Applying Deep Learning to Discover Highly Functionalized Nucleic Acid Polymers That Bind to Small Molecules

Applying Deep Learning to Discover Highly Functionalized Nucleic Acid Polymers that Bind to Small Molecules

A THESIS PRESENTED
BY
MICHAEL WORNOW
TO
THE DEPARTMENT OF COMPUTER SCIENCE
&
THE DEPARTMENT OF STATISTICS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
BACHELOR OF ARTS (HONORS)
IN THE SUBJECT OF
COMPUTER SCIENCE & STATISTICS

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
APRIL 2020

Thesis advisor: Professor David R. Liu Michael Wornow

Applying Deep Learning to Discover Highly Functionalized Nucleic Acid Polymers that Bind to Small Molecules

Abstract

Developing novel binders for small molecule and protein targets has been at the core of a number of recent medical breakthroughs. A popular developmental method focuses on monoclonal antibodies, currently a $98B industry. Antibodies, however, are costly to produce and difficult to manufacture. DNA aptamers – oligonucleotides that bind to a specific target – present a promising alternative, as they can be manufactured at scale for lower cost. Unfortunately, existing experimental methods for identifying functional aptamers typically take months, and can only sample a small fraction (1 in 10^{10} possibilities) of the theoretical search space. Deep learning techniques offer a novel solution to the challenge of aptamer discovery. In addition to developing a theoretical model for quantifying aptamer binding affinity, this thesis demonstrates that a conditional variational autoencoder (CVAE) can be used to generate novel high-binding aptamers for daunomycin, a chemotherapeutic agent. After training on eight rounds of experimental data, a CVAE was able to successfully generate entirely original aptamers that performed as well as, or better than, those generated through conventional selections ($K_D \approx$ 10-30 nM). This thesis shows the power of using deep learning techniques, coupled with intimate domain knowledge, to accelerate the process of screening aptamers, and thereby advance the field towards concrete therapies and diagnostics.

Contents

1 Introduction 1

1.1 Antibodies 2

1.2 Aptamers 4

1.3 SELEX 8

1.4 Problems with SELEX 10

1.4.1 Execution 10

1.4.2 Experimental Design 11

1.4.3 Data Analysis 14

1.5 Motivation 18

1.6 Related Work 19

1.6.1 *In Silico* Screening Methods 19

1.6.2 Deep Learning 22

2 Biological Problem 28

2.1 Highly Functionalized Nucleic Acid Polymers (HFNAPs) 29

2.1.1 Nucleotide Backbone 29

2.1.2 Functionalization 30

2.2 Dataset 32

2.2.1 Target: Daunomycin 32

2.2.2 HFNAP Synthesis 33

2.2.3 SELEX 34

2.2.4 Raw Data 36

2.2.5 Analysis Metrics 38

3 Computational Methods 42

3.1 Weak Supervised Learning 43

3.1.1 Enrichment Fold Change 43

3.1.2 Fitness Model 45

3.2 Predictive Models 49

3.2.1 Variational Autoencoder (VAE) 50

3.2.2 Conditional Variational Auto Encoder (CVAE) 59

3.3 Experimental Validation 64

4 Results 66

4.1 VAE 66

4.2 CVAE 69

4.3 Experimental Validation 73

5 Conclusion 77

A Appendix 80

A.1 Code 80

References 90

Listing of figures

2.1.1 Functional Groups 31

3.1.1 Distribution of Fitness Scores 45
3.2.1 Autoencoder Architecture 51
3.2.2 VAE Statistical Model 53
3.2.3 VAE Architecture 59
3.2.4 CVAE Architecture 64

4.1.1 VAE Performance Graphs 68
4.2.1 CVAE Performance Graphs 72
4.3.1 MST Dose-Response Curves of CVAE-Generated HFNAPs . . 75

Acknowledgments

I would like to thank my committee of readers, Professors David R. Liu, Michael Smith, and Samuel Kou. With a field as intersectional as computational biology and a motivating biological problem as specific as aptamer screening, I really appreciate their willingness to support my endeavors across disciplines and their consideration of my work. I would particularly like to thank Professor Liu for being my primary thesis advisor, and for hosting me in his lab over the past several years.

I also want to thank Jon C. Chen for being an incredible mentor, collaborator, and friend throughout this process. Quite literally none of the work described in this thesis could have been accomplished without him, and I owe him an incredible debt of gratitude. Jon was patient, generous with his time, and helpful with anything and everything that I needed as I worked my way through this project. Additionally, his ability to generate the raw data for the model, as well as experimentally validate its outputs in the lab, were instrumental in making this thesis feasible. He is an incredible scientist and immensely creative researcher, and it was an absolute privilege to be able to learn from him.

Finally, I would like to thank my family for being so supportive and encouraging of everything I do. I cannot even imagine where I would be without their love and support, and for that I am eternally grateful.

1
Introduction

Breakthroughs in DNA sequencing and synthesis technologies over the past decade have led to revolutionary advancements in virtually all fields of biology. Innovative experimental techniques, coupled with advances in computational modeling and analysis, have enhanced our understanding of the human genome, enabled the manufacture of innovative biologics and chemicals, and led to the development of novel therapies [46, 72].

The dramatically increased scale at which experimental data is being generated has substantially increased the need for computational tools that can efficiently process this data. Illumina's NovaSeq 6000, the most recent offering from its NovaSeq high-throughput sequencing series, can generate ~6 TB of sequencing data over the span of a couple days for roughly $72,000, despite being no larger than a typical office printer [41]. As a point of reference, that amount of data is roughly equivalent to 3,000 hours of HD video [19]. Thus, the interdisciplinary field of computational biology – which leverages computer science, statistics, and mathematical techniques to solve otherwise insurmountable challenges in analyzing biological data — has become an important driver for advancing a variety of subfields related to genomics, DNA synthesis, and biochemistry.

One area of research with particularly high therapeutic impact is the development of novel binders for small molecule and protein targets [70]. A “binder” is simply a substance that can chemically bind to another substance. If molecule *A* binds to protein *B*, then *A* is said to be a ”binder” for *B*. By developing binders for specific targets (e.g. toxins, pathogens, proteins, cell receptors linked to cancer, etc.), researchers can treat and diagnose disease, improve industrial chemical processes, and detect the presence of contaminants and toxins. Diagnostic applications for binders have become an increasingly important public health consideration, as the ongoing COVID-19 pandemic illustrates – the dearth of available diagnostic testing kits for COVID-19 has greatly hampered the efficacy of policies meant to curtail the spread of the virus, as well as make it nearly impossible to assess the true extent of the outbreak [13],

The discovery of novel binders has already been at the core of several recent breakthroughs in therapeutics and diagnostics, and represents a promising area of future research [43]. A key challenge that remains, however, is accelerating the laborious, time consuming, and costly process of discovering “hits,” i.e. strong binders that exhibit high affinity and specificity for their chosen targets.

1.1 Antibodies

A popular current method for hit discovery and applications is monoclonal antibodies. The monoclonal antibody market generated over $98B in global sales in 2017, and is expected to reach between $130-200B in annual revenue by 2022 [33].

Antibodies are Y-shaped proteins that have highly specific binding patterns [16]. This specificity allows them to be precisely targeted to particular biomarkers and cells, providing antibodies with many modalities of action. For example, antibodies can be used to disrupt signaling pathways associated with cancer by binding to key molecules in the signaling pathway, thus removing them from circulation [58]. Alternatively, antibodies can be used to directly target growth factors implicated in cancer, altering their structure once bound and thereby preventing them from functioning [58]. Antibodies can also be used to target the delivery of other non-specific therapies to certain cells or organs by chemically attaching them to the desired therapeutic agent [58]. The first monoclonal antibody therapy, Orthoclone OKT3, was approved by the FDA in 1986 [58]. Since then, monoclonal antibody therapies that have been approved have generally targeted oncology and hematology diseases [58].

Despite their prominence, however, antibodies are costly to produce and difficult to discover; these challenges have limited the ability of pharmaceutical companies to unlock the full potential of these binders [22]. Antibodies are difficult to mass manufacture because they can only be generated through *in vivo* processes (i.e. by being synthesized in living cells) rather than *in vitro* (i.e. created through a chemical reaction in a test tube) [89]. Having to synthesize antibodies *in vivo* dramatically increases their manufacturing complexity, as living cell lines must be incubated, maintained, and constantly monitored to ensure that they are kept in the optimal conditions for maximizing their yield [89]. This *in vivo* manufacturing process also increases the risk of contamination, as well as increased variability between batches in the quality of antibodies produced [89].

In addition to these manufacturing challenges, the chemical structure of antibodies themselves also severely limits their therapeutic potential, affordability, and accessibility [89]. Antibodies are large proteins (150-180 kDa), and injecting a high concentration of foreign protein into a patient as part of a targeted immunotherapy regime can sometimes trigger a variety of deleterious immunogenic responses from the patient's immune system [9, 23, 55]. Because the patient's body does not recognize the foreign antibodies injected, it tries to fight them off, thereby triggering an immune response [89]. The best case scenario when this happens is that the antibody therapy is rendered ineffective; this means other treatment options must then be pursued. In the worst case, however, the immunotherapy can lead to an allergic reaction and, potentially, death. Techniques like "humanization" have been developed to reduce the immunogenicity of antibodies; however, it is still nearly impossible to predict with certainty whether a given patient will exhibit an adverse immunogenic response to a given antibody treatment [17].

Even if immunogenicity were not an issue, the large size of antibodies reduces the rate at which tissues in the body can absorb them [89]. This requires higher doses than would otherwise be necessary, delaying the efficacy of the therapy while also increasing its costs. Antibodies' larger sizes also means that they have relatively low stabilities and shelf lives, and will permanently denature if stored at too high of a temperature [89].

In spite of these challenges, the development of monoclonal antibodies has been hailed as "one of the major medical breakthroughs of the 20th century...open[ing] endless possibilities...to diagnose, prevent, and treat a whole variety of diseases" [17]. Thus, it is exciting to imagine the medical potential of a technology that shares the powerful, highly-specific binding abilities of antibodies but resolves some of the challenges associated with their manufacture and chemical structure.

1.2 Aptamers

An "aptamer" is a single-stranded DNA or RNA molecule that binds to a specific target. The name "aptamer" comes from the Latin word *aptus*, which means "to fit," and the Greek word *meros*, meaning "part" [89].

DNA, or "deoxyribose nucleic acid," carries the genetic information for life. It is naturally found in all living cells in the form of a double-stranded helix of "nucleotides." A nucleotide is simply a chemical compound comprised of a nucleoside and a phosphate group. There are four nucleotides, also known as "bases," which can be found in a molecule of DNA – Adenine (A), Thymine (T), Cytosine (C), and Guanine (G). A strand of multiple nucleotides bound together is referred to as a "nucleic acid," hence the "NA" in the acronym "DNA."

The bases in a single-stranded molecule of DNA can bind with each other to form hydrogen bonds. More specifically, A binds with T while C binds with G. A is referred to as the "complement" of T, while C is the "complement" of G. This complementary base pairing is why DNA naturally forms a double-stranded molecule within cells – given a solution containing a single-strand of DNA and a high concentration of "free" nucleotides (i.e. nucleotides that are not already bound to a strand of DNA), the free nucleotides will tend to base pair with the unbound bases of the single-stranded DNA molecule. After a series of cellular processes for synthesizing DNA are then performed, these nucleotides will become fully integrated into that fragment of DNA, thereby making it a double-stranded molecule. For the purposes of this paper, the key concept is that while DNA is a natural substance and typically forms a double-stranded molecule in living cells, aptamers are single-stranded and are an entirely artificial construction.

RNA, or "ribonucleic acid," is also a nucleic acid composed of nucleotides. Unlike DNA, however, RNA is naturally found in living cells as a single-stranded molecule. Additionally, RNA has several chemical and structural differences from DNA – namely, the ribose sugar in RNA has one less hydroxyl group than the deoxyribose sugar present in DNA, and RNA uses Uracil (U) in place of Thymine (T). Functionally, DNA and RNA also serve quite distinct functions – whereas DNA simply serves as a storage device for genetic information, RNA performs a multitude of diverse roles to help express the information encoded in a cell's DNA through translation, transcription, and other processes.

While it is possible to construct aptamers from both RNA and DNA, because DNA aptamers are inherently more stable and easier to manufacture they have been the subject of most research in the field [89]. This thesis also focuses on DNA aptamers (henceforth referred to simply as "aptamers"), but it should be noted that the structural differences between RNA and DNA do not necessarily mean that the results of this paper would immediately generalize well to RNA aptamers.

Because aptamers are synthetically generated, they can theoretically target an incredibly wide range of substances, from ions to molecules to proteins, viruses, bacteria, and human cells [87]. Much like antibodies, it is their specific and targeted binding patterns which give aptamers their immense potential for therapeutic, diagnostic, and industrial applications.

In terms of size, aptamers are typically 20-100 nucleotides long (6-30 kDa) [89]. This makes them significantly smaller than antibodies (150-180 kDa) [89]. Due to their single-stranded structure, aptamers exhibit a high propensity for complementary base pairing among the nucleotides that make up their sequences. In other words, the unpaired nucleotides in an aptamer will tend to base pair amongst themselves, thereby causing the aptamer to fold into a highly unique 3D conformation. This structure is what determines the aptamer's binding "affinity" (i.e. how strongly it binds) and binding "specificity" (i.e. how precisely it binds only to its target) [63]. Thus, the sequence of A/T/G/C's which comprise an aptamer are sufficient to determine its function, albeit indirectly via the secondary and tertiary structures which that sequence encodes [35, 87].

The most widely utilized experimental method to generate aptamers was developed in 1990 by Tuerk and Gold through an *in vitro* process known as Systematic Evolution of Ligands by Exponential Enrichment, or "SELEX" [78].

It is this *in vitro* discovery process, as well as the smaller size, chemical composition, and ability to be synthesized *in vitro*, that fundamentally distinguishes aptamers from antibodies, and potentially offers a solution to the numerous challenges of monoclonal antibody development mentioned previously in 1.1 [4].

Unlike antibodies, aptamers do not require cell lines to be produced [89]. Instead, they can be chemically synthesized *in vitro* [89], allowing them to be mass manufactured at scale for lower cost [30]. With aptamers, researchers can avoid the laborious and expensive process of incubating and monitoring cell lines [30]. The avoidance of cell lines for manufacturing also reduces batch-to-batch variability between rounds of manufactured aptamers [30], as well as the risk of foreign cell colonies contaminating a vat full of nutritious growth media [89].

Because aptamers are generated through chemical reactions in a test tube, their production can therefore be more tightly controlled and optimized.

Because aptamers are also structurally much smaller than antibodies, they can be absorbed more quickly into the body's tissues [89]. Additionally, their smaller size enables aptamers to bind to smaller targets/binding domains that would otherwise not be possible for antibodies to reach [89]. This greatly broadens the range of targets to which an aptamer can bind; in some cases, aptamers can reach targets 10 times smaller than that which would be theoretically reachable by an antibody [4]. Additionally, aptamers have higher stabilities and longer shelf lives than antibodies, and can even be re-natured after denaturation [89]. This makes aptamers more resilient to environmental variability and substandard storage conditions [30].

Aptamers therefore offer a highly promising alternative to monoclonal antibodies, and their further development could help accelerate the field of targeted therapeutics and diagnostics while addressing the inherent obstacles faced by antibodies. Not only are aptamers cheaper to manufacture, but they can also target a wider range of molecules (expanding their therapeutic reach) and remain stable for longer in harsher environmental conditions (resolving supply chain difficulties).

Why, then, have aptamers not already replaced monoclonal antibodies?

Unfortunately, scientists face several lingering issues when working with aptamers. While the following three issues will not be addressed by this thesis (as they require more of a biological, economic, and legal perspective to resolve), they are worth briefly mentioning. First, the physiochemical properties of aptamers complicate their delivery in patients. Though aptamers are stable *in vitro*, they tend to degrade quickly *in vivo* (after roughly 10 minutes in living organisms) [89]. Additionally, aptamers are typically filtered out of the blood stream relatively quickly (after about 30 minutes versus 1 month for antibodies) [89]. While this is not an issue for many applications of aptamers, specifically those of diagnostic and industrial relevance, studies are currently being conducted to develop methods (such as cap modifications) to extend the time that aptamers can survive in patients [59]. Second, only one aptamer (pegaptanib) has yet been approved by the FDA for use in patients [89]. Thus, there is currently a lack of safety data/clinical precedence for their usage in humans, which has in turn made companies more hesitant to invest in the field. Third, patents on the SELEX procedure had initially slowed innovation in the field, although most of these protections have now expired [89].

The fourth, and most significant, issue facing the development and application of aptamer-based technologies is our relatively poor ability to discover and screen aptamers that bind well to specific targets [30]. As previously mentioned, once a researcher knows the sequence of the aptamer that the researcher wants to create, it is relatively easy to successfully manufacture that aptamer at scale *in vitro* [89]. **The challenge, however, is deciding which aptamer sequences should be manufactured** – in other words, once we know that a given aptamer binds well to a target protein, we can produce large quantities of it fairly quickly; however, the ability to first **discover** the appropriate aptamer for a given target is still a significant challenge.

The next section of this chapter provides more detail on SELEX, the current state-of-the-art experimental procedure for discovering aptamer sequences that bind well to a given target. Unfortunately, this procedure is error-prone, time intensive, low-throughput, and covers only a fraction of the total theoretical search space of aptamers that might bind to a target. Being able to accelerate this screening process could help chemists and biologists more fully unlock the potential of aptamers and increase the speed at which this promising technology is adopted.

1.3 SELEX

The current "gold-standard" experimental procedure for discovering aptamers that bind well to a specific target is Systematic Evolution of Ligands by Exponential Enrichment, or "SELEX" [89].

SELEX is an iterative process of directed selection that takes a pool of randomly generated aptamers, continuously increases the selective pressure on those aptamers' abilities to bind to a desired target (e.g. a protein associated with cancer), removes the aptamers that fail to bind to the target, then repeats this process over several rounds until the only aptamers that remain in the pool have an extremely high binding affinity for that target [20, 78]. Thus, in a nutshell, SELEX is essentially a "throw-everything-at-the-wall-and-see-what-sticks" style approach to aptamer screening.

Recall that an aptamer is simply a strand of nucleotides. And it is this unique sequence of A/T/G/C's which determines the binding affinity/specificity of that aptamer for its target (albeit indirectly through determining the physical structure of that aptamer) [35, 87].

Unfortunately, it is extremely difficult (if not impossible) to determine in advance which sequences of nucleotides will lead to 3D structures that bind well to a specific target [44]. Thus, the only way to discover high-binding aptamers is to combinatorially generate trillions of random aptamer sequences, test each one against the target, and hope that this combinatorial library contains an aptamer that binds well [4, 20, 78]. As this thesis demonstrates, deep learning models may offer a more rigorous means to explore this space of potential aptamer candidates.

Once the initial library of random aptamers is generated, SELEX works as follows: place the library of trillions of random aptamers in a pool, apply a selective pressure (i.e. the ability to bind to the protein target) to that pool, then wait until only sequences that were "fit" enough to survive (i.e. bind to the target) remain [78]. Isolate the fit sequences, place them in another pool with even higher selective pressure (e.g. increase the temperature, decrease the concentration of target available for binding, etc.), and repeat this process of removing unbound sequences and re-running the procedure with last round's surviving sequences [78]. This process continues for multiple rounds until the only sequences that are left in the final pool have extremely high "fitness" (i.e. binding affinity to the desired target) [15].

1.4 Problems with SELEX

Several obvious problems with SELEX make the discovery of high-binding aptamers difficult. Though not exhaustive, this section lists several of these core issues, grouping them into the three main levels at which they negatively affect the aptamer discovery process:

1. Execution

2. Experimental design

3. Data analysis

1.4.1 Execution

Like any *in vivo* or *in vitro* experimental process (as opposed to an *in silico* simulation), successfully running SELEX requires a trained chemist familiar with the laboratory environment as well as the necessary equipment for running the experiment.

Even with a well-equipped lab and trained personnel, however, SELEX is a very time- and labor-intensive process. A single run of SELEX typically consists of at least five rounds, where each round consists of isolating fit aptamers, sequencing them, amplifying them, and placing them in a pool configured to have a higher selection stringency [78]. Thus, one run of SELEX can take a month or more from start-to-finish [89]. This lengthy period between experimental design and results reduces the ability of chemists to iterate and improve on their aptamer designs, thereby slowing progress within the field.

This laborious process is also prone to technical error. Thus, tools that can help accelerate or side-step components of SELEX can help reduce the amount of time spent on executing experiments, repeating failed trials, and replicating results.

1.4.2 Experimental Design

The design itself of SELEX also creates several inherent limitations on the data that it generates. This severely restrains and complicates conclusions that can be drawn from it.

First, SELEX is intrinsically a low-throughput procedure. Each run of SELEX begins with a randomly generated pool of aptamers. Since no new aptamers are added to this pool during the experiment, the size of the starting pool is the upper theoretical limit on how much information can be gleaned from a single run of SELEX. For a specific type of aptamer (e.g. the set of aptamers with 45 nucleotides, the set of aptamers with 60 nucleotides and 40% GC content, etc.), these libraries tend to cover a small fraction of the total theoretical search space. For example, this paper will consider a set of aptamers that share a specific structure which, if one were to enumerate every possible sequence included in this set, would consist of roughly 10^{22} possible unique sequences. A single run of SELEX, however, can theoretically screen a library of up to 10^{15} sequences, although in practice this tends to hover around 10^{12}-10^{13} [89]. Thus, assuming in the best case that every sequence in our screening library were unique, each month-long run of SELEX would yield information about only 1 in every 10^{10} possible sequences, or one ten-billionth of the space of interest. Each run of SELEX therefore provides only a small amount of information about whether a better binding aptamer for a specific target might exist.

This lack of coverage is compounded by the fact that it is extremely difficult to generalize the binding performance of one aptamer sequence to another using statistical or computational analyses [44]. In other words, knowing that sequence *A* binds well to target *T* and that sequence *B* binds poorly to target *T* provides little reliable information as to whether sequence *C* will bind well to *T*, unless *A*, *B*, and *C* are virtually identical. Thus, enabling scientists to better generalize the results of a single run of SELEX could help amplify the power of this conventionally limited experimental procedure, thereby reducing the number of runs needed to cover a sufficient amount of an aptamer's theoretical search space.

Second, the SELEX procedure tends to be biased towards certain aptamer sequences independent of their binding affinities for the desired target [89]. Thus, weak-binding aptamers may be mistakenly identified as having high fitnesses [89]. Similar to how single-cell RNA-Seq generates an unprecedented wealth of valuable and high-resolution data but at the cost of inducing a number of significant biases and complications that must be taken into account before the data can be reliably utilized, the experimental procedures underlying SELEX also introduce a number of biases into the resulting dataset that must be controlled [61, 66, 86, 88]. There are biases introduced during the PCR amplification of sequences, biases introduced by the ligase, and biases introduced during the translation step of the experiment [89]. Furthermore, the addition of chemical functional groups bound to the DNA backbone of an aptamer (to create the "highly functionalized" aptamers that this paper specifically considers) adds another layer of potential biases, as these functional groups have different impacts on the translation/ligation efficiency of the underlying aptamer sequence to which they are bound [52].

Third, the rate of false negatives can be extremely high in SELEX due to the iterative nature of the procedure. A false negative in the context of aptamer screening would involve the inclusion, in one's initial library, of an aptamer sequence that binds very well to a target molecule but which, due to random chance, fails to bind to the target that round, and therefore gets filtered out of the pool of aptamers and thus permanently removed from consideration during later rounds of SELEX [89]. Each successive round of SELEX is seeded using the aptamer sequences that successfully bound to the target during the directly preceding round; thus, failing to bind to a target during an early SELEX round (i.e. 1st-4th rounds) effectively eliminates an aptamer from consideration as a top binder [78]. By solely considering the enrichment scores of later rounds, one may thus overlook sequences with favorable binding properties (making SELEX an even lower-throughput technology than its theoretical upper bound would suggest) [30].

The other key factor that makes the risk of false negatives substantial is the fact that the initial libraries used in SELEX tend to contain at most a handful of the same sequence [89]. Given some aptamer sequence *X*, if the one or two distinct fragments in your library representing that sequence *X* fail to bind to the target during the first round of SELEX, then sequence *X* will have a read count of zero not only for round 1 but also for every subsequent round. These types of events are referred to as ”drop-outs” in the RNA-Seq literature [64]. The stochastic failure of potentially only a handful of distinct oligonucleotides to bind to a target in an early round of SELEX, coupled with the fact that each successive round of SELEX uses the previous round's bound aptamers as the seed for its selection, means that the rate of false negatives with the procedure can be troubling and compounds with each successive iteration of the procedure [30].

Fourth, the issue of false positives can also complicate results. As with any physical experiment, there are multiple possible surfaces involved in the execution of the experiment that could be bound to – for example, the immobilization matrix to which the target molecule is attached during SELEX [87, 89]. An aptamer that bound to one of these non-target experimental components would appear to have "survived" the previous round of SELEX despite failing to bind to the desired target, and should have been filtered out of the aptamer pool. The solution is a slight variation on SELEX known as "counter-SELEX," in which a counter-selection against known potential environmental confounders is run during every round [4, 90]. This counter-selection will select for aptamers that bind to non-target surfaces, thereby enabling their removal from the pool. This method has been shown to yield aptamers with roughly 10 times more binding affinity to the target than traditional SELEX, and several other more involved variations of this core concept have since been developed [90].

The experimental procedure that generated the dataset used in this thesis therefore utilized counter-selections before every round to reduce the risk of false positives. Nonetheless, no counter-selection is perfect, and running a counter-selection also runs the risk of filtering out aptamers that bind well to both the target protein and the environmental confounder, thus potentially leading to additional false negatives.

1.4.3 Data Analysis

In addition to the aforementioned issues inherent in the very execution and experimental design of SELEX, there are also challenges that arise from the pre-processing, analysis, and validation of data generated by the procedure.

The first and most problematic issue is that data generated after each round of SELEX is in terms of the **relative** enrichment level of each aptamer sequence, not its true binding affinity for a given target. Thus, the data that SELEX generates doesn't actually measure the "ground truth" binding affinity of each aptamer being tested – rather, the data measures how well each aptamer binds to the target relative to the other set of randomly generated aptamers contained in the pool.

Ideally, one would be able to test each aptamer sequence individually by placing it in a test tube with the target and measuring how many rounds of increasingly stringent selections that aptamer were able to bind to the target. Unfortunately, it is simply impossible to do this sort of individualized experiment at scale, and thus the necessity and genius of SELEX in allowing scientists to test upwards of 10^{12} sequences at once.

The trade-off, however, comes in the relativistic nature of the data being generated, for each round of SELEX does not test how well a given aptamer can bind to a target, but rather tests how well an aptamer can out-compete the other aptamers in its library to bind to that target. Though this difference might seem pedantic (an aptamer that intrinsically binds well to a target should be able to out-compete other aptamers that don't bind as well for that target), the fact that the frequency of each aptamer at the end of every round depends heavily on how well every other aptamer in the library performed means that having a single extremely high-binding aptamer in a library can cause all other aptamers to appear to be poor binders (or, conversely, many poor binders can inflate the value of another weak but slightly stronger binder) and thus throw off an entire experiment.

By way of illustration, we can consider a library of 100 aptamers. Assume that 99 of these aptamers bind well to a target at roughly the same level. The remaining aptamer binds at roughly 10 times the level as these other aptamers. By the final round of SELEX, this highest-binding aptamer may account for over 80% of the total number of sequences left in the pool, for as its very name implies, SELEX is inherently an "exponential" enrichment procedure in which higher-binding sequences get amplified exponentially over the course of its many rounds. Thus, this single highest-binding aptamer will essentially crowd out the other 99 aptamers despite them also being inherently good binders for the target, leaving these 99 aptamers to split the remaining 20% of total binding activity amongst themselves. Thus, converting the relative enrichment scores yielded by SELEX into a "ground truth" binding fitness for each aptamer tested is a key open problem in the field. In addition to merely serving as a proxy for the true measurements we care about, the relativistic enrichment scores yielded by SELEX complicate our analysis by making every aptamer sequence's measured binding activity dependent on the performance of every other aptamer in the library. This also makes it extremely difficult to generalize the results from one run of SELEX to another, for the data generated by each run is, by definition, entirely dependent on the other unique sequences contained in that run.

Second, SELEX data is not a census of the aptamers that survive each round but rather a random sample that captures a small fraction of the set of surviving aptamers [89]. This can lead to the problem of "false 0's" (i.e. "drop-outs" in RNA-Seq parlance [64]) being reported for an aptamer sequence even though it was able to bind to the target and survive that round of selection, simply due to the fact that the sample of aptamers taken from that pool and submitted for high-throughput sequencing did not happen to contain that sequence [44]. Thus, even under the extremely generous assumptions that 1) every aptamer that binds well to a target managed to actually bind to that target, 2) every aptamer that does not bind well to that target failed to bind to the target and was successfully filtered out of the pool of aptamers, and 3) one had access to a perfect, 100% accurate high-throughput sequencing machine, because the input to that sequencing machine would be a relatively small, randomly sampled subset of the aptamers that successfully bound to the target, a researcher would still not be guaranteed to accurately measure either 1) all of the aptamers sequences that bound to the target (and thus there will necessarily be false negatives since most aptamers will occur at single-digit frequencies in our sample) or 2) accurately measure the relative frequencies of the aptamers that bound to the target protein (and thus not accurately capture the strength of each aptamer's binding affinity relative to the rest of the library tested)[89].

Both of these issues can be "solved" by simply making the modeling assumption that the random sample taken from the pool of aptamers reflects the true underlying distribution from which it was drawn – and this is the approach taken in this paper – however, it is important to note that the necessary experimental step of first translating and amplifying the random sample of aptamer sequences taken from the pool of surviving aptamers after a round of selection induces biases that will complicate this assumption [44, 89].

Third, each round of SELEX is entirely dependent on the results of the previous SELEX round, and thus every measurement has a high level of temporal dependencies that should be taken into account [78]. The enrichment level measured for each aptamer after every round is directly tied to the previous enrichment level measured for that aptamer sequence, for the previous round's surviving sequences are used as the seed for the following round. Failing to recognize this dependence can lead to misinterpretations of how well an aptamer performed [44].

For example, if aptamer *X* were measured to have 1,000 copies present in the pool after Round 4, one might assert that it would clearly be a better binder for the target than aptamer *Y*, which had only 200 copies remaining after Round 4. But what if you also knew that aptamer *X* was measured to have 2,000 reads in Round 3, whereas aptamer *Y* only had 20 reads in Round 3? Given this new information, it would appear that aptamer *Y* is actually a much better binder for the target, as the concentration of aptamer *X* was halved between rounds as the stringency of our selection increased, whereas aptamer *Y* actually experienced a

10-fold increase in its abundance between rounds. Thus, ignoring the temporal dependencies of SELEX data, as well as the fact that different sequences can begin with a different number of fragments due to random chance, can lead to misleading results.

One solution to this issue (which is utilized in this paper) is to simply consider the fold change between rounds of an aptamer's enrichment levels [49]. This accounts for differences in starting concentrations while also collapsing the multiple time points that represent each round's selection (before and after) into a single summary statistic. This method is detailed further in 3.1.1.

Fourth, in order to validate the conclusions of a run of SELEX, it is necessary to perform additional time-intensive, low-throughput assays like surface plasmon resonance (SPR), electrophoretic mobility shift, filter-binding, flow cytometry, or microscale thermophoresis (MST) in order to determine the "ground truth" binding fitness of an aptamer candidate [30].

SELEX is a highly stochastic procedure, and thus is extremely sensitive to even the tiniest environmental fluctuations (especially regarding false negatives at the earliest rounds). Additionally, as previously detailed, SELEX only measures how well an aptamer binds relative to the other aptamers contained in its library. In order to actually determine whether (and to what extent) a "top hit" – i.e. an aptamer with a high enrichment score – from a run of SELEX is actually a strong binder for the target, one must re-synthesis the corresponding aptamer and then individually test its binding ability using one of the aforementioned higher-accuracy assays [30].

In summary, there are many issues with SELEX that have limited the ability of researchers to screen aptamers at scale with consistency. **Perhaps the issue of greatest importance, though, is the low coverage of the theoretical search space yielded by each run of SELEX**. For even if all of the other issues were remedied and SELEX could yield error-free information on the "ground truth" fitnesses of every aptamer tested, each month-long run of SELEX would still only yield information on less than 1 in 10^{10} of the theoretical space of potential aptamers. This lack of coverage means that the vast majority of strong-binding aptamers will be missed by researchers, making it impossible to optimize the technology for medical and diagnostic applications.

Thus, the principal focus of this thesis centers on resolving the core issue of low coverage that has plagued the field of aptamer discovery. By constructing a deep learning model that can more fully explore the fitness landscape of aptamers and thereby generate entirely novel high-binding sequences *in silico* without the need for wet lab experimentation, this thesis presents a constructive approach towards removing a significant impediment that has delayed and constrained the progress of this promising field.

1.5 Motivation

Computational tools that can assist researchers in overcoming the limitations of SELEX could help accelerate the discovery of high-binding aptamers. This thesis aims to contribute to the emerging discourse at the intersection of biochemistry and computer science by applying deep learning techniques to a well-defined case of aptamer discovery, as well as addressing some of the broader conceptual issues raised when employing these methods.

It is extremely difficult to generalize the results of a single run of SELEX [4]. While a single run may provide accurate information on the sequences tested (as well as insight as to how aptamers which share extremely similar sequences might perform), generalizing these results (which cover only one ten-billionth of the total theoretical search space) to a broader range of possible aptamers is currently an unsolved problem.

Thus, this thesis sought to construct a generative deep learning model that could learn the complex fitness landscape of aptamer binding for a specific target based on training data from a single run of SELEX. Then, the model could generate novel aptamer sequences sampled from this fitness landscape, and thereby amplify the amount of information that could be harnessed from this widely used, decades-old experimental technique. By assisting in the discovery of novel and diverse aptamers that would otherwise go unconsidered, efforts like this thesis can help accelerate the development of more affordable, scalable, and precise diagnostic tool kits, vaccines, and targeting methods for therapeutics that leverage aptamers.

1.6 Related Work

1.6.1 *In Silico* Screening Methods

Prior work has attempted to use computational tools to accelerate the time-consuming process of screening aptamers. In general, the most widely adopted models leverage one of the following three techniques to compare and rank sequences against one another:

1. Sequence clustering
2. Structural motif-based clustering
3. Molecular dynamic force field simulations

Sequence Clustering

Sequence clustering tools identify and leverage similarities among the actual sequences (A/T/G/C's) of different aptamers in a SELEX pool in order to group them together, and thereby gain a more robust estimate for the binding performance of closely related sequences. These methods are computationally fast since they treat aptamers as simple strings, and therefore leverage previously developed highly efficient string comparison algorithms. For example, FASTAptamer and PATTERNITY-Seq are two popular tools which use Levenshtein distance to cluster sequences [2, 44, 44]. Levenshtein distance is a string-similarity measurement that is determined by calculating the minimum number of insertions/deletions/substitutions needed to convert one sequence of characters into another. Another commonly used tool is AptaCluster, which is part of the AptaSuite bioinformatics package [44]. AptaCluster leverages locality sensitive hashing and *k*-mer counting to assess the level of similarity between aptamer sequences [38].

By treating aptamers as mere strings of A/T/G/C's, these sequence clustering models are able to dramatically reduce the time needed to analyze large SELEX datasets. However, these simpler algorithms also greatly limit the accuracy of these methods, as they are too simple to detect many of the complex relationships between an aptamer's sequence and its binding performance. These clustering models do not incorporate domain knowledge about aptamer binding into their predictions, and they fail to capture any of the secondary structural information that plays a significant role in determining an aptamer's binding fitness.

Structural Motif-Based Clustering

Models that leverage structural motifs represent the next step in predictive sophistication. These tools attempt to predict the secondary structural conformation of each input aptamer sequence, cluster them based on shared structural motifs, and make predictions of a sequence's binding affinity based on its similarity to already-seen sequences. AptaTrace, which like AptaCluster is part of the AptaSuite package, is one of the most widely used such tools [44]. AptaTrace tries to associate each structural motif observed in a library of aptamers with its impact on enrichment levels [14]. It accomplishes this by segmenting each sequence into a series of *k*-mers for various values of *k*, simulating how each *k*-mer would be predicted to fold, and associating *k*-mers with similar conformations to each other [14]. APTANI and MFold are alternative programs that are able to cluster aptamers based on their predicted secondary structures [10, 92].

While these models offer more sophisticated analyses than those which can be provided by sequence clustering models, they tend to take significantly longer to run due to the increased computational costs of having to first predict secondary structures before being able to cluster aptamers based on structural motifs.

Additionally, the very fact that both sequence clustering and structural motif-based clustering models use **clustering** as the basis for prediction means that they will be inherently biased towards aptamers that are highly similar to sequences which were already observed in the training set. Thus, while valuable for analyzing and better understanding the performance of the set of aptamers contained within a run of SELEX, these tools are poorly designed for helping researchers generalize the findings of a run to a broader segment of the total theoretical aptamer search space, and thus cannot readily assist in the design/evaluation of entirely novel aptamer sequences. As mentioned previously, because a single run of SELEX covers roughly only 1 in 10^{10} of all possible sequences, this clustering approach greatly limits the ability of researchers to take advantage of the true power of aptamers – namely, the immense combinatorial space of unique structures (and thus functional properties) of aptamers. By limiting our search to only those aptamers which are highly similar – either sequence-wise or structurally – these clustering methods limit our ability to generalize SELEX results and access a significant portion of the total search space for aptamer sequences. This consideration was a key driver behind the decision to utilize an alternative approach, a generative deep learning model, for this thesis.

Molecular Dynamic Force Field Simulations

A third major approach to identifying high-binding aptamers applies molecular dynamic (MD) force field simulations to the target molecule and aptamer of interest. Two of the most popular simulation tools are CHARMM and AMBER, which estimate at an atomic level the potential energy of the various binding components of the target and aptamer in order to predict binding strength [8, 44, 68]. Additionally, 3D structural prediction tools like Rosetta have been used in conjunction with computational docking tools like DOVIS to identify high-binding aptamer constructs [15, 42, 67].

While the dynamics of protein-protein interactions have been extensively studied, modeled, and refined within these software packages, these tools unfortunately lack the same rigor when predicting DNA-protein interactions [44]. Unfortunately, this is precisely the scenario that occurs when aptamers (strands of DNA) bind to their targets (typically proteins). Thus, the accuracy of these tools is reduced when applied to the domain of aptamer screening. Additionally, in order to make their predictions, these models first require the full 3D structures of both the aptamer and target [44]. This is extremely computationally demanding and infeasible to conduct at the scale of SELEX experiments [44]. Given these issues, while MD force field simulations are valuable tools for fine-tuning specific aptamer sequences and testing interactions on a small scale, they are unfortunately ill-prepared for providing scientists with a robust method for exploring the fitness landscape of aptamers [44].

Finally, in addition to the aforementioned problems concerning computational speed, accuracy, generalizability, and the ability to fully capture the complex relationships between sequence and binding fitness, all three of these aforementioned conventional *in silico* approaches are currently incapable of modeling the specialized type of aptamer studied in this paper, highly functionalized nucleic acid polymers (HFNAPs). HFNAPs contain additional functional groups attached to their core nucleotide sequence of A/T/G/C's [36], a unique feature for which these existing models were not developed.

1.6.2 Deep Learning

Recent advances in deep learning may offer a solution to many of the challenges that hinder the aforementioned *in silico* screening methods.

Deep learning refers to the utilization of artificial neural networks containing many inter-connected layers of neurons to learn complex supervised, semi-supervised, or unsupervised tasks [56]. Deep neural networks have demonstrated the ability to achieve unprecedented performance on a variety of difficult tasks, from text translation [82] to autonomous driving [39] to protein structure prediction [6]. Deep networks are theorized to have achieved these feats by imitating the hierarchical nature of how we believe that the human brain processes low-level stimuli into abstract thoughts, ideas, and conceptualizations – by learning increasingly higher-level representations of their inputs through multiple layers of neurons, deep neural networks learn to understand complex higher-order relationships between features, and thereby achieve impressive performance on a broad range of difficult tasks [56].

Deep learning has already been successfully applied to the subfield of optimizing the design of biologics. For example, the field of protein engineering suffers from the same scarcity issues that plague aptamers. Proteins are simply sequences of amino acids chained together, and just like the vast majority of aptamer sequences will not bind well to a given target, the vast majority of protein sequences will lack the functional properties desired. Echoing some of the exact concerns discussed previously in 1.4 for aptamers, Yang et al. notes of proteins that, "highly functional sequences are vanishingly rare and overwhelmed by nonfunctional and mediocre sequences ... [yet] even the most high-throughput screening or selection methods only sample a fraction of the sequences that can be made" [84]. Yang et al. cites several machine learning approaches, ranging from Gaussian processes to random forests to neural networks, which have been employed within the field of protein engineering to resolve this issue [84]. In particular, the authors highlight the efforts of Sinai et al.[74] and Riesselman et al.[65] in employing a specific deep learning technique, a variational autoencoder (VAE), to successfully learn the functional landscape of protein sequences and thereby model the impact of specific mutations on protein function [84].

A VAE is a generative unsupervised learning model that was originally introduced in 2013 by Kingma et al. [45] A VAE is comprised of two parts – an "encoding" segment and a "decoding" segment. The "encoding" segment takes as input a protein or DNA sequence, then "encodes" that input sequence into an alternative, more compressed representation. This condensed representation of the input is known as its ”latent" representation, for we do not observe what the true values for these compressed representations of the target inputs should be (and thus they are latent variables). Thus, the challenge for the VAE is to learn, in an unsupervised manner, how to compress its inputs into a constricted latent space without losing information.

The "decoding" segment of the VAE does the exact opposite – it takes a latent representation as its input, then tries to "decode" that compressed representation back into the original sequence that was fed into the encoder to generate that latent encoding. The encoding and decoding segments are typically implemented as densely connected neural networks, and additional detail on the statistical underpinnings of this model are provided in 3.2.1.

In Sinai et al., the authors constructed a VAE in which both the encoding and decoding segments were comprised of three layers of 250 exponential linear units [74]. The inputs were compressed onto five latent variables, and the VAE was optimized with ADAM [74]. The authors' overarching goal was to train the VAE to predict the impact of protein sequence mutations on the protein's function [74]. In terms of performance, the trained VAE was able to outperform baseline methods which assumed independence between sequence locations, as well as (in some cases) "state-of-the-art" methods leveraging the inverse-Potts model [74]. One key limitation of the VAE that the authors note, however, was the lack of any recurrences built into its architecture which likely resulted in decreased accuracy [74]. Thus, unlike Sinai et al., this paper's final model involved recurrent features in the form of a long short-term memory network (LSTM), as detailed in 3.2.2.

In Riesselman et al., the authors built a slightly modified VAE called DeepSequence to also predict the impact of mutations on protein function [65]. The encoding segment of this model was comprised of three fully connected layers of size 1,500, 1,500, and 60 nodes using ReLU activations, while the decoding segment consisted of two hidden layers of size 100 with ReLU activations followed by a layer of 2,000 nodes with sigmoid activations [65]. Just as in the case of Sinai et al., the more sophisticated deep learning model utilized in this paper was able to capture non-linear dependencies between sequence locations better than models which assumed independence or only considered pairwise dependencies [65]. As the authors note, extending these simpler models to include higher-order interactions would be "statistically unfeasible," requiring over 1 billion parameters to model 3rd order interactions terms for proteins of only 100 amino acids [65]. Thus, deep learning models like DeepSequence which leverage latent variables to efficiently learn complex, higher-order dependencies uniquely enable researchers to broaden their search for optimized protein sequences, and thus take fuller advantage of the complex relational information contained within each sequence [65]. The VAE has thus proven to be a promising new method for the principled engineering of proteins.

Deep learning methods have also been successfully applied to the optimization of antibody targeting, which, as previously described in 1.1, is a subfield that is closely related to the challenge of screening aptamers. Similar to SELEX, existing experimental methods for discovering antibodies rely on combinatorially generated libraries [54]. Thus, methods that have been successful within the realm of antibody targeting may also perform well for aptamer targeting, as the two procedures seek to optimize the same basic objective – discovering strong binders – despite dealing with two very different types of molecules.

Liu et al. used an ensemble of convolutional neural networks (CNNs) to optimize the construction of Immunoglobulin G antibodies beyond the binding affinities achievable using the standard experimental discovery technique of phage display [54]. By training on data containing the enrichment levels of each antibody sequence for a particular target, the authors avoided the need to model the molecular structure of the antibody and the target itself [54]. This costly structural modeling step would have been required with other computational approaches [54]. By solely training on sequence instead of structure, the authors were thus able to greatly expand the size of their training set [54]. This architectural decision motivated the approach taken in this thesis of solely using an aptamer's sequence and not its structure as training data.

Liu et al. started with a library of 10^{10} antibodies whose CDR-H3 segments contained between 10 to 18 randomized amino acids [54]. The authors then ran three rounds of phage display panning and measured the enrichment level of each antibody sequence using high-throughput sequencing [54]. The authors decided to use the log fold change enrichment measured between Rounds 2 and 3 for each sequence as a measure of its binding fitness [54]. This paper also used log fold change as an enrichment metric, and a definition of the metric is provided in 2.2.5. By building an ensemble of 18 different neural networks of various architectures – 15 convolutional, 3 fully connected – which were collectively referred to as "Ens-Grad," the authors were able to successfully identify antibody sequences that performed better than those observed experimentally during their initial phage display panning [54].

Other labs have also had success at applying deep learning techniques to later stages in the antibody development pipeline. In general, once an antibody "hit" has been discovered and had its binding affinity optimized through phage display (plus whatever computational tools may have assisted, e.g. Ens-Grad), the selected antibodies must then be further optimized for therapeutic deployment in humans. This final step requires the selected antibodies to be incubated in mammalian cells, an experimental process that dramatically reduces the throughput of antibody screening – typically only 10^3 sequences can be screened at once [57]. This is a tiny fraction of the total throughput of phage display which, in turn, is a tiny fraction of the total search space of antibody sequences [57]. For example, the modest stretch of 18 randomized amino acids investigated in Liu et al. presents over 10^{23} total sequence possibilities [54].

Thus, Mason et al. decided to leverage deep learning methods to optimize this later stage in the development of antibodies [57]. Beginning with a library of identical copies of the antibody trastuzumab, the authors applied CRISPR-Cas9-mediated mutagenesis to generate a training set of $5 * 10^4$ variants, each containing a unique CDR-H3 sequence [57]. An LSTM with three hidden layers (each containing 40 nodes) was trained using RMSprop, while a separate CNN with one dense layer of 50 ReLUs was trained using ADAM [57]. The output of both models was a binary prediction of whether the input antibody was a "binder" or "non-binder" for the desired target, and both sought to optimize binary cross-entropy as their objective functions [57].

To ensure that these neural networks sufficiently captured the total theoretical search space (and were thus not simply returning sequences that appeared similar to sequences contained within the training dataset), the authors added the constraint that all generated sequences had to have a Levenshtein distance of at least 5 from the original antibody (trastuzumab) that had been used as the seed for their mutagenesis library [57]. This thesis would also utilize this technique of using Levenshtein distance to assess how well models had generalized. Mason et al. then selected 30 antibodies that both models had agreed were "binders," and experimentally determined their binding affinities using flow cytometry [57]. All 30 of these computationally generated antibodies were confirmed to be binders for the target, with one particular variant showing a 3-fold increase in binding affinity compared to the original trastuzumab sequence [57]. This promising result demonstrated the potential for deep learning methods to generate novel high-binding sequences and allow for a more comprehensive exploration of the fitness landscape for a binder.

In summary, deep learning techniques like LSTMs, CNNs, and VAEs have already been successfully applied to the optimization of protein and antibody sequences. Based on this success, deep learning appears well-suited for capturing the complexity of the relationship among an aptamer's sequence, sidechains, and binding performance.

Despite their promise, however, deep learning techniques have not yet been applied to the field of aptamer discovery. One key reason is that aptamers are strands of DNA, whereas proteins and antibodies are both peptides. Thus, the computational models already developed in other subfields of biochemical optimization do not easily translate across domains [44]. Additionally, there is simply less experimental research overall being conducted on aptamers, which has in turn slowed the development of computational tools to assist such research [89]. Given the potential promise of aptamers, however, this lack of development is a serious mistake. Building off the prior work discussed in this chapter, the unique contribution of this thesis is thus the utilization of deep learning techniques for the prediction of novel aptamers with high affinities for small molecule targets.

2

Biological Problem

The field of aptamer discovery is incredibly broad, as there are many different types of aptamers and dozens of variations on SELEX that have been developed [4]. This thesis considers one very specific instance of this broader area of research, and uses it as a proof-of-concept for the power of applying deep learning methods to aptamer screening. This chapter describes the set-up for that specific biological problem, as well as the actual dataset used as the foundation for this research.

The first section of this chapter describes the particular type of aptamer that was studied, highly functionalized nucleic acid polymers (HFNAPs). The second section describes the actual data that was considered by the model, as well as the experimental procedure used to generate it.

2.1 Highly Functionalized Nucleic Acid Polymers (HFNAPs)

An HFNAP is comprised of a nucleotide backbone (i.e. a traditional aptamer) that has had various chemical functional groups attached as sidechains [36]. These functional groups are what distinguish HFNAPs from traditional aptamers. The dataset analyzed in this thesis was comprised of fully functionalized 45-mer HFNAPs.

2.1.1 Nucleotide Backbone

Traditional DNA aptamers are single-stranded oligonucleotides composed of nucleotides (A/T/G/C's). In this paper, the specific type of aptamer considered was a "45-mer," which means that each aptamer contained exactly 45 such nucleotides. Importantly, however, the first nucleotide in every "trimer," or set of three nucleotides, was restricted to be either a T or a C. Thus, every trimer was of the form "YNN," where "Y" represents T/C and "N" represents A/T/G/C under the IUPAC bioinformatics code. Thus, the total number T of possible aptamer sequences that could constructed using this structural template was:

$$T = 2^{15} * 4^{45-15} = 2^{75} \approx 10^{22.57}$$

Given the fact that a single run of SELEX can test at most 10^{12} to 10^{13} sequences, this means that each month-long run of SELEX would cover only about 1 in 10^{10} possible sequences [89]. Additionally, since most sequences included in the initial library will drop-out (i.e. be removed from the pool after the first couple rounds) due to chance, we will not be able to collect any meaningful data on them. Thus, the actual coverage of the total theoretical search space that is provided by a single run of SELEX will actually be much smaller than 1 in every 10^{10} possible sequences for this particular experimental set-up.

2.1.2 Functionalization

Whereas traditional aptamers simply contain nucleotides, HFNAPs contain functional groups (e.g. phenols or alcohols) attached to a strand of nucleotides [36]. These functional groups are also sometimes referred to as "sidechains." Thus, one can think of an HFNAP as simply a traditional aptamer that has been accessorized by the addition of chemical sidechains to its "backbone" of nucleotides.

Existing literature on aptamers refers to HFNAPs as a type of "base-modified aptamer" [29]. Compared to other types of base-modified aptamers, however, the HFNAP system considered in this paper allows for more comprehensive and targeted testing of the impact of sidechain inclusion by allowing researchers to specify a broader range of sidechains to be included at every location of the molecule [11, 36, 52].

The theoretical advantage offered by HFNAPs (and base-modified aptamers more generally) over traditional aptamers is stronger binding affinity for target molecules, and thus more optimized functional properties [11]. The addition of functional groups also provides HFNAPs with significantly higher chemical diversity, and thus a wider range of possible structural conformations and binding properties [52]. For example, HFNAPs can bind to molecules that would not otherwise interact with DNA, and would thus be otherwise entirely ignored by traditional aptamers [29]. HFNAPs could thus serve as a valuable innovation allowing researchers to more fully bridge the gap that currently exists between the binding performances of aptamers and antibodies [29].

At the same time, however, the addition of functional groups makes predicting the binding activity of HFNAPs even more difficult than it already is for traditional aptamers, for these sidechains heavily impact the structural conformation of the nucleotide backbones to which they are bound [52]. Additionally, HFNAPs are more difficult to synthesize due to their increased complexity, and thus there remains an unresolved tension between better binding affinity and synthesis efficiency [29].

First developed by Hili et al. in 2013, HFNAPs are constructed using trimers in which the first nucleotide contains a nucleobase functionalization [36]. The term “functionalization” refers to the attachment of a sidechain to a nucleotide. By definition, every 45-mer contains 15 trimers, and since only the first nucleotide in each trimer can be functionalized, this means that every HFNAP considered in this thesis had 15 possible locations for sidechains.

As mentioned previously, the HFNAPs considered in this paper had every trimer begin with either a C or T for stability reasons. Trimers can be synthesized with a variety of sidechains. For this paper, 8 different sidechains were considered. Each of these 8 sidechains was associated with 4 distinct trimers. Thus, each of the 32 possible trimers would be associated with only one sidechain. Within a single run of SELEX, each trimer would be either always functionalized (i.e. always bound to its corresponding sidechain) or completely unfunctionalized (i.e. never associated with its sidechain). The mapping between each of these 8 sidechains and their 4 corresponding trimers, as well as the sidechains’ chemical names and structures, is illustrated in Figure 2.1.1.

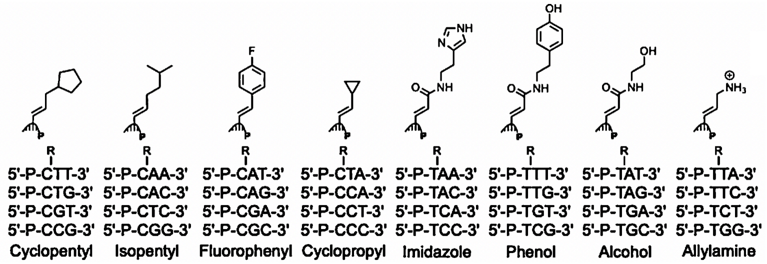
F
OH
NH
N
OH
O
NH
O
NH
O
NH
NH3
P
R
5'-P-CTT-3' 5'-P-CAA-3' 5'-P-CAT-3' 5'-P-CTA-3' 5'-P-TAA-3' 5'-P-TTT-3' 5'-P-TAT-3' 5'-P-TTA-3'
5'-P-CTG-3' 5'-P-CAC-3' 5'-P-CAG-3' 5'-P-CCA-3' 5'-P-TAC-3' 5'-P-TTG-3' 5'-P-TAG-3' 5'-P-TTC-3'
5'-P-CGT-3' 5'-P-CTC-3' 5'-P-CGA-3' 5'-P-CCT-3' 5'-P-TCA-3' 5'-P-TGT-3' 5'-P-TGA-3' 5'-P-TCT-3'
5'-P-CCG-3' 5'-P-CGG-3' 5'-P-CGC-3' 5'-P-CCC-3' 5'-P-TCC-3' 5'-P-TCG-3' 5'-P-TGC-3' 5'-P-TGG-3'
Cyclopentyl
Isopentyl
Fluorophenyl
Cyclopropyl
Imidazole
Phenol
Alcohol
Allylamine

Figure 2.1.1: Mapping of each sidechain (chemical structure shown in top row, name in bottom row) to its 4 corresponding trimers. All sidechains are attached to the first nucleotide of their respective trimers. Figure is taken from Chen et al. [11]

When synthesizing HFNAP molecules, a set of sidechains (a “sidechain set”) must first be selected for inclusion in the HFNAP library. Not every one of the 8 aforementioned sidechains must be included in this set. For this paper, however, the HFNAPs considered were all synthesized using the full set of available sidechains. Thus, every trimer was functionalized by its corresponding functional group.

Because every trimer is associated with only one functional group, one can reconstruct which sidechains were present in an HFNAP by simply sequencing the nucleotide backbone of that HFNAP. Following the mapping of Figure 2.1.1, if an HFNAP incubated using a fully functionalized sidechain set had the sequence "CTTCAATTA," then one would immediately know that this molecule contained a cyclopentyl sidechain followed by an isopentyl sidechain followed by allylamine.

2.2 Dataset

This section describes the actual dataset that was utilized by the machine learning models of this thesis, as well as the experimental procedure used to generate the relevant data. Understanding the generative process behind the dataset heavily informed the modeling choices discussed in this paper, and allowed for several simplifying assumptions to be made based on domain knowledge and empirical observations. The data, a single run of SELEX for eight rounds against the target molecule daunomycin, was generated entirely by Jon C. Chen and will be published in a paper that, as of April 3rd 2020, is currently in preparation.

While it was necessary to perform this initial run of SELEX in order to generate the training data for the machine learning models, the over-arching goal of building these predictive models was to amplify the strength of conclusions that could be drawn from this singular dataset, and thereby avoid the need to conduct additional lab work to broaden the scope of aptamers considered.

2.2.1 Target: Daunomycin

A molecule that would benefit from the development of a complementary binder is daunomycin.

Daunomycin is an anthracycline antibiotic, and is one of the most commonly utilized small-molecule chemotherapeutic agents [81]. The World Health Organization included the molecule on its 2019 List of Essential Medicines [60].

Unfortunately, however, daunomycin can be both toxic and carcinogenic to the human body at high enough doses [7]. It can also cause deleterious side effects ranging from nausea to vomiting to stomatitis [7]. The necessary dosage of the drug varies from patient-to-patient, and thus it can be difficult for doctors to precisely adjust their administration of the drug for each individual case [77]. Doctors must also continually monitor daunomycin levels in treated patients, in order to detect fluctuations in its concentration level and thereby minimize the risk of deleterious toxic or carcinogenic effects [81].

Traditional methods – e.g. capillary electrophoresis and high-performance liquid chromatography – for measuring daunomycin levels in patient samples, however, require "a high level of instrumentation and qualified handling skills" which "hampers fast and convenient monitoring" [81]. Thus, aptamers have been studied as a technology to create potentially cheaper and easier-to-use diagnostic tools for the detection of daunomycin in cancer patients, and thereby improve therapeutic outcomes while decreasing the cost of care [81].

While the lessons of, and methods taken in, this thesis reflect an approach that would also apply in general to the broader field of aptamer discovery, for concreteness this thesis specifically examines daunomycin as a target molecule of HFNAPs functionalized according to Figure 2.1.1.

2.2.2 HFNAP Synthesis

As detailed in the previous section, the HFNAPs considered in this paper were 45-mers attached to a series of sidechains. These sidechains were selected from a set of 8 functional groups, and each sidechain was associated with 4 distinct trimers.

The process to generate these HFNAPs, at a high-level, was as follows. First, a set of $6 * 10^{13}$ random 45-mer DNA templates was generated and placed in solution. Trimers that were bound to their associated functional groups were then introduced into the solution. These free-floating trimers would then hybridize to complementary sequences located on the DNA templates (e.g. the "TCC" trimer would bind to "AGG" on the DNA template) [24]. After ligating these hybridized trimers together with T3 DNA ligase and separating them from their DNA template using alkaline denaturation, the library of HFNAPs was created [36].

2.2.3 SELEX

The HFNAP library was run through 8 rounds of SELEX selections with daunomycin as the target. Streptavidin beads were used to immobilize the daunomycin in solution.

Before each round, a counter-selection was performed in which the HFNAPs were incubated with blank streptavidin beads and biotin-linkers. HFNAPs that bound to these substances were removed, thus helping to reduce the rate of false negatives, as discussed in 1.4.

The flow-through from the counter-selection was then isolated and incubated in a solution with streptavidin-immobilized daunomycin. The flow-through from this solution contained HFNAPs that did not bind to daunomycin, and thus they were removed. After eluting the bound HFNAPs, polymerase chain reaction (PCR) was used to "reverse translate" the HFNAPs into complementary strands of DNA. Polyacrylamide gel electrophoresis (PAGE) was then used to remove truncated amplicons.

Each HFNAP molecule contains, in addition to a nucleotide backbone of A/T/G/C's, several chemical modifications to the beginning and end of its backbone, in addition to whatever sidechains are attached. Reverse translating an HFNAP molecule into a strand of DNA essentially strips away these extraneous modifications, leading to a DNA molecule that reflects solely the nucleotide backbone of the original HFNAP.

Why is this reverse translation step necessary? First, it allows us to sequence the HFNAPs using conventional high-throughput DNA sequencers and therefore measure the relative frequencies of HFNAPs after running the positive selection against daunomycin. To borrow a metaphor from computer science – because high-throughput sequencing machines can only process pure DNA, this reverse translation step allows us to sequence our HFNAP molecules by first "compiling" them down to a language (DNA) that sequencing machines understand. Since we know that every trimer in a library will either be always bound to its associated sidechain or never bound to that sidechain, we don't actually lose any information about the sidechains attached to each HFNAP during this reverse translation step. Second, reverse translating HFNAPs back into DNA enables us to conduct future rounds of selection. Because of the increased selection stringency of each round of SELEX, the number of reads of HFNAPs in our pool will eventually converge to zero. Additionally, reads are removed from the pool after each round for sequencing, and there is an inevitable level of dilution and loss of sequences due to experimental conditions. These factors mean that in later rounds of selections, there will be significantly fewer sequences in the pool remaining than there were at the start, thereby making it harder to detect the binding activity of each HFNAP. The solution is to re-amplify every sequence after each round so that it is present in a high enough concentration to both bind to the target and be detected. Reverse translation allows us to amplify complementary DNA strands before translating them back into HFNAPs, thereby ensuring that there are enough HFNAPs each round to generate measurable data even as we filter out a substantial portion of the HFNAPs that do not bind. Note that, as the name implies, the process of reverse translation is the reversal of the process used to initially synthesize the HFNAPs – whereas before we converted DNA templates into HFNAPs, here we are doing the opposite.

Once the HFNAPs are reverse translated back into DNA, the DNA strands were then amplified using PCR so that we could 1) remove a subset of them for sequencing and thus get a measure of how frequent each HFNAP sequence was post-selection, and 2) create a sufficient concentration of DNA strands such that we could successfully translate them back into HFNAPs and conduct another round of selection.

After PCR amplification, a random sample of DNA strands were isolated from the overall pool of amplified DNA. These sequences were then run through a high-throughput sequencing machine to generate a file containing millions of individual sequences (referred to as "reads") as well as a count of how frequent each read appeared in the DNA sample.

The above process was repeated 8 times, yielding 8 distinct sets of sequencing results. Between each successive round, the "selection stringency" of the procedure was increased, meaning that it became harder for an HFNAP to bind to daunomycin in the $(i+1)$th round than it would have been for that same HFNAP to bind in the ith round. The first 4 rounds were conducted under relatively low stringency conditions, while the last 4 rounds were conducted under a much higher stringency. This approach was taken to reduce the probability of false negatives from high-binding sequences dropping out early in the SELEX process [50].

2.2.4 Raw Data

Unfortunately, the raw count of how many times each HFNAP appears in the random sample taken after every round of SELEX is not very meaningful. That's because these raw read counts depend heavily on the efficiency of the PCR step, which can vary widely between SELEX experiments, as well as the number of reads included in the sample. The solution is to convert each of these absolute read count scores into relative read counts [35]. This controls for experiment-to-experiment variation in PCR amplification, as well as the size of the sample taken from the overall pool of HFNAPs that remained post-selection.

Let us assume that there were n unique reads measured in a sample of DNA. Then the raw count score c_s for each sequence s where $1 \leq s \leq n$ can be converted into a "relative enrichment score" r_s using the following formula:

$$r_s = \frac{c_s}{\sum_{i=1}^{n} c_i}$$

Thus, each unique sequence s measured after a round of selection is associated with a score $0 \leq r_s \leq 1$, where r_s represented the frequency of that sequence appearing in the sample. Ideally, under the assumption that there are no experimentally induced sampling biases, this sample accurately reflects the distribution of sequences remaining in the pool after each round of selection. Thus, r_s represents how often HFNAP sequence s successfully bound to the target protein relative to the other sequences in the library.

An HFNAP sequence s that has a naturally higher binding affinity for the target should have a higher r_s than a sequence with a weaker binding affinity for the target [30]. However, due to noise and early-round-dropout (as discussed in 1.4), this relationship does not always hold. Additionally, just because a sequence has a high relative enrichment score does not necessarily mean it is a strong binder – the other sequences tested could simply be extremely poor binders for the target, thus artificially inflating these relativistic measurements. On the other hand, the presence of many strong binders will make it appear as if none of the sequences are strong binders, for they will each deflate the relative enrichment score of one another. The same issues would hold true had we simply used the raw read counts, since once we take our random sample of HFNAPs then the size of that sample will, by definition, be fixed. Thus, there will always be an issue of crowding out regardless of whether we look at relative or absolute scores. However, because we can't control in advance what the precise size of our sample will be, we need to convert to relativistic scores to control for round-to-round variation in this measurement.

At the end of each round, we are therefore left with a dataset containing tens of thousands of unique 45-mer sequences, as well as the relative frequencies r_s at which they were observed in the sample of DNA taken after that round.

The increasing stringency of each round, combined with measurement noise and random chance, means that the list of unique sequences that appear in each round's set of relative enrichment scores can be highly variable [44]. This complicates analysis, for sequences can disappear and then re-appear at later time points.

For example, a sequence s that appears with a relative frequency of 0.0001 in Round 1 may not appear at all in Round 2, then re-appear in Round 3 with a score of 0.0004, then disappear from Rounds 4-7 before re-appearing in Round 8 with a score of 0.0001. These fluctuations tend to be due to noise, but they could also reflect a sequence that simply was not captured in several sequencing rounds despite binding to daunomycin. Given its low starting frequency, this sequence could also have been an HFNAP which would have bound well to daunomycin but never had the sufficient concentration to do so. Generally, however, sequences which bind well will see their relative enrichment scores increase round-to-round, or at the very least stay constant while other stronger-binding HFNAPs increase their share of the overall pool of strong binders [44].

In total, 174,086 distinct HFNAP sequences were identified in the SELEX dataset analyzed by this thesis.

2.2.5 Analysis Metrics

In order to condense the information contained in a given aptamer sequence's round-by-round enrichment scores into a single summary statistic of that sequence's "ground truth" fitness, several approaches have been proposed in the literature.

The simplest method for summarizing the time series data associated with each aptamer sequence is to simply take the enrichment level achieved in the last round of SELEX and use that as the "true" fitness score for each sequence [48]. This approach makes immediate intuitive sense, as the performance of an aptamer in the last round depends on its performance in all previous rounds. Thus, the final enrichment of an aptamer in its last round of SELEX should implicitly contain the information revealed from its performance in previous rounds.

Using this metric, the sequence s with the highest raw read count c_s (or, equivalently, relative enrichment score r_s) in the last round of SELEX would thus be considered the best binding aptamer contained in the library that was screened.

Even ignoring the issues of measurement noise, false positives, and false negatives, however, this conclusion will only hold true if every aptamer sequence starts with the same initial number of fragments and receives the same level of amplification during PCR [30]. Otherwise, this fitness metric will be biased towards sequences that happen to appear at a higher frequency in the initial aptamer library, or which get preferentially amplified by PCR completely independent of their binding affinities for the target [49]. A more robust fitness metric would effectively control for the random and uneven starting concentrations of each sequence before each round.

An alternative metric that accomplishes this is the "enrichment fold change." This involves measuring the fold change of a sequence's relative frequency between each pair of rounds, then applying some function to reduce these fold change values into a single score [35].

For example, one can take the maximum of calculated fold changes to derive a "greatest enrichment fold change," or take the fold change between the relative enrichment scores of the earliest few rounds, or between non-consecutive rounds. Enrichment fold change has been largely accepted in the literature as a standard method for converting multiple rounds of enrichment scores into a single "ground truth" fitness metric, and represents the "most often used" metric in aptamer analysis pipelines due to its ease of calculation, straightforward interpretation, and successful track record in practice [63]. It has been experimentally demonstrated that aptamers identified using the method of greatest enrichment fold change outperform aptamers chosen simply based on their abundance in the last round of SELEX [30]. And, in general, measuring enrichment in terms of relative fold change has experimentally been shown to outperform absolute- or prevalence-based metrics [49]. Thus, there is an empirical justification for preferring relative fold change over the aforementioned metric of final round read counts when screening aptamers based on SELEX data.

Alternative metrics that have been proposed in the literature, but not been as widely adopted, include the following: clustering aptamers after each round by the edit distance between sequences to allow for the averaging of binding performance over similar aptamers [29] and observing how quickly sequences enrich at the earliest – rather than latest – rounds of SELEX to help counter-act the biases induced by PCR after every round [30].

Due to its wider acceptance in the literature, however, this paper used relative fold change as the fitness label for each sequence. This value was calculated by taking the multiplicative change in each HFNAP's relative frequency between the start and end of the higher stringency SELEX rounds (rounds 4 and 8, respectively), for these rounds demonstrated a more robust ability to filter for high-binding sequences than earlier rounds.

Once an aptamer has been selected based on its performance in SELEX, it must then be run through an additional assay to more rigorously assess its true binding affinity for the target [30]. Experiments that can provide a more accurate measurement of an aptamer's true binding abilities include flow cytometry, electrophoretic mobility shift, filter-binding, microscale thermophoresis (MST), and surface plasmon resonance (SPR) [30]. In particular, this paper's results were experimentally verified using MST.

This fold change metric can be used to treat aptamer binding as a weak supervised learning problem. Though the true label, i.e. "ground truth" fitness, for each aptamer sequence is unknown, the greatest enrichment fold change output by SELEX can serve as a noisy proxy for the true binding affinity of each aptamer. Though it is true that this metric lacks a rigorous theoretical underpinning, in light of its accepted usage in the literature as well as its fairly robust predictive performance in practice, it seemed justifiable to use as a training label given this existing domain knowledge [30]. The problem of predicting aptamer binding performance based on SELEX data can be effectively framed as a supervised learning problem despite lacking the true labels that a model would try to predict.

An alternative approach would be to generate fitness scores from scratch by applying a more complex function to the data that is output by SELEX. Fold change ignores a significant amount of information reflected in the overall up-and-down performance of an aptamer across rounds of SELEX. Thus, more fully taking into account the time series nature of the data, the increased selection stringency of each round, and the increased concentration of better-binding aptamers after each selection would likely enable a more accurate assessment of an aptamer's true binding fitness. This thesis outlines a theoretical approach to constructing such a model, and though not utilized for the daunomycin dataset at the core of this paper, presents a principled approach and implementation for use in future selections.

3
Computational Methods

Deep learning has already revolutionized a number of computational fields. Its application to the field of aptamer discovery, however, has remained relatively unexplored.

This chapter establishes a novel framework, methodology, and implementation strategy for utilizing deep learning techniques to screen aptamers. The methods detailed in this chapter can uniquely capture the complex relationships among an aptamer's sequence, sidechain, and binding performance, thereby allowing for the *in silico* generation of entirely novel, high-binding aptamer sequences.

First, this chapter will discuss how the task of aptamer screening was framed as a weak supervised learning problem. Second, the specific model – a conditional variational autoencoder (CVAE) – chosen to address this task will be detailed. Finally, the implementation of the model and approach used to train, test, and validate it will be described.

3.1 Weak Supervised Learning

3.1.1 Enrichment Fold Change

Unlike other areas of computational chemistry and genomics, the field of aptamer discovery lacks a unified framework for understanding how such methods should even be applied [26, 91].

In terms of applying machine learning techniques to the field of aptamer discovery, one of the main challenges is deciding what the true “label” assigned to each aptamer sequence should be.

Several conventional analytic methods have been utilized in the literature for converting SELEX relative enrichment scores into “ground truth” fitnesses, as described in 2.2.5. One of the more widely utilized of these methods, enrichment fold change, was selected as a way of labeling the SELEX data on which the CVAE was trained. Additionally, the performance of the aptamer sequence over the final five rounds of ”high stringency” selections (rounds 4-8) were included as inputs to the model.

There were three primary reasons for this decision. (1) Fold change is computationally simple to calculate and a relatively intuitive metric to understand. (2) The metric has been demonstrated in practice to achieve a relatively solid ability at discriminating strong- from weak-binding aptamers, and thus it seemed counter-productive to ignore this relevant domain knowledge [30]. (3) The fact that this metric is already utilized in the literature helps make the results of the CVAE more directly comparable to the results of the rest of the field and that of other *in silico* methods, as opposed to creating an entirely novel metric for evaluating the model. [49, 63]

The enrichment fold change of a sequence is calculated as follows. Let the integer s represent an aptamer sequence. Let c_{sk} represent the raw count of the number of reads of sequence s sampled from the eluted flow-through after the kth round of SELEX. Assume that there are K total rounds of SELEX performed. Then let n_k represent the total number of unique sequences measured after round

k, where $1 \leq k \leq K$, and label each sequence in round k with a unique integer i such that $1 \leq i \leq n_k$. Let us also assume that aptamer s appears in every round (so that the following calculations are not all simply equal to 0), and thus $1 \leq s \leq n_k.\forall k$.

Then the relative frequency score r_{sk} of sequence s in round k is given by:

$$r_{sk} = \frac{c_{sk}}{\sum_{i=1}^{n_k} c_{ik}}$$

Define the fold change $f_{s,j,k}$ of sequence s between rounds j and k where $j < k$ to be as follows:

$$f_{s,j,k} = \frac{r_{sk}}{r_{sj}}$$

Note that $f_{s,j,k}$ is only defined for $1 \leq j < K$.

A variety of fold-change-based metrics can now be calculated. For example, the greatest enrichment fold change g_s for an HFNAP s could be calculated as follows:

$$g_s = \max_{1 \leq j < K} \{f_{s,j,j+1}\}$$

For the purposes of training and testing the CVAE model used in this thesis, the fold change between the relative enrichment level of a sequence between rounds 4 and 8 was utilized (since these were the starting and ending rounds of the higher stringency portion of the selection), then scaled and normalized to exist on a continuum between 0 and 1. Thus, the following formula was used to calculate the "ground truth" fitness t_s for each sequence s:

$$t_s = \frac{f_{s,4,8}}{\max_{1 \leq p \leq n_8} \{f_{p,4,8}\}}$$

As a slight caveat, the top 650 enriched sequences observed in round 8 had their enrichment scores equalized before calculating $f_{s,4,8}$. This was done in order to avoid having several extremely highly enriched sequences from skewing the values assigned to the rest of the sequences in the pool (hence the slightly higher concentration of fitnesses around 1).

The overall distribution of fitness scores calculated through this process is shown in Figure 3.1.1. Overall, roughly 2,000 sequences had fitness scores > 0.25, reflecting the well-known fact that the vast majority of aptamer sequences will be poor binders for a given target [4].

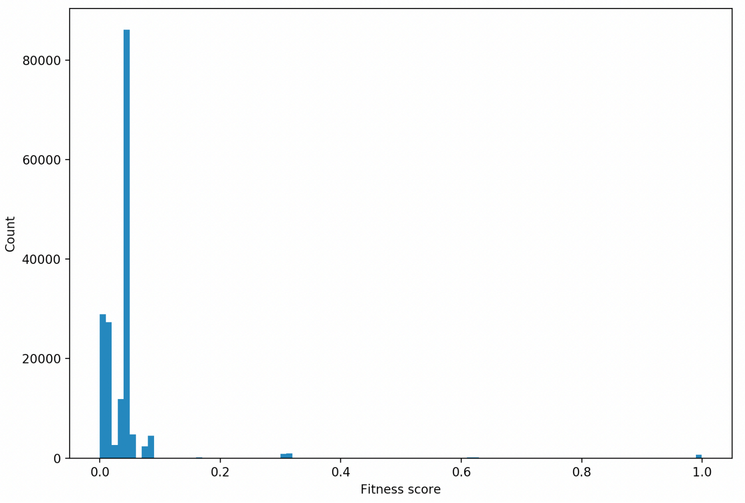

Figure 3.1.1: This graph shows the distribution of the fitness scores assigned to the HFNAPs in the SELEX dataset. These fitness scores were determined by calculating the fold change between the relative enrichment level of each sequence between rounds 4 and 8, then normalized to exist on a continuum between 0 and 1. The very top sequences had their fitness scores equalized to avoid having several extremely highly enriched sequences from skewing the values assigned to the rest of the sequences in the pool (hence the slightly higher concentration of fitnesses around 1).

3.1.2 Fitness Model

Compressing 8 rounds of SELEX data into a single metric based on only 2 rounds of activity, however, likely discards a significant amount of valuable information. Ignoring the overall pattern of a sequence's enrichment and de-enrichment may make it harder to successfully make predictions.

This section describes the development of a more general theoretical model to estimate the "ground truth" fitness scores of aptamers, and hopefully allow for more accurate fitness estimates in the future than the aforementioned fold change approach.

Assume once again that there are K total rounds of SELEX. Let S_k represent the set of unique aptamer sequences measured after round k, where $1 \leq k \leq K$.

Given an aptamer sequence s, let f_s represent the true binding fitness of that sequence. We will define f_s such that $0 \leq f_s \leq 1$, where 0 represents the weakest possible binding activity and 1 the strongest.

Thus, the overarching goal of SELEX is to determine f_s for all sequences s, so that we can then rank the aptamers by their binding affinities for the target and select the best ones.

Unfortunately, we are unable to observe f_s directly. Instead, we observe a relative frequency r_{sk} for each sequence, which is defined as the proportion of reads in the sample taken after round k that have sequence s. Thus,

$$\sum_{s \in S_k} r_{sk} = 1$$

This means that every sequence's r_{sk} is dependent on the r_{tk} of every other sequence $t \in S_k \setminus s$ in round k. This value is also dependent on the stringency of the selection for that round (e.g. environmental variables like temperature, elution conditions, etc.). Thus, we can express r_{sk} in terms of the following function R:

$$r_{sk} = R(f_s, E, f_{t \neq s}, b_s, r_{s,k-1})$$

where E represents environmental variables (e.g. temperature) that are identical for each sequence, $f_{t \neq s}$ represents the true fitness of every other sequence in round k that are competing with sequence s to bind to the target, $r_{s,k-1}$ represents the relative abundance of sequence s after the previous round's selection, and b_s represents the translation/amplification biases that may exist for that specific sequence.

This last term, b_s, will be ignored momentarily while modeling R in order to simplify calculations, although this value can be separately controlled by first pre-processing SELEX data based on the performance of unfunctionalized libraries or the PCR biases measured for each sequence through additional tests.

The actual mechanistic process through which $f_{t \neq s}$ will impact r_{sk} is through these other sequences $t \in S_k \setminus s$ competing with sequence s to bind to the finite amount of target binding locations. Here, another simplifying assumption will be made to make this model more computationally tractable – the impact of $f_{t \neq s}$ on r_{sk} will be the same for all s. In terms of the actual chemistry, this assumption can be justified by the fact that the overall level of aptamer v. aptamer competition exhibited in a pool of hundreds of thousands of distinct sequences should be relatively unchanged, even if a single one of those sequences is removed from the pool. Thus, since $f_{t \neq s}$ essentially represents the overall level of competition during a given round, and we are making the simplifying assumption that $f_{t \neq a} = f_{t \neq b}. \forall a, b \in S_k$, then we can roll this set of variables into an all-encompassing competition variable C that is identical for all sequences.

Conveniently, the impact of C on r_{sk} actually trends in the same direction as the environmental factors represented by E – later rounds of SELEX will have a higher concentration of stronger sequences and thus a higher level of competition, for the experimental design of SELEX requires the removal of non-binding sequences after each round. Thus, later rounds of SELEX will have higher overall selection stringencies, due to both the harshening of environmental variables like temperature and target concentration as well as the increased average competitiveness of the aptamers remaining in the pool [78].

We can now rewrite r_{sk} in terms of a new function R' based on E, C, and $r_{s,k-1}$:

$$r_{sk} = R'(f_s, E, C, r_i')$$

How, then, can we convert raw fitness scores f_s into relative scores r_{sk}?

A commonly used mathematical procedure for normalizing a set of real numbers into a probability distribution is the softmax function $\sigma : \mathbb{R}^m \rightarrow \mathbb{R}^m$ [28]. Given a vector $\mathbf{x} \in \mathbb{R}^m$ as input, the softmax function is defined as follows, where i represents the ith element of the corresponding vector:

$$\sigma(\mathbf{x})_i = \frac{e^{x_i}}{\sum_{j=1}^{m} e^{x_j}}$$

The base of the exponent (e in the above equation) can be changed to any positive real number in order to generate a probability distribution of higher/lower concentration. Choosing a higher valued base will yield a probability distribution that is more concentrated around the largest valued elements of the input $\mathbf{x}$ [28]. Choosing a smaller base, on the other hand, will more evenly spread out the resulting distribution generated by the softmax [28].

Increasing the selection stringency of a SELEX round will result in fewer sequences binding to the target, as only the strongest binders will be capable of successfully overcoming both the harsher environmental conditions and increased competition from being placed in a more competitive pool [78]. These selection pressures result in fewer unique sequences remaining after non-binding sequences have been removed post-selection. Thus, the overall distribution of the r_{sk}'s observed in round k should be more highly concentrated in higher stringency rounds, i.e. as k approaches K.

Returning to the original equation $r_{sk} = R'(f_s, E, C, r_{s,k-1})$, we can thus model the environmental stringency E, competition variable C, and relationship between "true" fitness f_s and relative fitness r_{sk} as a softmax function where the exponential base increases with each successive round. This allows for the generation of distributions with different concentrations to reflect the differences in selection stringencies among rounds. This yields the following model, where d_k represents the difficulty of binding, aka the stringency, of round k:

$$r_{sk} = r_{s,k-1} \frac{d_k^{f_s}}{\sum_{p \in S_k} d_k^{f_p}}$$

Note that the true fitness f_s of an aptamer does not depend on k, for an aptamer's true binding affinity for a target remains constant even as its relative enrichment scores vary round-by-round based on environmental and competitive dynamics.

By converting raw fitness scores into relative enrichment scores while taking into account the true fitnesses of other aptamers in the round through the summation in the denominator, this equation for r_{sk} allows us to fit a model to all rounds of SELEX data and utilize the information contained in each round to estimate what the true ground truth fitness f_s for each sequence should be. By fitting a d_k to the empirical concentration of the distribution of frequency scores generated by each round of SELEX, this model can be tailored to the specific experimental conditions exhibited during a given run of SELEX.

An implementation of this model was written in Python 3.6 using Pytorch v1.2.0. The program determines the ground truth fitness values f_s by fitting the aforementioned model for r_{sk} on all rounds of the input SELEX data through stochastic gradient descent. The loss function utilized is the KL divergence between the predicted distribution of relative frequencies based on the fitted f_i generated by the model and the actual distribution of relative frequencies observed in the SELEX dataset. See Appendix A for the Github repository where this model is stored.

3.2 Predictive Models

As previously detailed in 1.6, considerable effort has already been made towards utilizing computational approaches to more effectively screen aptamer sequences. Based on a comprehensive literature review, however, there does not currently appear to be any existing work on applying deep learning techniques to this field. Thus, this thesis uniquely applies deep learning to the challenge of generating novel high-binding aptamer sequences based on a single run of SELEX data.

3.2.1 Variational Autoencoder (VAE)

Inspired by the success that Sinai et al. had in utilizing a variational autoencoder for the task of predicting how changes to a protein's sequence impacted its function, a simple VAE was chosen as the first model to consider for this paper [74].

A VAE is an unsupervised generative model that seeks to model a joint distribution of the observed data and the unobserved "latent" variables which determine that data [45, 76]. By learning to model the actual landscape of aptamer sequences through these latent variables, instead of merely how to classify a given sequence as "strong" or "weak," this generative model will readily allow for the generation of novel aptamer sequences through sampling of the distribution learned by the model.

Before considering any fitness measurements, a simple VAE was thus constructed to take as input an aptamer sequence, compress it into a smaller set of latent variables, and reconstruct the original sequence of A/T/G/C's. This was done in order to first assess whether a VAE-type model would be able to identify and effectively capture the types of motifs present in aptamer sequences, and also to determine the size of the latent space needed to capture such motifs. Fitness scores were later factored in by extending this VAE to a conditional variational autoencoder (CVAE), as will be explained in the next section.

At a high level, an autoencoder is comprised of two segments – an "encoding" network that compresses its input down into some lower-dimensional representation of that input (the "latent" encoding), followed by a "decoding" network that attempts to reconstruct the original input given this compressed encoding [79].

This contraction of information into some fixed-length latent vector forces the autoencoder to extract as much information as densely as possible from its input

[79]. Figure 3.2.1 shows the high-level architecture of an autoencoder.

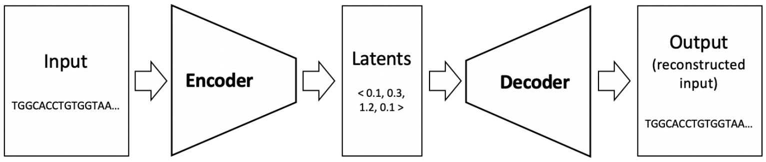

Figure 3.2.1: The basic structure of an autoencoder.

VAEs extend this traditional autoencoder architecture by forcing the learned encoding to represent a continuous distribution that can be sampled from [45]. This allows the VAE to generate novel outputs by sampling from its learned encoding distributions, and is thus better suited for generative modeling. By choosing a generative rather than discriminative model, the goal was to provide the ability for the end user to generate entirely original HFNAP sequences optimized for binding that had not been present in the VAE's training set.

Let $p(x)$ represent the distribution of aptamer sequences. Then let $\mathbf{x} \sim p(x)$ be the data that was actually observed during the HFNAP SELEX run for daunomycin.

Let $p(z)$ be the distribution for the latent variables, and $\mathbf{z}$ be the latent variables associated with the observed data $\mathbf{x}$. Then for each observation x_i, we assume the following generative process:

$$z_i \sim p(z)$$
$$x_i \sim p(x|z_i)$$

Thus, we sample each aptamer sequence from the latent variables that succinctly describe it. The goal of the encoder is to determine $p(z|x)$ – that is, the latent variables that best describe the observed data input to the VAE.

In the context of HFNAP reconstruction, this means determining what latent variable z_i is sufficient to describe the complete 45-mer sequence x_i input to the model. If z_i is k-dimensional, then the first dimension of z_i may represent a common sequence motif (e.g. "AAAAAA" repeated three times), the second dimension may be another sequence motif, the third dimension may contain information on the folding structure of the sequence inferred from the presence/absence of certain sequential patterns, etc., for all k dimensions. The guiding hope of this type of model is that these learned latent distributions will capture valuable, higher-order relationships within each sequence that could not be modeled through other less expressive models [65, 74].

In order to calculate $p(z|x)$, we can use Bayes Theorem:

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

In order to calculate the denominator $p(x)$, we can marginalize over the latent variables z:

$$p(x) = \int p(x|z)p(z)dz$$

Unfortunately, however, this integral is intractable, as it requires marginalizing over all possible values for the latent variable z, which takes exponential time. Thus, instead of calculating $p(z|x)$, we will instead use in its place an approximate distribution $q(z)$ that is tractable to compute. (Note: In the following pages, the symbol $\approx$ will be used to refer to an approximation of one distribution with another).

In terms of the "encoding" and "decoding" aspects of a VAE, this new distribution $q(z) \approx p(z|x)$ represents the encoding segment (as we are mapping an input x to its compressed latent representation z) while $p(x|z)$ represents the decoding segment (as we are mapping a latent representation z back to its original corresponding input x). This is depicted in Figure 3.2.2.

The goal now is to ensure that $q(z)$, our approximation for the posterior distribution of the latents z given the observed data x, is as close as possible to $p(z|x)$. But what does "close" mean in the context of probability distributions?

The Kullback-Leibler (KL) divergence D_{KL} measures the similarity between two different probability distributions [47]. Let A and B be probability density

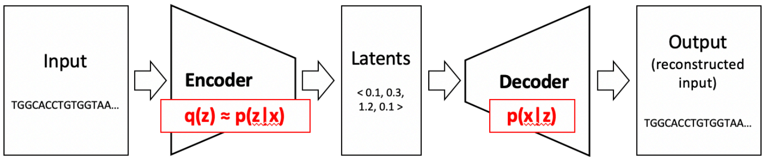

Figure 3.2.2: The encoder approximates $p(z|x)$ by modeling a distribution $q(z)$ that can be parametrized by a neural network.

functions. Then the formula for the KL divergence between the two distributions, $D_{KL}(A||B)$, is provided below [47]:

$$D_{KL}(A||B) = \mathbb{E}_{x \sim A}[\log(\frac{A(x)}{B(x)})] = \int \log(\frac{A(x)}{B(x)})A(x)dx$$

Intuitively, $D_{KL}(A||B)$ measures how much information is lost when using the distribution A to approximate the distribution B. (Note, however, that this metric is not symmetric, i.e. $D_{KL}(A||B) \neq D_{KL}(B||A)$). We thus want to minimize this value in order to ensure that our approximate distribution $q(z)$ does as well as possible at approximating the desired distribution $p(z|x)$.

The KL divergence is minimized when A and B are the same probability distribution, i.e. when $A(x) = B(x).\forall x$ then $D_{KL}(A||B) = 0$. That's because if $A = B$, then it must be true that $\log(\frac{A(x)}{B(x)}) = \log(1) = 0.\forall x$, and thus $\int \log(\frac{A(x)}{B(x)})A(x)dx = 0$. In terms of our VAE model, this means that ideally we would have $q(z)$ being an identical function to $p(z|x)$ and thus $D_{KL}(q||p) = 0$.

Additionally, the KL divergence is always non-negative, i.e. $D_{KL}(A||B) \geq 0$ for any A and B. We can prove this using Jensen's inequality, which states that the following must hold when f is a convex function [40]:

$$f(E[X]) \leq E[f(X)]$$

Since $\log(x)$ is a concave function, $-\log(x)$ must be convex. Thus, we have that:

$$\begin{aligned}
D_{KL}(A||B) &= \mathbb{E}_{x\sim A}[\log(\frac{A(x)}{B(x)})] \\
&= -\mathbb{E}_{x\sim A}[-\log(\frac{A(x)}{B(x)})] \\
&= \mathbb{E}_{x\sim A}[-\log(\frac{B(x)}{A(x)})] \\
&\geq -\log[\mathbb{E}_{x\sim A}[\frac{B(x)}{A(x)}]] \\
&= -\log[\int \frac{B(x)}{A(x)}A(x)dx] \\
&= -\log[\int B(x)dx] \\
&= -\log(1) \\
&= 0
\end{aligned}$$

Where we have that $\int B(x)dx = 1$ by the axioms of probability since $B(x)$ is a probability density function.

Thus, we have that the KL divergence of two probability distributions must be non-negative. This fact will provide some insight into the next several calculations as we move back to our original problem – finding a $q(z)$ which best approximates $p(z|x)$.

To approach this problem in a principled manner, let us return to the initial problem of modeling our observed data x. We want to maximize the likelihood of this observed data, which by definition is $p(x)$. Maximizing $p(x)$ is equivalent to maximizing the log-likelihood $\log(p(x))$ since log is monotonically increasing. Thus, we should aim to maximize:

$$\log(p(x)) = \log[\int p(x|z)p(z)dz]$$

Multiplying by $\frac{q(z)}{q(z)} = 1$ yields:

$$\begin{aligned}\log(p(x)) &= \log[\int p(x|z)p(z)dz] \\ &= \log[\int p(x|z)p(z)\frac{q(z)}{q(z)}dz] \\ &= \log[\mathbb{E}_{z\sim q(z)}[\frac{p(x|z)p(z)}{q(z)}]]\end{aligned}$$

By applying Jensen's inequality again, we thus have that:

$$\begin{aligned}\log(p(x)) &\geq \mathbb{E}_{z\sim q(z)}[\log(\frac{p(x|z)p(z)}{q(z)})] \\ &= \mathbb{E}_{z\sim q(z)}[\log(p(x|z)p(z)) - \log(q(z))] \\ &= \mathbb{E}_{z\sim q(z)}[\log(p(x|z)) + \log(\frac{p(z)}{q(z)})] \\ &= \mathbb{E}_{z\sim q(z)}[\log(p(x|z))] + \mathbb{E}_{z\sim q(z)}[\log(\frac{p(z)}{q(z)})] \\ &= \mathbb{E}_{z\sim q(z)}[\log(p(x|z))] + \int \log(\frac{p(z)}{q(z)})q(z)dz \\ &= \mathbb{E}_{z\sim q(z)}[\log(p(x|z))] - \int \log(\frac{q(z)}{p(z)})q(z)dz\end{aligned}$$

Note that by the definition of KL divergence, we can replace the last expression with $D_{KL}(q(z)||p(z))$. Thus, we have that:

$$\log(p(x)) \geq \mathbb{E}_{z\sim q(z)}[\log(p(x|z))] - D_{KL}(q(z)||p(z))$$

The term on the right-hand side of the above inequality is known as the Evidence Lower Bound, or the "ELBO," since it provides a lower bound on the log probability of the observed data.

$$ELBO = \mathbb{E}_{z\sim q(z)}[\log(p(x|z))] - D_{KL}(q(z)||p(z)) \leq \log(p(x))$$

This final inequality gives us a means of converting our original inference problem of determining $p(z|x)$ into an optimization problem. The $\mathbb{E}_{z\sim q(z)}[\log(p(x|z))]$ term is known as the "reconstruction" term, as the value for this expression comes from our attempt to reconstruct the observed data x from our latents z [53]. The $D_{KL}(q(z)||p(z))$ term is known as a "regularization" term since it penalizes estimates for $q(z) \approx p(z|x)$ that stray from whatever prior $p(z)$ we've assigned to the family of distributions that could represent $q(z)$ [53].

Thus, by **maximizing the ELBO through the maximization of the reconstruction term and minimization of the regularization term, we can maximize the likelihood of the observed data** x.

Another way to interpret the above results is as follows: Revisiting the previous note about the KL divergence being non-negative, one can also establish the following relationship between the marginal data likelihood and the *ELBO* [45].

$$\log(p(x)) = ELBO + D_{KL}(q(z)||p(z|x))$$

Since $\log(p(x))$ is fixed as a function of q, and the KL divergence is always non-negative, this means that we can also interpret **maximizing the ELBO** as **minimizing the KL divergence between** $q(z)$ **and** $p(z|x)$ (where $p(z|x)$ was the distribution we desired to approximate with $q(z)$). Since a smaller KL divergence means that two probability distributions are more similar, the maximization of the ELBO achieves precisely our original goal when minimizing the divergence between $q(z)$ and $p(z|x)$.

Having transformed our inference problem into one of optimization, we can now train our model by maximizing the ELBO which, as detailed previously, can be calculated by breaking it down into its two separate components (the reconstruction and regularization terms).

Returning to the original motivating biological problem, the goal of the VAE was to learn how to compress and reconstruct aptamer sequences (i.e. 45-mers) that were contained in the daunomycin SELEX dataset. Because this set-up simply treats each aptamer as a string of text, the VAE was expected to perform fairly well since VAEs have already been successfully demonstrated on a variety of text-based reconstruction tasks [51, 71, 83, 85].

For this thesis, the VAE model was implemented in Python 3.6 using Pytorch v1.2.0, an open-source machine learning library originally developed by Facebook AI Research [62]. A latent space Z comprised of 15 distinct Gaussian distributions was chosen for the VAE. Each Gaussian would need to be parametrized by a mean and variance, and thus there were 15 total pairs of latent variables of the form (μ_i, σ_i) needing to be fit, where $z_i \in Z$ and $z_i|x \sim N(\mu_i, \sigma_i)$ for $1 \leq i \leq 15$, where x is the training data. A prior of the standard normal was assumed, and thus $z_i \sim N(0, 1) \forall i$.

Let $q(z)$ again represent the approximation of the posterior $p(z|x)$. Then the regularization term from the ELBO optimization equation that was previously derived could be calculated as follows under these modeling assumptions [45]:

$$D_{KL}(q(z)||p(z)) = -\frac{1}{2} \sum_{i=1}^{15} 1 + \log(\sigma_i^2) - \mu_i^2 - \sigma_i^2$$

As input, the VAE took a 300x1 vector of binary values. This vector was comprised of (1) an 180x1 one-hot vector encoding the 45 nucleotides of the aptamer, and (2) a 120x1 vector encoding which of the 8 possible sidechains was located at each of the 15 trimer locations of the sequence. Though this latter information was technically redundant, it was provided to the VAE to hopefully make identifying and modeling the secondary structural information contained in a sequence more obvious.

This input was then fed through the encoder of the VAE in order to model $q(z)$. The encoder was comprised of 300 fully connected nodes with $tanh()$ activations followed by another fully connected layer of 60 nodes with $tanh()$ activations. The network then branched off into two separate sets of 15 nodes (each fully connected to the previous layer of 60 nodes) in which each set of 15 nodes generated one set of the desired (μ_i, σ_i) parameters for the latent space. Thus, the final layer was comprised of 30 total nodes which collectively provided a compressed probabilistic representation of the input.

These latent encodings were then fed through the decoding portion of the

VAE. The decoder aims to model the distribution $p(x|z)$, i.e. the reconstruction of the original 45-mer sequence x based on the 15 Gaussians z that were learned as the latent representation for that sequence. During training, a random sample was taken from each of these Gaussians to yield a 15x1 vector of real numbers as the latent representation for that sequence. During inference, however, the mean of each distribution (μ_i) was simply spit out by the VAE to represent our most likely estimate for the input [45].

This 15x1 vector was fed through a fully connected layer of 480 nodes with $tanh()$ activations. The output of this layer was then split into 15 different non-overlapping vectors of size 32x1, and each of these vectors was fed through a separate set of the following sequence of layers: a fully connected layer of 32 nodes with $tanh()$ activations, another fully connected layer of 32 nodes with $tanh()$ activations, and another fully connected layer of 32 nodes. The final outputs from each of these 15 separate networks were then softmaxed. This final 32x1 vector represented the probability that each of the 32 possible trimer-sidechain pairings was located at a specific location in the aptamer sequence. Recall that each trimer is associated with a unique sidechain, and that there were 32 possible trimers of the form "YNN" (since, by design, they must start with a C or T). Thus, specifying the trimer at each location in an HFNAP was sufficient information to fully reconstruct that sequence. Since each HFNAP has 45 nucleotides, there are thus 15 total trimers per aptamer. Thus, the final output of the VAE was a 32x15 vector. In order to convert this back into a sequence and enable a direct comparison with the original input, an argmax was taken over each component 32x1 vector to identify the most likely trimer associated with each sequence position. These numbers could then be mapped back to the actual string of nucleotides that they encoded and directly compared to the originally input HFNAP.

A more refined version of the previous VAE diagrams which demonstrates the full architecture of this model is shown in Figure 3.2.3.

The loss function for the model was based on maximizing the ELBO, as

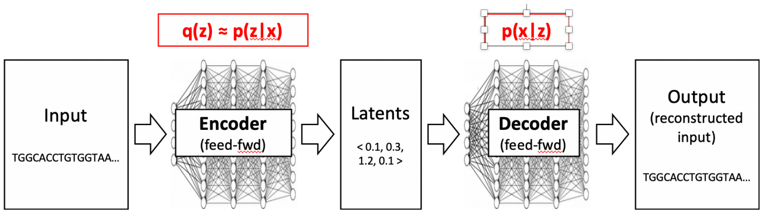

Figure 3.2.3: A feed-forward neural network is used in both the encoding and decoding segments of the VAE to approximate their associated distributions.

previously derived. This entailed calculating the cross entropy of the reconstructed sequence, based on the 32x15 output of the model, and adding it to the KL divergence between the learned latents distributions and a standard normal distribution.

3.2.2 Conditional Variational Auto Encoder (CVAE)

After it was successfully demonstrated that a VAE could compress and reconstruct aptamer sequences, the next goal was to associate each sequence with a fitness value.

A VAE, however, will not easily permit this, as there is no way to control what type of abstraction will be learned by each of the latent variables. Given the set of 15 Gaussians that comprised the latent space of the previously trained VAE, how would one know in advance which values should be sampled from each of these Gaussians in order to get the VAE to output a high-binding aptamer, without simply iterating over the entire latent space (which would be computationally infeasible)? While it is easy to ask a VAE to "generate any arbitrary aptamer sequence," it is essentially impossible to ask it to "generate an aptamer sequence with *x* binding affinity" [75].

The solution is a conditional variational autoencoder (CVAE), which will allow us to specify a "condition" associated with each input sequence – i.e. binding fitness [75]. We will then be able to sample from the latent space learned by this model to generate outputs **conditional on the output having the**

desired binding fitness. The CVAE is essentially a semi-supervised modification to the VAE that still retains many of its desirable unsupervised generative properties [75].

More concretely, a "condition" $c \in [0, 1]$ would be associated with each HFNAP sequence. This would represent its "ground truth" binding fitness. When an HFNAP sequence is input to the encoder of the CVAE, we also append to it the associated condition c. Once this sequence is compressed into a vector of latent variables z, we then append the condition c to z before feeding the combined vector through the decoder of the model. Thus, the decoder trains on both a set of latents z and the desired condition c. This key step of feeding the condition c back into the decoder will allow us to later specify the binding fitness of the novel aptamers that we wish to generate from our model when we sample from the learned latent space, and thus identify only high-binding sequences.

The previously calculated formula for the ELBO can be re-derived by simply conditioning all of the distributions on the condition c. By repeating the simplification of $\log(p(x|c))$ using Jensen's inequality, we are left with the following slightly modified version of our VAE's original ELBO [75]:

$$\log(p(x|c)) \geq \mathbb{E}_{z\sim q(z|c)}[\log(p(x|z, c))] - D_{KL}(q(z|c)||p(z|c))$$

As the equation above shows, we will once again be optimizing the traditional reconstruction and regularization terms of the ELBO – this time, however, everything is conditional on the class c specified for each input.

Once this model is trained, a user could then specify a desired fitness (e.g. 1 to get the best possible binder), feed that into the trained decoder along with some sample taken from the latent space, and then get as output an aptamer sequence predicted to have that fitness.

Like the VAE, the CVAE model was implemented in Python using Pytorch. Given the success of the VAE's architecture at capturing sequence motifs, a latent space Z comprised of 15 distinct Gaussians z_i was again chosen for this model, each parametrized such that $z_i|x, c \sim N(\mu_i, \sigma_i)$ for $1 \leq i \leq 15$, where x is the training data and c is the condition. In this case, the condition was the "ground truth" fitness value of each input sequence, as measured on a scale of $c \in [0, 1]$ following the calculations described in 3.1.1. A prior of the standard normal for $z_i \sim N(0, 1)$ was again assumed, allowing us to re-use the aforementioned results of Kingma et al. in calculating the KL divergence between $q(z)$ and $p(z)$ [45].

As input, the VAE took a 306x1 binary vector. The vector was comprised of (1) a 180x1 one-hot vector encoding the 45 nucleotides of the aptamer, (2) a 120x1 vector encoding which of the 8 possible sidechains was located at each of the 15 trimer locations of the sequence, (3) a 5x1 vector containing the fold change enrichment of the HFNAP over the final rounds of SELEX, and (4) a 1x1 vector containing the "ground truth" fitness score estimated for the sequence based on its SELEX performance calculated per the equation of 3.1.1. This final value, the "ground truth" fitness score, represented the condition c for the model. Though this fitness score would ideally contain all of the information revealed by the 5x1 vector of in-round enrichment fold changes, because the fitness score c is inherently unstable and fairly difficult to assess the accuracy of (as detailed previously), it was decided to include some raw data from the high stringency rounds of the SELEX run itself in order to provide the model with a more complete picture of how each HFNAP performed.

This 306x1 vector was then fed through the encoder of the CVAE in order to model $q(z|c)$. Unlike the VAE, however, the encoder for the CVAE was chosen to be a recurrent neural network (RNN) instead of a feed-forward neural network. This change was made after early testing showed that the initial feed-forward encoder network lacked the ability to effectively capture both the HFNAP sequence and its associated fitness score.

Traditional RNNs are able to model recurrent data like sequences better than other neural network architectures [21]. By maintaining an internal representation of past characters in an input sequence, RNNs can better detect how one part of a sequential input impacts future elements of that sequence [73].

These benefits, however, tend to be fairly short lived, as traditional RNNs tend to forget information that was observed more than 10 time steps ago [25]. This is problematic for HFNAP modeling, as HFNAPs can fold in extremely complex shapes and have sidechain interactions many bases apart [12, 35].

As a result, the choice was made to utilize an LSTM, which is a type of RNN known to be better capable of retaining and keeping track of long-term information [25, 37].

The basic procedure performed by an LSTM is as follows: while moving through the input sequence, the LSTM maintains two states which are referred to as the "cell" and "hidden" states [37]. The cell state essentially functions as the long-term memory of the LSTM, allowing it to "remember" past parts of a sequence that it has already seen [25]. The hidden state, on the other hand, is essentially the working or short-term memory of the LSTM [25]. The actual structure of the LSTM is composed of three gates: an "input," "output," and "forget" gate [25]. The forget gate takes as its input the current hidden state and the next value in the overall sequence that was input to the model [25]. It then decides which information should be discarded, and thus "forgotten," by the model [25]. The input gate also takes the current hidden state and the next value in the sequence being processed as its inputs [37]. It then decides whether this information will enter the cell state, and thus get incorporated into the long-term memory of the LSTM [37]. The new cell state is generated from the outputs of these two gates. Finally, the output gate takes as its input the original hidden state, the original cell state, and the new cell state calculated by the forget and input gates [25]. It then generates a new hidden state that will be passed to the LSTM during the next time step, along with the new cell state [25].

As previously detailed in 1.6, Mason et al. was able to successfully utilize an LSTM for the closely related field of predicting how sequence mutations impact protein function [57]. More generally, Agarwal et al. showed that bidirectional LSTMs more effectively "summarize[] the input sequence and capture[] important motif information" than alternative methods for learning compressed representations of DNA sequences [1]. Agarwal et al. specifically found that feeding the data both forwards and backwards through the LSTM, i.e. "bidirectionally," improved performance by allowing the network to leverage both past and future context when processing each character in a sequence [1]. Previous machine learning research using bidirectional LSTMs has also shown that greater accuracy and modeling performance is achievable with this method than unidirectional LSTMs [31, 32, 80].

Thus, a bidirectional LSTM was chosen to parametrize $q(z|c)$. Given the sequential nature of the dataset (literally referred to as HFNAP "sequences"), this choice made intuitive architectural sense to better allow the CVAE's encoder to capture the underlying properties of the input sequences.

The bidirectional LSTM implemented in the encoder consisted of a layer of 200 nodes followed by another layer of 200 nodes. Once the encoder output the appropriate latent variables (again of the form $(\mu_i, \sigma_i).1 \leq i \leq 15$), these latents were then fed through the decoder of the CVAE in addition to the original binding fitness condition c associated with that sequence.

The structure of the CVAE's decoding section was similar to the structure of the decoder of the VAE. At a high-level, the decoder took as input the 31x1 vector of 30 latent variables concatenated to the condition c. During training, the decoder would take a random sample from these latent Gaussians to get a vector of scalars. During inference, however, the mean of each distribution was utilized to represent our most likely estimate for the input [75].

The CVAE then passed this latent encoding through a fully connected layer of 480 nodes with $tanh()$ activations. The output of this layer was then split into 15 separate 32x1 vectors. Each 32x1 vector was fed through a separate network consisting of three fully connected layers of 32 nodes, where the first two layers had $tanh()$ activations and the final layer was softmaxed. Each of the 15 separate 32x1 vectors output by these parallel networks thus represented the probability that any of the 32 possible trimer-sidechain pairings could be found at a specific location in the HFNAP sequence. By taking the argmax over this vector, an actual sequence of A/T/G/C's that could be synthesized in the lab was generated.

In order to generate the predicted fitness that the model associated with this output sequence, this 32x15 vector was then fed into another set of 15 separate parallel networks that each contained two layers of 32 fully connected nodes with *tanh*() activations. The final layer in these 15 separate networks were then all connected to a single node with sigmoid activation which generated the reconstructed fitness value associated with the sequence generated by the CVAE.

A diagram which illustrates the updates that were made to transform the VAE it into the CVAE is shown in Figure 3.2.4.

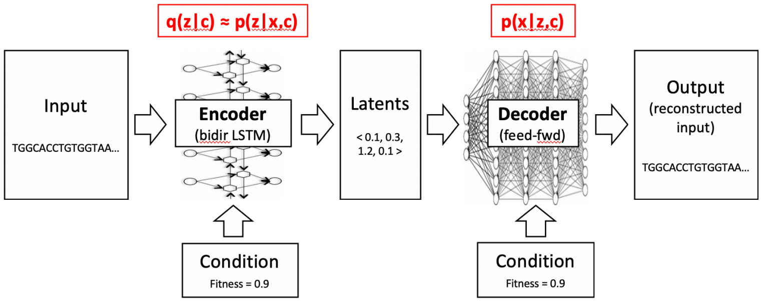

Figure 3.2.4: The conditional variational autoencoder used a bidirectional LSTM as its encoder and a series of parallel feed-forward networks as its decoder. The condition fed into the model at both the encoding and decoding stages was the estimated binding fitness of the input sequence.

The loss function for the model was again based on maximizing the ELBO through minimizing the cross entropy of the reconstructed sequence and the KL divergence of the learned Gaussians with respect to the prior of a standard normal. Unlike the VAE, however, an additional term was added to the loss of the CVAE in order to measure the ability of the model to reconstruct the fitness score – this reconstruction ability was measured via the mean squared error between the predicted fitness and true "ground truth" fitness label associated with the input HFNAP.

3.3 Experimental Validation

Once the CVAE model output a set of sequences predicted to have high binding fitness for the desired target, the next step was to experimentally validate the performance of each sequence in the lab.

This was achieved through both 1) running a high-stringency round of competitive selection pitting sequences generated from the CVAE against the top 2,000 sequences from the original training library generated through conventional SELEX, and 2) conducting an MST assay to measure the precise binding affinity for daunomycin of each of the top 5 CVAE-generated aptamers. These two steps were conducted in the wet lab by Jon C. Chen.

4
Results

This chapter details the computational results of training the VAE and CVAE on the daunomycin training dataset, as well as the experimental results generated through an MST assay of the top 5 sequences generated by the CVAE.

4.1 VAE

The VAE was trained on Harvard's FAS Research Computing Cannon cluster using one Nvidia V100 GPU, taking roughly two days to run for 300 epochs. The total size of the daunomycin SELEX dataset was 174,086 distinct sequences. This was split 80/20 to create a training set of 139,269 sequences and a testing set of 34,817 sequences. Hyperparameters including batch size, learning rate, drop out, and applying linear scaling factors to the terms of the ELBO in order to modify the relative influence of the reconstruction/regularization terms on the loss function (as re-weighting these terms has been shown to enable higher performance [3]) were chosen via grid search [45]. The final model was trained over 300 epochs with a batch size of 256, and optimized using ADAM with a learning rate of 0.001.

Overall, the VAE proved successful at reconstructing HFNAP sequences. Figure 4.1.1 demonstrates the performance attained over 300 epochs by the best VAE in terms of final testing error, and the rest of this section provides a brief analysis of these graphs.

"Accuracy" is defined as the percentage of identical bases between an input sequence and the output sequence of the VAE. For example, if the input sequence were "TAATGG" and the output of the VAE were "TAATCT," then the accuracy of the model would be 66%. The VAE seems to plateau in performance around the 100th epoch, achieving a slight performance gain of roughly 2% by the 300th epoch, but primarily oscillating around an accuracy of 92% between the 100th and 300th epochs. This means that when reconstructing an HFNAP of 45 nucleotides, we can expect the VAE to miss an average of roughly 3.6 nucleotides. This high but imperfect accuracy is actually a desirable feature of the VAE, for if the model simply reconstructed all sequences with 100% accuracy, then we would never be able to generate novel aptamer sequences as desired by the original goal of this paper. Rather, the variance in output demonstrated by this level of accuracy indicates that the model will be able to generate novel sequences not seen in our training set, but still achieve a fairly strong internal model of how to represent and generate HFNAP sequences.

"CE" stands for cross entropy, while "KLD" stands for KL divergence. Summed together (along with whatever scaling factors were applied [3]), these two values represent the loss of the VAE. The cross entropy measures the reconstruction error of the model $\mathbb{E}_{z\sim q(z)}[\log(p(x|z))]$, i.e. how well its outputs match its inputs. The KL divergence measures $D_{KL}(q(z)||p(z))$, which based on the modelling assumptions made behind the VAE, measures how similar the learned latent Gaussians are to the standard normal $N(0,1)$. As expected, the KL

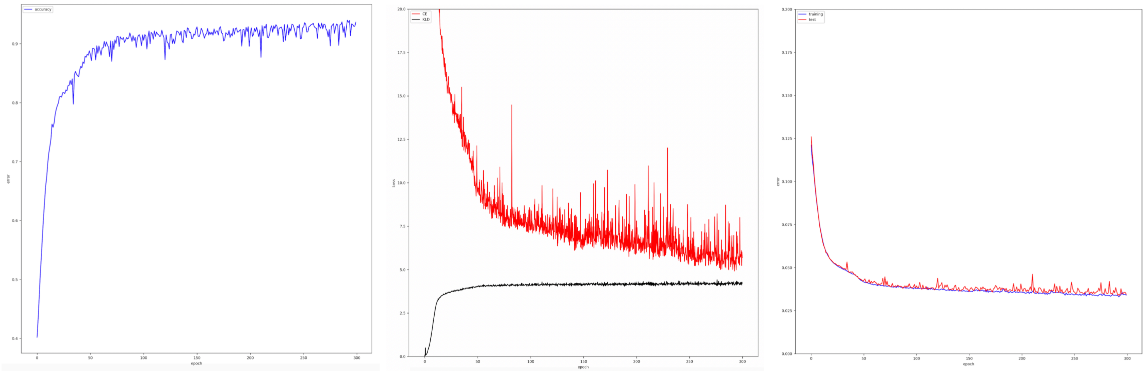

Figure 4.1.1: (A) "Accuracy" is defined as the percentage of identical bases between an input sequence and the output sequence of the VAE. The values depicted in the graph are calculated by averaging over the accuracies measured for each sequence per batch. The accuracy achieved by the VAE steadily increased until the 100th epoch, at which point it plateaued at around 92% accuracy. This means that, on average, the output of the VAE would have 3.6 nucleotides different than the input sequence. **(B)** "CE" (black line) stands for cross entropy, while "KLD" (blue line) stands for KL divergence. As the VAE learns to better represent its inputs in terms of a latent space of Gaussians by adjusting their distributions, the cross entropy goes down. The KL divergence, however, increases since the latent distributions drift farther away from the standard normal as they are adjusted by the VAE to better fit the data. **(C)** The training (blue line) and testing error (black line) decrease as the number of training epochs increases. The values in the graph were generated by running the model on both the training and testing set after each successive epoch of training. The error, a linear combination of the KL divergence and reconstruction error, plateau at around the 100th epoch. The downward trajectory of the error and tight fit between the training and testing errors indicate that the model is learning how to effectively compress the input without over-fitting to the training data.

divergence starts at zero since we initialize each latent to the standard normal, while the cross entropy loss peaks at the start of training since the VAE has not had time to learn how to effectively compress each input. As the VAE learns to better represent each sequence in terms of its latent space by adjusting its latent Gaussian distributions, however, the cross entropy loss is reduced. The KL divergence, on the other hand, increases as the VAE adjusts its latent distributions, and these distributions begin to drift farther away from the standard normal. The KL divergence eventually plateaus while the cross entropy continues to decrease, indicating that the VAE has learned to stop "cheating" by distorting its Gaussian distributions in order to represent its input, and is instead learning to associate a set of fixed distributions with abstractions that capture motifs in the input sequence.

The "training error" represents the overall loss of the VAE on the training set after each epoch, while the "testing error" measures the loss of the VAE on the testing set after each epoch. As expected, the testing error is higher than the training error, although the two appear to be extremely close in most instances, potentially indicating that the model is not suffering from over-fitting. The performance of the model seems to plateau around the 100th epoch, which mirrors what was observed in the plot of accuracy.

4.2 CVAE

The CVAE was trained on both Harvard's FAS Research Computing Cannon cluster using one Nvidia V100 GPU and Uber AI Lab's computing cluster (through Uber AI Labs research scientist Jon P. Chen), taking several hours to run for 100 epochs on Cannon and an unspecified amount of time on Uber's cluster. An 80/20 split was again used, thereby leading to a training set of 139,269 sequences and a testing set of 34,817 sequences. Again, hyperparameters including batch size, learning rate, drop out, and applying a linear scaling factor to the terms in the ELBO [3] were chosen via grid search [45]. The final model was trained over 100 epochs with a batch size of 256, and optimized using ADAM with a learning rate of 0.001.

An early obstacle encountered during training was the fact that very few randomly generated aptamer sequences will bind strongly to a given target. Thus, the space of fitness scores (the condition for our CVAE) was extremely sparse – as noted in 3.1.1, less than 2% of sequences in the dataset had a fitness score > 0.25. In addition to making it more difficult for the CVAE to learn what a strong sequence looked like amidst a flood of weak sequences, it was also difficult to test the model's ability to generate strong binders, since a random sample of only 20% of our dataset was (1) unlikely to contain many high-binding aptamers to test, and (2) would contain so many weak-binding aptamers that they would simply drown out the signal of the model's performance on high-binding sequences. Since the overarching goal of this project was to generate novel *high-binding* aptamers (the ability to generate weak binders is of little scientific interest), the model needed to somehow internalize that it should prioritize learning from high-binding training samples. Additionally, a testing metric that only evaluated the model's performance on high-binding inputs was needed.

The naïve solution of simply artificially restricting the number of weak-binders that the model trained on in order to increase the ratio of high to weak binders, however, would both dramatically reduce the size of our training set (thus decreasing model performance) and distort the landscape of aptamer fitness that was learned by the model. Thus, a compromise of withholding 50% of the top 500 enriched HFNAP sequences from the SELEX experiment was struck. These sequences were separated into a "high fitness" test set that would be utilized in addition to the normal test set to assess the CVAE's performance, and thus evaluate how well the model had specifically generalized for the space of high-binding HFNAPs. The trade-off with holding out such a large percentage of high-binding sequences for testing was that the CVAE would have less information to learn what high-binding sequences looked like. While this is certainly a potential drawback, this was not as much of an issue in practice as the plots of the model's performance illustrate. In fact, by preferentially selecting for trained models which had achieved the best scores on the "high fitness" test set, the final model that was selected ended up having higher performance on predicting high-binding sequences than on the overall test set. Thus, the benefit of gaining additional clarity into the CVAE's ability to model high-binding fitness space did not seem to come at too high a cost of harming the model's ability to train.

The performance of the best run of the CVAE is shown in Figure 4.2.1. As defined previously in 4.1, the term "Accuracy" represents the number of identical bases between the input and output sequences averaged over each batch, while "training error" is the model's loss on the training dataset, "high fitness test" is the model's loss on the held-out set of 250 top-binding sequences, and "general fitness test" is the loss of the CVAE on the 34,817 sequences on the entire testing set.

Once the CVAE was trained, it was time to generate completely novel aptamer sequences that would be predicted to have high binding fitness for daunomycin. The top 3,000 sequences from the original experimental SELEX run (as measured by their computed fitness scores) were fed into the CVAE to seed it for novel high-binding aptamer generation. Random samples were taken from the latent distributions generated by the model to represent these 3,000 high-binding HFNAPs. These latent representations were then fed through the decoder of the network along with a fitness condition $c \in [0, 1]$. In order to err on the conservative side given the uncertain confidence in the CVAE's true ability to generalize to untested spaces of the overall HFNAP fitness landscape, the condition c was varied from 0.5 to 0.95 in increments of 0.05, with 1,000 sequences generated at each value. This process was used to generate 10,000 total novel HFNAPs.

In order to assess how well the CVAE had generalized, the Levenshtein distance between every pair of CVAE-generated HFNAPs was calculated. Levenshtein distance represents the minimal number of insertions/deletions/substitutions needed to transform one string into another.

The Levenshtein distance $L(A, B)$ between two strings A and B can be

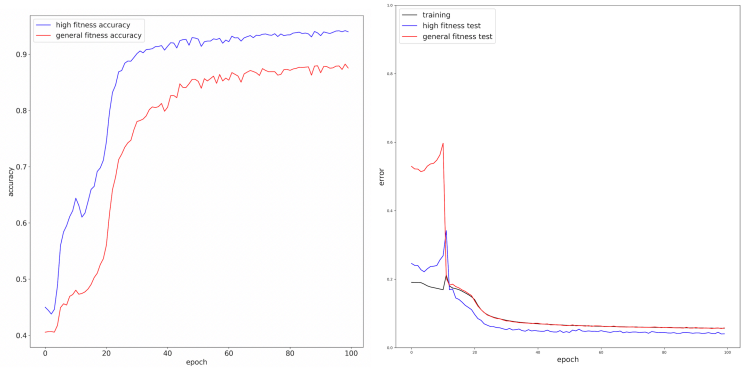

Figure 4.2.1: **(A)** "Accuracy" is defined as the percentage of identical bases between an input sequence and the output sequence of the CVAE. The values depicted in the graph are calculated by averaging over the accuracies measured for each sequence per batch. The "high fitness" measurement reflects the model's accuracy on the held-out set of 250 top-binding sequences, while the "general fitness" measurement reflects the model's accuracy on the overall testing set. The CVAE is able to reconstruct the high fitness sequences with roughly 91% accuracy by the 100th epoch, while the accuracy achieved on the entire testing set was roughly 86%. **(B)** The training (black line), high fitness testing error (blue line), and overall testing error (red line) decrease as the number of training epochs increases. The error is comprised of a linear combination of the KL divergence between the learned latent Gaussian distributions and the standard normal, the cross entropy between the reconstructed sequence and the original sequence input to the model, and the mean squared error of the reconstructed fitness value with the "ground truth" value associated with the input sequence. The model appears to achieve better performance on the high fitness testing set than it achieves on the overall training and testing sets. This is somewhat encouraging given the over-arching goal of optimizing the model to generate high-binding aptamers, even if that comes at the expense of being slightly worse at generating moderate- to low-binding sequences.

calculated using the following recursive formula, where $|x|$ is the length of string x, the expression $\mathbb{I}(A_i \neq B_j)$ is the indicator random variable equal to 1 when $A_i \neq B_j$ and $L(A, B) = d_{A,B}(|A|, |B|)$ [34]:

$$d_{A,B}(i,j) = \begin{cases} \max(i,j) & \min(i,j) = 0 \\ \min \begin{cases} d_{A,B}(i-1,j) + 1 \\ d_{A,B}(i,j-1) + 1 \\ d_{A,B}(i-1,j-1) + \mathbb{I}(A_i \neq B_j) \end{cases} & \min(i,j) \neq 0 \end{cases}$$

Thus, the lower the average minimal pairwise Levenshtein distance measured for the set of CVAE-generated HFNAPs, the better the CVAE will have generalized, for this indicates that the generated sequences were distinct from one another and thus covered a broader range of the theoretical search space. Encouragingly, only ~1 % of the CVAE-generated sequences had a minimum pairwise Levenshtein distance of < 10. Given that when calculating pairwise Levenshtein distances among the 3,000 top-binding HFNAPs used to seed the CVAE, over 70% possessed a minimum pairwise distance of ≤ 2 bases (reflecting poor coverage of the total theoretical space), this result indicates that the vast majority of sequences generated by the CVAE were substantially more different from each other than were the top sequences generated by SELEX. **Thus, the CVAE had covered a much wider range of the theoretical search space than its experimentally generated training input.**

4.3 Experimental Validation

The experimental validation of the novel HFNAPs generated by the CVAE was conducted by Jon C. Chen.

At a high-level, the validation process was as follows: DNA templates were synthesized for all 10,000 novel sequences, then translated into HFNAPs using the same procedure described in 2.2.2. These sequences were then pooled with 2,000 of the top-binding sequences from the original training set generated through conventional SELEX, then collectively run through one high-stringency round of selection against daunomycin. After removing HFNAPs that failed to bind to daunomycin, the remaining sequences were identified using high-throughput sequencing, and the overall results showed that a number of novel sequences had enriched to levels at or better than those achieved by the 2,000 top sequences from the training set.

The five CVAE-generated HFNAPs that were most enriched during this high-stringency selection round were then isolated. Their binding affinities for daunomycin were determined through MST, and all five were found to bind with $K_D = 10\text{-}30$ nM affinity, roughly in-line with the binding affinities of $K_D = 10\text{-}100$ nM that had been measured for the best sequences generated through the conventional SELEX used to generate the training dataset.

The MST dose-response curve for the five CVAE-generated aptamers is shown in Figure 4.3.1. The affinities of the CVAE-generated HFNAPs were also within an order of magnitude of the affinities measured for daunomycin binders generated through other experimental methods [18, 69]. Importantly, the minimum Levenshtein distance measured between each of these five CVAE-generated sequences and the most similar of the 3,000 sequences used to seed the CVAE were all > 12, indicating that the CVAE had successfully generalized to areas of the search space that had not been covered by the training set.

What do these measurements actually mean? K_D stands for the "dissociation constant," a measurement that is used in chemistry to measure a substance's binding affinity [44]. In the context of this experiment, the interpretation of the K_D value is as follows. Imagine you have a solution consisting solely of daunomycin. How much HFNAP do you need to add to that solution such that half of the daunomycin will get bound to HFNAP? The K_D represents the concentration needed of HFNAP such that half of the daunomycin molecules will be bound to that HFNAP. Thus, a lower K_D is better, as it means that less of an HFNAP will be required to bind to a given concentration of daunomycin

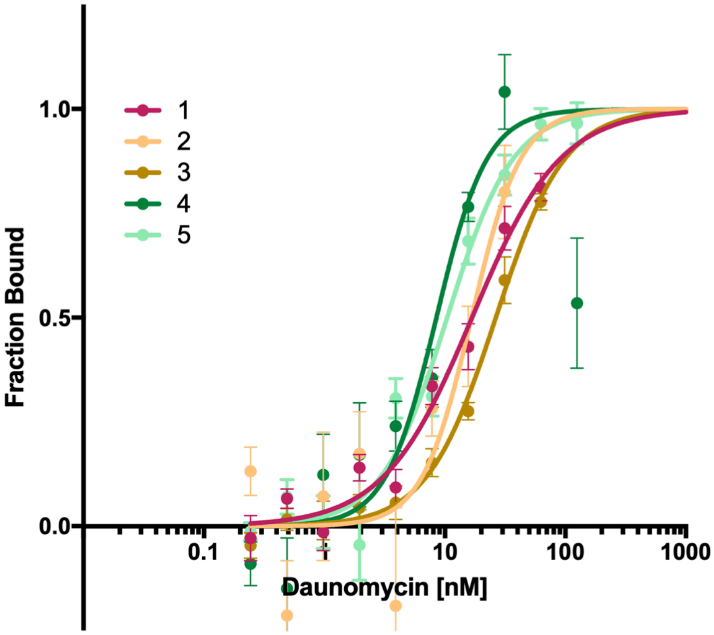

Figure 4.3.1: Dose-response curves for each of the five top-performing HF-NAPs generated by the CVAE. This data was generated by running each sequence through a microscale thermophoresis assay. Image and data generated by Jon C. Chen.

[44]. Thus, a concentration of roughly 10-30 nM would be required for the top five HFNAPs generated through the CVAE model to bind to half of the daunomycin in a shared solution.

5

Conclusion

By building a computational model that can better generalize the results of the low-throughput *in vitro* selections required to discover novel high-binding aptamers, this thesis presents a promising proof-of-concept of the application of deep learning methods to aptamer screening.

Though robust machine learning methods had not yet been applied to this subfield of biochemistry, a CVAE trained on eight rounds of SELEX data showed the ability to effectively capture sequence motifs and associate them with binding fitness for a small molecule target, daunomycin. As demonstrated experimentally, novel sequences generated by the model that were predicted to have strong binding fitness for daunomycin did exhibit affinities that were in-line with sequences discovered using conventional wet lab techniques. In particular, the five top sequences generated by the model all had $K_D =$ 10-30 nM, as measured by MST. Importantly, all of these sequences were entirely novel, and represented distinct elements of the theoretical search space for HFNAPs that were not covered by the original SELEX dataset.

This thesis shows how computational biology can help to vastly expand the search space for potentially life-saving aptamers while reducing the time, effort, and cost needed to identify them. Machine learning can hasten the speed at which aptamers are screened and subsequently deployed for therapeutic and diagnostic applications.

Unfortunately, progress of the field has been severely hampered by the low-throughput of experimental techniques used to discover aptamers with the necessary functional properties [89]. Improvements offered by deep learning methods to the process of aptamer screening can help accelerate the transformation of aptamers from a still relatively unproven technology in the lab to the front lines of therapeutics and diagnostics [63].

There still, however, remains substantial work to be accomplished in the field. Several areas of improvement and possible extensions for the models discussed in this paper abound. First, permitting arbitrary trimer-sidechain pairings would greatly expand the expressive power of the model and allow it to identify a much broader range of potentially high-binding HFNAPs with unique functional properties [52]. Second, adding the ability to factor in the secondary/tertiary structure of an aptamer sequence may increase the accuracy of the model at the expense of speed [44]. Such a trade-off will need to be more closely studied, as one of the main advantages of the CVAE discussed in this paper was its relatively low computational cost. However, secondary/tertiary structure is known to be highly predictive of an aptamer's binding affinity, and this type of modeling is already leveraged by other aptamer screening computational prediction engines [35, 44]. Thus, leveraging these existing structural prediction tools to label aptamers with their 3D structures before feeding them into a modified CVAE may yield superior results. Finally, a completely different modeling approach could have also been pursued instead of utilizing a VAE. For example, generative adversarial networks (GANs), another commonly utilized unsupervised generative model, could have been utilized to learn the fitness landscape revealed by SELEX data [27]. GANs are known to generate sharper distinctions between their learned latent distributions than VAEs, and thus could improve the quality of generated HFNAPs by reducing "blurring" between motifs [5].

More broadly, this thesis hopes to contribute to a wider discourse surrounding the field of aptamers, computational biology, and the intersection of machine learning and wet lab experimentation. While this thesis attempted to construct a computational model around an already-established experimental procedure in order to amplify its ability to generate novel insights, tighter coupling of deep learning analyses and experimental design in the future should yield improved results and further accelerate advancements in chemistry and healthcare.

A
Appendix

A.1 Code

Code for the machine learning models, data processing, diagrams, and analyses is located across the following two Github repositories: (1) HFNAP-Binding and (2) Small-Molecule-Binding. The former is public. The latter is private as of April 3rd 2020, due to its connection to a paper that is still in preparation.

References

[1] Vishal Agarwal, N Reddy, and Ashish Anand. Unsupervised representation learning of dna sequences. *arXiv preprint arXiv:1906.03087*, 2019.

[2] Khalid K Alam, Jonathan L Chang, and Donald H Burke. Fastaptamer: a bioinformatic toolkit for high-throughput sequence analysis of combinatorial selections. *Molecular Therapy-Nucleic Acids*, 4:e230, 2015.

[3] AA Alemi, I Fischer, JV Dillon, and K Murphy. Deep variational information bottleneck int. In *Conf. on Learning Representations*, 2017.

[4] Mohamed H Ali, Marwa E Elsherbiny, and Marwan Emara. Updates on aptamer research. *International journal of molecular sciences*, 20(10):2511, 2019.

[5] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Cvae-gan: fine-grained image generation through asymmetric training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2745–2754, 2017.

[6] Yoshua Bengio. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 17–36, 2012.

[7] Ronald H Blum and Stephen K Carter. Adriamycin: a new anticancer drug with significant clinical activity. *Annals of internal medicine*, 80(2):249–259, 1974.

[8] Bernard R Brooks, Charles L Brooks III, Alexander D Mackerell Jr, Lennart Nilsson, Robert J Petrella, Benoît Roux, Youngdo Won, Georgios Archontis, Christian Bartels, Stefan Boresch, et al. Charmm: the biomolecular simulation program. *Journal of computational chemistry*, 30 (10):1545–1614, 2009.

[9] Laura C Cappelli, Ami A Shah, and Clifton O Bingham. Immune-related adverse effects of cancer immunotherapy—implications for rheumatology. *Rheumatic Disease Clinics*, 43(1):65–78, 2017.

[10] Jimmy Caroli, Cristian Taccioli, A De La Fuente, Paolo Serafini, and Silvio Bicciato. Aptani: a computational tool to select aptamers through sequence-structure motif analysis of ht-selex data. *Bioinformatics*, 32(2): 161–164, 2016.

[11] Zhen Chen, Phillip A Lichtor, Adrian P Berliner, Jonathan C Chen, and David R Liu. Evolution of sequence-defined highly functionalized nucleic acid polymers. *Nature chemistry*, 10(4):420–427, 2018.

[12] Yaroslav Chushak and Morley O Stone. In silico selection of rna aptamers. *Nucleic acids research*, 37(12):e87–e87, 2009.

[13] Jon Cohen and Kai Kupferschmidt. Labs scramble to produce new coronavirus diagnostics. *Science*, 367(6479):727–727, 2020. ISSN 0036-8075. doi: 10.1126/science.367.6479.727.

[14] Phuong Dao, Jan Hoinka, Mayumi Takahashi, Jiehua Zhou, Michelle Ho, Yijie Wang, Fabrizio Costa, John J Rossi, Rolf Backofen, John Burnett, et al. Aptatrace elucidates rna sequence-structure motifs from selection trends in ht-selex experiments. *Cell systems*, 3(1):62–70, 2016.

[15] Mariia Darmostuk, Silvie Rimpelova, Helena Gbelcova, and Tomas Ruml. Current approaches in selex: An update to aptamer selection technology. *Biotechnology advances*, 33(6):1141–1161, 2015.

[16] David R Davies and Susan Chacko. Antibody structure. *Accounts of chemical research*, 26(8):421–427, 1993.

[17] Erik Doevendans and Huub Schellekens. Immunogenicity of innovative and biosimilar monoclonal antibodies. *Antibodies*, 8(1):21, Mar 2019. ISSN 2073-4468. doi: 10.3390/antib8010021.

[18] Benjamin Doughty, Yi Rao, Samuel W Kazer, Sheldon JJ Kwok, Nicholas J Turro, and Kenneth B Eisenthal. Binding of the anti-cancer drug daunomycin to dna probed by second harmonic generation. *The Journal of Physical Chemistry B*, 117(49):15285–15289, 2013.

[19] Dropbox. How much is 1 tb of storage? https://www.dropbox.com/features/cloud-storage/how-much-is-1tb, 2020.

[20] Andrew D Ellington and Jack W Szostak. In vitro selection of rna molecules that bind specific ligands. *nature*, 346(6287):818–822, 1990.

[21] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2): 179–211, 1990.

[22] Suzanne S Farid. Process economics of industrial monoclonal antibody manufacture. *Journal of Chromatography B*, 848(1):8–18, 2007.

[23] Tara C Gangadhar and Robert H Vonderheide. Mitigating the toxic effects of anticancer immunotherapy. *Nature Reviews Clinical Oncology*, 11(2):91, 2014.

[24] Bharat N Gawande, John C Rohloff, Jeffrey D Carter, Ira von Carlowitz, Chi Zhang, Daniel J Schneider, and Nebojsa Janjic. Selection of dna aptamers with two modified bases. *Proceedings of the National Academy of Sciences*, 114(11):2898–2903, 2017.

[25] Felix A. Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. *Neural Computation*, 12(10): 2451–2471, 2000. doi: 10.1162/089976600300015015.

[26] Garrett B. Goh, Nathan O. Hodas, and Abhinav Vishnu. Deep learning for computational chemistry. *Journal of Computational Chemistry*, 38(16): 1291–1307, 2017. doi: 10.1002/jcc.24764.

[27] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[28] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[29] Chelsea KL Gordon, Diana Wu, Anusha Pusuluri, Trevor A Feagin, Andrew T Csordas, Michael S Eisenstein, Craig J Hawker, Jia Niu, and Hyongsok Tom Soh. Click-particle display for base-modified aptamer discovery. *ACS chemical biology*, 2019.

[30] Michael R Gotrik, Trevor A Feagin, Andrew T Csordas, Margaret A Nakamoto, and H Tom Soh. Advancements in aptamer discovery technologies. *Accounts of chemical research*, 49(9):1903–1910, 2016.

[31] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2047–2052 vol. 4, 2005.

[32] A. Graves, A. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649, 2013.

[33] António L Grilo and A Mantalaris. The increasingly human and profitable monoclonal antibody market. *Trends in biotechnology*, 37(1):9–16, 2019.

[34] Rishin Haldar and Debajyoti Mukhopadhyay. Levenshtein distance technique in dictionary lookup methods: An improved approach. *arXiv preprint arXiv:1101.1232*, 2011.

[35] Michiaki Hamada. In silico approaches to rna aptamer design. *Biochimie*, 145:8–14, 2018.

[36] Ryan Hili, Jia Niu, and David R Liu. Dna ligase-mediated translation of dna into densely functionalized nucleic acid polymers. *Journal of the American Chemical Society*, 135(1):98–101, 2013.

[37] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[38] Jan Hoinka, Alexey Berezhnoy, Zuben E Sauna, Eli Gilboa, and Teresa M Przytycka. Aptacluster–a method to cluster ht-selex aptamer pools and lessons from its application. In *International Conference on Research in Computational Molecular Biology*, pages 115–128. Springer, 2014.

[39] Brody Huval, Tao Wang, Sameep Tandon, Jeff Kiske, Will Song, Joel Pazhayampallil, Mykhaylo Andriluka, Pranav Rajpurkar, Toki Migimatsu, Royce Cheng-Yue, Fernando A. Mujica, Adam Coates, and Andrew Y. Ng. An empirical evaluation of deep learning on highway driving. *CoRR*, abs/1504.01716, 2015.

[40] Johan Ludwig William Valdemar Jensen et al. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta mathematica*, 30:175–193, 1906.

[41] Sol A Jeon, Jong Lyul Park, Jong-Hwan Kim, Jeong Hwan Kim, Yong Sung Kim, Jin Cheon Kim, and Seon-Young Kim. Comparison of the mgiseq-2000 and illumina hiseq 4000 sequencing platforms for rna sequencing. *Genomics & informatics*, 17(3), 2019.

[42] Xiaohui Jiang, Kamal Kumar, Xin Hu, Anders Wallqvist, and Jaques Reifman. Dovis 2.0: an efficient and easy to use parallel virtual screening tool based on autodock 4.0. *Chemistry Central Journal*, 2(1):18, 2008.

[43] H Kaur, JG Bruno, A Kumar, and TK Sharma. Aptamers in the therapeutics and diagnostics pipelines. theranostics 8 (15): 4016–4032, 2018.

[44] Andrew B Kinghorn, Lewis A Fraser, Shaolin Liang, Simon Chi-Chin Shiu, and Julian A Tanner. Aptamer bioinformatics. *International journal of molecular sciences*, 18(12):2516, 2017.

[45] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[46] Daniel C Koboldt, Karyn Meltz Steinberg, David E Larson, Richard K Wilson, and Elaine R Mardis. The next-generation sequencing revolution and its impact on genomics. *Cell*, 155(1):27–38, 2013.

[47] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 03 1951. doi: 10.1214/aoms/1177729694.

[48] Gillian V Kupakuwana, James E Crill, Mark P McPike II, and Philip N Borer. Acyclic identification of aptamers for human alpha-thrombin using over-represented libraries and deep sequencing. *PloS one*, 6(5), 2011.

[49] Agata Levay, Randall Brenneman, Jan Hoinka, David Sant, Marco Cardone, Giorgio Trinchieri, Teresa M Przytycka, and Alexey Berezhnoy. Identifying high-affinity aptamer ligands with defined cross-reactivity using high-throughput guided systematic evolution of ligands by exponential enrichment. *Nucleic acids research*, 43(12):e82–e82, 2015.

[50] Bennett Levitan. Stochastic modeling and optimization of phage display. *Journal of molecular biology*, 277(4):893–916, 1998.

[51] Ruizhe Li, Xiao Li, Chenghua Lin, Matthew Collinson, and Rui Mao. A stable variational autoencoder for text modelling. *arXiv preprint arXiv:1911.05343*, 2019.

[52] Phillip A Lichtor, Zhen Chen, Nadine H Elowe, Jonathan C Chen, and David R Liu. Side chain determinants of biopolymer function during selection and replication. *Nature chemical biology*, 15(4):419–426, 2019.

[53] Shuyu Lin, Stephen Roberts, Niki Trigoni, and Ronald Clark. Balancing reconstruction quality and regularisation in elbo for vaes. *arXiv preprint arXiv:1909.03765*, 2019.

[54] Ge Liu, Haoyang Zeng, Jonas Mueller, Brandon Carter, Ziheng Wang, Jonas Schilz, Geraldine Horny, Michael E Birnbaum, Stefan Ewert, and David K Gifford. Antibody complementarity determining region design using high-capacity machine learning. *Bioinformatics*, 11 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz895. btz895.

[55] Yu-Hui Liu, Brian Giunta, Hua-Dong Zhou, Jun Tan, and Yan-Jiang Wang. Immunotherapy for alzheimer disease—the challenge of adverse effects. *Nature Reviews Neurology*, 8(8):465–469, 2012.

[56] Gary Marcus. Deep learning: A critical appraisal. *CoRR*, abs/1801.00631, 2018.

[57] Derek M Mason, Simon Friedensohn, Cédric R Weber, Christian Jordi, Bastian Wagner, Simon Meng, Pablo Gainza, Bruno Correia, and Sai T Reddy. Deep learning enables therapeutic antibody optimization in mammalian cells by deciphering high-dimensional protein sequence space. *bioRxiv*, page 617860, 2019.

[58] Diane R Mould and Bernd Meibohm. Drug development of therapeutic monoclonal antibodies. *BioDrugs*, 30(4):275–293, 2016.

[59] Shuaijian Ni, Houzong Yao, Lili Wang, Jun Lu, Feng Jiang, Aiping Lu, and Ge Zhang. Chemical modifications of nucleic acid aptamers for therapeutic purposes. *International journal of molecular sciences*, 18(8):1683, 2017.

[60] World Health Organization. Model list of essential medicines: 21st list, 2019, 2019.

[61] Alicia Oshlack and Matthew J Wakefield. Transcript length bias in rna-seq data confounds systems biology. *Biology direct*, 4(1):14, 2009.

[62] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.

[63] Karlis Pleiko, Liga Saulite, Vadims Parfejevs, Karlis Miculis, Egils Vjaters, and Una Riekstina. Differential binding cell-selex method to identify cell-specific aptamers using high-throughput sequencing. *Scientific reports*, 9(1):1–12, 2019.

[64] Peng Qiu. Embracing the dropouts in single-cell rna-seq analysis. *Nature Communications*, 11(1):1–9, 2020.

[65] Adam J. Riesselman*, John B. Ingraham*, and Debora S. Marks. Deep generative models of genetic variation capture mutation effects. *Nature Methods*, 15:816–822, 2018.

[66] Adam Roberts, Cole Trapnell, Julie Donaghey, John L Rinn, and Lior Pachter. Improving rna-seq expression estimates by correcting for fragment bias. *Genome biology*, 12(3):R22, 2011.

[67] Carol A Rohl, Charlie EM Strauss, Kira MS Misura, and David Baker. Protein structure prediction using rosetta. In *Methods in enzymology*, volume 383, pages 66–93. Elsevier, 2004.

[68] Romelia Salomon-Ferrer, David A. Case, and Ross C. Walker. An overview of the amber biomolecular simulation package. *WIREs Computational Molecular Science*, 3(2):198–210, 2013. doi: 10.1002/wcms.1121.

[69] Stephan Sass, Walter FM Stöcklein, Anja Klevesath, Jeanne Hurpin, Marcus Menger, and Carsten Hille. Binding affinity data of dna aptamers for therapeutic anthracyclines from microscale thermophoresis and surface plasmon resonance spectroscopy. *Analyst*, 144(20):6064–6073, 2019.

[70] Stuart L Schreiber. A chemical biology view of bioactive small molecules and a binder-based approach to connect biology to precision medicines. *Israel journal of chemistry*, 59(1-2):52, 2019.

[71] Stanislau Semeniuta, Aliaksei Severyn, and Erhardt Barth. A hybrid convolutional variational autoencoder for text generation. *arXiv preprint arXiv:1702.02390*, 2017.

[72] Jay Shendure, Shankar Balasubramanian, George M Church, Walter Gilbert, Jane Rogers, Jeffery A Schloss, and Robert H Waterston. Dna sequencing at 40: past, present and future. *Nature*, 550(7676):345–353, 2017.

[73] Alex Sherstinsky. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *CoRR*, abs/1808.03314, 2018.

[74] Sam Sinai, Eric Kelsic, George M Church, and Martin A Nowak. Variational auto-encoding of protein sequences. *arXiv preprint arXiv:1712.03346*, 2017.

[75] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*, pages 3483–3491, 2015.

[76] Tianbao Song, Jingbo Sun, Bo Chen, Weiming Peng, and Jihua Song. Latent space expanded variational autoencoder for sentence generation. *IEEE Access*, 7:144618–144627, 2019.

[77] Paul AJ Speth, Reinier AP Raijmakers, Jan BM Boezeman, Peter CM Linssen, Theo JM de Witte, Hans MC Wessels, and Clemens Haanen. In vivo cellular adriamycin concentrations related to growth inhibition of normal and leukemic human bone marrow cells. *European Journal of Cancer and Clinical Oncology*, 24(4):667–674, 1988.

[78] Craig Tuerk and Larry Gold. Systematic evolution of ligands by exponential enrichment: Rna ligands to bacteriophage t4 dna polymerase. *science*, 249(4968):505–510, 1990.

[79] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[80] Di Wang and Eric Nyberg. A long short-term memory model for answer sentence selection in question answering. In *Proceedings of the 53rd Annual*

Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 707–712, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-2116.

[81] Aniela Wochner, Marcus Menger, Dagmar Orgel, Birgit Cech, Martina Rimmele, Volker A Erdmann, and Jörn Glökler. A dna aptamer with high affinity and specificity for therapeutic anthracyclines. *Analytical biochemistry*, 373(1):34–42, 2008.

[82] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016.

[83] Yijun Xiao, Tiancheng Zhao, and William Yang Wang. Dirichlet variational autoencoder for text modeling. *arXiv preprint arXiv:1811.00135*, 2018.

[84] Kevin K Yang, Zachary Wu, and Frances H Arnold. Machine-learning-guided directed evolution for protein engineering. *Nature methods*, 16(8):687–694, 2019.

[85] Zichao Yang, Zhiting Hu, Ruslan Salakhutdinov, and Taylor Berg-Kirkpatrick. Improved variational autoencoders for text modeling using dilated convolutions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3881–3890. JMLR. org, 2017.

[86] Matthew D Young, Matthew J Wakefield, Gordon K Smyth, and Alicia Oshlack. Gene ontology analysis for rna-seq: accounting for selection bias. *Genome biology*, 11(2):R14, 2010.

[87] Yang Zhang, Bo Shiun Lai, and Mario Juhas. Recent advances in aptamer discovery and applications. *Molecules*, 24(5):941, 2019.

[88] Wei Zheng, Lisa M Chung, and Hongyu Zhao. Bias detection and correction in rna-sequencing data. *BMC bioinformatics*, 12(1):290, 2011.

[89] Jiehua Zhou and John Rossi. Aptamers as targeted therapeutics: current potential and challenges. *Nature reviews Drug discovery*, 16(3):181, 2017.

[90] Zhenjian Zhuo, Yuanyuan Yu, Maolin Wang, Jie Li, Zongkang Zhang, Jin Liu, Xiaohao Wu, Aiping Lu, Ge Zhang, and Baoting Zhang. Recent advances in selex technology and aptamer applications in biomedicine. *International journal of molecular sciences*, 18(10):2142, 2017.

[91] James Zou, Mikael Huss, Abubakar Abid, Pejman Mohammadi, Ali Torkamani, and Amalio Telenti. A primer on deep learning in genomics. *Nature genetics*, 51(1):12–18, 2019.

[92] Michael Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic acids research*, 31(13):3406–3415, 2003.