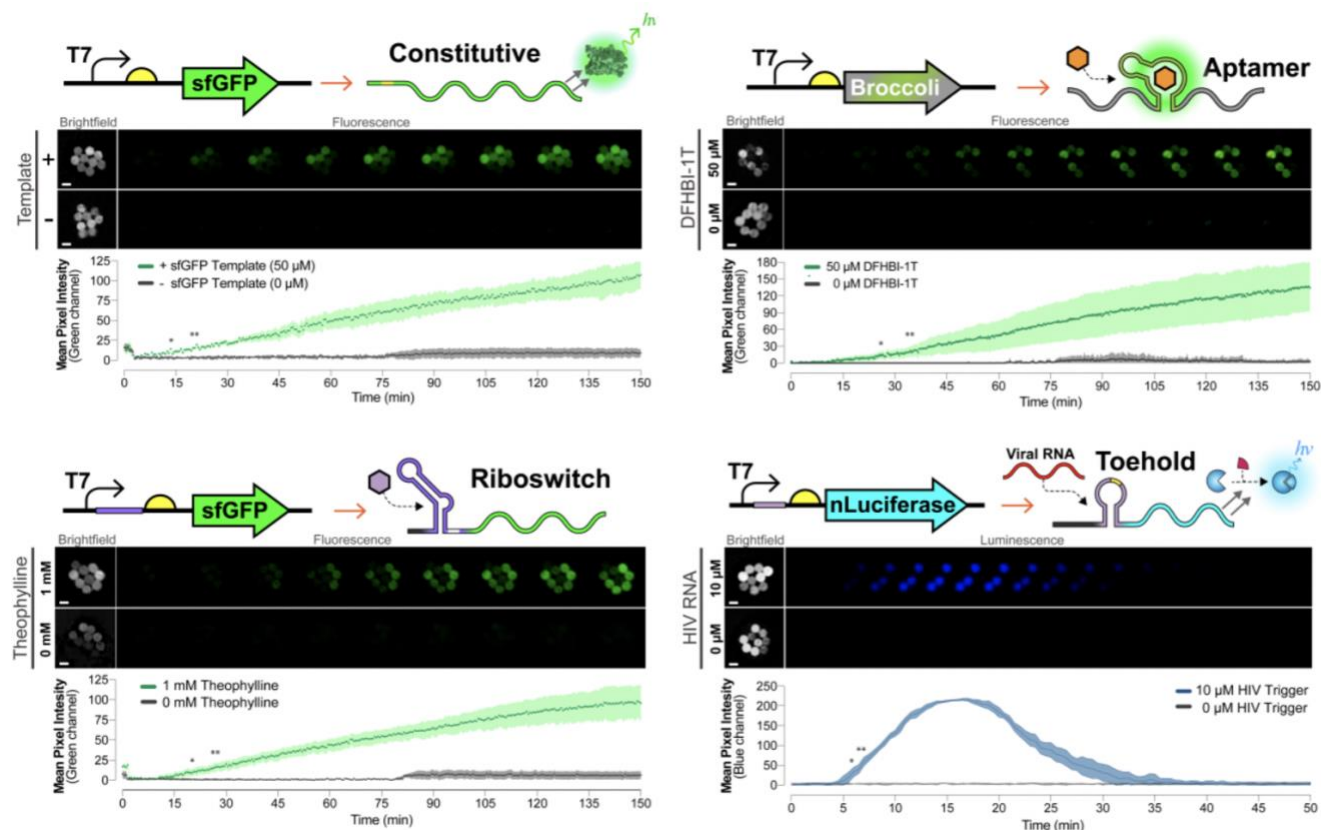
Figure 3-10. Integration of fluorescent wearable patches into wireless digital sensing platform. Top left: Rapid fluorescent signal after rehydration of wFDCF constitutive sfGFP template as compared to control. Fluorescent signal in-device is statistically distinguishable from the control after 11 min (P<0.05). Bottom left: Activation of FDCF riboswitch with 1 mM theophylline in a wearable device as compared to 0 mM theophylline control. Fluorescent signal in-device is statistically distinguishable from the control after 19.5 min (P<0.05). Top right: Wearable demonstration of fluorescent Broccoli aptamer being activated by the presence of 50 µM DFHBI-1T substrate as compared to 0 µM DFHBI-1T control. Fluorescent signal in-device is statistically distinguishable from the control after 24.5 min (P<0.05). Bottom right: Luminescence output detected from an HIV toehold sensor with NanoLuciferase operon. HIV RNA trigger was added at 10 µM and was statistically distinguishable from the control after 6 min (P<0.05) post-rehydration. All images above graphs correspond to time sequences of the recorded POF images in each sensor demonstration with bundle pictures synchronized with reaction profiles. Each experiment is from three independent wells, each having three fiber optic sensors for a total of 9 fiber optic outputs. Scale bar in brightfield images is 250 microns.

While improving on the sensitivity of the NAT assays in our colorimetric prototypes proved intractable, with a fluorescence readout we were able to deploy a CRISPR-based DETECTR/SHERLOCK assay that relies on quenched-fluorophore reporters for signal generation. As mentioned in chapter two, the advantages of CRISPR-based systems over existing biosensors include high sensitivity, specificity, freeze-drying compatibility, and the notable programmability to target any DNA or RNA sequence through interchangeable guide RNAs (gRNAs). Thus, we integrated CRISPR-based sensors into our fluorescence wFDCD platform to demonstrate this detection technique in wearable applications. We used Cas13a and Cas12a for the detection of RNA and DNA, respectively. For DNA detection, we used a Cas12a ortholog from Lachnospiraceae bacterium (LbaCas12a)[74,75] that displays a non-specific collateral cleavage activity towards single-stranded DNA (ssDNA) after detection of a gRNA-defined double-stranded DNA (dsDNA) target. This Cas12a-based sensor was paired with recombinase polymerase amplification (RPA)[76] and freeze-dried into a one-pot reaction to demonstrate state-of-the-art detection limits for wearable clinical applications. In the presence of a target dsDNA sequence, isothermally generated RPA amplicons activate Cas12a-gRNA complexes. Then, active Cas12a engages in trans-ssDNase activity and cleaves quenched ssDNA fluorophore probes, resulting in a fluorescence output (see chapter two). For our wearable CRISPR-based demonstrations, we designed gRNAs against three common resistance markers in Staphylococcus aureus: specifically, the mecA gene common in methicillin-resistant S. aureus (MRSA) [77], the spa gene which encodes the protein A virulence factor[78], and the ermA gene conferring macrolide resistance[79]. When tested in wFDCF format, our RPA-Cas12a sensors displayed detectable signals within 56-78 min ($P<0.05$) with femtomolar limits of detection (Figure 3-11). Moreover, using our mecA wFDCF sensor (Figure 3-11), we were able to confirm single-digit femtomolar sensitivity

of 2.7 fM. Compatibility with RNA inputs and other CRISPR enzymes such as Cas13a, the ortholog from Leptotrichia wadei bacterium (LwaCas13a) was also confirmed (data not shown), exhibiting similar in-device activation dynamics as that of cell-free reactions conducted in a plate reader. These results suggest that our wearable textile platform could be adapted to achieve sensitivities rivaling that of current laboratory diagnostic tests such as qPCR for monitoring contamination or spread of bacteria and viruses.
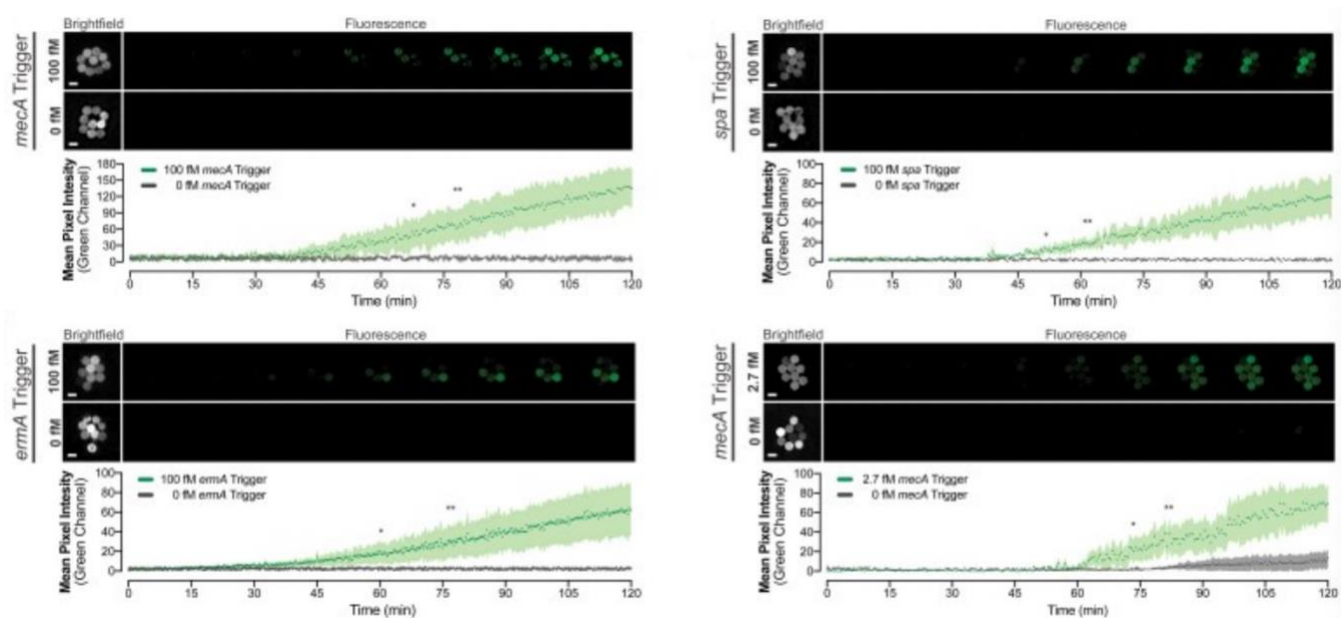
Figure 3-11. Validation of SHERLOCK detection assays in wearable platform. Top left: Freeze-dried mecA CRISPR-based sensor exposed to sample containing 100 fM mecA trigger or water. Top right: Freeze-dried spa CRISPR-based sensor exposed to 100 fM spa trigger or water. Bottom left: ermA CRISPR-based sensor exposed to 100 fM ermA trigger or water. Statistically distinguishable signals (P<0.05) were observed after 72, 56 and 78 min for mecA, spa and ermA sensors respectively. Bottom right: Detection of mecA CRISPR-based sensor at 2.7 fM trigger was statistically distinguishable after 75 min (P<0.05), corresponding to ~10,000 copies/µL. Each experiment is from three independent wells, each having three fiber optic sensors, for a total of 9 fiber optic outputs. Statistical significance is indicated for specific time points (*P ≤ 0.05 and **P ≤ 0.01).

To further demonstrate the modularity of our CRISPR-Cas12a wearable sensors, we tested wFDCF devices containing three orthogonal Cas12a-gRNA complexes in isolated reaction wells (Figure 3-12). In this experiment, each device was splashed with dd-H2O containing different targets, each specific to only one Cas12a-gRNA complex. The orthogonal behavior of our CRISPR-based wearable sensors is robust, where higher fluorescence was observed only for the cases in which the dsDNA trigger matched the pre-defined Cas12a-gRNA complex at each sensor location. These results suggest the broad applicability of CRISPR-based synthetic biology sensors for multiplexing or logic-gating in wearable synthetic biology applications.
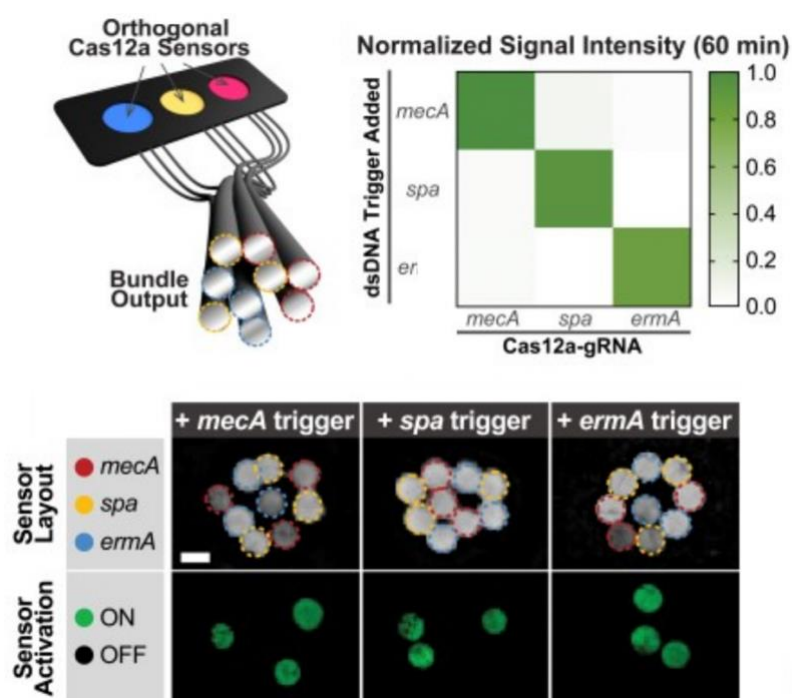
Figure 3-12. Three-fold multiplexed SHERLOCK detection assays. Top: The orthogonality of SHERLOCK assays is demonstrated for three targets with a mecA/spa /ermA CRISPR-based multi-sensor wearable. Rehydration only yielded activation of sensors when the Cas12a-gRNA sensor was in the presence of its programmed trigger dsDNA. Bottom: Activation of each individual sensor was distinguishable by assessing the corresponding bundle fibers. Scale bars in brightfield images are 250 µm.

In addition, the fluorescence optical sensor allows for facile fluorescent output multiplexing simply by using fluorescent proteins with orthogonal emission profiles (Figure 3-13). In this example, wFDCF reactions for three constitutively expressed fluorescent output proteins (eforRed30 , dTomato31, and sfGFP32) [80,81,82] were used to demonstrate detection of distinguishable output signals in a single bundle. We also found that the wFDCF POF system is fully compatible with an integrated lyophilized lysis components, allowing for release and detection of a plasmidborne mecA gene when challenged with intact bacterial cells (data not shown). Finally, to develop a complete data feedback cycle between the platform and the user, we integrated the detector system with a custom wireless mobile application that enables continuous cloud-based data logging, signal processing, geolocation tracking, and on-the-fly control of various detector components through a smart phone or other networked digital device (Figure 3-9). All images and spectral data presented were collected and processed using wFDCF devices fully integrated with our wearable spectrometer and mobile phone application.

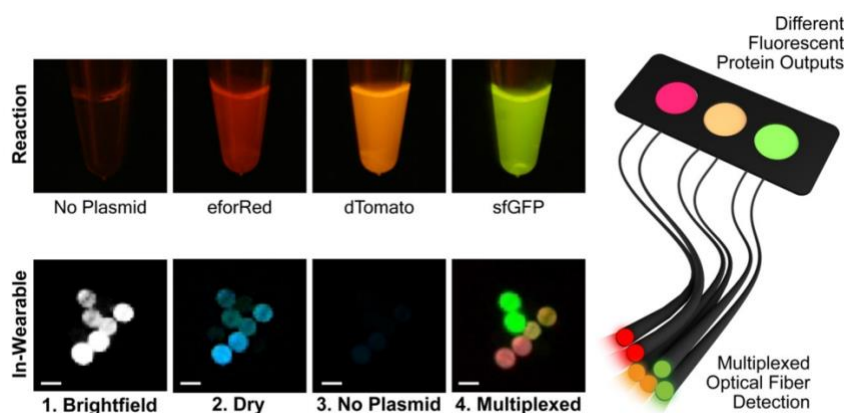**Figure 3-13. Sensor multiplexing using fluorescent proteins.** Top row, cell-free reactions demonstrating different fluorescent proteins synthesized after 30 min at 30C. Bottom row, sensor images of fiber topic bundles in (1) brightfield (2) excitation when the sensor is dry, (3) rehydrated reactions without plasmid, and (4) rehydrated cell-free reactions with fluorescent protein plasmids (30 min incubation at 30C).

Lastly, though not a NAT diagnostic application, we tested the operation of our fluorescent wearable diagnostic platform for the detection of chemical threats such as organophosphate nerve agents used in chemical warfare and the pesticide industry, both of which constitute prime targets for wearable detection. To achieve this, we modified our POF platform optics for excitation and detection at near-infrared (NIR) fluorescence, generated from a lyophilized acetylcholinesterase (AChE)-choline oxidase (ChOx)-HRP coupled enzyme reaction (Figure 3-14). In the presence of acetylcholine, this reaction can produce NIR fluoresence that is readily detectable with our wearable prototype. When exposed to an organophosphate AChE inhibitor, the sensor fluorescence is ameliorated as compared to unexposed controls. Our wearable nerve agent sensor was validated using paraoxon-ethyl as a nerve agent simulant at levels that are four orders of magnitude lower than the reported lethal dose (LD50) by dermal absorption in mammals [83].
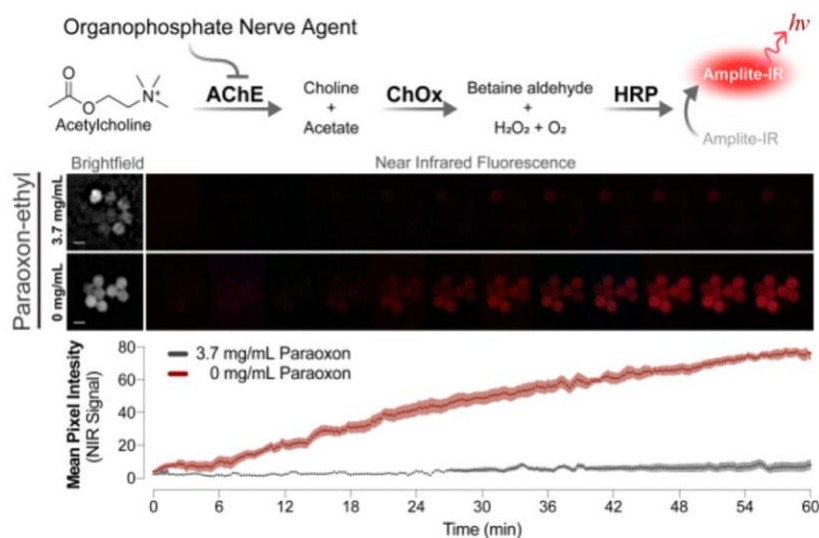

Figure 3-14. Wearable detection of organophosphate nerve agents. A lyophilized HRP-coupled enzyme sensor was rehydrated with 3.7 mg/mL paraoxon-ethyl (acetylcholinesterase inhibitor). An Amplite IR substrate is oxidized to generate near-IR fluorescence emission. Each experiment is from three independent wells, each having three fiber optic sensors, for a total of 9 fiber optic outputs.

3.4 A Face Mask-Embedded Wearable Sensor for SARS-Cov2

The wearable freeze-dried synthetic biology sensors demonstrated here thus imbue programmable and highly sensitive diagnostic sensing to general apparel. With the current SARS-CoV-2 pandemic that has led to significant strain on the medical system of all impacted countries and considerable delays in diagnostic testing, we explored whether our wFDCF system could be adapted to a key wearable gear, face masks, that has been shown to be critical in reducing the transmission of this highly infectious virus [84,85]. Although face masks are placed on all incoming patients that are presumptive SARS-CoV-2 carriers, confirmation through burdened laboratory diagnostics may result in delays that could negatively impact rapid triaging or effective contact tracing of patients [86,87,88]. Here, we demonstrate that our freeze-dried synthetic biology sensors can be adapted for a rapid point-of-care diagnostic integrated into a face mask, which may take advantage of the accumulation of SARS-CoV-2 virus on the inside of the mask as a result of coughing, talking and normal respiration, as demonstrated in numerous studies [89-99]. A large surface area (1.3 in^2) collection sample pad is positioned inside of the mask in front of the mouth and nose area (Figures 3-15, 3-16). Unlike other current SARS-CoV-2 NATs (chiefly qPCR assays) that require laboratory equipment and trained technicians [84, 85-88], all of the steps in our face mask NAT sensor are freeze-dried, shelf-stable, and passively actuated – all the user has to do is press a button on the outside of the mask, which pierces a plastic blister containing nuclease-free water. Capillary action then wicks fluid and viral particles through a folded wax-printed μPAD (microfluidic paper-based analytical device) arrangement containing various reaction zones with freeze-dried lysis and detection components (Figure 3-15). Between each of the reaction zones are polyvinyl alcohol (PVA) time delays that enable tunable incubation times between each reaction, greatly improving the efficiency of the sensor. Lysis reagents, similar in composition to those

previously used by others to lyse viral membranes, release the viral genome [100,101,102]. Next, the sample moves to an RT-RPA reaction zone containing an optimized isothermal amplification reaction developed to target a non-overlapping region of the SARS-CoV-2 S gene. The final reaction zone contains a Cas12a SHERLOCK/DETECTR sensor with an optimized gRNA for detection of the amplified dsDNA amplicon. In the presence of SARS-CoV-2 derived amplicons, the activated Cas12a enables trans-cleavage of a 6-FAM-(TTATTATT)-Biotin ssDNA probe, which is detected using an integrated lateral flow assay (LFA) strip threaded to the outside of the mask for a simple visual readout. The output strip is positioned to preserve patient confidentiality (Figures 3-15, 3-16).
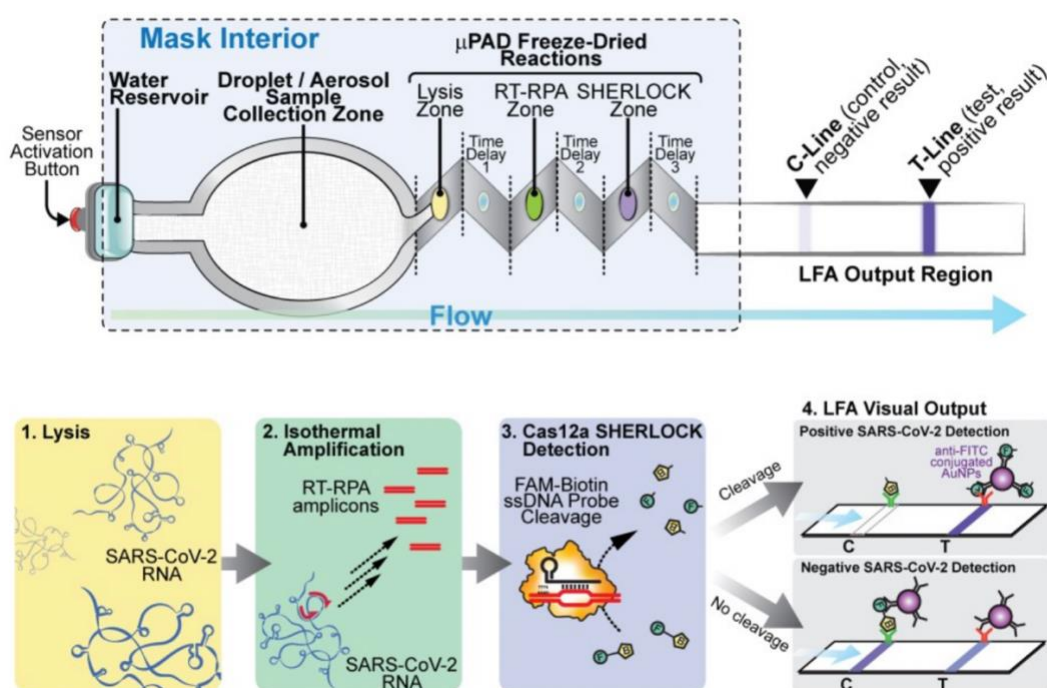
Figure 3-15. Schematic of SARS-Cov2 mask assay. Top: Puncture of the water blister reservoir results in flow through wicking material, moving viral particles collected from the wearer's respiration from the sample collection zone to downstream freeze-dried reactions integrated into a µPAD device. Bottom: Lysis releases SARS-CoV-2 vRNA, RT-RPA amplification proceeds at room temperature, and finally Cas12a detection results in collateral cleavage of FAM-Biotin ssDNA probes visualized on an LFA strip.

The total reaction time from activation of the face mask sensor to a final readout only takes ~30-40 min. The limit of detection observed for our sensors is 500 copies of SARS-CoV-2 in vitro transcribed (IVT) RNA, which is comparable to other laboratory-based diagnostics [85,86,87,88] (Figure 3-17, left panel). The sensors also do not cross react to RNA from other commonly circulating human coronavirus strains (HCoV) (Figure 3-17, right panel).

Figure 3-16. Photographs of the SARS-CoV-2 sensor integrated into a face mask. Shown from the exterior angle (left) and the interior angle (right). Details of device function can be seen in Figure 3-15.

Most critically as compared to the other prototypes of a wearable freeze-dried synthetic biology diagnostic presented in this thesis, the hands-off diagnostic reaction proceeds to completion at room temperature, which to our knowledge is the first SARS-CoV-2 NAT that is able to achieve high sensitivity and specificity at ambient temperatures. This crucial ability obviates the need for any heating instruments and reduces evaporative loss, which ultimately allows for significantly improved integration into a wearable format. However, it should be noted that this ability was largely achieved through brute force screening of RPA primer pairs

(unpublished data). As described previously, most isothermal amplification methods generally do not perform well at ambient temperatures. Although the Cas12a used in DETECTR/SHERLOCK operates well at ambient temperature, RPA operates robustly only at elevated temperatures of 30-37C. At lower temperatures of 20-25C, only a very small fraction of RPA primer pairs will give efficient enough amplification to achieve clinical sensitivity. In our hands this was a nearly intractable challenge, and it is likely that for many pathogenic targets, it may not be possible to find a primer pair that performs acceptably at room temperature. In addition to posing a significant design obstacle (reducing the modular plug-and-play nature of the platform), robustness in terms of day-to-day reproducibility was also adversely impacted by attempting to use RPA at ambient temperature, posing a significant obstacle for product development. Nonetheless, we were able to obtain impressive data from the optimized prototype.
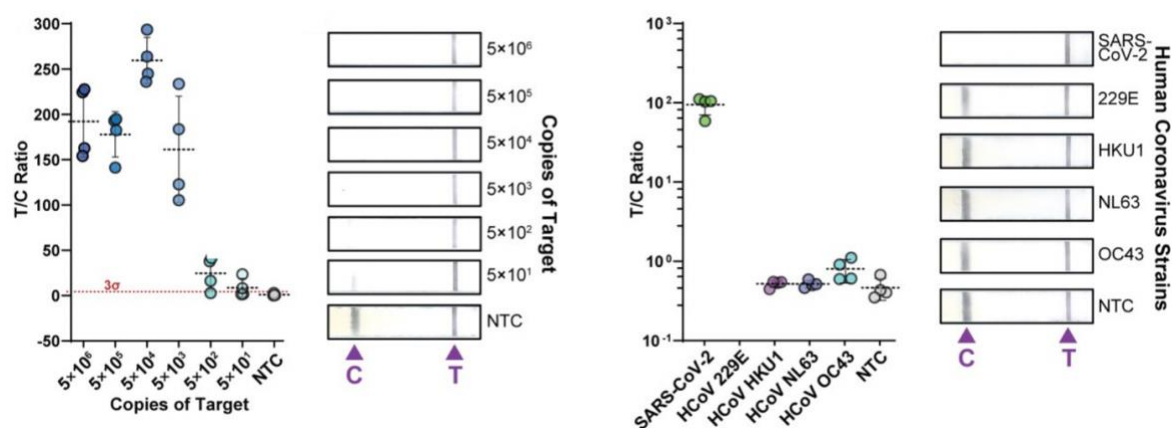
Figure 3-17. Limit of detection and cross-viral-strain specificity of SARS-Cov2 facemask. Left: Sensitivity of the face mask sensors at various inputs on the sensor zone of synthetic SARS-CoV-2 S-gene RNA. The limit of detection threshold, +3 S.D. of the no-template control (NTC), is shown as a red dotted line. Right: Specificity demonstration of face mask sensors shows no cross-reactivity with synthetic RNA from other commonly circulating human coronaviruses. SARS-CoV-2 RNA was added at 100,000 copies. All other HCoV RNAs were tested at 1,000,000 copies. Representative images of LFA outputs from the sensitivity measurements are shown for both sets of summarized data.

It is possible that, despite questions of robustness and scalability, a rapid face mask-integrated SARS-CoV-2 diagnostic as presented here could relieve saturated medical systems by combining protection and sensing into a simple and easy-to-deploy wearable system, greatly improving patient outcomes. Our face mask system could also potentially be further developed to discriminate between SARS-CoV-2 and other respiratory viruses, such as influenza, for rapid triaging of patient populations.

3.5 Concluding thoughts on Wearable Cell-Free Synthetic Biology Diagnostics

In summary, cell-free synthetic biology systems can be used to build practical wearable biosensors that are shelf-stable, genetically programmable, and highly sensitive. We designed and validated wFDCF sensors that are responsive to external rehydration events, such as splashes with contaminated fluids, and that withstand inhibitory evaporative and dilutive effects in open-environment conditions (30-40% RH and ~25-30C). We showed that these freeze-dried systems generate measurable colorimetric, fluorescence, or luminescence outputs upon exposure to relevant real-world targets such as SARS-CoV-2. We also demonstrated the integration of our device designs into garments that are compatible with wireless sensor networks to provide real-time dynamic monitoring of exposure using custom smartphone applications.

To our knowledge, the presented platform is the first wearable technology demonstrated to detect nucleic acids from potential viral or bacterial pathogens in contaminant fluid samples with sensitivities rivaling those of traditional laboratory tests at ambient temperatures. Our wFDCF platform evinces a number of distinct advantages over existing POC diagnostics, which similarly attempt to eliminate the need for time-consuming laboratory tests. Current field-portable POC systems typically use a swabbed or directly applied sample to provide a readout. Our wearable

platform accomplishes field sensing in one of the most critical environmental spaces for testing – that is, the surface of the user or areas that collect patient samples, such as the inside of a face mask. Moreover, in contrast to a batch-mode POC sensor, our wFDCF synthetic biology sensors can be networked to provide sensing arrays of lyophilized reactions and lightweight polymer fabrics, thus cloaking the user and continuously generating high-density, real-time outputs without sacrificing comfort or agility in the field. Our platform is also designed to operate autonomously, unlike most current POC instruments that require training for use and multiple operations by the user to acquire the final results. This feature removes the need to perform regular exposure checks, freeing those in the field to focus on their core tasks. Finally, the wFDCF components and optical fiber textiles are inexpensive, allowing the wearables to be utilized as disposable protective garments with advanced sensing technology.

Field applications that would greatly benefit from our wFDCF synthetic biology platform include warfighters and first-responders operating in environments where a specific chemical or biological threat is suspected. In this situation, our apparel of disposable wFDCF sensors could be used to maintain situational awareness, with continuous spatio-temporal monitoring of exposure and bodily resolution down to centimeters. Another set of potential uses for our platform involves the environmental awareness of clinicians, health workers, and researchers working in high-risk areas. Our wearable sensing platforms could enable rapid responses to contagion so that any exposed users could begin decontamination and neutralization procedures immediately. Similarly, wFDCF-enabled coats and gowns in hospitals could provide alerts to prevent the spread of nosocomial infections to vulnerable populations, such as immune-compromised patients or newborns. An additional promising application is patient-worn sensor-enabled wearables that can provide inexpensive, shelf-stable, and labor saving POC diagnostics to rapidly inform clinicians

in outbreak events, such as the current COVID-19 pandemic that has rapidly overwhelmed the resources of our medical infrastructure.

On a final note, there remains room for improvement in the freeze-dried isothermal amplification reaction itself. As mentioned, the ease of design and robustness of the RT-RPA reaction used in the face mask was adversely affected by operation at room temperature, and to address this in chapter five I will highlight efforts to advance the efficiency of isothermal amplification methods at low ambient temperatures.

Chapter 4

4.1 A deep learning approach to designing programmable RNA switches

Thus far in this thesis I have described the design and implementation of several engineered RNA-based synthetic biology tools, including the toehold switch, leaderless riboswitch (LL-Riboswitch), and gated Cas13 crRNA. More broadly, RNA-based tools have shown promise in recent years as programmable response elements that can be induced by small molecules, proteins, and nucleic acids, many of which hold great promise for a variety of in vitro and in vivo applications, including potential use in NAT diagnostics [39,103].

Unfortunately, despite their appeal the design and validation of this emerging class of synthetic biology modules have proven challenging due to variability in function that remains difficult to predict [38,39,104,105,106,107,108,109]. Notable thus far throughout this thesis is a focus on low-throughput experimentation in the design process, with rational design rules used as primary guide. Current efforts that aim to unveil fundamental relationships between RNA sequence, structure, and behavior have focused on mechanistic thermodynamic modeling and low-throughput experimentation (Figure 4-1). However, these often fail to deliver sufficiently predictive information to aid in the design of complex RNA tools, and in general designs that would be predicted to function based on rational thermodynamic modeling often fail to perform as expected [104,105,106,107,108,109]. Deep learning, by contrast, constitutes a set of computational techniques well suited for feature recognition in complex and highly combinatorial biological problems [110,111,112,113,114], such as the sequence design space of synthetic RNA tools. However, the application of deep learning to predicting function in RNA synthetic biology has been limited by a notable scarcity of datasets large enough to effectively train deep neural networks (DNN). Toehold switches, in particular, represent a benchmark RNA element in synthetic biology that

could greatly benefit from deep-learning approaches to better predict function and elucidate useful design rules [38,39].
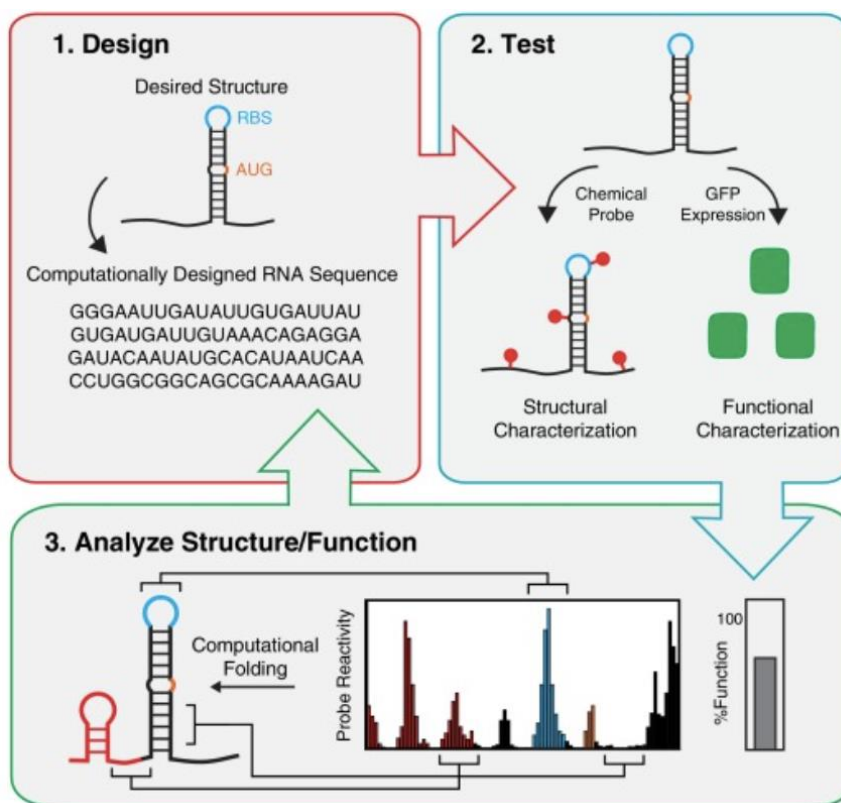
Figure 4-1. A low-throughput computational design paradigm for RNA synthetic biology. A representative scheme for common design efforts in RNA synthetic biology. 1) Novel RNA structures are designed based on rational design rules, 2) are synthesized and assayed in low throughput with typically fewer than dozens of variants, and 3) computational models based on rational biological insight are built to predict new variants. Figure from Chappel et al, 2015.

As described in chapter two, toehold switches are a class of versatile prokaryotic riboregulators inducible by the presence of a fully programmable trans-RNA trigger sequence (Figure 2-2) [38,39,40,104,105,106,115]. These RNA synthetic biology modules have displayed impressive dynamic range and orthogonality when used both in vivo as genetic circuit components [39,105,106],

and in vitro as nucleic acid diagnostic tools utilizing cell-free protein synthesis (CFPS) systems [103,104,115]. Similar to other RNA synthetic biology tools, a substantial fraction of toehold switches show poor to no measurable function when tested experimentally, and while efforts have been made to establish rational, mechanistic rules for improved performance based on low-throughput datasets [39,40,104-109,115], the practical utility of these approaches remains inconclusive. Thus, considering the wide applicability and general challenges of toehold-switch design, we chose to develop a deep-learning platform to predict toehold-switch function as a canonical RNA switch model in synthetic biology.

To achieve this goal, we first expanded the size of available toehold datasets using a high-throughput DNA synthesis and sequencing pipeline to characterize over 10^5 toehold switches. We then used this comprehensive dataset to demonstrate that deep neural networks trained directly on switch RNA sequences can outperform rational thermodynamic and kinetic analyses to predict toehold-switch function. Furthermore, we enhanced the transparency of our deep-learning approach by utilizing a nucleotide complementarity matrix input representation to visualize important learned secondary-structure patterns in selected models. This attention-visualization technique, which we termed VIS4Map (Visualizing Secondary Structure Saliency Maps), allowed us to identify RNA module success and failure modes by discovering secondary structures that our deep-learning model uses to accurately predict toehold-switch function. The resulting dataset, models, and visualization analysis represent a substantial step forward for the validation and interpretability of high-throughput approaches to designing RNA synthetic biology tools, surpassing the limits of current mechanistic RNA secondary-structure modeling.
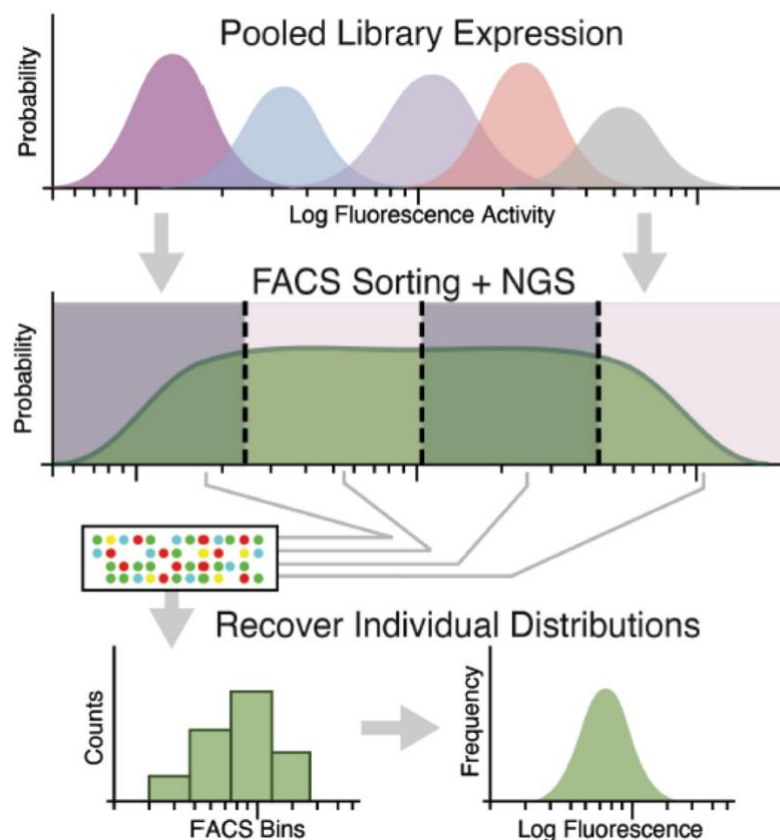
Figure 4-2. Summary of flow-seq assay pipeline. A flow-sequence (seq) pipeline was used to characterize the fluorescence signal of individual toehold switches in a pooled sequential assay. Pooled induction of a library of variants was followed by fluorescence-activated cell sorter (FACS) sorting, and then next-generation sequencing (NGS). Count frequency analysis of the resulting NGS dataset was used to recover the original fluorescence distributions of the data. Figure adapted from Peterman et al. [122].

4.2 Synthesis and Validation of a Toehold Switch Library

A fundamental hurdle in applying deep-learning techniques to RNA synthetic biology systems is the limited size of currently published datasets, which are notably smaller than typical dataset sizes required for the training of deep network architectures in other fields [110,116-119]. For example, to date, <1000 total toehold switches have been designed and tested [38,39,40,103-105,115]. While a recent attempt was made to apply deep learning to a bacterial riboswitch dataset with 263

variants [120], the lack of high-throughput datasets has generally limited the synthetic biology community's ability to analyze this type of response molecule using deep-learning techniques. High-throughput assays that utilize deep sequencing to analyze fluorescence-sorted bacteria have previously been used to characterize the translation of Escherichia coli mRNA [121-125]; in this part of the thesis, in order to improve our understanding and ability to predict new functional RNA-based response elements, we synthesized and characterized an extensive in vivo library of toehold switches using a high-throughput flow-seq (also known as sort-seq)[121,122] pipeline for subsequent exploration using various machine-learning and deep-learning architectures.
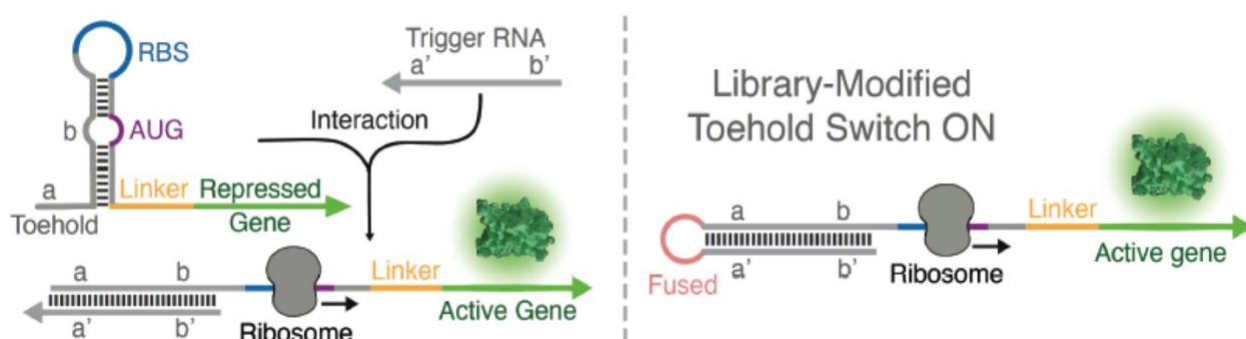
Figure 4-3. Modified library-compatible toehold switch design. Left: While the original toehold-switch architecture from Green et al.[39] was used, containing a 12-nucleotide toehold (a/a′) and an 18-nucleotide stem (b/b′) fully unwound by the trigger, it needed to be modified to be compatible with high-throughput DNA synthesis. Right: We selected to fuse the RNA trigger to the 5′ end of the switch by an unstructured linker to facilitate this.

Our toehold-switch library was designed and synthesized based on a large collection (244,000) of putative trigger sequences, spanning the complete genomes of 23 pathogenic viruses, the entire coding regions of 906 human transcription factors, and ~10,000 random sequences. From a synthesized oligo pool, we generated two construct libraries, for ON and OFF states, which were subsequently transformed into BL21 E. coli (Figure 4-2). The first library contained OFF toehold-

switch constructs that lacked a trigger, while the second library of ON constructs contained the same toeholds with complementary triggers fused to their corresponding switches (Figure 4-3). The two libraries were then sorted on a fluorescence-activated cell sorter (FACS) using four bins (Figure 4-4), and the toehold-switch variants contained in each bin were quantified using next-generation sequencing (NGS) to recover their individual fluorescence distributions from raw read counts. To accomplish this, frequencies of each variant were tabulated for each cell-sorted bin and normalized to the total reads per bin. Each variant's functional value was computed as the weighted mean of its normalized frequencies across all bins. Because each library was sorted using the same gates, and since each library spanned a remarkably similar range of minimum and maximum GFP intensity (Figure 4-4), we scaled the ON and OFF values for each variant to fall between [0, 1]. A value of 0 was given to a variant if all corresponding reads were found only in the lowest bin and a value of 1 if all corresponding reads were found only in the highest bin. An ON/OFF metric was calculated by subtracting these individuals ON and OFF signal metrics (Fig. 1), which resulted in values between [−1, 1] (to clarify to the reader, this was done because fluorescence measurements taken by the FACS sorting instrument were on a log scale and log-ratios are calculated by subtracting differences) (Figure 4-4).
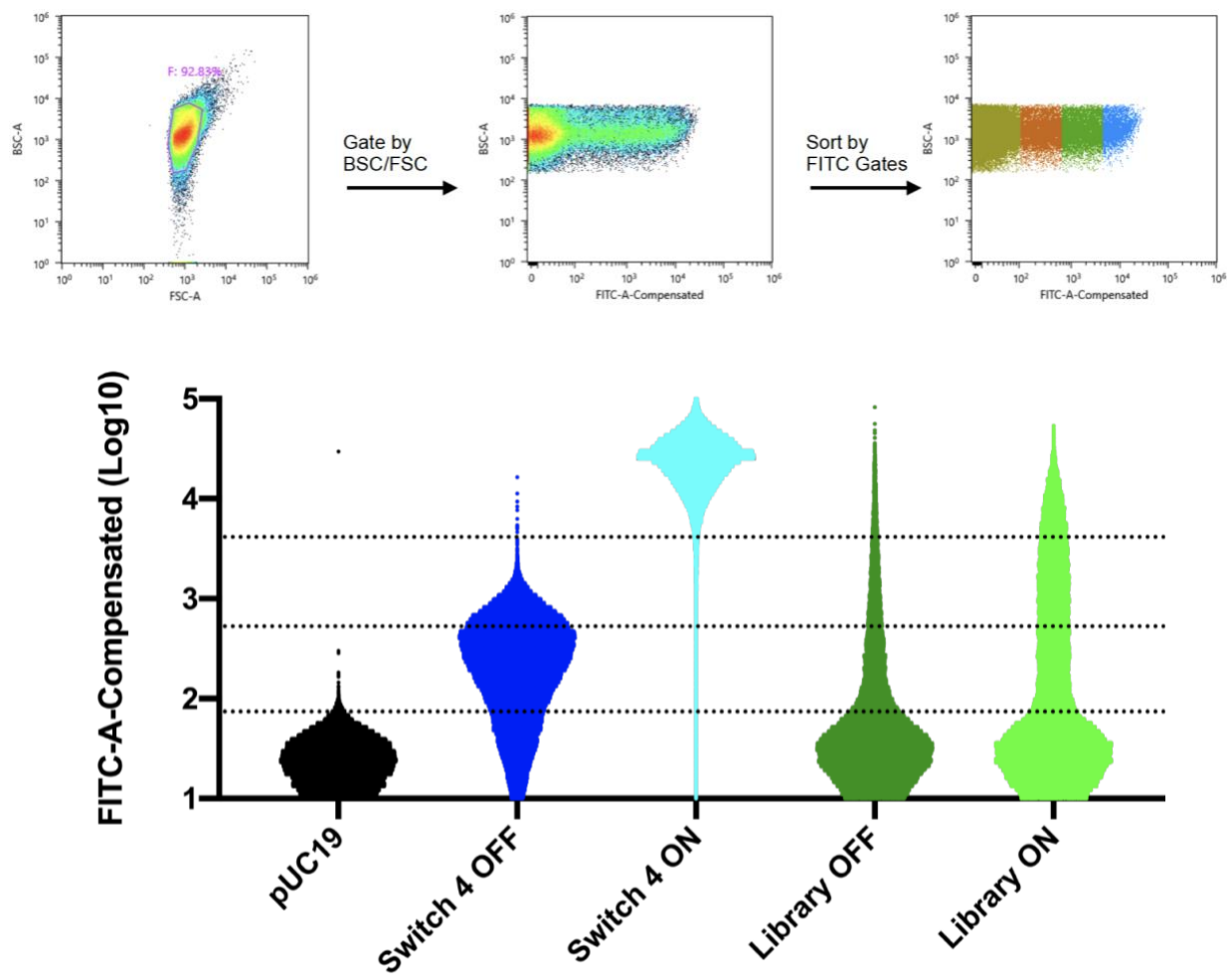
Figure 4-4. Library FACS cell population distributions and their empirically-derived sorting gates. Top: The gating strategy for sorting IPTG-induced *E. coli* BL21-star cells by GFP fluorescence is shown. Bottom: FITC distribution plots from the three control conditions, the complete ON and OFF libraries, and the pooled panel of twenty switches from Green et al. [39], with the boundaries of the four final sorting bins shown as dotted lines. The resulting measurements obtained for ON, OFF, and ON/OFF using our flow-seq pipeline are shown as raw fluorescence values or raw fold change, rather than normalized to the range of [0,1] or [-1, 1]. To determine the boundaries of the sorting gates for our high-throughput toehold switch pipeline, we used Switch #4 from Green et al. [39] in ON and OFF conformations as a positive control, and a pUC19 plasmid lacking a GFP gene as a negative control.

Interestingly, the distribution of GFP signal in the flow-sorting data displayed in Figure 4-4 is highly imbalanced for both the ON and OFF libraries. Frustratingly, a large fraction of the oligo library pool contained incorrectly synthesized oligomers. These were largely truncated products lacking a start codon, lacking an SD sequence, or containing a frameshift that we would expect to lead to low GFP signal. We estimate that at least 50% of the cells that we sorted contained such a truncated variant, and most of these ended up in the lowest bin. Thus, only sequences matching our intended designs were retained for further analysis. For the ON and OFF libraries, respectively, 10,390,207 reads and 20,788,966 reads were mapped to a correct switch sequence. After quality control, the toehold-switch library contained 109,067 ON-state measurements, 163,967 OFF-state measurements, and 91,534 ON/OFF paired ratios, where both ON and OFF states were characterized for a given switch (Figure 4-5). These switches were representative of all of the *in silico* designed categories of trigger RNA targets including human transcriptome and viral genome targets (Figure 4-6). Both ON and OFF data spanned the full range of measured GFP signals, meaning that some ON switches failed to induce and expressed no measurable GFP signal, while some OFF switches failed to repress ribosome binding and leaked the maximum measurable GFP signal. The final ON and OFF datasets seen in Figure 4-5 are notably less skewed than the flow-sorting data seen in Figure 4-4, thanks to the exclusion of reads corresponding to incorrectly synthesized switches, but it should be noted that OFF data still remained skewed towards low-signal variants.
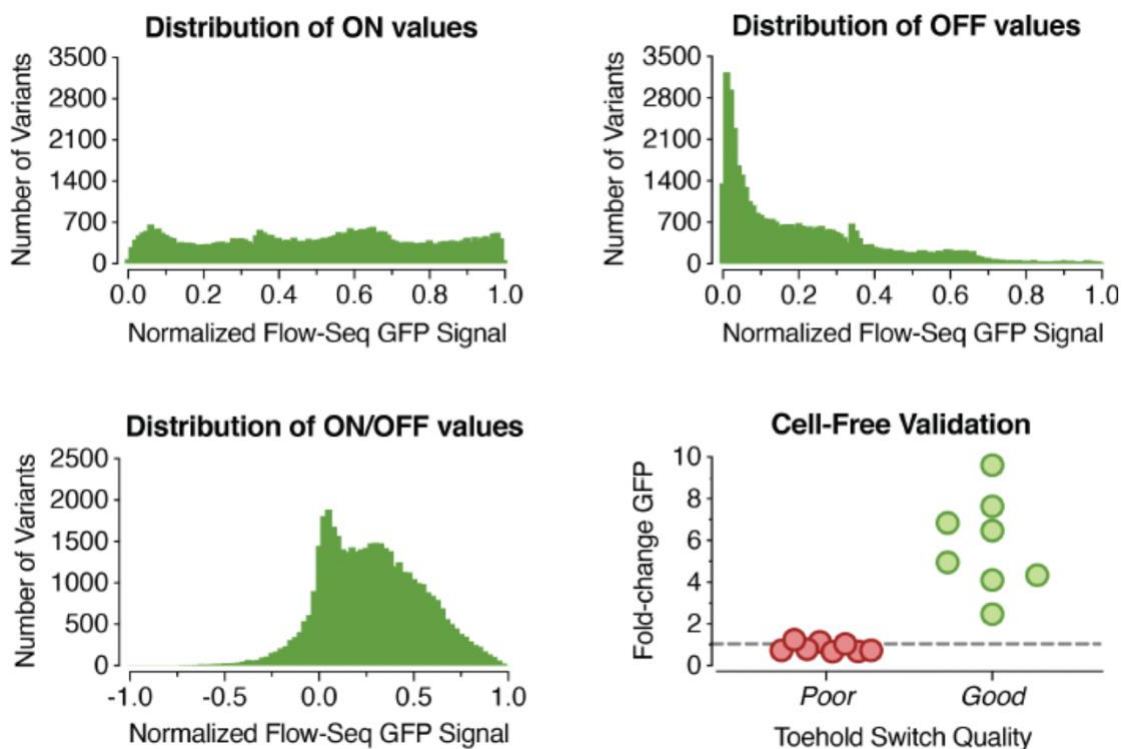
Figure 4-5. Quality-controlled toehold switch library distributions. The distribution of recovered toeholds for ON-state signals (top left), OFF-state signals (top right), and calculated ON/OFF ratios (bottom left) are shown. In the bottom right, validation results are shown for toehold switches expressed in a PURExpress cell-free system with un-fused-trigger RNA, including eight low-performing (poor, ON/OFF < 0.05) and eight high-performing (good, ON/OFF > 0.97) samples. This obtained in vivo flow-seq data shows competency in classifying switch performance for this in vitro cell-free biological context ($P < 0.0001$ between high and low switches, two-tailed t test) with $n = 3$ biologically independent samples each for both ON and OFF measurements.

Moving on after assaying our switch library, we wanted to validate our in vivo ON/OFF measurements in an in vitro setting to ensure they were reasonable indicators of switch performance in a CFPS system, since as seen in much of this thesis RNA synthetic biology tools are often used within in vitro cell-free systems [40,103,104,115.] To achieve this, we selected eight high-performance switches and eight low-performance switches, and individually cloned and

characterized each one in a CFPS (Figure 4-5). All low-performance switches showed no induction, while the high-performance switches showed a spread of ON/OFF ratios between 2 and 13 ($P < 0.0001$ between high and low switches, two-tailed t test). The wide range of GFP expression seen from the high-performance switches results from a relatively weak rank-order correlation we have observed between the performance of our toeholds in vivo and in vitro, which differs from other work comparing RNA actuators in living cells and cell-free systems [126]. The effect may stem from differences in trigger-toehold interactions between the in vivo cellular environment and the in vitro cell-free environment. Nonetheless, these results indicate that while the performance of toehold switches in vivo and in vitro may differ, in vivo measurements can still be used to classify categorically whether a switch will function in vitro.
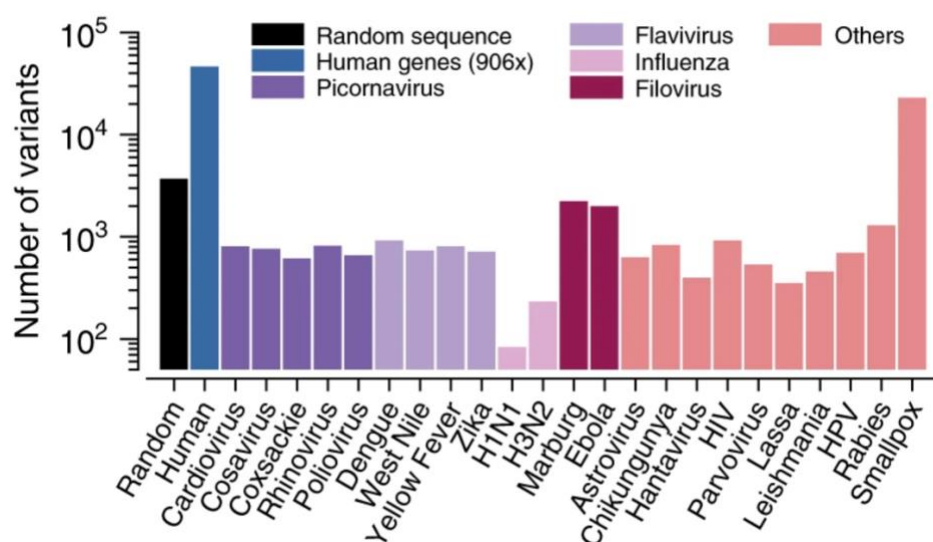

Figure 4-6. Quality-controlled toehold switch library representation. After quality control, tested switch/trigger variants were representative of each in-silico designed origin category, including randomly generated sequences, from 906 human transcription factor transcripts, and from 23 pathogenic viral genomes. The number of variants contained in each category is shown.

4.3 Rational Analysis of Library Data using RNA Secondary-Structure Models

Before initiating the exploration of deep-learning models to predict function in our large-scale toehold-switch library, we sought to determine whether traditional tools for analyzing synthetic RNA modules could be used to accurately predict toehold-switch behavior, including k-mer searches and mechanistic modeling utilizing thermodynamic and kinetic parameters. K-mer searches of biological sequence data are often used to discover motifs, but while certain overrepresented motifs were found in our dataset (Figure 4-7), methods for utilizing these did not readily facilitate functional predictions of switch behavior [127].
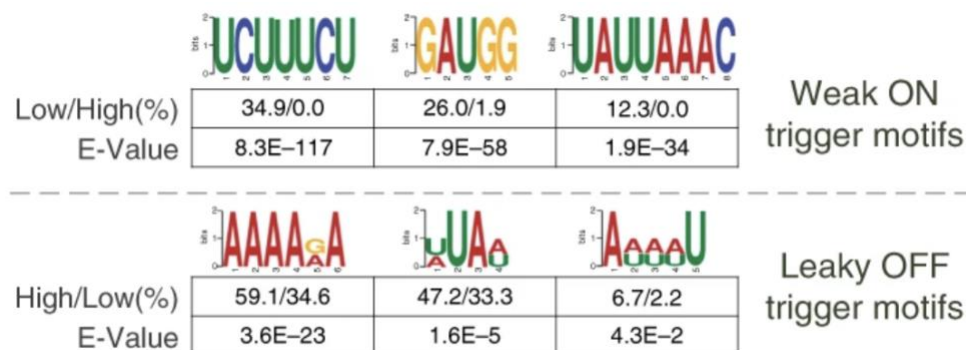
Figure 4-7. K-mer search for overrepresented sequence motifs. Sequence logos for k-mer motifs discovered to be disproportionately represented in weakly induced switches (low ON, top panel) and leaky switches (high OFF, bottom panel) are shown. In both cases the rest of the dataset was used as a background. Shown are functional proportions (percentage frequency of target group compared to background group), and the E-value of the finding.

Other current state-of-the-art approaches for designing RNA synthetic biology tools primarily analyze secondary structure using thermodynamic principles [128,129,130]. Following such prior works, we used NUPACK[128] and ViennaRNA[130] software packages to calculate a total of 30 rational features for our entire library, including the minimum free energy (MFE), ideal ensemble

defect (IED), and native ensemble defect (NED) of the entire toehold-switch library as well as various sub-segments in each sequence. A number of these parameters had previously been reported to correlate with experimental toehold-switch ON/OFF measurements for smaller datasets[2], and NUPACK's design algorithm, in particular, is set to optimize IED when proposing target RNA secondary structures and is by far the most commonly used tool in our lab to predict RNA secondary structure during day-to-day tool design [103,104,115,128]. However, when analyzing these rational features with our larger dataset, we found them to be poor predictors of toehold-switch function (Figure 4-8). In modest agreement with the findings of Green et al.[39], the MFE of the RBS-linker region showed the highest correlation of this feature set for ON/OFF (R2: ON = 0.14, OFF = 0.06, ON/OFF = 0.04), with NUPACK's IED also showing above-average correlation (R2: ON = 0.07, OFF = 0.02, ON/OFF = 0.03). While measurable, these correlation metrics were too weak for practical use in computer-aided design of this specific RNA synthetic biology tool [103,104,115,128]. Based on unpublished negative data in the lab, this was not a surprising result. Most often we find that, at a typical low-throughput scale of design iteration, using NUPACK and similar tools to design optimal versions of a desired RNA secondary structure or folding mechanism is not sufficiently better than random guesses that we notice a distinct benefit over time. Nonetheless these tools have continued to be used for years out of inertia and familiarity.

In addition to NUPACK and ViennaRNA, we next explored the use of more complex thermodynamic models that take into account well-established hypotheses for translation initiation and the ribosome docking mechanism in combination with multiple thermodynamic features to improve their predictions [131-137]. One of the most developed of these models is the ribosome-binding site (RBS) calculator (v2.1; Salis Lab), which is a comprehensive regression model parameterized on thousands of curated RBS variants [131, 134]. We used the RBS calculator to predict

the ON and OFF translation initiation rates for our toehold switches, but also found low predictive

performance comparable to other rational features (Figure 4-8) when tested on our database (R^2:
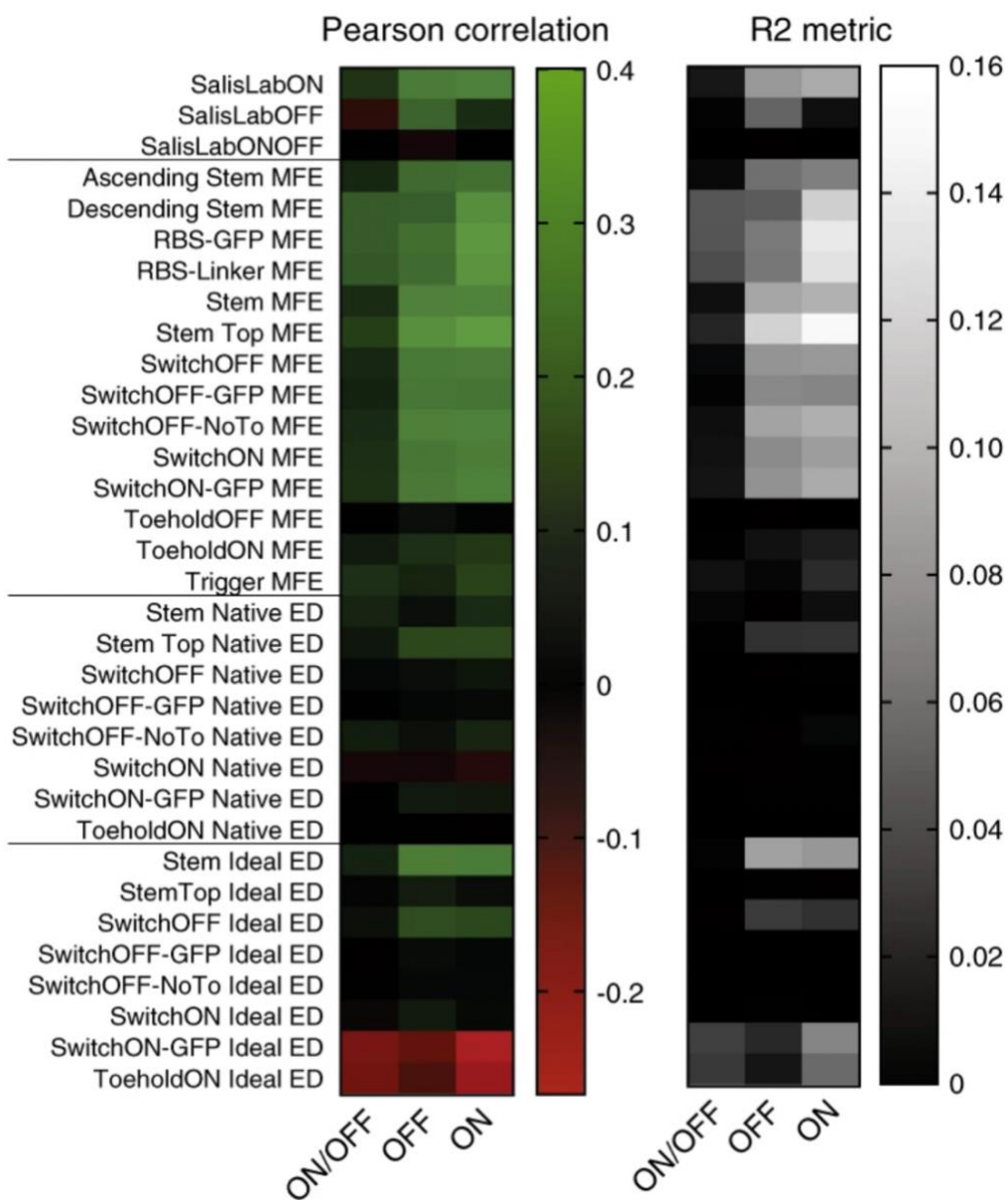
ON = 0.09, OFF = 0.05, ON/OFF = 0.0001).

Figure 4-8. Analysis of toehold switch dataset using rational thermodynamic modeling. The
Pearson correlation (left, |max| = 0.4) and R2 metric (right, |max| = 0.16) are shown for 30 state-
of-the-art thermodynamic features and obtained RBS Calculator v2.1 outputs when used to predict
either the ON/OFF, OFF, or ON of our toehold switch dataset.

One potential explanation for the limited predictive power of current thermodynamic models for RNA folding tasks concerns the influence of kinetically stable secondary-structure intermediates that may compete with thermodynamic equilibrium states [134,137,138]. To determine whether a kinetic analysis of toehold-switch folding dynamics could help explain our experimental results, we calculated four additional features based on kinetic trajectories using the Kinfold package [39] (Figure 4-9). Folding trajectories were run using the Kinfold package for the OFF-state switch sequence. Most trajectories took longer than 10^3 au, compared to the Kinfold analyses in Borujeni et al. [134] where average trajectory times fell in the range of 10^1 -10^3 au, and 10^4 au was the longest allowed trajectory time. For each switch in a higher-quality dataset (total 19,983 variants), 100 trajectories were run and the following measurements taken: the mean and negative standard deviation of the trajectories' average energy during the first 10^3 au, the fraction of trajectories that completed folding of the MFE structure before 10^3 au, and the ratio of average trajectory energy to the minimum possible MFE energy. As with predictions obtained using other thermodynamic models, these kinetic features showed poor correlations (R^2: ON $=0.04$, OFF $=0.04$, ON/OFF $=0.001$ for the best feature) to our empirical dataset (Figure 4-9). This suggested that models taking into account local energy minima calculated by rational mechanistic models did not perform better than similar models which only seek to predict the global energy minimum.
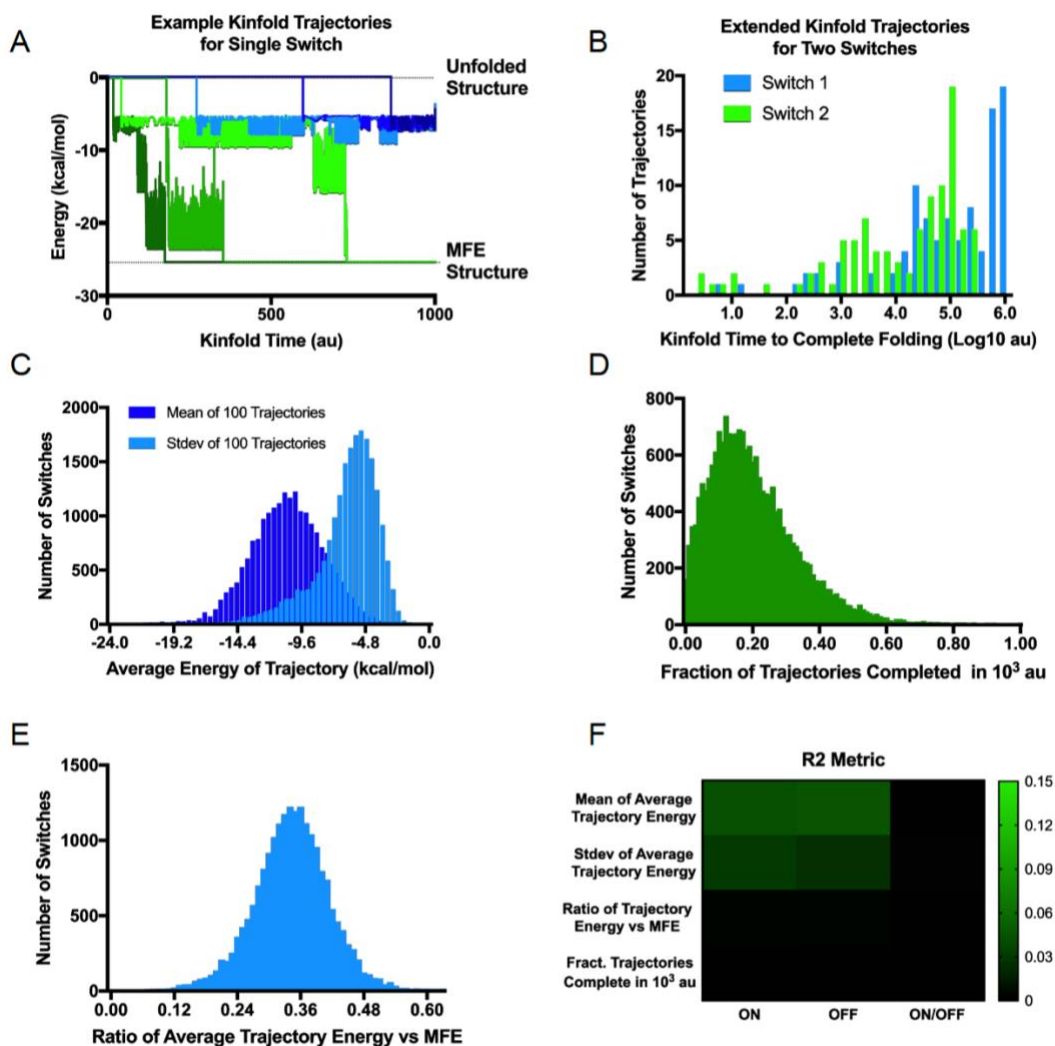
Figure 4-9. Kinetic toehold switch folding analysis using Kinfold. Folding trajectories were run using the Kinfold package [134] for the OFF-state switch sequence. (A) For a single representative toehold switch, six example trajectories are shown. Trajectories in green reached the MFE structure within 10^3 arbitrary time units (au), while those in blue did not. (B) For two representative toehold switches, 100 trajectories were run for a maximum time of 10^6 au. Histograms of the time required for a trajectory to reach the MFE structure are shown. (C) histograms of the mean and negative standard deviation of the trajectories' average energy during the first 10^3 au, (D) the fraction of trajectories that completed folding of the MFE structure before 10^3 au, (E) the ratio of average trajectory energy to the minimum possible MFE energy, and (F) the R2 correlation between the metrics in C,D,E and the empirical measurements in our toehold switch dataset. For comparison with previous rational features the heatmap axis is set similarly to Figure 4-8.

Considering these results, the cause of limited functional predictions from thermodynamic and kinetic RNA secondary-structure models remains unclear but may stem from the use of potentially incomplete energetic models, incorrect mechanistic hypotheses, or interference from the in vivo context of the bacterial cell. Regardless of the source of error, we next sought to explore deep learning as a machine-learning paradigm to develop models with higher predictive abilities than previously reported, with the hope of allowing useful computer-aided systems for the design of RNA synthetic biology tools.

4.4 Improved Prediction using Multilayer Perceptron Models

Given that simple regression models based on previous state-of-the-art RNA thermodynamic and kinetic calculations were ineffective at predicting toehold-switch performance, we next tested the use of feed-forward neural networks, also known as multilayer perceptron (MLP) models, as a baseline architecture for our investigation (Figure 4-10). We first trained a three-layer MLP model on our dataset with an input consisting of the 30 previously calculated thermodynamic rational features (see Figure 4-8). When trained in regression mode, this MLP model was able to deliver better predictions than any of the individual rational features or the RBS calculator based on R^2 and mean absolute error (MAE) (R^2: ON = 0.35, OFF = 0.25, ON/OFF = 0.20) (Figure 4-11). Similarly, when this model was trained in classification mode (ON/OFF: binarized at +/− 0.7), it achieved a 0.76 area under the receiver–operator curve (AUROC) and 0.18 area under the precision-recall curve (AUPRC), as seen in Figure 4-12. The MLP model slightly outperformed a logistic regressor trained on the same rational features (Figures 4-11 and 4-12), suggesting that the MLP architecture was able to abstract higher-order patterns from these features as compared to simpler non-hierarchical models.
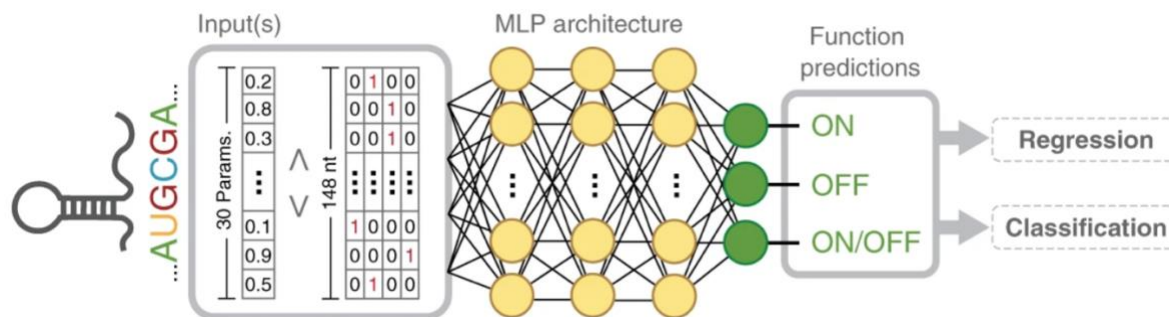
Figure 4-10. Multilayer perceptron model architecture. Base architecture of our MLP models, featuring three fully connected layers and an input consisting of thirty calculated rational features, a one-hot encoding of the switch sequence, or a concatenation of both. For training in regression mode, three different outputs were predicted (ON, OFF, ON/OFF), whereas for classification training, only a single binary output based on ON/OFF (threshold at 0.7) was predicted.

While these results already constitute an improvement compared to the current state-of-the-art analysis of RNA synthetic biology tools, we hypothesized that the use of pre-computed rational features as network input led to information loss that could inherently limit the predictive power of these models. Considering that possibility, we trained an MLP model solely on one-hot encoded sequence representations of our toehold switches, eliminating potential bias introduced by a priori mechanistic modeling. We found that this sequence-based MLP delivered improved functional predictions based on R^2 and MAE metrics (R^2: ON = 0.70, OFF = 0.53, ON/OFF = 0.43) (Figure 4-11). These values represent a doubling of R^2 performance as compared to the MLP trained on rational features and a tenfold improvement in ON/OFF R^2 over the best individual rational feature used for previous linear models. When training for classification, our one-hot sequence MLP produced similarly improved AUROCs and AUPRCs of 0.87 and 0.36, respectively (Figure 4-12).

Figure 4-11. Multilayer perceptron model performs better than simpler models with regression tasks. Top: Box-and-whisker plots for R^2 accuracy between experimental and regression-based predictions for best-performing rational features, logistic regression models and MLPs using tenfold cross-validation (dotted line shows the best individual rational feature). Bottom: Box-and-whisker- plots for mean absolute error (MAE) between experimental and predicted values for these same models. Model inputs were either thirty calculated rational features, a one-hot encoding of the switch sequence, or a concatenation of both.
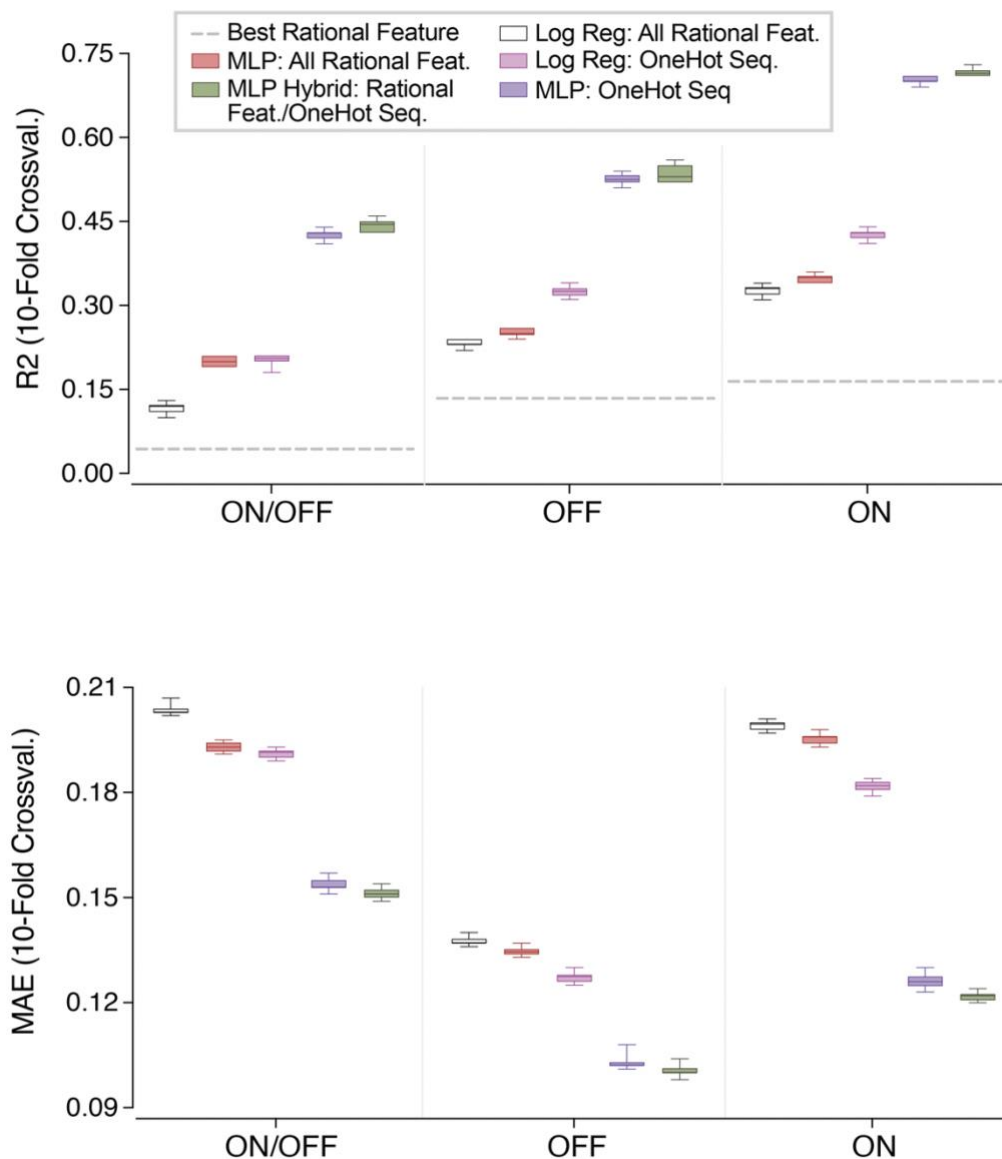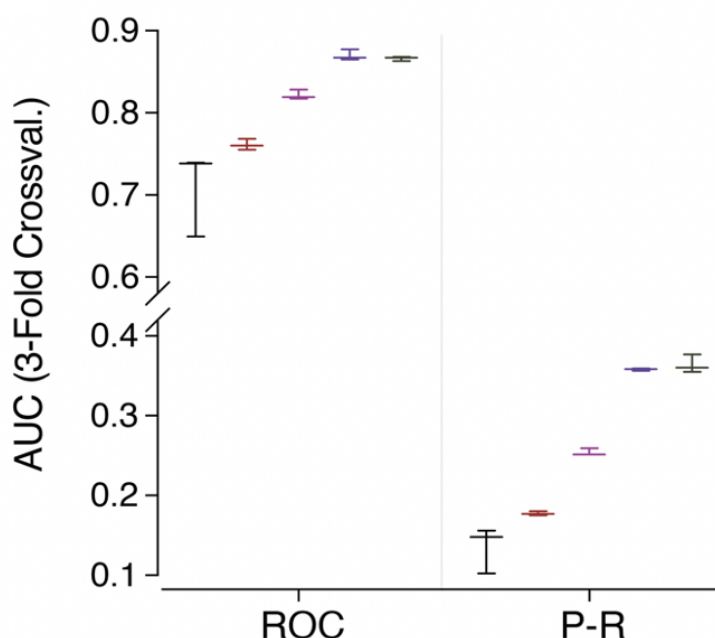
Figure 4-12 Multilayer perceptron model performs better than simpler models with classification tasks. Box- and-whisker plots for the area under the curve (AUC) of the receiver–operator curve (ROC) and the precision-recall curve (P–R) in classification-mode predictions compared to experimental values using threefold cross-validation. Model inputs were either thirty calculated rational features, a one-hot encoding of the switch sequence, or a concatenation of both.

The improvement in performance when training on sequence-only inputs compared to rational features suggests that significant information loss occurs when performing thermodynamic calculations on toehold-switch sequences, a problem that may extend to other RNA synthetic biology tools in use today. The sequence-only MLP model dramatically outperformed a logistic regressor model trained on the same one-hot sequence input (Figures 4-11 and 4-12), further supporting the hypothesis that improved accuracy of our sequence-based MLP arises from learned hierarchical nonlinear features extracted directly from RNA sequences. Concatenating both the rational features and the one-hot representation into a combined input gave a small but significant improvement in regression mode ($\Delta R2 \approx 0.025$ and $\Delta MAE \approx -0.0025$,

P < 0.05 for all six comparisons, two-tailed t test), but no significant improvement for AUROC or AUPRC when in classification mode (Figures 4-11 and 4-12). These results suggest that while the use of rational features may facilitate the abstraction of potentially relevant information of toehold-switch function, the one-hot sequence-only MLP model can recover such information without a priori hypothesis-driven assumptions built into the model if given a sufficiently large training dataset.

In order to evaluate the degree of biological generalization in our sequence-only MLP model, we performed two additional rounds of validation. First, we iteratively withheld each of the 23 tiled viral genomes in the dataset during training and predicted their function as test sets, resulting in a 0.82–0.98 AUROC range (Figure 4-13), similar to previous results from our sequence-only MLP. We then carried out an external validation on unseen data from a previously published dataset of 168 characterized toehold switches[2] that had been collected under different experimental conditions. Our MLP models achieved an AUROC of 0.70, 0.81, and 0.79, when trained on rational features, one-hot sequence, and concatenated inputs, respectively (Figure 4-14). The improved performance observed when training the models directly on nucleotide sequence rather than thermodynamic features, even for an external dataset, suggest a competent degree of biological generalization and supports the value of modeling RNA synthetic biology tools using deep-learning and high-throughput datasets, removing the current assumptions of mechanistic rational parameters.
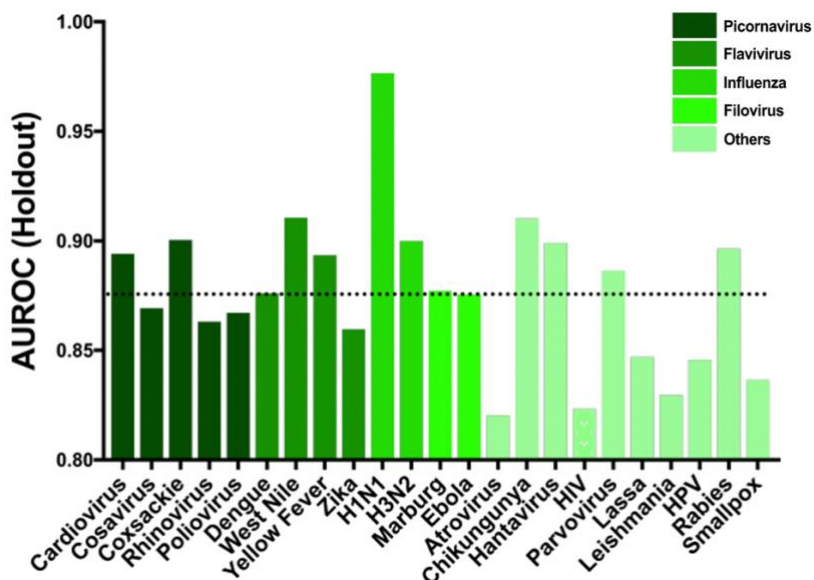
Figure 4-13. Multilayer perceptron model performance on withheld data. For each virus in our toehold switch dataset, variants targeting it were withheld, and an MLP model was trained with the remaining sequences classifying for ON/OFF ratio from a one-hot sequence input. The model performance was then evaluated on the switches of the withheld viral genome as a test set.
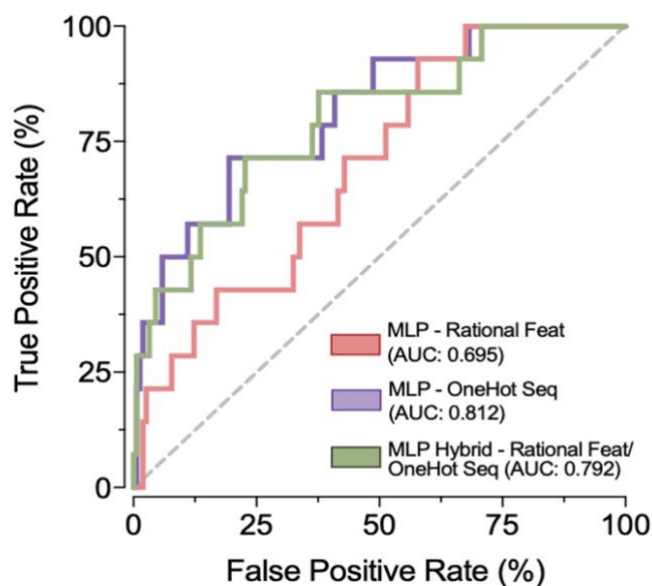
Figure 4-14. Multilayer perceptron model generalization on unseen data. ROC curves are plotted of pre-trained MLP classification models validated with an unseen external dataset from Green et al. [39] containing 168 switch variants.

4.5 Predictive Performance of Higher-Capacity Models

Having explored a baseline deep-learning architecture, we next sought to determine whether training our dataset on higher-capacity convolutional neural networks (CNN) and long short-term memory (LSTM) recurrent neural networks could increase our predictive ability. CNN and LSTM models have been applied to a variety of biological datasets in recent years, and have been cited as being particularly adept at recognizing motifs and long-range interactions in nucleotide sequence data [110,115-119,139-143]. We trained a CNN on a one-hot sequence input, an LSTM on a one-hot sequence input, and a CNN on a two-dimensional, one-hot complementarity map representation input (see the section below on VIS4Map for a description of the complementarity map). Upon evaluating both the R^2 and MAE in regression mode and the AUROC and AUPRC in classification mode for these models (Figures 4-15 and 4-16), we concluded that these neural network architectures did not lead to superior predictive models, as compared to the sequence-based, three-layer MLP described previously. In these cases, increased model capacity led to under- or over-fitting, potentially requiring additional training examples or improved fine-tuning to allow for effective trainings. This result is notable because many applications of deep learning to biological datasets report the performance of such more complex architectures as CNN or LSTM models, without first comparing them to simpler models such as the MLP [110,139-143]. It is therefore possible that in many such cases the deep learning approach taken is unnecessarily costly in terms of computation time, and likely serves simply to increase the impact of a publication rather than to provide actual improvements performance.

Figure 4-15. Multilayer perceptron model performs better than more complex model architectures with regression tasks. Performance metrics for convolutional neural networks (CNN) and long short-term memory (LSTM) networks trained on one-hot encoded toehold sequences, as well as a CNN trained on a two-dimensional, one-hot encoded sequence complementarity map. All models are compared to the previously reported MLPs trained on the 30 pre-calculated thermodynamic features and one-hot toehold sequences. Top: Box-and-whisker plots for R^2 accuracy between experimental and regression-based predictions (dotted line shows the best individual rational feature). Bottom: Box-and-whisker- plots for mean absolute error (MAE) between experimental and predicted values for these same models.
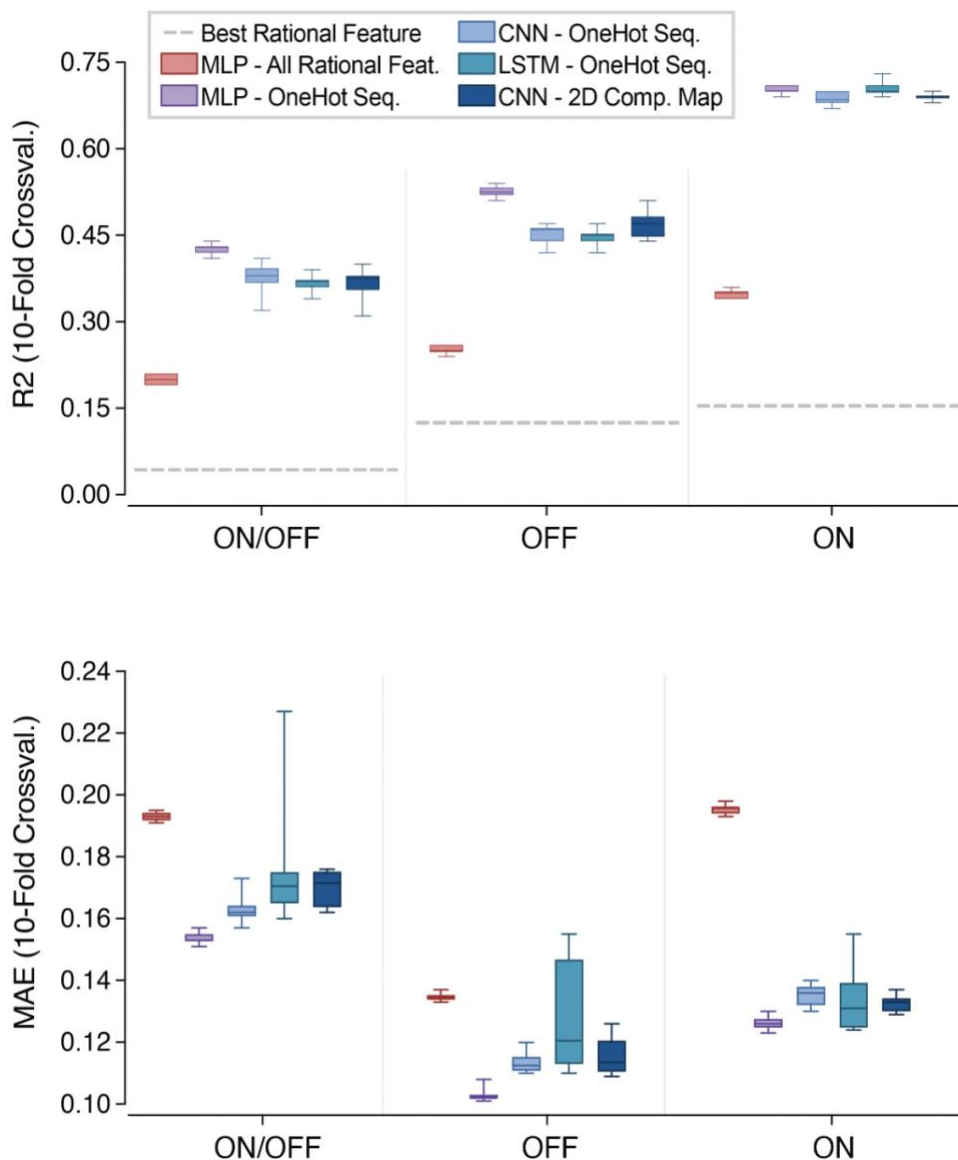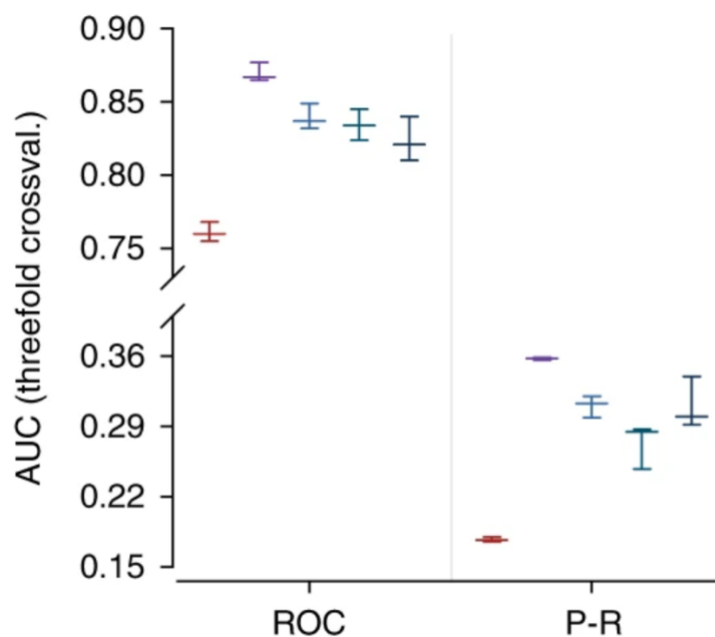
Figure 4-16. Multilayer perceptron model performs better than more complex model architectures with classification tasks. Performance metrics for convolutional neural networks (CNN) and long short-term memory (LSTM) networks trained on one-hot encoded toehold sequences, as well as a CNN trained on a two-dimensional, one-hot encoded sequence complementarity map. All models are compared to the previously reported MLPs trained on the 30 pre-calculated thermodynamic features and one-hot toehold sequences. Box- and-whisker plots for the area under the curve (AUC) of the receiver–operator curve (ROC) and the precision-recall curve (P–R) in classification-mode (using threefold cross-validation).

4.6 Improving Model Transparency by Visualizing RNA Secondary Structures

One significant challenge of using deep learning to predict biological function is the inherent difficulty in understanding learned patterns in a way that helps researchers to elucidate biological mechanisms underlying model predictions. Recent work has been developed to visualize sequence features by mapping learned convolutional filters to biologically relevant sequence motifs [144,145]. Additional methods have been established to address how models link

biological theory, including alternative network architectures [146], and the use of saliency maps [147,148], which reveal the regions of input that deep-learning models weigh most heavily and therefore pay the most attention to when making predictions. While saliency maps have been previously used to visualize model attention in one-hot representations of sequence data [110,115,147], such implementations focus only on the primary sequence and have not been developed to identify salient secondary-structure interactions. When used, we found that one-hot saliency maps gave only high-level information, for example that the higher region of the toehold stem had more impact than the bottom (Figure 4-17). Furthermore, prior work related to RNA secondary structure prediction using deep learning [149] has not utilized saliency techniques to highlight relevant secondary-structure regions that lead to improved function in RNA sensors. Instead, visualized representations have been constrained to predetermined structures based on thermodynamic model predictions [142,143], whose abstractions we have found cause significant information loss.
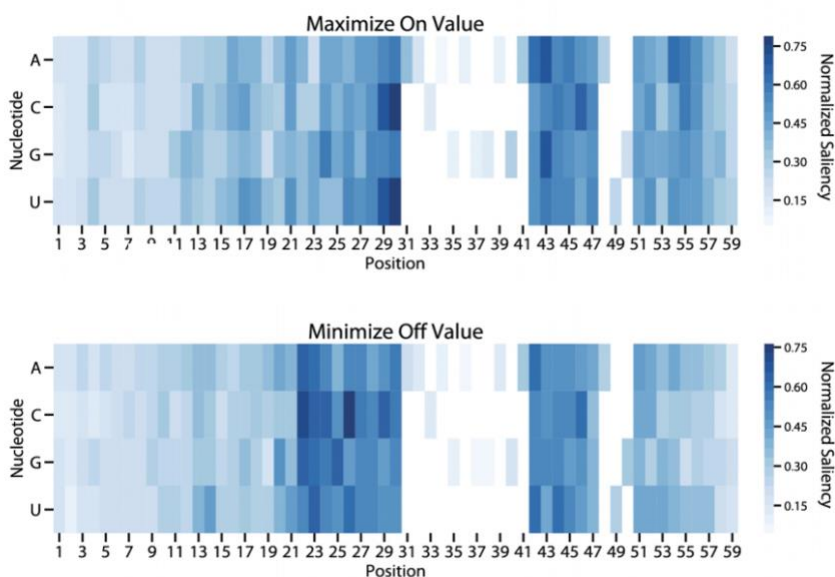

Figure 4-17. Saliency maps of one-hot encoded sequence inputs. An MLP model was trained on a one-hot sequence input representation of the toehold switch from the toehold to the bottom of the stem. A gradient-weighted activation mapping is performed to visualize saliency for high and low predicted ON and OFF values, respectively.

We sought to visualize important RNA secondary structures learned by our neural networks as it relates to biological function. To achieve this visualization, we trained a CNN on a two-dimensional nucleotide complementarity map representation (Figure 4-18) to allow for attention pattern visualization in this secondary-structure space. Each position in this complementarity map corresponds to the potential pair between two nucleotides, indicating its identity with a one-hot encoding (G–C, C–G, A–U, U–A, G–U, U–G, or a canonically unproductive pair). We hypothesized that by training deep networks on such a representation of RNA sequences, it would be possible to generate saliency maps revealing learned secondary structures as visually intuitive diagonal features. Importantly, because the complementarity map is unconstrained by a priori hypotheses of RNA folding (similar to our sequence-based MLP models), we anticipated this approach to be able to identify secondary structures that might be overlooked by commonly used thermodynamic and kinetic algorithms, such as NUPACK and Kinfold.

To first validate whether our visualization approach could capture any meaningful RNA structure features, we trained a CNN to predict NUPACK MFE values from a complementarity map representation of a randomly selected in silico RNA sequence dataset. Because NUPACK's calculated MFE is directly determined by a predicted RNA secondary structure, we anticipated that a CNN undergoing this training would likely pay attention to secondary-structure features, a situation that was confirmed through visualization of individual attention maps (Figure 4-19). Indeed, the saliency maps generated from a CNN trained on a complementarity map input contained primarily diagonal features that showed a statistically significant degree of agreement with the predicted MFE structures from which NUPACK based its MFE calculations (Figure 4-19). Additionally, we found that the use of a complementarity map input improved the CNN's

predictions of MFE from $R^2 = 0.6$ to $R^2 = 0.74$ compared with a one-hot sequence input (Figure 4-20). Hence, without prior knowledge of the algorithm or parameters NUPACK uses to calculate MFE, our CNN was able to learn similar abstractions as NUPACK, which we then used to intuitively visualize underlying relevant RNA secondary structures utilizing our complementarity map input representation. We named this approach for interpreting RNA deep-learning models Visualizing Secondary Structure Saliency Maps or VIS4Map.
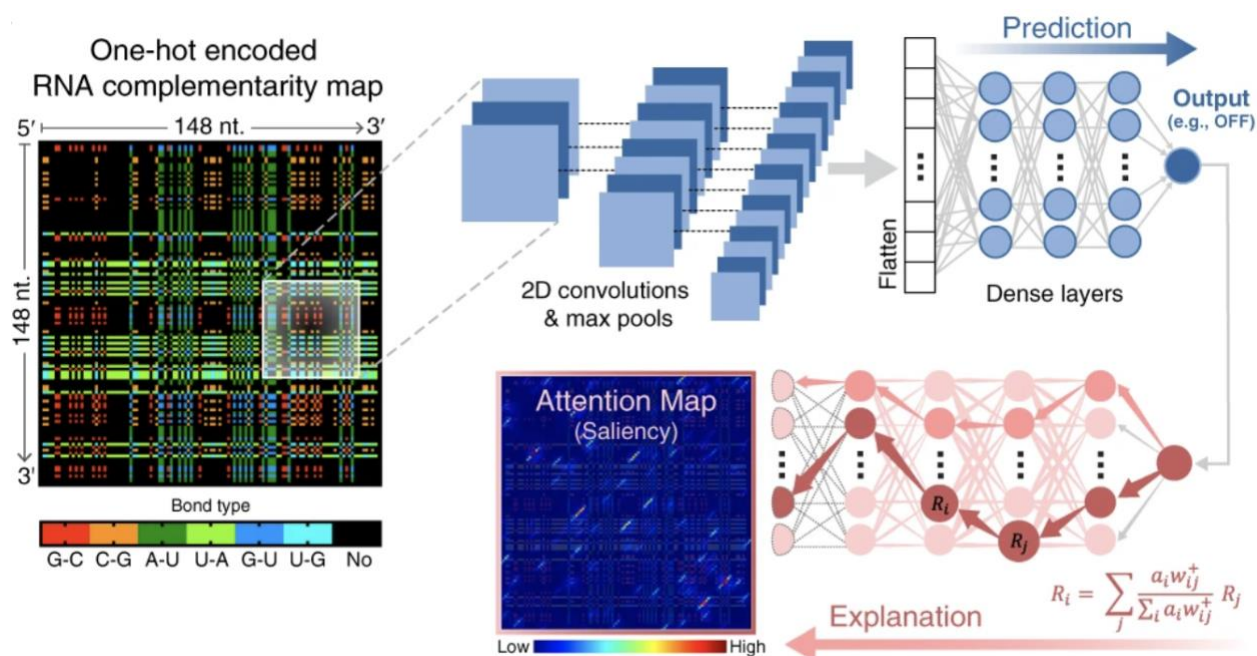
Figure 4-18. Summary of 2D base-complementarity maps and VIS4Map technique. A simplified schematic of the convolutional neural networks (CNN)-based architecture used to generate toehold functional predictions with network attention visualizations is shown. The system receives a one-hot encoded, two-dimensional (2D) sequence complementarity map as input. This matrix enumerates with a seven-dimensional vector at each position i,j the type of base pair that can be formed between nucleotides i and j in an RNA sequence. This input is fed into three 2D convolutional/max-pooling layers, a flattening step, and finally a set of dense layers for regression tasks. After output generation (e.g., OFF), a gradient-weighted activation mapping is performed to visualize activation maximization regions responsible for delivered predictions (VIS4Map).
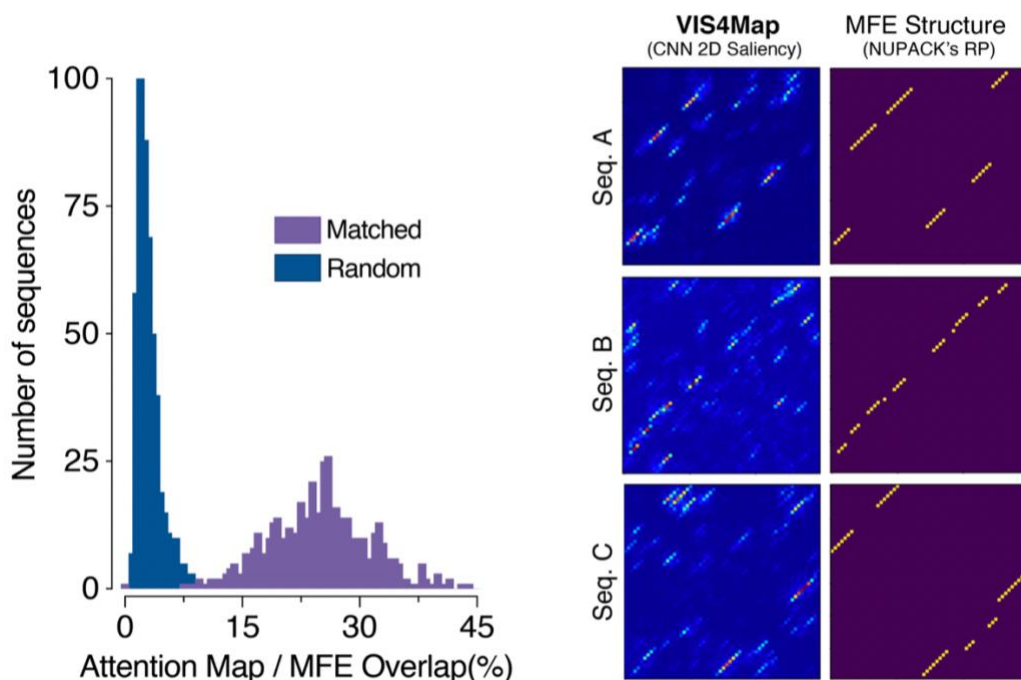
Figure 4-19. VIS4Map validation on in silico NUPACK-generated data. Left: Histograms of overlap between VIS4Maps generated by a CNN and MFE maps generated by NUPACK. When analyzed using 500 random test-set sequences, the distributions of correctly matched and randomly assigned maps are distinct with increased percentage overlap from matched samples as compared to unmatched. Right: Examples of saliency VIS4Maps compared with their corresponding MFE structures as predicted by NUPACK for three randomly selected 60-nt RNA sequences.
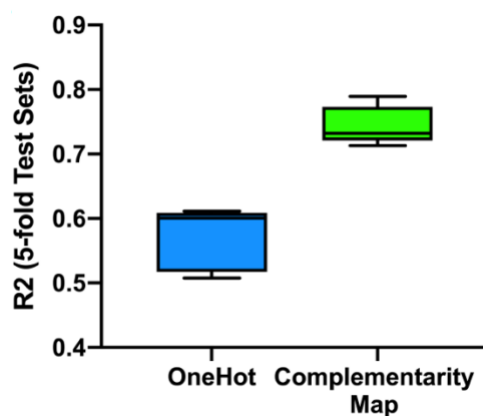
Figure 4-20. VIS4Map improves prediction accuracy on silico NUPACK-generated data. R^2 coefficients between NUPACK-calculated MFE values and the predictions of a CNN model trained either on a one-hot representation or a complementarity matrix representation of RNA sequences. Box and whisker plots summarize n=5 shuffled test sets.

Encouraged by our CNN's ability to elucidate putative RNA secondary-structure features directly from in silico-generated training data, we applied VIS4Map to our entire experimental toehold-switch function dataset. When trained on a complementarity map representation both in regression mode and classification mode (Figures 4-15 and 4-16), VIS4Map significantly outperformed an MLP trained on rational thermodynamic features. However, VIS4Map did not significantly outperform an MLP trained on a one-hot sequence input (as was the case when predicting NUPACK MFE). The failure of VIS4Map to improve predictions compared with a simpler three-layer MLP model likely results from over- or under-fitting of the higher-dimensional input, similar to the case of our other higher-capacity models (Figures 4-15 and 4-16). Encouragingly, nonetheless, we found that saliency maps produced by this CNN model displayed clear diagonal secondary-structure features (Figure 4-21). These structures appear to span from hybridization between the toehold and the ascending stem, to hybridization between the descending stem and the downstream linker. We confirmed the biological relevance of these features by averaging saliency maps and finding that the shared structures corresponded to the designed on-target structure of the switch hairpin (Figure 4-22). We further analyzed learned features outside of the designed equilibrium structure by sorting saliency maps using the toehold-switch OFF signal (Figure 4-23). We found that for leakier (high OFF) switches, the CNN identified a high degree of salient off-target secondary structures that could compete with the main hairpin stem and thereby expose the RBS, whereas for tight (low OFF) switches, the CNN identified fewer competing off-target secondary structures. In the context of general riboregulator behavior, these findings support the hypothesis that leaky expression from an RBS repressed by secondary structures can be caused by the misfolding of the repressive structure into less stable kinetic intermediate conformations [137,138] (Figure 4-23).
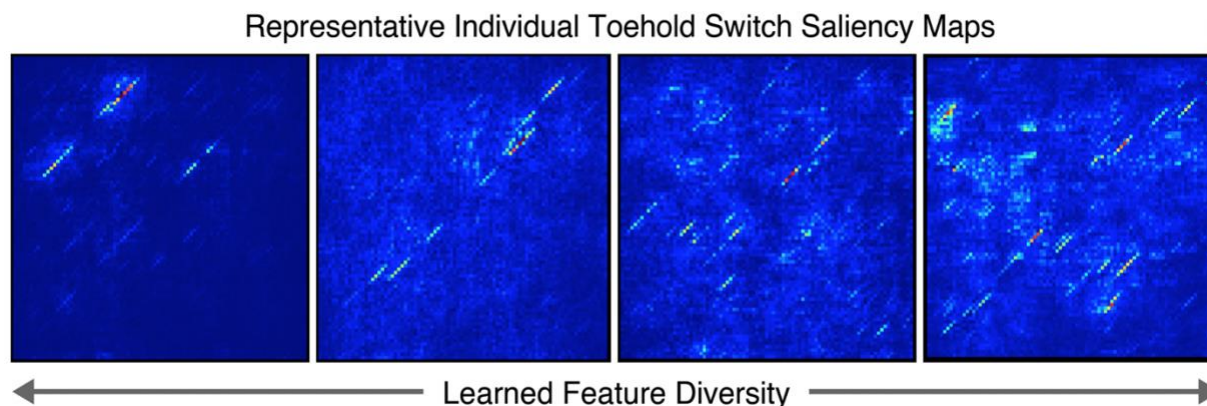
Figure 4-21. VIS4Map visualizes a diversity of learned secondary structure features from empirical toehold switch data. Four representative VIS4Map examples of randomly selected 118-nt RNA toehold-switch sequences from an OFF-predictive CNN model. Structures span interactions between the hairpin and non-hairpin regions, as well as on-target hairpin structures.
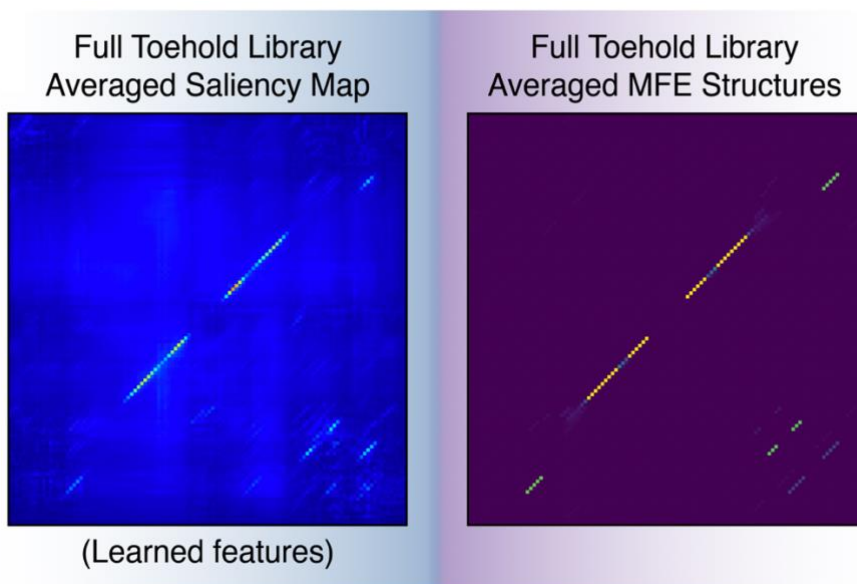
Figure 4-22. VIS4Map confirms model learning of on-target designed secondary structure. Averaged VIS4Maps of 10,125 randomly selected toehold-switch RNA sequences from our library test set processed with our OFF-predicting CNN model (left) and compared their corresponding averaged MFE maps obtained using NUPACK (right).

The fact that VIS4Map was able to identify both equilibrium and kinetically stable RNA secondary structures indicates a remarkable ability to uncover biologically relevant information, which in this case supports currently postulated hypotheses on prokaryotic translation initiation. Importantly, the identified secondary-structure features could not have been visualized using the one-hot sequence representation commonly associated with saliency maps [110,115]. These findings compound to the advantage of using sequence-only deep-learning approaches for analyzing RNA synthetic biology tools. Outside of toehold switches and other synthetic RNA systems, we anticipate VIS4Map will be broadly useful for the discovery of previously unknown equilibrium or kinetically stable structures contributing to RNA biology that are not predicted by current mechanistic RNA structure models.
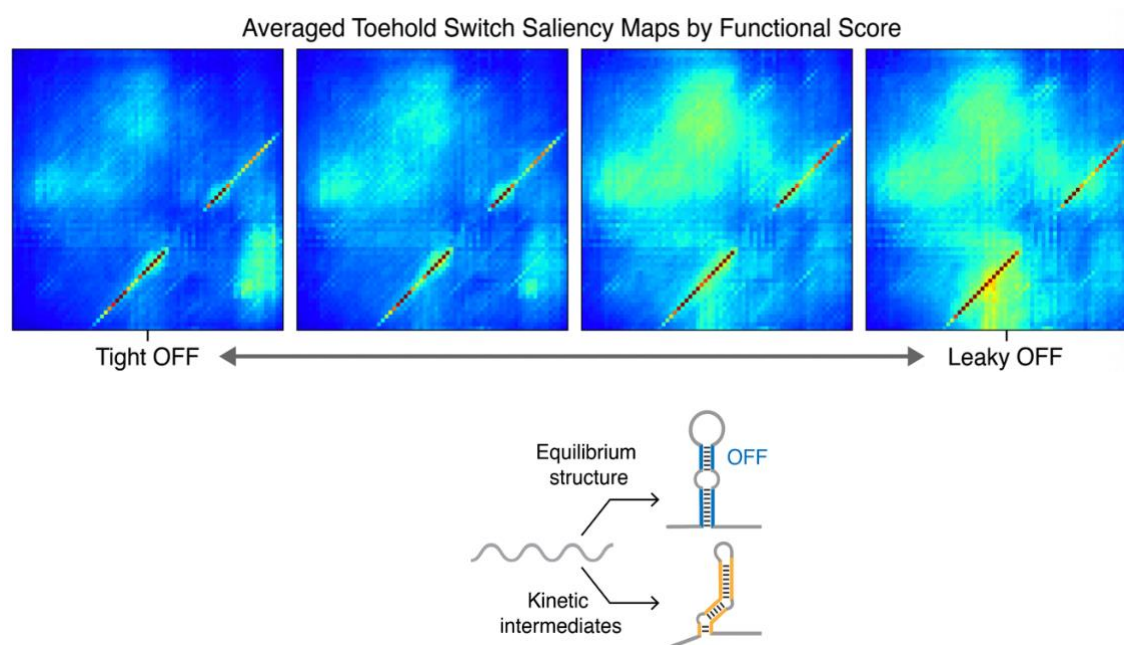
Figure 4-23. Averaged VIS4Maps sorted by quartile from lowest OFF to highest OFF. A noticeable increase in structures outside of the prominent equilibrium-designed switch hairpin structure appears to correlate with increased toehold leakiness. A toehold-switch schematic (bottom) is shown to denote how incorrectly folded and potentially weaker kinetically stable intermediate structures might compete with the correctly folded structure that is designed to be reached at equilibrium.

4.7 Concluding Thoughts on Deep Learning for RNA Synthetic Biology Design

Having produced a toehold-switch dataset ~100-fold larger than previously published as a model system for investigating synthetic RNA response elements [38,39,40,103-105], were able to conclusively demonstrate the benefits of using deep-learning methods that directly analyze sequence rather than relying on calculations from mechanistic thermodynamic and kinetic models. This approach resulted in a tenfold improvement in functional prediction R^2 over an ensemble of commonly used thermodynamic and kinetic features, and demonstrated robust biological generalization.

As with most work in RNA synthetic biology, all previous attempts to improve toehold-switch functionality had relied on the guidance of mechanistic thermodynamic modeling and low-throughput datasets [38,39,40,103-105,115]. Too frequently, rational design rules fail to give meaningful predictions of function for RNA-based synthetic systems. The results presented here suggest that the biological processes underlying RNA biology may be more complex than current state-of-the-art analyses take into account and that high-throughput DNA synthesis, sequencing, and deep-learning pipelines can be more effective for modeling said complexity. Combining improved predictions with enhanced understanding, our VIS4Map method further allowed us to visualize the equilibrium and kinetic secondary-structure features that our deep-learning models identified as important to the leakage of the switch OFF state. While secondary structures identified by NUPACK, Kinfold, and other rational mechanistic models are limited by predefined abstractions, which may cause significant information loss, our approach explored sequence space in an unrestricted manner and analyzed all possible RNA secondary structures. VIS4Map could prove useful for identifying complex secondary-structure information that might otherwise be ignored by simplified physical energetic models of RNA folding.

The dataset reported here also represents an extensive repository of characterized toehold switches, tiling the entire genomes of 23 pathogenic viruses of high clinical importance, as well as hundreds of human transcripts, including many that are differentially expressed in cancerous phenotypes [150]. The total cost of our flow-seq pipeline equated to ~\$0.08 per measurement, suggesting that the benefits of high-throughput design and assaying of RNA synthetic biology tools could be made widely accessible. We hope that this work will encourage the use of high-throughput data collection for the training of deep-learning systems, paired with more interpretable neural network architectures unrestricted by thermodynamic or kinetic secondary-structure models for improved prediction and insight generation in RNA synthetic biology.

Chapter 5

5.1 Motivating a Cold-Temperature Robust Isothermal Amplification Method

Among various discussed applications for so-called "next generation" NAT diagnostic technologies, the need for at-home point-of-care tests that can be marketed directly to consumers has become increasingly apparent due to the ongoing SARS-Cov2 pandemic [29,100,101,102]. While a handful of such products will soon reach the market at varying degrees of scale [25,26], significant room for improvement over available underlying technologies remains. Some of the major outlined challenges to creating an at-home POC NAT test have been addressed by others as well as by work presented in this thesis.

First, the need for expensive equipment has in part been alleviated by the simplification of heating devices thanks to isothermal amplification, the simplification or elimination of sample purification workflows, and the adoption of low-cost fluorescence readers or visual colorimetric outputs [2,3,4,9,10,11,16]. Second, the manufacturing cost of enzymatic and chemical reagents has continued to fall as scalability issues are addressed in industry, to the point where in most cases the major price components are device cost and overhead [13]. Third, the high sensitivity of NAT assays has been maintained as technology has shifted from cycling-based qPCR to isothermal amplification, with typical LOD's in the range of 1-10cp/uL [4,9,10,11,16,39,40]. The stringent requirement of sensitivity and specificity for many applications has been particularly apparent during the ongoing pandemic, as the false negative rate for detection of SARS-Cov2 infections can rise dramatically even with small loses in sensitivity (a 10-fold loss in sensitivity results in a ~12% loss in coverage due to low-viral load infections) (Figure 5-1). Fourth and lastly, while the turnaround time for NAT assays has broadly continued to fall, the need for rapid turnaround in an at-home POC NAT may not be as essential as previously considered. Compared to tests intended

to be used in a doctor's office or a clinic where long wait times can compromise patient outcomes [86,87,88], a test purchased and operated directly by the consumer can be used at a location and time of the user's choosing, meaning that longer turnaround times (potentially 1-2 hours) are less problematic, as long as device operation remains convenient.
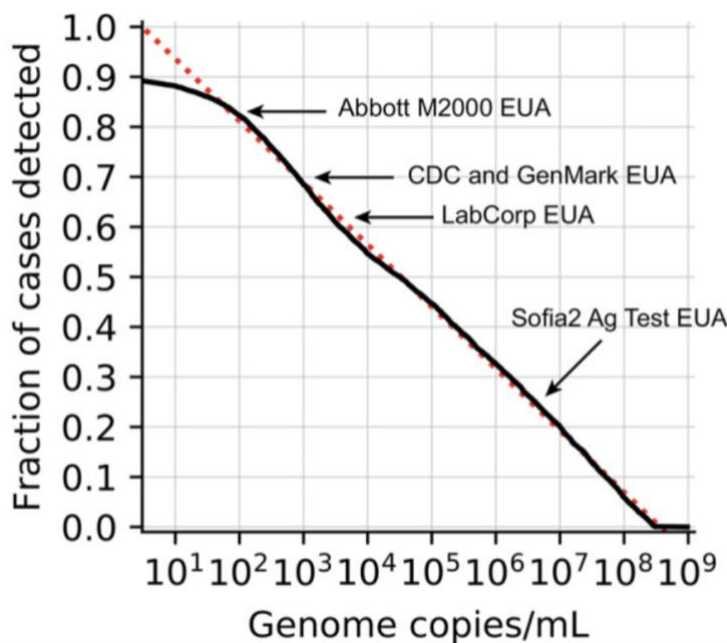
Figure 5-1. Viral load distribution and LoD. Cumulative histogram distribution of viral loads showing percent detected as a function of limit of detection - actual, solid line, and trend-line, dotted line. Slope of fitted line is approximately 12.5% loss in detected cases for every order of magnitude decrease in LOD. Figure from Arnaout et al. [151].

Nonetheless despite the aforementioned technological advancements, POC NAT technologies have not found widespread use in consumer homes in the same way that antigen-based rapid tests have. While some of this may be a result of the shorter timeframe during which such devices have been on the market, a larger obstacle is likely to be their continued high price tag [25,26]. While reagent costs for isothermal amplification have fallen to the range of $1-4, the total price tag for available at-home NAT diagnostics continues to fall in the range of $50-100, even

requiring a doctor's prescription in the case of the Lucira diagnostic, making them more a luxury item and precluding widespread use. The larger part of this price tag comes from the continued need for instrumentation, even if it has been minimized. This minimal instrumentation can be broken down into three components: i) Isothermal amplification typically requires a heating element with temperature control in the range of 37-65C, ii) most assay workflows require multiple liquid handling steps that must be automated within the device cartridge, and iii) many technologies require a fluorescent readout with a simplified fluorescence detector.

5.2 Prior art for Ambient Temperature Isothermal Amplification

In order to facilitate lowering assay costs by further simplifying instrumentation, I sought to develop a method for isothermal amplification of nucleic acid targets that could be robust to low temperatures. Whereas SHERLOCK/DETECTR has helped to partially address the second and third problems by extreme simplification of lysis steps with one-pot compatibility of all reaction components and the use of colorimetric LFA-based outputs [42,100,101,102], the first problem of eliminating the high-temperature incubation step has largely been unsolved. To my knowledge only two works have cited attomolar-level sensitivity of nucleic acid detection using isothermal amplification at ambient temperature: our own work detecting SARS-Cov2 with RPA in the wearable facemask (see chapter 3), and a CRISPR-Cas9 nickase enhanced strand-displacement amplification (SDA) scheme known as CRISDA [152].

In the case of our own wearable facemask, though we were able to detect a total of 500 copies of RNA (~mid-attomolar depending on input volume), the design process was challenging and in general the performance of RPA at room temperature in our hands fell far short of reaching

gold-standard sensitivity for most primer pairs. Robustness of RPA was similarly an issue, with day-to-day variability leaving significant room for improvement.

In the case of the CRISDA platform [152], only a single reaction is shown in the publication's supplementary information indicating the amplification of 220aM of dsDNA target at 25C (Supplementary Figure 4). The rest of the work describes reactions carried out at 37C, similar to the temperature optimum of RPA. The work also only describes a method for the amplification of dsDNA or ssDNA, not RNA targets. Furthermore, when we attempted to replicate this single experiment, the effective limit of detection in our hands of a ssDNA target turned out to be approximately three orders of magnitude worse than claimed, despite optimization of both primer concentration and single-stranded binding protein concentration (Figure 5-2). Although it is possible the discrepancy in results is due to our own error, we did not generally find SDA as described in the CRISDA paper to be robust at lower temperatures, faring similarly to RPA.
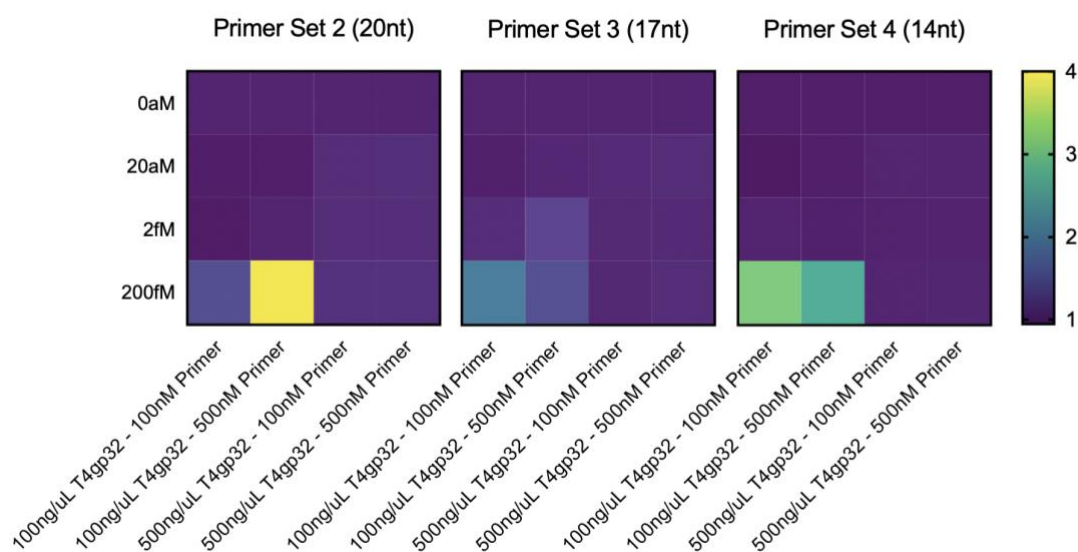
Figure 5-2. Room temperature SDA amplification using canonical primer design. Amplification reactions identical to those described in Zhou et al. [152] were caried out with primer length, primer concentration, and single-strand binding protein (T4gp32) concentration varied as indicated. Fold-change in signal from a second-step Cas12a detection reaction is shown at different concentrations of an input ssDNA trigger (0aM, 20aM, 2fM, 200fM).

5.3 Optimizing Strand Displacement Amplification for Low-Temperature Use

The strand-displacement amplification reaction relies on a three step process starting from a dsDNA product containing opposing nickase sites (Figure 5-3). First a nickase cleaves at a target site, leaving a free 3' OH end available. Second, a strand-displacing DNA polymerase initiates polymerization from the freed 3' OH and displaces the top strand of the dsDNA duplex from the nicked site onward. Third, a primer binds in the reverse complementary direction to the opposing nickase target site, and the DNA polymerase extends it, producing a copy of the initial dsDNA. In this fashion the dsDNA is exponentially amplified, with each nicking and extension event producing a ssDNA copy that can then be made double-stranded again.
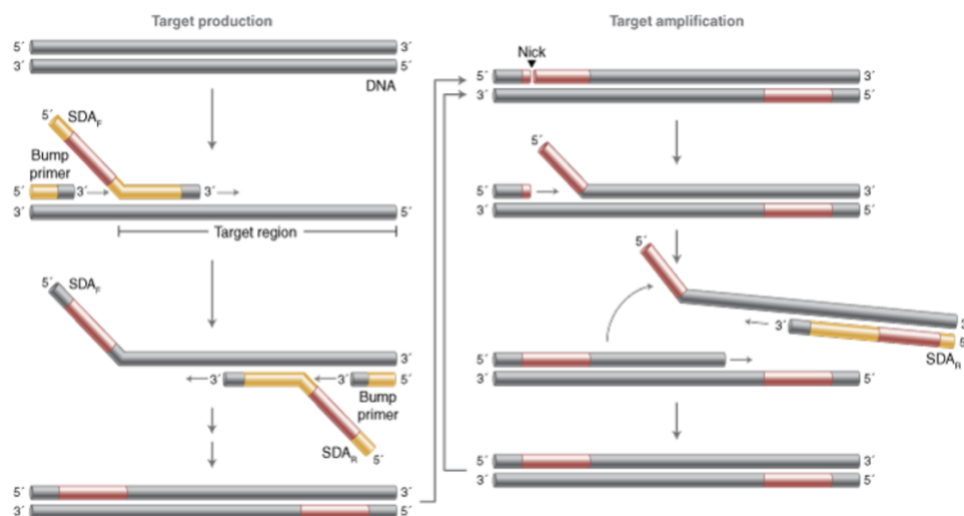
Figure 5-3. Strand displacement amplification reaction scheme. Starting at the top left from an initial target nucleic acid lacking nickase sites (dsDNA in this case, but ssDNA for most of this work and ideally also RNA), nickase sites in red are added to the target by the use of forwards and reverse bump primers. For non-dsDNA targets, different methods can be used to facilitate addition of nickase sites. Following through at the top right, a nickase site is cleaved by the nickase, and a DNA polymerase then extends and displaces the top strand of the duplex from the nick site. A reverse primer then binds to the opposing nickase site, producing again a dsDNA product containing two nickase sites (one that is partial but can still bind to a primer after being displaced). Figure is adapted from [153].

The failure mode of SDA reactions at lower temperatures was not immediately apparent. We initially suspected that the reaction speeds of the two enzymes involved had slowed too much to provide efficient amplification (a nickase and a strand-displacing polymerase, Nb.BbvCI and Klenow LF, respectively). However upon assessing the rate of nicking and strand displacement carried out at room temperature compared with higher temperatures, we did not believe this alone could explain the dramatic reduction in sensitivity observed (data not shown). Instead it became clear rather that somewhat reduced reaction kinetics were exacerbating the degree to which primer dimer products would form before sufficient levels of amplicon could be generated (Figure 5-4). While in theory the single-stranded binding protein (SSB, T4gp32) in the reaction is meant to prevent this by artificially raising the melting temperature of nucleic acids, we had observed that higher concentrations of SSB protein did not seem to improve the limit of detection (Figure 5-2).
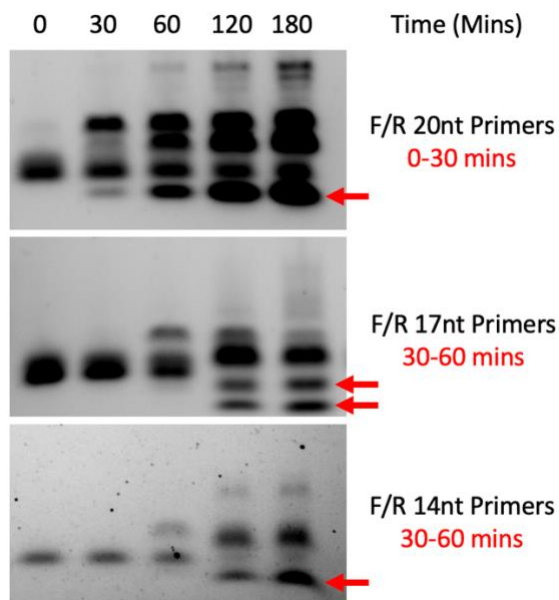

Figure 5-4. Timecourse of SDA primer dimer formation at room temperature. Agarose gels (4%) of SDA reactions stopped at indicated timepoints (by heat killing at 80C) are shown for three pairs of primers of varying length. Red arrows and red text indicate the timing of formation of nickase-extended products smaller than the initial starting primers.

To our knowledge, this issue inherent in the SDA reaction has not been overcome, and while at higher temperatures (typically 37-55C [152,153]) primer dimer formation is of lesser consequence, it seemed likely to be the key bottleneck to lower-temperature performance of the reaction. In particular I hypothesized that shortening the primer length would reduce or eliminate the fraction of the primer dimer pool containing correctly paired nickase sites that might induce exponential amplification of the dimer. The standard design for an SDA primer necessitates their length being in excess of 14nt, and typically in the range of 20-40nt (Figure 5-5) [152,153,154]. Instead a novel design that only has 9nt of hybridization region was found to produce only linear and no exponential primer dimer, ultimately taking much longer to be consumed (Figure 5-6).
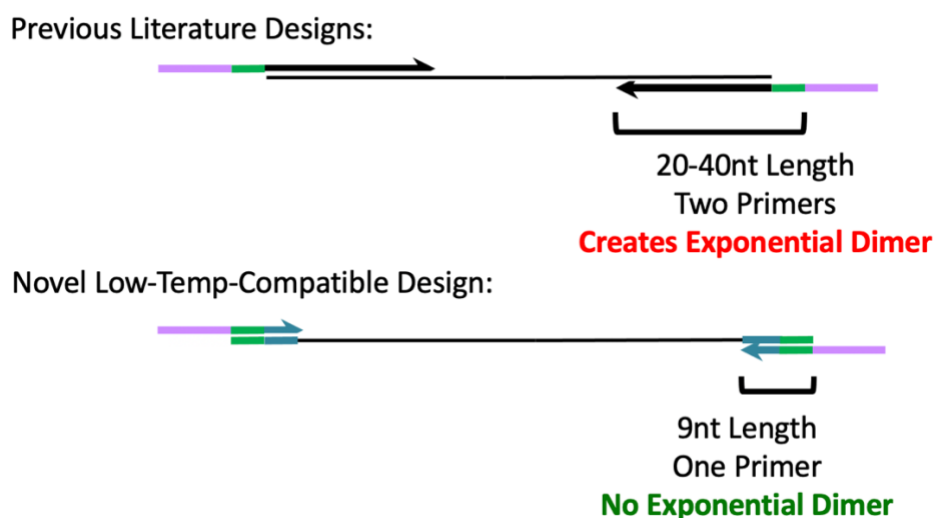
Figure 5-5. Novel SDA primer design. Top: Previous designs for strand displacement amplification described in the literature employ two primers with target-specific regions (black) as well as nickase sites (green) and stabilization regions (pink). Bottom: The novel primer design described in this work uses a single primer for both forwards and reverse directions. This is made possible by incorporating an identical primer-hybridization sequence (blue) when the nickase site (green) is added to the target by a variety of means, including the use of bump primers, ligase splints, etc. The removal of the target-specific region and the consolidation of two primers into one allows for very short primers to be used.
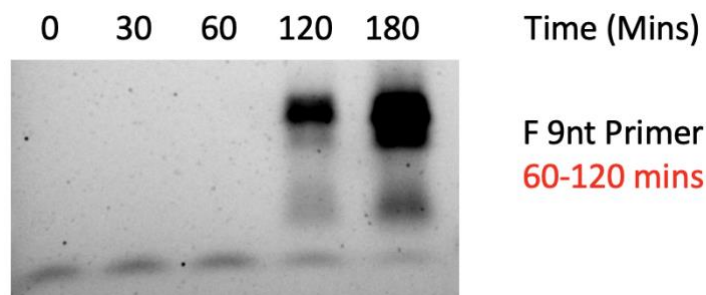
Figure 5-6. Timecourse of reduced-length SDA primer dimer formation at room temperature. Agarose gels (4%) of SDA reactions stopped at indicated timepoints (by heat killing at 80C) are shown for a single shortened 9nt forwards primer. Red text indicate the timing of formation of primer-dimer, which is notably never nickase-generated (as those are smaller).
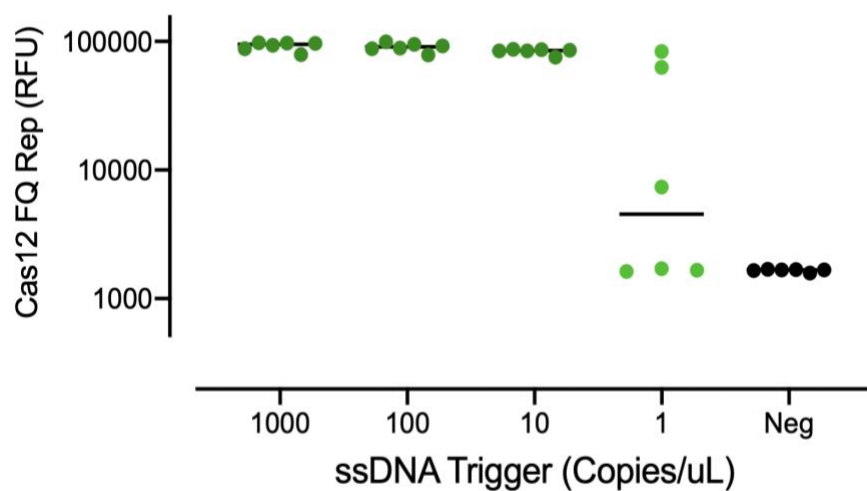
Figure 5-7. Ambient-temperature detection of ssDNA targets using Cas12a and dimer-resistant primers. An ssDNA target was amplified for two hours at room temperature using novel SDA primer design and optimized reaction conditions. After amplification, a gRNA:Cas12a complex targeting a PAM-containing site on the amplicon was added to the reactions and allowed to cleave a fluorophore-quencher reporter for thirty minutes. Individual values and mean of six replicates are shown (n=6).

Using this optimized reaction condition, up to single-digit copies of a ssDNA target were detectable within a two hour reaction time using Cas12a as an amplicon detector. This low-attomolar LOD (~5aM) surpassed even the sensitivity of the covid facemask described previously in this work, and was a significant improvement over the CRISDA platform [152] both as the authors originally reported and particularly as carried out in our hands.

5.4 Concluding Thoughts on Low-Temperature SDA for POC NAT Diagnostics

The challenge of adapting isothermal amplification technologies to lower ambient temperatures has largely gone unaddressed since their first design iterations two decades ago. Here I have shown that while lower catalytic activity of enzymes at ambient temperature compared to 37-55C temperature range may have a deleterious effect on amplification, the true bottleneck for sustaining amplification at such temperatures is the production of primer dimer. If this can be avoided through modification of primer design, then I have shown here amplification can be sustained even at low temperatures with limits of detection approaching single-digit copy number. Such gold-standard sensitivity is comparable to qPCR and has, to my knowledge, not been shown for an ambient-temperature isothermal amplification technique.

Looking forward, further innovations must be incorporated for this low-temperature SDA approach to have widespread utility. Importantly, the analyte I have detected is an ssDNA target, and in most cases of clinical diagnosis the target of interest will be dsDNA or RNA. Of particular interest due to the ongoing pandemic is the detection of RNA to diagnose the presence of the RNA-genome virus SARS-Cov2. In order to convert an RNA target into an amplifiable ssDNA molecule, a pair of RT primers with appropriate partial nickase sites as tails would need to be used along with bump primers, as outlined in Figure 5-3. Another key innovation would be the collapse

of the current two-pot system into a one-pot system, such that the Cas12a detection and SDA amplification steps can function in the same solution [42]. This crucial improvement would simplify liquid-handling on board the eventual diagnostic cartridge, increasing robustness, speed, and ease-of-use. Combined with existing LFA-based colorimetric outputs for Cas12 detection [42] and simplified sample processing techniques outlined in this work and elsewhere [100,101,102] (particularly for the covid facemask in chapter three) a fully instrument-free "sample-in-results-out" device could be constructed that would require no electronics. To our knowledge, no such product is currently on the market or in preparations to be launched.

An affordable, user-friendly, gold-standard POC NAT technology would dramatically expand access to informative clinical diagnosis. The ongoing SARS-Cov2 pandemic has made clear the need for rapid turnaround in testing, and while reducing the time a test takes to complete can aid in this goal, bringing testing directly to the consumer in order to facilitate the integration of testing into varied personal schedules may arguably have a greater impact [84-88]. I am therefore optimistic that with the aid of synthetic biology techniques and advanced high-throughput prototyping as described in this thesis, a variety of next-generation nucleic acid diagnostic tools will become available on the market in the coming decade.

Bibliography

1. Dobson, Mark G., Paul Galvin, and David E. Barton. "Emerging technologies for point-of-care genetic testing." Expert Review of Molecular Diagnostics 7.4 (2007): 359-70. Web.

2. O'Kane, Maurice J. "Point of Care Testing—Current and Emerging Quality Perspectives." Point of Care: The Journal of Near-Patient Testing & Technology 13.1 (2014): 1-5. Web.

3. Shaw, Julie L.v. "Practical challenges related to point of care testing." Practical Laboratory Medicine 4 (2016): 22-29. Web.
<http://www.sciencedirect.com/science/article/pii/S2352551715300056>.

4. Everitt, Micaela L., et al. "A Critical Review of Point-of-Care Diagnostic Technologies to Combat Viral Pandemics." Analytica Chimica Acta, vol. 1146, 2021, pp. 184–199., doi:10.1016/j.aca.2020.10.009.

5. Roberts, Teri, and Cara Kosack. "Prices analysis of essential HIV CD4 and VL technologies." MSF Diagnostics Network. World Health Organization, 2014. Web.
<http://apps.who.int/hiv/events/MSF_AMDS.pdf>.

6. Philippidis, Alex. "Challenges Loom for Point-of-Care Diagnostics | GEN Genetic Engineering & Biotechnology News - Biotech from Bench to Business." GEN. Genetic Engineering & Biotechnology, 2016. Web. <http://www.genengnews.com/gen-exclusives/challenges-loom-for-point-of-care-diagnostics/77900690>.

7. "FAQs: Global Poverty Line Update." World Bank. N.p., 2016. Web.
<http://www.worldbank.org/en/topic/poverty/brief/global-poverty-line-faq>.

8. Yetisen, Ali Kemal et al. "Paper-based microfluidic point-of-care diagnostic devices". Lab on a Chip, 13, 2210-2251, (2013). Web.
<https://pubs.rsc.org/en/content/articlelanding/2013/lc/c3lc50169h#!divAbstract>

9. Zanoli, Laura, and Giuseppe Spoto. "Isothermal Amplification Methods for the Detection of Nucleic Acids in Microfluidic Devices." Biosensors, vol. 3, no. 1, 2012, pp. 18–43., doi:10.3390/bios3010018.

10. Crannell, Zachary Austin, et al. "Equipment-Free Incubation of Recombinase Polymerase Amplification Reactions Using Body Heat." PLoS ONE, vol. 9, no. 11, May 2014, doi:10.1371/journal.pone.0112146.

11. Moore, Matthew D, and Lee-Ann Jaykus. "Recombinase Polymerase Amplification: a Promising Point-of-Care Detection Method for Enteric Viruses." Future Virology, vol. 12, no. 8, 2017, pp. 421–429., doi:10.2217/fvl-2017-0034.

12. Obermeyer, Ziad, et al. "Individual Differences in Normal Body Temperature: Longitudinal Big Data Analysis of Patient Records." Bmj, 2017, doi:10.1136/bmj.j5468.

13. "How Low Can We Go? Pricing for HIV Viral Load Testing in Low- and Middle-Income Countries." Médecins Sans Frontières Access Campaign . MSF, 2014. Web. <http://www.msf.org/sites/msf.org/files/how_low_can_we_go_vl_pricing_brief.pdf>.

14. Cook, Jackie, et al. "Loop-Mediated isothermal amplification (LAMP) for point-of-Care detection of asymptomatic low-Density malaria parasite carriers in Zanzibar." Malaria Journal, vol. 14, no. 1, 2015, p. 43., doi:10.1186/s12936-015-0573-y.

15. Mohon, Abu Naser, et al. "A novel SNP-LAMP assay for detection of artemisinin-Resistant Plasmodium falciparum malaria." Open Forum Infectious Diseases, Sept. 2018, doi:10.1093/ofid/ofy011.

16. Howson, Emma L.a., et al. "Defining the Relative Performance of Isothermal Assays That Can Be Used for Rapid and Sensitive Detection of Foot-and-Mouth Disease Virus." Journal of Virological Methods, vol. 249, 2017, pp. 102–110., doi:10.1016/j.jviromet.2017.08.013.

17. Layden, J. E. et al. "Hepatitis C in Sub-Saharan Africa: Urgent Need for Attention." Open Forum Infectious Diseases 1.2 (2014): n. pag. Web. <https://academic.oup.com/ofid/article/doi/10.1093/ofid/ofu065/1466359/Hepatitis-C-in-Sub-Saharan-Africa-Urgent-Need-for>.

18. Masimba, Pax et al. "Prevalence of Drug-Resistance Mutations and HIV-1 Subtypes in a HIV-1 Infected COHORT in rural Tanzania." AIDS Research and Human Retroviruses (2013): 130627192948006. Web. <https://www.ncbi.nlm.nih.gov/pubmed/23806135>.

19. Coll, Francesc et al. "Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences." Genome Medicine 7.1 (2015): n. pag. Web. <https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-015-0164-0>.

20. "Xpert MTB/RIF assay for the diagnosis of pulmonary and extrapulmonary TB in adults and children." WHO. World Health Organization, 2013. Web. <http://www.who.int/tb/publications/xpert-mtb-rif-assay-diagnosis-policy-update/en/>.

21. Gamell, Anna et al. "A Case Series of Acquired Drug Resistance-Associated Mutations in Human Immunodeficiency Virus-Infected Children: An Emerging Public Health Concern in Rural Africa: Table 1." Open Forum Infectious Diseases 3.1 (2015): n. pag. Web. <https://academic.oup.com/ofid/article-lookup/doi/10.1093/ofid/ofv199>.

22. Hawkins, Claudia et al. "HIV virological failure and drug resistance in a cohort of Tanzanian HIV-infected adults." Journal of Antimicrobial Chemotherapy 71.7 (2016): 1966-974. Web. <https://academic.oup.com/jac/article-lookup/doi/10.1093/jac/dkw051>.

23. Brady, Oliver J. et al. "Refining the Global Spatial Limits of Dengue Virus Transmission by Evidence-Based Consensus." PLoS Neglected Tropical Diseases 6.8 (2012): n. pag. Web. <http://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0001760>.

24. Haug, Charlotte J., Marie Paule Kieny, and Bernadette Murgue. "The Zika Challenge." New England Journal of Medicine 374.19 (2016): 1801-803. Web. <http://www.nejm.org/doi/full/10.1056/NEJMp1603734#t=article>.

25. Regalado, Antonio. "We Reviewed Three at-Home Covid-19 Tests. Here's What Happened." MIT Technology Review, MIT Technology Review, 4 May 2021, www.technologyreview.com/2021/05/04/1024450/at-home-covid-test-review-accuracy-binaxnow-lucira-ellume/.

26. US Food and Drug Administration, Lucira COVID-19 All-In-One Test Kit Emergency Use Authorization Letter of Authorization., (available at https://www.lucirahealth.com/wp-content/uploads/2020/11/Lucira-EUA-LOA.pdf).

27. H. C. Yang, C. H. Chen, J. H. Wang, H. C. Liao, C. T. Yang, C. W. Chen, Y. C. Lin, C. H. Kao, M. Y. J. Lu, J. C. Liao, Analysis of genomic distributions of SARS-CoV-2 reveals a dominant strain type with strong allelic associations. *Proc Natl Acad Sci U S A*. **117**, 30679–30686 (2020).

28. Public Health England, Investigation of novel SARS-COV-2 variant, (available at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/947048/Technical_Briefing_VOC_SH_NJL2_SH2.pdf).

29. CDC, Coronavirus Disease 2019 (COVID-19). *Centers for Disease Control and Prevention* (2020), (available at https://www.cdc.gov/coronavirus/2019-ncov/more/science-and-research/scientific-brief-emerging-variants.html).

30. Lineage B.1.525, (available at https://cov-lineages.org/global_report_B.1.525.html#table2link).

31. P. Wang, L. Liu, S. Iketani, Y. Luo, Y. Guo, M. Wang, J. Yu, B. Zhang, P. D. Kwong, B. S. Graham, J. R. Mascola, J. Y. Chang, M. T. Yin, M. Sobieszczyk, L. Shapiro, Z. Sheng, M. S. Nair, Y. Huang, D. D. Ho, Increased Resistance of SARS-CoV-2 Variants B.1.351 and B.1.1.7 to Antibody Neutralization, *bioRxiv*, in press.

32. F. Tian, B. Tong, L. Sun, S. Shi, B. Zheng, Z. Wang, X. Dong, P. Zheng, Mutation N501Y in RBD of Spike Protein Strengthens the Interaction between COVID-19 and its Receptor ACE2, *bioRxiv*, in press.

33. Z. Liu, L. A. VanBlargan, L.-M. Bloyet, P. W. Rothlauf, R. E. Chen, S. Stumpf, H. Zhao, J. M. Errico, E. S. Theel, M. J. Liebeskind, B. Alford, W. J. Buchser, A. H. Ellebedy, D. H. Fremont, M. S. Diamond, S. P. J. Whelan, Landscape analysis of escape variants identifies SARS-CoV-2 spike mutations that attenuate monoclonal and serum antibody neutralization,

bioRxiv, in press.

34. C. K. Wibmer, F. Ayres, T. Hermanus, M. Madzivhandila, P. Kgagudi, B. E. Lambson, M. Vermeulen, K. van den Berg, T. Rossouw, M. Boswell, V. Ueckermann, S. Meiring, A. von Gottberg, C. Cohen, L. Morris, J. N. Bhiman, P. L. Moore, SARS-CoV-2 501Y.V2 escapes neutralization by South African COVID-19 donor plasma, *bioRxiv*, in press.

35. Z. Liu, L. A. VanBlargan, L.-M. Bloyet, P. W. Rothlauf, R. E. Chen, S. Stumpf, H. Zhao, J. M. Errico, E. S. Theel, M. J. Liebeskind, B. Alford, W. J. Buchser, A. H. Ellebedy, D. H. Fremont, M. S. Diamond, S. P. J. Whelan, Identification of SARS-CoV-2 spike mutations that attenuate monoclonal and serum antibody neutralization. *Cell Host Microbe* (2021)

36. A. J. Greaney, T. N. Starr, P. Gilchuk, S. J. Zost, E. Binshtein, A. N. Loes, S. K. Hilton, J. Huddleston, R. Eguia, K. H. D. Crawford, A. S. Dingens, R. S. Nargi, R. E. Sutton, N. Suryadevara, P. W. Rothlauf, Z. Liu, S. P. J. Whelan, R. H. Carnahan, J. E. Crowe, J. D. Bloom, Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain that Escape Antibody Recognition. *Cell Host & Microbe*. **29**, 44-57.e9 (2021).

37. Gootenberg, Jonathan S., et al. "Nucleic acid detection with CRISPR-Cas13a/C2c2." Science, vol. 356, no. 6336, 2017, pp. 438–442., doi:10.1126/science.aam9321.

38. Pardee, Keith, Alexander A. Green, Melissa K. Takahashi, Dana Braff, Guillaume Lambert, Jeong Wook Lee, Tom Ferrante, Duo Ma, Nina Donghia, Melina Fan, Nichole M. Daringer, Irene Bosch, Dawn M. Dudley, David H. O'Connor, Lee Gehrke, and James J. Collins. "Rapid, Low-Cost Detection of Zika Virus Using Programmable Biomolecular Components." Cell 165.5 (2016): 1255-266. Web.

39. Green, Alexander A., Pamela A. Silver, James J. Collins, and Peng Yin. "Toehold Switches: De-Novo-Designed Regulators of Gene Expression." Cell 159.4 (2014): 925-39. Web.

40. Pardee, Keith, Alexander A. Green, Tom Ferrante, D. Ewen Cameron, Ajay Daleykeyser, Peng Yin, and James J. Collins. "Paper-Based Synthetic Gene Networks." Cell 159.4 (2014): 940-54. Web.

41. Cai, Qi et al. "A simplified and robust protocol for immunoglobulin expression in Escherichia coli cell-free protein synthesis systems." Biotechnology Progress 31.3 (2015): 823-831. Web.

42. Gootenberg, Jonathan S., et al. "Multiplexed and Portable Nucleic Acid Detection Platform with Cas13, Cas12a, and Csm6." Science, vol. 360, no. 6387, (2018), pp. 439–444., doi:10.1126/science.aaq0179.

43. East-Seletsky, Alexandra, et al. "Two Distinct RNase Activities of CRISPR-C2c2 Enable Guide-RNA Processing and RNA Detection." Nature, vol. 538, no. 7624, (2016), pp. 270–273. doi:10.1038/nature19802.

44. Abudayyeh, Omar et al. "C2C2 is a single-component programmable rna-guided rna-targeting CRISPR effector." Science. (2016), 353, 6299. Web.

45. J. S. Chen, E. Ma, L. B. Harrington, M. Da Costa, X. Tian, J. M. Palefsky, J. A. Doudna, CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. *Science*. **360**, 436–439 (2018).

46. Salis, Howard M, et al. "Automated Design of Synthetic Ribosome Binding Sites to Control Protein Expression." Nature Biotechnology, vol. 27, no. 10, 2009, pp. 946–950., doi:10.1038/nbt.1568.

47. Udagawa, Tsuyoshi, et al. "Evidence for the Translation Initiation of Leaderless MRNAs by the Intact 70 S Ribosome without Its Dissociation into Subunits in Eubacteria." Journal of Biological Chemistry, vol. 279, no. 10, Nov. 2003, pp. 8539–8546., doi:10.1074/jbc.m308784200.

48. Brock, J. E., et al. "Ribosomes Bind Leaderless MRNA in Escherichia Coli through Recognition of Their 5-Terminal AUG." Rna, vol. 14, no. 10, 2008, pp. 2159–2169., doi:10.1261/rna.1089208.

49. Yamamoto, Hiroshi, et al. "70S-Scanning Initiation Is a Novel and Frequent Initiation Mode of Ribosomal Translation in Bacteria." Proceedings of the National Academy of Sciences, vol. 113, no. 9, 2016, doi:10.1073/pnas.1524554113.

50. Krishnan, K. M., et al. "Proximity of the Start Codon to a Leaderless MRNAs 5 Terminus Is a Strong Positive Determinant of Ribosome Binding and Expression in Escherichia Coli." Journal of Bacteriology, vol. 192, no. 24, 2010, pp. 6482–6485., doi:10.1128/jb.00756-10.

51. Duval, Mélodie, et al. "Multiple Ways to Regulate Translation Initiation in Bacteria: Mechanisms, Regulatory Circuits, Dynamics." Biochimie, vol. 114, 2015, pp. 18–29., doi:10.1016/j.biochi.2015.03.007.

52. Milón, Pohl, and Marina V. Rodnina. "Kinetic Control of Translation Initiation in Bacteria." Critical Reviews in Biochemistry and Molecular Biology, vol. 47, no. 4, 2012, pp. 334–348., doi:10.3109/10409238.2012.678284.

53. Studer, Sean M., and Simpson Joseph. "Unfolding of MRNA Secondary Structure by the Bacterial Translation Initiation Complex." Molecular Cell, vol. 22, no. 1, 2006, pp. 105–115., doi:10.1016/j.molcel.2006.02.014.

54. Dixon, Andrew S., et al. "NanoLuc Complementation Reporter Optimized for Accurate Measurement of Protein Interactions in Cells." ACS Chemical Biology, vol. 11, no. 2, 2015, pp. 400–408., doi:10.1021/acschembio.5b00753.

55. Stevens, Adam J., et al. "A Promiscuous Split Intein with Expanded Protein Engineering Applications." Proceedings of the National Academy of Sciences, vol. 114, no. 32, 2017, pp. 8538–8543., doi:10.1073/pnas.1701083114.

56. Fozouni, Parinaz, et al. "Amplification-Free Detection of SARS-CoV-2 with CRISPR-Cas13a and Mobile Phone Microscopy." Cell, vol. 184, no. 2, 2021, doi:10.1016/j.cell.2020.12.001.

57. Liu, Xinyue, et al. "Stretchable Living Materials and Devices with Hydrogel–Elastomer Hybrids Hosting Programmed Cells." Proceedings of the National Academy of Sciences, vol. 114, no. 9, 2017, pp. 2200–2205., doi:10.1073/pnas.1618307114.

58. Moser, Felix, et al. "Light-Controlled, High-Resolution Patterning of Living Engineered Bacteria Onto Textiles, Ceramics, and Plastic." Advanced Functional Materials, vol. 29, no. 30, 2019, p. 1901788., doi:10.1002/adfm.201901788.

59. Kim, J., Campbell, A.S., de Avila, B.E. & Wang, J. Wearable biosensors for healthcare monitoring. Nat Biotechnol 37, 389-406 (2019).

60. Khalil, A.S. & Collins, J.J. Synthetic biology: applications come of age. Nat Rev Genet 11, 421 367-379 (2010).

61. Tao, X. Smart fibres, fabrics and clothing: fundamentals and applications. (Elsevier, 2001).

62. Liu, X. et al. Stretchable living materials and devices with hydrogel-elastomer hybrids hosting programmed cells. Proc Natl Acad Sci U S A 114, 2200-2205 (2017).

63. Liu, X. et al. 3D printing of living responsive materials and devices. Adv Mater 30 (2018).

64. Wang, W. et al. Harnessing the hygroscopic and biofluorescent behaviors of genetically tractable microbial cells to design biohybrid wearables. Sci Adv 3, e1601984 (2017).

65. Slomovic, S., Pardee, K. & Collins, J.J. Synthetic biology devices for in vitro and in vivo diagnostics. Proc Natl Acad Sci U S A 112, 14429-14435 (2015).

66. English, M.A. et al. Programmable CRISPR-responsive smart materials. Science, 780- 785 (2019).

67. Pardee, K. et al. Portable, on-demand biomolecular manufacturing. Cell 167, 248-259 e212 442 (2016).

68. Huang, A. et al. BioBits Explorer: A modular synthetic biology education kit. Sci Adv 4, eaat5105 (2018).

69. Stark, J.C. et al. BioBits Bright: A fluorescent synthetic biology education kit. Sci Adv 4, eaat5107 (2018).

70. Stark, J.C. et al. BioBits Health: Classroom activities exploring engineering, biology, and human health with fluorescent readouts. ACS Synth Biol 8, 1001-1009 (2019).

71. Choi, J.K., Miki, K., Sagawa, S. & Shiraki, K. Evaluation of mean skin temperature formulas by infrared thermography. Int J Biometeorol 41, 68-75 (1997).

72. Filonov, G.S., Kam, C.W., Song, W. & Jaffrey, S.R. In-gel imaging of RNA processing using broccoli reveals optimal aptamer expression strategies. Chem Biol 22, 649-660 (2015).

73. Hall, M.P. et al. Engineered luciferase reporter from a deep sea shrimp utilizing a novel imidazopyrazinone substrate. ACS Chem Biol 7, 1848-1857 (2012).

74. Harrington, L.B. et al. Programmed DNA destruction by miniature CRISPR-Cas14 enzymes. Science 362, 839-842 (2018).

75. Li, S.-Y. et al. CRISPR-Cas12a has both cis-and trans-cleavage activities on single-stranded DNA. Cell Res 28, (2018).

76. Piepenburg, O., Williams, C.H., Stemple, D.L. & Armes, N.A. DNA detection using recombination proteins. PLoS Biol 4, e204 (2006).

77. Lee, A.S. et al. Methicillin-resistant Staphylococcus aureus. Nat Rev Dis Primers 4, 18033 (2018).

78. Falugi, F., Kim, H.K., Missiakas, D.M. & Schneewind, O. Role of protein A in the evasion of host adaptive immune responses by Staphylococcus aureus. MBio 4, e00575-00513 (2013).

79. Westh, H., Hougaard, D.M., Vuust, J. & Rosdahl, V.T. Prevalence of erm gene classes in erythromycin-resistant Staphylococcus aureus strains isolated between 1959 and 1988. Antimicrob Agents Chemother 39, 369-373 (1995).

80. Alieva, N.O. et al. Diversity and evolution of coral fluorescent proteins. PLoS One 3, e2680 (2008).

81. Shaner, N.C. et al. Improved monomeric red, orange and yellow fluorescent proteins derived from Discosoma sp. red fluorescent protein. Nat Biotechnol 22, 1567-1572 (2004).

82. Pedelacq, J.D., Cabantous, S., Tran, T., Terwilliger, T.C. & Waldo, G.S. Engineering and characterization of a superfolder green fluorescent protein. Nat Biotechnol 24, 79-88 (2006).

83. Handbook of Pesticide Toxicology. (Academic Press, London, UK; 2010).

84. Wang, X., Ferro, E.G., Zhou, G., Hashimoto, D. & Bhatt, D.L. Association Between Universal Masking in a Health Care System and SARS-CoV-2 Positivity Among Health Care Workers. JAMA (2020).

85. Leung, N.H.L. et al. Respiratory virus shedding in exhaled breath and efficacy of face masks. Nat Med 26, 676-680 (2020).

86. Tang, Y.W., Schmitz, J.E., Persing, D.H. & Stratton, C.W. Laboratory Diagnosis of COVID 19: Current Issues and Challenges. J Clin Microbiol 58 (2020).

87. Rong, X.M., Yang, L., Chu, H.D. & Fan, M. Effect of delay in diagnosis on transmission of COVID-19. Math Biosci Eng 17, 2725-2740 (2020).

88. Kretzschmar, M.E. et al. Impact of delays on effectiveness of contact tracing strategies for COVID-19: a modelling study. Lancet Public Health 5, e452-e459 (2020).

89. Gouma, P.I., Wang, L., Simon, S.R. & Stanacevic, M. Novel Isoprene Sensor for a Flu Virus Breath Monitor. Sensors (Basel) 17 (2017).

90. Maier, D. et al. Toward Continuous Monitoring of Breath Biochemistry: A Paper-Based Wearable Sensor for Real-Time Hydrogen Peroxide Measurement in Simulated Breath. ACS Sens 4, 2945-2951 (2019).

91. Gould, O., Ratcliffe, N., Krol, E. & de Lacy Costello, B. Breath analysis for detection of viral infection, the current position of the field. J Breath Res 14, 041001 (2020).

92. Ruszkiewicz, D.M. et al. Diagnosis of COVID-19 by analysis of breath with gaschromatography-ion mobility spectrometry - a feasibility study. EClinicalMedicine 29, 100609 (2020). 12.

93. Tromberg, B.J. et al. Rapid Scaling Up of Covid-19 Diagnostic Testing in the United States - The NIH RADx Initiative. N Engl J Med 383, 1071-1077 (2020).

94. Liu, Y. et al. Aerodynamic analysis of SARS-CoV-2 in two Wuhan hospitals. Nature 582, 557-560 (2020).
95. Fennelly, K.P. Particle sizes of infectious aerosols: implications for infection control. Lancet Respir Med (2020).

96. Van Doremalen, N. et al. Aerosol and Surface Stability of SARS-CoV-2 as Compared with SARS-CoV-1. N Engl J Med 382, 1564-1567 (2020).

97. Santarpia, J.L. et al. Aerosol and surface contamination of SARS-CoV-2 observed in quarantine and isolation care. Sci Rep 10, 12732 (2020).

98. Liu, L., Li, Y., Nielsen, P.V., Wei, J. & Jensen, R.L. Short-range airborne transmission of expiratory droplets between two people. Indoor Air 27, 452-462 (2017).

99. Ma, J. et al. Exhaled breath is a significant source of SARS-CoV-2 emission. medRxiv, 2020.2005.2031.20115154 (2020).

100. Baek, Y.H. et al. Development of a reverse transcription-loop-mediated isothermal amplification as a rapid early-detection method for novel SARS-CoV-2. Emerg Microbes Infect 9, 998-1007 (2020).

101. Joung, J. et al. Point-of-care testing for COVID-19 using SHERLOCK diagnostics. medRxiv (2020).

102. Arizti-Sanz, J. et al. Integrated sample inactivation, amplification, and Cas13-based detection of SARS-CoV-2. bioRxiv (2020).

103. F. J. Isaacs, D. J. Dwyer, J. J. Collins, RNA synthetic biology. *Nature biotechnology* **24**, 545 (2006).

104. M. K. Takahashi *et al.*, A low-cost paper-based synthetic biology platform for analyzing gut microbiota and host biomarkers. *Nature communications* **9**, 3347 (2018).

105. A. A. Green *et al.*, Complex cellular logic computation using ribocomputing devices. *Nature* **548**, 117 (2017).

106. S.-J. Kim, M. Leong, M. B. Amrofell, Y. J. Lee, T. S. Moon, Modulating responses of toehold switches by an inhibitory hairpin. *ACS synthetic biology* **8**, 601-605 (2019).

107. M. Krishnamurthy *et al.*, Tunable riboregulator switches for post-transcriptional control of gene expression. *ACS synthetic biology* **4**, 1326-1334 (2015).

108. J. Kim *et al.*, De-Novo-Designed Translational Repressors for Multi-Input Cellular Logic. *bioRxiv*, 501783 (2018).

109. A. C.-Y. To *et al.*, A comprehensive web tool for toehold switch design. *Bioinformatics* **34**, 2862-2864 (2018).

110. H. K. Kim *et al.*, Deep learning improves prediction of CRISPR–Cpf1 guide RNA activity. *Nature biotechnology* **36**, 239 (2018).

111. S. Webb, Deep learning for biology. *Nature* **554**, (2018).

112. C. Angermueller, T. Pärnamaa, L. Parts, O. Stegle, Deep learning for computational biology. *Molecular systems biology* **12**, (2016).

113. M. Wainberg, D. Merico, A. Delong, B. J. Frey, Deep learning in biomedicine. *Nature biotechnology* **36**, 829 (2018).

114. D. M. Camacho, K. M. Collins, R. K. Powers, J. C. Costello, J. J. Collins, Next-generation machine learning for biological networks. *Cell* **173**, 1581-1592 (2018).

115. D. Ma, L. Shen, K. Wu, C. W. Diehnelt, A. A. Green, Low-cost detection of norovirus using paper-based cell-free systems and synbody-based viral enrichment. *Synthetic Biology* **3**, ysy018 (2018).

116. G. Chuai *et al.*, DeepCRISPR: optimized CRISPR guide RNA design by deep learning. *Genome biology* **19**, 80 (2018).

117. S. Zhang, H. Hu, T. Jiang, L. Zhang, J. Zeng, TITER: predicting translation initiation sites by deep learning. *Bioinformatics* **33**, i234-i242 (2017).

118. J. Zuallaert, M. Kim, Y. Saeys, W. De Neve, in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. (IEEE, 2017), pp. 1233-1237.

119. E. C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi, G. M. Church, Unified rational protein engineering with sequence-only deep representation learning. *bioRxiv*, 589333 (2019).

120. Groher, Ann-Christin, et al. "Tuning the Performance of Synthetic Riboswitches Using Machine Learning." ACS Synthetic Biology, vol. 8, no. 1, Apr. 2018, pp. 34–44., doi:10.1021/acssynbio.8b00207.

121. D. B. Goodman, G. M. Church, S. Kosuri, Causes and effects of N-terminal codon bias in bacterial genes. *Science* **342**, 475-479 (2013).

122. Peterman, Neil, and Erel Levine. "Sort-Seq under the Hood: Implications of Design Choices on Large-Scale Characterization of Sequence-Function Relations." BMC Genomics, vol. 17, no. 1, Sept. 2016, doi:10.1186/s12864-016-2533-5.

123. Cambray, Guillaume, et al. "Evaluation of 244,000 Synthetic Sequences Reveals Design Principles to Optimize Translation in Escherichia Coli." *Nature Biotechnology*, vol. 36, no. 10, 2018, pp. 1005–1015., doi:10.1038/nbt.4238.

124. Kinney, J., McCandlish, D. (2019). Massively Parallel Assays and Quantitative Sequence-Function Relationships. *Annual Review of Genomics and Human Genetics.* 20(1), 99-127. doi: 10.1146/annurev-genom-083118-014845

125. Kinney, J., Murugan, A., Callan, C., Cox, E. (2010). Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence/ *Proceedings of the National Academy of Sciences*. 107(20), 9158-9163. doi: 10.1073/pnas.1004290107

126. François-Xavier Lehr, Maleen Hanst, Marc Vogel, Jennifer Kremer, H. Ulrich Göringer, Beatrix Suess, and Heinz Koeppl Cell-Free Prototyping of AND-Logic Gates Based on Heterogeneous RNA Activators ACS Synthetic Biology 2019 8 (9), 2163-2173

127. T. L. Bailey, DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **27**, 1653-1659 (2011).

128. J. N. Zadeh, B. R. Wolfe, N. A. Pierce, Nucleic acid sequence design via efficient ensemble defect optimization. *Journal of computational chemistry* **32**, 439-452 (2011).

129. R. M. Dirks, M. Lin, E. Winfree, N. A. Pierce, Paradigms for computational nucleic acid design. *Nucleic acids research* **32**, 1392-1403 (2004).

130. R. Lorenz *et al.*, ViennaRNA Package 2.0. *Algorithms for molecular biology* **6**, 26 (2011).

131. H. M. Salis, E. A. Mirsky, C. A. Voigt, Automated design of synthetic ribosome binding sites to control protein expression. *Nature biotechnology* **27**, 946 (2009).

132. A. Espah Borujeni *et al.*, Precise quantification of translation inhibition by mRNA structures that overlap with the ribosomal footprint in N-terminal coding sequences. *Nucleic acids research* **45**, 5437-5448 (2017).

133. A. Espah Borujeni, A. S. Channarasappa, H. M. Salis, Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. *Nucleic acids research* **42**, 2646-2659 (2013).

134. A. Espah Borujeni, H. M. Salis, Translation initiation is controlled by RNA folding kinetics via a ribosome drafting mechanism. *Journal of the American Chemical Society* **138**, 7016-7023 (2016).

135. B. Reeve, T. Hargest, C. Gilbert, T. Ellis, Predicting translation initiation rates for designing synthetic biology. *Frontiers in bioengineering and biotechnology* **2**, 1 (2014).

136. M. M. Meyer, The role of mRNA structure in bacterial translational regulation. *Wiley Interdisciplinary Reviews: RNA* **8**, e1370 (2017).

137. S. Badelt, S. Hammer, C. Flamm, I. L. Hofacker, in *Methods in enzymology*. (Elsevier, 2015), vol. 553, pp. 193-213.

138. B. Sauerwine, M. Widom, Kinetic Monte Carlo method applied to nucleic acid hairpin folding. *Physical Review E* **84**, 061912 (2011).

139. V. I. Jurtz *et al.*, An introduction to deep learning on biological sequence data: examples and solutions. *Bioinformatics* **33**, 3685-3690 (2017).

140. X.-Q. Liu, B.-X. Li, G.-R. Zeng, Q.-Y. Liu, D.-M. Ai, Prediction of Long Non-Coding RNAs Based on Deep Learning. *Genes* **10**, 273 (2019).

141. J. Baek, B. Lee, S. Kwon, S. Yoon, Lncrnanet: long non-coding rna identification using deep learning. *Bioinformatics* **34**, 3889-3897 (2018).

142. G. Aoki, Y. Sakakibara, Convolutional neural networks for classification of alignments of non-coding RNA sequences. *Bioinformatics* **34**, i237-i244 (2018).

143. A. Fiannaca, M. La Rosa, L. La Paglia, R. Rizzo, A. Urso, nRC: non-coding RNA Classifier based on structural features. *BioData mining* **10**, 27 (2017).

144. Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks Genome Res. 2016 Jul;26(7):990-9. doi: 10.1101/gr.200535.115.

145. Simon Höllerer, Laetitia Papaxanthos, Anja Cathrin Gumpinger, Katrin Fischer, Christian Beisel, Karsten Borgwardt, Yaakov Benenson, Markus Jeschek (2020) Large-scale DNA-based phenotypic recording and deep learning enable highly accurate sequence-function mapping. bioRxiv 2020.01.23.915405; doi: https://doi.org/10.1101/2020.01.23.915405

146. N. Frosst, G. Hinton, Distilling a neural network into a soft decision tree. *arXiv preprint arXiv:1711.09784*, (2017).

147. P. K. Koo, S. R. Eddy, Representation Learning of Genomic Sequence Motifs with Convolutional Neural Networks. *BioRxiv*, 362756 (2018).

148. K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, (2013).

149. Singh, J., Hanson, J., Paliwal, K. et al. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. Nat Commun 10, 5407 (2019).

150. A. Dhawan, J. G. Scott, A. L. Harris, F. M. Buffa, Pan-cancer characterisation of microRNA across cancer hallmarks reveals microRNA-mediated downregulation of tumour suppressors. *Nature communications* **9**, 5228 (2018).

151. Arnaout, Ramy, et al. "SARS-CoV2 Testing: The Limit of Detection Matters." 2020, doi:10.1101/2020.06.02.131144.

152. Zhou, Wenhua, et al. "A CRISPR–Cas9-Triggered Strand Displacement Amplification Method for Ultrasensitive DNA Detection." Nature Communications, vol. 9, no. 1, 2018, doi:10.1038/s41467-018-07324-5.

153. Biolabs, New England. "Isothermal Amplification & Strand Displacement." NEB, www.neb.com/products/pcr-qpcr-and-amplification-technologies/isothermal-amplification-and-strand-displacement/isothermal-amplification-and-strand-displacement.

154. Shi, Chao, et al. "Exponential Strand-Displacement Amplification for Detection of MicroRNAs." Analytical Chemistry, vol. 86, no. 1, 2013, pp. 336–339., doi:10.1021/ac4038043.