SemiSynBio-III

Moving Millions of Droplets at Megahertz Speeds: DNA Computing, DNA Storage, and Synthetic Biology on an Industrial Platform for Digital Microfluidics

1 Introduction

1.1 Context: DNA Storage

Ever since Watson and Crick first described the molecular structure of DNA, its information-bearing potential has been apparent to computer scientists. With each nucleotide in the sequence drawn from the four-valued alphabet of $\{A, T, C, G\}$, a molecule of DNA with n nucleotides stores 4^n bits of data. Indeed, this information storage underpins life as we know it: all the instructions on how to build and operate a life form are stored in its DNA, honed over eons of evolutionary time.

In a highly influential Science paper in 2012, the renowned Harvard genomicist George Church made the case that we will eventually turn to DNA for information storage, based on the ultimate physical limits of materials [10]. He delineated the theoretical storage **capacity** of DNA: 200 petabytes per gram; the read-write **speed**: less than 100 microseconds per bit; and, most importantly, the **energy**: as little as 10^{-19} joules per bit, which is orders of magnitude below the femtojoules/bit (10^{-15} J/bit) barrier touted for other emerging technologies. Moreover, DNA is stable for decades, perhaps even millennia, as DNA extracted from the carcasses of woolly mammoths can attest. In principle, DNA could outperform all other types of media that have been studied or proposed.

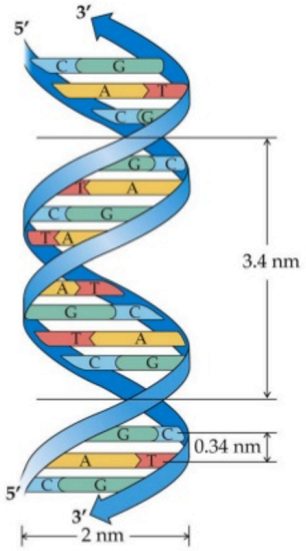

Figure 1: Physical Dimensions of DNA. If data were to be packed at these dimensions, with 2 bits per nucleotide, DNA could store 10 exabytes (one billion gigabytes) of data per cubic centimeter [10].

Of course, no one has yet built a DNA storage system that comes close to beating existing media (magnetic, optical, or solid-state storage). Church's capacity numbers are based on the physical dimensions of DNA (see Fig. 1). No one knows how to pack DNA in single-copy form at this density. His read-write speed estimates are based on molecular chemistry: the time it takes for a single letter of DNA to bind to the end of a strand. His energy estimates are based on the chemical energy required to form such bonds. An actual system has to *move* and *mix* chemicals to synthesize DNA, at scale. The space, time, and power to do this must be part of the equation.[1] While such details might be dismissed as engineering challenges, they are formidable.

Of course, DNA technology is not exotic. Spurred by the biotech and pharma industries, the technology for both sequencing (*reading*) and synthesizing (*writing*) DNA has followed a Moore's law-like trajectory for the past 20 years. Sequencing 3 billion nucleotides in a human genome can be done for less than $1,000. Synthesizing a megabyte of DNA data can be done in less than a day. Inspired no doubt by Church's *first-principles* thinking, but also motivated the trajectory of sequencing and synthesis technology, there has been a ground-swell of interest in DNA storage. (Surely a good fraction of the proposals submitted to this NSF program focus on this topic.) The leading approach is synthesis of DNA based on phosphoramidite chemistry [6]. However, many other creative ideas and novel technologies, ranging from nanopores [8] to DNA origami [11], are being deployed.

[1] Existing state-of-the-art DNA storage systems use liquid-handling robotics. The power consumption of these is on the order of *hundreds* of watts, so *hundreds* of joules/sec [6]. This for a system that only synthesizes *kilobytes*/sec. So, at present, the state-of-the-art is 18 orders of magnitude off of Church's theoretical limit when it comes to energy.

This proposal pertains to a state-of-the-art system for DNA storage being developed by our industrial partner, Seagate. (This is a GOALI proposal, so an academic-industrial partnership.) As we discuss in Section 1.3, the research objectives of the academic team are to explore applications of the technology beyond DNA storage itself.

1.2 Our Technology: Digital Microfluidics

Seagate is developing a platform for DNA storage based on **digital microfluidics** (DMF) that is unique both in its capability and its scale. DMF is a fluid-handling technology that precisely manipulates small droplets on a grid through electrical charge. It works on the principle of *electrowetting* which refers to the ability of an applied voltage to modulate the "wettability" of a surface. Aqueous droplets naturally bead-up on a hydrophobic surface, but a voltage applied between a droplet and an insulated electrode can cause the droplet to spread on the surface, as shown in Fig. 2. DMF harnesses electrowetting to control droplets. Electrical signals are applied to an array of electrodes to define the size and position of each droplet. Droplets are moved by turning the voltage on and off in succession across adjacent electrodes. The same mechanism can be used to dispense, merge, and mix droplets using electrical signals. These basic operations become the building blocks to perform biochemical reactions. *So chemical lab work becomes electronic hardware.* Fig. 3 shows a photo of a digital microfluidics system in action.

DMF technology is not new. It has been studied extensively in academia [12], and in recent years applied for specific tasks in industry [25]. However, it is fair to say that it remains a niche technology. Scaling the size of the electronic grid and so the capability of a DMF device is an expensive proposition [50]. The difference here: our industrial partner and its ambitions. Seagate is the world-leader in storage technology, with annual revenues of $10 billion. Its business centers on both electronics and material science, so it has exactly the requisite expertise to develop DMF technology. Most importantly, it has the resources to develop a DMF solution that other entrants in the DNA storage space do not.

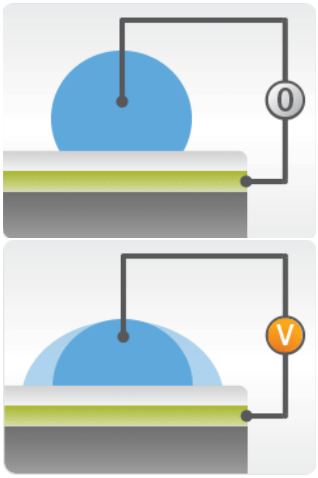

Figure 2: Electrowetting: aqueous droplets spread on a hydrophobic surface when a high voltage is applied. Top: no voltage. Bottom: high voltage

Contrast DMF with other forms of technology for chemical lab work, such as liquid-handling robotics [42]. Such systems have precise servo and stepper motors to move liquid between wells. Although very capable, such systems require precision engineering. The equipment is expensive to build, expensive to maintain, and expensive to operate – in terms of dollars and electricity. Moreover, such robotics simply do not scale the same way that electronics does. Arguably, electronics scale better than any technology humans have ever invented. With resources, we expect Seagate's technology to follow a Moore's Law-type trajectory: in 2 to 5 years, their DMF system will have millions of femtoliter-sized droplets moving reliably across a miniaturized electronic grid, moving at kilohertz speeds (so movement across thousands of grid points per second). Whereas existing DMF systems only have hundreds of electrodes, **Seagate is designing a system that will have millions.** Fig. 3 illustrates the dimensions of Seagate's product.

Seagate's corporate interest in this technology is narrow: they are aiming for a DNA storage system as a commercial product for archival storage. Although narrow, this goal is ambitious, since DNA storage in this form would have to displace existing technology, namely magnetic tape media and hard-drive storage. Such storage is commoditized. Accordingly, Seagate's target is to deliver a DNA storage system that can write a terabyte of data in one hour, for a cost of $50. In the process, of course, Seagate will create technology with entirely new capabilities, in terms of its scale. In the view of the academic team, even a preliminary prototype of the device would enable exciting, transformative science. For those with wetlab experience,

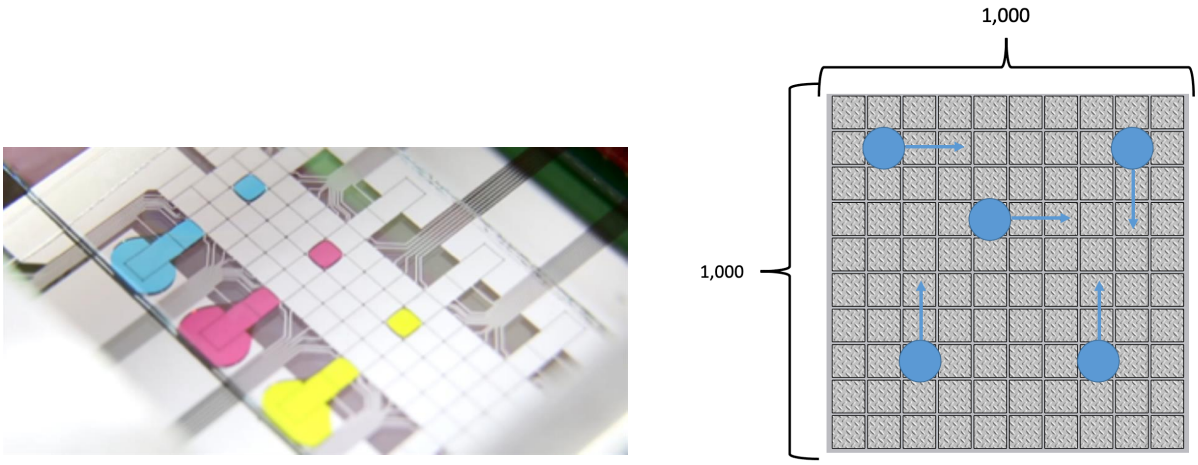

Figure 3: Left: A digital microfluidics (DMF) platform dispensing and moving droplets on a grid of electrodes. Right: Dimensions of Seagate's DMF device, developed for DNA Storage: it will have 1,000 × 1,000 = 1 million grid points, moving picoliter-sized droplets, at kilohertz speeds (so across thousands of grid points per second).

the proposition is: *what kind of science could one do, given a single device that can perform as much liquid-handling per second as 1,000 robots*[2]?

1.3 Our Research

The goal of this proposal is to explore alternate applications of the world-class technology that Seagate is developing for DNA storage. These range from novel computational paradigms on data stored in DNA – "in-memory" computing – to manipulating synthetic cells and engineering bioreactors. Our team of PIs is highly interdisciplinary. Here we first summarize the background and expertise of the PIs and how these integrate into the research. Then we summarize the themes from the SemSynBio III program that we address, including the leadership roles of the PIs. Details of the research plan are given Sections 2–4.

1.3.1 SemiSynBio-III Discipline Areas Covered

- PI Riedel's background is in digital circuit design. Accordingly, Prof. Riedel contributes expertise in **engineering**, specifically electronic design automation (EDA). As a faculty member, he has been pursuing research into novel paradigms for computation, including molecular computing in general, and DNA computing in particular.

- PI Soloveichik's background is in theoretical **computer science**. As a faculty member, his primary field of study has been DNA computing. He performs wet lab experiments with DNA. So, in addition to computer science, he brings expertise in **molecular biology**.

- PI Adamala's background is in molecular biology and neurobiology. She is among the most prominent researchers in the field of synthetic life. She brings expertise in **experimental biology**.

- PI Reddy's background is in mechanical engineering, specifically in **continuum mechanics**. He earned his Ph.D. from UC Berkeley, studying the shape stability and translation of acoustically-driven microbubbles. His expertise in **fluid dynamics**, and work experience in fabrication of micron-scale electronic components, are directly applicable to the task of digital microfluidics. Currently an Engineering Director in Seagate Research, Dr. Reddy manages a broad range of projects related to the tribology of the transducer-magnetic media interface, next generation optical communications, and archival storage systems.

[2]Say 1,000 "Echo" acoustic liquid handlers [36].

1.3.2 SemiSynBio-III Theme Areas Covered

Theme: Addressing fundamental research questions at the interface of biology and semiconductors.
Theme Leader: PI Adamala: Scaling up chemical lab work with robotics has been a trend in industry, spurred largely by the pharmaceutical business. Some tasks are straight-forward to automate simply by replacing manual labor with equivalent mechanized labor, for instance pipetting. However, automating research in synthetic biology is still an open challenge. Here the experimental protocols are often nuanced; automating and parallelizing them is not straightforward. In her research, PI Adamala is building and studying synthetic cells, with applications in biomanufacturing, medicine and space exploration. Synthetic cells are liposomal bioreactors that can mimic natural biology, as well as create new functionalities from biochemical parts. Building cells from scratch, she studies the origin and early evolution of life. She uses synthetic cell technologies to make tools for metabolic engineering, drug development, and biosensing. The challenge that she will address in this research is how to deploy this expertise on the digital microfluidics substrate. The goal: to explore the possibilities of using synthetic cells as biocomputers and as bioreactors. She will design and perform experiments on cell-free protein expression, liposome encapsulation, and RNA engineering on the substrate.

Theme: Designing sustainable bio-materials for novel bio-nano hybrid architectures and circuits that test the limits in transient electronics.
Theme Leader: PI Reddy: The technology that Seagate is developing is pushing the limits of what electronics can do in the context of fluidic handling for bio applications. No one has manufactured a digital microfluidics grid with thousands, let alone millions of grid points. To move milliliter-sized droplets requires a charge of 300 volts. For smaller droplets, the voltage level will come down. Nevertheless, creating a silicon substrate that can switch high voltages at high speeds is a significant electronic design problem. Dissipating heat, scaling competing forces (e.g., dielectrophoretic, drag, capillary), managing droplet dispense, and preventing contamination are significant additional challenges that will be addressed by the industrial partner.

Theme: Scaling-up and characterization of integrated hybrid synthetic bio-electronic storage systems.
Theme Leader: PI Soloveichik: For the DNA storage system, synthesizing large datasets entails routing hundreds of thousands of droplets across a miniaturized silicon grid. Given an arbitrary list of data symbols to be encoded, droplets must be created, destinations on the grid chosen, and routes calculated. For "in-memory" computing on data stored in DNA, we must choose DNA sequences that do not cross-hybridize with each other or form secondary structures. This is a perennial problem in the field of DNA computing, one that PI Soloveichik has addressed successfully in prior research. However, the scale of in-memory computing that we are proposing – with terabytes of data, and hundreds of thousands of parallel operations per second – is unprecedented. Addressing the combinatorics of DNA sequence design for storage and for in-memory computing on DNA is a significant thrust of this proposal.

2 Synthesizing DNA on Digital Microfluidics Device

Before discussing applications with the DMF platform, we discuss the strategy for synthesizing DNA on the device that Seagate has developed. This material is a prerequisite for the applications of in-memory computing that we are proposing in Section 3. Note that Seagate's aim is to synthesize one terabyte's worth of data (1 billion gigabytes) in the form of DNA per hour. The only way to achieve this sort of data rate is:

1. With massive *parallelization* in terms of the operations. Perform many, many steps in the synthesis process in parallel.
2. By synthesizing significant data in *each operation*, thereby increasing the bitrate.

In addition to the problem of speed, any strategy for DNA synthesis must confront the problem of fluid

volume. Even if the initial reagents are combined in volumes as low as milliliters, the final volume of waste liquid produced in writing one terabyte would be measured in *billions* of liters. The solution is to move to digital microfluidics, as discussed in Section 1.2. This technology not only allows for precise handling of droplets in a massively parallel manner, it also leverages the phenomenon of electrowetting to allow moving, merging, mixing and splitting of droplets that are *nanoliters* or smaller in volume. In this section, we discuss a system for writing large datasets into DNA with this platform. We describe a novel writing protocol based on **Gibson assembly**, which will significantly increase the bitrate per operation compared to extant methods.

In 2009, Gibson proposed a method for joining multiple DNA fragments in a single reaction [17]. The matched fragments must have overlapping ends of a few dozen base pairs in length, leaving room for their longer interior portions to be unique. Besides the fragments to be assembled, three enzymes must also be mixed in. Exonuclease is required to cut back the strands on the end; DNA polymerase to fill in gaps; and finally DNA ligase to join the matched strands. Temperature must be controlled during the multistep process to ensure reactions proceed properly. This assembly method is general-purpose; with slight modifications it can be adapted to serve as a method of constructing very large data-storage strands. Abusing biological terminology, we will call these data-storage units "genes."

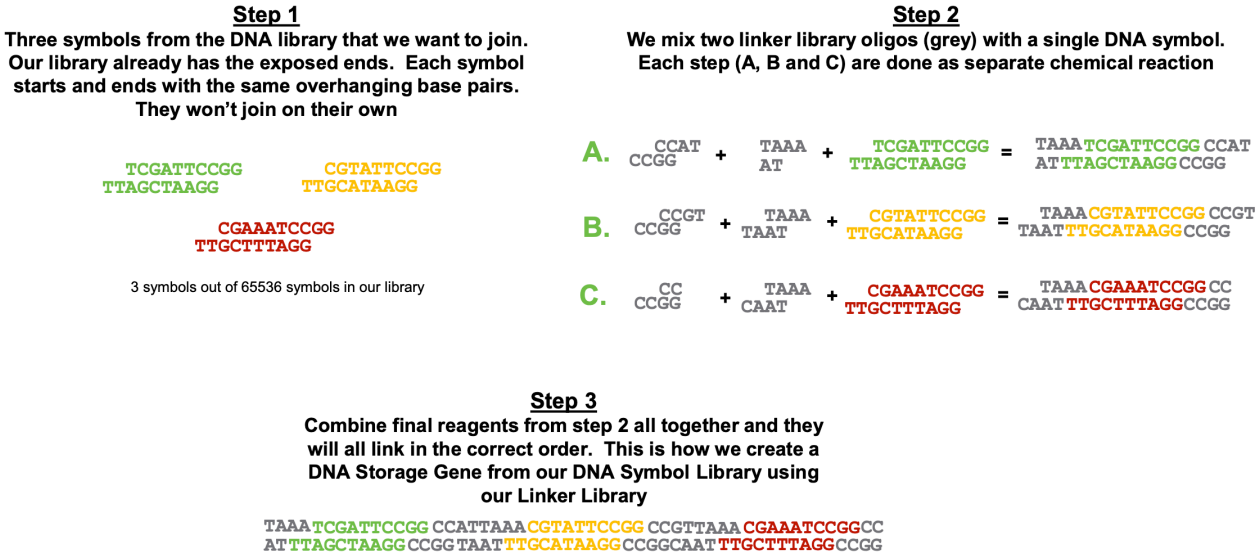

Figure 4: An example of the symbol-linker assembly process.

The challenge lies in building a library of DNA oligonucleotides (or "oligos" – short single strands of synthetic DNA) that meet the requirements:

1. They can be assembled in any order to represent arbitrary data.
2. Multiple oligos from the library can be attached simultaneously in a Gibson assembly process without risk of misalignment.

The first requirement necessitates segments that can be linked with any of their peers in any order, but the second requirement contradicts this. To ensure desired ordering, segments must be uniquely matched with one another.

The solution is a dual library of oligos we call "symbols" and "linkers". The symbols and linkers respectively cover the first and second requirements. Data genes consist of long chains of alternating symbols and linkers, with relevant information contained in the symbols. All symbols have unique interior segments,

composed of 8 base pairs, allowing each to encode 16 bits. By using multi-bit symbols instead of assembling one base pair at a time, we exchange much of the fabrication time for overhead in maintaining the library. All symbols share the same 5' and 3' sequences, one for each end. The 3' and 5' ends are not complementary, preventing inadvertent direct Gibson assembly of two symbols.

Complementary ends are shared by all linkers, but each linker will only have one end matched with those of the symbols. The other end binds with its unique, complementary linker. Thus, any desired chain of symbols can be assembled by first using Gibson assembly to separately attach each symbol to the appropriate linkers and then bringing all attached symbol-linker pairs together, in another Gibson assembly process. The linkers will naturally order themselves according to their unique matches and the symbols will automatically fall into the appropriate order. This process is demonstrated in Fig. 4. Following assembly, the new string of symbols will undergo purification and polymerase chain reaction (PCR) multiplication to reduce the risk of errors.

Any gene requiring more symbols than can be assembled in a single process can be constructed by repeated assembly processes, each consisting of a manageable number steps linking individual segments. This will allow reliable assembly of arbitrary sequences of symbols, re-using the same linker library. Readout can be assisted by a "bookend" sequence shared by every linker that marks the space between each symbol. This process does not require macroscopic chemical reagent droplets, and can in theory be miniaturized to the domain of microscopic droplets found on a DMF device.

Assembling extremely large datasets of millions or billions of symbols will require vast numbers of individual Gibson assembly operations. This presents a non-trivial problem in the form of droplet traffic. Given an arbitrary list of symbols to be encoded in a gene, droplets must be created, destinations chosen, and routes carefully controlled. This multistep process necessitates an automated system capable not only of routing and traffic control but also deciding what Gibson assembly operations to perform, and when, in order to build the desired gene. The UMN team has been tackling this computer science problem in a collaboration with Seagate.

3 In-Memory Computing on DNA

Beginning with the seminal work of Adelman in 1994 [1], DNA computing has promised the benefits of massive parallelism in operations. However, it is fair to say that in the three decades since, the practical impact of research in this field has been modest. Operations are typically performed on the *concentration* of DNA strands in solution. For instance, with DNA strand displacement cascades, single strands displace parts of double strands, releasing single strands that can then participate in further operations [26, 37, 51]. The inputs and outputs are the concentration values of specific strands. With a focus on concentration, the application space for this research has been limited to computing that is embedded in chemical systems.

A practical DNA storage system, particularly one that is inherently programmable such as our platform, changes this. The scheme that we discuss below operates on data stored not in the sequence of nucleotides, but rather in topological modifications to the strands: breaks in the phosphodiester backbone of DNA that we call "nicks" and gaps in the backbone that we call "toeholds." In prior work, we have proposed performing the nicking enzymatically with a system such as

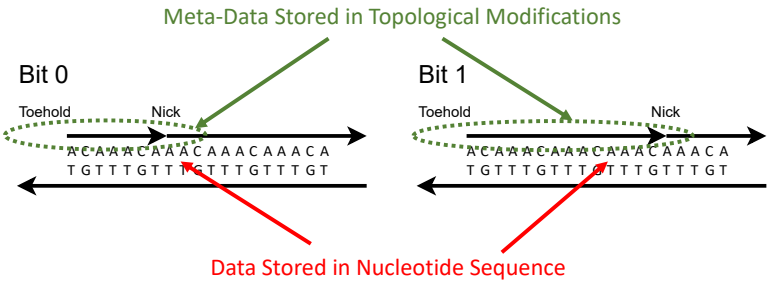

Figure 5: Data is stored in multiple dimensions. The sequence of nucleotides stores data in the form of the A's, C's, T's, and G, with 2 bits per letter. Superimposed on this, we store data via topological modifications to the DNA, in the form of nicks and exposed toeholds. This data is **rewritable**, with techniques developed for DNA computation.

CRISPR/Cas9 [19, 40]. The key concept here is that we can assemble DNA using Seagate's digital microfluidics technology, at scale, with the requisite nicks and toeholds; and then perform computation on the DNA assembled this way.

Note that the data that we operate on with this form of DNA computing is encoded in a different dimension than the data encoded in the sequence data of the DNA. The **underlying data** – perhaps terabyte's worth of it – is stored as the sequence of A's, C's, T's, and G's in synthesized strands. Superimposed on this, we store **metadata** via topological modifications. This is illustrated in Fig. 5. This metadata is rewritable with the techniques that we describe here. Accordingly, it fits the paradigm of "in-memory" computing [18]. The paradigm that we will deploy on our digital microfluidics substrate, dubbed "SIMD||DNA", was recently introduced by PIs Soloveichik [47], with follow-on work by PI Riedel [9].[3]

3.1 SIMD||DNA structure

Our implementation of SIMD provides a means to transform stored data, perhaps large amounts of it, with a single parallel instruction. We divide stretches of double-stranded DNA into "domains", where each domain is a contiguous sequence of nucleotides of some small specified length (typically 5 to 20). A sequence of several (typically 5 to 7) domains maps to a "cell" storing one binary bit. Whether a cell stores a 0 or a 1 depends upon topological variations, specifically the location of "nicks", i.e., breaks in the DNA backbone. The nicks always occur on one strand of a double-stranded complex (generally the top strand in our examples); the other remains untouched. We create these nicks in the assembly process; these occur between the edges of the smaller fragments that we snap together, as described in Section 1.1.

Strand displacement is used to implement computation on the stored values. It is predicated on the encoding scheme for data, shown in Figure 6. Each cell stores a single binary value (a "bit"). Each cell consists of 7 domains. We do not specify the actual nucleotide sequence of the domains here for simplicity. While preparing this cell, the top DNA strand must be nicked before and after domain 1. This strand can then be displaced by denaturing, creating an exposed toehold. Domain 1 is always exposed as a toehold in this representation. Domains 2 through 7 are covered. When storing a bit 0, we will nick the top strand between domains 3 and 4; when storing a bit 1, we will nick between domains 5 and 6.

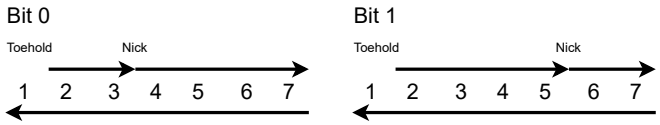

Figure 6: Bit representation in the encoding scheme. Horizontal lines represent DNA strands. Integers represent "domains": specific sequences of nucleotides. Arrowheads represent nicked positions: places where the phosphodiester bond in the backbone of the DNA strand has been broken, via gene-editing techniques. Cells store binary values. Each cell consist of 7 domains. Domain 1 is always exposed, forming a toehold.

The computation is carried out by a sequence of "instructions", where each instruction implements DNA strand displacement reactions on cells. Instructions are initiated by single-stranded "instruction strands" added to the solution. After the strand displacement cascades complete, any single-strand fragments in the solution are washed away; the original strand is kept and separated via a magnetic bead. After a sequence of instructions, the data is transformed to its final state. The readout can be performed via fluorescence or with Oxford nanopore devices [2, 23].

We omit details for rewriting data, as they are intricate. We refer the reader to [9, 47]. Figure 7 illustrates part of the process of rewriting the contents of a cell. By exposing toeholds across domains 2 through 7 in a cell, we can rewrite the content of that cell – so change a 1 to 0 or a 0 to 1 – with three instructions. The idea is that, since there are exposed domains, we can displace the content of the cell with a single strand covering all these domains. Then we can remove the covering strand through the exposed "tag" domain (S

[3] SIMD is a computer engineering acronym for Single Instruction, Multiple Data [13], a form of computation in which multiple processing elements perform the same operation on multiple data points simultaneously. It contrasts with the more general class of parallel computation called MIMD (Multiple Instructions, Multiple Data). Much of the modern progress in electronic computing power has come by scaling up SIMD computation with platforms such as graphical processing units (GPUs).

in Figure 7) using a complementary strand. The cell is now completely exposed. We can write a new bit to it by hybridizing the strands according to our encoding scheme, leaving domain 1 as a toehold and placing the nick at the desired location.

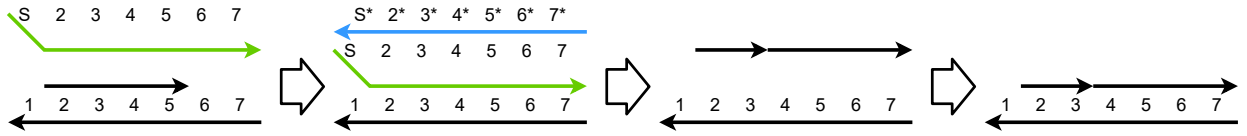

Figure 7: Example of Rewriting: through strand displacement, we change the location of nicks. These encode values of 0 or 1, as "meta-data." Here we change a 1 to a 0.

3.2 Applications

We have demonstrated "SIMD||DNA" implementations for a variety of algorithms [9, 47], operating on binary data, stored with the encoding shown in Fig. 6. All are fundamental operations in computer science. These include:

- **Sorting**: performed in only N parallel steps, where N is the number of bits to be sorted.
- **Shifting**: performed in a single parallel step. (This implements multiplication or division by powers of 2 on binary data.)
- **Searching**: returns an answer as to whether a query substring is present in a target string in $\log(n)$ steps where n is the length of the query string.
- **Counting**: incrementing and decrementing binary data.

With this grant, we will demonstrate all of these applications experimentally on the DMF substrate. We will generalize searching in the form of queries on data stored in DNA:

1. *Yes/No Queries*. Supply a query string: DNA-based computation returns "yes", if the string is found in data; "no" if it is not.
2. *Content-Addressable Storage*. Supply a "key": DNA-based computation returns the data associated with the key.
3. *Similarity Search*. Supply a string: DNA-based computation returns data similar to it.

As discussed above, our "in-memory" computing will be performed on the metadata layer, via topological modifications. We will explore aspects of computing with metadata such as *access counts*, *watermarks*, and *error checking* via computation on this layer.

3.3 Experimental implementation of SIMD||DNA algorithms

3.3.1 Multiplexing and Random Access

A key aspect of SIMD||DNA is that multiple data registers can be manipulated in parallel each holding different input data. (This gives the paradigm its name: Single Instruction Multiple Data.) An important goal of this proposal is to experimentally realize such parallel computation, on the digital microfluidics substrate. Another is to be able to selectively address or read out only certain of these data registers – so implement *random access*.

In our unpublished preliminary results, we have experimentally demonstrated parallel computation and random access for the *binary counting* application, in a conventional wetlab (so not yet on the target DMF

substrate). We pooled registers with unique initial values in the same test tube and performed parallel computation. To achieve random access, each register had a unique "adaptor" sequence, so only a unique strand for that register can perform strand displacement and release that register for post-computation processing from the magnetic bead.

3.3.2 Scaling up: larger register sizes, increasing program complexity

While the preliminary data that we have is limited in scope, the goal with this grant is to demonstrate that scaling up SIMD||DNA computation is possible on the DMF substrate. To achieve this, we will first perform computation for multiple *sequential* steps. We also plan to show that SIMD||DNA is capable of performing computation on registers with an order of magnitude more bits than the preliminary experiments.

Then, we will show *parallel computation* and *random access* of multiple registers at different locations, each with a unique sequence space. In addition, we will show parallel computation and random access of multiple programs in parallel. We will conduct different programs simultaneously on registers at different locations. This adds yet another level of parallelism to SIMD||DNA.

An important problem of strand displacement systems is that undesired reactions could cause false computation results, which is called *leak*. In electrical and computer engineering, error correction strategies include adding redundancy to the system design: making one mistake in the overall computation requires multiple leak events to occur [46]. Our recent work has shown that the idea of introducing redundancy can also reduce leak in chemical systems [43, 49]. We will study how such error correction method can be applied in the context of SIMD||DNA algorithms.

3.3.3 Generating registers via microfluidics symbol-linker assembly

Prior work proposes to construct the registers used by SIMD||DNA via enzymatic nicking [41] or by annealing chemically synthesized DNA [47]. In this project, we will adapt the symbol-linker assembly process (Fig. 4) to construct the SIMD||DNA registers with specific initial inputs. Importantly, by omitting the 5' phosphate group on the "top" strands, we can ensure that only the "bottom" strands are ligated to form one long strand, while the nicks on the top strands are preserved. Thus, an alphabet of symbols and linkers can be assembled into registers with a prescribed input (setting of which cells are 0 and which are 1 by the choice of the corresponding symbol). This assembly would allow high-throughput assembly of information in the form compatible with SIMD||DNA computation via the microfluidics platform, which should allow orders of magnitude larger systems than possible with prior assembly methods.

4 Engineering Enzyme Pathways

Here we describe a core application in synthetic biology that we will implement on the digital microfluidics (DMF) substrate: engineering and optimizing a complex enzyme pathway. A task such as this entails many successive steps of chemical lab work. This, of course, is what the DMF platform excels at. It also entails gene manipulation and DNA recovery. To our knowledge, such biology has never been attempted with digital microfluidics.

In nature, the origin of cellular individuality marked the origin of biological life. In metabolic engineering, evolving pathways encapsulated in live cells comes with the constraints and limitations that live cell-metabolism places on artificially constructed pathways [21]. To overcome most of those limitations, cell-free protein expression systems have been used to engineer enzymes [3]. However, the activity of the entire enzymatic pathway is often a multidimensional space: optimizing it is not a question of optimizing a single enzyme but rather the task of optimizing *all* the enzymes simultaneously – in a complete pathway.

Pathway engineering of this sort has traditionally been done on live, natural cells. This sort of biology would be challenging on the platform. However, PI Adamala's expertise is working with **synthetic cells** and **"cell-free" protein** expression systems. These are ideal applications for our testbed. Compartmentalization

is a key concept in engineering pathways. Multiple genes can be kept together, with selection applied to the cumulative product of activity of all those genes.

Fig. 8 demonstrates our approach. A model metabolic pathway consists of four steps, catalyzed by three different enzymes, with one of the enzymes being a two-protein complex. To evolve and optimize the pathway with independent control over mutations and stoichiometry of each enzyme, first we will prepare libraries of each of the proteins in the pathway. We will encapsulate a single copy of each of the enzyme genes in a single synthetic cell. Each DNA strand will be amplified inside it. Synthetic cell populations containing copies of each individual gene will be fused, to complete the construction of a 3-enzyme, 4 protein-pathway. The addressable microfluidic system will be used to control and sort the liposomes.

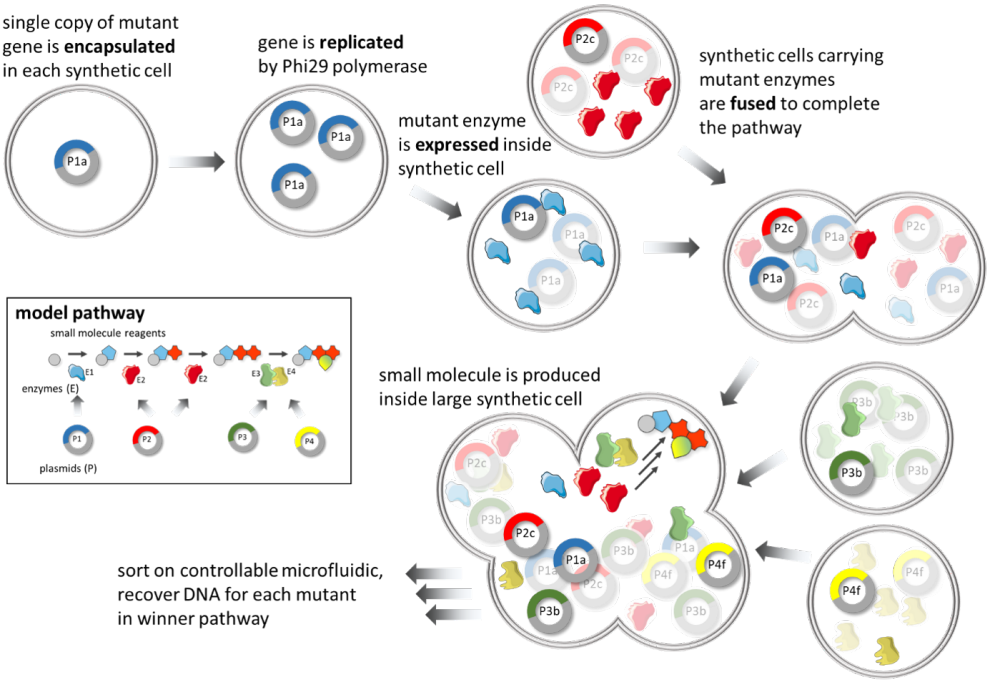

Figure 8: Evolving Synthetic Pathways on a Digital Microfluidics Platform

4.1 Synthetic Cells

Synthetic cells are liposome bioreactors containing cell-free protein expression systems [15]. Compartmentalization of translation inside phospholipid liposomes delivers all advantages of natural compartmentalization. In addition, it provides the flexibility of cell-free systems: keeping reaction pathways together, coupling activity of all genes in a pathway to each other, and coupling the final product to all the genes needed to complete the pathway [4, 7, 38]. Physical co-localization of reactants inside a lipid compartment increases local concentration of reagents.

We propose using synthetic cells, here understood as phospholipid compartments with cell-free translation system and libraries of genes for the desired pathway, as a platform for evolving whole gene pathways. We will use libraries of each of the genes, with each variant of the pathways co-encapsulated with variants of all other enzymes. Each synthetic cell with a variant of the metabolic engineering pathway will then be tested for presence of the final product, with the selection done on the complete pathway instead of on single enzymes.

A synthetic cell pathway evolution platform will help us optimize whole enzymatic pathways, as well as create new pathways. These experiments will be a model application, demonstrating the utility of the ad-

dressable microfluidic platform. Cell-free translation is, in many ways, lineage agnostic. There is no codon bias, and synthetic cells have no endogenous metabolism that could re-direct metabolites from the designed pathway. It is therefore possible to engineer pathways with enzymes coming from different species [5].

While optimizing a pathway, there is no need to balance flux with the needs of a natural cell's metabolic processes. Metabolic engineering pathways built in synthetic cells can be optimized for maximizing production of the desired metabolite, and many parameters of the pathway can be freely explored. We will be able to iterate parameters such as scaffolding of enzymes and/or reagents, or copy number of each enzyme. It will also be possible to build systems operating under non-physiological reaction conditions, including the presence of unnatural amino acids and reagents possibly toxic to live cells.

Synthetic cells are the only system that enables optimizing whole enzymatic pathways, while offering all the advantages of cell-free protein engineering. Encapsulation will also allow high-throughput testing of novel modifications to the active centers of enzymes. We will make a library of pathways with various amino acids in the active site replaced by a single unique codon. Each of the synthetic cells will contain that unique codon tRNA with a different unnatural amino acid. With this library, we will develop novel functionalities in enzymatic pathways based on novel, unnatural chemical modifications of enzymes in that pathway.

The key challenges in developing the pathway evolution system will be in the following areas: 1) Expression of the enzymes in cell-free system; 2) encapsulation of the appropriate mutants of each pathway in single synthetic cell; 3) the physical act of selection based on pathway activity; and 4) the pathway design. The first and fourth problems will be addressed with the help of collaborators on this grant, while the second and third will be addressed using technologies recently developed in the Adamala lab.

1. **Expression**. The first thing to address will be if each of the desired enzymes expresses and is active in a cell-free system. We will troubleshoot expression problems in genes, including changing the design of DNA templates and changing the composition and type of the cell-free translation system used.

2. **Encapsulation of the correct mutant pathway**. To build libraries of genetic pathways inside synthetic cell liposomes, and to be able to select based on a performance of specific pathway with identifiable mutants of each gene, it will be necessary to encapsulate a single copy of each pathway enzyme. One way to achieve this would be to put all enzymes of the pathway on a single, multicistronic plasmid. The megaplasmids created this way are known to be fragile, difficult to propagate, and difficult to use. This approach would also negate one of the key advantages of cell-free systems: the ability to control the stoichiometry of proteins by simply controlling the copy number of the DNA for each gene. To maintain full control of stoichiometry of gene products and the composition of the final pathway, and to avoid encapsulating more than one mutant of each enzyme in the pathway inside the same synthetic cell, we will utilize two technologies recently developed in our group: controllable synthetic cell mating and replication of DNA inside synthetic cells.

   The Adamala lab developed a programmable synthetic cell fusion system, allowing us to selectively "mate", or fuse, synthetic cells containing specific fusion tags. This system allows orthogonal fusion of a theoretically unlimited number of different populations of liposomes (we have demonstrated fusion of 9 different populations to complete a 12-gene pathway). We have demonstrated the metabolic engineering utility of this system to reprogram the violaceic acid pathway [16].

   For this project, we will deploy a universal method for replication of DNA in synthetic cells. This method is independent of the identity of the gene, and allows amplification of plasmid or linear DNA in cell-free protein expression systems without limiting translational activity of the genes, and without using up protein-expression resources [22]. The replication of DNA can be controlled by sequences encoded in the gene. Therefore, we can make libraries of a gene, where the final post-replication copy number depends on the strength of the replication signal – enabling the creation of synthetic

cell libraries by varying by the amount of gene, to optimize stoichiometry of enzymes in the pathway. Combining the DNA replication and the programmable synthetic cell mating technologies, we can first create a library of synthetic cells containing a single copy of each variant of each of the genes needed for the desired pathway. Then, we will replicate that single copy inside synthetic cells, and we will fuse the compartments in the desired stoichiometry, creating complete pathways. This should satisfy the requirement of low copy number encapsulation and solve problems with creation of unique pathway libraries.

3. **Selection & Pathway design**. Optical sorting of liposomes is notoriously difficult. We will take advantage of the digital microfluidic system developed in this project to greatly expand our ability to design and optimize metabolic engineering pathways that do not produce optically identifiable products. The Adamala lab demonstrated the use of synthetic cells to optimize the violacein pathway [16], as well as antibiotic resistance pathways (manuscript under consideration). Those will be the first two pathways we will use to validate the experimental system. While attempting to select completely novel enzyme pathways would be beyond the scope of this project, we will use the demonstration of this particular pathway to showcase the capabilities of the technology.

5 University-Industry Interaction

The GOALI partner, Seagate Technology has an existing research relationship with PI Riedel on the topic of DNA storage. Specifically, the University of Minnesota team is analyzing and optimizing the choice of chemical protocols, the parameters, the layout, and the routing of droplets on the digital microfluidics chip that Seagate is developing for this purpose. The goal is to produce a parametrized design that maximizes throughput, minimizes the complexity, and minimizes the cost of the storage system.

With this funding, the academic team will explore applications of the technology. Seagate will provide scientific, engineering and technical support for the research in this grant. All physical experiments will be conducted on the university campuses in Minnesota and Texas.

- Seagate will communicate to the University PIs and co-PIs details about the capabilities and design specifications for the digital microfluidics (DMF) platform that it is developing.
- When a working prototype is available, it will allow the University PIs and co-PIs to perform experiments with it at their universities.
- It may also perform some of the experiments that the university PIs have designed at its engineering facilities.

6 Deliverables and Timeline

We summarize the deliverables of this grant.

1. **DNA Storage**
   (a) Deliver a prototype DMF board with the ability to mix, merge, split, thermocycle, and perform PCR. [year 1]
   (b) Demonstrate a prototype that hybridizes genes 8 symbols long, from a set of 8 symbols, each 8 nucleotides in length. [year 2]
   (c) Deliver a prototype that assembles, in parallel, 256 such genes. [year 3]
2. **In-Memory Computing**
   (a) Implement basic "SIMD||DNA" operations on the DMF prototype. [year 1]

(b) Implement basic algorithms on small data sets (128 binary bits) on the prototype: counting, sorting, shifting, and searching. [year 2]

(c) Implement the algorithm on large data sets (1024 binary bits). [year 3]

3. **Synthetic Biology.** Design and perform experiments on the DMF substrate, demonstrating:

   (a) Cell-free protein expression. [year 1]

   (b) Liposome encapsulation. [year 2]

   (c) RNA engineering on the substrate. [year 3]

7 Impact

The case for transformative synthetic biology with the DMF substrate is clear: it will enable experimental biologists to scale what they do by orders of magnitude. The case for DNA storage, outlined in Section 1, is self-evident. We conclude this proposal with a speculative argument about DNA computing, as outlined in Section 3. With the "SIMD||DNA"framework, we note that there are, in fact, two layers of parallelism possible:

1. **Bit-level Parallelism**: instructions applied to all bits in an array at once.
2. **Data-level Parallelism**: the same instructions applied to *multiple* arrays at once.

With these levels of parallelism, DNA computation can – in principle – scale to a truly impressive regime. Consider the following back-of-an-envelope estimates. Suppose:

– we have encoded 10^{12} bits on distinct DNA strands in a single droplet on the DMF substrate;

– a single "in-memory" operation takes 15 minutes to complete as droplets move across the substrate;

This means that we can perform approximately 10^9 operations per second. Now suppose:

– we have 100 independent DMF devices (in a rack).

This means we can compute at 100,000 MIPS (million instructions per second). This is comparable to what very respectable existing silicon systems can achieve. The key conceptual difference between the SIMD DNA approach and other forms of DNA computing is that it exploits a substrate on which data is stored. This enables the SIMD parallelism.

8 Results from Prior NSF Support

For PI **Adamala**, the most relevant items are:

1. NSF award #1844313,1/2020–12/2023, *RoL: RAISE: DESYN-C3: Engineering multi-compartmentalised synthetic minimal cells*. Intellectual merit: this project aims at construction of a synthetic minimal cell that can produce its own lipid bilayers, including both the external membrane and interior organelle membranes of distinct composition and properties, and that can respond to chemical signals from the environment. This will allow construction of synthetic minimal cells with varying membrane composition, programmed by the genes inside, with coacervate cytoplasm and genetic circuits responding to signals from the environment. Broader impacts include facilitating DIY bio community outreach, collaboration with K12 teachers through PSU sponsored lab residencies, and mentoring undergraduate students by both PI. Publications: [35, 45] and papers under consideration with journals: Heili et al., submitted 2021, and Cash et al., submitted 2022.

2. NSF award #1807461, 8/2018–9/2023, *SeMiSynBio Very Large scale genetic circuit design and automation*. Intellectual merit: We demonstrated engineering complex genetic circuits in distributed populations of synthetic minimal cells. Broader impacts include public talks, popular science articles, undergraduate mentoring. Publications: [16, 24, 35]; and papers under consideration with journals: Heili et al., submitted 2021, and Sato et al., submitted 2022.

3. NSF award #1901145, 4/2019–5/2024, *RCN Build-a-Cell: An Open Community Considering & Advancing the Construction of Synthetic Cells*. Intellectual merit: The Build-a-Cell RCN will facilitate studies on understanding and engineering a diverse range of synthetic cells. Broader impact: student education and outreach to the US public through public media, articles and workshops. Publications: [14, 24, 35, 39].

For PI **Soloveichik** the most relevant items are:

1. *NSF award #1652824 (CAREER: Robust Molecular Computation: Error-Correcting Reaction Networks and Leakless DNA Circuits, 2017-2022, $500,000) PI: Soloveichik*; **Intellectual Merit:** This grant funded theoretical and experimental work on improving the fidelity of DNA molecular networks by developing systematic means to reduce leak in strand displacement systems including [48, 49] (published in *PNAS*). **Broader Impacts:** This grant supported a female graduate student (Boya Wang) who transferred to the ECE PhD program from biochemistry. This award also supported her travel to conferences (winning a best student paper award at the 25th International Conference on DNA Computing and Molecular Programming). Further, the award funded conference travel of an undergraduate student (Niels Kornerup), winning a best paper award at the 16th International Conference on Computational Methods in Systems Biology, as well as the department's Best Undergraduate Thesis award.

2. *NSF award #1901025 (FET: Medium: Collaborative Research: Engineerable Molecular Computing: Flying like an Airplane, not like a Bird, 2019-2023, $1,000,000) PIs: David Doty (UC Davis), Soloveichik*: **Intellectual Merit:** This grant funded the theoretical and experimental implementation of a class of chemical reaction networks whose computational power stems entirely from stoichiometry rather than reaction rates. This class was shown to be surprisingly powerful, including in-principle capable of implementing ReLU neural networks widely used in machine learning [44]. **Broader Impacts:** This award supported a female postdoctoral researcher in her transition from a graduate student, helped support undergraduate work, and supported graduate course development.

For PI **Riedel** the most relevant item is:

1. is: (a) NSF award #1423407, $300K, 08/2014–07/2017. (b) Title: *"Advanced Digital Signal Processing with DNA."* (c) Intellectual merit: The PIs developed circuits that perform advanced digital signal processing operations such as finite-impulse response (FIR) and infinite impulse response (IIR) digital filters, fast Fourier transforms (FFT), and power spectral density (PSD) computations, using molecular reactions in general, and DNA-based reactions in particular. Broader Impact: Ahmad Salehi was funded throughout his PhD studies by this grant. He is now a faculty member at the University of Kentucky. (d) The project led to nine publications: [20, 27, 28, 29, 30, 31, 32, 33, 34]. (e) Data from this project was presented publicly at conferences and all source code has been made publicly available through the UMN Conservancy Project.

9 Broader Impacts

We propose an ambitious set of synergistic activities for our broader impacts. Our ultimate goals are to create a well-trained, diverse workforce, with future leaders in the field. Also, to communicate scientific

discoveries to the public at large. There are three main threads in our activities: (1) engagement in public policy discussions, particularly regarding artificial life and biosafety; (2) mentoring activities to improve the mental health of students; and (3) engaging high-school students in biochem "maker" culture.

– Biosafety and Biosecurity Considerations:
Engineering entirely novel biocomputing systems creates potential for both accidental and malicious misuse. All work proposed in this project will be toward engineering synthetic cellular systems that do not fit into any known current lineages of life, but all building blocks (DNA and proteins) will be compatible with natural life. Therefore, biosafety and biosecurity considerations need to be taken into account from the very beginning of the experimental planning phase.

PI Adamala is the leader of the Engineering Biology Research Consortium Security working group. She is also the leader of the Build-a-Cell Consortium Ethics and Biosafety working group. In both of those roles, she has the capacity to enter problems into the consideration of both groups, and at Build-a-Cell as a virtue of being the lead PI of the Build-a-Cell network with the authority to make that decision individually. The premise of this project will be presented for the malice analysis to the both EBRC and Build-a-Cell security groups. All principles of design of engineered biocomputing systems, the role in sequestering natural and engineered nucleic acid sequences and the system of response to external stimuli will be analyzed both for accidental issues and for potential for bad actor use. This analysis will be done at the beginning of the work on this project, to enable informed changes in the overall strategy of designing types of plasmids, and any modification to the mitochondria system. The broader impact of this strategy will be the use of this project as a template for future collaboration with researchers: both safety working groups will be able to provide basic malice analysis to the groups looking to consider other bioengineering projects with potential biosafety and biosecurity implications.

– Mentoring Students with Mental Health Issues:
Since January 2022, PI Soloveichik has been co-leading a multidisciplinary effort (collaboration with Nursing and the Department of Psychology faculty) to help improve the mental health of graduating ECE seniors by developing an intervention to decrease anxiety due to decision making. The ongoing study focuses on using ideas from quantum physics (the many worlds interpretation) to challenge the participants' view of the world, with the hypothesis that such challenges may decrease harmful rumination.

– Bio/Chem "Maker" Fairs:
The "maker" culture is taking a new generation of technology enthusiasts by storm. It is both a response to and an outgrowth of digital culture, made possible by new tools and electronic components that let people integrate the physical and digital worlds simply and cheaply. Hobbyists everywhere are building every manner of gizmo and gadget out of cheap components. Two significant technologies have fueled the interest: 3-D printing and cheap Wi-Fi enabled microcontrollers. People with no prior knowledge or experience with programming or engineering can easily work with these technologies. 3-D printers offer a powerful way to connect the digital and the physical realms: they take a digital model of an object and print it out by building it up, one layer at a time, using plastic extruded from a nozzle.

Building on this interest, we will organize "maker" fairs, both in Minnesota and Texas. Unlike existing events of this sort, there will be an explicit emphasis on **bio** and **chemical computation**. Students will be given materials and guidance to pursue creative projects in a competitive format. All projects will entail either some chemistry or biology, as well interfacing this with electronic sensors, microcontrollers, and actuators.

These technologies represent a unique opportunity for researchers in our field to engage students – including students that might otherwise be hard to engage, such as students from poorer and disadvantaged high schools. Unlike most such outreach efforts, the focus will not be on *elite* high schools, but rather

disadvantaged ones. These include the Red Lake Public School District, Richfield Public School District, and the St. Cloud Public School District in Minnesota, as well as the Travis County and Austin Independent School Districts. In our experience, students respond well to demos with hands-on gizmos and gadgets. We expect that such interactions will encourage students to pursue studies in engineering at four-year colleges.

Project Summary

Overview

The information-bearing potential of DNA is apparent. With each nucleotide in the sequence drawn from the four-valued alphabet of $\{A, T, C, G\}$, a molecule of DNA with n nucleotides stores 4^n bits of data. In principle, DNA could provide a storage medium that is many orders of magnitude denser than conventional media. Spurred by the biotech and pharma industries, the technology for both sequencing (*reading*) and synthesizing (*writing*) DNA has followed a Moore's law-like trajectory. Nevertheless, a large gap remains between what is theoretically possible in terms of read/write speed and what has been demonstrated for DNA storage systems. In order to be competitive with conventional media, a DNA storage drive must achieve a write speed of 1 terabyte per hour and cost no more than $50 per unit. Scaling any existing technology for synthesis to this speed, at this cost, is a formidable challenge.

The industrial partner in this GOALI proposal is developing a platform for DNA storage that will achieve this. It is based on **digital microfluidics** (DMF), a fluid-handling technology that precisely manipulates small droplets on a grid through electrical charge. Everything is choreographed by software that controls the switching of the voltages on the grid. Droplets are moved by turning the voltage on and off in succession across adjacent electrodes. The same mechanism can be used to dispense, merge, and mix droplets using electrical signals. These basic operations become the building blocks to perform biochemical reactions. *So chemical lab work becomes electronic hardware.*

The goal of the academic team on this proposal is to explore alternate applications of the world-class technology that Seagate is developing for DNA storage. These range from novel computational paradigms on data stored in DNA – so "in-memory" computing – to manipulating synthetic cells and engineering bioreactors. Our team of PIs is highly interdisciplinary, with expertise in electronics, material science, fluid dynamics, computer science, and molecular biology.

Keywords: DNA storage, DNA computing, molecular computing, in-memory computing

Intellectual Merit

Arguably, electronics scales better than any technology humans have ever invented. With resources, we expect Segate's technology to follow a Moore's Law-type trajectory: in 2 to 5 years, their DMF system will have millions of femtoliter-sized droplets moving reliably across a miniaturized electronic grid, moving at kilohertz speeds (so movement across thousands of grid points per second.) Whereas existing DMF systems have only tens or perhaps hundreds of electrodes, **Seagate is designing a system that will have millions.** In the process, of course, Seagate will create technology with entirely new capabilities, in terms of its scale, namely the digital microfluidics platform discussed above. The academic team is poised to exploit this new technology, applying it to challenging problems in computing and synthetic biology. The computer science community has unique expertise that can be brought to bear on the challenging design problems encountered in engineering and molecular biology. Applications in these domains, in turn, offer a wealth of interesting problems in algorithmic development. With its cross-disciplinary emphasis, this project will bring new perspectives to both fields.

Broader Impacts

The project will communicate the goals and the impetus for interdisciplinary research to a wide audience, including to students in the sciences and in engineering, as well as to the general public. The PIs have an excellent track record with such public engagement. The research in this proposal points to many applications that the general public can appreciate. There are three main threads in our activities: (1) engagement in public policy discussions, particularly regarding artificial life and biosafety; (2) mentoring activities to improve the mental health of students; and (3) engaging high-school students in biochem "maker" culture.

Data Management Plan

I. The types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project.

This project will result in datasets of results of wet lab experiments such as: microscopy, HPLC, nucleic acid gels, fluorescent plate reader and UV-vis data.

All research results will be promptly disseminated on a website jointly hosted by University of Minnesota and the University of Texas. The front-end will exploit modern software infrastructure for data analytics and visualization. The back-end will consist of a MySQL database, directly linked to the computational software, running on a distributed platform. The website will be dynamically updated as the results of simulation trials complete. It will provide all raw simulation results, in plain text, as a spreadsheet, or through structured SQL queries. It will feature:

- An interactive Chemical Reaction Network (CRN) simulation tool, with options to simulate stochastically or via differential equations. This will build off existing CRN software developed at the University of Minnesota and the University of Texas.
- An interactive CRN synthesis tool, with a Verilog-like syntax. Again, this will build off existing software developed at the University of Minnesota.
- A portal for creating and displaying molecular structures, generated with the package "PyMol": `https://pymol.org/2/`
- A curated, dynamically-updated set of links to related websites and related publications.

The data will be dynamically updated as simulations are completed. There will be no embargo periods. The original creators will not retain any rights prior to releasing the data to wider use. The data created will be freely available under Creative Commons, GNU, MIT, or other open licenses for such datasets. While data will be distributed freely, proper security software and protocols will be maintained on all computing infrastructure.

II. Data and Metadata Standards

The data collected from this project will be stored in the original forms in which the data was collected. The HPLC traces (ca. 1,000/yr) will be stored in its original Agilent file format, with metadata describing sample, analysis conditions, and full detector readout traces. All fluorescence and UV-vis data (ca. 10,000 spectra per year) will be stored in the original format generated by the instrument. All imaging data (ca. 1,000 images per year) will be stored in the original Nikon file format with metadata describing the microscope, light source and objective settings. Gel images will be stored in lossless TIFF format (ca. 500 per year). Plasmids (5-10 over the project period) will be archived with Addgene for public distribution (see OpenMTA commitment below). The water and soil samples from various environments will be recorded with geolocation data and all relevant

Molecular data will be maintained in standard formats including those specified in METLIN Gen2:`https://metlin.scripps.edu/`. This data will be managed as Unix files and in SQL databases. The contextual details we need to make our data meaningful are visualizations and graphs. Graphs will be made using the "Matplotlib" python package. 3D visualizations of molecular interactions will be created using the package "3dMol.js." When graphs are presented on the website, links will be provided for the raw plotting

commands, as well as for downloading the underlying data.

III. Policies for access and sharing and provisions for appropriate protection/privacy

Project results will be published in peer-reviewed journals, and disseminated at scientific conferences for an international audience. Efforts will be made to publish in open access journals or to select open access publication options in order to ensure that research results are freely available.

Plasmid and DNA distribution: The plasmids will be distributed on Addgene with the standard academic use license allowing for modification and production of derivative constructs. The DNA constructs created in this project, particularly the viroid chloroplast imaging and all gene transfer constructs, will be distributed under the OpenMTA, Open Material Transfer Agreement. The basic principles of OpenMTA are: - Materials available under the OpenMTA are free of any royalty or fees, other than appropriate and nominal fees for preparation and distribution. - Providers may request attribution and reporting for materials distributed under the OpenMTA. - Materials available under the OpenMTA may be modified or used to create new substances. - The OpenMTA does not restrict any party from selling or giving away the materials, either as received or as part of a collection or derivative work. - The OpenMTA supports the transfer of material between researchers at all types of institutions, including academic, industry, government and community laboratories.

V. Policies and provisions for re-use, re-distribution

There will be no permissions or restrictions required to access data.

VI. Plans for archiving and Preservation of access

We will make regular offline data backups of all generated datasets using Google Drive, GitHub, and other cloud services. Code will be hosted on GitHub, and there will be a copy of it on each developer's computer. The University of Minnesota will be the central host of the databases and will also host the web interface to our software. Curated data, for instance all data referenced in publications, will be archived indefinitely through the University of Minnesota Digital Conservancy service.

We intend to publish in highly read, broad-interest journals, as we have done in the past. When space permits, raw data (such as gels) will be available within papers. When data are important to reproducing results, but space or formatting constraints do not permit their presentation within the main text of a paper (such as large raw datasets), they will be presented in the supplementary material where necessary. All data will be provided in standard file formats. The PI's past publications have contained extensive supplementary documentation of this nature.

Some materials are not required to reproduce the authors' results, but will be helpful to investigators in related fields. If such materials cannot be presented in a publication or in supplementary information, they will be reproduced on the author(s)' website(s), or they will be made available upon request from the author(s). All data will be provided in standard file formats. Similarly, the PI's past publications as a graduate student and postdoc have been made available on their websites (aaronengelhart.com and protobiology.org), along with abstracts tailored to a lay audience to better communicate the work.

We will archive all the results obtained in the course of this project, positive results published in peer review journals as well as negative results that will not be suitable for publication, in the Data Repository for University of Minnesota. The database developed in the course of this project will be available online, and backed up on the GitHub account, as well as backup copy will be kept on the University of Minnesota Google Drive account. The source code will be available on the GitHub account.

Award Abstract

Ever since Watson and Crick first described the molecular structure of DNA, its information-bearing potential has been apparent to computer scientists. With each nucleotide in the sequence drawn from the four-valued alphabet of $\{A, T, C, G\}$, a molecule of DNA with n nucleotides stores 4^n bits of data. In principle, DNA could provide a storage medium that is many orders of magnitude denser than conventional media. Spurred by the biotech and pharma industries, the technology for both sequencing (*reading*) and synthesizing (*writing*) DNA has progressed rapidly. Nevertheless, a large gap remains between what is theoretically possible in terms of read/write speed and what has been demonstrated for DNA storage systems. In order to be competitive with conventional media, a DNA storage drive must achieve a write speed of 1 terabyte per hour and cost no more than $50 per unit. Scaling any existing technology for synthesis to this speed, at this cost, is a formidable challenge. The industrial partner in this research, Seagate, is developing a purely electronic platform that will meet this challenge. The goal of the academic team on this proposal is to explore alternate applications of the world-class technology that Seagate is developing. These range from novel computational paradigms on data stored in DNA – so-called "in-memory" computing – to manipulating synthetic cells and engineering bioreactors. The academic team will strive for a high level of public engagement. This includes initiating policy discussions, particularly regarding artificial life and biosafety; mentoring activities to improve the mental health of students; and engaging high-school students in biochem "maker" culture.

This proposal pertains to a state-of-the-art system for DNA storage being developed by our industrial partner, Seagate, based on **digital microfluidics**. Chemical reactions for DNA storage, DNA computing, and synthetic biology are effected by manipulating droplets via electric charge. This entails not only moving droplets, but a variety of complex operations at the interface between the electronics and the droplets: heating and cooling (for instance for polymerase chain reactions); mixing; and purifying (with magnetic beads). The academic team will focus on developing compute applications for the Seagate's DNA storage platform. They will explore a scheme for computing on data stored not in the sequence of nucleotides, but rather in topological modifications to the strands: breaks in the phosphodiester backbone of DNA called "nicks" and gaps called "toeholds." In prior work, such computation has been demonstrated by nicking DNA enzymatically with enzymatic systems such as CRISPR/Cas9. In this work, DNA with the requisite nicks and toeholds will be assembled directly using Seagate's digital microfluidics technology. The academic team will also explore applications in synthetic biology with the device that Seagate is developing, including using synthetic cells as biocomputers and as bioreactors. They will design and perform experiments on cell-free protein expression, liposome encapsulation, and RNA engineering on the electronic grid.

The project was jointly funded by Division of Molecular and Cellular Biosciences (MCB) in the Directorate for Biological Sciences (BIO); Division of Computing and Communication Foundations (CCF) in the Directorate for Computer and Information Science and Engineering (CISE); Division of Electrical, Communications and Cyber Systems (ECCS) in the Directorate for Engineering (ENG) and the Division of Materials Research (DMR) in the Directorate for Mathematical and Physical Sciences (MPS).

10 Timeline, Risk, and Mitigation Strategies

A timeline for our proposed technical tasks is given in Figure **??**. We point to several risks and the mitigation strategies that we will adopt.

A potential risk in using multicomponent molecules for data storage is the complexity associated with *in situ* synthesis by a nanofluidic liquid handler. Although the Ugi four-component condensation is a well-established reaction, the synthetic yield of Ugis can vary depending on the starting components. Undesired byproducts left in the library can react with each other to further contaminate the encoding mixtures. This cross-reactivity can be minimized by excluding the components that lead to low Ugi production. In such a case, we will employ synthetic analogs of high-yielding components to maintain the overall library size.

A risk is that cross-reactivity among compounds in a synthesized chemical library may hamper design efforts. To mitigate this risk, we plan to explore a wide range of potential models, and fit the computation to the chemistry – thereby exploiting the cross-reactivity to do useful computations – instead of trying to fit the chemistry to the computation.

A risk that is intertwined in all the computation tasks is that the measurement accuracy, i.e., the signal-to-noise ratio, may not be sufficient to distinguish the results from the computations. A possible mitigation strategy is to use additional readout modalities in order to identify subtle differences. We will also apply machine learning on noise models in order to calibrate and recover target signals from readouts. Measurement noise can also lead to readable results that are inaccurate or irreproducible. To mitigate this risk, we will refocus our experiments on stochastic and approximate computing techniques, areas of expertise of PIs Riedel and Reda, respectively. These techniques produce computing systems that can compute efficiently, yet inherently tolerate inaccuracies and errors.

References

[1] Leonard M Adleman. Molecular computation of solutions to combinatorial problems. *Science*, 266(5187):1021–1024, 1994.

[2] Nagendra Athreya, Olgica Milenkovic, and Jean-Pierre Leburton. Detection and mapping of dsDNA breaks using graphene nanopore transistor. *Biophysical Journal*, 116(3):292a, 2019.

[3] Erik D Carlson, Rui Gan, C Eric Hodgman, and Michael C Jewett. Cell-free protein synthesis: applications come of age. *Biotechnol. Adv.*, 30(5):1185–1194, September 2012.

[4] Filippo Caschera and Vincent Noireaux. Compartmentalization of an all-e. coli cell-free expression system for the construction of a minimal cell. *Artif. Life*, 22(2):185–195, March 2016.

[5] Arturo Casini, Fang-Yuan Chang, Raissa Eluere, Andrew M King, Eric M Young, Quentin M Dudley, Ashty Karim, Katelin Pratt, Cassandra Bristol, Anthony Forget, Amar Ghodasara, Robert Warden-Rothman, Rui Gan, Alexander Cristofaro, Amin Espah Borujeni, Min-Hyung Ryu, Jian Li, Yong-Chan Kwon, He Wang, Evangelos Tatsis, Carlos Rodriguez-Lopez, Sarah O'Connor, Marnix H Medema, Michael A Fischbach, Michael C Jewett, Christopher Voigt, and D Benjamin Gordon. A pressure test to make 10 molecules in 90 days: External evaluation of methods to engineer biology. *J. Am. Chem. Soc.*, 140(12):4302–4316, March 2018.

[6] Luis Ceze, Jeff Nivala, and Karin Strauss. Molecular digital data storage using DNA. *Nature Reviews Genetics*, 20(8):456–466, Aug 2019.

[7] Anna H Chen and Pamela A Silver. Designing biological compartmentalization. *Trends Cell Biol.*, 22(12):662–670, December 2012.

[8] Kaikai Chen, Jinbo Zhu, Filip Bošković, and Ulrich F. Keyser. Nanopore-based dna hard drives for rewritable and secure data storage. *Nano Letters*, 20(5):3754–3760, 2020. PMID: 32223267.

[9] Tonglin Chen, Arnav Solanki, and Marc Riedel. Parallel Pairwise Operations on Data Stored in DNA: Sorting, Shifting, and Searching. In Matthew R. Lakin and Petr Šulc, editors, *27th International Conference on DNA Computing and Molecular Programming (DNA 27)*, volume 205 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 11:1–11:21, Dagstuhl, Germany, 2021. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[10] George Church, Yuan Gao, and Sriram Kosuri. Next-generation digital information storage in DNA. *Science (New York, N.Y.)*, 337:1628, 08 2012.

[11] George D. Dickinson, Golam Md Mortuza, William Clay, Luca Piantanida, Christopher M. Green, Chad Watson, Eric J. Hayden, Tim Andersen, Wan Kuang, Elton Graugnard, Reza Zadegan, and William L. Hughes. An alternative approach to nucleic acid memory. *Nature Communications*, 12(1):2371, Apr 2021.

[12] R. B. Fair. Digital microfluidics: is a true lab-on-a-chip possible? *Microfluidics and Nanofluidics*, 3(3):245–281, Jun 2007.

[13] Michael J. Flynn. Some computer organizations and their effectiveness. *IEEE Trans. Comput.*, 21(9):948–960, September 1972.

[14] Caroline Frischmon, Carlise Sorenson, Michael Winikoff, and Katarzyna P Adamala. Build-a-cell: Engineering a synthetic cell community. *Life (Basel)*, 11(11):1176, November 2021.

[15] Nathaniel J Gaut and Katarzyna P Adamala. Reconstituting natural cell elements in synthetic cells. *Adv Biol (Weinh)*, 5(3):e2000188, March 2021.

[16] Nathaniel J. Gaut, Jose Gomez-Garcia, Joseph M. Heili, Brock Cash, Qiyuan Han, Aaron E. Engelhart, and Katarzyna P. Adamala. Programmable fusion and differentiation of synthetic minimal cells. *ACS Synthetic Biology*, 11(2):855–866, Feb 2022.

[17] Daniel G Gibson, Lei Young, Ray-Yuan Chuang, J Craig Venter, Clyde A Hutchison, 3rd, and Hamilton O Smith. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods*,

6(5):343–345, May 2009.

[18] Daniele Ielmini and H-S Philip Wong. In-memory computing with resistive switching devices. *Nature electronics*, 1(6):333–343, 2018.

[19] Fuguo Jiang and Jennifer A Doudna. Crispr–cas9 structures and mechanisms. *Annual review of biophysics*, 46:505–529, 2017.

[20] Hua Jiang, Sayed Ahmad Salehi, Marc D Riedel, and Keshab K Parhi. Discrete-time signal processing with dna. *ACS synthetic biology*, 2(5):245–254, 2013.

[21] Christopher P. Kempes, M. A. R. Koehl, and Geoffrey B. West. The scales that limit: The physical boundaries of evolution. *Frontiers in Ecology and Evolution*, 7, 2019.

[22] K. Libicher, R. Hornberger, M. Heymann, and H. Mutschler. In vitro self-replication and multicistronic expression of large synthetic genomes. *Nature Communications*, 11(1):904, Feb 2020.

[23] Ke Liu, Chao Pan, Alexandre Kuhn, Adrian Pascal Nievergelt, Georg E Fantner, Olgica Milenkovic, and Aleksandra Radenovic. Detecting topological variations of DNA at single-molecule level. *Nature communications*, 10(1):3, 2019.

[24] Rebecca Mackelprang, Katarzyna P. Adamala, Emily R. Aurand, James C. Diggans, Andrew D. Ellington, Samuel Weiss Evans, J. L. Clem Fortman, Nathan J. Hillson, Albert W. Hinman, Farren J. Isaacs, June I. Medford, Shadi Mamaghani, Tae Seok Moon, Megan J. Palmer, Jean Peccoud, Elizabeth A. Vitalis, India Hook-Barnard, and Douglas C. Friedman. Making security viral: Shifting engineering biology culture and publishing. *ACS Synthetic Biology*, 11(2):522–527, 2022. PMID: 35176864.

[25] David Millington, Scott Norton, Raj Singh, Rama Sista, Vijay Srinivasan, and Vamsee Pamula. Digital microfluidics comes of age: high-throughput screening to bedside diagnostic testing for genetic disorders in newborns. *Expert review of molecular diagnostics*, 18(8):701–712, Aug 2018. 30004274[pmid].

[26] L. Qian and E. Winfree. A simple DNA gate motif for synthesizing large-scale circuits. *Journal of the Royal Society Interface*, February 2011.

[27] Sayed Ahmad Salehi, Hua Jiang, Marc D Riedel, and Keshab K Parhi. Molecular sensing and computing systems. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, 1(3):249–264, 2015.

[28] Sayed Ahmad Salehi, Keshab K Parhi, and Marc D Riedel. Chemical reaction networks for computing polynomials. *ACS synthetic biology*, 6(1):76–83, 2016.

[29] Sayed Ahmad Salehi, Keshab K. Parhi, and Marc D. Riedel. Computation of mathematical functions using dna via fractional via fractional coding. *Nature Scientific Reports*, 8(8312), 2018.

[30] Sayed Ahmad Salehi, Marc D Riedel, and Keshab K Parhi. Asynchronous discrete-time signal processing with molecular reactions. In *Signals, Systems and Computers, 2014 48th Asilomar Conference on*, pages 1767–1772. IEEE, 2014.

[31] Sayed Ahmad Salehi, Marc D Riedel, and Keshab K Parhi. Markov chain computations using molecular reactions. In *Digital Signal Processing (DSP), 2015 IEEE International Conference on*, pages 689–693. IEEE, 2015.

[32] Sayed Ahmad Salehi, Marc D Riedel, and Keshab K Parhi. Computing polynomials using chemical reaction networks. In *IEEE Globecom Symposium*. IEEE, 2016.

[33] Sayed Ahmad Salehi, Marc D Riedel, and Keshab K Parhi. Computing polynomials with positive coefficients using stochastic logic by double-nand expansion. In *ACM Great Lakes Symposium on VLSI*. ACM, 2017.

[34] Sayed Ahmad Salehi, Marc D Riedel, and Keshab K Parhi. Molecular computation of complex markov chains with self-loop state transitions. In *Asilomar Conference on Signals, Systems and Computers*, 2017.

[35] Wakana Sato, Tomasz Zajkowski, Felix Moser, and Katarzyna P Adamala. Synthetic cells in biomedical applications. *Wiley Interdiscip. Rev. Nanomed. Nanobiotechnol.*, 14(2):e1761, March 2022.

[36] Beckman Coulter Life Sciences. Echo 650 series next-generation acoustic liquid handlers. 2022.

[37] David Soloveichik, Georg Seelig, and Erik Winfree. DNA as a universal substrate for chemical kinetics. *Proceedings of the National Academy of Sciences*, 107(12):5393–5398, 2010.

[38] Pasquale Stano. Gene expression inside liposomes: From early studies to current protocols. *Chemistry – A European Journal*, 25(33):7798–7814, 2019.

[39] Oskar Staufer, Jacqueline A De Lora, Eleonora Bailoni, Alisina Bazrafshan, Amelie S Benk, Kevin Jahnke, Zachary A Manzer, Lado Otrin, Telmo Díez Pérez, Judee Sharon, Jan Steinkühler, Katarzyna P Adamala, Bruna Jacobson, Marileen Dogterom, Kerstin Göpfrich, Darko Stefanovic, Susan R Atlas, Michael Grunze, Matthew R Lakin, Andrew P Shreve, Joachim P Spatz, and Gabriel P López. Building a community to engineer synthetic cells and organelles from the bottom-up. *Elife*, 10, December 2021.

[40] S. Tabatabaei, Boya Wang, Nagendra Athreya, Behnam Enghiad, Alvaro Hernandez, Christopher Fields, Jean-Pierre Leburton, David Soloveichik, Huimin Zhao, and Olgica Milenkovic. DNA punch cards for storing data on native DNA sequences via enzymatic nicking. *Nature Communications*, 11, 12 2020.

[41] S Kasra Tabatabaei, Boya Wang, Nagendra Bala Murali Athreya, Behnam Enghiad, Alvaro Gonzalo Hernandez, Christopher J Fields, Jean-Pierre Leburton, David Soloveichik, Huimin Zhao, and Olgica Milenkovic. DNA punch cards for storing data on native DNA sequences via enzymatic nicking. *Nature communications*, 11(1):1–10, 2020.

[42] Houriiyah Tegally, James Emmanuel San, Jennifer Giandhari, and Tulio de Oliveira. Unlocking the efficiency of genomics laboratories with robotic liquid-handling. *BMC Genomics*, 21(1):729, Oct 2020.

[43] Chris Thachuk, Erik Winfree, and David Soloveichik. Leakless DNA strand displacement systems. In *DNA Computing and Molecular Programming*, Lecture Notes in Computer Science, pages 133–153. Springer, 2015.

[44] Marko Vasic, Cameron Chalk, Sarfraz Khurshid, and David Soloveichik. Deep molecular programming: a natural implementation of binary-weight ReLU neural networks. In *International Conference on Machine Learning (ICML)*, pages 9701–9711. PMLR, 2020.

[45] Orion M Venero, Wakana Sato, Joseph M Heili, Christopher Deich, and Katarzyna P Adamala. Liposome preparation by 3d-printed microcapillary-based apparatus. *Methods Mol. Biol.*, 2433:227–235, 2022.

[46] John Von Neumann. Probabilistic logics and the synthesis of reliable organisms from unreliable components. In *Automata Studies.(AM-34), Volume 34*, pages 43–98. Princeton University Press, 2016.

[47] Boya Wang, Cameron Chalk, and David Soloveichik. SIMD||DNA: single instruction, multiple data computation with DNA strand displacement cascades. In *DNA25: International Conference on DNA Computing and Molecular Programming*, volume 11648, pages 219–235. Springer, LNCS, 2019.

[48] Boya Wang, Chris Thachuk, Andrew D Ellington, and David Soloveichik. The design space of strand displacement cascades with toehold-size clamps. In *DNA Computing and Molecular Programming*, volume 10467 of *Lecture Notes in Computer Science*, pages 64–81. Springer, 2017.

[49] Boya Wang, Chris Thachuk, Andrew D Ellington, Erik Winfree, and David Soloveichik. Effective design principles for leakless strand displacement systems. *Proceedings of the National Academy of Sciences*, 115(52):E12182–E12191, 2018.

[50] Ya-Tang Yang and Tsung-Yi Ho. Conquering the tyranny of number with digital microfluidics. *Front. Chem.*, 9:676365, May 2021.

[51] Bernard Yurke. A DNA-fuelled molecular machine made of DNA. *Nature*, 406(6796: 605), 2000.