digitalSELEX:
A Novel Oligonucleotide Design Platform

Stephen Gunther Hummel

A dissertation

submitted to the Faculty of

the department of Biology

in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Boston College
Morrissey College of Arts and Sciences
Graduate School

March 2023

digitalSELEX: A Novel Oligonucleotide Design Platform

Stephen Gunther Hummel

Advisor: Tim van Opijnen, Ph.D.

Molecules that have high affinity and specificity for their target are critical for functioning biosensors and effective therapeutics. Aptamers, or single-stranded oligonucleotides, are one type of molecule capable of both high affinity and specificity. Systematic Evolution of Ligands by EXponential enrichment (SELEX) is the iterative *in vitro* process for identifying aptamers with high affinity and specificity from an initial pool of approximately 10^{15} randomized nucleotide molecules. There have been a multitude of SELEX variations developed over the years to include incorporation of machine learning algorithms to address the limited success (~30%), cost, and time required to identify high affinity and specific aptamers.

While some SELEX variations have been more successful than others in addressing some of the challenges, issues remain. To confront these challenges, the digitalSELEX platform introduces a novel *de novo* design approach. The platform has two main components. The first component analyzes the target molecule identifying clusters of amino acids along the molecule's surface based on their accessibility and proximity of atoms relevant to target-aptamer binding. The platform then proposes aptamers built from sequences of nucleotides that paired to the amino acids in the clusters. The second component improves these aptamers sequentially. This is done via simulation-based optimization procedure which uses molecular docking and stochastic optimization techniques. It explores small adjustments made on the starting aptamer that

increase the affinity and specificity that is calculated extracting binding related features from the output of the docker. Once *in silico* counter-selection is complete, the best possible sequences are extracted for *in vitro* validation.

To validate digitalSELEX, aptamers were designed for four different target molecules of varying size ranging from 18 – 140 kDa. Some of the aptamers were designed with specific counter-targets while others did not have counter-target molecules. In total, 19 oligonucleotides were chemically synthesized, and their affinity and specificity tested for five explicit validation problems. All 19 aptamers demonstrated high affinity for their respective target molecules. Sixteen of the 19 oligonucleotides were tested for specificity with nine meeting the 4-times K_d-value difference specificity criteria. Depending on the computational capacity being employed for each problem, the approximate time required from initiating the *de novo* design to the point of validation was 170 hours. The cost of *in silico* oligonucleotide design is negligible while validation of a few aptamers is few hundred dollars.

The digitalSELEX platform was comprehensively tested examining the initial *de novo* design through affinity and specificity determination. The digtalSELEX platform is a prototype that has the opportunity for further development such as employing different molecular simulators.

TABLE OF CONTENTS

Table of Contents v
List of Tables viii
List of Figures ix
List of Terminology xii
List of Abbreviations xiv
Acknowledgements xv
Disclaimer xvii
1.0 Chapter 1 – Introduction 1
1.1 The Problem 1
1.2 Aptamers 2
1.3 Oligonucleotide Selection 3
1.3.1 In Vitro SELEX 3
1.3.2 In Silico Selection 6
1.3.2.1 Machine Learning 6
1.3.2.2 Molecular Simulations 7
1.3.3 Hybrid Selection Strategies 8
1.4 The MeasureMent Problem 9
1.5 Challenges with EXISTING Selection Methods 10
1.5.1 Challenges: Affinity and Specificity 10
1.5.2 Challenges: Success Rate 11
1.5.3 Challenges: Cost 11
1.5.4 Challenges: Time 12
1.6 Proposed Method 13
1.7 Road Map 15
2.0 Chapter 2 – The Measurement Problem 20
2.1 Defining Good and Bad Aptamers 20
2.1.1 Oligonucleotide Structure Stability 21
2.1.2 Oligonucleotide-Target Binding 21
2.2 Facets of Specificity 25
2.2.1 Cofactor Influence on Binding 28
2.2.2 Application Based Design Constraints 29
2.3 Measuring Affinity and Specificity 30
2.3.1 Experimental Design 31
2.3.2 Potential For Non-Specific Binding 33
2.3.3 Measurement Options 34
2.4 Experimental Set-Up Details 34
2.4.1 Magnetic Bead Preparation 35
2.4.2 Protein-Fluorophore Preparation 37
2.4.3 Flow Cytometry 38
2.4.4 Negative Controls – Unbound Fluorophore Contribution 40

2.4.5 Flow Cytometry Data Analysis 41
2.4.6 Calculating K_d-value 41

3.0 Chapter 3 – Cold Start Module 55
3.1 Design Concept 55
3.2 Concept Algorithms 56
3.2.1 Algorithm 1: Concept 57
3.2.2 Algorithm 2: Concept 58
3.2.3 The Core Sequence 59
3.3 Novel Components 59
3.4 Implementation Details 60
3.4.1 pdb file import 60
3.4.2 Identification of Biologically Relevant and Accessible Atoms 62
3.4.3 Algorithm 1: K-means Clustering 64
3.4.3.1 Nucleotide Assignment 64
3.4.3.2 Algorithm 2: Probabilistic Application Algorithm 65
3.4.4 Initial Sequence Generation 67
3.4.5 Designated Sequence Sources 68
3.4.5.1 Basic Constraints 69
3.4.5.2 Unique Constraints 70
3.5 Potential Limitations 71

4.0 Chapter 4 – Warm Start Module 82
4.1 Central Concept 82
4.2 Generating Small Changes 83
4.2.1 Evaluating Small Changes 84
4.2.2 Sequence to Structure 85
4.2.2.1 Secondary Structure Generation 85
4.2.2.2 Tertiary Structure Generation 86
4.2.3 Molecular Docking 87
4.2.4 Molecular Dynamics 89
4.2.5 Molecular Docker Selection 89
4.2.6 Selecting Mutations that Improve the Sequence 91
4.3 Novel Sections 92
4.4 Implementation Details 93
4.4.1 Module Options 93
4.4.1.1 Number of Counter-Targets 93
4.4.1.2 Protonation Option 94
4.4.1.3 Scoring Function 94
4.4.1.4 Fixed Sequence Portion 94
4.4.1.5 Number of Mutated Sequences 95
4.4.1.6 Number of Mutations 95
4.4.2 Small Sequence Change Implementation 95
4.4.2.1 Sequence to Structure 96
4.4.3 Evaluating Small Changes 97
4.4.3.1 Scoring Function 99
4.4.4 Final Sequence Selection 102
4.4.5 Sequence Extraction 103
4.5 Potential Limitations 103

5.0 Chapter 5 – digitalSELEX Platform Validation 115
5.1 Platform Validation Problem Sets 115
5.1.1 Problem 1: *de novo* Spike with HA Specificity Aptamer 115

5.1.1.1 Spike: Prior Knowledge for Design 117
5.1.1.2 Spike Aptamer Goal 117
5.1.2 Problem 2: *de novo* ACE2 aptamer 118
5.1.2.1 ACE2: Prior Knowledge for Design 119
5.1.2.2 ACE2 Aptamer Goal 119
5.1.3 Problem 3: *de novo* HA Aptamer 120
5.1.3.1 HA: Prior Knowledge for Design 121
5.1.3.2 HA Aptamer Goal 121
5.1.4 Problem 4: *de novo* PD1 Aptamer 121
5.1.4.1 PD1: Prior Knowledge for Design 122
5.1.4.2 PD1 Aptamer Goal 123
5.1.5 Problem 5: *de novo* Spike Aptamer with ACE2 Specificity 123
5.1.5.1 Spike-ACE2: Prior Knowledge for Design 123
5.1.5.2 Spike-ACE2 Aptamer Goal 124
5.2 Platform Validation Results 124
5.2.1 Problem 1: *de novo* Spike with HA Specificity Results 124
5.2.1.1 *Cold Start* Module 124
5.2.1.2 *Warm Start* Module 126
5.2.1.3 Validation 127
5.2.2 Problem 2: *de novo* ACE2 Results 128
5.2.2.1 *Cold Start* Module 128
5.2.2.2 *Warm Start* Module 129
5.2.2.3 Validation 130
5.2.3 Problem 3: *de novo* HA Results 132
5.2.3.1 *Cold Start* Module 132
5.2.3.2 *Warm Start* Module 133
5.2.3.3 Validation 134
5.2.4 Problem 4: *de novo* PD1 Results 135
5.2.4.1 *Cold Start* Module 135
5.2.4.2 *Warm Start* Module 136
5.2.4.3 Validation 136
5.2.5 Problem 5: *de novo* Spike Results with ACE2 Specificity Results 137
5.2.5.1 *Cold Start* Module 137
5.2.5.2 *Warm Start* Module 137
5.2.5.3 Validation 138
5.3 Results Discussion 139
5.3.1 Sustains 139
5.3.2 Improves 140
6.0 Chapter 6 – Discussion / Conclusions 177
6.1 Platform Validation Overview 177
6.2 Challenges Addressed 180
6.3 Opportunities for Improvement 182
6.4 Lessons Learned 183
6.5 Future Applications 184
7.0 References 187

LIST OF TABLES

Table 1. de novo Spike Warm Start Options..... 148

Table 2. de novo ACE2 Warm Start Options..... 155

Table 3. de novo HA Warm Start Options..... 164

Table 4. de novo PD1 Warm Start Options..... 172

Table 5. de novo Spike with ACE2 Specificity Warm Start Options. 174

Table 6. Summary of Oligonucleotides form Validation Problems..... 186

LIST OF FIGURES

Figure 1-1. Basic G-FET. 16

Figure 1-2. General SELEX Method. 17

Figure 1-3. Generic Oligonucleotide Library 18

Figure 1-4. Proposed digitalSELEX Model. 19

Figure 2-1. Intermolecular Forces. 44

Figure 2-2. Disrupting Forces. 45

Figure 2-3. Aptamer 1C and 4C Affinity and Specificity. 46

Figure 2-4. Bead Configurations. 47

Figure 2-5. Visualization of Validation Configuration Tests. 48

Figure 2-6. Effect of Non-Specific Binding. 49

Figure 2-7. BSA Effect on Binding Measurements 50

Figure 2-8. Experimental Set-up 51

Figure 2-9. Dose Response Curve with Negative Control. 52

Figure 2-10. FlowJo Fluorescent Positive Population Analysis. 53

Figure 2-11. Dose Response Curve Analysis 54

Figure 3-1. Cold Start Process. 73

Figure 3-2. PDB File Structure. 74

Figure 3-3. Delaunay Triangulation and Solid Angle. 75

Figure 3-4. Algorithm 1 Process. 76

Figure 3-5. Nucleotide - Amino Acid Frequency. 77

Figure 3-6. Algorithm 2 Process. 78

Figure 3-7. Genetic Optimization Algorithm Process. 79

Figure 3-8. Genetic Algorithm Biological Operators. 80

Figure 3-9. Cold Start Output. 81

Figure 4-1. Aptamer-Target Binding Energy Landscape..... 105

Figure 4-2. Sequence to Structure Process..... 106

Figure 4-3. Molecular Docking Methodology. 107

Figure 4-4. HDOCK Output Format. 108

Figure 4-5. Warm Start Process. 109

Figure 4-6. RNAComposer Submission. 110

Figure 4-7. RNAComposer Strucutre Example. 111

Figure 4-8. Scoring Function Options..... 112

Figure 4-9. Sequence Selection Process. 113

Figure 4-10. Final Sequence Extraction..... 114

Figure 5-1. Aptamers 1C and 4C Binding Spike, ACE2, and HA..... 142

Figure 5-2. Aptamer 6 Binding ACE2 and Spike. 143

Figure 5-3. Aptamer 1 Binding HA and Spike. 144

Figure 5-4. de novo Spike Cold Start Results. 145

Figure 5-5. de novo Spike Cluster Visualization. 146

Figure 5-6. de novo Spike Initial Sequence. 147

Figure 5-7. Sequence Extraction for de novo Spike. 149

Figure 5-8. de novo Spike Dose Response Plots. 150

Figure 5-9. Final de novo Spike Oligonucleotides. 151

Figure 5-10. de novo ACE2 Cold Start Results. 152

Figure 5-11. de novo ACE2 Cluster Visualization. 153

Figure 5-12. de novo ACE2 Initial Sequence. 154

Figure 5-13. Sequence Extraction for de novo ACE2. 156

Figure 5-14. de novo ACE2 Extracted Sequences for Validation. 157

Figure 5-15. de novo ACE2 Dose Response Plots..... 158

Figure 5-16. de novo ACE2 With Control Aptamers. 159

Figure 5-17. Final de novo ACE2 Aptamer. 160

Figure 5-18. de novo HA Cold Start Results. 161

Figure 5-19. de novo HA Cluster Visualization. 162

Figure 5-20. de novo HA Initial Sequence. 163

Figure 5-21. Sequence Extraction for de novo HA. 165

Figure 5-22. de novo HA Dose Response Plots. 166

Figure 5-23. AlphaFold PD1 Predicted Structure. 167

Figure 5-24. Truncation of PD1 Structure. 168

Figure 5-25. de novo PD1 Cold Start Results. 169

Figure 5-26. de novo PD1 Cluster Visualization. 170

Figure 5-27. de novo PD1 Initial Sequence. 171

Figure 5-28. de novo PD1 Affinity Validation. 173

Figure 5-29. Sequence Extraction for de novo Spike with ACE2 Specificity. 175

Figure 5-30. de novo Spike with ACE2 Specificity Dose Response Plots. 176

LIST OF TERMINOLOGY

Affinity	The binding of an aptamer to its target molecule. A high affinity aptamer has a K_d-value between 1 – 99 nM.
Alpha Shape	Creates a bounding volume along the surface of the molecule of interest.
Aptamer	A single-strand oligonucleotide, either ssDNA or RNA, with both high affinity and specificity
Biosensor	A device that generates a measurable signal proportional to the concentration of the target molecule
Cold Start	The *in silico de novo* design process of a single-stranded oligonucleotide using the target molecule as a guide assigning nucleotides
Delaunay Triangulation	A mathematical and computational geometry method for identifying relationships between points so that given a set of points there are no points in the set inside the circumcircle of any triangle. This relationship allows connections between points to be determined.
De Novo	Latin for starting from the beginning, from new
En Toto	Latin for completely or wholly
In Silico	Latin for conducted or produced by means of computer modeling or computer simulation
In Vitro	Performed or taking place in a test tube, a culture dish, or outside of a living organism
K-Means Clustering	A data partitioning algorithm that assigns n observations to exactly one of k clusters defined by centroids and minimization of distance between the observation and the centroid.
Measurement Problem	Identifying an *in vitro* method for measuring both affinity and specificity that accurately mimics application conditions.

Molecular Docking	A method that examines the potential interaction between two molecules based on their orientation and proximity.
Molecular Dynamics	A simulation method that analyzes the physical movements of atoms and molecules
Molecular Simulators	A category of computational tools employed to study the interactions between molecules and predict structures.
Probabilistic Application	An algorithm that maximizes the probability of interaction between oligonucleotide and the target molecule while minimizing the interaction with the counter-selection target
SELEX	The *in vitro* process identifying single-stranded oligonucleotides which is also known as the Systematic Evolution Ligands by EXponential enrichment.
Solid Angle	A measure of the amount of the field of view from some point that a given object covers. This measure is used to determine accessibility such that an atom which is surrounded has an angle of 4π.
Specificity	The preferential binding of an aptamer to its target over a counter-target molecule. The affinity for the target is at least four times better than the counter-target.
Warm Start	The *in silico* counter-selection module where the initial designed oligonucleotide undergoes perturbations and is evaluated against both the target molecule and a counter-target molecule

LIST OF ABBREVIATIONS

ACE2	ANGIOTENSIN CONVERTING ENZYME 2
AMBER	ASSISTED MODEL BUILDING WITH ENERGY REQUIREMENT
BSA	BOVINE SERUM ALBUMIN
CPU	CENTRAL PROCESSING UNIT
CRP	C-REACTIVE PROTEIN
DNA	DEOXYRIBONUCELID ACID
GA	GENETIC OPTIMIZATION ALGORITHM
GFET	GRAPHENE FIELD EFFECT TRANSISTOR
GROMACS	GRONINGEN MACHINE FOR CHEMICAL SIMULATIONS
HA	HEMAGGLUTININ PROTIEN
KA	AFFINITY CONSTANT
KD	DISSOCIATION CONSTANT
ML	MACHINE LEARNING
NLP	NATURAL LANGAUGE PROGRAMMING
NM	NANOMOLAR
PCR	POLYMERASE CHAIN REACTION
PD1	PROTEIN DEATH 1 RECEPTOR
PDB	PROTEIN DATA BANK
RBD	RECEPTOR BINDING DOMAIN
RNA	RIBONUCLEIC ACID
RSV	RESPIRATORY SYNCYTICAL VIRUS
SELEX	SYSTEMATIC EVOLUTION OF LIGANDS BY EXPONENTIAL ENRICHMENTS
SSDNA	SINGLE-STRANDED DEOXYRIBONUCELID ACID

ACKNOWLEDGEMENTS

The opportunity to pursue a doctoral degree and this research would not have been possible without the support of several individuals and entities. I do not take this opportunity lightly especially since I began the program with not only a condensed timeline due to needs of the U.S. Army, but also during a global pandemic.

First, I need to recognize the United States Department of Defense, the United States Army specifically the Advanced Civil Schooling (ACS) program which has enabled me to previously earn a Masters degree and now a PhD. The United States Military Academy at West Point for the opportunity to return as faculty to the Department of Chemistry and Life Science. I also need to thank Colonel F. John Burpo for our discussions and his guidance over the years.

I need to thank the Boston College Department of Biology for enabling this pursuit. I also specifically thank Dr. Welkin Johnson for his guidance and Ms. Dina Goodfriend for her expert assistance with navigating the University and the Army paperwork requirements.

The Boston College core facilities have also been outstanding. I appreciate the hard work and efforts of Bret Judson in the Imaging core who assisted me in early validation steps as well as Dr. Patrick Autissier in the Flow Cytometry core. I appreciate the time and effort it took to run the validation samples particularly when I would drop off 75 tubes at a time. Also, thank you Michael Geiwitz for our discussion G-FET devices and applications.

My committee has been especially helpful over the past three years. The ACS is a great opportunity but requires a condensed timeline. The guidance from Drs. Jose Bento (Computer Science), Kenneth Burch (Physics), Jianmin Gao (Chemistry), Michelle Meyer (Biology), and Tim van Opijnen (Biology) has been incalculable.

I have special appreciation for Drs. van Opijnen and Bento. First to Dr. Opijnen, I appreciate the opportunity to join your lab and to deviate from the original plan. And to Dr. Bento, I appreciate our long conversations, zoom calls, and coding sessions and we worked to mix MATLAB and Python scripts with biological operators. I think our work on digitalSELEX has disrupted the established SELEX paradigm.

And finally, to Sam, Hannah, and of course, Krissy, it has been a long and simultaneously short three years. Disrupting your lives and moving during a pandemic is never easy and you all rose to the challenge. You gave me the opportunity and time to succeed, and my success is a direct reflection of your efforts. Thank you

DISCLAIMER

The work presented in this document does not reflect any endorsement, policies, or perspectives of the United States Government, Department of Defense, the United States Army, and the United States Military Academy.

1.0 CHAPTER 1 – INTRODUCTION

This chapter establishes the foundation for understanding aptamers, their general application, and the current selection process along with associated challenges. The subsequent chapters layout a novel *in silico* aptamer design platform beginning with the *Measurement Problem*, the *Cold Start* design module, and the *Warm Start* module. The *de novo* design platform is then validated across several unique challenges in terms of affinity and specificity.

1.1 THE PROBLEM

Molecules that are both high affinity and specific for a given target have tremendous potential as either a therapeutic or as a biosensor component (*e.g.*, probe). These molecules can range from chemical compounds (*e.g.*, drugs), antibodies, and aptamers.

Chemical compounds are small (~10-40 atoms) and easy to synthesize.[1] The cost of discovery, however, particularly with respect to therapeutics can be extremely high. One published report showed the investment for developing new chemical compounds ranging from $314 million to $2.8 billion.[2] Antibodies consistently demonstrate high affinity and specificity, however, they take months to generate and purify.[3] These

molecules are also subject to batch-to-batch variations due to the process by which antibodies are generated.[4] Aptamers, or single-strand oligonucleotides, have demonstrated both high specificity and affinity, are approximately one-tenth the size of an antibody (~25,000 atoms)[5], and are not subject to batch-to-batch variations since these can be chemically synthesized.[6-7] The *in vitro* identification process for aptamers, however, presents multiple challenges.

The elemental challenge for SELEX regardless of being *in vitro* or a combination *in vitro / in silico* methods is ensuring the aptamers are both high affinity and specific. The different SELEX variations to one degree or another address specific concerns such as initial pool size, time, cost, low success rate, and data dependence but no method mitigates them all.[8-9] To overcome these inherent challenges, we propose a novel *in silico* platform that is capable of *de novo* design. This platform has the capacity to also improve both the affinity and specificity of existing single-stranded oligonucleotides. Since there are a variety of applications for high affinity and specific molecules, the initial design and validation process focused on biosensors, for example, **Figure 1 – 1** illustrates a graphene-based biosensor that employs aptamers.

1.2 APTAMERS

The word aptamer is from the Latin words *aptus* meaning to fit and *meros* meaning part.[10] Aptamers are short, single-stranded oligonucleotides, that are intended to bind to a specific target.[11] The oligonucleotides are composed of either single-stranded deoxyribonucleic acids (ssDNA) or ribonucleic acids (RNA). The concept of using

aptamers as affinity molecules for various target compounds was first proposed in 1990.[12-13] Similar to proteins, the unique sequence of the single-strand oligonucleotide dictates is structure, and the structure dictates its capacity to bind.[14]

Aptamers are analogous to antibodies for their target recognition and range of applications. Typically, these single-stranded nucleotide polymers are one-tenth the molecular weight of antibodies. Aptamers can be chemically synthesized and are not subject to batch variations like antibodies, which require host animals.[15] Chemical synthesis enables easy modification to include biotinylation, fluorophore attachment, phosphorylation, and others which enable aptamers to be employed in a variety of biosensors. Long term storage of aptamers is also possible at the correct temperature and pH allowing them to be design, synthesized, and then employed as need in a variety of biosensors.[16-17]

1.3 OLIGONUCLEOTIDE SELECTION

The current method for identifying high affinity and specificity aptamers is through either *in vitro* selection or a hybrid *in vitro / in silico* selection process.

1.3.1 In Vitro SELEX

The "gold standard" for identifying aptamers with both high affinity and specificity towards their target protein is achieved through a process known as a SELEX.[18-19] This *in vitro* selection process is illustrated in **Figure 1 – 2**.[20] There are

multiple SELEX variations[21,22,23,24], but the process generally consists of an initial library containing approximately 10^{15} randomized single-stand oligonucleotide molecules undergoing several selection rounds to identify high affinity and specific sequences.

The libraries are designed with specific primers of approximately 15-20 nucleotides on the 5'- and 3'-end of the oligonucleotide sequence. The remaining nucleotides in each sequence are randomly selected. This randomization creates aptamers with unique structures. Chemical synthesis makes the randomization process easy. The exact number of random nucleotides varies depending on the size of the primers and desired length of the oligonucleotide. For example, aptamer V46 for detection of H1N1 variants is 40 nucleotides in length while gD-HSV-1 aptamer for HSV is 114 nucleotides in length.[25]

Selection occurs by exposing the initial library sequential to the target and counter-target molecules. In the first selection step, the initial library is incubated typically for an hour or less with the target molecule which can be a specific protein or cell (*e.g.*, E. *coli*). The supernatant, or unbound aptamer population in solution, is then discarded. The bound aptamer is separated from the target molecule via thermal or chemical (*e.g.*, alter pH) dissociation. The separated aptamers are resuspended prior to exposure to a non-target cell or protein in a negative, or counter-selection step. Depending on the desired application, the specific details regarding the incubation of the aptamer library with target and non-target molecules or cells can vary.

The aptamers that bind to the non-target molecules are discarded. While the bound oligonucleotides are discarded, the supernatant containing the remaining population of aptamers is retained. The retained aptamers are then amplified in a

polymerase chain reaction (PCR) using complementary primers to the library specific primers. The amplification process does not alter the remaining oligonucleotides, rather increases the number of molecules for the next round. The PCR product is purified to remove enzymes, incomplete sequences, and unincorporated nucleotides. The purified library is utilized in the next round of selection and counter-selection. The SELEX process usually consists of 10-15 iterative rounds of positive and negative selection. Following the final amplification step, the oligonucleotides are sequenced.

The classic SELEX concept has been advanced and modified since 1990 to incorporate new technologies. Some of these advancements include capillary-electrophoresis (CE) – SELEX[26], micro free flow electrophoresis (μFFE) – SELEX,[27] and capture – SELEX[28] to reduce time, cost, and improve the success rate of the selection process.

The SELEX method can produce high affinity aptamers (dissociation constants in the low nanomolar range), the repetitive nature and stringent counter-selection steps however make the process both time consuming and expensive. The *in vitro* selection makes successful use of selected aptamers only likely in applications whose conditions directly mimics those of the selection process. These conditions not only include temperature, pH, and ion concentration, but also the number of counter-target molecules.

For example, an aptamer selected for the receptor binding domain (RBD) of the SARS-CoV-2 virus may have high affinity, but the specificity is limited to the counter-selection molecules present in the SELEX process. While it is reasonable to use the ACE2 receptor molecule as the counter-selection molecule in this SELEX process, other molecules could bind with the aptamer. In a biosensor, interaction with the target may be

limited or generate a false positive due to the aptamer binding with other proteins such as Influenza Hemagglutinin (HA), Respiratory Syncytial Virus (RSV) Glycoprotein, or C-Reactive Protein (CRP) that are potentially present in the patient sample.

It is feasible to expose an aptamer library to multiple counter-selection molecules, there are however trade-offs with respect to time, cost, and the initial library. The randomized aptamer library could easily be exhausted since specificity is not absolute. This means that the oligonucleotides in the initial library bind to the multiple counter-targets prior to their affinity being determined.

1.3.2 In Silico Selection

Parallel to the incorporation of new technologies and techniques into in vitro SELEX, numerous *in silico* methods were investigated to also improve aptamer-binding affinity and identify structural patterns for the aptamer-protein interactions. These methods can be viewed as hybrid SELEX which are developed incorporating both *in silico* methods with *in vitro* SELEX. These hybrid SELEX methods generally fall into one or both of two general categories: machine learning and molecular simulations.

1.3.2.1 Machine Learning

Natural Language Processing (NLP), Deep Neural Networks, and Variational Autoencoder are just a few examples of algorithms that have been employed to predict potential interactions between oligonucleotides and sequences as well as identify key structural motifs on the target molecules.[29,30,31,32]

A critical component of all machine learning algorithms is the data required to train the model and several databases, *Riboapt DB*,[33] *RPINBASE*,[34] Apta-Index,[35] PDBBind,[36] and aptamer free base,[37] have been established for that purpose. These datasets however are limited by quality of experiments that examine known aptamer-target interactions. There is a dearth of information in these datasets regarding specific nucleotide-amino acid interactions derived from structural experiments. For example, PDBBind data set lists the protein and the sequence of the oligonucleotide but provides no additional information about the where the oligonucleotide is interacting with the protein.[38]

Additionally, many entries in the datasets also involve proteins that specifically interact with DNA or RNA (*e.g.,* polymerases, ligases, and so forth). These entries introduce a bias to the dataset towards specific structures and domains on both the protein and oligonucleotide that may not be useful for the application of the aptamer. Within these datasets there is also little information regarding the experimental conditions for the aptamer selection such as pH, counter-selection targets, incubation time, and buffer ion concentrations.

As previously mentioned, ML algorithms are only as good as the data used to generate the model. Consequently, identification of an oligonucleotide sequence using a ML algorithm may fail during *in vitro* testing as well as in the desired application.

1.3.2.2 Molecular Simulations

There are a variety of molecular simulators that can be used to study different molecular characteristics such as structure (*e.g.*, protein and nucleic acid folding) and interaction via molecular docking or molecular dynamics.

For determining secondary structure, studies either employed viennafold, rnafold (MATLAB version of the server based viennafold), or unafold.[39] To generate the tertiary or three-dimensional structure, a common application used is RNAComposer.[40]

There are a variety of molecular docking programs available for oligonucleotide-protein interactions to include AutoDock Vina, Rosetta, Lighdock, HDOCK, and ZDOCK.[41] These software applications vary significantly in their input files requirements. This variation includes protein data bank (pdb) file preprocessing requirements to remove or add hydrogen atoms, naming convention, and file structure (*e.g.,* convert pdb to mol).

Molecular dynamic software, such as AMBER (Assisted Model Building with Energy Requirements) and GROMACS (GROningen MAchine for Chemical Simulations), have also been employed to examine the proposed temporal electrostatic interactions between the aptamer and protein as determined by the molecular docking programs.[42-43]

1.3.3 Hybrid Selection Strategies

Hybrid selection incorporates *in silico* methods, such as machine learning algorithms or molecular simulations, with *in vitro* SELEX methods. This combination has led to the development of two generalized strategies.

The first strategy uses sequencing data generated through *in vitro* SELEX and applies machine learning algorithms to identify small clusters within the sequence population. Both SMART-Aptamer and RaptRanker are two such algorithms.[44-45] These identified clusters or fragments (~3-5 nucleotides) are then stitched together in random

order to generate a full-length sequence.[46] These sequences are then docked against the target molecule and the sequences with the best scores are selected for validation.[47]

The second strategy builds the aptamer *in toto*.[48] In this method, a single nucleotide is initially docked with the target and grown a single nucleotide at a time. This process selects the best docked nucleotide at each step and is repeated until the desired oligonucleotide length is achieved.[49] The sequence is converted into a three-dimensional structure and docked after the addition of each new nucleotide. The relationship between the structure of the oligonucleotide and the sequence means that the docking of an elongated aptamer may not improve the score in a linear or monotonic manner.

While the various SELEX methods and *in vitro / in silico* combinations are designed to find structural patterns, provide critical information regarding nucleotide-amino acid interactions, improve success rate, and reduce the overall SELEX time, these methods in general have several challenges.[50,51,52,53,54,55]

1.4 THE MEASUREMENT PROBLEM

The critical outcome of any selection method is the identification of high affinity and specific aptamers. There are a variety of quantitative *in vitro* measurement methods to determine the affinity and specificity of an oligonucleotide sequence such as flow cytometry, confocal microscopy, surface plasmon resonance, and so forth. These methods vary in precision and accuracy. The measurement method should also reflect the desired use of the aptamer to include configuration and environmental conditions. The measurement problem is explored in more detail in Chapter 2.

1.5 CHALLENGES WITH EXISTING SELECTION METHODS

There are four critical challenges that existing selection methods, whether *in vitro* or *in silico*, must overcome. The most basic and inherent challenge of all oligonucleotide selection methods is the identification of high affinity and specific sequences. Coupled to the inherent challenge of affinity and specificity are the issues with success rate, cost, and time.[56-57] The various SELEX and selection methods developed over the years have attempted to address these challenges, but these issues remain in an interconnect equilibrium.

1.5.1 Challenges: Affinity and Specificity

The amino acids on the target that interact with the aptamer nucleotides are also the same amino acids, in terms of hydrophilicity and biologically relevant atoms, on the non-target molecule. Consequently, specificity and affinity are different sides of the same coin. This poses a challenge with respect to distinguishing between target and non-target binding.

Imperative in all selection methods is the identification of high affinity and specific oligonucleotides which is directly related to the types of interactions between the nucleotides and the amino acids.[58] The oligonucleotides sequence dictates its three-dimensional structure which in turn directs how the nucleotides interact with either the target or counter-target amino acids. This challenge of affinity and specificity along with the various of interactions are explored in more detail in Chapter 2.

1.5.2 Challenges: Success Rate

There are a variety of selection methods used to identify aptamers, however the rate of success remains relatively low, at less than 30%. This value is determined by saying that only three or less SELEX runs out of ten will generate aptamers with the desired specification.[59] There are two definitions of success. One is success with respect to the selection process itself while the other definition focuses on the use of the oligonucleotides.[60] For the selection process, success directly relates to identifying aptamers with both high affinity and specificity, while the other definition reflects the aptamers being able to function (*e.g.*, biosensor probe or therapeutic) in the environment and conditions. There are several factors that contribute to this limited success rate to include initial oligonucleotide library size, selection of targets / counter-targets, *in vitro* conditions, and *in silico* database information. The effect of the initial library size contributing to this challenge is illustrated in **Figure 1 – 3**.

1.5.3 Challenges: Cost

While a single SELEX run has limited costs, the low success rate of SELEX suggests that multiple SELEX trials are required. Chemical synthesis of a random library of approximately 60-nucleotides costs about $100. The exact cost of the target and counter-target molecules can vary greatly depending on if the process uses proteins fixed to a substrate (*e.g.*, magnetic beads) or live cell (*e.g.*, bacteria). If the fix protein is fixed to a substrate, the target and counter-target proteins typically cost about $1,000 each

target and counter-target per SELEX. The amplification process requires standard PCR reagents and library specific primers which will total about $500 per SELEX.

A sequences kit from Illumina costs approximately $1,000 per sequencing run; generating a list of 100 – 1000 oligonucleotide sequences. *In silico* validation, such as molecular docking and structural analysis may help reduce the potential number of sequences that require *in vitro* testing.

The *in vitro* validation of the SELEX generated sequences however becomes the bulk of the cost. Chemical synthesis of an oligonucleotide costs about $1.25 per nucleotide so the synthesis of a single 60-nucleotide aptamer costs $75. This cost estimate of aptamer synthesis does not include any desired modifications such as biotinylation or fluorophore tags. Subsequently, synthesis of 100 60-nucleotide aptamers would cost at least $7,500 while 1000 aptamers would be $75,000. The total a single SELEX run with a single target and counter-target, as described above, is $11,100 ($100 + $1,000 (target) + $1,000 (counter-target) + $500 + $1000 + $7,500).

With the historic success rate of approximately 30% or three out ten, a moderate planning factor of four full SELEX runs is used to estimate the cost of identifying high affinity and specific aptamers. At a cost of $11,100 per run, the total estimated cost is at least $44,400. This estimate does not include other reagents such as PBS, tubes, pipette tips, or man-hour salary costs.

1.5.4 Challenges: Time

The selection process can range from approximately 2 – 8 weeks.[61] While it is possible for a single SELEX run to identify a high affinity and specific oligonucleotide, it

is probable that the process needs to be repeated multiple times. At a general success rate of 30%, in the worst-case scenario it could be assumed that seven out of ten SELEX runs fail; meaning only selection runs eight, nine, and ten would be successful. Subsequently, a planning factor of eight SELEX runs is utilized for estimating the maximum time required. The time range could be as little as two weeks where a single SELEX run generates the desired aptamers. The worst possible scenario the time required could be as long as 64 weeks (8 * 8 weeks) (16 months). Comparatively, antibody generation typically takes approximately 6 months.

The time range of 2 – 64 weeks is a simple planning factor, but it does indicate time required for an individual or group to be either directly involved in the work or supervising robotic machines.

1.6 PROPOSED METHOD

The challenges existing selection methods are not insignificant. SELEX variants (*e.g.,* CRISPR-mediated SELEX) to include hybrid methods (*e.g.*, SMART-Aptamer, APTIANI) do not break the selection paradigm established over 30 years ago. While advancements in both SELEX and *in silico* methods mitigate some challenges; these challenges persist. This selection paradigm relies on the precept that one or more high affinity and specific oligonucleotides resides in a randomized library.

Our proposed digitalSELEX platform, **Figure 1 – 4,** disrupts the existing selection paradigm. Instead of employing randomized oligonucleotide populations, our method focuses on the target molecule and the opportunities for interaction. The

digitalSELEX platform can be viewed as two functional modules (*Cold Start* and *Warm Start*) plus validation.

The *Cold Start* module is the *de novo* design process. This module identifies clusters of accessible, binding-relevant atoms based on the three-dimensional data presented in a protein databank file. Nucleotides are then assigned to the corresponding amino acids to generate a core sequence. This core sequence is then optimized for stability and application using a genetic optimization algorithm.

The user defined constraints can include but are not limited to structural features such unpaired nucleotides, cytosine-guanine content, and structure (*e.g.*, loops). The optimization algorithm continues until the number of generations, either total or stall (number of generations with no change), has been achieved or the desired stability score for the oligonucleotide has been achieved. For the stability score, the free energy of folding a sequence. The more negative the value then the more stable the corresponding structure for the sequence. This optimization algorithm ensures the output sequence is stable. [62,63]

The nucleotide sequence is then converted to a three-dimensional structure in the sequence to structure module. This module uses a combination of MATLAB function, rnafold, and the webserver, RNAComposer.[64] A Python script is used to automate this function.

The *Warm Start* module improves the interaction between an oligonucleotide and target protein using a molecular docker as a guide. A series of random mutations are added to the initial sequence some of which might improve the *in silico* interaction. Each oligonucleotide is then docked against one or many targets and/or counter-targets, while

examining if the sequence perturbations diminish or augment the interaction between the target and/or counter-target. This method is also capable of handling multiple targets and counter-targets simultaneously. These perturbations and tests are performed repetitively as part of a stochastic optimization simulation driven loop.

The sequences demonstrating improved interaction *in silico* are identified from the *Warm Start* module for validation to determine affinity and specificity. The number of sequences requiring validation is less than ten.

There are three critical components: *Cold Start* module, *Warm Start* module, and the validation.

1.7 ROAD MAP

The subsequent chapters focus on detailing the two novel processes, *Cold Start* and *Warm Start,* of the digitalSELEX platform.

To frame these modules and how their processes address the challenges of existing SELEX methods, Chapter 2 details the *Measurement Problem.* Chapters 3 and 4 detail the *Cold Start* and *Warm Start* module, respectively. Chapter 5 highlights the validation of the digitalSELEX platform and data demonstrating how the platform overcomes the current challenges. Finally, Chapter 6 provides a brief discussion of conclusions and future applications.

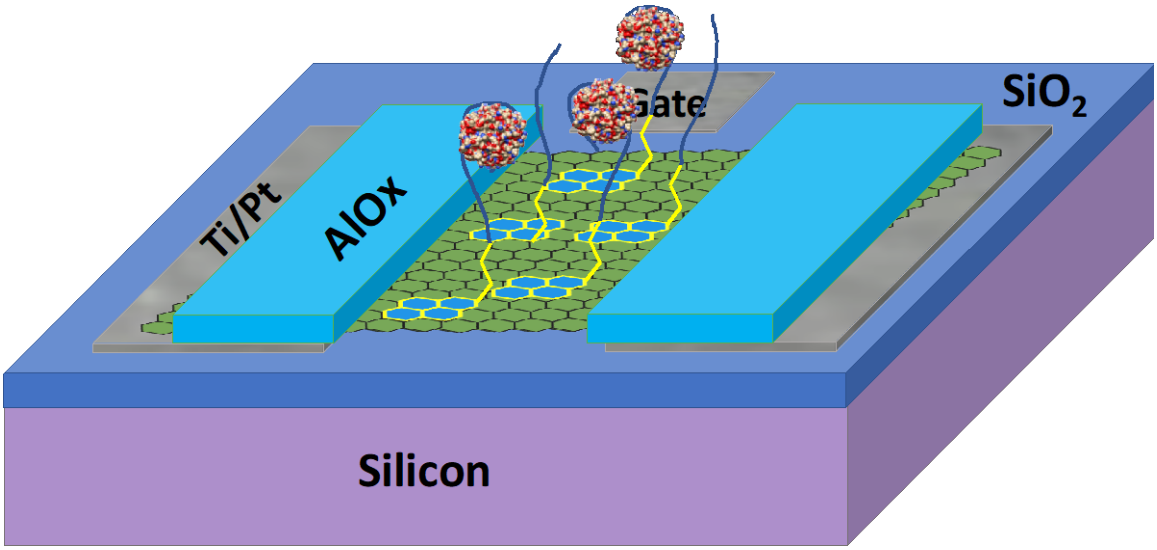


Figure 1-1. *Basic G-FET.*
A representative illustration of a Graphene Field Effect Transistor (GFET) device. The single layer of carbon or graphene (green) is layered on a silicon oxide chip. The aptamer probes are drawn in dark blue and attached to the graphene using the PBASE linker (yellow). A voltage is applied to the device and the resistance across the graphene is measured. When a target molecule attaches to the aptamer a change in resistance occurs. The *pyrene moiety* on the PBASE linker creates π–π stacking with the graphene for the linker-aptamer complex to remain attached to the device.

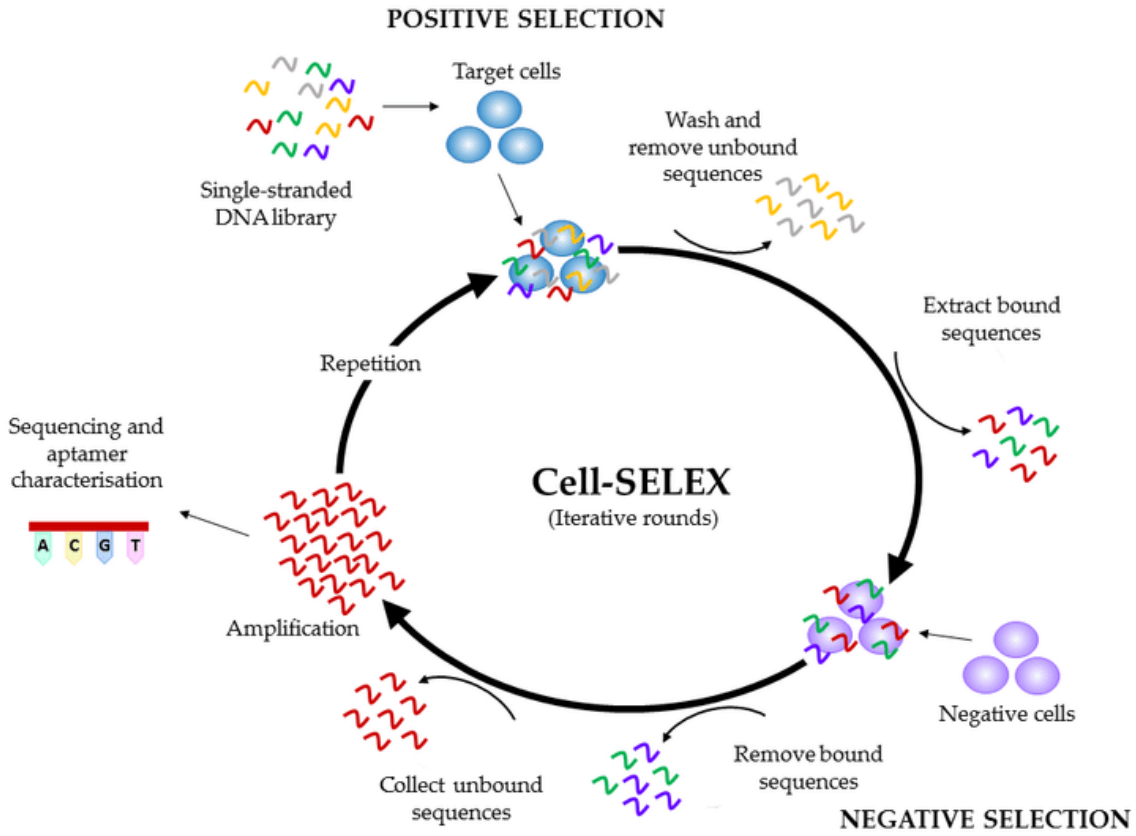


Figure 1-2. General SELEX Method.
A representative illustration of Systematic Evolution of Ligands by EXponential enrichment. An initial library of oligonucleotides (~10^{15} unique molecules) is exposed to the target molecule either cell or protein. The unbound nucleotides are removed from the system. The sequences bound to the target are isolated from the target before being exposed to a non-target molecule. The molecules that do not bind to the non-target molecule are retained and then amplified in a PCR reaction. The process typically repeats for 10-15 rounds before the oligonucleotides are sequenced. Typically, there are 100 to 1000 sequences at the end of the SELEX rounds to be further tested and evaluated to ensure the oligonucleotide has the desired specificity and affinity in the application conditions. Figure is adopted from Hays *et al*.[65]

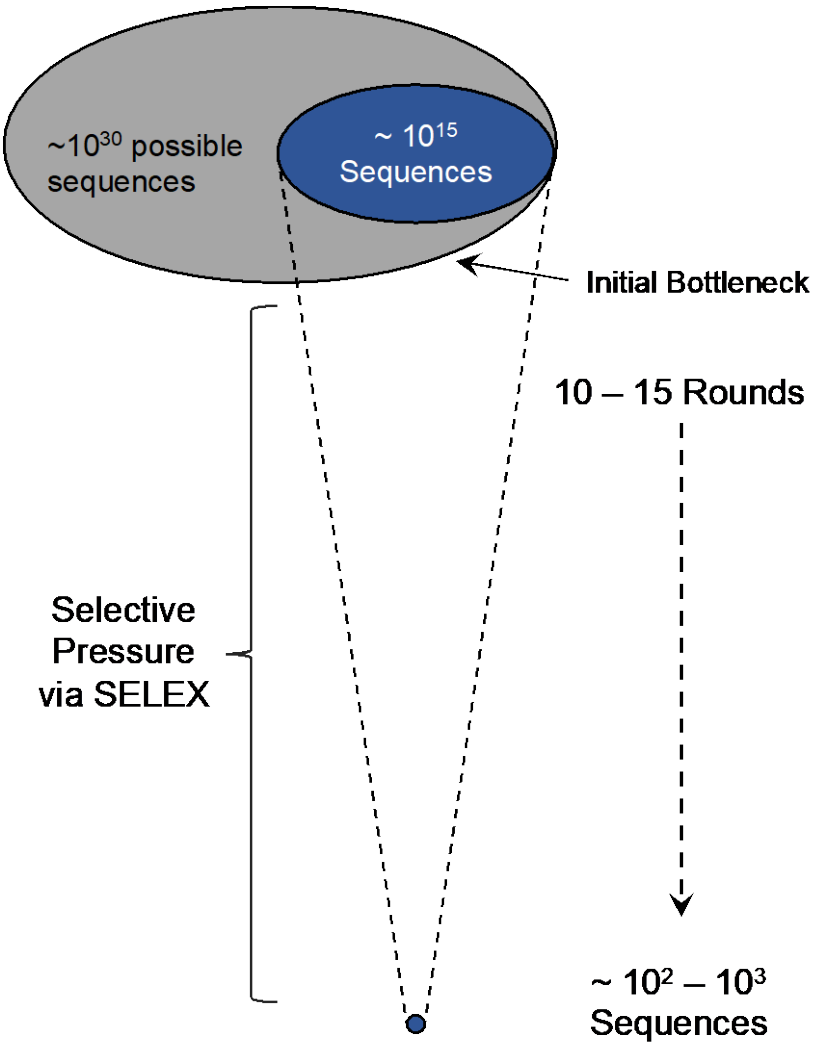


Figure 1-3. *Generic Oligonucleotide Library*
Illustrated the diminishing initial oligonucleotide pool during SELEX. For a 50-nucleotide aptamer, there are 10^{30} possible unique molecules, however only an initial library of 10^{15} molecules are used: shown as the initial bottleneck. Selective pressure between positive and negative selection further shrinks the pool over 10 – 15 successive rounds. At the end of each round an amplification step occurs to increase the number of molecules but this does not change the type of unique molecules in the pool.

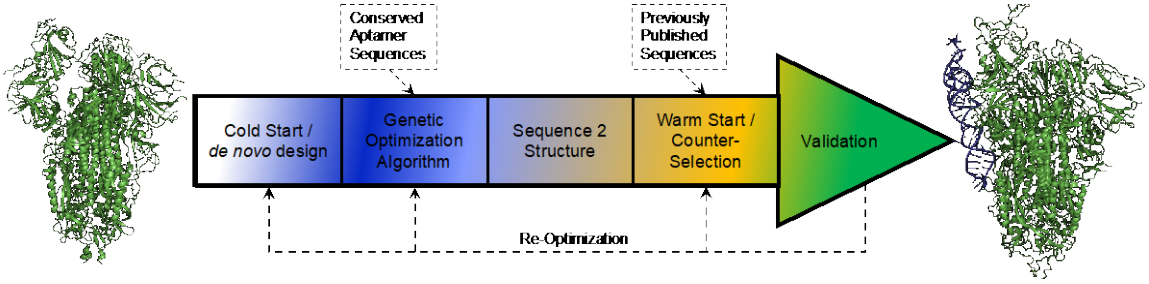


Figure 1-4. Proposed digitalSELEX Model.
Illustration of the *in silico* digitalSELEX platform. The starting point is a target molecule structure that is dissected and analyzed in the *Cold Start* module. The output of this module is an oligonucleotide sequence that is optimized towards the application constraints and for stability with the Genetic Optimization Algorithm. The optimized oligonucleotide sequence is then converted into a three-dimensional structure. This structure can then be imported into the *Warm Start* module that simulates its interactions with the target and counter-target molecules. This module is designed to introduce mutations into the sequence to improve the potential interaction between the oligonucleotide and the target while degrading the interaction between the aptamer and counter-target. The finalized oligonucleotide is then validated *in vitro* using conditions that mimic the application.

2.0 CHAPTER 2 – THE MEASUREMENT PROBLEM

This chapter defines the notion of "good aptamers" as those with high affinity and specificity. It details a procedure for determining the affinity and specificity of a given set of oligonucleotides with a given set of targets and to determine if an aptamer is good or bad.

2.1 DEFINING GOOD AND BAD APTAMERS

The simplest definition of a good oligonucleotide is that it demonstrates both high affinity and specificity for its targets. A bad aptamer lacks either affinity or specificity for its target. We define high affinity as the target-aptamer pair as having a dissociation constant (K_d) value in the low nanomolar range. The dissociation constant is a measure of affinity where half of the target molecules have bound aptamers. The K_d -value is related to the affinity constant (K_a) where the K_a is equal to one over the K_d.

Specificity is defined as a meaningful difference between the K_d -value of the aptamer-target pair and the K_d -value of the aptamer-counter-target pair. The difference can vary depending on the desired application of the aptamer. For biosensor, the affinity of the aptamer for the target must be equal to or better than four times the K_d-value for the aptamer-counter-target pair. There are two factors that contribute to affinity and specificity. These factors are stability aptamer-target binding and the type of interaction between the aptamer and target.

2.1.1 Oligonucleotide Structure Stability

A good aptamer is stable. Its three-dimensional structure does not fluctuate much. The oligonucleotide structure needs to be stable because structural uncertainty leads to functional uncertainty and the aptamer-target binding needs to last of over time. The chemical synthesis of oligonucleotides enables any combination of nucleotides, however not all combinations provide a stable three-dimensional structure. Environmental factors, such as temperature, pH, and the presence of ions, can have a stabilizing or destabilizing effect on the structure. Since these influencing conditions are derived from the application of the oligonucleotide, the stability calculation should incorporate these factors in the most feasible manner. Consequently, a good aptamer is one that is stable in the application environment.

2.1.2 Oligonucleotide-Target Binding

Oligonucleotide-target binding occurs through non-covalent intermolecular physical interactions. These interactions are electrostatic forces (point charges and dipoles), van der Waals, and hydrogen bonding.

A point charge is the most fundamental non-covalent interaction. This interaction is guided by Coulomb's law where the energy between two atoms in a vacuum is simply proportional to their two charges divided by the distance between them.[66] The interactions between proteins and nucleic acids however do not occur in a vacuum.

Coulomb's law is subsequently modified to incorporate the dielectric constant of the media / solution.

A dipole is another type of electrostatic force, shown in **Figure 2 – 1**, and a molecule does not require a net charge. Instead, the electron density of a molecule can be localized if the atoms of the molecule have varying *electronegativity*, which is the tendency of an atom to attract electrons based on atomic number and distance to valence electrons. Atoms with the largest electronegativity have an excess of negative charge, while others have an excess positive charge.[67] The separation of the charge in a molecule creates a *dipole moment* which is proportional to the distance and magnitude of charge between positive and negative charges. Dipoles interact with point charges and other dipoles. The complexity of the interaction is determined by the relative orientation of the atoms.[68]

The ubiquitous interaction between all molecules regardless of the presence of charge are known as van der Waals interactions. These interactions can occur between two molecules with permanent dipoles, one molecule with a permanent dipole with a molecule with an induced dipole, or two molecules with mutually induced dipoles.[69]

While the first two van der Waals scenarios are easy to understand due to the presence of a permanent dipole, the mutually induced dipole is more complex. If orbiting electrons are perfectly spherical in their trajectory, then there is no dipole. A temporary asymmetry in the movement of electrons around the nucleus can induce a dipole which can polarize a neighboring neutral atom, creating an attraction between them.[70] The energy potential of van der Waals interactions is frequently visualized by a Lennard-

Jones plot which relates the interaction energy between two atoms with the distance between the centers.

Hydrogen bonding is a special type of dipole where two electronegative atoms are attracted to the same hydrogen atom. The hydrogen is covalently bonded, shares an electron pair, to one of the atoms, which is known as the donor, and interacts with another electronegative atom, known as the acceptor. The donor creates a dipole where it has a partial negative charge, and the hydrogen has a partial positive charged. The positively charged hydrogen atom will then interact with the electronegative acceptor atom.

Hydrogen bonds are formed when a hydrogen atom is bound to a fluorine, oxygen, or nitrogen. These are the three most electronegative atoms. The bond length depends on the electronegative of the donor and acceptor. The shorter the bond length then the greater the electronegativity.[71]

The intermolecular forces of electrostatic interactions, van der Waals interactions, and hydrogen bonding are critical for interactions between proteins and nucleic acids regardless of the presence or absence of a specific binding domain. It has been noted that per complex van der Waals interactions comprise near 75% of all protein-DNA interactions, while the remaining 25% are comprise of the electrostatic forces (~10%) and hydrogen bonding (~15%).[72] While hydrogen bonding only comprises about 15% of all protein-nucleic acid interactions, the strength of their interactions contributes greatly to the overall energy of interaction.[73] It is estimated that the energy contribution from hydrogen bonding is approximately 5 – 6 kcal per mole[74] whereas the energy contribution from van der Waals interaction is approximately 0.5 – 1 kcal per mole.[75] The

significant energy contribution from hydrogen bonds directs identification of binding relevant atoms in the *de novo* design process highlighted in Chapter 3.

The strength of the interaction between the aptamer-target pairing is a function of the non-covalent interactions, however, there are molecular forces that can break these interactions. These forces can be visualized through the movement of atoms in a larger structure. There are three forms of molecular movements: vibrations, rotations, and translation.

As noted by Glaser in Biophysics, "vibrations are oscillations in the binding distances between the atoms in a molecule. The term rotation means not only the rotation of the whole molecule, but additionally, the spin of the individual atoms or atomic groups around the axis of their bonds. The full translocation of a molecule or a part of it in space is meant by translation."[76] The frequency of these movements differs with rotation occurring at $10^{10} – 10^{12}$ Hz (s^{-1}), while the frequency of vibration is higher at approximately 10^{14} Hz, depicted in **Figure 2 – 2**. The quantum energies of both rotation and vibration can be estimated and are near the thermal energy of the Boltzmann constant times temperature, kT. In this equation, room temperature of 300 K corresponds to thermal energy of 2.6×10^{-2} eV. By contrast, translation energy is much smaller at 10^{-16} eV at room temperature (22-25 C or 295-298 K). This thermal energy means that molecular vibrations and rotations contribute greatly to the forces that break the non-covalent interactions between the aptamer-target pair.[77]

Tools have been developed to study the molecular movement of atoms in binding interactions. This is important to study pair stability due to temporal fluctuations of shapes and distances. These tools are broadly referred to as molecular dynamic

simulators. Molecular dynamics attempt to predict how the atoms vibrate and move over time to understand the dynamics of interatomic interactions.[78] Even though these simulation tools can provide high resolution insight into electrostatic interactions (*e.g.*, hydrogen bonding) between molecules during a temporal period, these processes are computationally expensive. Initial testing demonstrated a molecular dynamic simulation between an aptamer and target molecule took approximately 120 hours for GROMACS to run a 10-nanosecond temporal window on a 32-cluster system. Additionally, there specific environment definitions constraining the interactions that must be identified for running the simulator. Molecular dynamic tools are not employed in the digitalSELEX prototype and is addressed in further detail in Chapter 4.

For an interaction between an oligonucleotide and its target to be stable and demonstrate affinity, the sum of the energies of the non-covalent interactions must greater than the energy of the disrupting forces. With respect to specificity, the oligonucleotide interactions with the counter-target must be minimized such that the non-covalent interactions are less than the disrupt forces.

2.2 FACETS OF SPECIFICITY

At an atomic level affinity and specificity are both driven by non-covalent interactions. At the molecular level, affinity and specificity are determined by the three-dimensional each structure which dictates the physical opportunities for nucleotide-amino acid non-covalent interactions in the aptamer-target pair. There are two types of aptamer-target interactions at the molecular level.

The first interaction type is when the aptamer is larger than the target molecule (*e.g.*, ion) and the second is when the target is larger than the aptamer. Due to the flexibility of the aptamer, in the former scenario, the oligonucleotide tends to incorporate the smaller target molecule into its structure. Chemically, these interactions are stabilized through stacking (*e.g.*, π–π stacking), complementary electrostatic interactions, and the formation of hydrogen bonds.[79,80,81,82,83,84]

Steric hinderance helps facilitate specificity when the aptamer is larger than the target molecule. One notable example of this specificity is the theophylline aptamer which can bind to the bronchodilator theophylline but does not bind to caffeine which differs from the bronchodilator by the presence of a single methyl group.[85] In this example, the aptamer has two conserved loop structures and when the methyl group is present, the hairpin secondary structure, particularly the C32 residue, cannot interact with the caffeine molecule. The additional atoms of the methyl group prevent the nucleotide atoms from being within interaction proximity.[86]

In the second scenario, the target molecule is larger (*e.g.*, protein), the aptamer is integrated into its structure or attaches to the targets surface.[87] Proteins tend to demonstrate high structural complexity and the mechanisms by which aptamers interact with proteins are more diverse than those of aptamer-small molecule interactions. As highlighted by Kohlberger and Gadermaier, "hydrogen bonds still play a role, but often only in combination with polar interactions and structural complementarity."[88]

The negative charges created by the phosphodiester bonds between the phosphate and oxygen atoms in the oligonucleotide backbone interact with the positively charged surfaces of a target protein and are non-specific. These negative charges on the

oligonucleotide great opportunities for polar interactions. From a structural perspective, the specific interactions between the oligonucleotide bases and amino acids are directed by the side chains on the protein. For example, x-ray crystallography of the transcription factor nuclear factor - κB with a high affinity RNA aptamer illustrate both the specific and non-specific electrostatic interactions between the aptamer and protein.[89] The crystallography shows seven nucleotides in the guanine rich loop of the RNA oligonucleotide that form hydrogen bonds the nuclear factor - *k*B p50 homodimer amino acids.[90]

The structural complementarity between aptamer and target molecule are also found in the structure of a thrombin-aptamer complex. The aptamer interacts with the same amino acid residue on the thrombin protein that naturally bind to heparin, an anticoagulant medicine.[91] The crystal structures of multiple aptamer-target protein interactions, collectively, indicate that electrostatic interactions are the driving force for high affinity between the aptamer and target protein.[92] The electrostatic interactions however are driven by the three-dimensional structure of both the aptamer and protein.

While the oligonucleotide structure can generate higher affinity for one target over a counter-target, the possible magnitude of the difference is often unstudied and underdetermined. For example, the SELEX method that identified the oligonucleotides 1C and 4C that bind to the SARS-CoV-2 Spike protein used the ACE2 protein as a counter-selection target. The objective was to generate aptamers solely specific for the Spike protein and not its host-cell target receptor, the ACE2 protein.[93] For aptamers 1C and 4C, Song *et al* report K_d-values for the Spike protein of 5.8 nM and 19.9 nM respectively. For the ACE2 counter-target, no K_d-value was provided.

Instead of using the same technique for determining a K_d-value between both 1C and 4C with the ACE2 molecule, a competition assay was used. The change in aptamer 1C and 4C binding to Spike in the presence of the ACE2 protein was employed as a means of demonstrating specificity. This data is shown in **Figure 2 – 3**. The normalized binding efficiency, as measured by fluorescence levels, decreased by 44% for 1C and decreased by 56% for 4C in the presence of ACE2 protein. Even though, the ACE2 protein was used in the selection process as a counter-target, there is still interaction with the between the Spike aptamer and the counter-target. The aptamers do not exhibit absolute specificity. The lack of aptamer-counter-target K_d-values makes it difficult to determine if specificity exists as defined by the K_d-value for the counter-target is 4 times greater than the K_d-value for the target.

2.2.1 Cofactor Influence on Binding

Cofactors are non-protein compounds that bind to either an enzyme or other protein molecules and facilitate binding of one molecule with another.[94] The compounds have been found to favorably affect the binding between proteins and oligonucleotides. For example, transcription activator proteins typically contain two functional domains which are a DNA binding domain and an activation domain. One such transcription factor is NF-*k*B and its binding interaction was found to be "significantly lower" when using purified recombinant protein instead of extracts from activated cells. This indicates binding was induced when cofactors were present.[95]

As mentioned in Chapter 1, the databases used to train in silico machine learning algorithms have entries containing K_d-values derived from experiments in the presence of

cofactors not described in the dataset. By eliminating the presence or potential use of cofactors, our affinity and specificity are derived from the direct interaction between the nucleotides and the amino acids of the target molecule.

An agnostic environment is also supported by the elimination of cofactors. The agnostic environment focuses the application of the oligonucleotides for use as a biosensor probe. However, if the intent is to use the aptamers for therapeutic purposes, the presence and role of cofactors when determining affinity and specificity should be further examined.

2.2.2 Application Based Design Constraints

The oligonucleotide application environment and set-up are another aspect to be considered when measuring affinity and specificity.

For a graphene biosensor, if the aptamer is too short, then the charge on the target molecule could erroneously skew the resistance measurement across the graphene biosensor. Conversely, if the oligonucleotide is too long then an oligonucleotide could potentially interact with adjacent oligonucleotides diminishing the sensitivity of the graphene biosensor.

There are a variety of techniques available to measure the oligonucleotide affinity and specificity such as surface plasmon resonance, flow cytometry, and confocal microscopy to identify just a few. The procedure and techniques employed need to properly encapsulate the application-based constraints and conditions in order to ensure validation of the oligonucleotides supports the intended application.

2.3 MEASURING AFFINITY AND SPECIFICITY

As previously mentioned, the goal is to design high affinity and specific oligonucleotides. High affinity is defined as the target-aptamer pair as having a K_d-value in the low nanomolar range. Specificity, on the other hand, is defined as a meaningful difference between the K_d -value of the aptamer-target pair and the K_d -value of the aptamer-counter-target pair. The affinity of the aptamer for the target must be equal to or better than four times the K_d-value for the aptamer-counter-target pair.

The K_d-value is determined as the concentration at which half of the target molecules have bound ligands. To determine the K_d-value, various concentrations of the ligand are used, and dose-response plot is generated. Further details on this process are provided later in this Chapter. The lower the K_d-value, then the higher the affinity of the aptamer for a specific target.

Flow cytometry was used to determine the K_d-value for the aptamer-molecule interactions. Flow cytometry can measure cells, beads, and other particles as the items flow singly passed a detector. A flow cytometer provides information about the structure of each item, the fluorescence of each item, as well as counts of the items of interest.[96] The value of using flow cytometry to measure the interaction between the oligonucleotide and the target molecule is the ability to measure large numbers of individual cells and particles within a short period of time. The interaction of the aptamer-target pair is determined by the presence of fluorescence during an interaction and absence when no interaction is occurring.

2.3.1 Experimental Design

The use of flow cytometry enables several unique experimental configurations. The first configuration is the set-up of the streptavidin magnetic beads which can have either a biotinylated protein or a biotinylated oligonucleotide as shown in **Figure 2 – 4**. These configurations are referred to “Protein on the Bead” and “Aptamer on the Bead” respectively.

The second configuration is the use of an additional washing step following the incubation period. The options for this configuration are washed versus non-washed samples. The third configuration is the use of fluorescence calibration. The use of the calibration information could provide further detail about the specific number of bindings per bead in the fluorescence positive population. The fourth configuration is the employment of a joint or single model to fit the dose response curves to the data.

To identify the experimental the correct flow cytometry configuration, a series of experiments were conducted using previously published oligonucleotides to compare K_d-values and molecular simulator scores to test all configurations. The selected aptamers were 1C plus five corresponding mutants as well as 4C and three corresponding mutants for a total of 10 oligonucleotides. Both 1C and 4C aptamers were initial identified to bind with Spike protein.[97]

Three different target proteins (Spike, HA, and CRP) were selected for the experimental design experiments. Molecular simulator data was generated using the molecular docker, HDOCK[98], with the 10 oligonucleotides and three proteins in both the aptamer on the bead and protein on the bead configurations.

Six different concentrations were used for both the protein-fluorophore complex and the aptamer-fluorophore complex in order to estimate the K_d-value of the aptamers and mutants. The fluorescence positive population at each concentration was determined for each configuration. The results were then compared using both parametric (Pearson) and non-parametric (Spearman and Kendall) methods which examined either mean and standard deviation or ordinal position respectively.

Using plotly in the statistical environment R[99], an interactive analysis of both the experimental data with the different configurations is examined for the three correlation methods.

The MATLAB analysis and plotly visualization[100] illustrate, **Figure 2 – 5**, the best correlation, both parametric and non-parametric, between the aptamers, docking scores, and published data is achieved using the aptamer on the bead configuration without the washing step along with the individual fit model and no fluorescence calibration.

While the experimental correlation data illustrates the aptamer on the bead configuration, the added benefit is this configuration more closely mimics the G-FET biosensor where the aptamer is bound to the graphene. Consequently, the subsequent K_d-values are reported using the aptamer on the bead configuration. The following information details the process for the aptamer on the bead experimental set-up including preparation of the aptamer on the beads, attaching the fluorophore to the protein, and the overall experimental design.

The experimental set-up uses a biotinylated aptamer that is attached to a streptavidin magnetic bead and then exposed to the target protein with an attached

fluorophore. This method is employed to mimic the aptamer configuration on the G-FET devices.

2.3.2 Potential For Non-Specific Binding

All experimental configurations have the potential to generate non-specific binding where ligand binds to unintended target molecules. For example, the protein-fluorophore complex interacting directly with the streptavidin magnetic bead. The possibilities of non-specific binding in this experimental is illustrated in **Figure 2 – 6**.

The molecule, bovine serum albumin (BSA), is used to prevent non-specific binding by blocking spaces over a solid surface once immobilization of the aptamer or protein on the bead.[101] To determine the potential effects of non-specific binding with the aptamer on the bead configuration, four aptamers that bind the SARS-CoV-2 Spike protein were tested with and without 100 ng /ml bovine serum albumin (BSA). The concentration of the Spike-fluorophore molecule was set to 0.72 nM which preliminary data showed was near the K_d-value of the previously published aptamer 1C. The four aptamers tested were 1C and 4C (previously published) and dnSpike-1 and 2.

If non-specific binding were occurring, then BSA would block the interaction and there should be a significant decrease in the amount of fluorescent positive population of streptavidin beads. The data, **Figure 2 – 7**, shows that there is no significant difference in fluorescence between the treat with BSA and untread populations. Aptamers were tested four times over multiple days and the figure shows the standard error for each aptamer and category. Our non-specific binding results are comparable to previous studies

showing minimal contribution from non-specific beading using a similar experimental set-up.[102,103,104]

2.3.3 Measurement Options

Successful validation is results from an oligonucleotide sequence that demonstrates both high affinity and specificity. The validation method is closely tied to the end-state application of the oligonucleotide. The geometric configuration of the aptamer on the bead matches the graphene biosensor application by linking the 5'-end of the oligonucleotide to a streptavidin bead. On the graphene biosensor the 5'-end is covalently linked to the PBASE linker molecule. Additionally, the flow cytometry conditions match the biosensor conditions of pH and temperature by both using the same phosphate buffer solution and incubations being done at room temperature (22-25^{o}C).

2.4 EXPERIMENTAL SET-UP DETAILS

Detection of interaction between aptamer and target molecules in the low nanomolar range is required for determining affinity and specificity. There are several methods capable of meeting this detection requirement to include enzyme-linked immunosorbent assay (ELISA), capillary electrophoresis, and surface plasmon resonance (SPR). Flow cytometry was employed in our validation process to measure binding between our aptamers and target molecules. While multiple methods exist for generating the same on graphene biosensor configuration, flow cytometry has two distinct

advantages. These advantages include the relatively low cost and the ability to process multiple samples with multiple concentrations in a short amount of time. A flow cytometer can analyze over 50,000 beads in just a few seconds.

The following sections provide experimental details for measuring the binding of the oligonucleotides to target and non-target molecules.

2.4.1 Magnetic Bead Preparation

To determine the K_d -value using the aptamer on the bead configuration, the aptamer needs to be bound to the streptavidin magnetic bead. The following details the process for linking the aptamer to the bead.

The aptamers are ordered through Integrated DNA Technologies (IDT, Coralville, IA) with a biotin tag on the 5'-end of the oligonucleotide sequence. The aptamers are then diluted to 100 μM concentration using milliQ water in accordance with the IDT specification sheet.

An aliquot of AcroBiosystem streptavidin magnetic beads is transferred to a 1.5 ml Eppendorf tube, prior to being placed into a magnetic Eppendorf tube holder. The magnetic beads are separated from the suspension milliQ water. The water is then removed and discarded. The beads are then re-suspended in an equal volume of binding buffer (PBS plus 0.55 mM $MgCl_2$), prior to magnetic separation and discard of the solution. The magnetic beads are then resuspended again in PBS and prepared for the addition of the biotinylated aptamer.

Per 100 μl of streptavidin magnetic beads, 2 μl of 100 μM biotinylated aptamer is used, equating to 2 μM final concentration of biotinylated aptamer. With 150,000 beads

per µl, this concentration of biotinylated aptamer provides approximately 1.204 x 10^{14} aptamer molecules per bead: ensuring maximum oligonucleotide coverage on the beads.

The necessary amount of biotinylated aptamer is transferred to a 200 µl PCR tube and incubated in at 98°C in a thermocycler (Applied Biosystems, Serial Number: 297806787) for 5 minutes. This step breaks the intermolecular forces (hydrogen bonding) holding the secondary and tertiary structure of the aptamer together. The aptamers straighten and then refold to their natural confirmation while at room temperature (22-25°C). Once the thermocycler incubation is complete, the biotinylated aptamer is added to the washed streptavidin magnetic beads. The Eppendorf tubes containing the magnetic bead and aptamer are vortexed for 3-5 seconds before being placed in a rotational Labquake Shaker (Serial Number: L-1237) from Lab Industries for 2-hours at room temperature (22-25°C).

Following the 2-hour incubation, the Eppendorf tube is returned to the magnetic holder for separation. After 3 minutes, the binding buffer containing unbound aptamer is removed while the beads are retained. The beads with bound aptamer are resuspended in the original volume of binding buffer before being returned to the magnetic holder. After 3 minutes to allow for separation, the binding buffer is then removed. The remaining beads are resuspended again in the original volume of binding buffer.

To confirm attachment of the oligonucleotide to the magnetic beads, the concentration of the single-strand aptamer is determined using the Life Technologies Qubit 3.0 Fluorometer (Serial Number: 2321602342). The assay uses the Qubit single-strand DNA reagents to include buffer, fluorophore, and standards. In addition to the assay for the prepared oligonucleotide on the beads, the concentration assay is also used

on blank streptavidin magnetic beads to determine any possible interaction with the Qubit fluorophore and the streptavidin to ensure accurate concentration of the attach aptamer.

The aptamer-magnetic bead complexes are then stored at -20°C until use in a flow cytometer validation experiment.

2.4.2 Protein-Fluorophore Preparation

For the aptamer on the bead configuration experiment to determine K_d-value, the target or counter-target molecule needs to have a fluorophore for detection by the flow cytometer. The following section details the process for attaching the fluorophore to the molecule.

A biotinylated protein is tagged with a streptavidin conjugated Alexa Fluorophore 488 (Jackson ImmunoResearch Laboratories, Product Number: 016-540-084) for use in the flow cytometry experiments. The concentration of the Alexa Fluorophore 488 is 1.5 μg / ml. This stock solution is created by rehydrating with milliQ water in accordance with the product specification sheet.

For incubation with each biotinylated protein, at least 10 μl of streptavidin-Alexa488 is added to the protein in a 200 μl PCR tube for a final concentration of 0.5 μg / ml. The exact volumes of the protein and fluorophore solution are used to calculate the concentration of the dilute fluorophore-biotinylated protein. The tube is then taped to the inside of an empty pipette tip and covered in aluminum foil. The box is placed on a Boekel Scientific Orbitron Rotator II (Model: 260250, Serial Number: 012202554) at room temperature (22 – 25°C) for 120 minutes.

Once the incubation is complete, the volume is aliquoted into smaller PCR tubes for use during the flow cytometry experiments. The exact volume of the protein-fluorophore molecule used in the flow cytometry experiments is determined by the final concentration of the stock solution and the required concentrations for calculating the K_d-value.

2.4.3 Flow Cytometry

The validation experiments for the digitalSELEX platform were completed using the aptamer on the bead configuration. This configuration, shown in **Figure 2 – 8**, has two advantages. First, the aptamer on the bead mimics the placement and configuration of the aptamer on the G-FET devices. In the case of the magnetic bead, the nucleotide's 5'-end has a biotin molecule, whereas for the G-FET device there is an amine group on the 5'-end on the nucleotide to covalently bind the aptamer to the PBASE linker molecule.

The second advantage is the binding of a protein to an aptamer on a bead will not increase or decrease the likelihood of another protein binding to another bead. This configuration reduces the ability of cooperative binding on the target molecule to increase the fluorescence signal. The size of the protein could prevent multiple protein-fluorophore molecules from binding to a single bead, however only a single protein-fluorophore molecule needs to be bound to generate a positive signal.

As previously discussed, the K_d-value is the mid-point in a dose response curve that is generated using multiple concentrations. The concentration range of 1 pg / ml to 10 μg / ml was used but can be modified depending on the determined K_d -value. For the

validation experiment, a calculated aliquot of aptamer-magnetic beads is warmed to 98°C for 5 minutes in a 200 µl PCR tube. Typically, the amount warmed is 5 µl more than the required experimental amount in order to account for physical loss and pipette errors.

Once the incubation is complete, then 15 µl of aptamer-magnetic beads is added to 35 µl of binding buffer for a total volume of 50 µl in a 200 µl PCR tube. There is one tube per concentration. The PCR tubes are placed in a 96-well magnetic ring plate for the beads to separate from the solution. The supernatant is removed and discarded leaving the beads. The PCR tubes are then removed from the magnetic plate. The beads are then resuspended in a 50 µl solution composed of binding buffer (PBS with 0.55 mM $MgCl_2$) a designated concentration of the protein with attached Alexa488 fluorophore. The PCR tubes are then moved to an empty pipette tip box and secured using masking tape on their side. Aluminum foil is then placed over the PCR tubes containing the aptamer on the beads and protein-fluorophore solution to prevent ambient light from exciting the fluorophore. The box is then placed on a Boekel Scientific Orbitron Rotator II (Model: 260250, Serial Number: 012202554) at room temperature (22 – 25°C) for 45 minutes.

Following the incubation period, the PCR tubes are returned to the 96-well magnetic plate to separate the beads from the solution. After 3 minutes of separation time, the supernatant is removed from the PCR tube and discarded. The beads are then resuspended in 50 µl of binding buffer and returned to the magnetic plate. Following another 3-minute separation, the supernatant is removed and discarded. The beads are then resuspended in 125 µl binding buffer and transferred to the properly labeled flow cytometry tube.

The samples are then analyzed using a BD Systems FACS Aria IIIu in accordance with Boston College Flow Cytometry Core procedures as established by Dr. Patrick Autissier. Each experimental run includes both an untreated sample and the desired range of concentrations. The flow cytometer acquires at least 50,000 beads per sample tube.

2.4.4 Negative Controls – Unbound Fluorophore Contribution

To ensure the percent fluorescent positive magnetic bead population is not affected by unbound streptavidin fluorophore, a negative control experiment was generated.

Non-aptamer bound streptavidin magnetic beads were incubated with 0.5 μg / ml streptavidin fluorophore. This concentration is greater than the possible exposure to unbound streptavidin fluorophore when using the 10 μg / ml protein-fluorophore concentration. Following the 45-minute incubation, the fluorescence was measured using the flow cytometry.

The contribution of the free streptavidin-Alexa488 fluorophore on the magnetic beads is shown in **Figure 2 – 9**. The figure shows the dose response curve for the previously published aptamer 1C using the aptamer on the bead configuration. The dose response curve shows the concentration of the protein-fluorophore molecule on the x-axis and the percent fluorescent positive on the y-axis. The green point in the lower right-hand corner illustrates the negligible contribution of unbound streptavidin fluorophore binding to the streptavidin magnetic beads.

2.4.5 Flow Cytometry Data Analysis

Each flow cytometry sample generates a corresponding flow cytometry file (fcs) which are then analyzed using FlowJo (version 10.8.1). The FlowJo software is used to determine the fluorescence (mean, median, and standard deviation) as well as the distribution of fluorescence across the magnetic beads. The distribution of fluorescence in each treated sample is compared to the untreated control to determine the percent fluorescence positive beads. This is analysis is illustrated in **Figure 2 – 10**.

To extract the data, the fcs files are gated to remove any doublets from the population. The fcs file is visualized using the FITC-histogram mode in log scale. The area beyond the untreated population then becomes a subset of the overall population and is the fluorescent positive population. While in the untreated control this percentage of the overall population is close to zero, this subset gate is propagated through the samples with increasing protein-fluorophore to identify FITC positive beads. The fluorescent positive population for each sample is recorded.

This percent of the fluorescent positive data is used to calculate the K_d-value for the aptamer for both its target and non-target proteins.

2.4.6 Calculating K_d-value

There are two predominant methods for calculating K_d-value of the aptamer-protein interaction: using the fluorescence intensity, which is the binding per bead, or the population that is fluorescent positive which the number of beads showing fluorescence.

Due to our configuration with the aptamer on the bead to resemble the G-FET, our dissociation constant (K_d) value analysis utilizes the second method. The K_d-value is the effective dose based on the midpoint between the lower and upper asymptotes. This becomes an issue when the lower asymptote is not close to zero or the upper asymptote does not approach one hundred percent. Physical loss and incomplete reactions account for error with the upper asymptote. The K_d-value was only calculated if the upper asymptote on a designed aptamer was greater than 80%. Experiments were repeated at least three times to reduce potential error.

The analysis was done in R[105] using the dose response model package.[106] An example of the dose response model and analysis is illustrated in **Figure 2 – 11**. The concentration is on the x-axis and the positive population is in the y-axis. There are two common approaches to represent the positive population on the y-axis. The first method is showing the raw values, while the second method is to normalize the data to the maximum value. The non-normalized data representation was used for the digitalSELEX validation. This method was chosen to highlight oligonucleotides that fail to reach 100% fluorescent positive beads. The low binding of these aptamers can be attributed to a variety of factors to include misfolding of the oligonucleotide or the length of the aptamers causing them to interact with adjacent aptamers on the bead. Regardless of the specific cause, aptamers that fail to achieve over 80% of the fluorescent positive beads are not considered good for application purposes. The maximum binding is an additional data point used to ensure aptamers are good.

Dose response plots were generated for each aptamer using the non-normalized percent population versus concentration. The concentrations ranged from zero to the low

hundred nanomolar protein concentration range since the goal was to identify aptamers with a K_d-value in the single digit to low double nanomolar concentration. The dose response model package in are enables the selection of the best function however it is typically the Weibull 2.4 which is a four-parameter dose response function.[107-108]

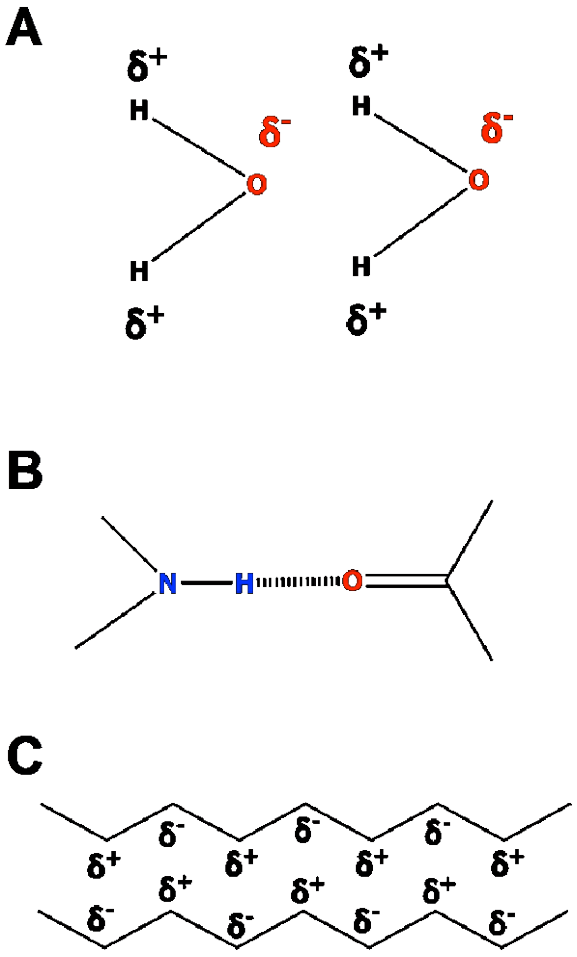


Figure 2-1. Intermolecular Forces.
Diagram illustrates three examples of intermolecular forces that are biologically relevant. These interactions are a dipole-dipole (A), hydrogen bond (B), and dispersion interactions (C). The dipole-dipole interaction occurs when the electron density of the molecule is localized. In this illustrate the electrons are localized around the oxygen atom. This gives the oxygen a partial negative charge while the hydrogen atoms have a partial positive charge. The energy distance dependence of approximately $1/r^3$. A dipole-dipole, either temporary or permanent, induce Van der Waals forces which occurs when other molecules and atoms align. The hydrogen bond (B) is a special type of dipole where the two electronegative atoms compete for the same hydrogen atom. The energy dependence of the hydrogen bond is approximately $1/r^2$. Panel C illustrates a dispersion interaction which is a sub-type of Van der Waals and is the only intermolecular force present in non-polar molecules. The energy distance dependence is approximately $1/r^6$.

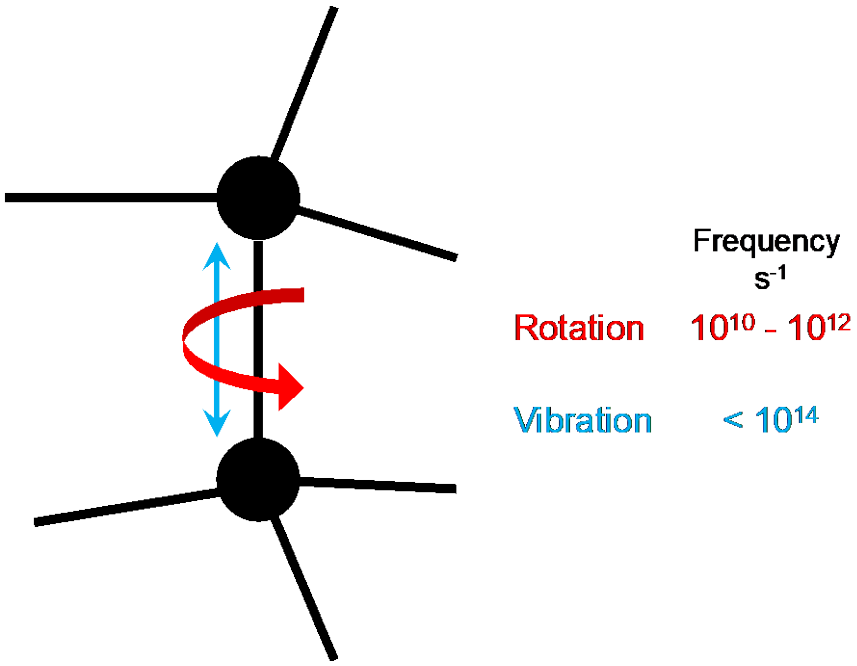


Figure 2-2. Disrupting Forces.
Diagram of vibration and rotation of two carbon atoms in a covalent C – C bond. The frequency of these movements is given in Hz or s^{-1}. The corresponding thermal energy of vibration and rotation requires ~2.6 x 10^{-2} eV at 300 k whereas translational movement requires small amounts of energy (~10^{-16} eV) and can be considered continuous. Adopted from Glaser, Biophysics.

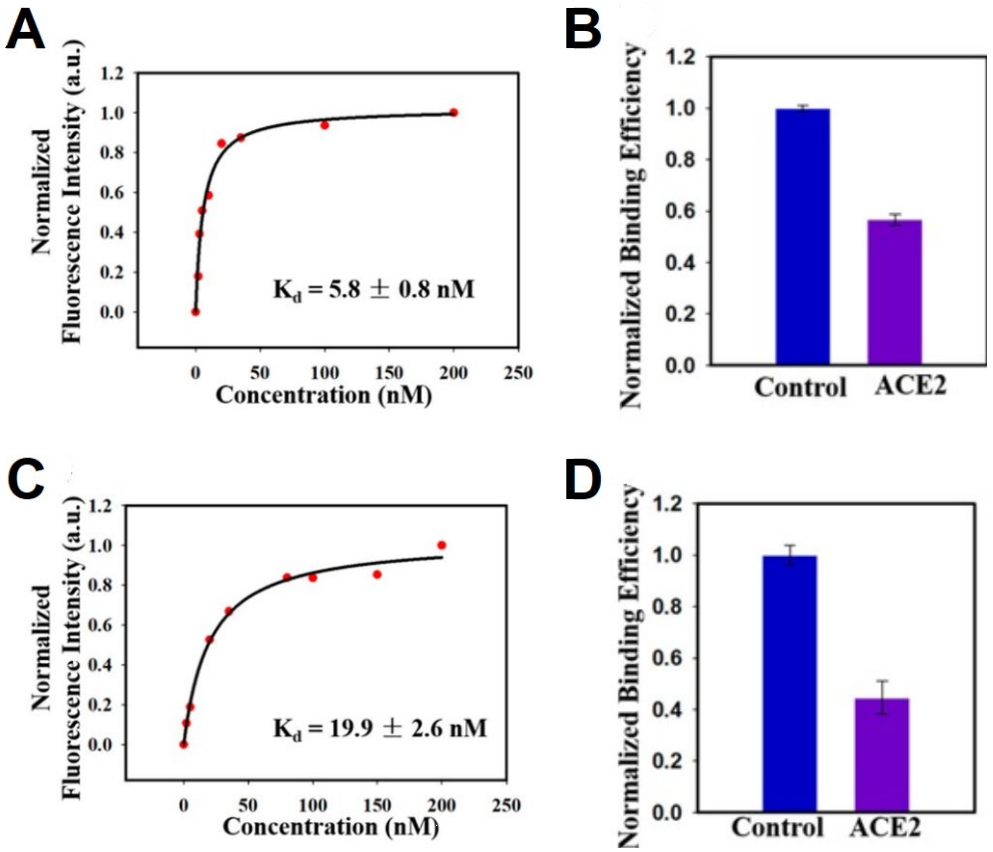


Figure 2-3. Aptamer 1C and 4C Affinity and Specificity.
This figure is compiled from figures 4 and 5 published by Song *et al*. These plots were used by the authors to illustrate affinity and specificity for aptamers 1C and 4C. The aptamers were selected to be specific for the SARS-CoV-2 Spike protein over the ACE2 receptor. The authors experimentally determined the K_d-value of both 1C (A) and 4C (C) binding Spike protein. A competition assay with a mixture of both Spike and ACE2 proteins was however employed to demonstrate specificity for 1C (B) and 4C (D). The control (blue) is the measure of the aptamer binding to the Spike protein alone. The ACE2 bar plot (purple) shows decrease in the relative fluorescent on the Spike protein when 4-times the amount of ACE2 was added compared to Spike protein. No K_d-value was determined for the aptamers with ACE2 protein using the same method as the affinity assay.

Figure 2-4. Bead Configurations.
Illustration of the two tested flow cytometry configurations. Panel A shows the aptamer on the bead configuration where an oligonucleotide is chemically synthesized with a biotin tag on the 5' – end and is linked to a streptavidin magnetic bead. Panel B shows the protein on the bead configuration where a biotinylated protein is bound to a streptavidin bead. This is the most common configuration as the protein being the larger molecule with bound to the bead. The aptamer on the bead configuration however mimics the graphene biosensor since the oligonucleotide is fixed.

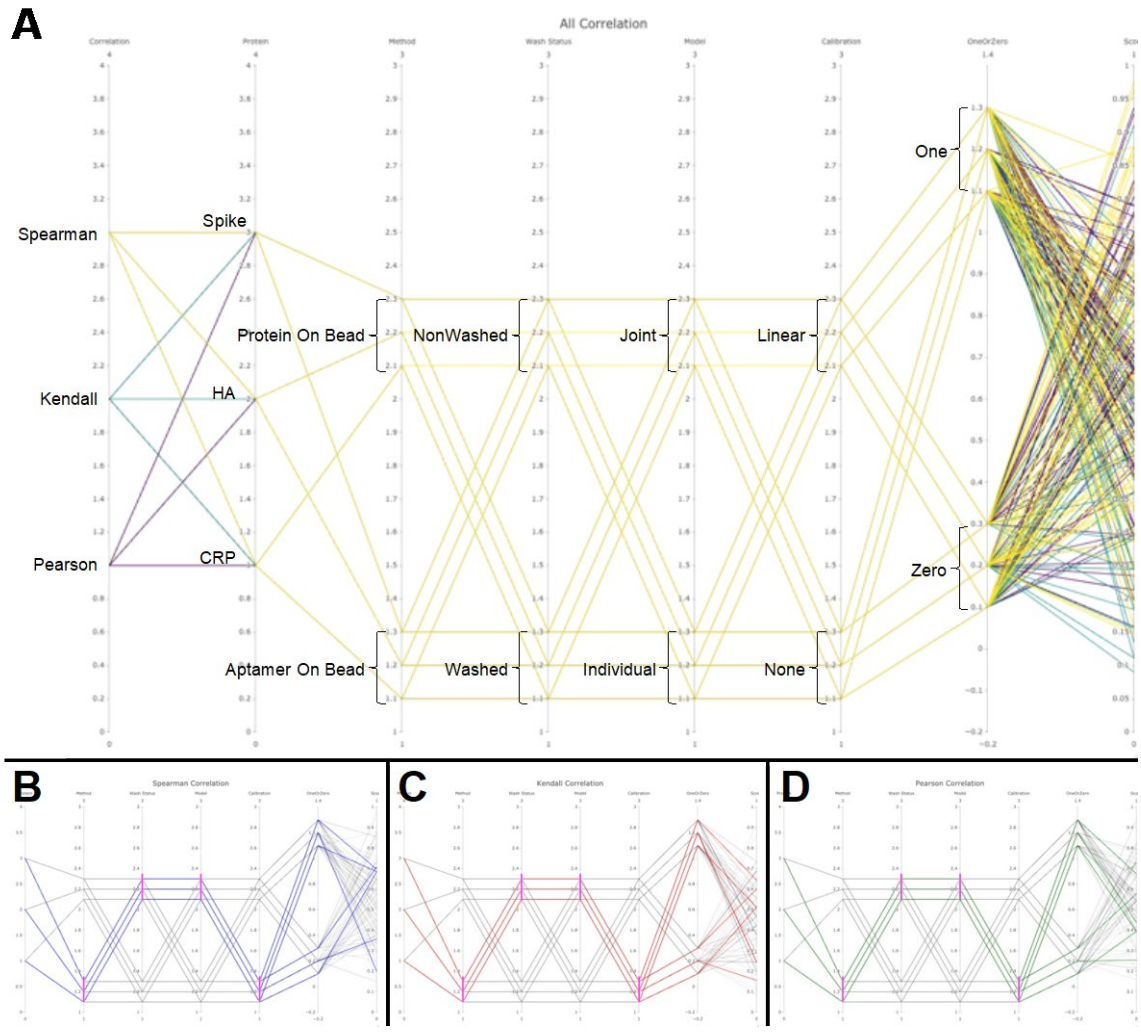


Figure 2-5. Visualization of Validation Configuration Tests.
The plotly visualization of the MATLAB correlation analysis. Panel A illustrates all three correlation methods (Spearman, Kendall, and Pearson) with the different proteins (Spike, HA, and CRP) along with the different experimental configurations (aptamer on the bead, protein on the bead, and washing step). The analysis method was also varied with joint or individual fit models and using the fluorescence calibration data. Panels B-D highlight the individual correlation methods of Spearman, Kendall, and Pearson respectively.

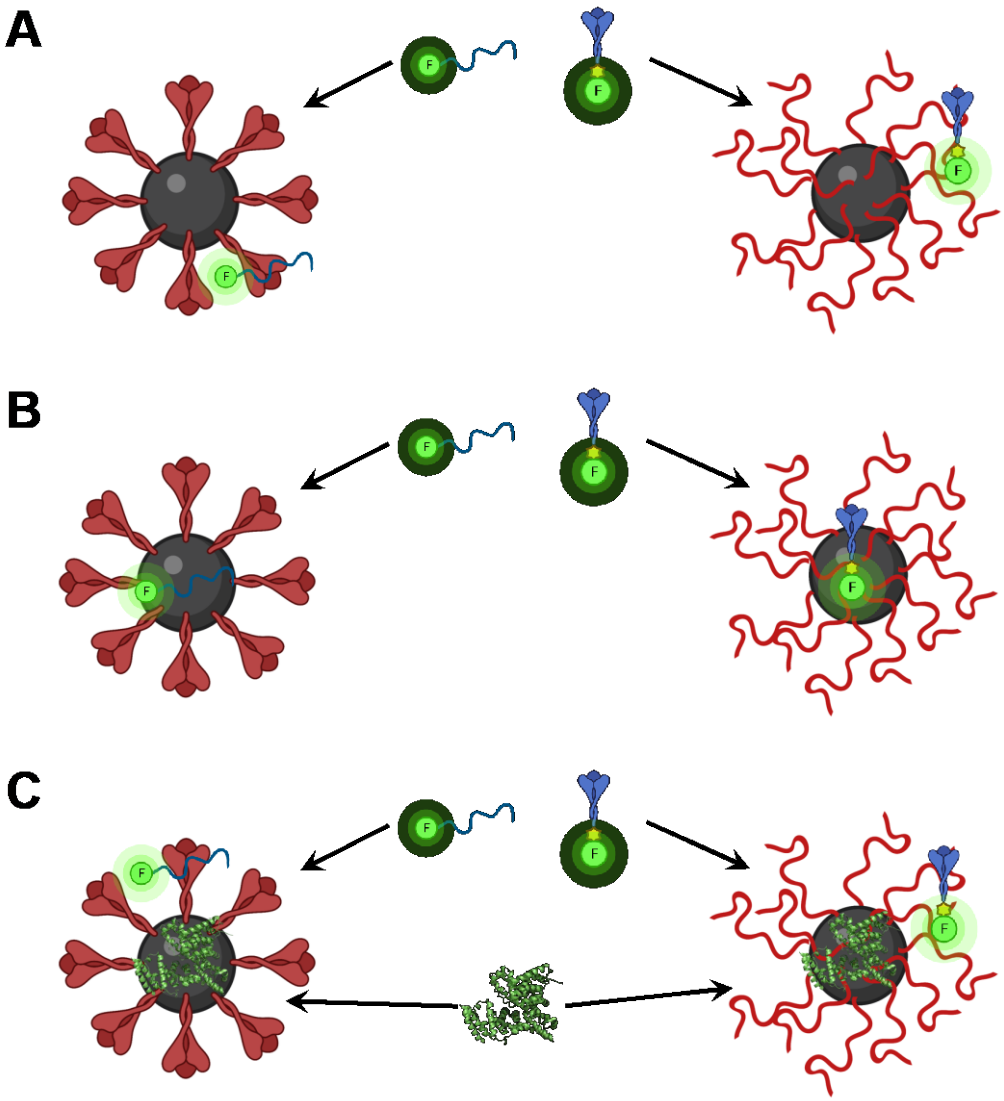


Figure 2-6. Effect of Non-Specific Binding.
Illustration of potential non-specific binding on the two bead configurations. Panel A shows the desired scenario where the aptamer- fluorophore (right column) and protein-fluorophore (left column) bind to their respective targets in the two configurations. Panel B shows the potential for non-specific binding where the molecule with the fluorophore binds to the streptavidin bead instead of the target molecule. This potential is great with the protein on the bead configuration where steric hinderance can limit coverage across the bead. Since the aptamer is approximately one-tenth the size of the protein, it could interact with the bead directly and not the protein. The aptamer on the bead configuration has better coverage across the bead due to less steric hindrance and non-specific binding is less likely. Panel C shows that Bovine Serum Albumin (BSA) is small enough to interact with the beads and block non-specific binding though not necessary in the aptamer on the bead method.

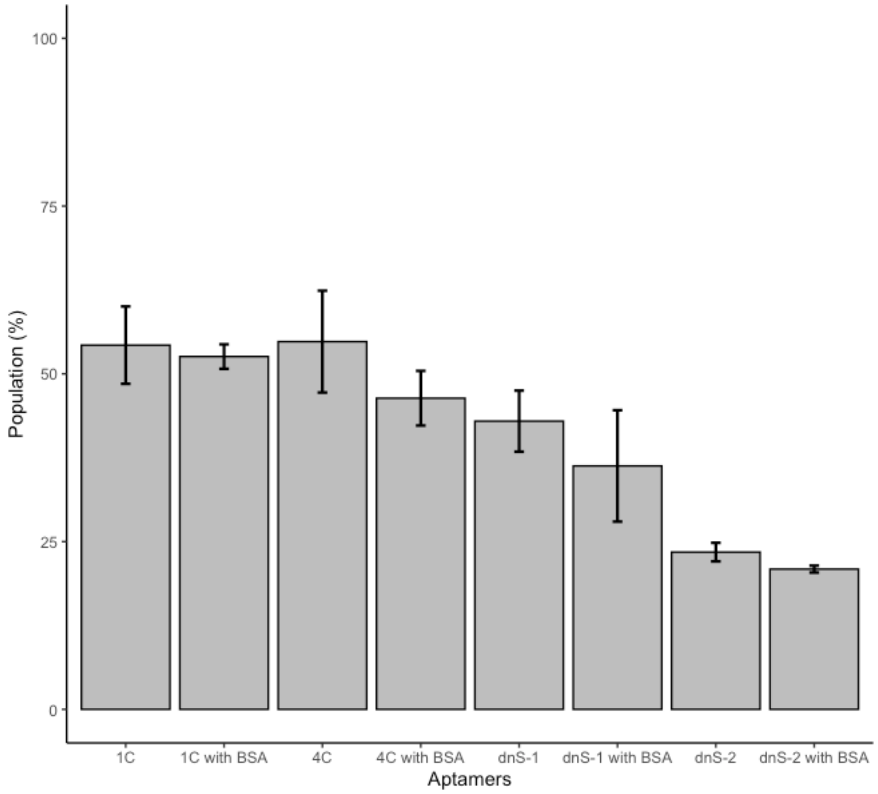


Figure 2-7. BSA Effect on Binding Measurements
This bar plot shows the results of the fluorescent positive population with four different aptamers when binding 0.72 nM SARS-CoV-2 Spike fluorophore with and without 100 ng /ml of BSA. There is no significant difference for any of the aptamers binding the Spike protein when BSA is present.

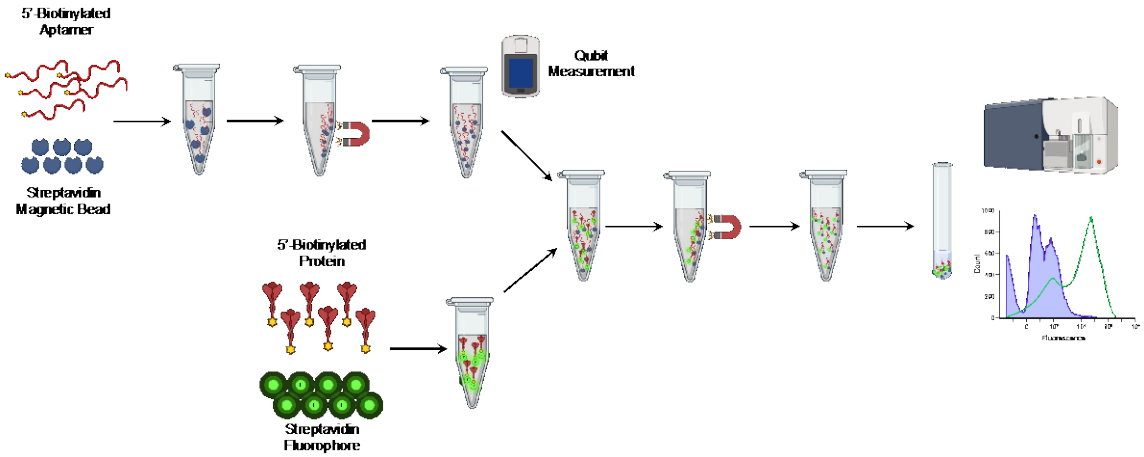


Figure 2-8. Experimental Set-up
Illustration of the aptamer on the bead experimental configuration used to measure the binding affinity of the aptamers to the target protein. The concentration of the aptamers is measured on the beads using the Qubit Fluorometer and corresponding ssDNA kit. The flow cytometry provides data on the fluorescent positive population corresponding to the number of protein-fluorophore complexes bound to the aptamers on the beads.

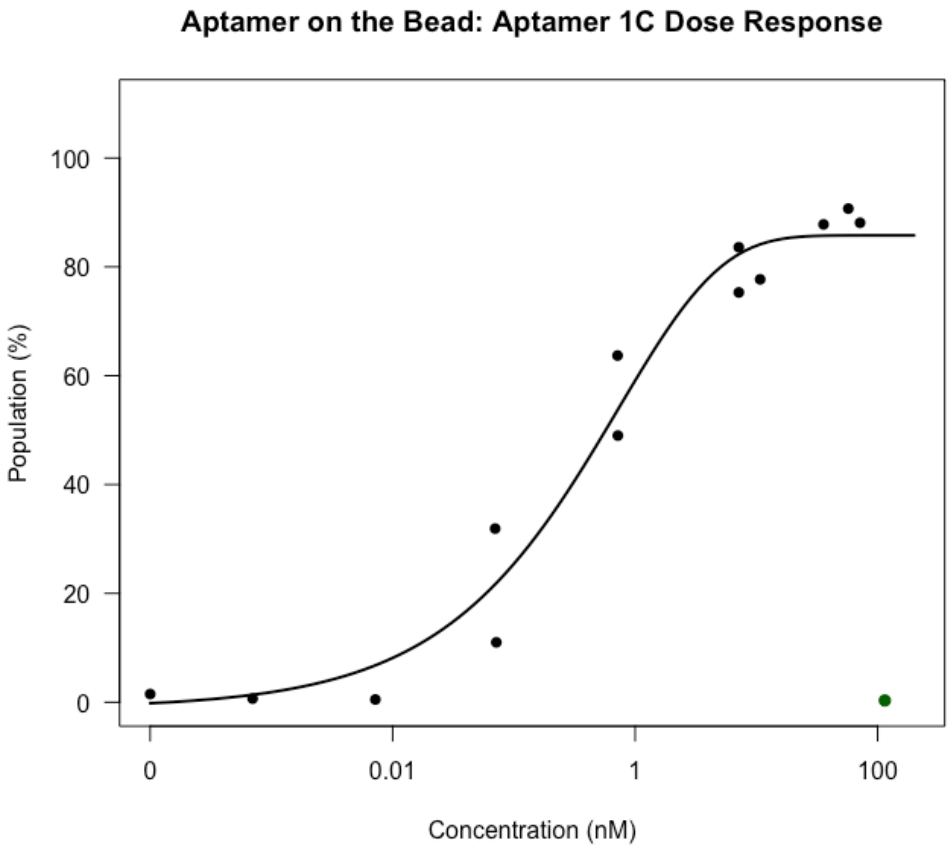


Figure 2-9. Dose Response Curve with Negative Control.
This depicts the dose response curve of the previously published aptamer 1C with the SARS-CoV-2 Spike protein using the aptamer on the bead configuration. This oligonucleotide was selected *in vitro* to bind the Spike protein. The green point in the lower right-hand corner is the positive population of streptavidin magnetic beads when incubated for 45 minutes at room temperature with streptavidin-Alex488 fluorophore. This control was tested three times with no discernible contribution from unbound fluorophore.

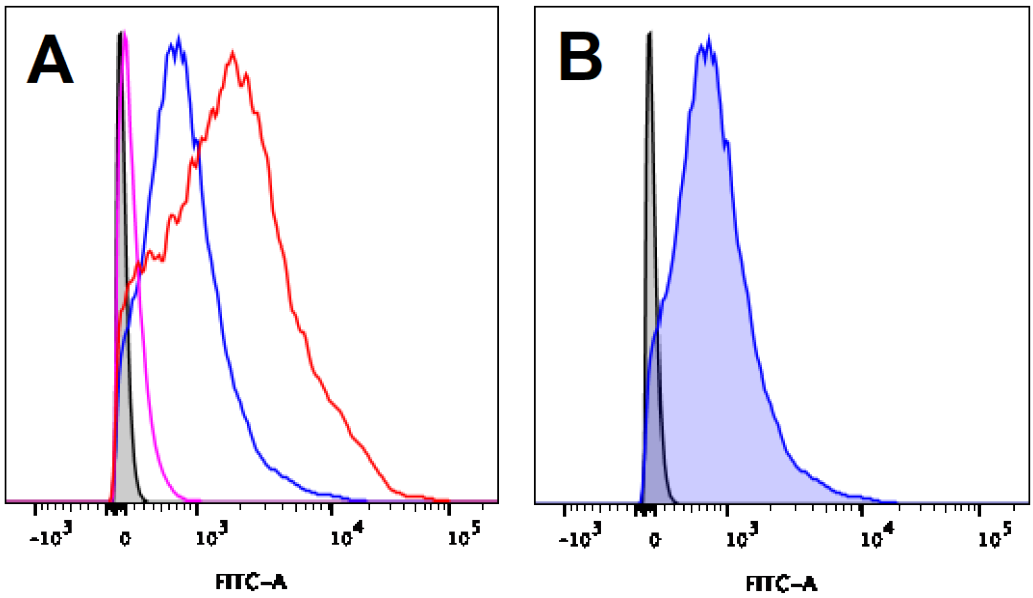
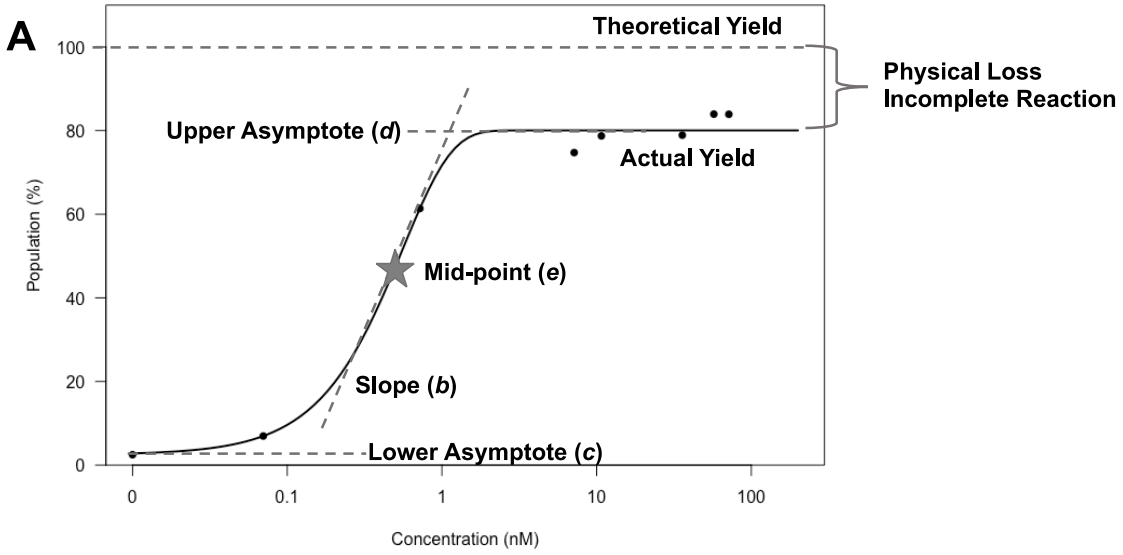


Figure 2-10. FlowJo Fluorescent Positive Population Analysis.
Example figures showing comparative fluorescence positive populations. Panel A shows the control sample (grey) (beads and aptamer only) relative to increasing concentrations of green fluorophore tagged SARS-CoV-2 Spike protein. The percent positive population is calculated by determining the percentage of the beads with aptamer that are fluorescent positive compared to the control. Panel B shows just a single concentration (blue) compared to the control (grey). FlowJo analysis software calculates the difference between the population. These percent positive quantities are recorded for each tested concentration and used to determine the K_d-value.



B

$$f\{x,(b,c,d,e)\} = c + (d - c)\{1 - \exp(-\exp[b\{\log(x) - \log(e)\}])\}$$
$$= c + (d - c)\{1 - \exp\left(-\left(\frac{x}{e}\right)^{b}\right)\}$$

Figure 2-11. Dose Response Curve Analysis.
The dose-response curve in Panel A depicts a typical plot of the Weibull 2.4 model of aptamer binding to a target protein at varying concentrations (solid black line). Highlighted on the plot are the lower and upper asymptotes, the slope, and mid-point of the plot. Notice the actual yield is less than the theoretical yield. This difference can be attributed to physical loss (*e.g.*, protein interacting with walls of the vessel) or incomplete reactions. Since the flow cytometry data is a snapshot, it could be that the incomplete reactions are a result of motion / vibrations dislodging weakly bound protein to aptamers. Panel B is the four-parameter equation for the Weibull type 2 model. In this model, parameter *b* reflects the steepness of the slope, while *c* and *d* reflect lower and upper asymptote limits, respectively. The fourth parameter *e* is the half-way point between the lower and upper limit and reflects the dissociation constant (K_d). This model was developed by Piegorsch and Bailer in 2005.

3.0 CHAPTER 3 – COLD START MODULE

This chapter describes the process of single-stranded oligonucleotide design by identifying clusters of atoms that are both accessible and binding-relevant on the target molecule. Nucleotides are then assigned to the corresponding amino acids of the atom cluster prior to an initial sequence being generated through a genetic optimization algorithm.

3.1 DESIGN CONCEPT

Challenges of existing selection methods are Initial pool size, time, cost, and relatively low success rate, and data dependence. The *Cold Start* module in the digitalSELEX platform is our approach to address these inherent challenges through *de novo* design. The central concept, **Figure 3 – 1**, behind our *Cold Start* algorithm is to: (a) identify regions on a target molecule that can interact with potential oligonucleotides and (b) build an initial aptamer that has a high likelihood of binding with the target molecule. Potential regions have atoms that are both physically accessible by the aptamer and capable of interacting with the atoms of the oligonucleotide. The aptamer is built using tabled data that indicate specific nucleotides are likely to bind with specific amino acids. Non-covalent interactions between nucleotides and amino acids were discussed in Chapter 2.

3.2 CONCEPT ALGORITHMS

The primary assumption of the Cold Start module is that three-dimensional structure of an aptamer and target molecule are global. Local regions on both the aptamer and target molecule are structurally simple and where the interaction between the two molecules will occur.

To identify local regions capable of interacting with potential oligonucleotides, we propose two algorithms, a *k-means clustering* and *probabilistic application algorithm*. Both algorithms use the same primitives to assess the structure of the target molecule and determine the atoms that might interact with the aptamer. The two algorithms differ in their methods for assembling nucleotides for corresponding amino acids.

The shared function of the *Cold Start* module imports the three-dimensional structure of a target molecule, parses out biologically relevant atoms, specifically oxygen and nitrogen, and determines the atoms accessibility through a solid angle calculation. Once this information has been derived, the *Cold Start* module deviates into the two algorithms.

The two algorithms generate only a section of the final oligonucleotide sequence. The part generated by the algorithms is known as the “core” sequence. Afterwards, nucleotides are added before and after the core sequence, which are referred to as the prefix and suffix respectively. A genetic optimization algorithm is used to identify the nucleotides that compose the prefix and suffix sequences to generate the most stable aptamer that satisfies application specific constraints.

3.2.1 Algorithm 1: Concept

Algorithm 1, at this point, clusters the accessible atoms so that the number of amino acids in the cluster is approximately 10 – 15. This clustering employs MATLAB's k-means algorithm using the squared Euclidean distance so that the centroid of the cluster is the mean distance away from the points. The specific number can be varied by altering the number of clusters to be generated. This variable is useful depending on the size of the target molecule and desired number of amino acids in the cluster. The solid angle for the biologically relevant atoms is then summed within each cluster. The clusters are then rank ordered based accessibility with the top cluster being the most accessible cluster. Nucleotides are then assigned to each amino acid based on frequency of interaction data.[109] The order of the nucleotides is based on the primary sequence of the amino acids.

While it is possible to determine an order of nucleotides based on the three-dimensional organization of the amino acids, there are potential drawbacks to implementing this additional step. This process makes alignment of the nucleotides of the aptamer and amino acids more rigid. The binding of the aptamer to the protein is not a rigid configuration, like a key into a lock. Instead, the protein and the aptamer are moving and vibrating as discussed in Chapter 2. Additionally, there is a size difference between the nucleotides and most amino acids. Using specific three-dimensional information would require filler nucleotides to generate perfect alignment which then adds greater complexity to the overall structure. By keeping the nucleotide arrangement simple, there is greater flexibility in binding which allows some nucleotides to be bound while others are not.

3.2.2 Algorithm 2: Concept

Algorithm 2, probabilistic application, assumes that on small-scale amino acids of the target molecule are linear in the three-dimensional space. Therefore, it is reasoned, small subsequences of amino acids on the target molecule, counter-target molecule, and even the assigned nucleotides can be arranged in a quasi-linear three-dimensional fashion. Algorithm 2 applies the data derived from the molecular structure (*e.g.,* solid angle and proximity to other atoms) in a different manner.

The probabilistic application algorithm first generates a search map of atoms with a maximal and minimal distance. This search map is an exhaustive string of atoms which is further delineated to all possible combinations of orders for a given size. The given size is the number of amino acids. This process is done for both the target molecule and counter-target molecule. All unique strings for both the target and counter-target are combined into a single list of unique strings. Each unique string is then counted in each molecule.

The application of probabilities occurs as a list of all possible nucleotides sequences is generated for the number of amino acids in the string. For example, if the number of amino acids in the search string is 6, then there is 4^6 or 4,096 possible nucleotide combinations. For each possible nucleotide in each possible combination, a probability / frequency of occurrence is assigned based on the corresponding amino acids. The probabilities are then multiplied by the occurrence of the sequence on the target and counter target separately. The possible nucleotide combination probabilities are then summed across all the unique amino acid strings. This generates two lists (target and counter-target) containing probabilities for each possible nucleotide combination.

3.2.3 The Core Sequence

After the “core” sequences are identified via either Algorithm 1 or 2, the remaining nucleotides of the sequence need to be identified. With respect to Algorithm 1, the nucleotides are identified using a genetic optimization algorithm. The specific number of nucleotides before and after the core sequence are user identified. Algorithm 2 can use the same optimization method as Algorithm 1, however there is an additional option. Specifically, Algorithm 2 identifies multiple small sequence fragments which can then be stitched together and then optimized.

3.3 NOVEL COMPONENTS

The *Cold Start* module is novel from previous *in vitro* / *in silico* SELEX work in several ways. First, no previous work identified biologically relevant and accessible atoms on the target or counter-target molecule. Secondly, no other work has clustered the atoms based on solid angle, proximity or used a frequency calculation to rank potential amino acid substructures.

The use of a Genetic Optimization Algorithm coupled with rnafold for examining the secondary structure of the oligonucleotide sequence is also novel. The combination of using an optimization algorithm with rnafold to identify the most stable structure under specific constraints has not been applied in other previous SELEX variations.

3.4 IMPLEMENTATION DETAILS

Essential to both Algorithm 1 and 2 is the identification of accessible, binding-relevant atoms. Accessible atoms are defined as atoms that are not surrounded by other atoms as determined by calculating the solid angle which is a measure of the three-dimensional angular volume akin to the plane angle of two dimensions. Binding-relevant atoms are those atoms capable of hydrogen bonding, specifically oxygen and nitrogen. While other electrostatic interactions contribute to the binding of an oligonucleotide to the target molecule, hydrogen bonding requires specific distance to support the special dipole-dipole interaction and is an indicator of a stable interaction. These non-covalent interactions were detailed in Chapter 2.

3.4.1 pdb file import

The first function of the *Cold Start* module uses the MATLAB function, pdbread[110], to import a protein databank (pdb) file. The pdbread function can import a pdb file from the protein databank using the four-character identifier or the file can be imported from a specific folder. A pdb file is a text file and the pdbread function imports the file as a structure, delineating the components of the into accessible arrays. An example of the protein structure data in a pdb file is shown in **Figure 3 – 2**.

The primary component of a pdb file is the experimentally determined three-dimensional structure information. The experiments used to elucidate the macromolecule structure vary but typically the structure information is derived from x-ray crystallography, NMR spectroscopy, or cryo-electron microscopy.[111] The experimental

data is contained in the experimental data structure of the text file, while the header section which details information about the protein, citation information, and the structure resolution.

The primary amino acid of the structure is listed. The amino acids listed are only those present in the experimental structure and not of the entire macromolecule. The three-dimensional cartesian coordinates (x, y, and z coordinates) of each atom and corresponding amino acid are listed atom section.[112] The first atom listed in the atom section corresponds to first atom of the amino-terminus amino acid that is visible in the three-dimensional structure. The first atom will be the nitrogen since it the furthest most atom on the amino terminus followed by the alpha carbon.[113] The atoms are labeled and numbered in accordance with the International Union of Pure and Applied Chemistry (IUPAC) nomenclature.[114] Hydrogen atoms are omitted in structures derived from x-ray crystallography due to resolution issues, however these are present in pdb files derived from NMR and theoretical models. While hydrogen atoms are critical to forming non-covalent hydrogen bonds, the *Cold Start* process focuses on both oxygen and nitrogen atoms as biologically relevant and accessible.

Not all pdb files have the same quality or resolution since the structure information is experimentally derived. The protein databank defines resolution as, "a measure of the quality of the data that has been collected on the crystal containing the protein or nucleic acid. If all proteins in the crystal are aligned in an identical way, forming a very perfect crystal, then all the proteins will scatter X-rays the same way, and the diffraction pattern will show the fine details of crystal. On the other hand, if the proteins in the crystal are all slightly different, due to local flexibility or motion, the

diffraction pattern will not contain as much fine information. So, resolution is a measure of the level of detail present in the diffraction pattern and the level of detail that will be seen when the electron density map is calculated. High-resolution structures, with resolution values of 1 Å or so, are highly ordered and it is easy to observe every atom in the electron density map. Lower resolution structures, with resolution of 3 Å or higher, show only the basic contours of the protein chain, and the atomic structure must be inferred. Most crystallographic-defined structures of proteins fall in between these two extremes. As a general rule of thumb, we have more confidence in the location of atoms in structures with resolution values that are small, called 'high-resolution structures'."[115]

Resolution of the structures presented in the pdb file is consequently critical for enabling the *Cold Start* module to identify and then cluster biologically relevant atoms that are accessible. It is recommended that users, select high resolution (~ 1 Angstroms) pdb files.

3.4.2 Identification of Biologically Relevant and Accessible Atoms

Proteins and macromolecules naturally differ in their size as do their corresponding pdb file. For example, the receptor binding domain (Spike protein) the SARS-CoV-2 virus in the 6XM3 pdb file has 3143 amino acids for a total of 25,383 atoms, whereas human C-reactive protein in its monoclinic form, depicted in 3PVO pdb file, has 4120 amino acids and 33,077 atoms. The data in the pdb files also differ for the same proteins depending on the methods and resolution. The resolution of the Spike protein is 3.46 A in 6VSB only has 2905 amino acids depicted compared to the 6XM3 model with a resolution of 2.46 A which has 3143 amino acids. The amount of

information in the pdb file can consequently affect the computational time of the function. Consequently, the initial step is parse out the amino acids into two categories: hydrophobic and hydrophilic. While the hydrophobic amino acids have both oxygen and nitrogen, these amino acids typically are not accessible and reside at the core of the molecule. The accessibility calculations incorporate all atoms in the molecules since even hydrophobic amino acids can hinder access to an atom.

The Delaunay triangulation function in MALTAB is then employed to create three-dimensional array for atoms. The connections in the Delaunay triangulation are listed in a four-column matrix. This method avoids the generation of low angle, or sliver, triangles. An example of the Delaunay triangulation is shown in **Figure 3 – 3A**.

The all-atom, four-column matrix is then subjected to an alpha shape algorithm in MATLAB, which generates a bounding volume that envelopes the three-dimensional points. The volume of the alpha shape can be manipulated by increasing or decreasing the alpha term. This manipulation increases or decreases the number of points to be considered. The product of this step includes all atoms on the periphery of the target molecule. A filter against all atoms in the alpha shape is then incorporated to retain all biologically relevant atoms. Biologically relevant atoms are oxygen and nitrogen on hydrophilic amino acids due to their propensity to form hydrogen bonds.

The solid angle is then calculated for all biologically relevant atoms designated after the alpha shape calculation using the Delaunay triangulation matrix of connections, **Figure 3 – 3B**. For each atom in the matrix, the angle of all surrounding atoms is determined. A completely surrounded, or internal, atom would have a sphere of other atoms around it with a calculated solid angle of 4pi whereas a highly accessible atom

would have a solid angle calculation of less than 2pi. The biologically relevant atoms are then filtered for their accessibility.

3.4.3 Algorithm 1: K-means Clustering

The accessible biologically relevant oxygen and nitrogen atoms are then clustered according to their proximity using a k-means nearest neighbor search in MATLAB.[116,117] The search algorithm searches for the closest neighbors of the highly accessible atoms. This process is illustrated in **Figure 3 – 4**. The number of amino acids in the cluster should be exceed 20 and ideally is between 10 – 15. The exact number can vary depending on the size of the target protein. Small proteins have less amino acids so the cluster number should be adjusted to reflect the size of the protein.

The generated clusters are then prioritized based on the total solid angle of all the atoms in the cluster. The output of this function is a list of 10 amino acid clusters with the smallest solid angle (most accessible). The list includes the cluster ID, atom, atom ID, amino acid, and amino acid id. An example of the output of this clustering process.

3.4.3.1 Nucleotide Assignment

The identified cluster amino acids are paired with an oligonucleotide based on historic nucleotide / amino acid interactions.[118,119,120] This function consists of a look-up table where every amino acid has a corresponding nucleotide based on the frequency of interaction data from Luscombe *et al*. This frequency data is shown in **Figure 3 – 5**.

The top ten clusters from the previous function are all assigned nucleotides. The nucleotide sequence is the primary amino acid sequence order in the target molecule.

While it is known that the primary sequence of amino acids in the target does not directly correspond to the three-dimensional structure of the macromolecule, the nucleotides are based on sequence due to simplicity of the code. The orientation of the aptamer to target can be fixed computationally, however molecular dynamics demonstrate that a fix orientation is not realist. The interaction between nucleotides and amino acids are transient and distances change in fractions of a second. The assumption is such that the order in the core sequence is less important than the availability of the nucleotides to interact. These assigned oligonucleotides then become the core sequence for optimization of the remaining sequence.

3.4.3.2 Algorithm 2: Probabilistic Application Algorithm

Like the *k-means clustering* algorithm, the probabilistic application algorithm utilizes the solid angle and three-dimensional location of biologically relevant atoms. Algorithm 2 deviates from the previous algorithm by employing a minimum and maximum search radius to find neighboring atoms and to build strings of atoms. These atom strings are then converted to corresponding amino acids. This algorithm is illustrated in **Figure 3 – 6**.

Substrings of the amino acids are then generated for a given length of *kappa*, which is the desired number of amino acids. The substrings are generated for both the target and counter-target molecule. The substrings of amino acids for both the target and counter-target are combined into a single list of unique strings. Counts of the unique strings on the target and counter-target are then generated. The string counts for the target versus counter-target can then be plot or sorted to identify strings of amino acids uniquely present and accessible in the target and not in the counter-target. The solid angle of atoms

in each string is also provided to further clarity on the accessibility of the identified strings. The output of this option is a substring of amino acids and corresponding nucleotides assigned using the same method as Algorithm 1. This is option one of Algorithm 2.

The second option of Algorithm 2 builds upon the unique substrings and counts of the amino acid strings in the target and counter-target. Based on the size of kappa, all possible nucleotide combinations are generated. For example, if kappa is equal to 6, then 4^6 or 4,096 combinations are generated. For each possible nucleotide, a probability is assigned based on the specific amino acid in that position. The probabilities are derived from the measured frequencies of interaction. The product of the probability for each nucleotide combination is calculated. The probabilities for each unique string are then multiplied frequency of occurrence in the target and counter-target. This process is repeated for every unique amino acid string.

The output of this option is a probability of nucleotide sequences of length kappa interacting with both the target and counter-target molecule. This is done since the combination of amino acids to nucleotides is degenerative. A nucleotide sequence for a unique string on the counter-target could be the same as unique string in the target molecule. By determining the probabilities of the possible strings, we further delineate potential nucleotide combinations that can interact with the target molecule with greater affinity than the counter-target molecule.

The output of Algorithm 2 is a small core sequence of nucleotides that is then further optimized using the genetic optimization algorithm.

3.4.4 Initial Sequence Generation

Once either Algorithm 1 or 2 identifies the core nucleotide sequence, the remainder of the larger oligonucleotide must to be identified. To ensure the overall oligonucleotide maintains the desired structure of the core sequence (*e.g.*, unpaired) and is stable to maintain the structure under the application conditions, an optimization algorithm is employed. This process is highlighted as part of the overall digitalSELEX platform in **Figure 3 – 7**.

A genetic algorithm (GA) is an optimization approach inspired by the biological evolution process of survival of the fittest concept.[121] The GA is one of several metaheuristic algorithms, that includes particle swarm optimization (PSO) and ant colony optimization (ACO), that have been applied different fields such as economics, engineering, politics, and management.[122] The GA optimization was first proposed in 1992 by J.H. Holland[123] with the basic elements consisting of chromosome representation, fitness selection, and biological-inspired operators.[124]

The chromosomes typically take a binary string format and are considered points in the solution space. A fitness function is then assigned to each chromosome in the population.[125] The biological-inspired operators are selection, mutations, and crossover, illustrated in **Figure 3 – 8**. The selection chromosomes are sometimes referred to as elites indicating the highest fitness and are maintained in the subsequent generation for processing. In crossover, a random chromosome changes subsequences with another randomly selected chromosome to create off-spring. The mutation operator selects bits of a chromosome, since binary, and randomly flips it (*e.g.*, 0 → 1 or 1 → 0).[126,127]

The biological operators are used to repeatedly modify the population of individual solutions / chromosomes. At each step or generation, the GA selects individuals from the current population to be parents and uses them to produce offspring / children for the next generation. Over successive generations, the population "evolves" toward an optimal solution.[128]

One of the powerful aspects of the GA and other metaheuristic algorithms is the ability to impose guidelines / constraints on both the selection process and operators. These algorithms are subsequently suited for optimizing oligonucleotides around a designated sequence to achieve a specific endpoint.

3.4.5 Designated Sequence Sources

Regardless of whether the oligonucleotide core sequence is derived from Algorithm 1 or 2, the remaining part of the sequence still needs to be identified. The nucleotide sequence before the core sequence (prefix) and the sequence at the core sequence (suffix) directly impact structure of the core sequence. This process schematic is shown in **Figure 3 – 9**. Collectively, these sequences direct the overall three-dimensional structure of the oligonucleotide via complementary base pairing. As mentioned in Chapter 2, the length of the oligonucleotide can also affect the application of the oligonucleotide. Consequently, The genetic optimization algorithm is constrained to find the most stable aptamer sequence within the application / user-defined requirements.

3.4.5.1 Basic Constraints

The optimization employs the genetic algorithm function from the MATLAB 2021a global optimization toolbox.[129] There are several basic user-defined parameters that can be employed or modified. The first parameter is the population size. The population size default is set to 50 if the number of variables is less than or equal to 5, otherwise the default is 200. With a large population, the GA search is more thorough, and reduces the chance of that a local minimum found is not the global minimum.[130] The large the population size however increases the computational time. The global minimum for design purposes is the stability score of the oligonucleotide based on the secondary structure of the sequence.

The second basic parameter is the number of generations. The user can determine the number of generations required based on the population size and the number of variables. For example, if there are 6 variables (nucleotides locations) and 4 possible nucleotides, then there are 4096 possible sequence combinations. With a population size of 200, the GA will be able to explore all possible combinations in less than 21 generations.

Analogous to the number of generations is the number of stall generations. Since the GA randomly generations mutations and crossover, there is a possibility of identifying the global minimum early in the number of generations. If the score does not improve over a designated number of stall generations then the optimization is terminated. The user can set the maximum number of stall generations to be equal to the number of generations to ensure an exhaustive search but may not be necessary.

The fourth basic parameter is defining the ideal fitness of the offspring; essentially the global minimum or best possible score. The GA optimization in digitalSELEX is designed to identify the most stable sequence that meets the required application constraints. The stability is defined by the free energy of folding of the oligonucleotide and is a finite value based on the number of nucleotides in the sequence. However, for ease of implementation and to ensure all possible sequences are explored, setting the fitness value to negative infinity is also a viable option and method.

3.4.5.2 Unique Constraints

The digitalSELEX genetic algorithm employs several unique constraints to limit the population to sequences that conform to specific application characteristics. These characteristics include length, CG content, sequence limitations, and unpaired nucleotides. The length of the aptamer is determined by the user prescribed values of prefix and suffix with are the length / number of oligonucleotides before and after the core sequence. The Cytosine-Guanine (CG) content can be modulated by the user but should comprise over 40% of the total oligonucleotide population.[131,132] Some of the sequence limitations are no quad nucleotides (*e.g.*, GGGG).

The final applied constraint to the genetic algorithm is unpaired nucleotides. The unpaired nucleotides occur in two locations in the sequence. First, due to the desired application of the aptamers to be used on the G-FET, the number of unpaired nucleotides on the 5'-end of the aptamer are unpaired. This specific constraint enables the attachment of the linker molecule, 1-pyrenebutyric acid N-hydroxysuccinimide ester (PBASE), without disrupting nucleotide pairing that could alter the overall structure. The second unpaired requirement is direct to the core sequence. The user can define the percentage of

unpaired nucleotides in the core to balance stability with capacity for interaction with the target.

Violations of these constraints cause a penalty to be incurred during the optimization. The summation of the penalties is balanced against the Gibb's Free Energy of folding for the oligonucleotide sequence which is the indicator of aptamer stability.[133] The genetic algorithm optimizes to towards the best score which corresponds to the kcal per mole free energy of folding. The more negative the value then the more stable the structure.

3.5 POTENTIAL LIMITATIONS

There are a few potential limitations with the *Cold Start* module. The first limitation stems from the degradation of options between the number of amino acids (22) and nucleotides (4). There are more unique combinations of amino acids then nucleotides combinations. From a design perspective, the assignment of nucleotides in the core sequence for the target could be the same for a counter-target. While this phenomenon is present in every SELEX variation, it is particularly acute when designing nucleotide sequences for specific amino acid sequences / clusters when using Algorithm 1.

A second potential limitation results from the identification of accessible and biological relevant atoms. The atoms capable of binding in the target molecule are the same atoms capable of binding in the counter-target molecule. This limitation does not affect the affinity of the oligonucleotide for the target but is critical towards specificity. The counter-selection / Warm Start module was developed to mediate this limitation.

A third limitation is the selection of the target molecule protein databank file. Many proteins have structures under different confirmations (*e.g.*, native versus bound ligand). Using an accurate protein that represents the configuration of the molecule during the application phase is essential. The solid angle of the atoms can vary when the molecule configuration changes which alters the clustering process.

A final potential limitation is the stability of the oligonucleotides generated by the optimization algorithm. Stability of the structure does not equate to increased stable interactions with the target molecule, however non-stable structure will have low affinity. The *Warm Start* module introduces random perturbations which will help identify the best binding sequence at least *in silico*.

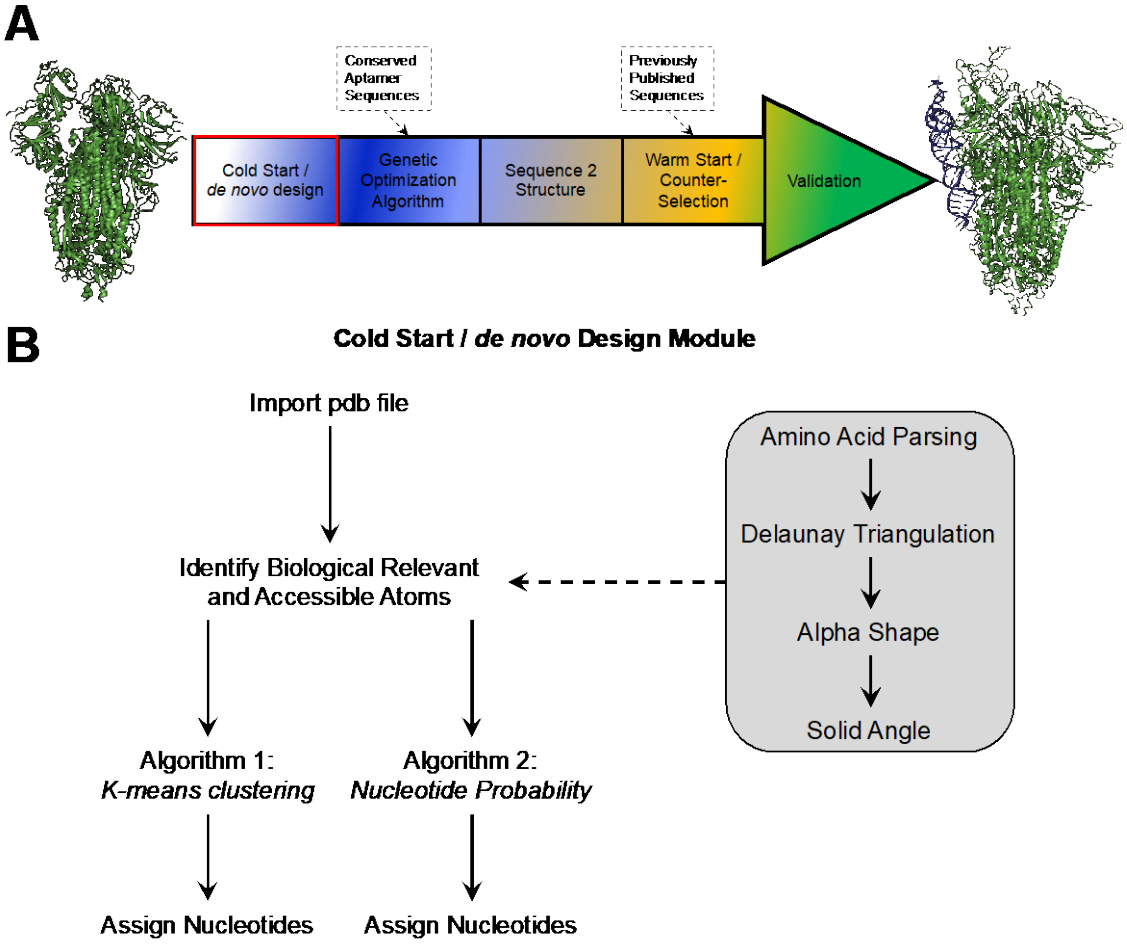


Figure 3-1. *Cold Start* Process.
A schematic showing the *Cold Start* module in digitalSELEX. This novel module initially identifies biological relevant and accessible atoms. The module then can employ two possible algorithms for clustering the amino acids. These algorithms either employ a k-means clustering or nucleotide probability. Finally, the module assigns nucleotides according to the clustering algorithm.

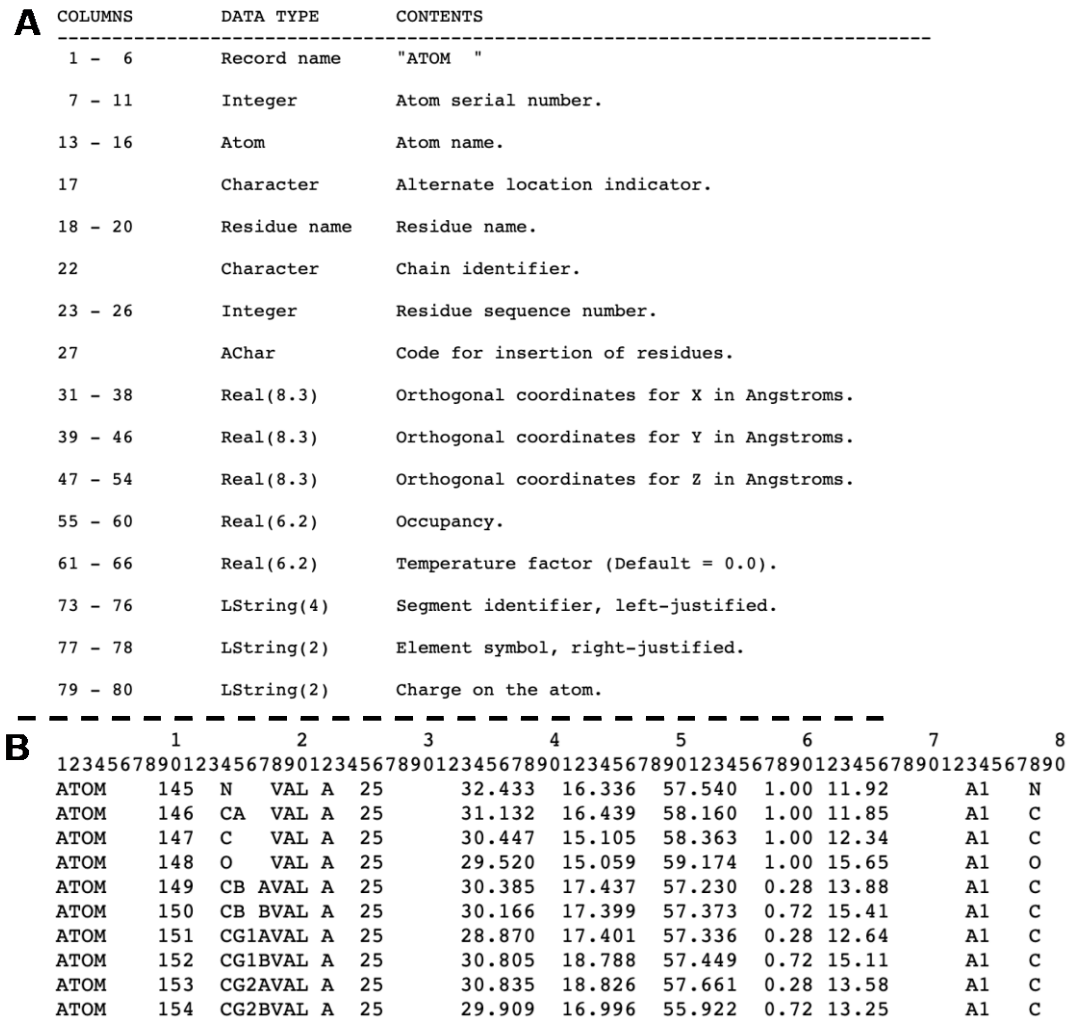


Figure 3-2. PDB File Structure.
Highlights the structural information in a protein databank file. Panel A indicates the columns of text file allotted to the specific data type and contents. Panel B provides an example of the structural data of a protein. This figure is adapted from the Zhang Group.

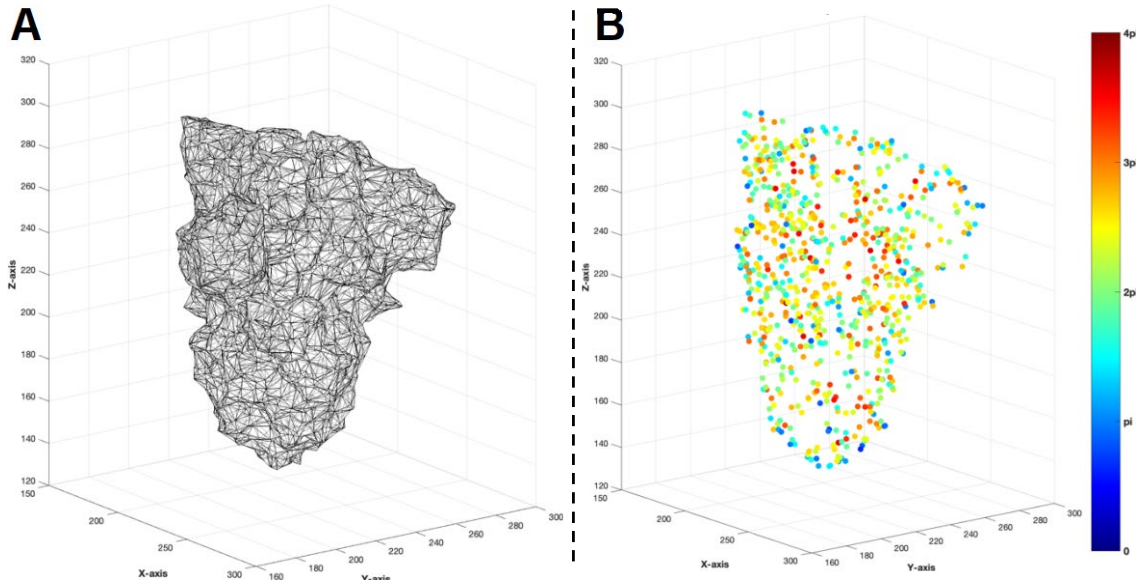


Figure 3-3. Delaunay Triangulation and Solid Angle.
Illustrates examples of the Delaunay Triangulation connections (panel A) and the Solid Angle of each atom (panel B) in the SARS-CoV-2 Receptor Binding Domain from PDB ID: 6vsb. The Solid Angle is calculated using the alpha shape function for the Delaunay triangulations connections. The highly accessible atoms in shown in blue and the color transitions to red showing in accessible atoms. The range of solid angle is from highly accessible (0-2pi) to inaccessible (4pi).

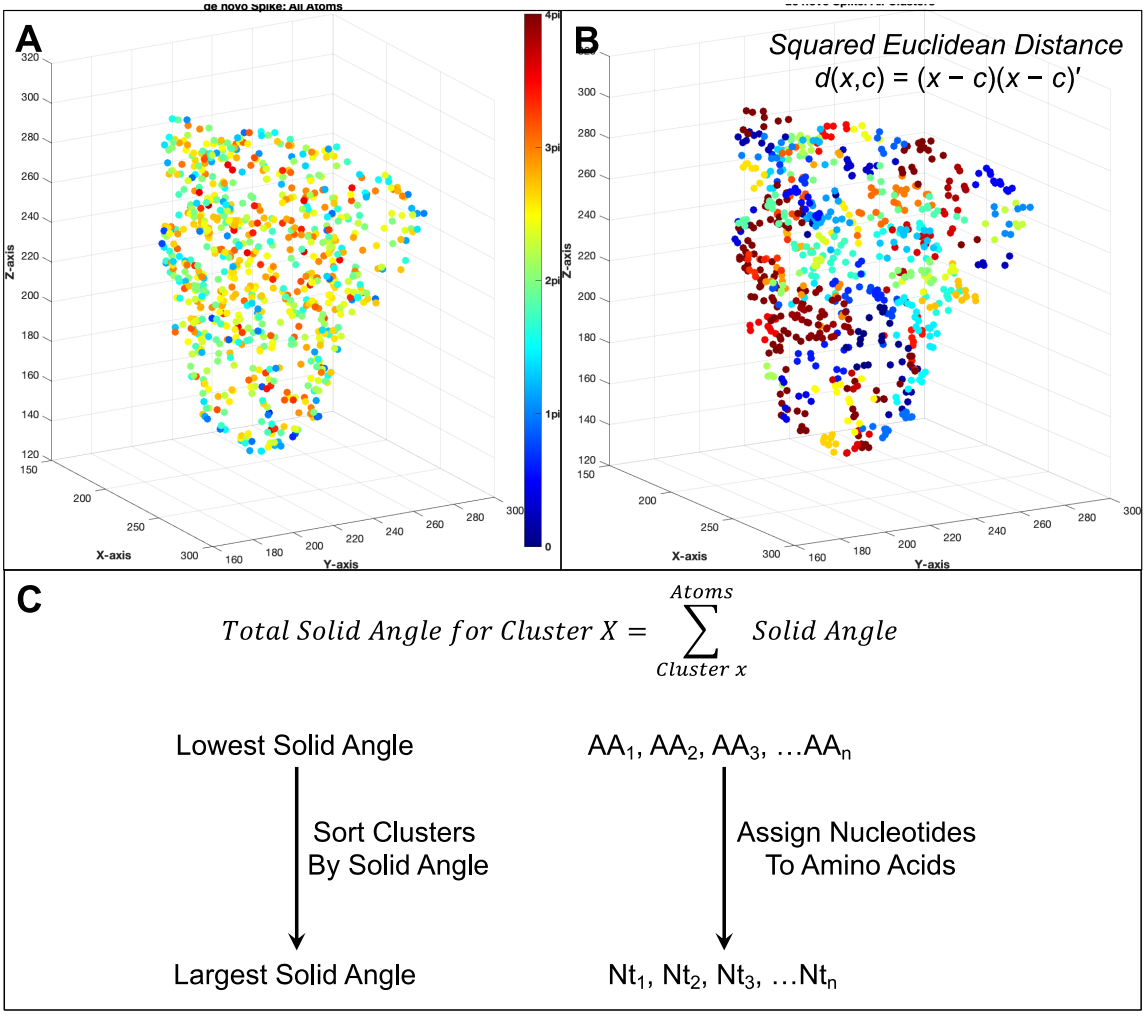


Figure 3-4. Algorithm 1 Process.
This figure highlights the process for Algorithm 1. Panel A shows indicates with a heat map the solid angle for every atom in the target molecule. The atoms are clustered using the k-means clustering by the squared Euclidean distance of the atoms. This output of clusters is color coded in panel B along with the k-means equation. Once clustered, the solid angle of every atom in the cluster is summed. The clusters are then sorted from lowest total solid angle to the largest. The amino acids of the clusters are assigned nucleotides. This process in panel C generates the core nucleotide sequence.

	Adenosine	Cytidine	Guanosine	Thymidine	Total

Figure 3-5. Nucleotide - Amino Acid Frequency.
Relative number of interactions between amino acid and nucleic acids with the percentage of interactions in parentheses. A total of 13,956 interactions were observed. The relative interaction information was used to create the core nucleotide assignment function for nucleotides to the biologically relevant and exposed atoms. This table is adopted from Hoffman *et al*, 2004.

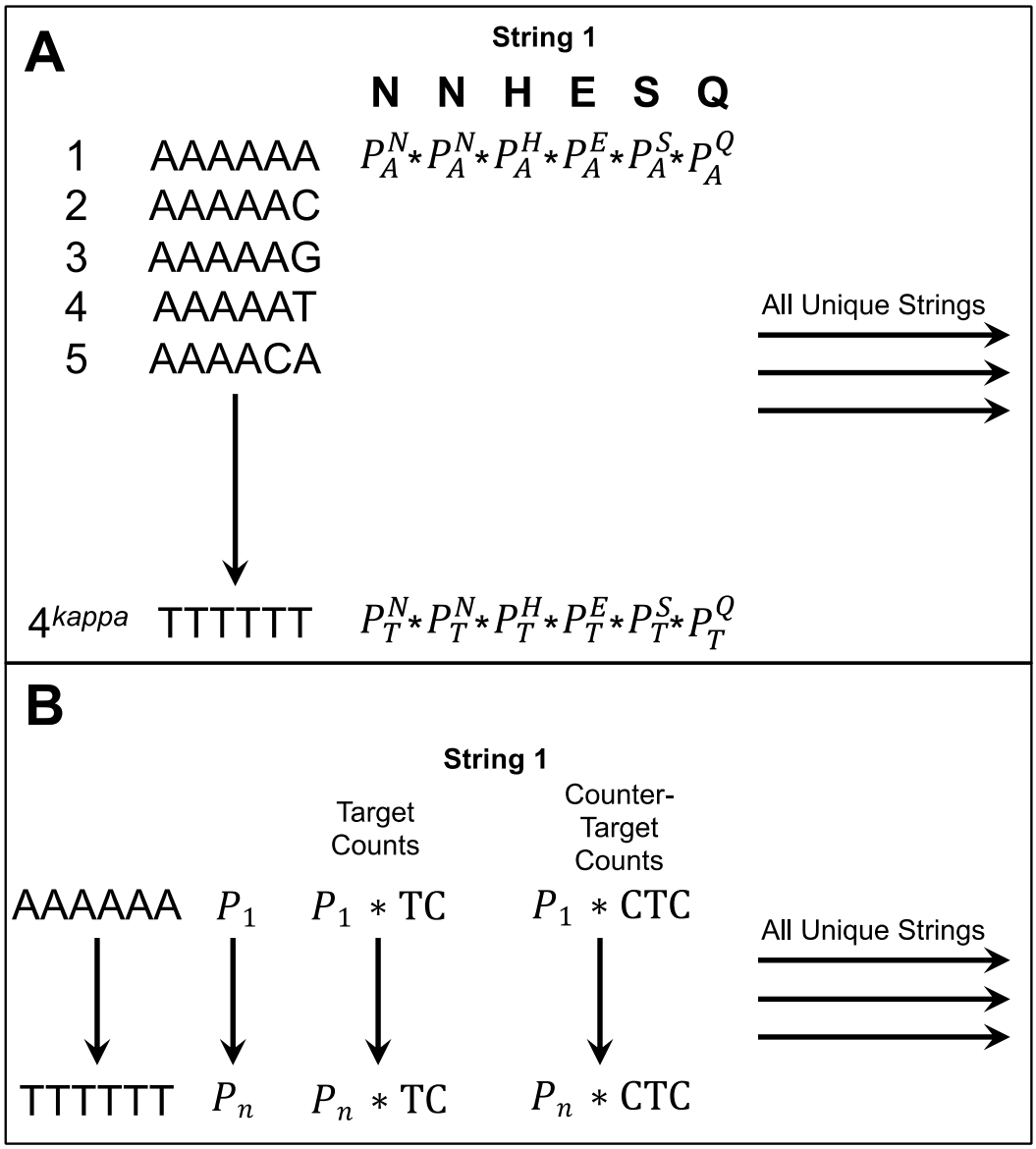


Figure 3-6. Algorithm 2 Process.
This figure details how the nucleotide probability using Algorithm 2 is determined. Panel A shows a string of amino acids of length *kappa*. All possible nucleotide combinations for length kappa are generated. The probability of each specific nucleotide in the combinations is assigned based on the specific amino acid. The probabilities within each sting are multiplied together. Panel B shows the next step in Algorithm 2. The probability corresponding to each nucleotide combination is multiplied by the counts in the target and counter-target of the string. The probabilities for each nucleotide combination for all strings are then summed. This provides a probability for the specific nucleotide strings in the target and counter-target. The nucleotide string that maximizes probability in the target and minimizes in the counter-target can be selected.

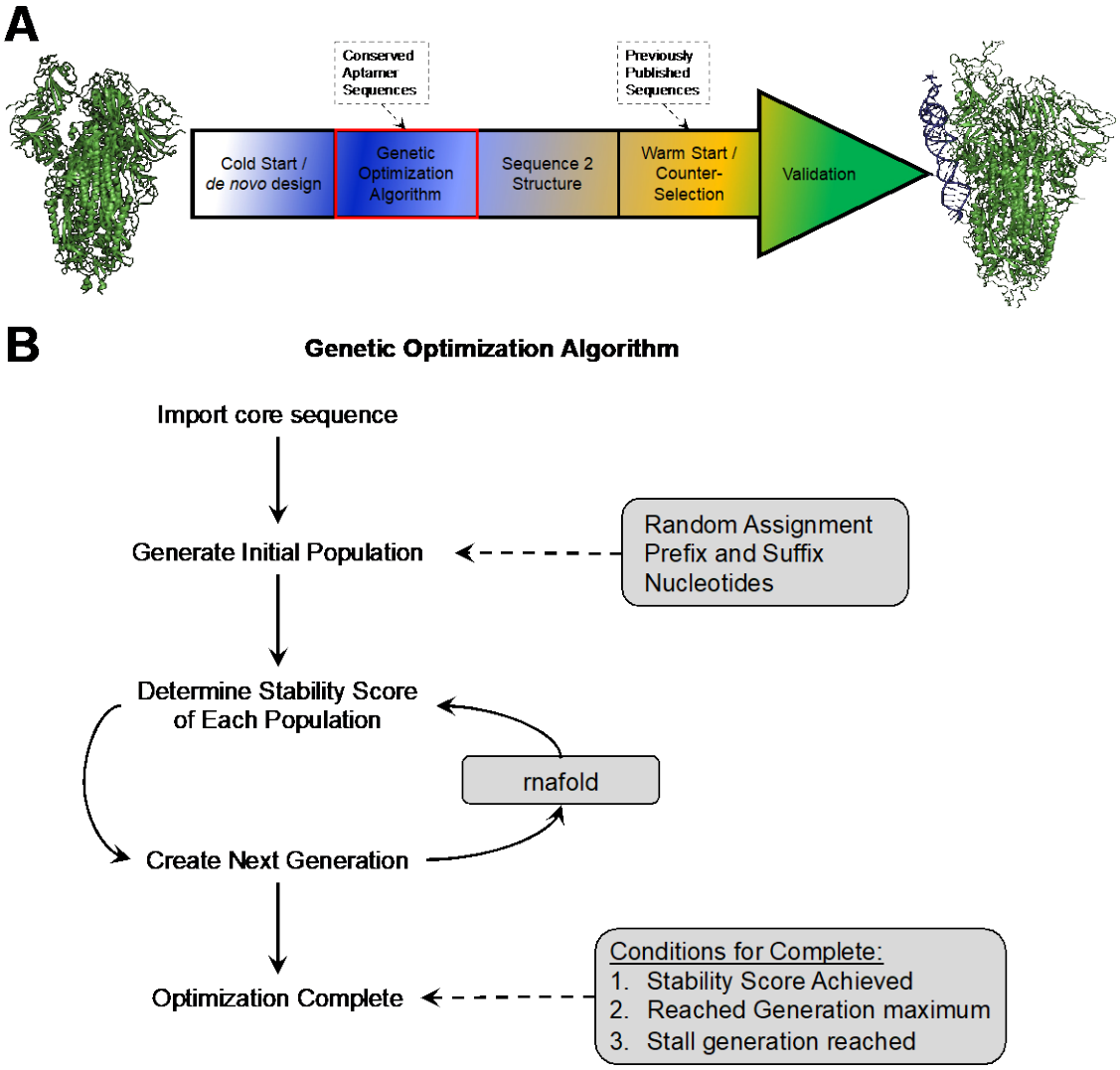


Figure 3-7. Genetic Optimization Algorithm Process.
A schematic showing the Genetic Optimization algorithm in the overall digitalSELEX process (panel A). While genetic algorithms are not novel, the incorporation of the rnafold function enables the algorithm to optimize around the stability of the oligonucleotide. The process in panel B can be utilized with and without constraints. The constrains allow optimization for specific applications such as a biosensor where a tail is required for binding. The generation of each generation comes from mutations, cross-over, and elites.

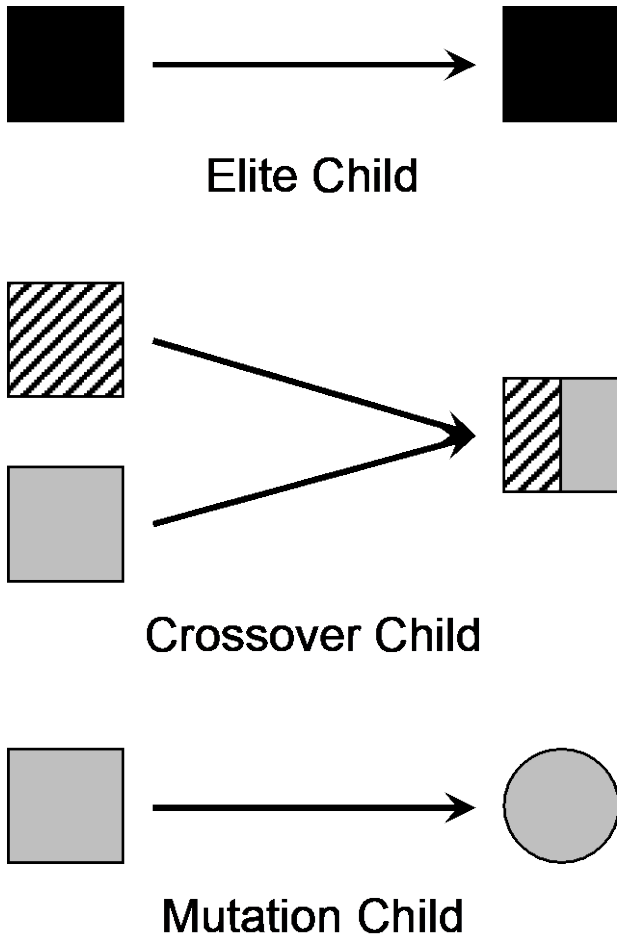


Figure 3-8. Genetic Algorithm Biological Operators.
The biological operators used in the genetic optimization algorithm are illustrated to demonstrate the potential effects on the offspring chromosomes in subsequent generations for the elites, crossover, and mutations. This figure is adapted from the MATHWORKS description of genetic algorithm.

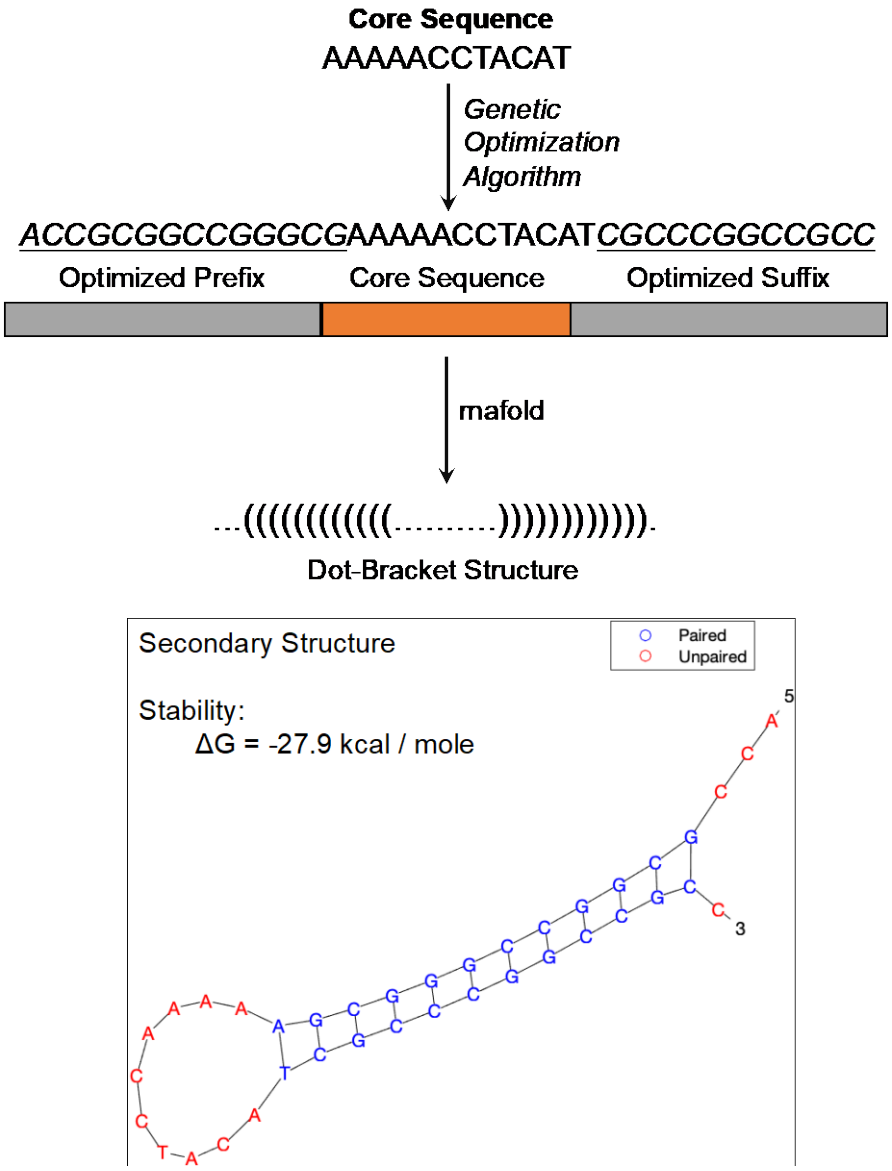


Figure 3-9. *Cold Start* Output.
This figure illustrates the output of the *Cold Start* process and the sequence stability optimization. The Cold Start process, using either algorithm 1 or 2, generates the core sequence. The prefix and suffix portions of the sequence then added to give the oligonucleotide the desired length. The nucleotides of both the prefix and suffix are optimized using the genetic optimization algorithm to generate the most stable structure under the specified constraints. The sequence can be visualized as a diagram or dot-bracket structure.

4.0 CHAPTER 4 – WARM START MODULE

This chapter explores the process of improving an initial oligonucleotide sequence. Improvement can be focused solely on affinity for the target, specificity, or both. Unlike the *Cold Start* process, this process is less agnostic about the details of the structure for both the protein and aptamer since the structure of both affects their interaction. To examine the effects on sequence and corresponding structure, we employ molecular simulators, specifically molecular dockers.

4.1 CENTRAL CONCEPT

The goal of the counter-selection module is to improve both the specificity and affinity of an oligonucleotide sequence via incremental improvements. This sequence can come from an existing aptamer or from the *de novo* design process. Despite the counter-selection nomenclature, this module consists of both *in silico* positive and negative selection steps. This module relies on molecular docking to test perturbed sequences. Sequences are converted into three-dimensional structures for molecular simulators to evaluate potential affinity and specificity.

The *Warm Start* module can be delineated into three sections. The first section is the generation of small changes, or mutations, to the oligonucleotide. The second section is the evaluation of the small changes to determine if the changes improved or worsened the binding of the oligonucleotide to the target molecule. This evaluation requires the

employment of molecular simulators. Finally, the third section is selection of mutations that improve oligonucleotide interaction during each iteration. The goal of this section is to find the potential global minimum of interaction between the oligonucleotide and target molecule as shown in **Figure 4 – 1**.

Conversion of a primary oligonucleotide sequence to a three-dimensional structure is essential to not only validate a *de novo* aptamer *in silico*, but to also improve existing aptamers. The process allows a user to mutate a sequence and generate the corresponding three-dimensional structure. The sequence to structure conversion occurs in two steps, shown in **Figure 4 – 2**. The first step uses the MATLAB function rnafold to identify the secondary structure of the oligonucleotide. The secondary structure is then converted into a three-dimensional structure using the webserver, RNAComposer. The sequence to structure module is a combination of both MATLAB and python scripts.

4.2 GENERATING SMALL CHANGES

Simultaneous random perturbation is a method for optimizing a system with multiple unknown parameters. The random perturbations are in essence sequence mutations. This process enables exploration of the potential interactions to find the global minimum. The potential energy of interaction is driven by the number of atoms, pairing of nucleotides, and the three-dimensional configuration of the atoms between the two molecules. Deliberate perturbations could similarly be used to achieve the same global minimum. The search across the global interaction energies would be difficult to find when little is known about how specific atoms are interacting.

Each oligonucleotide sequence perturbation is evaluated using a scoring function that evaluates the improvement in the interaction between the aptamer and target or counter-target molecule. There are multiple scoring options. Details of these scoring options are described in the implementation details section. The overall score is the score of the oligonucleotide interacting with the target minus the interaction score of the aptamer and counter-target. If there is no counter-target molecule, then the counter-target score is zero. There are multiple scoring options and details are described in the implantation details.

4.2.1 Evaluating Small Changes

Molecular dockers are employed to evaluate the potential binding between an oligonucleotide and target molecule. Molecular dockers are fast and semi-reliable tools that rely on the molecules structure. A molecular docker score cannot reveal a specific experimentally derived K_d-value. The scores however are a beneficial comparator to evaluate improving or worsening interactions between molecules.

In terms of identifying improved oligonucleotide binding for the target molecule versus a counter-target molecule, the general formula is the sum of the target partial scores minus the sum of the counter-target partial scores. These scores can be weighted depending on the influence of the size / number of atoms in the molecules influencing the scores.

4.2.2 Sequence to Structure

While perturbations in a sequence are easy to generate, the evaluation of small changes requires the structure of the molecules. The conversion of a primary oligonucleotide sequence to a three-dimensional structure is a two-step process requiring the MATLAB function rnafold and the website RNAComposer.

4.2.2.1 Secondary Structure Generation

The MATLAB rnafold function predicts the secondary structure of an RNA or ssDNA molecule by free energy minimization.[134] The MATLAB function is based on the establishment of a database in 1999 of 151,503 nucleotides in 955 structures. The function has a search algorithm to compare an input sequence to previously identified sequences and structural motifs.[135] The algorithm identifies the secondary structure with the most stable structure which has the minimum free energy of folding.[136] Similar structures, however, with slightly worse free energy of folding in the energy landscape can exist. These are akin to rotamers and means that multiple structures for the same sequence potentially exist.

The database has been updated since originally published in 1999 and is currently part of the RNAFOLD webserver which is maintained by Department of Theoretical Chemistry at the University of Vienna.[137]

The rnafold function in MATLAB was first introduced in version 2007b and accepts a string of characters using the single letter nucleotide identifiers: Adenine (A), Thymine (T), Cytosine (C), Guanine (G) and Uracil (U).[138] The structure in bracket format and the energy of folding are the two important outputs from the rnafold function.

The energy is given in kilocalories per mole (kcal / mole). The function denotes unpaired nucleotides by representing them with a dot and the paired nucleotides with a bracket. The direction of the bracket indicates the direction of the pairing. This process is visualized in **Figure 4 – 2**.

In necessary, the secondary structure can be further visualized using the MATLAB function rnaplot.[139] The visualization of the two-dimensional structure however is not required to generate the three-dimensional structure.

The sequence, the generated secondary structure in dot-bracket format, and the corresponding free energy of folding are written to a matrix. This matrix is exported as a comma separated value (csv) file used to generate the three-dimensional structure.

4.2.2.2 Tertiary Structure Generation

While the secondary structure provides the interaction between nucleotides, the tertiary structure provides the relative positioning of each atom in cartesian coordinates. The tertiary structure in a pdb file format is generated using RNAComposer.[140-141] This program applies the machine translation principle to relate RNA secondary structure and tertiary structure elements from the RNA FRABASE database. These structures can be generated either iteratively (one at a time) or in batches up to 10 sequences at a time.

The three-dimensional structures generated from RNAComposer depict the atoms in neutral conditions. Application conditions could be different and need to be considered since changes in the pH for example could lead to protonation and deprotonation of key atoms. RNAComposer also only adds the hydrogens required for internal folding. Some molecular dockers (*e.g.,* Autodock Vina) require all the hydrogens to be present for their internal scoring algorithms. While RNAComposer does not entirely protonate the

molecule, once the structures are generated the hydrogens can be added using programs such as PyMol and openbabel.

4.2.3 Molecular Docking

The concept of exploring potential alignments between ligands and receptors, which became known as molecular docking, was first explored by Kuntz *et al* in 1982.[142] Molecular dockers normally evaluate the relationship between biological macromolecules (*e.g.,* protein, DNA/RNA, peptide) and small molecules (*e.g.*, endogenous ligands or drugs). Dockers such as GOLD,[143] Surflex-Dock,[144] Autodock,[145] and GLIDE[146] are regularly used in structure-based drug discovery where a large library of chemical compounds is docked against a specific target.[147] This process is a first-order approach to reduce a library of several million compounds down to a few thousand potential compounds for further analysis.

The potential alignment between receptors and ligands is tested by changing the orientation of one molecule with respect to the other molecule, **Figure 4 – 3, panel A**. The change in the orientation of molecule, typically the ligand since it is smaller, varies from docker to docker. Some dockers require users to specify a binding box around the receptor to reduce the number of orientations required and reduce computational time.

A free energy of binding / interaction score is calculated at each orientation. For each atom, an energy score is determined using the Lennard-Jones plot for potential energy, illustrated in **Figure 4 – 3, panel B,** as a function of distance between atoms. The distance between a specific atom and all atoms within a specified radius is determined and the potential energy is summed based on the distances. This process considers the

repulsion forces with the atoms are too close and the London dispersion forces for the larger distances.

Dockers are estimation tools. Their output models and scores do not necessarily correspond with what is seen biologically. Molecular dockers are models and are subsequently only as good as the information inputted into the system. For example, some dockers require all molecules to be fully protonated, however, *in vivo*, and *in vitro* conditions have pH conditions that reduces the number of hydrogens on the molecules; altering the predicted affinity. Since dockers are a best approximation of interaction, and not absolute, it is difficult to directly estimate a K_d-value based on a molecular docking score.

Molecular dockers, however, can be used to estimate improving or worsening K_d-values, however this can be difficult at times based on the noise / error in the K_d-value determination, experimental design, and ensuring conditions are properly mimicked in both the molecular docker and *in vitro* experiment.

The output from dockers generally consists of the score, a root means square deviation (RMSD), and a protein data bank file showing the correspond configuration. Even though the docker searches the alignment between the ligand and receptor using multiple orientations typically only the top few models and data are output. Autodock, for example, only provides the top ten models while HDOCK webserver provides the top 100 models, and the stand alone HDOCK-lite version provides all 4392 generated models. In another example, the maximum number of models that can be generated by LightDock is equal to the number of "glow worms" that are employed in the glowworm swarm optimization.[148,149]

Additionally, dockers rank their models not on the likelihood of interaction occurrence, but by the score. The top model with the corresponding top score means that it is the model with the most atomic interactions between the ligand and the receptor. This model and score in all probability not feasible in either and *in vitro* or *in vivo* setting.

4.2.4 Molecular Dynamics

Another genre of molecular simulators is molecular dynamics. The two most prominent molecular dynamic simulators are GROMACS (GROningen MAchine for Chemical Simulations) and AMBER (Assisted Model Building with Energy Requirements). These tools are useful in identifying specific hydrogens bond formations in each region over time. These simulators are dependent on the input models specifically from a molecular docker. If the data from the docker is incorrect, then the data in the dynamic simulator will not accurately model the specific interactions. Additionally, these simulators are time consuming (~48 hours) unless a user specifies a small region of interest. Due to computational requirements and accurate input models, molecular dynamic simulators are not useful in the current counter-selection module.

4.2.5 Molecular Docker Selection

The latest round of the European Molecular Biology Laboratories (EMBL) CRitical Assessment of PRediction of Interaction (CAPRI) challenge had 34 different molecular docking programs / software.[150] To identify a molecular docker for the counter-selection software, there were several requirements that needed to be met. First,

the docker needed to be capable of predicting the interaction between proteins and nucleic acids. (Some dockers are only capable of protein-protein interactions or protein-chemical compound interactions.) Second, the docker should not have a limit on the number of atoms. For example, the QUASI-RNP dockers cannot handle molecules with more than 10,000 atoms which would immediately eliminate several targets of interests such as the SARS-CoV-2 Spike protein and the Hemagglutinin (HA) protein from Influenza. Third, the docker should provide as many possible configurations / predicted models as possible. Fourth, the docker should be relatively fast and require minimal modifications to generate the input files and /or extract the output data.

Based on the molecular docking criteria, HDOCK was selected for testing.[151,152] HDOCK is a capable of docking protein-protein and protein-nucleic acid interactions. This docker does not limit the number of atoms. Additionally, the docker does not stipulate which molecule must be the receptor or ligand. The input files are into HDOCK are the standard pdb format and while the molecules can be fully protonated it is not a requirement. HDOCK generates an output file that consists of the score, RMSD, and specific angles for both the receptor and ligand for each of the 4392 models tested which can then be used to generate the specific configuration pdb file. The total number of 4392 models is based on the specific rotation angles of 15 degrees and translation of 1.2 degrees in Euler coordinate space.[153] An example of this data is shown in **Figure 4 – 4**. Each simulation of the dockers takes approximately 45 minutes but can vary depending on the processor and the ability to parallelize the function.

As previously mentioned in Chapter 2, the correlation between the molecular docker, HDOCK, data and experimentally derived K_d-values was examined. While the

flow cytometry experiments focused on generating K_d-values for the different configurations, HDOCK was utilized to determine the top score and RMSD. The docking simulations for 10 aptamers against Spike, HA, and CRP proteins in both the aptamer on the bead and protein on the bead configurations generated a total of 60 simulations. The scores were rank ordered and then correlated against the rank order of the experimentally determined K_d-values using Spearman, Kendall, and Pearson algorithms. The results, previously illustrated in Chapter 2, showed that we could correlate the HDOCK scores to the K_d-values. As mentioned earlier in this Chapter, the limitation of molecular dockers is we cannot predict the exact K_d-value but get a relative understanding of improving or worsening values.

4.2.6 Selecting Mutations that Improve the Sequence

The *Warm Start* module generates multiple perturbed sequences during each iteration. More than one sequence can show improved potential binding with the target molecule and it because necessary to identify the best sequence. To evaluate the mutated sequences, a function was developed to examine how each mutation contributes to the improved score. Since there can be more than one mutated sequence, the mutations that improve the score the most are selected for an additional mutated sequence. This new sequence is then docked and scored using the same scoring function.

All the sequences are then evaluated and the sequence with the most improved score over the previous best sequence is selected. If no sequence is better than the previous, then no sequence from that iteration is selected. This is the failsafe of the sequence selection step.

4.3 NOVEL SECTIONS

There are three unique novel elements implemented in the *Warm Start* module. First, the implementation of simultaneous random perturbations to explore the potential energy of interaction between an oligonucleotide sequence and a target molecule is a novel component not previously explored in other SELEX methods. This process enables the identification of the global minimum of interaction.

The second novel element in this module is the robust scoring function; detailed in the implementation details section. A user can employ the scoring function to manipulate the potential interaction between the oligonucleotide and the target molecule. While most users will solely focus on improving affinity and specificity, the function enables a user to direct the interaction, or rather the level of contacts between the two molecules. A more negative value means more models have greater number of interacting atoms, while a less negative number means less interaction.

The third novel application is the process itself. The ability to perturb a character string, convert it to a three-dimensional structure, and then determine the potential level of interaction is also novel. While some previous *in silico* SELEX methods have built small aptamers (< 20 nucleotides) at a time, this method enables design of large nucleotides as well as the ability to improve existing ones.

4.4 IMPLEMENTATION DETAILS

The following provides details the implementation of the *Warm Start* module, **Figure 4 – 5**. A critical aspect of the module is the conversion of character string sequences into three dimensional structures. This process is then followed by the molecular docker is used to identify sequences that both improve with respect to the target molecule and worsen when interacting with the counter-target.

4.4.1 Module Options

Prior to initiating the *Warm Start* module, several options should be considered. These options include number of counter-targets, protonation of molecules, scoring functions, fixed portion of the sequence, number of sequences per iteration, and the number of mutations per sequence.

4.4.1.1 Number of Counter-Targets

The MATLAB code enables a user to identify multiple counter-selection target molecules. These molecules are identified to the MATLAB script via the path to the pdb files. Multiple counter-selection targets are written into a cell array of files. Each counter-selection target is docked against each sequence so multiple counter-selection targets can dramatically elongate the computational time required.

4.4.1.2 Protonation Option

If desired, the module employs PyMol to protonate the oligonucleotides and the target proteins. PyMol is an open-source program developed and maintained by the Schrodinger Institute. The program is generally used for visualization but can also be used to modify biological molecule files in a variety of ways. This program enables command line inputs for the modification of the files to include the addition of hydrogens. This option adds the hydrogens without changing file and takes less than a second per file to alter.

4.4.1.3 Scoring Function

There are 9 different scoring functions developed for the *Warm Start* module. The selection of a specific function can vary depending on the desired effect in the docking output. These functions are detailed later under the *Evaluating Small Changes* Section.

4.4.1.4 Fixed Sequence Portion

Depending on the desired goal, there may be an impetus to leave a section of the sequence unperturbed. For example, if a user is trying to improve binding of a known oligonucleotide to a target without modifying a core sequence. In the script the variable is labelled, FixedNucleotides, and a range of nucleotides is input (*e.g.*, (20:28)). If not no fixed position is desired, then the variable is left empty (*e.g.*, [ ]). The sequence range is based on the nucleotide positions numerically from the 5' to 3' – end of the sequence.

4.4.1.5 Number of Mutated Sequences

The number of mutated sequences, which can be considered the number of opportunities to explore the interaction space per iteration, is an important variable to consider. This option is identified in the script as the MaxInternalReps and is simply a numeric input. Since each perturbed sequence undergoes molecular docking, the limitation on the number of mutated sequences may be computational capacity.

4.4.1.6 Number of Mutations

Simultaneous random perturbation is an optimization method. This method is dependent on two factors. The first factor is the number of mutations or perturbations per sequence. This option is dependent on the sequence length as a short sequence may not require many perturbations to explore the potential energy of interaction between molecules. This option is identified in the script as MaxMutationRate and is numeric value. Deceasing or increasing this value will not alter the computational time per iteration but could affect the optimization towards finding the global minimum of interaction. Too few mutations and the success rate is slowed and too many mutations per iteration could potential cause the algorithm to jump across the interaction space too much.

4.4.2 Small Sequence Change Implementation

An oligonucleotide sequence is typically expressed in its letter string of A, C, G, and T (U). To make perturbations / mutations to the sequence, the letters are converted to a numeric string one through four representing the alphabetical order of the nucleotides.

Once in a numeric string, the location of the random mutations is randomly identified within the length of the string using a random number generator for the given number of mutations. Once the locations are identified, then a random number between one and four, but not the current number, is selected for that position. Once all numeric mutations are made, the numeric string is converted back to the oligonucleotide letters. The sequence then needs to convert to a secondary structure and then tertiary structure for molecular docking.

4.4.2.1 Sequence to Structure

Each initial or perturbed sequence is converted to its three-dimensional structure via a two-step process. First, the letter string is input into MATLAB's rnafold function. While this function can generate several data points about the sequence including the stability energy, the most critical for generating the structure is the dot-bracket structure. This structure identifies how the nucleotides are paired or unpaired. This process is shown in **Figure 4 – 2**. The sequence are sequence and dot-bracket structure are written to a csv file.

RNAComposer, which generates the three-dimensional structures, is a website that does not have an Application Programming Interface (API). Therefore, the selenium library in Python coupled with both Chrome driver and Chrome browser were used to automate this step. A python script, written by Boston College Undergraduate Qingwei Meng, logs into the RNAComposer website, uploads the sequences and structure as a batch file, and then downloads the three-dimensional structure files when complete. The script generates a random 6 letter name for each oligonucleotide. The username and password for RNAComposer are input arguments; enabling a user to submit multiple

batches of sequences and secondary structures without interfering other submissions. An example of this process is shown in **Figure 4 – 6**.

The construction of the three-dimensional structures in usable protein data bank (pdb) file format can take several minutes. Every 30 seconds, the python script re-logs into RNAComposer and searches the downloadable files in the user workspace. When complete, the python script initiates the download of the files and extracts the pdb files. Each file is identifiable by the random generated name. An example molecule generated by RNAComposer is depicted in **Figure 4 – 7.** The files are saved in a folder labelled, "aptamers" and subsequently extracted by the counter-selection module for molecular docking.

4.4.3 Evaluating Small Changes

This procedure details how the molecular docker is employed for exploring the potential free energy of interaction between an oligonucleotide and either the target (positive selection) and/or a non-target (negative selection). The code for this process is written in MATLAB however there are system calls for using the molecular docker as well as implementing the python code for the conversion of perturbed sequences during each iteration.

The input sequence can be derived from the *de novo* design process or user specified from a previously published study or *in vitro* SELEX. The initial sequence is character string and can represent either single-stranded DNA or RNA. The target or non-target, depending on the desired outcome, is also specified in the protein databank (pdb) file format.

There are several facets of the module that a user can regulate depending on the desired application of the oligonucleotide. These characteristics discussed in the Options section include protonation of the both the oligonucleotide and the protein, the scoring function, and the number of perturbations.

The initial sequence is converted to its three-dimensional structure using the sequence to structure module generating a pdb file. The interaction between the oligonucleotide structure against the target and counter-target molecules is then explored using the molecular docker, HDOCK.[154,155,156,157,158] This first docking simulation generates the baseline level of potential interaction. HDOCK generates 4392 scores and RMSD values per docking simulation. A typical, HDOCK molecular docking simulations takes approximately 45 minutes and multiple simulations can be run in parallel when utilizing multiple cores.

The character string sequence then undergoes simultaneous random perturbation / mutations. The number of mutations per sequence and the number of sequences per interaction can be specified. Each sequence is then converted to its three-dimensional structure in the pdb file. Each of the perturbed sequence is docked against the target and counter-target molecules. The score of interaction for each sequence is calculated using the user desired method listed in the scoring function.

Once the scores for each sequence against the target and counter-target molecule have been determine, the module identifies all the sequences where the interaction scoring was better than the baseline score. The improved score of that specific sequence is then divided by the number of mutations. Since the perturbations are random more than one sequence may have a mutation at the same nucleotide location. The attributed score

for a specific mutation is then summed together. The algorithm then generates a new sequence based on the previous perturbations that generated the best scores. This new sequence is also docked against the target and counter-target molecules.

The module then identifies the sequence with the best interaction based on the selected scoring function. If no, sequence provides an improved interaction relative to the baseline (previous best sequence) then the baseline sequence is maintained. If the interaction is better than the baseline sequence, then that sequence is selected and becomes the new baseline for the next iteration.

Improving the potential interaction between the oligonucleotide and the target molecule is an iterative process. There may be iterations, particularly during the initial rounds, where new sequences are selected every iteration. As the process advances, it becomes more and more difficult to identify improved sequences, so the baseline becomes more stable. Changes to the baseline sequence then become more difficult.

The sequence character string, scores, and docking data for each sequence is saved following every iteration. The sequence evolution from its initial interactions to the new energy minimum can be examined.

4.4.3.1 Scoring Function

There are 9 different scoring methodologies (TopScore, TopPercentile, ScoreCount, TopScoreMixedWithRMSD, Median, Range, Probability, EnergyAtMaxProbability, and AreaUnderTheCurve) that can be selected and modified depending on the desired effects on the molecular docking output distribution. These scoring functions are graphically depicted in **Figure 4 – 8**.

The TopScore method utilizes the top score of the 4392 possible scores and models generated by HDOCK. It refers to the top score because HDOCK organizes the scores and models by most negative free energy of interaction score to the most positive. This model with the top score would be indicative of the configuration between the two molecules with the most points of interaction.

The TopPercentile score allows a user to select a specific percentile value and that score to be used. For example, to select the top 10^{th} percentile score then the 439 value of the 4392 total values will be used in the comparative process.

The ScoreCount function sums all the models with a score better than a specific value. This allows the module to select mutated sequences with scores beyond a specific threshold. This ensures there are a greater number of interactions between the oligonucleotide and the target protein.

The TopScoreMixedWithRMSD generates a weighted score between the free energy of interaction score and the RMSD. The score used in this process is the top score which corresponds to the model with the most interactions between the oligonucleotide and the protein. The RMSD for the model is the summation of the deviation of the ligand to fit with the model. The lower the RMSD value then the lower the strain on the molecule to interact with the receptor. The weights are preset with the score being 0.8 and the RMSD being 0.2, however this could be adjusted in the scoring function if necessary.

Like both the TopPercentile and TopScore methods, the Median uses the median score as the marker for selecting sequences.

The Range method allows a user to select a range of scores with the goal being to increase the number of interactions / models within that specified scoring range. This scoring method is designed on the concept that the top model will not be achievable in a biological context. The most likely interaction between the receptor and the ligand is a partial binding, hence the user can identify the range to shift more of the models towards.

The remaining three scoring functions, Probability, EnergyAtMaxProbability, and the AreaUnderTheCurve, employ the histcounts function in MATLAB.[159] This function uses the probability normalization option with auto binning. This generates a distribution based on relative probability such that the sum of all bins is equal to 1. The auto bin option is selected to identify the minimum number of bins for the underlying distribution. Consequently, if the models generate scores relatively close together few bins are required compared to models that are more dispersed. For the Probability scoring option, the score is set to select the maximum probability of the distribution.

The EnergyAtTheMaxProbability option, identified the potential energy interaction at the point of maximum probability.

The AreaUnderTheCurve is a combination of the previous two options. First, this function sums the probability of the distribution at the max as well as the two bin points on either side of the maximum probability. This sum is then multiplied by the energy at the maximum probability. Here the goal is to not only improve the energy of binding but also ensure there are many models around the maximum point.

4.4.4 Final Sequence Selection

Following the evaluation of each perturbed sequence it is necessary to identify a best sequence in each round. The best sequence in each iteration is one with improved binding towards the target and worse binding for the counter-target.

To choose the best sequence, an additional mutated sequence is generated based on all the evaluated sequences that did better than the previous round sequence. For each sequence, the difference in score between the previous sequence and better sequence is determined. This value is then divided by the number of mutations and assigned to the mutations. This process is depicted in **Figure 4 – 9**.

Since multiple sequences could have improved scores and with the same random mutations in the same location, the scores attributed to each mutation are summed for the location in the sequence. The top mutations are accepted, and the new sequence is generated.

The new sequence is then converted into the three-dimensional structure and docked against the target and counter-target molecules. Using the same scoring function, the score of the combined mutation sequence is determined. The sequence with the most improved score for the target and worse for the counter-target across all mutated sequences is selected. If no perturbed sequence demonstrates improved potential binding, then the previous sequence is maintained for the next iteration.

4.4.5 Sequence Extraction

Each iteration of the *Warm Start* module generates multiple perturbed sequences. Some sequences lead to improved characteristics against the target and counter-target molecule. At the end of each round, the sequence and molecular docking data against the target and counter-target is written to an output file. This data is parsed with a simple MATLAB script that identifies changes in the baseline. This script not only extracts the sequences but generates a probability distribution versus potential energy of binding plot. An example distribution plot of this data binding both the target and counter-target along with the corresponding sequences is shown in **Figure 4 – 10**.

The few sequences identified through this data extraction method, compared to the total number of sequences examined, can be chemically synthesized, and validated against the target and counter-target molecules in the *in vitro* validation process described in Chapter 2.

4.5 POTENTIAL LIMITATIONS

There are several potential limitations to *Warm Start* module. The first limitation resides in the three-dimensional structures used in the molecular docker. The structure of the oligonucleotide is generated using RNAComposer. The accuracy of the three-dimensional oligonucleotide structures continues to be debated. This structure uncertainty means there is uncertainty in the molecular docker predictions.

The second structural limitation comes from the selection of the target molecule protein databank file. The method (*e.g.,* pH, bound ligand and such) for generating the protein databank file can influence the generated structure. The better the resolution and native configuration of the molecule then the better the molecular docker predictions.

The next limitation is that molecular dockers are prediction tools and do not provide absolutes. Consequently, there may be noise in the output where minor variations in sequence are indistinguishable in the molecular docker scores.

The last potential limitation is computational time. To explore the entire potential energy of interaction space between oligonucleotides and a molecular target requires large computational capacity. This limitation is easily overcome when employing multiple cores and servers, however it is a resource requirement.

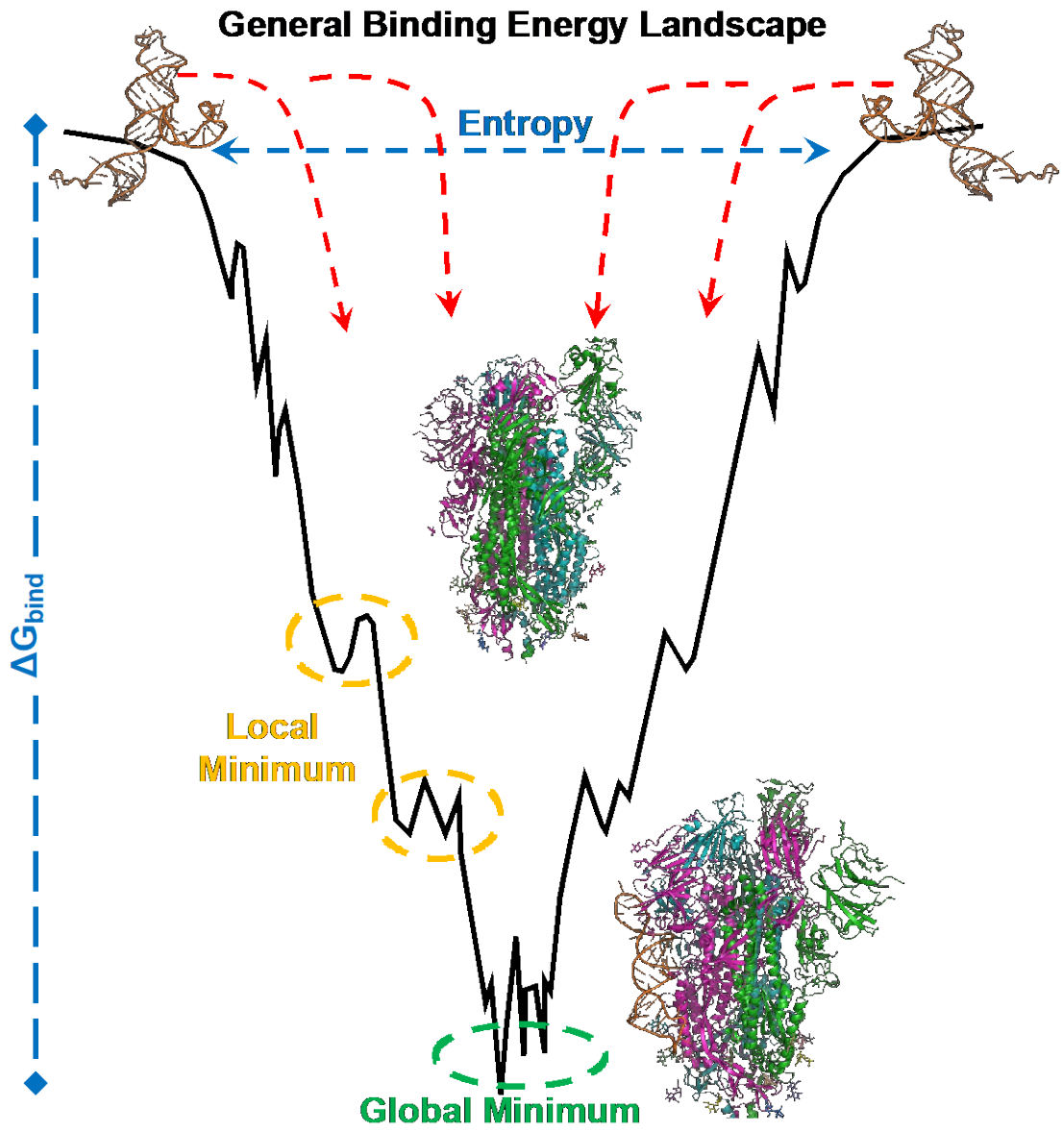


Figure 4-1. Aptamer-Target Binding Energy Landscape.
Illustrates the possible energy interactions between an oligonucleotide and target molecule. The energy level of interaction between protein and oligonucleotide is dependent on the orientation of the interaction between the two molecules.

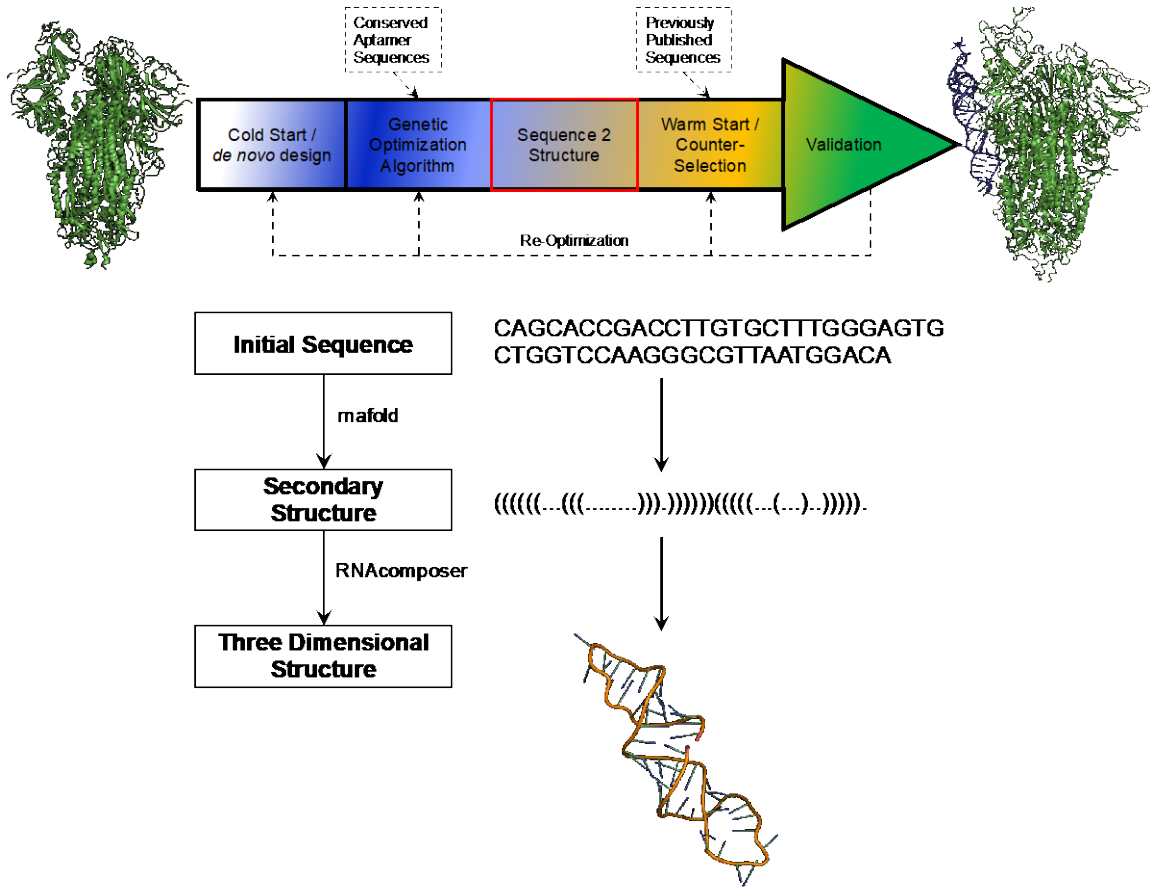


Figure 4-2. Sequence to Structure Process.
Illustrates the Sequence to Structure process. A sequence is converted to its secondary structure using rnafold in MATLAB. The secondary structure is shown as a dot-bracket structure. The sequence and dot-bracket structure are uploaded to the webserver RNAComposer using a python script. This same script also downloads the predicted three-dimensional structure.

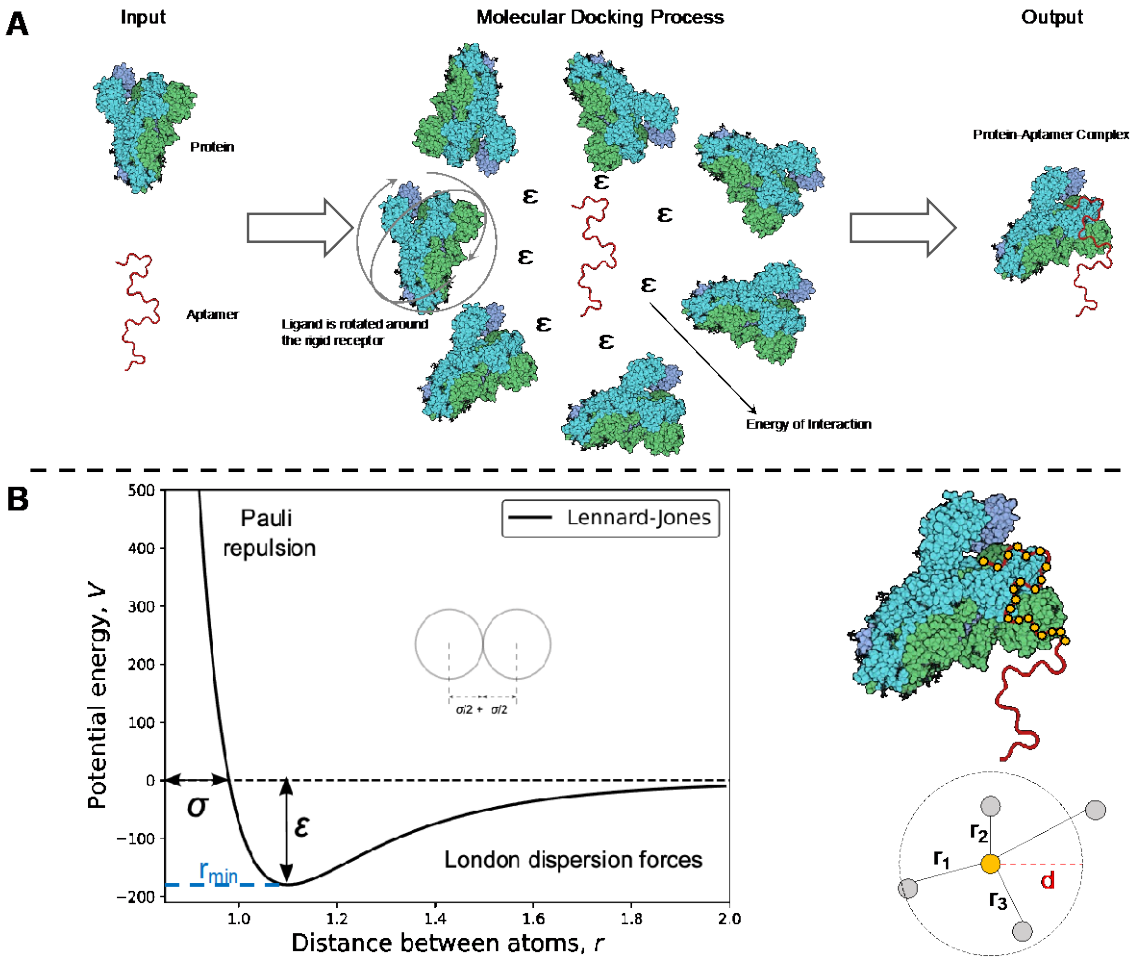


Figure 4-3. Molecular Docking Methodology.
Panel A illustrates the basic methodology of a molecular docker. The input consists of two molecules a receptor and ligand. In the figure, the aptamer is depicted as the receptor and the protein is the ligand. This configuration was chosen due to our downstream goal of using the aptamers on biosensor. The protein is rotated on multiple axis around the oligonucleotide. At each rotation leads to a different configuration between the nucleotides and the protein and a corresponding score. Panel B depicts the use of a Lennard-Jones plot to show the relationship between the potential energy of interaction and the distance between the atoms. The orange atom in panel B represents on atom on the receptor and the grey atoms within a preset radius (d). The potential energy between the orange atom and each grey atom is determined using the Lennard-Jones plot and summed. This process is repeats for each atom within a given distance between the receptor and ligand. The better the free energy score of corresponds to the more interactions between the two molecules.

All Information

Scores	RMSD	Angle Data for Model Generation					
1.0e+03 *							
-1.1753	0.1026	0.0005	0.0024	0.0045	-0.0568	-0.0470	-0.0501
-1.1692	0.1267	0.0001	0.0005	0.0025	-0.0393	0.1107	-0.0165
-1.1619	0.0900	0.0043	0.0009	0.0018	-0.0372	-0.0788	0.0052
-1.1251	0.0964	0.0007	0.0025	0.0048	-0.0507	-0.0457	-0.0464
-1.1173	0.0964	0.0007	0.0025	0.0047	-0.0546	-0.0438	-0.0455
-1.1161	0.0997	0.0005	0.0025	0.0047	-0.0507	-0.0477	-0.0493
-1.1099	0.0910	0.0043	0.0010	0.0019	-0.0354	-0.0806	0.0023
-1.1021	0.1029	0.0005	0.0025	0.0045	-0.0546	-0.0486	-0.0509
-1.0969	0.1164	0.0030	0.0012	0.0055	-0.0638	-0.0967	0.0240
-1.0920	0.0919	0.0043	0.0009	0.0019	-0.0381	-0.0811	0.0055

nrows = 4392

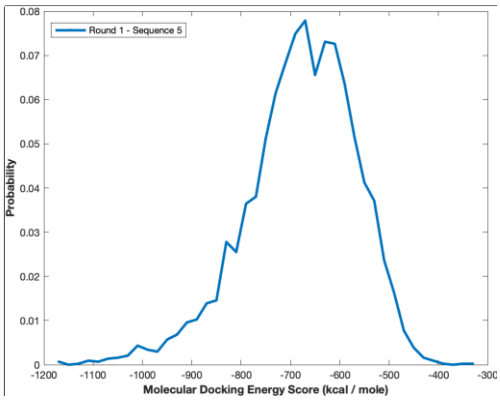


Figure 4-4. HDOCK Output Format.
HDOCK generates 4392 configurations with each configuration having a score, RMSD, and angle data for model generation. The score is the potential energy of interaction in kcal / mole. The scores can be plotted using the histcounts function with the normalized probability to bin the scores for a probability versus potential energy plot.

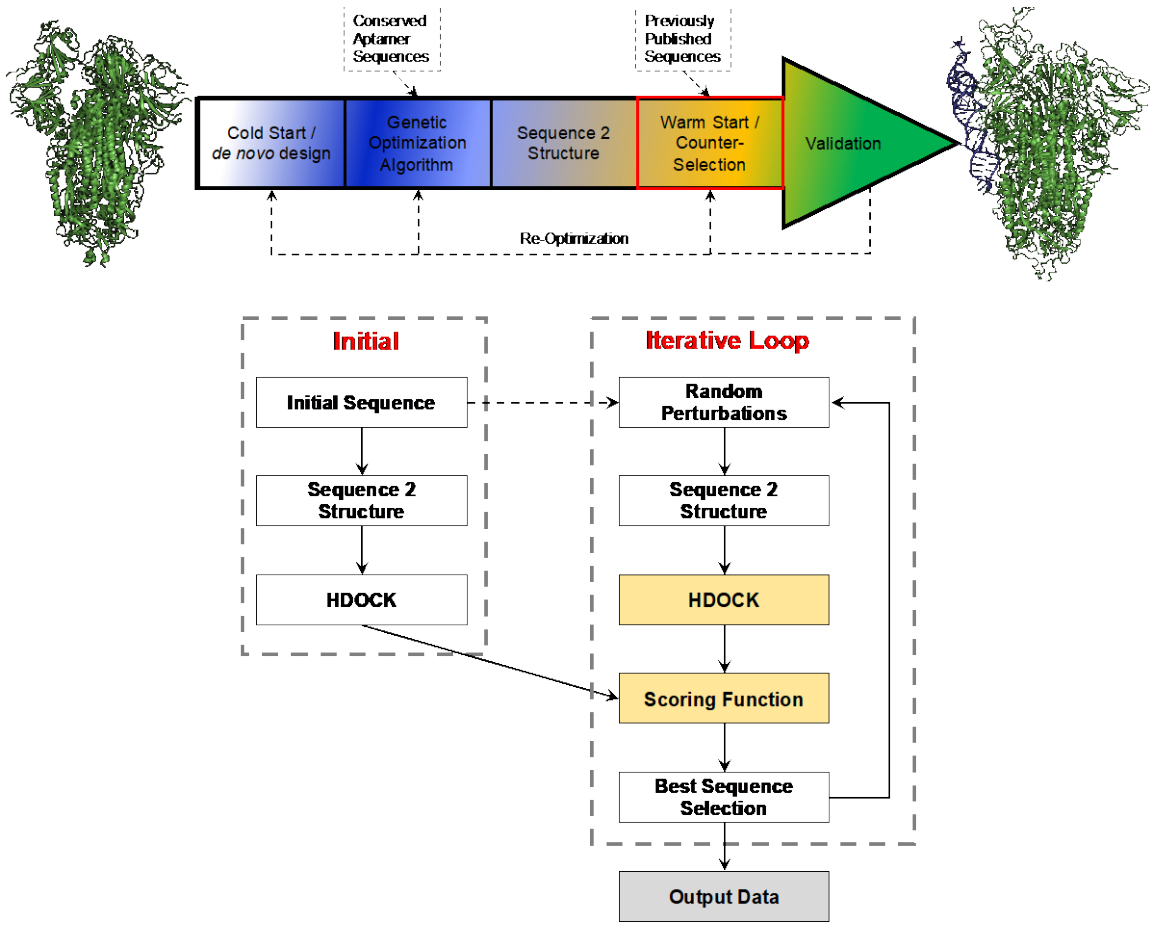


Figure 4-5. *Warm Start* Process.
The schematic shows the overall *Warm Start* module. The input or initial sequence is converted to its three-dimensional structure and then docked using HDOCK. The initial sequence then undergoes a series of random mutations. These generated sequences are then converted to their three-dimensional structure. These structures are docked and scored which is run in parallel using the parallel computing toolbox in MATLAB. These two parallel processes are shown in light orange. When the series of sequences is complete, the scores for each sequence against the target is analyzed selecting the best sequence which may be the initial sequence or the previously best selected sequence. At the end of each round, the data and workspace are saved. There are multiple scoring functions which are discussed in the Option section.

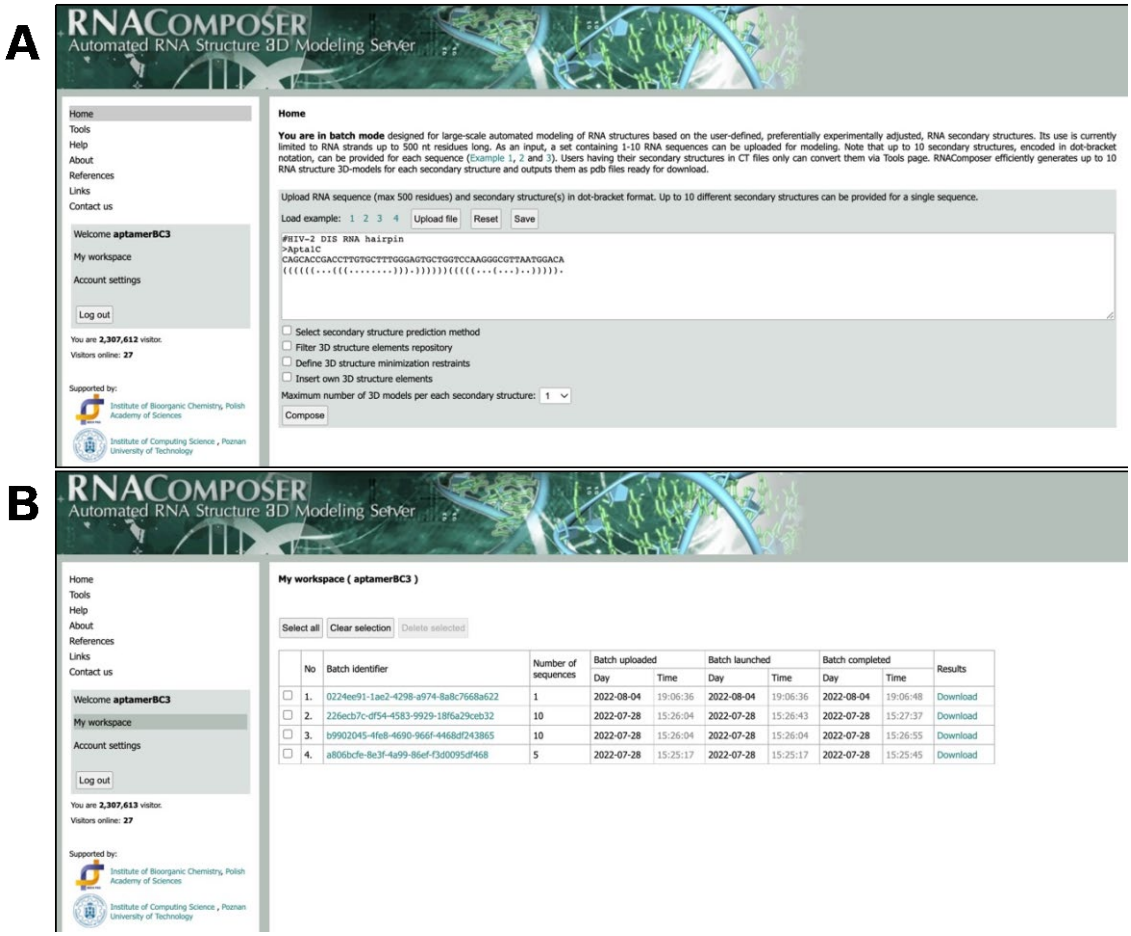


Figure 4-6. RNAComposer Submission.
Depicts the submission to RNAComposer of the sequence and secondary structure to generate the three-dimensional pdb file. Panel A shows the submission of the Aptamer 1C in the batch submission portion of the workspace. Panel B shows the download section of the workspace where the python script identifies the files to be downloaded.

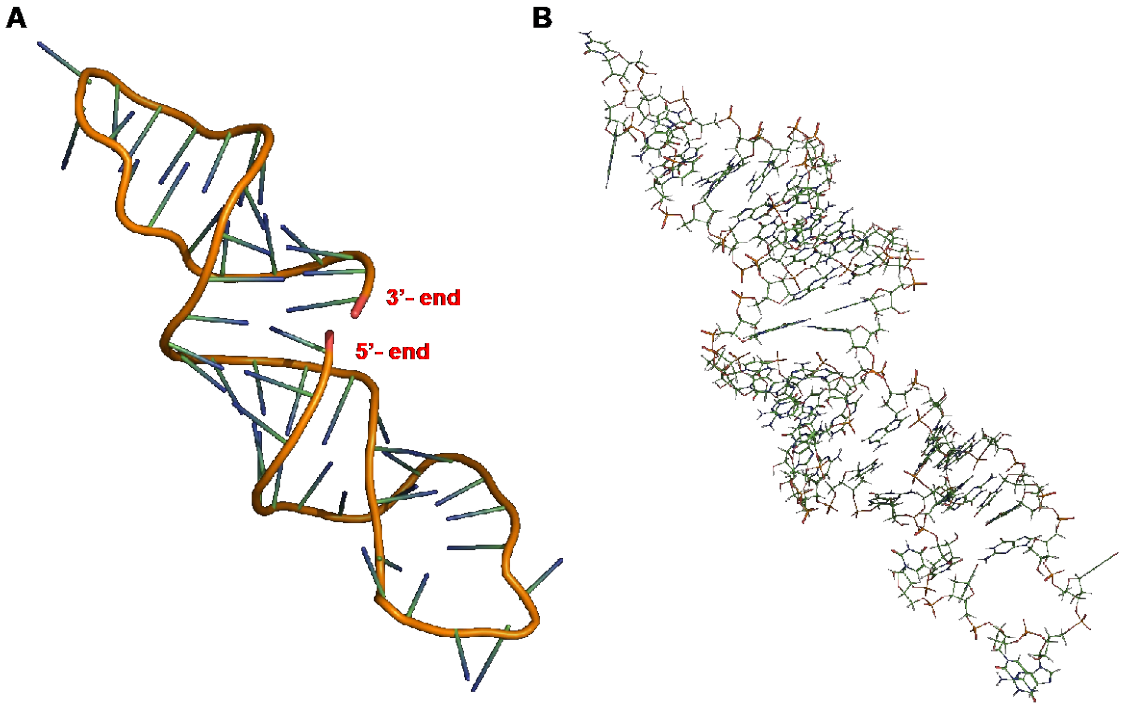


Figure 4-7. RNAComposer Strucutre Example.
The three-dimensional structure of Aptamer 1C generated using RNAComposer. Panel A shows the PyMol rendered cartoon structure with the 5' – and 3' – end. Panel B is a PyMol render of the structure showing the atoms and their orientation according to the RNAComposer created pdb file.

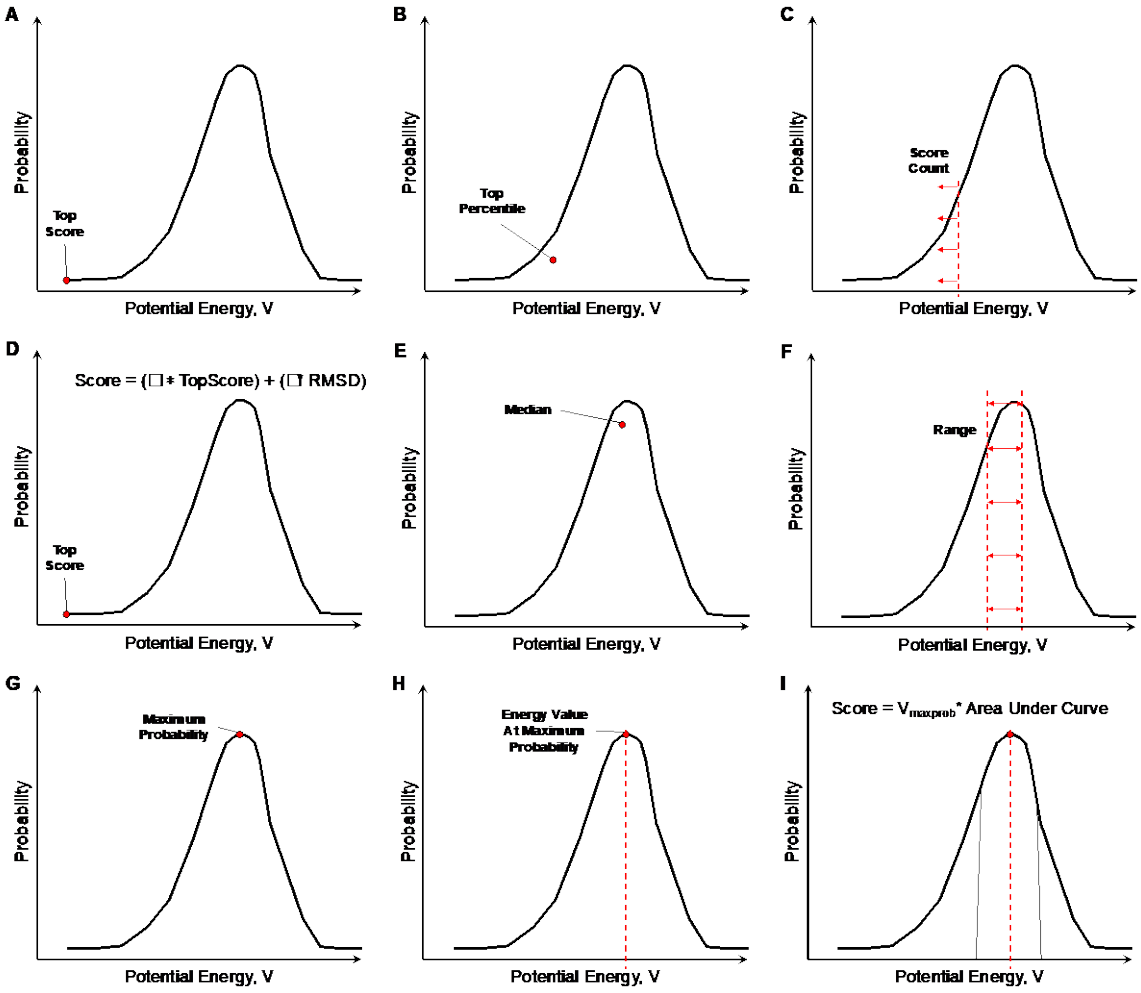


Figure 4-8. Scoring Function Options.
The scoring function options are depicted in panels A – I. The scoring options are include using the Top Score (A), the Top Percentile Score (B), the Score Count (C), the combination of the Top Score with RMSD (D), the Median Score (E), Range (F), the maximum probability (G), the energy value at the maximum probability (H), and the area under the curve times the energy value at the maximum probability (I). Score count identifies sequences that increase the number of scores beyond a designated threshold. The range function maximizes the number scores between two values.

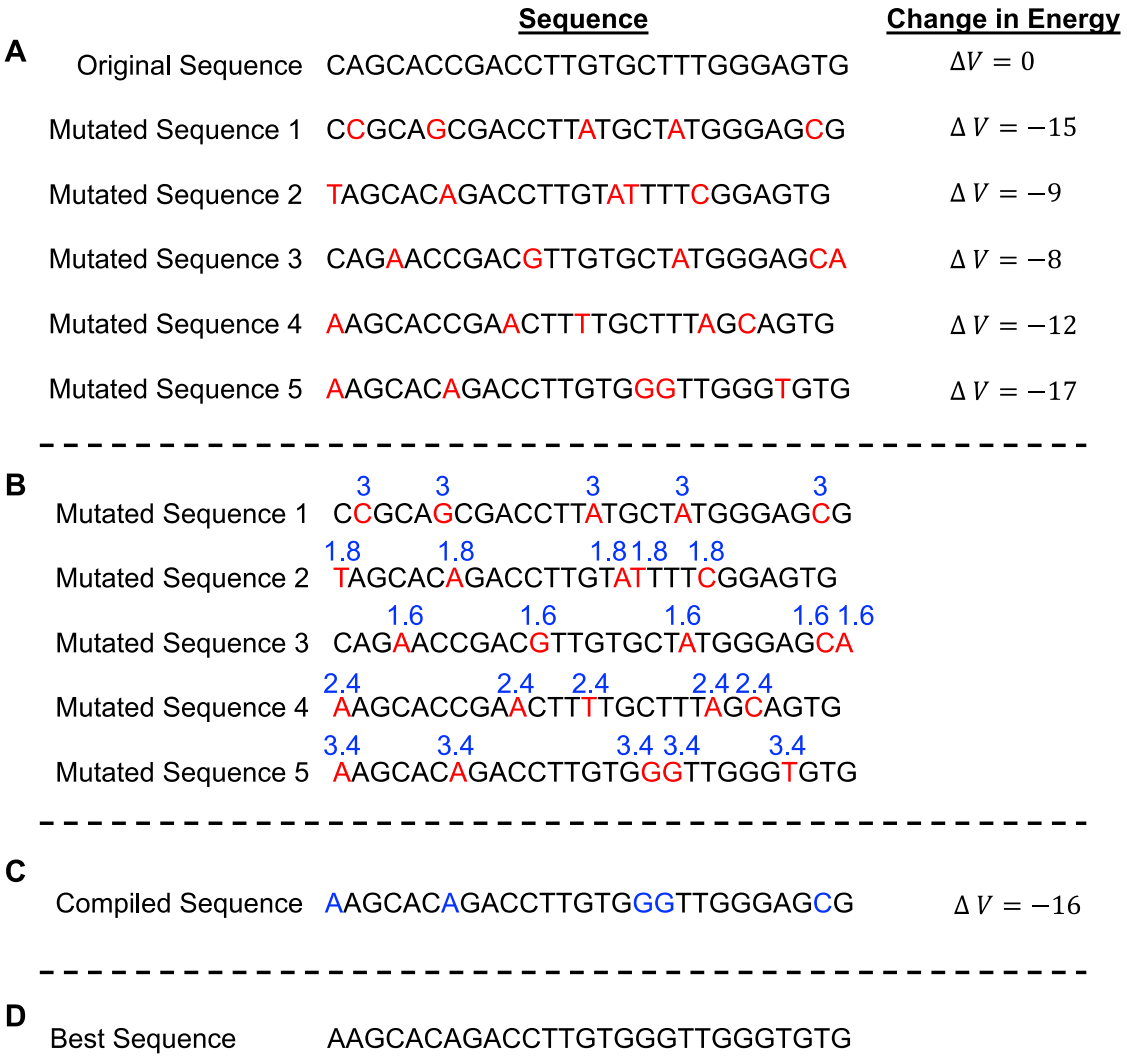
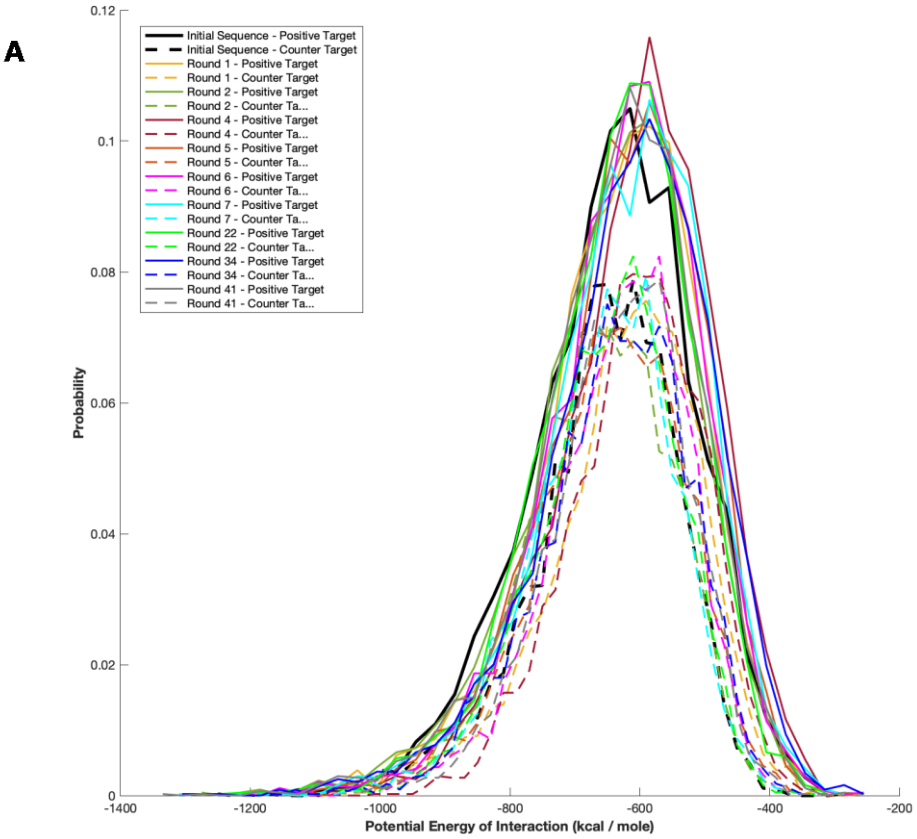


Figure 4-9. Sequence Selection Process.
Once each sequence is scored, the sequences with better scores than the previous best sequence are identified (panel A). Each mutation in the sequence is assigned an equal portion of the improved score, shown in panel B. Mutations with the greatest attributed score are used to generate a compiled sequence (panel C). The compiled sequence is converted into its three-dimensional structure and then docked. The max score of the original / previous best sequence, mutated sequences, and compiled sequence is identified (panel D). This corresponding sequence then becomes the best sequence for the next iteration. If no perturbed sequence leads to an improved score, then the previous round best sequence is maintained.



B

Initial	GCGGCGCCCGGCCGCCATGCATGTAACGCGGCGGCCGGGCG
Round 1	TCTGCGCCCGGCCGCGATGCAGGTAACGCGGCGGACGGGCG
Round 2	CCTGCGCCCGGCCGCGATACAGGTAGCGCGTCGGACGTGCG
Round 4	CCTGCGCCCGCCCGCGATATACGTAGCCCGTCGGACATGCG
Round 5	CCCGCGGCCGCCCGCGATATACGTAGCGGGTCGGAGATGCG
Round 6	GCCGCCGCCGCCCGCGATATACGTAGTGGGTCGGTCATGCG
Round 7	GCTGCCGCCCCCCGCGTTATACGTAGTGGGTCGGACATACG
Round 22	GCTGCCGCCCCCCGCGGGAAGCGTAGTGGGTCTGACATACG
Round 34	GCTGCCTCCCCCCGGAGGAAGCTTAGTGGGTCGGACATACG
Round 41	GCTGCCTCCCCCCAGAGGAAGCTGAGTTAGTCGGACGTACG

Figure 4-10. Final Sequence Extraction.
The molecular docking data for each perturbed sequence from every round is written to an output file at the end of each round. The sequences with improved scores for the target molecule and worse interaction towards the counter-target are easily extracted from the output file. Panel A is a probability versus potential energy of interaction plot. Each line represents a sequence identified to have improved potential binding. The solid line is for the sequence against the target molecule and the dotted line is against the counter-target.

5.0 CHAPTER 5 – DIGITALSELEX PLATFORM VALIDATION

This chapter outlines the results of multiple investigations of the digitalSELEX platform from *de novo* design through *in vitro* validation. This validation initially examined rapid generation of high affinity aptamers prior to examining oligonucleotides that are both high affinity and specific. These investigations covered multiple target proteins (Spike, ACE2, HA, and PD1) with counter-targets for specificity. The K_d-values for the target and counter-target molecules showing affinity and specificity were determined using the flow cytometry method highlighted in Chapter 2.

5.1 PLATFORM VALIDATION PROBLEM SETS

5.1.1 Problem 1: *de novo* Spike with HA Specificity Aptamer

The global COVID-19 pandemic caused by the novel coronavirus, SARS-CoV-2, illustrates the necessity for rapid identification to not only properly treat a patient but to also reduce the spread. A unique characteristic of the SARS-CoV-2 virus is its receptor binding domain, particularly the Spike protein. Studies have shown that the viral Spike protein interacts with the angiotensin converting enzyme 2 (ACE2) protein on the host target.[160]

In June 2020, Song *et al* published the sequence of the first aptamers, 1C and 4C, that bind Spike protein. The desired application for these aptamers was as a

therapeutic.[161] These aptamers were identified using a hybrid SELEX method that combined standard SELEX procedures with machine learning. The Spike protein was modified and bound to protein A beads forming what was referred to as a Spike-Protein A bead complex. A random library of single strand oligonucleotides was then exposed to the beads for 30 minutes at room temperature. The bound oligonucleotides and bead mixture was then incubated in a PCR mixture to amplify the oligonucleotides. After the fourth round, a counter-selection step was added to the process with ACE2 being the negative target molecule. The positive selection step incubation time was decreased every iteration from 30 to 8 minutes over 12 rounds.[162]

After the final round, the remaining aptamers were sequenced. The identified sequences were analyzed using SMART-Aptamer 2.0 to evaluate and cluster recurring groupings of nucleotides. This machine learning process helped reduce the population for oligonucleotide sequences from thousand to smaller clusters for *in vitro* testing. While the authors tested several aptamers, the results published show that aptamers 1C and 4C which are truncated (*e.g.,* primers removed from original version) oligonucleotides had the greatest affinity for the Spike protein. The authors determined the K_d-values to be 5.8 and 19.9 nM respectively using a flow cytometry assay.[163] No K_d-value against the counter-target ACE2 protein was given but the authors used a competition assay.

Even though aptamers 1C and 4C were identified for potential therapeutic purpose, it is necessary to consider their specificity towards other molecules such as Hemagglutinin (HA). The symptoms of COVID-19 and Influenza are indistinguishable.[164] Consequently, for aptamers 1C and 4C to be employed beyond their

intended application, it is necessary to ensure the oligonucleotides are specific. Currently there are no published oligonucleotides for Spike with specificity against HA.

5.1.1.1 Spike: Prior Knowledge for Design

There is little prior knowledge put forth by the Song *et al* publication that is directly applicable to the *de novo* design process. The authors did employ molecular docker, Rosetta, with aptamers 1C and 4C with the Spike protein to show potential interaction sites on the target. For consistency, the *de novo* design utilized the same Spike protein pdb file (PDB ID: 6vsb). Additionally, the final length of the 1C aptamer is 51 nucleotides is a starting approximation length for the designed aptamer.

5.1.1.2 Spike Aptamer Goal

Even though aptamers 1C and 4C demonstrate high affinity for the Spike target molecule, their specificity remains uncertain. In order to confirm specificity for application as a biosensor probe, the Aptamer on the Bead configuration flow cytometry experiments were conducted. The data visualized in **Figure 5 – 1** illustrates there is no difference in the dose response between being of Spike or HA protein with either aptamer 1C or 4C.

A *de novo* Spike aptamer for use as a biosensor probe, will not only demonstrate high affinity for Spike but also the affinity for Spike will be at least 4 times greater than that of HA protein.

5.1.2 Problem 2: *de novo* ACE2 aptamer

While aptamers 1C and 4C were identified to bind to the SARS-CoV-2 Spike protein as a potential therapeutic, another potential therapeutic oligonucleotide target is the ACE2 receptor. In November 2021, Villa *et al* published two aptamers, aptamers 6 and 14, that bind with affinity to the ACE2 receptor.[165] These aptamers were identified using an *in vitro* SELEX method where oligopeptides for the ACE2 receptor binding region with Spike protein was chemically synthesized with biotin tags. The oligopeptides were linked to streptavidin beads. A randomized library consisting of 40 random nucleotides surrounded by 5'- and 3'-end primers was used in 10 rounds of positive selection and negative selection (empty streptavidin beads). Following the rounds of selection and counter-selection the remaining oligonucleotide population was sequenced.

The sequences were chemically synthesized and subjected to an additional single round SELEX followed by a qPCR. This process identified 14 potential aptamers for validation. Following an ELISA assay, aptamers 1, 6 and 14 demonstrated the best potential to block the binding of the Spike protein to the ACE2 receptor. Using a calorimetric assay, aptamers 6 and 14 demonstrated the best K_d-value for the ACE2 receptor at 29 and 94 nM respectively.[166]

While an ELISA assay was used to illustrate that the aptamers blocked the interaction between the ACE2 receptor and the Spike protein, there was no determination of specificity. No K_d-value was calculated for binding of the aptamer towards the Spike protein for specificity comparison.

5.1.2.1 ACE2: Prior Knowledge for Design

There is no knowledge presented by the Villa *et al* publication of aptamers 6 and 14 that is directly relevant to the *de novo* design an ACE2 aptamer. The authors employed the molecular docker, HADDOCK, with the ACE2 protein (PDB ID: 6vw1) to show possible interactions between the two aptamers and the target molecule.[167] Neither the molecular docker or the PDB files are applicable to the digitalSELEX design process. The PDB file employed by the authors is derived from the x-ray crystallography structure showing a chimeric SARS-CoV-2 binding domain interacting with the ACE2 receptor.[168] Since the ACE2 receptor is not in its native confirmation then a design process would be skewed by a pre-bound molecule.

Additionally, the length of the identified aptamers is 76 nucleotides.[169] This length could demonstrate greater structural variation when chemically synthesized for validation. The designed aptamers will consequently be closer to 50 nucleotides in an effort to reduce structural variation.

5.1.2.2 ACE2 Aptamer Goal

The current aptamers 6 and 14 have published K_d-values of 29 and 94 nM. However, our validation procedure illustrates that the K_d-value for ACE2 is lower at 3.48 nM, as shown in **Figure 5 – 2**, and lacks specificity. Due to its high K_d-value, Aptamer 14 was not chemically synthesized or tested. The goal is to use the digitalSELEX platform for design of an oligonucleotide with affinity for ACE2 that is on par or better than aptamer 6. While specificity is critical for application, only positive selection will be employed in the digitalSELEX platform. This is done for two reasons. The first rationale is to accurately estimate the minimum time required to generate a high affinity aptamer.

The second reason is to see if the design process potential generates inherent specificity. Specificity for the *de novo* ACE2 oligonucleotides will determine specificity against Spike.

5.1.3 Problem 3: *de novo* HA Aptamer

While both Influenza and COVID-19 are caused by different viruses, both are contagious respiratory maladies with similar symptoms. As reported by the United States Centers for Disease Control (CDC), it is not possible to distinguish between the two illnesses by symptoms alone.[170] Specific testing is required to identify the specific virus. The Spike protein is a distinguishing molecular marker of the SARS-CoV-2 virus, and the Hemagglutinin (HA) is a unique marker on the influenza virus.

There have been two published aptamers for the HA molecule, aptamer 1 and V46. Both aptamers were published prior to the COVID-19 pandemic in 2016 and early 2019 respectively. The aptamer V46 was published for the application of sub-typing H1N1 variants. To this end, Bhardwaj *et al* generated their own system expressing a single chain of the HA protein for the different variants.[171] Consequently, the aptamer can distinguish between variant chains, but cannot detect the influenza virus with intact HA protein.

Aptamer 1 is a single-stranded oligonucleotide with a reported affinity for HA of 78 nM.[172] This aptamer was identified through a 13 round SELEX with positive selection towards the HA. The HA protein was complete protein from a recombinant system expressed in Madin-Darby canine kidney cells or commonly known as MDCK cells.[173] Following the SELEX method and sequence, the Li *et al* identified to potential candidates

that inhibited HA binding to sialic acid reporters on red blood cells. Both aptamers are 78 nucleotides in length with Aptamer 1 demonstrating the greatest affinity.

5.1.3.1 HA: Prior Knowledge for Design

A key element of prior knowledge is to use the entire HA molecule and not a single chain of the target molecule. This also includes employing the native structure of the molecule. Neither study that produced V46 or Aptamer 1 employed molecular docking so there is no initial protein databank reference file. Like the ACE2 aptamer, the length of Aptamer 1 at 78 nucleotides is too long for design consideration. This length oligonucleotide can have greater structural variation and is not ideal for implantation on a biosensor.

5.1.3.2 HA Aptamer Goal

The goal of the *de novo* design of an HA oligonucleotide is to emulate previous HA aptamers, **Figure 5 – 3**, and use only positive selection to generate a high affinity oligonucleotide. Specificity of the oligonucleotide with respect to Spike will also be determined but is not the goal.

5.1.4 Problem 4: *de novo* PD1 Aptamer

The Protein Death Receptor 1 (PD1) plays a vital role in the regulation of T cell responses. During an immune response, the ligands for the PD1 receptor are upregulated. These ligands bind to PD1 for the down regulation or suppression of the immune response. Studies have shown that the PD1 receptor is upregulated in tumor infiltrating

and peripheral T cells which leads to immune response evasion in cancers such as non-small cell lung cancer and Hodgkin's lymphoma.[174]

As a potential therapeutic, Gao *et al* identified aptamers via 10 rounds of SELEX that bind to PD1 in CHO-K1 cells that overexpress PD1. The initial library for the SELEX contained a randomized 40-nucleotide core surrounded by primers on either end that were 18 nucleotides in length. The negative selection target for this SELEX method was CHO-K1 cells that did not overexpress the PD1 receptor. The incubation process for the SELEX methods was at 4°C.[175]

At the end of the SELEX and validation process, four aptamers were found to be suitable (PD2, PD4, PD27, and PD4S). The first three aptamers were found to have K_d-values of 19.95, 19.80, and 54.28 nM respectively. Since PD4 had the lowest *Kd*-value, the authors removed the primers and reduced the length of the aptamer from 76 nucleotides to 40 nucleotides. The final dissociation constant of PD4S was determined to be 10.3 nM.[176] To date, these aptamers have not been chemically synthesized and validated in the aptamer on the bead configuration.

5.1.4.1 PD1: Prior Knowledge for Design

Based on the validation data presented by Gao *et al*, the shorter oligonucleotide had greater affinity for the target. This could be due to the small size of the PD1 receptor where the oligonucleotide is larger than the target molecule. *Steric hinderance*, discussed in Chapter 2, can lead to diminished interaction with the target molecule. Consequently, the designed aptamer should have similar length.

5.1.4.2 PD1 Aptamer Goal

The goal is to *de novo* design an aptamer for PD1 while using the predicted structure of the target molecule. The predicted molecular structure is generated using Google's AlphaFold which is an artificial intelligence program that predicts the structure of a protein when applying the data in its 100,000-molecule training dataset.[177-178] There is no specificity requirement for this problem as implementing the AlphaFold structure is itself novel.

5.1.5 Problem 5: *de novo* Spike Aptamer with ACE2 Specificity

The *in vitro* SELEX method used to identify aptamers 1C and 4C included a counter-selection step against the ACE2 protein. The authors Song *et al*, do not provide a K_d-value for the aptamers with the counter-target molecule.[179] The competition assay shows diminished binding of the aptamers to Spike in the presence of ACE2, but the data present lacks quantifiable values. To confirm or deny specificity, the validation process detailed in Chapter 2 was employed on aptamers 1C and 4C with both Spike and ACE2. The data, shown in **Figure 5-1**, indicate the aptamers are not specific. Therefore, there is a need to design an oligonucleotide that binds with high affinity to Spike with low affinity to ACE2.

5.1.5.1 Spike-ACE2: Prior Knowledge for Design

For consistency, the *de novo* design of an oligonucleotide with specificity against ACE2 protein will employ the same pdb file (PDB ID: 6vsb) that was used by the Song *et al* and in Problem 1.[180] There was no information provided regarding a molecular

docking against ACE2. Subsequently, the pdb file 1r42, which is the native structure of the ACE2 protein, was selected.[181] Since the clustering in the *Cold Start* module utilized Algorithm 1 which does not rely on a counter-target, then the same initial sequence was employed for the *Warm Start* module.

5.1.5.2 Spike-ACE2 Aptamer Goal

As previously shown in **Figure 5-1** neither aptamer 1C nor 4C demonstrate a 4-times difference in K_d-value for binding Spike versus ACE2. These aptamers do not meet the specificity criteria. Hence, the design goal is an oligonucleotide that has high affinity for Spike and is specific against ACE2. This *de novo* design will employ the initial aptamer sequence in Problem 1 *Cold Start* module.

5.2 PLATFORM VALIDATION RESULTS

5.2.1 Problem 1: *de novo* Spike with HA Specificity Results

5.2.1.1 *Cold Start* Module

The *Cold Start* module initializes with the PDB file: 6vsb. This structural file is for the cryogenic electron microscopy structure of a prefusion configuration of the SARS-CoV-2 virus.[182] This file has 22,854 atoms at a resolution of approximately 3.5 Å. The authors note that the predominant structure of the trimer has one of the three receptor-binding domains (RBDs) rotated up in the receptor accessible confirmation.[183]

For clustering accessible and biological relevant atoms, the *K-means* clustering (Algorithm 1) was used. The cluster number was set to produce clusters of approximately 10 – 15 amino acids. The output for the Delaunay triangular, solid angle determination, clustering, and top clusters are illustrated in **Figure 5 – 4**. The initial breakdown of the target molecule, while size dependent, only takes approximately 1 minute when employing Algorithm 1.

The amino acids of the top clusters are also written into a csv file. If necessary, these clusters can be visualized in other pdb file visualization software (*e.g.,* pymol), **Figure 5 – 5**, to confirm the selected clusters are accessible with respect to the larger molecule.

Once the top clusters are identified, the remainder of the sequence is generated using the genetic optimization algorithm. The initial population selects random nucleotides with 15 before and after the core sequence. In the cluster with the smallest solid angle, there are 11 nucleotides in the core sequence making the overall length 41 nucleotides. The maximum number of generations was established at 200 with 100 stall generations. The best stability score for the algorithm to achieve is negative infinite, ensuring the algorithm explores all options.

As discussed in Chapter 3, the general constraints of the genetic optimization algorithm are focused on biosensor application. The 5'-end must have at least 4 unpaired nucleotides, quad nucleotides are not allowed, and the CG content is at least 60% of the oligonucleotide. The initial sequence from the top or more open cluster is depicted in **Figure 5 – 6**. Using the current number of generations and population size of 100, the genetic optimization algorithm takes approximately 10 minutes to complete.

5.2.1.2 *Warm Start* Module

Even though a sequence is generated using the optimization algorithm for each of the top 10 clusters, the sequence from top cluster is the initial sequence in the *Warm Start* module. There are several options that can be adjusted based on computational capacity and design constraints. For the *de novo* Spike oligonucleotide these options are listed in **Table 5 – 1**. The positive selection pdb file is 6vsb which is the same structure used in the *Cold Start* module. The counter-selection molecule is pdb file: 3lzg which is the crystal structure of a 2009 H1N1 influenza virus HA molecule.[184] The target, counter-target molecules and all oligonucleotide structures are protonated via pymol. The scoring function used in this process is the maximum probability which selects the sequences that maximizes the probability difference between the potential binding of an oligonucleotide to the target versus counter-target. There was no fixed portion of the sequence used in this process. For each iteration, only 7 sequences were generated with a maximum of 5 mutations per sequence. The number of sequences was low due to the computation capacity as it required two CPUs per sequence to run the docking.

The module was allowed to complete 108 iterations before being stopped since the last selected sequence occurred at round 43. The scoring function identified 6 sequences that maximized the probability between the oligonucleotide binding to Spike protein versus HA. The probability distribution plot versus potential binding energy and corresponding sequences is highlighted in **Figure 5 – 7**. The limited number of sequences (6) that demonstrated via the molecular docker selective binding towards the target molecule versus the counter-target molecule vastly reduces the potential number of

sequences needing *in vitro* validation. The overall time for the 108 rounds of module took approximately 72 hours.

5.2.1.3 Validation

Four oligonucleotides were selected for validation using the aptamer on the bead configuration flow cytometry. These oligonucleotides were the initial sequence (dnSpikeI), and the oligonucleotides from rounds 1 – 3 which were denoted as dnSpike1, dnSpike2, and dnSpike3. The sequences from rounds 8, 34, and 43 could have also been selected. Each validation experiment indicated in **Figure 5 – 8** was completed at least 3 times. The four-parameter dose response model plot from the *drc* library in R defaults to plotting the mean response at each concentration. This helps reduce clutter on the plots.

The affinity of the four tested oligonucleotides was in the low single digit nanomolar range, **Figure 5 – 8A**. All four aptamers met the affinity requirement.

Regarding specificity, **Figure 5 – 8B** shows the dose response interaction of each oligonucleotide with HA protein. Of the four tested, only aptamers dnSpike1 and dnSpike2 met the 4-fold difference in affinity for the target versus the counter-target. The specificity for dnSpike1 and dnSpike2 was 4.5-fold and 19-fold difference which are purposely highlighted in **Figure 5 – 8C**.

The previously published aptamers 1C and 4C were not identified in their SELEX method to be specific against HA. The K_d-value difference between 1C and 4C with Spike and HA indicate these oligonucleotides would bind both molecules with nearly identical affinity. **Figure 5 – 8D**, however, indicates that dnSpike1 and dnSpike2 can distinguish between the two molecules especially when compared to the published aptamers.

The sequences, two-dimensional, and three-dimensional structures of the oligonucleotides, dnSpike1 and dnSpike2 are shown in **Figure 5-9**. These structures highlight the paired and unpaired nucleotides in the sequence which can interact with the target molecule with high affinity and specificity.

5.2.2 Problem 2: *de novo* ACE2 Results

5.2.2.1 *Cold Start* Module

The *Cold Start* module initializes with the PDB file: 1r42. This structural file is native human angiotensin converting enzyme-related carboxypeptidase (ACE2). This file contains 5,511 atoms at a resolution of 2.20 Å which was achieved through x-ray crystallography.[185]

The clustering of the accessible and biological relevant atoms was achieved using Algorithm 1: the *K-means* clustering method. The cluster number was adjusted to produce clusters of approximately 10 – 15 amino acids, like the *de novo* Spike aptamer. There are less atoms and amino acids in the ACE2 target molecule than that of the SARS-CoV-2 target. The number of potential clusters decreased to achieve the requisite number of amino acids. The Delaunay triangular for all connections, solid angle determination, clustering, and top clusters are visualized in **Figure 5-10**. The initial breakdown of the target molecule, while size dependent, only takes approximately 1 minute when employing Algorithm 1.

The amino acids of the top clusters are also written into a csv file during the Cold Start process. If necessary, these clusters can be visualized in other pdb file visualization

software (*e.g.,* pymol), **Figure 5-11**, to confirm the selected clusters are accessible with respect to the larger molecule.

After the top ten clusters are identified by rank ordering the sum of the solid angles, the remainder of the sequence is generated using the genetic optimization algorithm. The initial population selects random nucleotides with 14 before and 12 after the core sequence. The total length of the initial sequence from the top cluster is 38 nucleotides with 12 nucleotides in the core sequence, 14 in the prefix, and 12 in the suffix. The maximum number of generations was established at 300 with 100 stall generations. The best stability score for the algorithm to achieve is negative infinite, ensuring the algorithm explores all options.

As highlighted in Chapter 3, the general constraints of the genetic optimization algorithm are focused on biosensor application and have not been altered. The 5'-end must have at least 4 unpaired nucleotides, quad nucleotides are penalized, and the CG content is at least 60% of the oligonucleotide. The initial sequence from the top or more open cluster is depicted in **Figure 5-12**. This sequence does have a single penalty from quad nucleotides; however, this violation is in the core sequence and cannot be modified by the algorithm. Using the current number of generations and population size of 100, the genetic optimization algorithm takes approximately 10 minutes to complete.

5.2.2.2 *Warm Start* Module

The initial aptamer sequence in the *Warm Start* module is the best optimized sequence from the top cluster even though there was a penalty. Since the previously published aptamers 6 and 14 had no counter-selection molecules in their SELEX method, no counter-selection molecule was utilized for the *de novo* ACE2 aptamer. The module

options employed are listed in **Table 5 – 2**, but there are some key findings. First, there were 25 sequences generated per iteration since no counter-selection step was utilized. And secondly, the scoring option was different. This option used the top 10 percentile point as the measurement marker.

The module was stopped after 74 iterations. This process generated and docked 1,924 sequences with 10 generated sequences that possessed an improved molecular docking score at the top 10 percentile position. The probability distribution versus potential energy of interaction as well as corresponding sequences are illustrated in **Figure 5-13**.

Due to server maintenance, the computational time for the *de novo* ACE2 aptamers was slower than the *de novo* Spike aptamers at approximately 144 hours.

5.2.2.3 Validation

While it is possible to validate all 10 sequences plus the initial sequence, it was decided to focus on the sequences the greatest maximum probability in the probability versus molecular docking potential energy of interaction. The scoring function that identified these sequences was not related to the probability rather a specific point in the distribution. The rationale for selecting these high probability sequences was that the sequences interacted in a narrower capacity with the target molecule. This reduces the number of conformations with varying amounts of interactions. Consequently, these sequences have greater potential of interacting. The list of sequences, as reduced in **Figure 5-14**, to validate was reduced from 10 to 4 plus the initial sequence.

With the goal of using only positive selection to design an aptamer that interacts with the ACE2 protein with equal or better affinity than the previously published

oligonucleotides, the four sequence plus initial were chemically synthesized by IDT. Following the aptamer on the bead flow cytometry process described in Chapter 2, the affinity of these oligonucleotides was determined. The affinity for these sequences is illustrated **Figure 5-15A** dose response curves. All the tested aptamers demonstrated single digit nanomolar K_d-values, with dnACE2-I, dnACE2-7, and dnACE-9 were closest to the previously published Aptamer 6 which had a K_d-value of approximately 3.5 nM.

Even though specificity of the oligonucleotides was a secondary goal, the results of the *de novo* aptamers illustrate a greater degree of specificity towards the ACE2 molecule compared to Spike, **Figure 5-15B**. The published Aptamer 6 using the aptamer on the bead configuration, **Figure 5-2**, did not show specificity for ACE2 protein versus Spike. This however is not surprising since the SELEX method for identifiying aptamer 6 did not include Spike as a counter-target. The *de novo* aptamer, dnACE2-7, did demonstrate specificity with a fold difference of 4.8 and is shown in **Figure 5-15C**.

The comparison dose response curves, **Figure 5-15D**, illustrate that both Aptamer 6 and dnACE2-7 have comparable affinity for the ACE2 protein. The *de novo* oligonucleotide, unlike Aptamer 6, does demonstrate specificity for ACE2 protein over Spike protein.

Two additional controls were introduced in the *de novo* ACE2 aptamer validation process. The first control was testing a sequence generate from the cluster of amino acids with the largest solid angle. The core sequence of the worst cluster was used in the genetic optimization algorithm and following optimization it was synthesized without the *Warm Start* module. The second control was the generation of a completely random oligonucleotide sequence. The random nucleotide was generated using a random number

generator to select a number between 1 and 4 at total of 60 times using the MATLAB random number generator. The number 1 corresponds to Adenine, 2 to Thymine, 3 to Cytosine, and 4 to Guanine. The random sequence of numbers was then converted to a nucleotide sequence using int2nuc function in MATLAB. The dose response curve binding ACE2 for the worst cluster oligonucleotide and the random oligonucleotide are shown in **Figure 5-16** along with Aptamer 6 and dnACE2 – 7 for reference. This response curve illustrates the binding of the ACE2-fluorophore complex is greatly diminished with the two control oligonucleotides.

The sequences two-dimensional, and three-dimensional structure for dnACE2-7 oligonucleotide is shown in **Figure 5-17**. These structures highlight the paired and unpaired nucleotides in the sequence which can interact with the target molecule with high affinity and specificity.

5.2.3 Problem 3: *de novo* HA Results

5.2.3.1 *Cold Start* Module

The *Cold Start* module initializes with the PDB file: 3lzg. This file is the structure of crystal structure of the 2009 H1N1 influenza virus hemagglutinin generated via x-ray diffraction.[186] This file contains 24,137 atoms at a resolution of 2.26Å. The structure is shown as a hetero 6-mer with two unique side chains. For design purposes, the hetero 6-mer was reduced to a single two chain heteromer to prevent repetition of atom clusters.

The clustering of the accessible and biological relevant atoms was achieved using Algorithm 1: the *K-means* clustering method. The cluster number was adjusted to produce clusters of approximately 10 – 15 amino acids. The number of potential clusters

decreased to achieve the requisite number of amino acids. The Delaunay triangular for all connections, solid angle determination, clustering, and top clusters are visualized in **Figure 5-18**.

The location of the top cluster, which was used to generate the initial aptamer sequence for the *Warm Start* molecule, is depicted in both the Cold Start visualization and on the actual HA molecule in **Figure 5-19**. This cluster contains several amino acids of the three conserved sequences that composed the receptor binding domain of HA.[187]

5.2.3.2 *Warm Start* Module

The previously published Aptamer 1 did not have a counter-selection molecule in its SELEX process so there was no counter-selection molecule utilized in the *Warm Start* module. The goal of this process is to modify the initial sequence to have either equal or better affinity for the HA molecule. The module options employed are listed in **Table 5 – 3**, but there are some key notes. First, there were 25 sequences generated per iteration since no counter-selection step was employed. Secondly, the scoring function employed the maximum probability.

The module using the initial sequence shown in **Figure 5-20** for the most open cluster was stopped after 5 iterations due to power fluctuations the Boston College server at the time. Due to the short duration of the module, this process only generated and docked 130 sequences. Sequences in 2 rounds were selected in the sequence identification process as having improved molecular docking score. The probability distribution versus potential energy of interaction as well as corresponding sequences are illustrated in **Figure 5-21**.

The computational time for the *de novo* HA aptamers without counterselection took approximately 10 hours.

5.2.3.3 Validation

The goal of the *de novo* HA aptamer was to rapidly design an oligonucleotide with similar high affinity to HA as Aptamer 1. Since there were only 2 accepted sequence changes, three sequences (initial, dnHA-1, and dnHA-4) were chemically synthesized by IDT. Employing the Aptamer on the Bead Flow Cytometry procedure detailed in Chapter, the affinity of the oligonucleotides was determined and is shown in **Figure 5-22A**. The previously published aptamer for HA (Aptamer 1) has an experimentally determined K_d-value for HA of 2.47 nM (**Figure 5-3**).

Even though specificity of the oligonucleotides was a secondary goal, the results of the *de novo* aptamers illustrate a slight degree of specificity towards the HA molecule compared to Spike, **Figure 5-22B**. The published Aptamer 1 using the aptamer on the bead configuration, showed a similar level of affinity for Spike. Neither the previously published aptamer or designed aptamers demonstrated the 4 times affinity difference between HA and Spike to be considered specific.

Figure 5-22C illustrates the lack of difference between the binding of the designed aptamers to HA as well as Spike. As previously stated, these oligonucleotides cannot be considered specific at least with respect to Spike, however the aptamers can be considered high affinity.

The comparison dose response curves, **Figure 5-22D**, illustrates that Aptamer 1 and dnHA aptamers have comparable affinity for the HA protein. None of the

oligonucleotides meet the guidelines for specificity, however there was no counter-selection in the design process for *de novo* aptamers.

5.2.4 Problem 4: *de novo* PD1 Results

5.2.4.1 *Cold Start* Module

Unlike previous design methods, the *de novo* design of a PD1 aptamer used the AlphaFold AI predicted three-dimensional structure of the target instead of the experimentally derived crystal structure. The predicted structure can come in several formats including the pdb file format which allowed the Cold Start import function to remain unchanged.

The PD1 structure in AlphaFold utilizes is listed as Q15116 and is a single chain with only 288 amino acids. The Q15116 structure from AlphaFold as well as the model confidence legend is shown in **Figure 5-23**. Visible in this AlphaFold is a large hydrophobic tail that is used to anchor the receptor into the plasma membrane. To facilitate the *Cold Start* process, the hydrophobic tail was removed to reduce prevent the identification of accessible atoms from being generated from the tail. The full structure and the truncated PD1 structure are shown in **Figure 5-24**.

The clustering of the accessible and biological relevant atoms was achieved using Algorithm 1: the *K-means* clustering method. The cluster number was greatly reduced to 5 total to generate clusters of approximately 10 – 15 amino acids. The Delaunay triangular for all connections, solid angle determination, clustering, and top clusters are visualized in **Figure 5-25**.

The location of the top cluster, which was used to generate the initial aptamer sequence for the *Warm Start* molecule, is depicted in both the Cold Start visualization and on the actual HA molecule in **Figure 5-26**. The PD1 molecule is relatively small and the truncation process removing the hydrophobic tail further reduced the number of accessible amino acids.

5.2.4.2 *Warm Start* Module

The goal of the *de novo* PD1 process is to design high affinity aptamers using the AlphaFold model instead of a PDB file generated from a crystal structure. Specificity is ideal, however previous work and demand for a PD1 aptamer does not elucidate a counter-selection molecule for specificity. The module using the initial sequence shown in **Figure 5-27** and was run for 60 iterations. While the small size of the receptor imposes some constraints on the design process, the small size does improve the molecular docking time. The computation time for 60 iterations was approximately 2 hours.

The module options employed are listed in **Table 5 – 4**, but there are some key notes. First, there were 25 sequences generated per iteration since no counter-selection step was employed. Secondly, the scoring function employed the maximum probability.

5.2.4.3 Validation

Since the *Warm Start* module was limited due to the goal of the process, the validation step was also reduced. To date, the previously published aptamers have not been generated for comparison with the design aptamers. Only the affinity of the designed aptamers for PD1 was tested using the initial plus two generated sequences. The dose response curve showing the affinity of the oligonucleotides for PD1 are shown in

Figure 5-28. The dissociation constants indicate that the oligonucleotides all have high affinity for the target. More work is required to fully develop a PD1 oligonucleotide that is both high affinity and specific.

5.2.5 Problem 5: *de novo* Spike Results with ACE2 Specificity Results

5.2.5.1 *Cold Start* Module

The *Cold Start* process detailed in Problem 1 employed the Algorithm 1 K-means clustering and is counter-target agnostic unlike Algorithm 2. Consequently, the cluster identified in Problem 1, **Figure 5-4**, and its corresponding initial sequence, **Figure 5-6**,were used in this problem. The use of the existing sequence negated the necessity to re-initiate the Cold Start module and simply optimize the sequence with respect to the ACE2 counter-target.

5.2.5.2 *Warm Start* Module

The initial sequence developed in the Problem 1 *Cold Start* module was employed and is shown in **Figure 5-6**. There are several options that were adjusted based on the computational capacity and design constraints. For the *de novo* Spike oligonucleotide these options are listed in **Table 5 – 5**. The positive selection pdb file is 6vsb which is the same structure used in the previous *Cold Start*. The counter-selection molecule is pdb file: 1r42 which is the crystal structure for the native human angiotensin converting enzyme-related carboxypeptidase (ACE2).[188] The target, counter-target, and all generated oligonucleotide structures are protonated via pymol. The scoring function used in this process was the “area under the curve” option. This option sums the probability of the

two points on either side of the maximum probability. This value is then multiplied by the potential binding energy at the maximum point. The total score for each sequence is the value for the target minus the counter-target value. No sequence positions were fixed. For each iteration, 25 sequences were generated with a maximum of 5 mutations per sequence.

The module was set to and completed 60 iterations with the last sequence change occurring in round 41. Nine sequences out the over 1500 evaluated were identified by the scoring function for improved target versus counter-target interaction. The extracted sequences and probability distribution plot versus potential binding energy are highlighted in **Figure 5-29**. The overall time for the 60 iterations was approximately 145 hours.

5.2.5.3 Validation

Five oligonucleotides were selected for validation using the aptamer on the bead configuration flow cytometry detailed in Chapter 2. These oligonucleotides were the initial sequence (dnSpikeI), and the oligonucleotides from rounds 1, 4, 6, and 22, which are labeled dnSpike-AUC-1, dnSpike-AUC-4, dnSpike-AUC-6, and dnSpike-AUC-22, respectively. The AUC represents the scoring function employed in the Warm Start module while the following number indicates the round the sequence was identified. Each validation experiment indicated in **Figure 5-30** was completed at least 3 times. The four-parameter dose response model plot from the *drc* library in R defaults to plotting the mean response at each concentration.

The affinity of the five tested oligonucleotides was in the low single digit nanomolar range, **Figure 5-30A**. All five aptamers met the affinity requirement.

For specificity, **Figure 5-30B** shows the dose response for each oligonucleotide with ACE2 protein. All five oligonucleotides tested had at least a four times greater K_d-value when binding ACE2 with respect to the target molecule, Spike. The comparison between target and counter-target binding is shown in **Figure 5-30C** where the fold difference is 13.1, 5.1, 4.3, 4.9, and 4.9 respectively.

The previously published aptamers 1C and 4C did not have published K_d-values against ACE2, however **Figure 5-1** shows these aptamers are not specific, according to the 4-times difference in K_d-value to ACE2. **Figure 5-30D** compares the published aptamers with the oligonucleotides extracted from the Warm Start process. While Aptamer 1C demonstrated a better affinity for Spike protein it is not specific with respect to the ACE2 molecule, nor was Aptamer 4C. The four of the five *de novo* aptamers demonstrated K_d-values for their targets that were high affinity and were also significantly ($p < 0.001$) specific with respect to the ACE2 molecule.

5.3 RESULTS DISCUSSION

5.3.1 Sustains

The results of the digitalSELEX validation demonstrate some common successes. First, the *Cold Start* design can identify biologically relevant and accessible atoms. These atoms are then clustered and identified by their amino acids. Second, the *Cold Start* module is fast. The breakdown and analysis of the target molecule using Algorithm 1 takes approximately one minute per molecule.

Third, each designed oligonucleotide that was chemically synthesized demonstrated high affinity. The K_d-value for 18 out of the 19 total oligonucleotides was in the single digit nanomolar range. The aptamer that was not in the single digit range had a K_d-value less than 15 nanomolar. The fourth sustain focuses on the employment of a counter-selection target. When the counter-selection step was employed eight out of nine oligonucleotides demonstrated specificity.

Another sustain is the low cost. There is no sequencing, random library construction, primers, PCR reagents, or waste of target and counter-target proteins in the *de novo* design process. The only cost incurred results from the validation which all SELEX methods experience. The cost here is further minimized by few specific sequences generated by digitalSELEX compared numerous sequences generated by *in vitro* SELEX.

Finally, the last sustain is the speed of the process. The generation of an initial stable sequence from the Cold Start module and genetic optimization algorithm is complete in less than 10 minutes. The *Warm Start* process can take several days to a week to complete depending on the computational capacity. However, this timeframe is less than several weeks associated with *in vitro* SELEX methods.

5.3.2 Improves

The initial data of the digitalSELEX platform does demonstrate several opportunities for improvements. First, specificity is always a concern. When the *Warm Start* process occurred without a counter-selection molecule, fewer specific oligonucleotides were selected. It may be possible to generate higher specificity by

selecting a cluster not associated with a binding domain where the atoms inherently have a smaller solid angle. Regardless of the initial cluster, employing a counter-selection molecule in the Warm Start module helps guide specificity.

Another improve is the employment of Algorithm 2 in the clustering method. Due to time constraints, Algorithm 2, to date, has not been tested. It is possible that this algorithm will help not only improve specificity, but validation of both affinity and specificity is required.

When employing Algorithm 1, the cluster size needed to be adjusted to facilitate the identification of clusters of appropriate number of amino acids. This process can be automated by determining the number of amino acids with accessible atoms during the Cold Start process. The number of clusters required can be determined internally, reducing the user *a priori* input of a cluster number value.

Even though the speed of the digitalSELEX platform is listed as a sustain compared to *in vitro* SELEX, the speed can still be improved. Computational capacity does not prevent the execution of the platform, but it does limit its performance time. The MATLAB code runs the molecular docking simulations in parallel and is limited by the number of CPUs. Better and more CPUs will further minimize the challenge of time and enable the algorithm to more thoroughly examine the interaction landscape between the oligonucleotides and target molecules.

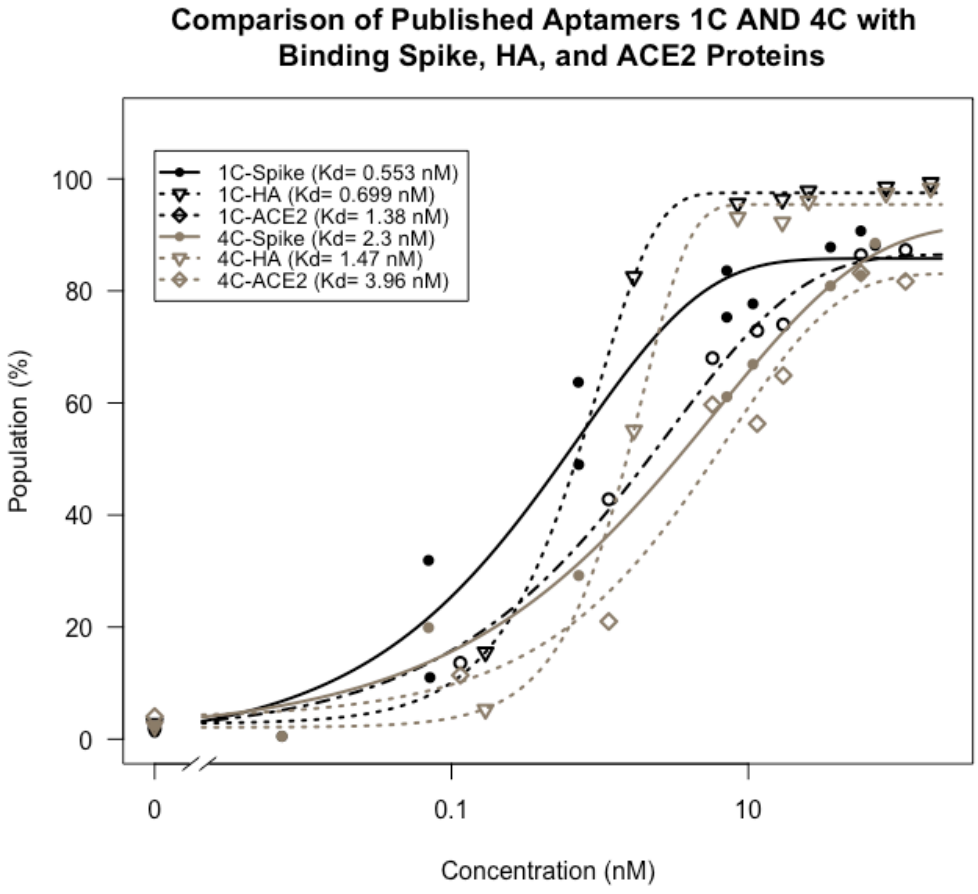


Figure 5-1. Aptamers 1C and 4C Binding Spike, ACE2, and HA
Dose response plot of Aptamers 1C and 4C binding not only the original Spike protein but also ACE2 and HA proteins. Validation experiments using the Aptamer on the Bead configuration detailed in Chapter 2 were done to confirm binding using our methodology. Aptamers 1C and 4C were identified via an *in vitro* SELEX for their interaction with Spike with the ACE2 counter-target molecule. The HA protein was not considered in the SELEX process. There were no K_d-values the aptamer binding to ACE2 provided in the 1C and 4C publication. Each point represents at least an n = 3 and generated in R using the dose response model library.

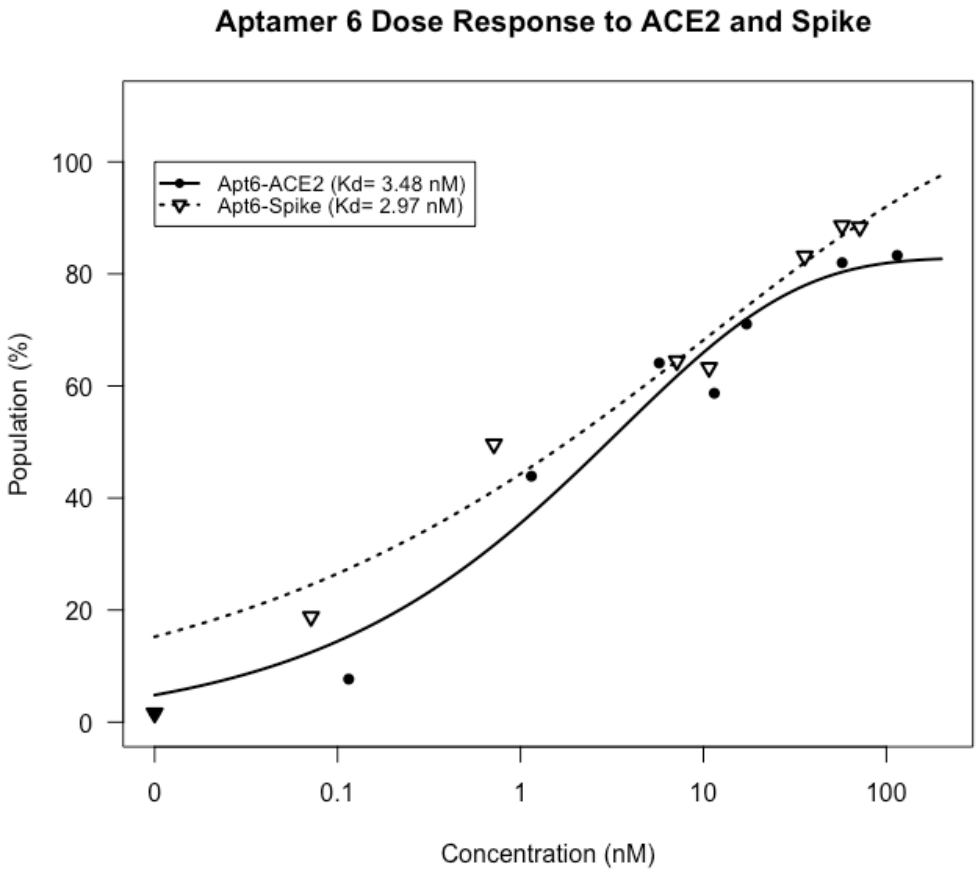


Figure 5-2. Aptamer 6 Binding ACE2 and Spike.
Dose response plot of Aptamer 6 interacting with both ACE2 and Spike. Aptamer 6 was identified to be selective for ACE2 and to function as a therapeutic however there was no counter-selection with Spike in the SELEX method. The plot indicates that higher concentrations of Spike are required to flatten the upper asymptote, but the trend indicated better binding for Spike than its target protein, ACE2.

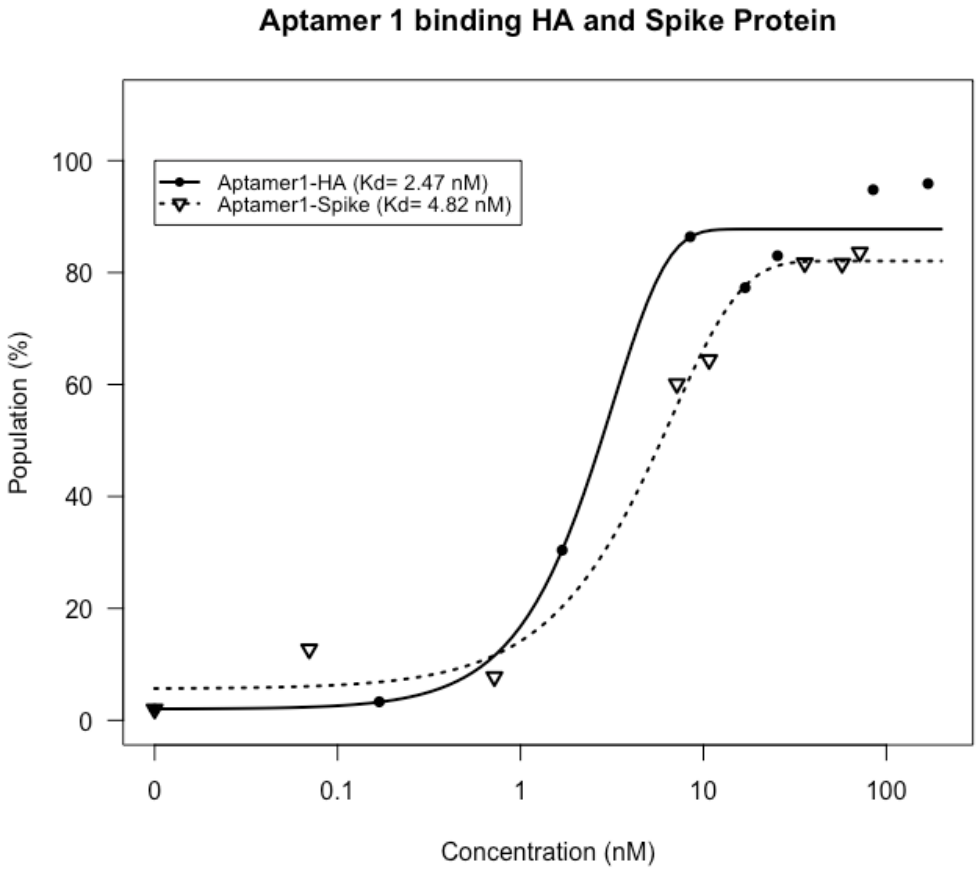


Figure 5-3. Aptamer 1 Binding HA and Spike.
Dose response plot of Aptamer 1 interacting with both HA and Spike. Aptamer 1 was identified to be selective for HA. This aptamer was identified prior to the COVID-19 pandemic and was not specific towards HA relative to Spike.

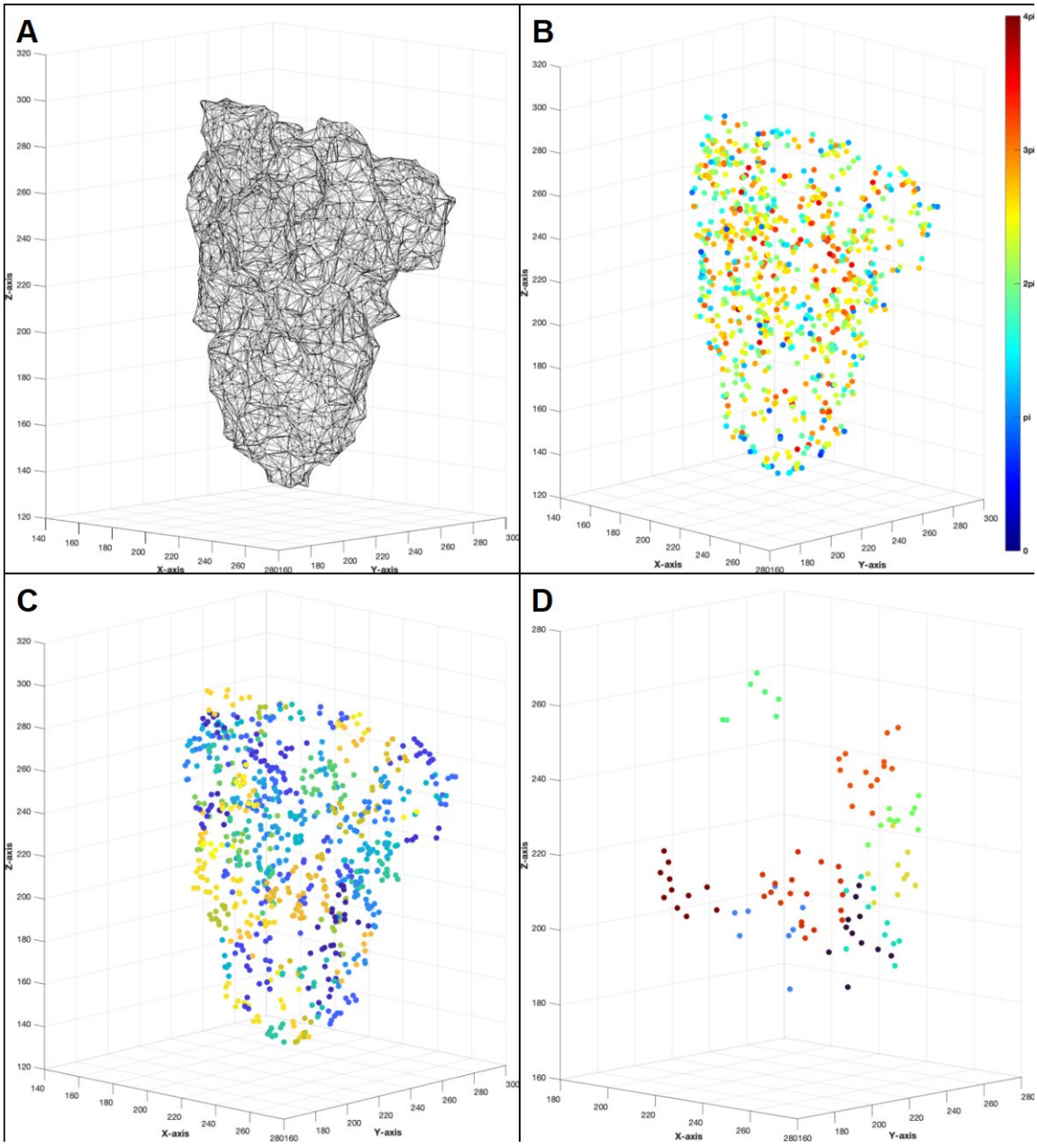


Figure 5-4. *de novo* Spike Cold Start Results.
Each panel illustrates the *Cold Start* analysis pf the SARS-CoV-2 Spike protein from the PDB file: 6vsb. Panel A shows the Delaunay triangulation of all the atoms. Based on the atom triangulation, panel B provides the solid angle of the alpha shape atoms. Each atom is color coded to represent the solid angle. Panel C depicts the clusters of the biologically relevant and accessible atoms. Panel D illustrates the top ten (smallest solid angle) clusters for the molecule.

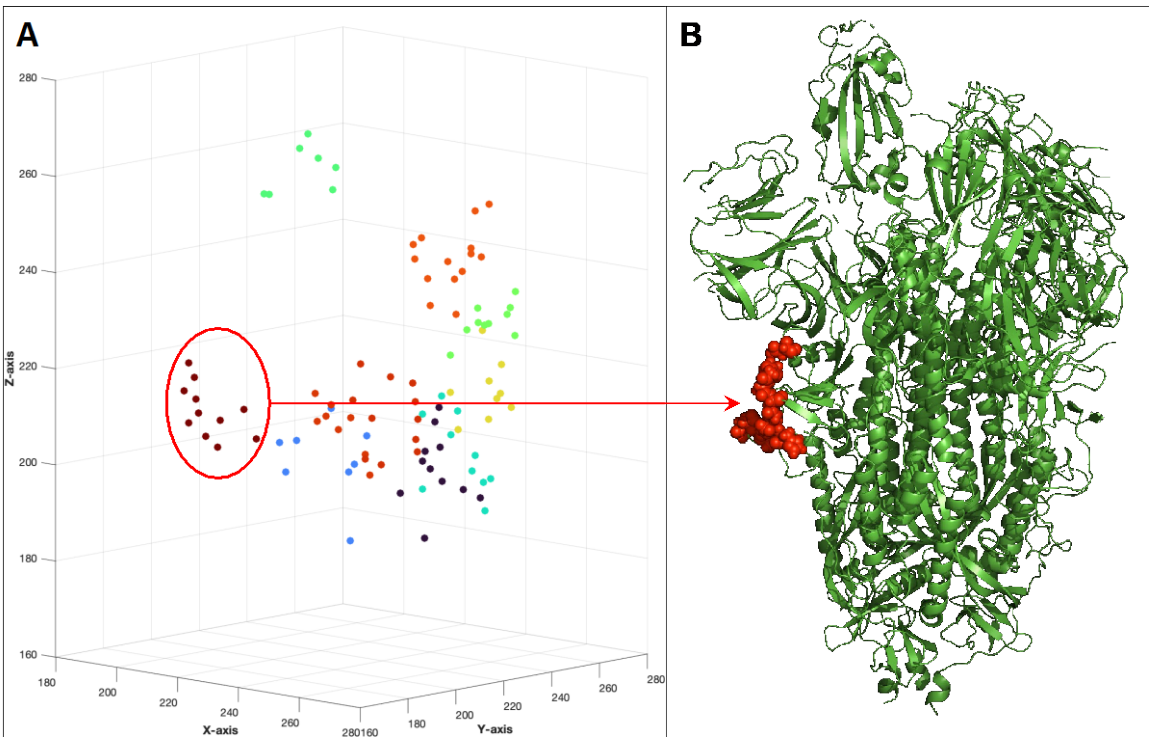


Figure 5-5. *de novo* Spike Cluster Visualization.
While the final clusters are difficult to visualize in MATLAB (panel A), the amino acids of the clusters can be highlighted and shown using additional pdb file visualization software such as pymol. Panel B highlights one of the top clusters identified in the cold start module with respect to the larger molecule.

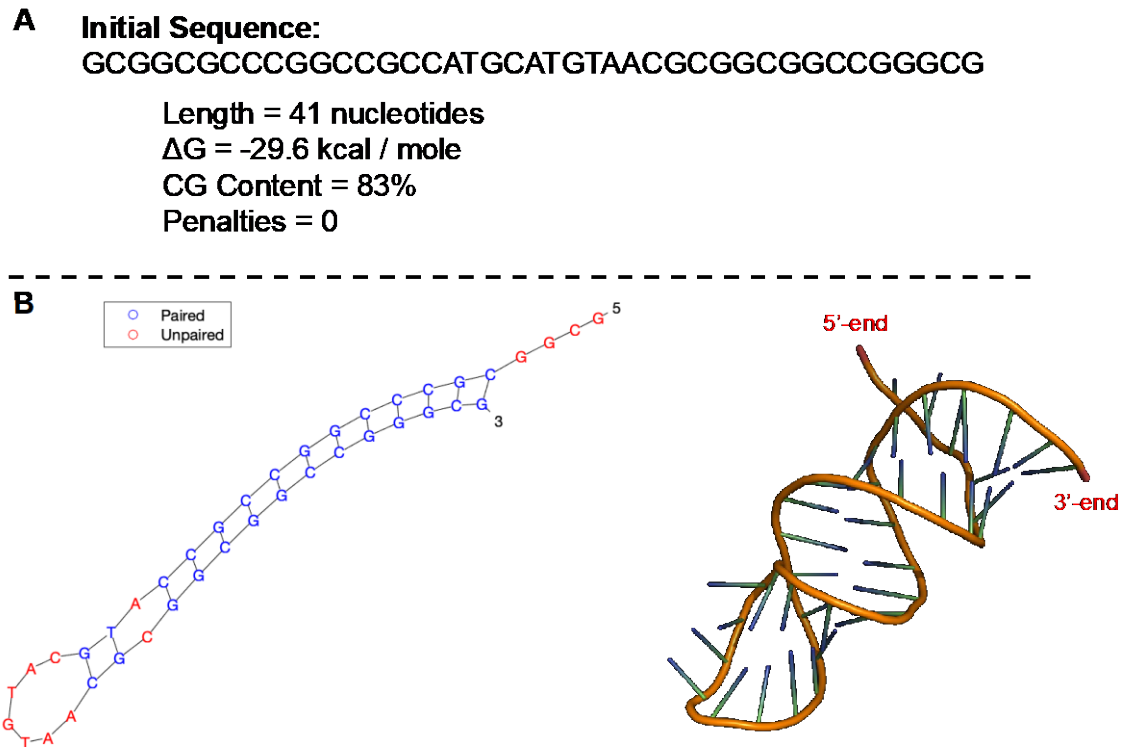
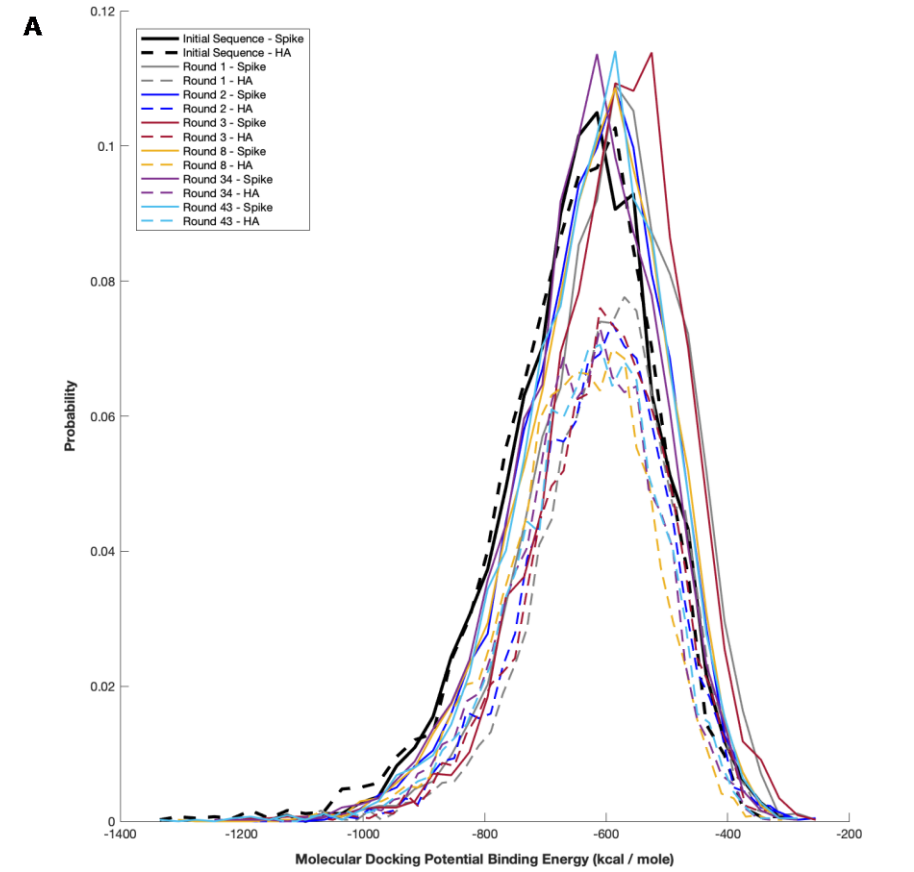


Figure 5-6. *de novo* Spike Initial Sequence.
The genetic optimization algorithm generates population of sequences and optimizes the sequences using biological operators such as mutation and crossover. The output of the optimization is an initial sequence that is the most stable confirmation of the oligonucleotide with minimal to no penalties for violating defined constraints. An example of the sequence for the Spike *de novo* design aptamer is shown in Panel A. This sequence is visualized as a two- or three-dimensional configuration in Panel B.

***de novo* Spike Options**	**Values**
Positive Target PDB	6vsb
Negative Target PDB	3lzg
Protonation	All molecules
Scoring Option	Max Probability
Fixed Portion	None
Number of Sequences per Iteration	7
Number of Mutations per Sequence	5

Table 1. *de novo* Spike Warm Start Options.
List of the options employed in the *de novo* Spike aptamer design *Warm Start* module.



B

Initial
GCGGCGCCCGGCCGCCATGCATGTAACGCGGCGGCCGGGCG

Round 1
GCGGAGCCTTGCCGCCATCCATGTAACGAGGCGGCCGGGCG

Round 2
GCGGAGCCTTGCCGCCATCCATGTAACCAGGCCCCCGTGTG

Round 3
GCGGAGCCTTGCCGGCTTCCATGTAACCAGTCTACCGTGTG

Round 8
GCGGACCCCTGCCGGCTTCCATGTAACCAGTCCACCGTTGG

Round 34
GCGGACCCCTGCTTGTTTTCATGTAACCAGTCCACCGTGGG

Round 43
GGGGACCCCTGCTTGTTTTTATGTAACCAGTCGAGCGTGGC

Figure 5-7. Sequence Extraction for *de novo* Spike.
The *Warm Start* module employs a molecular docker to evaluate perturbed sequences against the target and counter-target molecule. Panel A illustrates the probability distribution of the generated configurations for binding both the target molecule (Spike)(solid line) and counter-target (HA)(dotted line). Each line represents the best selected sequence and the round when the sequence led to an improved score. The sequences are listed in panel B.

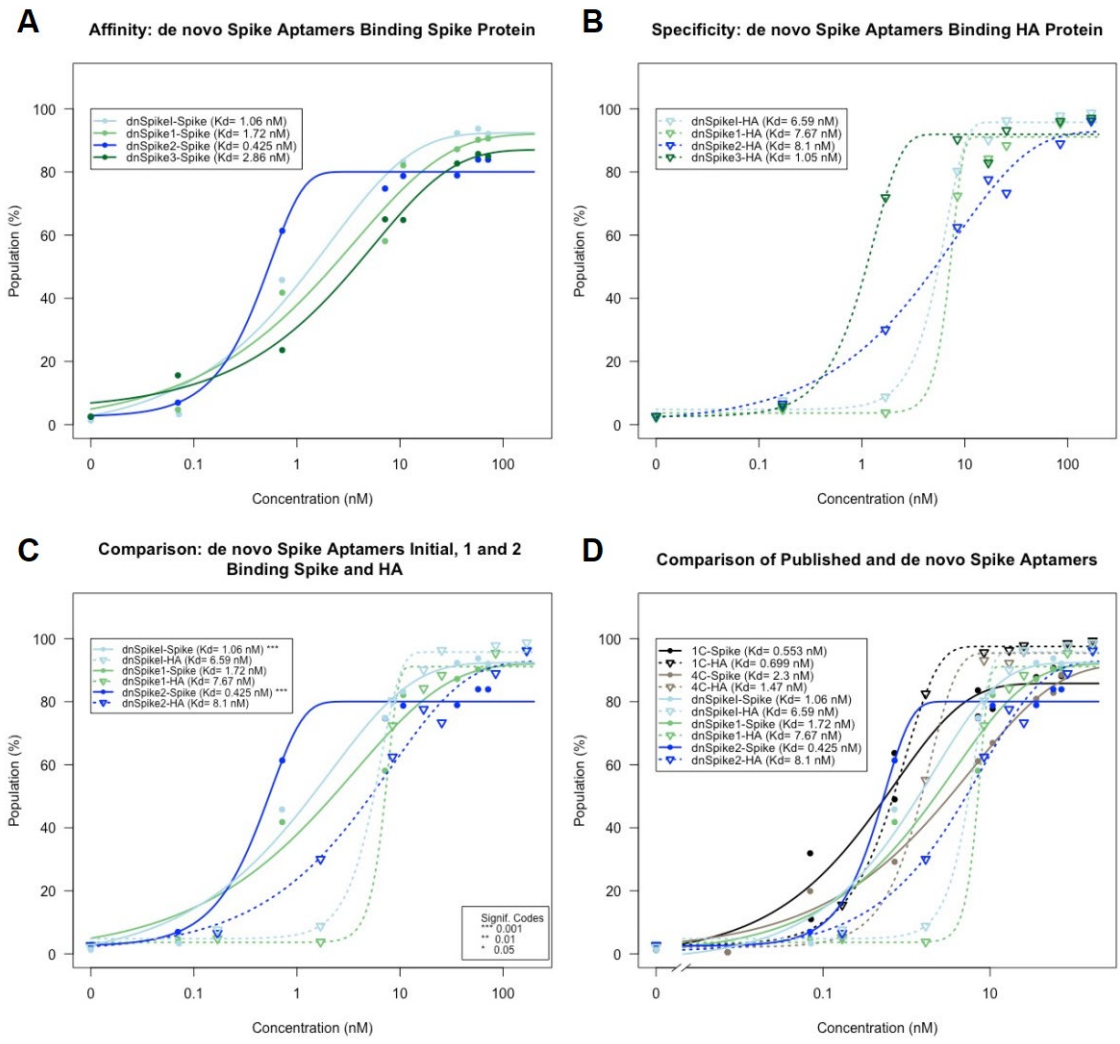


Figure 5-8. *de novo* Spike Dose Response Plots.
Four *de novo* Spike oligonucleotides (Initial plus 3 perturbed) were chemically synthesized and tested for affinity and specificity using the aptamer on the bead configuration. Panel A is the dose response curves for the four oligonucleotides to ensure high affinity (K_d-value in nanomolar range). Panel B is the dose response curves to examine specificity to ensure there is at least a 4 times difference in affinity for the target versus counter-target molecule. Three of the four oligonucleotides demonstrated both high affinity and specificity which are comparatively shown in panel C. To visualize the affinity and specificity of the *de novo* aptamers to the previously published oligonucleotides, panel D is the dose response curves for aptamers 1C, 4C, dnSpikeI, dnSpike1, and dnSpike2 binding both Spike and HA.

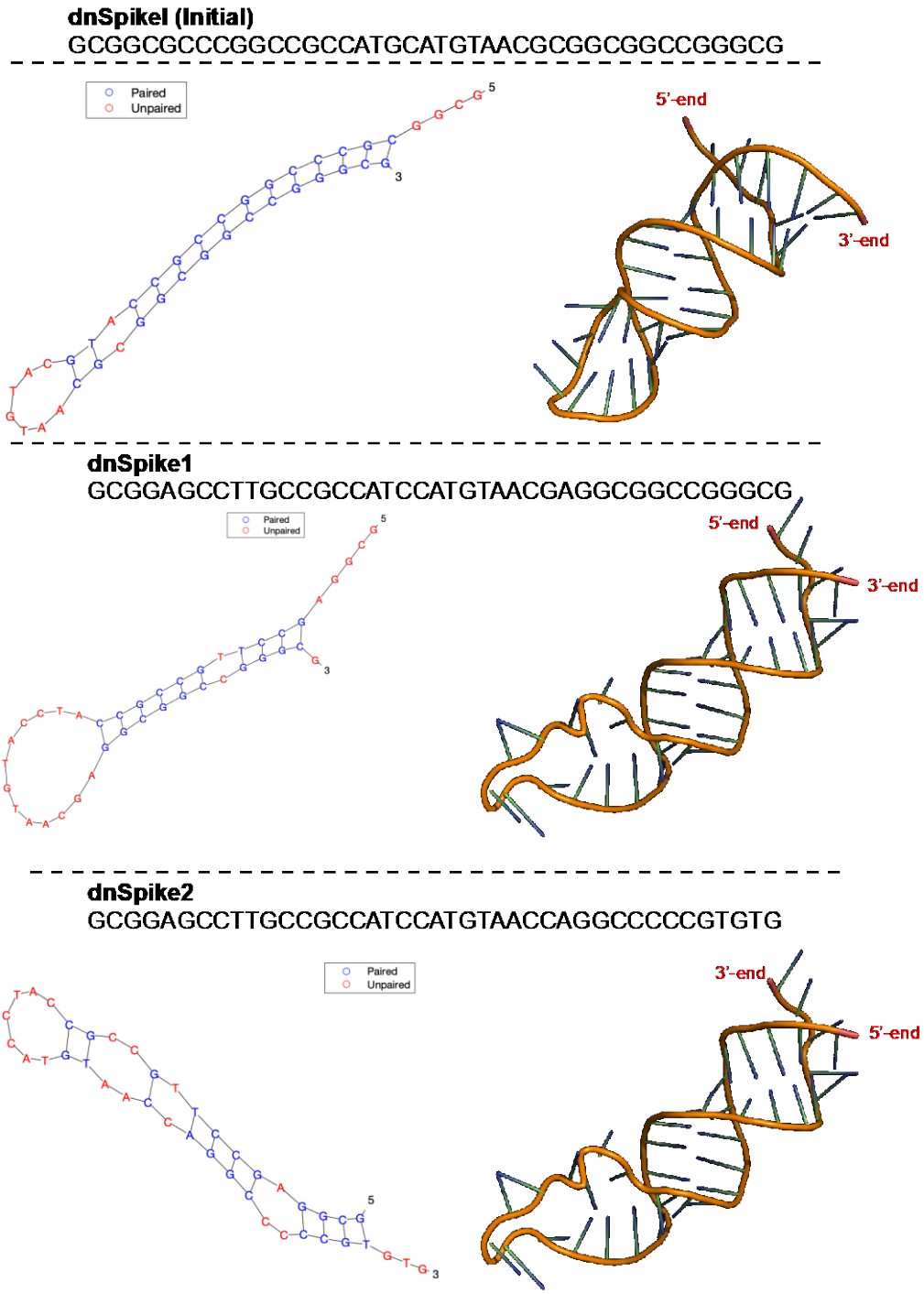


Figure 5-9. Final *de novo* Spike Oligonucleotides.
These three oligonucleotides are the result of Problem 1 demonstrating both high affinity and specificity for their target, Spike, versus their counter-target molecule, HA.

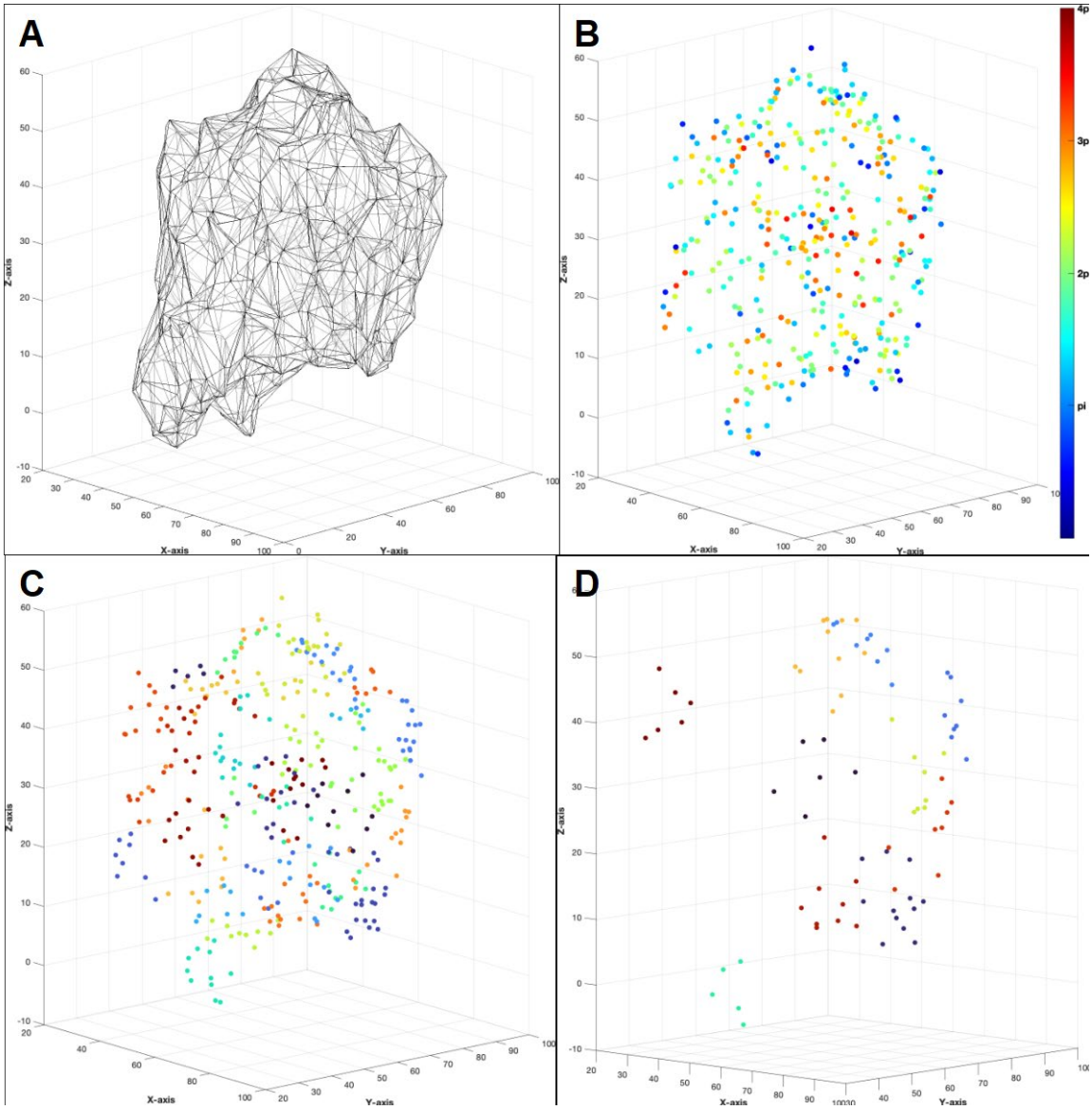


Figure 5-10. *de novo* ACE2 Cold Start Results.
Each panel illustrates the *Cold Start* analysis of the Angiotensin Enzyme 2 (ACE2) protein from the PDB file: 1r42. Panel A shows the Delaunay triangulation of all the atoms. Based on the atom triangulation, panel B provides the solid angle of the alpha shape atoms. Each atom is color coded to represent the solid angle. Panel C depicts the clusters of the biologically relevant and accessible atoms. Panel D illustrates the top ten (smallest solid angle) clusters for the molecule.

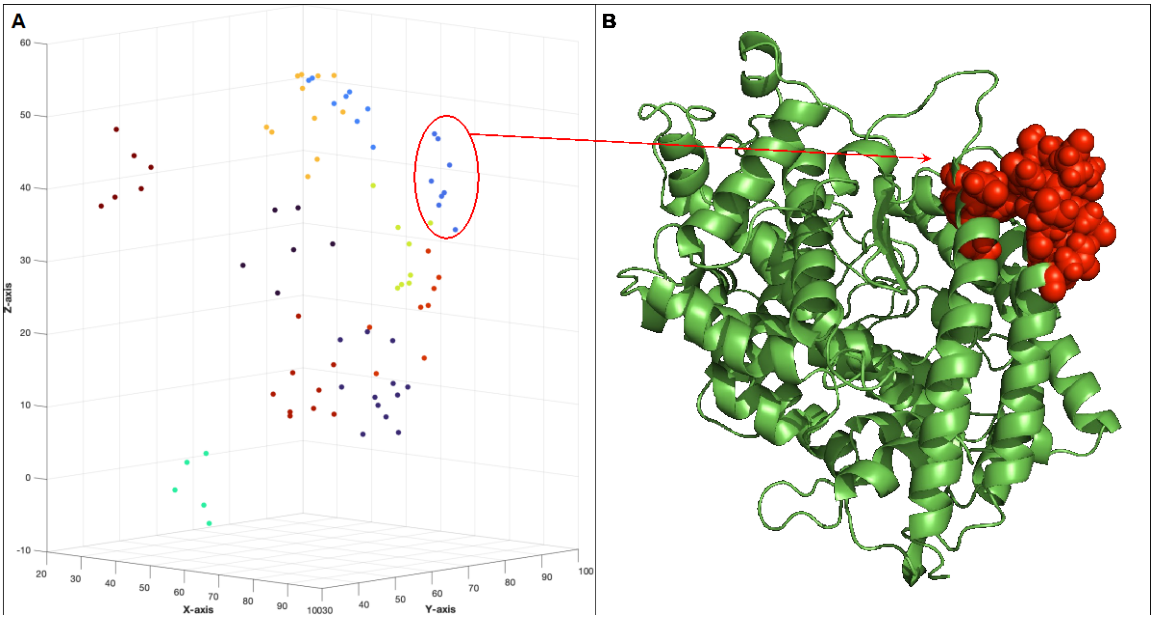


Figure 5-11. *de novo* ACE2 Cluster Visualization.
The top ten clusters are difficult to visualize with respect to the larger molecular structure in MATLAB (panel A). The amino acids of the clusters can however be highlighted and shown using additional pdb file visualization software such as pymol. Panel B highlights one of the top clusters identified in the cold start module with respect to the larger molecule.

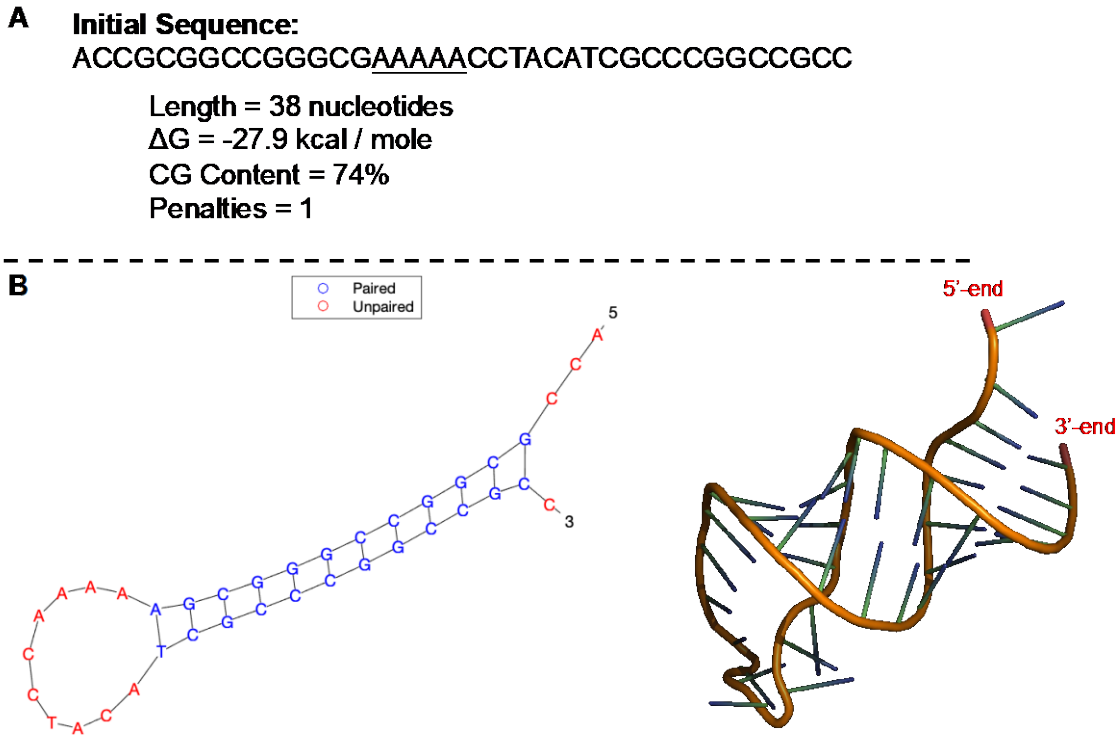


Figure 5-12. *de novo* ACE2 Initial Sequence.
The genetic optimization algorithm generates population of sequences and optimizes the sequences using biological operators such as mutation and crossover. The output of the optimization is an initial sequence that is the most stable confirmation of the oligonucleotide with minimal to no penalties for violating defined constraints. This sequence has one penalty due to a quad nucleotide complex in the core sequence which cannot be altered by the optimization algorithm. The initial sequence for the top cluster for the *de novo* ACE2 aptamer along with its characteristics is shown in panel A while panel B shows the structure in two- and three-dimensions.

de novo ACE2 Options	Values
Positive Target PDB	1r42
Negative Target PDB	None
Protonation	All molecules
Scoring Option	Top 10 Percentile
Fixed Portion	None
Number of Sequences per Iteration	25
Number of Mutations per Sequence	5

Table 2. *de novo* ACE2 Warm Start Options.
The options selected for the *de novo* design of an ACE2 oligonucleotide. This problem was focused on high affinity molecules, so no counter-selection target was employed. Additionally, the top 10 percentile point was used as the scoring metric. Since there was no counter-selection step, the number of sequences per iteration was increased.

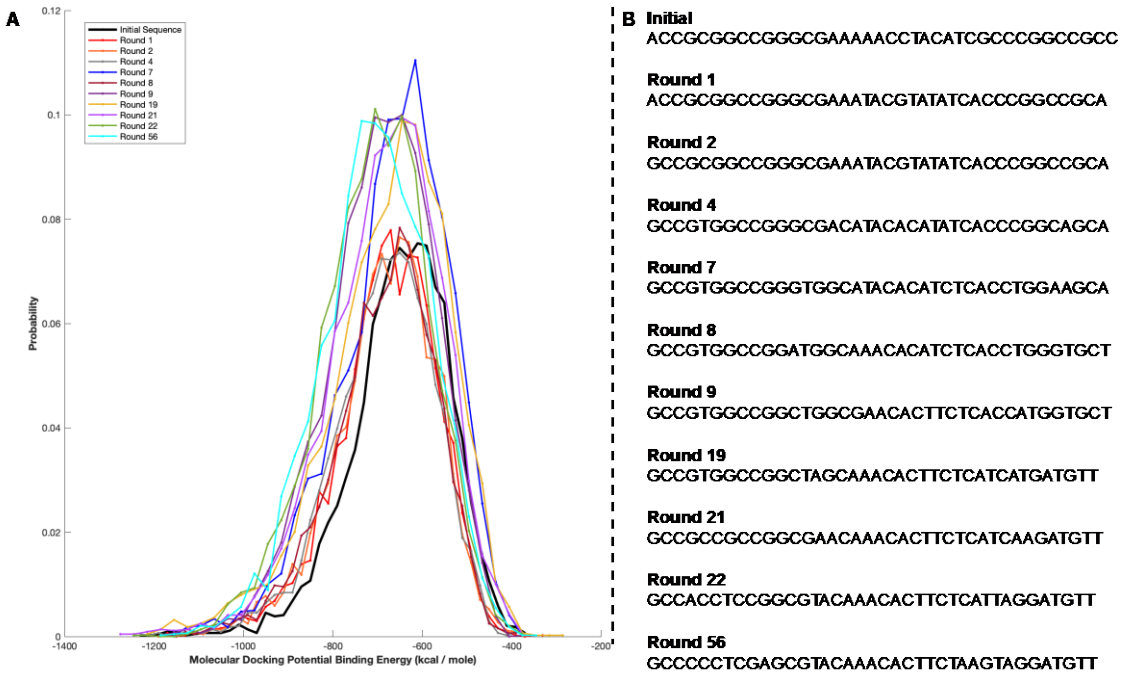


Figure 5-13. Sequence Extraction for *de novo* ACE2.
The *Warm Start* module completed 74 iterations and identified 10 sequences with improved scores out of the 1,924 sequences evaluated. Panel A illustrates the probability distribution of the generated configurations for binding both the target molecule (ACE2). The top 10 percentile score was the scoring metric consequently, the probability distribution plot does not indicate a clear trend with respect to the y-axis however the score distributions do shift to improved binding values. Panel B lists the improved sequences generated over the iterations.

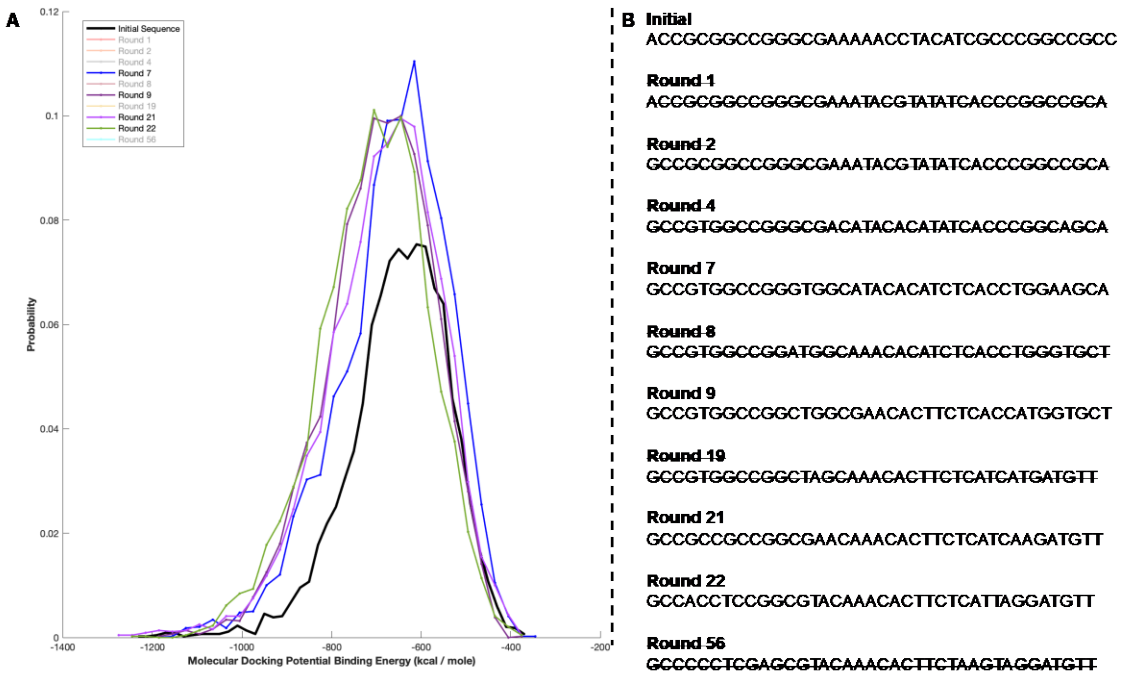


Figure 5-14. *de novo* ACE2 Extracted Sequences for Validation.
Panel A is the probability versus potential energy of interaction plots of the reduced number of sequences that were selected for validation. While Panel B lists the corresponding sequences. Beyond the initial sequence, the ones selected (sequences not lined out) demonstrated the highest probability according to the plot in Panel A.

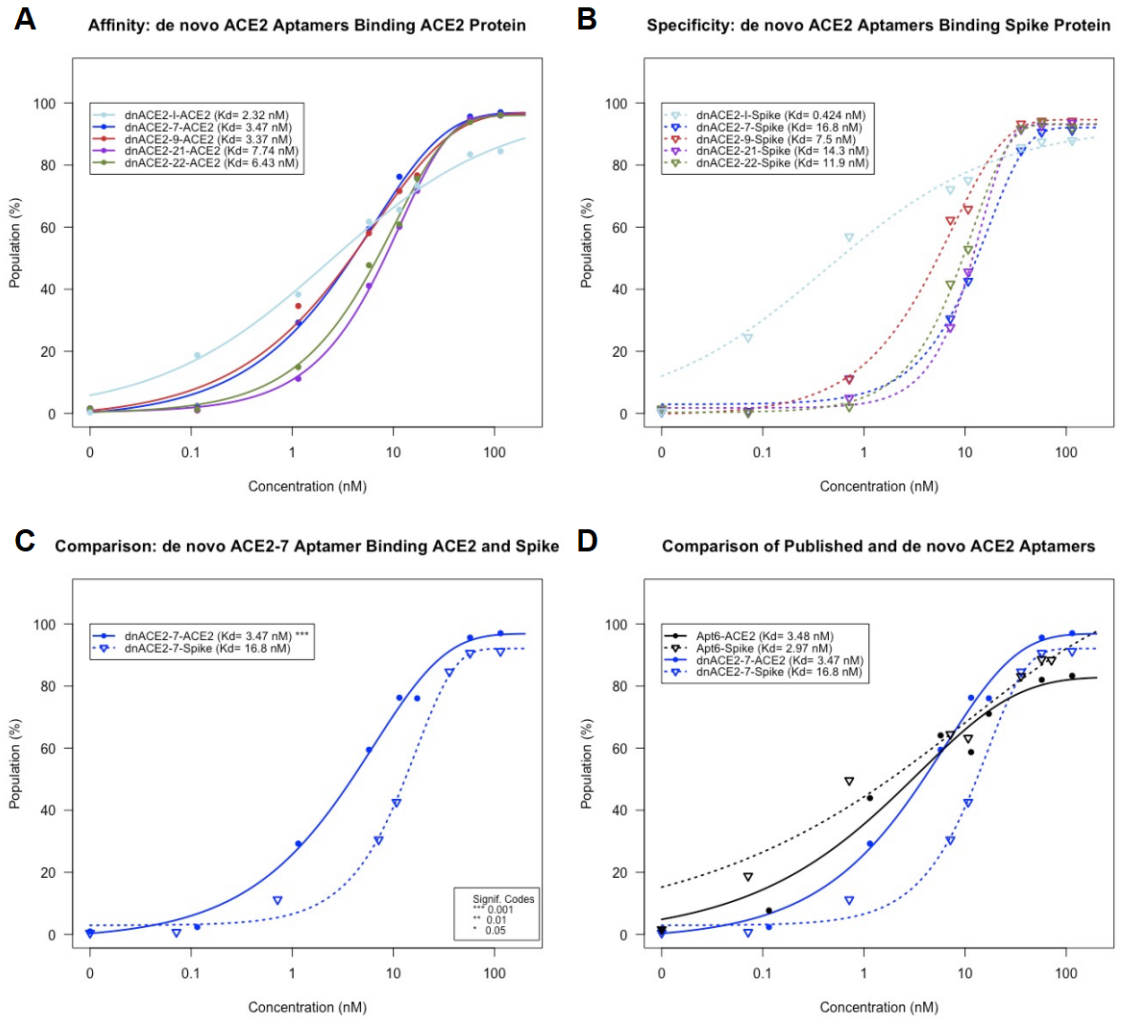


Figure 5-15. *de novo* ACE2 Dose Response Plots.
Five *de novo* ACE2 oligonucleotides (Initial plus 4 perturbed) were chemically synthesized and tested for affinity and specificity using the aptamer on the bead configuration. Panel A is the dose response curves for the four oligonucleotides to ensure high affinity (K_d-value in nanomolar range). Panel B is the dose response curves to examine specificity towards ACE2 versus Spike protein. Only dnACE2-7 oligonucleotide achieved the criteria to be both high affinity and specific as shown in panel C.
To visualize the affinity and specificity of the *de novo* aptamers to the previously published oligonucleotides, panel D is the dose response curves for aptamer 6 and dnACE2-7 binding both ACE2 and Spike.

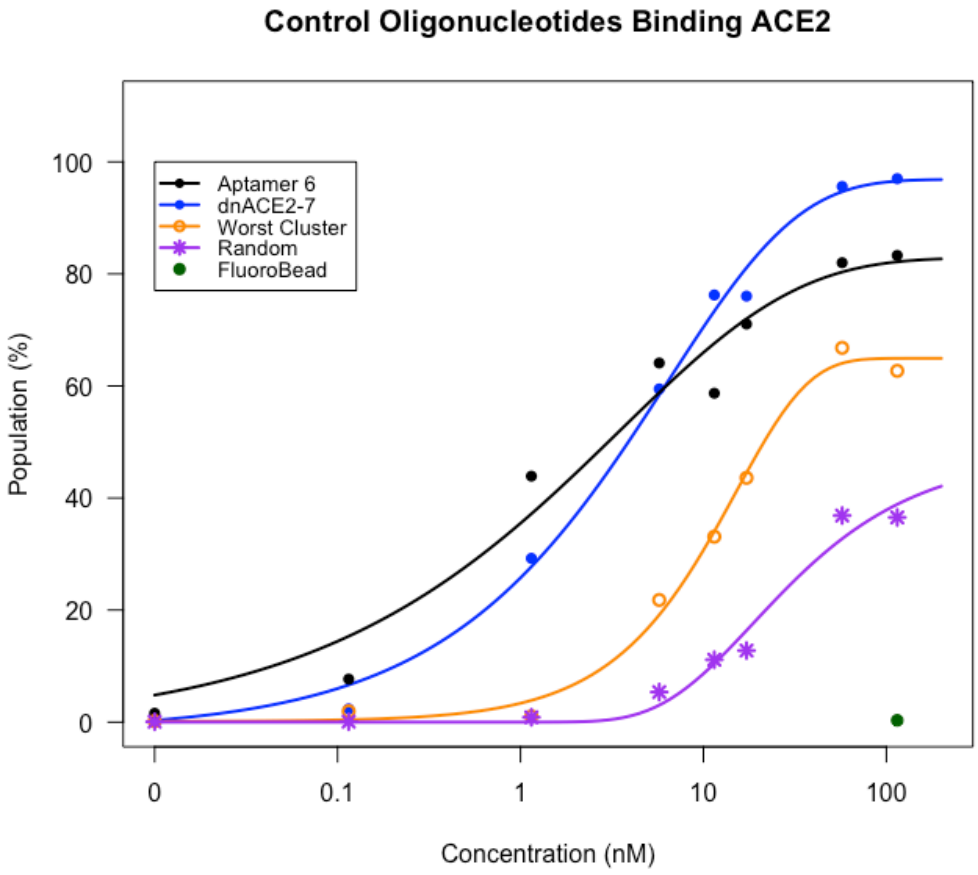


Figure 5-16. *de novo* ACE2 With Control Aptamers.
This dose response curve compares the two control oligonucleotides to the published Aptamer 6 and the dnACE2-7 oligonucleotide. The worst cluster aptamer (orange) was generated from the cluster of amino acids with the largest solid angle making it the least accessible on the target molecule. The random oligonucleotide (purple) is a randomized nucleotide sequence. The FluoroBead complex demonstrates the negligible contribution to the fluorescence positive bead population generated by a streptavidin fluorophore and streptavidin bead interaction.

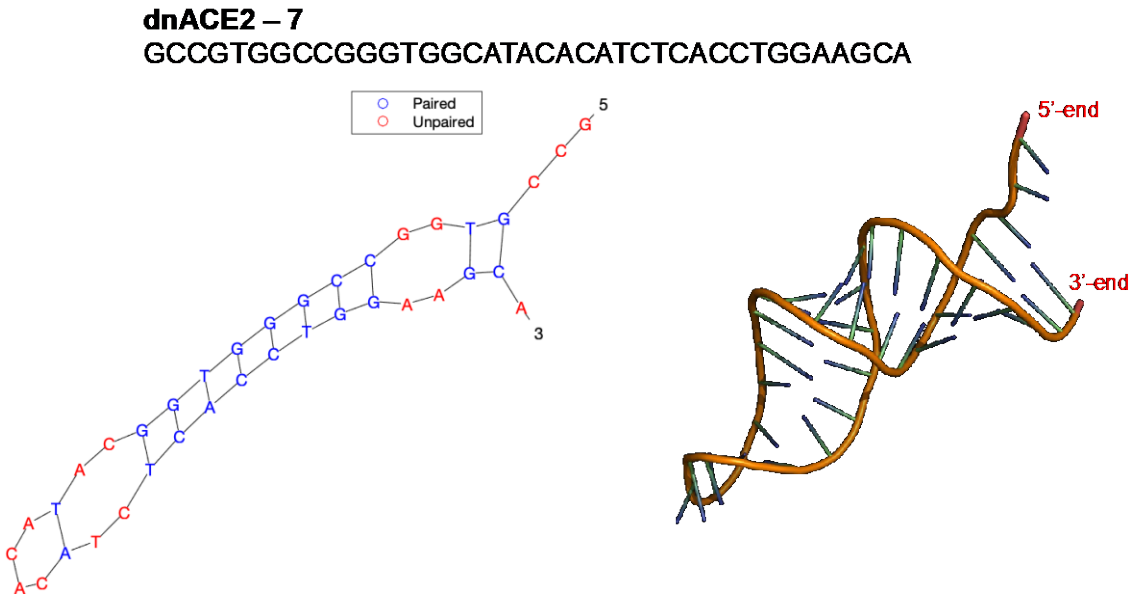


Figure 5-17. Final *de novo* ACE2 Aptamer.
This is the sequence along with both the two-dimensional and three-dimensional structure of dnACE2-7 oligonucleotide. While this molecule did not have the highest affinity for the ACE2 molecule, its affinity was comparable with improved specificity with respect to Spike protein.

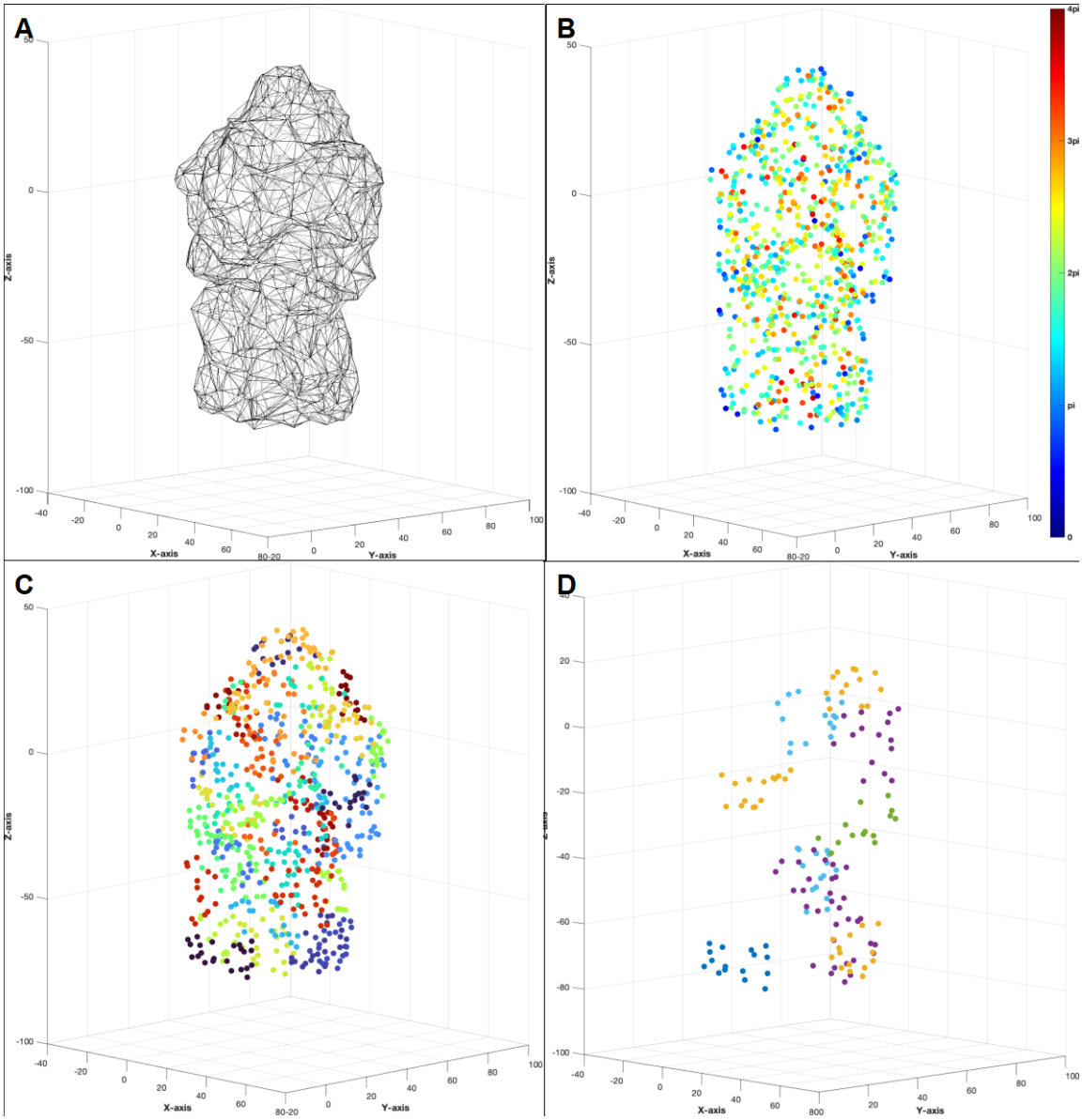


Figure 5-18. *de novo* HA Cold Start Results.
Each panel illustrates the *Cold Start* analysis of the Hemagglutinin (HA) protein from the PDB file: 3lzg. Panel A shows the Delaunay triangulation of all the atoms. Based on the atom triangulation, panel B provides the solid angle of the alpha shape atoms. Each atom is color coded to represent the solid angle. Panel C depicts the clusters of the biologically relevant and accessible atoms. Panel D illustrates the top ten (smallest solid angle) clusters for the molecule.

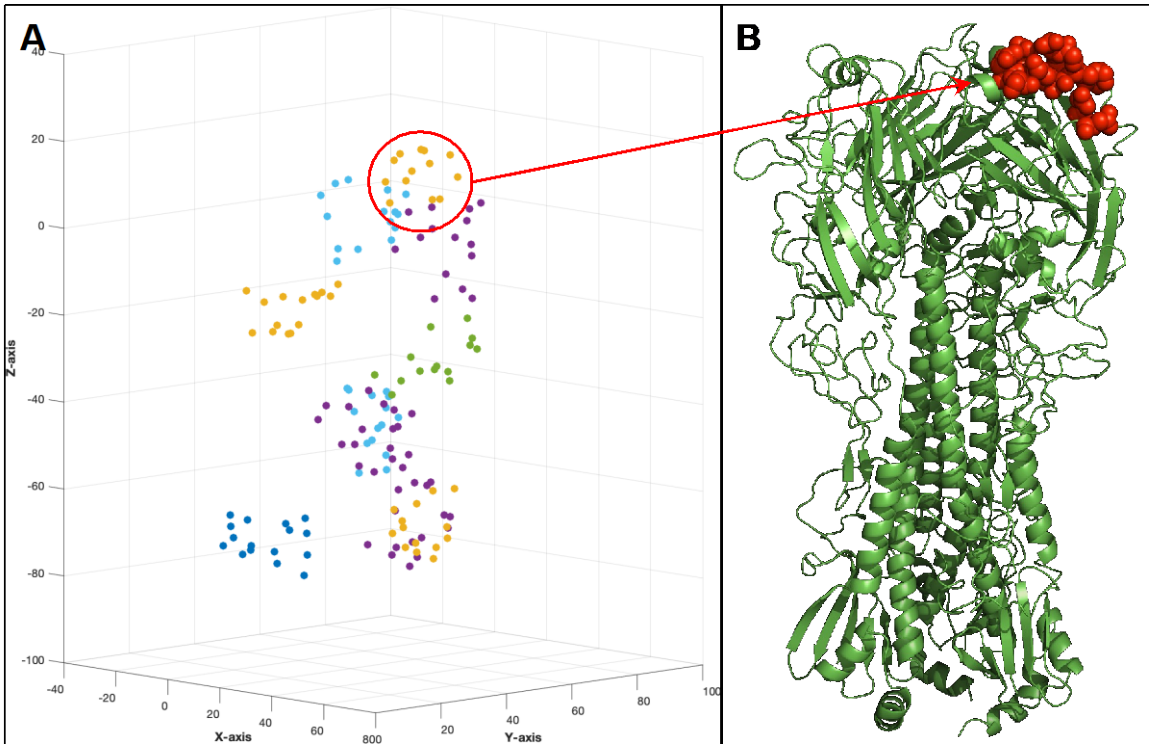


Figure 5-19. *de novo* HA Cluster Visualization.
The top ten clusters are difficult to visualize with respect to the larger molecular structure in MATLAB (panel A). The amino acids of the clusters can however be highlighted and shown using additional pdb file visualization software such as pymol. Panel B highlights the top cluster, one with the smallest total solid angle, as identified in the Cold Start module with respect to the larger molecule. This cluster corresponds contains several amino acids that were previously identified as part of the HA receptor binding domain.

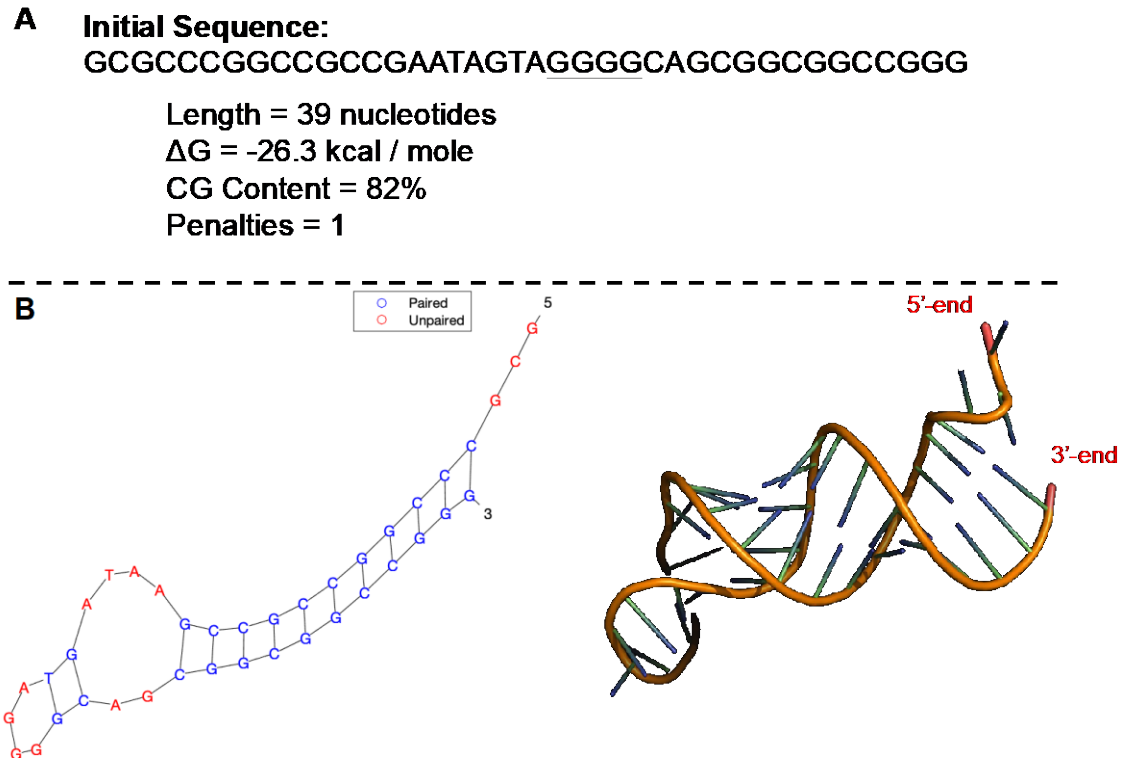
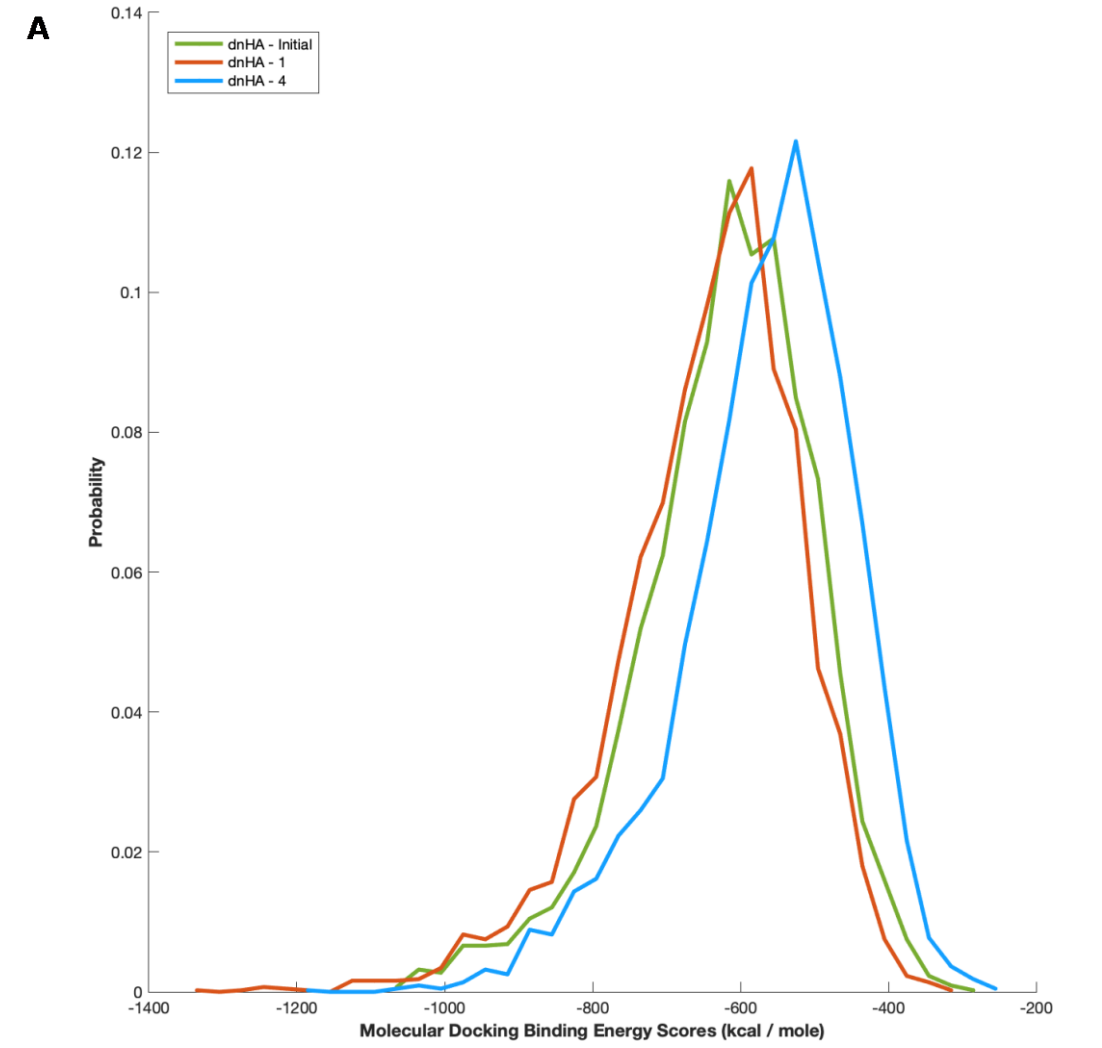


Figure 5-20. *de novo* HA Initial Sequence.
The genetic optimization algorithm generates population of sequences and optimizes the sequences using biological operators such as mutation and crossover. The output of the optimization is an initial sequence that is the most stable confirmation of the oligonucleotide with minimal to no penalties for violating defined constraints. This sequence has one penalty due to a quad nucleotide complex in the core sequence which cannot be altered by the optimization algorithm. The initial sequence for the top cluster for the *de novo* HA aptamer along with its characteristics is shown in panel A while panel B shows the structure in two- and three-dimensions.

***de novo* HA Options**	**Values**
Positive Target PDB	3lzg
Negative Target PDB	None
Protonation	All molecules
Scoring Option	Max Probability
Fixed Portion	None
Number of Sequences per Iteration	25
Number of Mutations per Sequence	5

Table 3. *de novo* HA Warm Start Options.
List of the options employed in the *de novo* HA aptamer design *Warm Start* module.



B

Initial
GCGCCCGGCCGCCGAATAGTAGGGGCAGCGGCGGCCGGG

Round 1
GCGCCCGGCTGCCAAATATTAGGGCCAGCGGCTGCCGGG

Round 4
GCGCCCGGCTCCCAAAGAGTAAGGCCAGCGGCCGCCGGG

Figure 5-21. Sequence Extraction for *de novo* HA.
The *Warm Start* module completed only 5 iterations and identified 2 sequences with improved scores out of the 130 sequences evaluated. Panel A illustrates the probability distribution of the generated configurations for binding the target molecule (HA). The maximum probability score was the scoring metric. Panel B lists the improved sequences generated over the iterations.

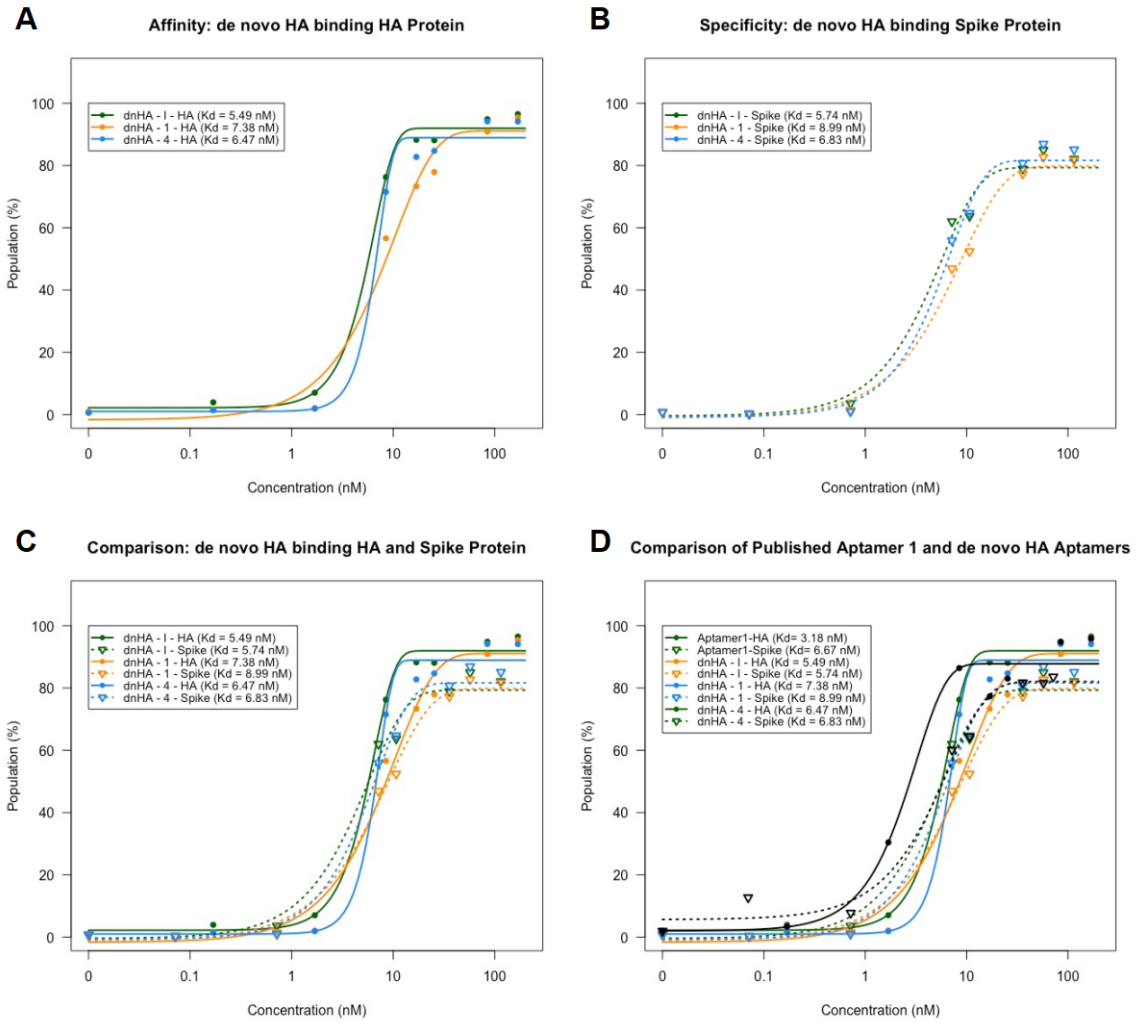


Figure 5-22. *de novo* HA Dose Response Plots.
Three *de novo* HA oligonucleotides (Initial plus 2 perturbed) were chemically synthesized and tested for affinity and specificity using the aptamer on the bead configuration. Panel A is the dose response curves for the four oligonucleotides to ensure high affinity (K_d-value in nanomolar range). Panel B is the dose response curves to examine specificity towards HA versus Spike protein. No oligonucleotide met the threshold to be considered specific. Panel C compares the binding of the *de novo* HA aptamers to both HA (solid line) and Spike (dotted line). To visualize the affinity and specificity of the *de novo* aptamers to the previously published oligonucleotides, panel D is the dose response curves for all the aptamers plus the published Aptamer 1.

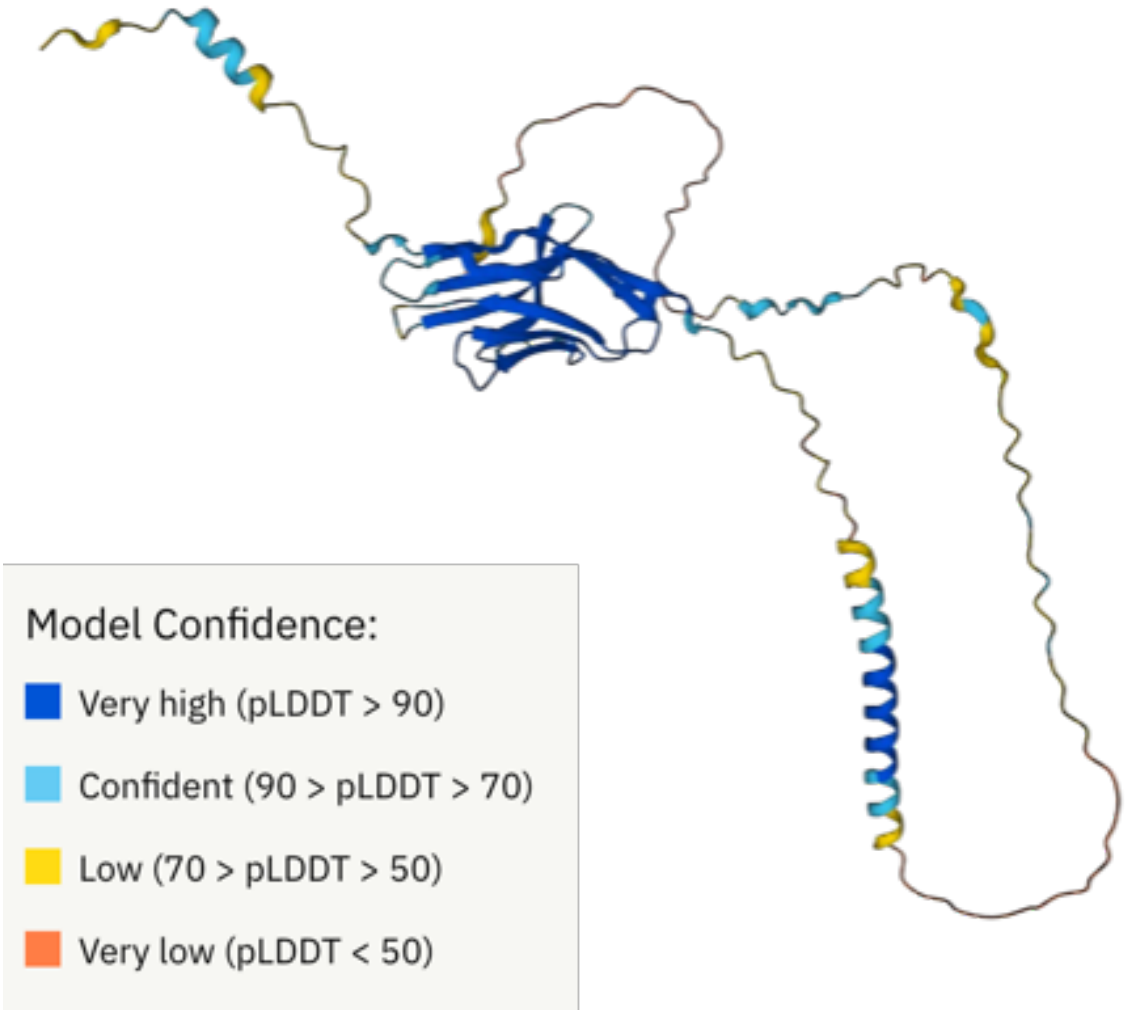


Figure 5-23. AlphaFold PD1 Predicted Structure.
This is the AlphaFold predicted structure for the Protein Death 1 (PD1) receptor (Q15116) with the corresponding model confidence. The receptor is 288 amino acids in length. The region identified for binding is labelled as very high confidence according to the pLDDT score. The pLDDT score is a confidence estimate (0 – 100) based on the per-residue local distance difference test.

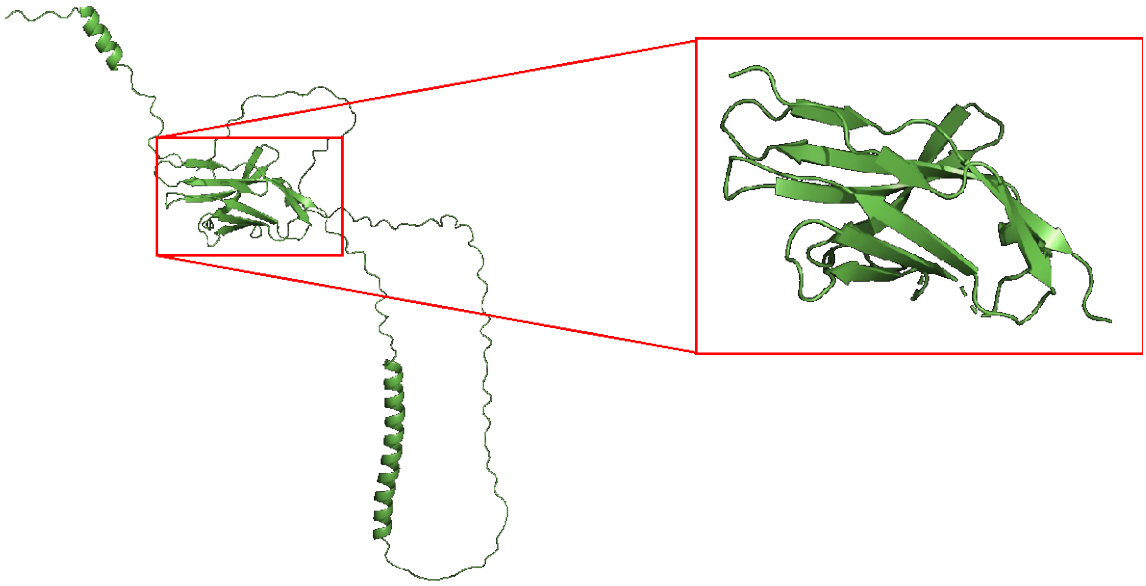


Figure 5-24. Truncation of PD1 Structure.
Due to the hydrophobic anchors, the PD1 was truncated using PyMol to only the extracellular portion of the receptor. This truncation process reduced the structure from 288 amino acids to only 116 for the actual design process.

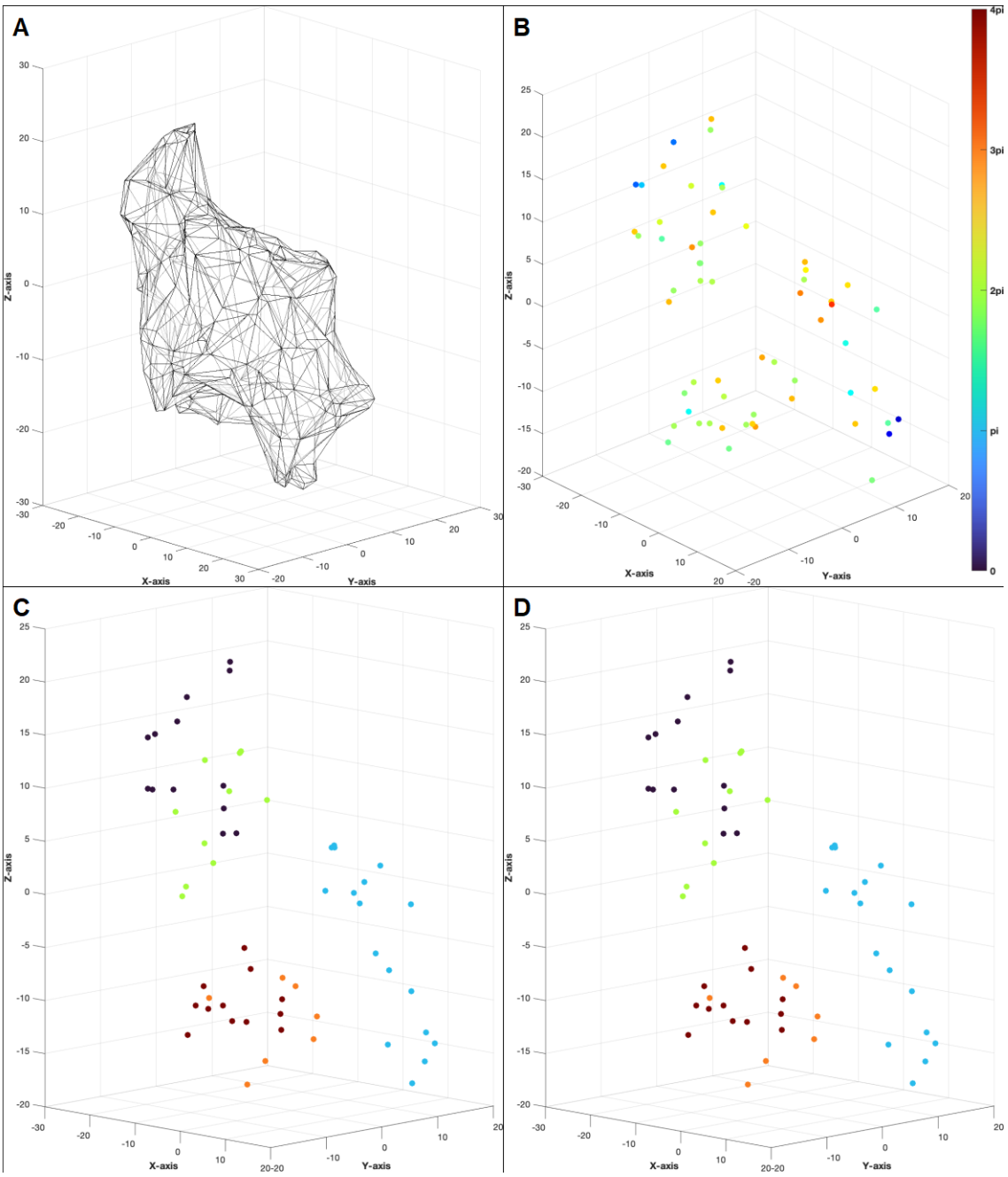


Figure 5-25. *de novo* PD1 *Cold Start* Results.
Each panel illustrates the *Cold Start* analysis of the Protein Death 1 (PD1) receptor from the AlphaFold generated PDB file: Q15116. Panel A shows the Delaunay triangulation of all the atoms. Based on the atom triangulation, panel B provides the solid angle of the alpha shape atoms. Each atom is color coded to represent the solid angle. Panel C depicts the clusters of the biologically relevant and accessible atoms, which is the same as the top clusters in Panel D since there are less than 10 clusters.

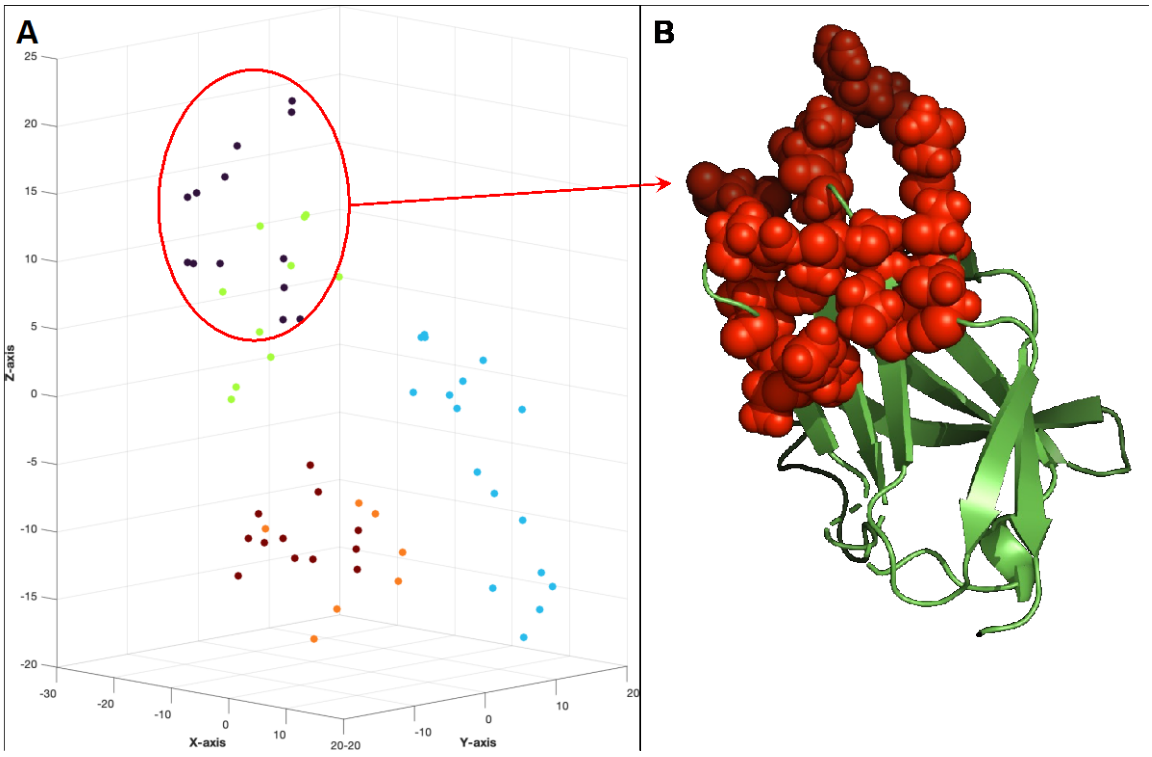


Figure 5-26. *de novo* PD1 Cluster Visualization.
The top clusters are difficult to visualize with respect to the larger molecular structure in MATLAB (panel A) since the rest of the structure is not visible. The amino acids are visualization software such as pymol. Panel B highlights the top cluster, one with the smallest total solid angle, as identified in the Cold Start module with respect to the larger molecule. The molecule rotated so Panel A and B have the same orientation.

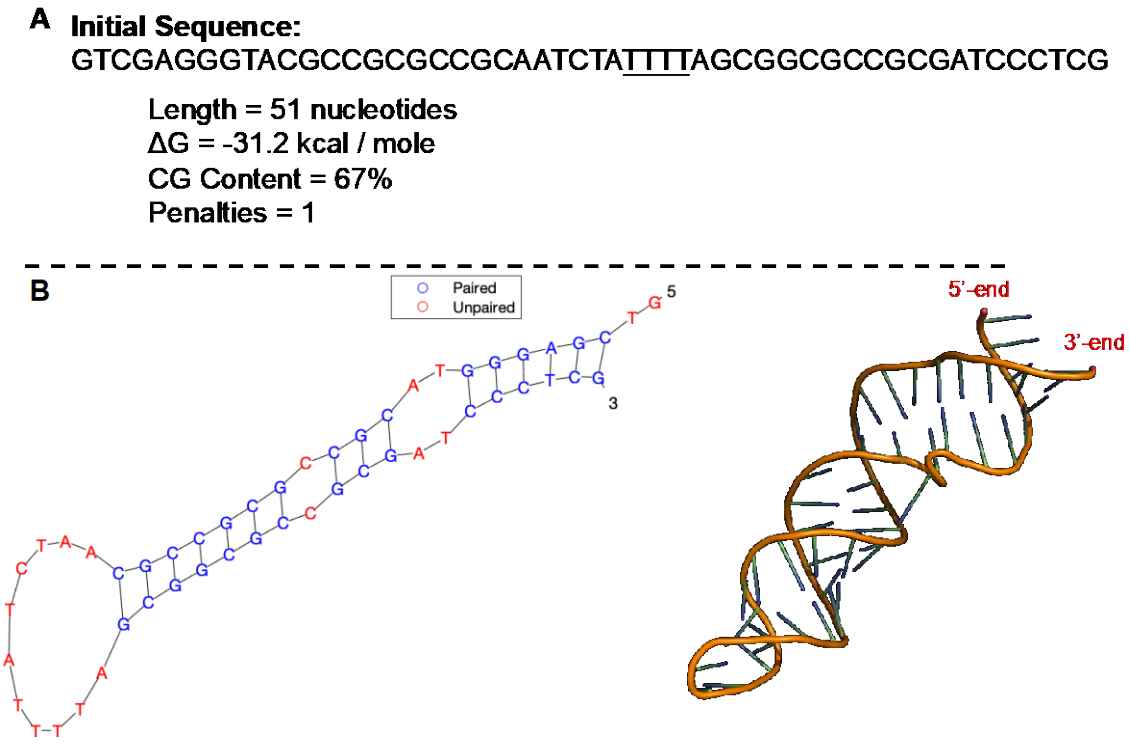


Figure 5-27. *de novo* PD1 Initial Sequence.
The genetic optimization algorithm generates population of sequences and optimizes the sequences using biological operators such as mutation and crossover. The output of the optimization is an initial sequence that is the most stable confirmation of the oligonucleotide with minimal to no penalties for violating defined constraints. This sequence has one penalty due to a quad nucleotide complex in the core sequence which cannot be altered by the optimization algorithm. The initial sequence for the top cluster for the *de novo* PD1 aptamer along with its characteristics is shown in panel A while panel B shows the structure in two- and three-dimensions.

***de novo* PD1 Options**	**Values**
Positive Target PDB	AF-Q15116
Negative Target PDB	None
Protonation	All molecules
Scoring Option	Max Probability
Fixed Portion	None
Number of Sequences per Iteration	25
Number of Mutations per Sequence	5

Table 4. *de novo* PD1 Warm Start Options.
List of the options employed in the *de novo* PD1 aptamer design *Warm Start* module.

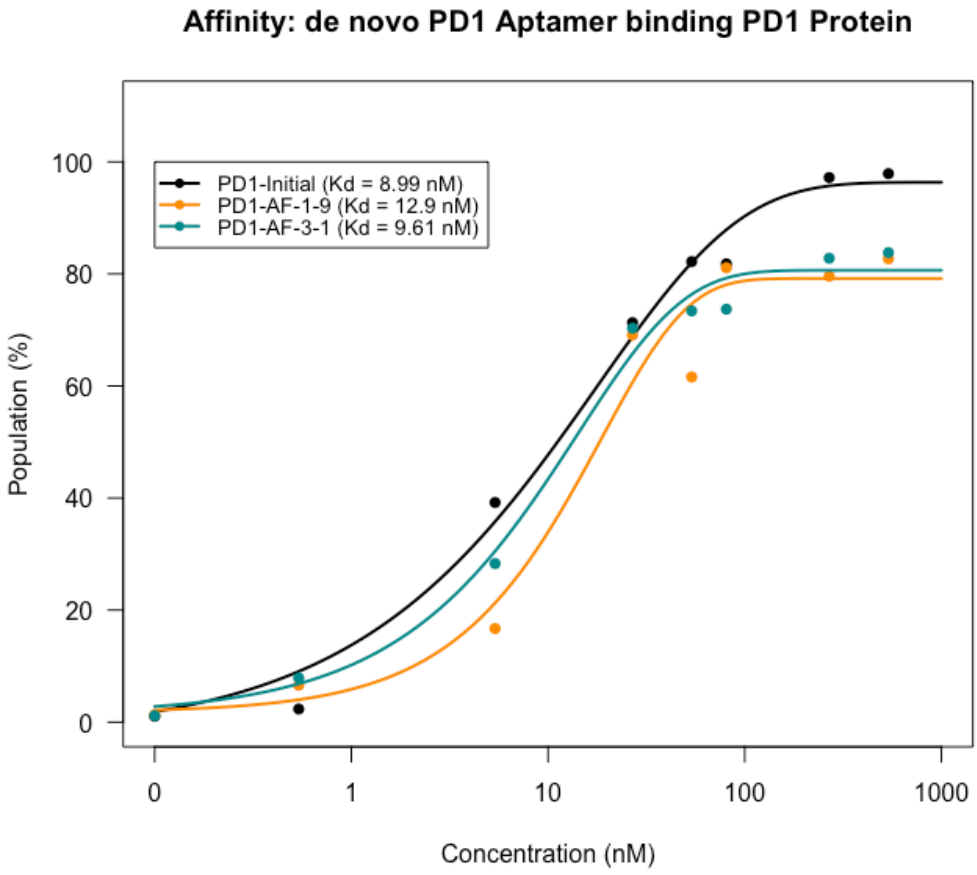
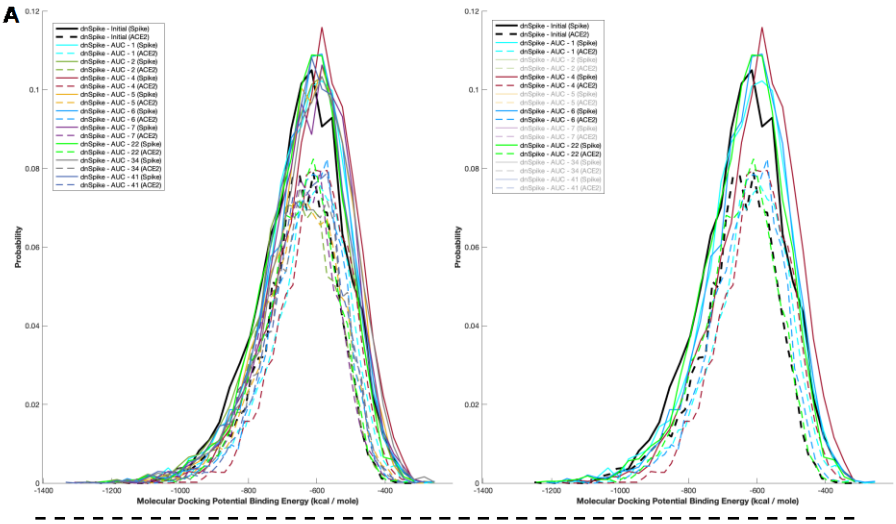


Figure 5-28. *de novo* PD1 Affinity Validation.
The dose response curve shows the response of the initial aptamer and two additional aptamers generated in the *Warm Start* module. Since the goal was to test the concept of employing the structure from AlphaFold, only affinity was initially tested. More *in vitro* validation experiments are required.

de novo Spike – ACE2 Options	Values
Positive Target PDB	6vsb
Negative Target PDB	1r42
Protonation	All molecules
Scoring Option	Area Under Curve
Fixed Portion	None
Number of Sequences per Iteration	25
Number of Mutations per Sequence	5

Table 5. *de novo* Spike with ACE2 Specificity Warm Start Options.
List of the options employed in the *de novo* Spike with specificity against ACE2 aptamer design *Warm Start* module.



Initial Sequence:
GCGGCGCCCGGCCGCCATGCATGTAACGCGGCGGCCGGGCG

dnSpike-AUC-1
TCTGCGCCCGGCCGCGATGCAGGTAACGCGGCGGACGGGCG

dnSpike-AUC-2
CCTGCGCCCGGCCGCGATACAGGTAGCGCGTCGGACGTGCG

dnSpike-AUC-4
CCTGCGCCCGCCCGCGATATACGTAGCCCGTCGGACATGCG

dnSpike-AUC-5
CCCGCGGCCGCCCGCGATATACGTAGCGGGTCGGAGATGCG

dnSpike-AUC-6
GCCGCCGCCGCCCGCGATATACGTAGTGGGTCGGTCATGCG

dnSpike-AUC-7
GCTGCCGCCCCCCGCGTTATACGTAGTGGGTCGGACATACG

dnSpike-AUC-22
GCTGCCGCCCCCCGCGGGAAGCGTAGTGGGTCTGACATACG

dnSpike-AUC-34
GCTGCCTCCCCCCGGAGGAAGCTTAGTGGGTCGGACATACG

dnSpike-AUC-41
GCTGCCTCCCCCCAGAGGAAGCTGAGTTAGTCGGACGTACG

Figure 5-29. Sequence Extraction for *de novo* Spike with ACE2 Specificity.
The *Warm Start* module employs a molecular docker to evaluate perturbed sequences against the target and counter-target molecule. Panel A illustrates the probability distribution of the generated configurations for binding both the target molecule (Spike)(solid line) and counter-target (ACE2)(dotted line). Each line represents the best selected sequence and the round when the sequence led to an improved score. The sequences are listed in panel B.

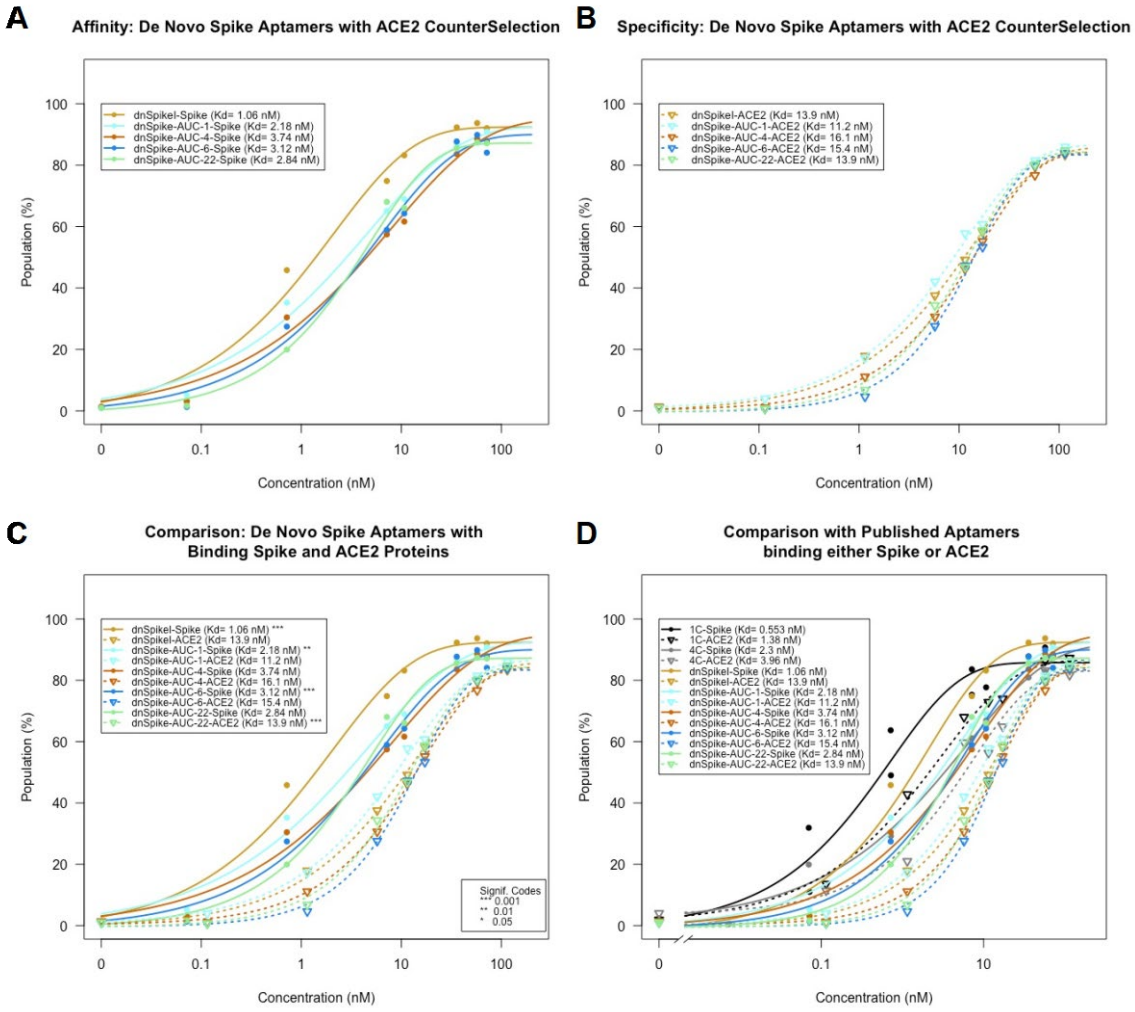


Figure 5-30. *de novo* Spike with ACE2 Specificity Dose Response Plots.
Five *de novo* Spike oligonucleotides (Initial plus 4 perturbed) were chemically synthesized and tested for affinity and specificity using the aptamer on the bead configuration. Panel A is the dose response curves for the five oligonucleotides to ensure high affinity (K_d-value in nanomolar range). Panel B is the dose response curves to examine specificity to ensure there is at least a 4 times difference in affinity for the target versus counter-target molecule. All five oligonucleotides demonstrated both high affinity and specificity which are comparatively shown in panel C. To visualize the affinity and specificity of the *de novo* aptamers to the previously published oligonucleotides, panel D is the dose response curves for aptamers 1C, 4C, dnSpikeI, dnSpike-AUC-1, and dnSpike-AUC-4, dnSpike-AUC-6, and dnSpike-AUC-22 binding both Spike and ACE2.

6.0 CHAPTER 6 – DISCUSSION / CONCLUSIONS

This chapter summarizes the digitalSELEX validation results. These results emphasize how the platform confronts the existing challenges of other selection methods. This chapter further discusses opportunities for improvements, lessons learned, and future applications.

6.1 PLATFORM VALIDATION OVERVIEW

The digitalSELEX validation process led to the chemical synthesis of 19 oligonucleotides from five proposed problems. The K_d -value for the target (affinity) and the counter-target (specificity) for each oligonucleotide along with the previously published aptamers was determined. The details are summarized in **Table 6** and discussed below. In summary, the *de novo* design process produced 9 oligonucleotides which met the criteria to be both high affinity and specific.

The goal of Problem 1 was to design an oligonucleotide with both high affinity for Spike and specificity with respect to HA. The purpose of such an oligonucleotide is to be able to distinguish between Spike and HA for detection of either SARS-CoV-2 virus or Influenza. The digitalSELEX process designed 3 oligonucleotides, dnSpike-I, dnSpike-1 and dnSpike-2, that met the high affinity and specificity requirements. The K_d -values for Spike were 1.06 nM, 1.72 nM and 0.425 nM respectively, while their K_d-

value for HA were 6.59 nM, 7.67 nM and 8.1 nM. These values met the fold difference definition for specificity with fold differences of 6.2, 4.45, and 19 respectively.

The goal of Problem 2 was to design an oligonucleotide with affinity for ACE2 on par or better than the published aptamer (aptamer 6).[189] This published aptamer was not selected with Spike as the counter-target molecule. Empty streptavidin beads were used as the counter-target. No counter-selection target was employed in the Counter-Selection module, but specificity against Spike was still experimentally assessed. This was done to examine if the initial sequence or scoring function for affinity would inherently lead to specificity. Four oligonucleotides were synthesized and all four met the benchmarks for high affinity. Of the four oligonucleotides, dnACE2-7, had a K_d -value for the ACE2 molecule of 3.47 nM and was specific. This designed aptamer had a K_d-value for Spike of 16.8 nM whereas the K_d-value for aptamer 6 against Spike is 2.97 nM.

The goal of Problem 3 was to design an aptamer with similar K_d-value as Aptamer 1 toward the HA protein. No counter-selection target was employed during the design or counter-selection process, but specificity against Spike was once again assessed. The three designed oligonucleotides have single digit nanomolar K_d-values which corresponds to Aptamer 1 affinity. The K_d-values for the aptamers against Spike did not meet the specificity threshold. *Warm Start* was terminated prematurely due to power fluctuation resetting the server. The terminated simulation has since been restarted, but the validation experiments have not been completed.

Problem 4 had a unique design goal to generate high affinity oligonucleotides for PD1 using the predicted AlphaFold structure. No counter-selection target was employed either during the *Warm Start* module or during *in vitro* validation. In the end, all three

oligonucleotides synthesized and validated met the high affinity criteria. Two aptamers had K_d-values in the single digit nanomolar range while the third was less than 15 nM. These results show potential for using predictive structures for design of high affinity oligonucleotides. Specificity is still an objective that needs to be examined when using predictive structures.

Problem 5 built upon Problem 1 and sought to design an oligonucleotide that was high affinity for Spike yet specific against ACE2 protein. This problem used the initial sequence generated in the Cold Start module and genetic optimization algorithm in Problem 1. After completing 60 iterations, four of the possible nine oligonucleotides plus the initial sequence were synthesized for validation. All five oligonucleotides tested demonstrated both high affinity and specificity.

There were four negative controls implemented throughout the digitalSELEX development and validation process. The first negative control demonstrated that unbound streptavidin fluorophore did not interact with the streptavidin magnetic beads. The unbound fluorophore does not contribute to the positive fluorescent bead population, as shown in **Figure 2-9**. Another control sought to determine the effect of non-specific binding between the aptamer-bead complex and the protein-fluorophore complex. Four different aptamer-bead complexes were tested against their target molecules with and without BSA (100 ng /ml). There was no significant difference between the two conditions for the four different aptamer-bead complexes, shown in **Figure 2-7**. These two controls suggest that the percent positive population as determined by flow cytometry comes from the interaction between the aptamers and the protein targets and not unbound fluorophore or non-specific binding.

The final set of controls were implemented during the platform validation. These two controls were specific aptamers; either a random generated sequence or a sequence derived from the worst cluster identified in the Cold Start module. The results illustrated in **Figure 5-16** show reduced affinity and percent binding populations for these two aptamers. This suggests the affinity and specificity of the *de novo* sequences are derived from the process of the digitalSELEX platform and are not unintended.

6.2 CHALLENGES ADDRESSED

As stated in Chapter 1, molecules that are both high affinity and specific for a given target have tremendous application as either a therapeutic or as a biosensor probe. Aptamers, or single-strand oligonucleotides, are molecules that can be both specific and high affinity. The Systematic Evolution of Ligands by EXponential enrichment (SELEX) has been the gold standard for aptamer identification for over 30 years. While a multitude of selection variations, both *in vitro* and with *in silico*, have been developed over the years, challenges remain. These challenges are initial pool size, time, cost, relatively low success rate, data dependence and ensuring the product is both high affinity and specific.

The digitalSELEX platform disrupts the selection paradigm and confronts these challenges. This *in silico* methodology does not rely on the hope that a sequence exists in partial or incomplete randomized library. The digitalSELEX platform instead designs oligonucleotides by focusing on potential interactions.

The *Cold Start* module identifies clusters of amino acids based on their atom accessibility and binding relevance. Nucleotides are assigned to clusters of accessible, binding relevant amino acids. These assigned nucleotides become the core sequence of the oligonucleotide. The core oligonucleotide sequence has additional nucleotides added to the 5'- and 3'-end to achieve the desired aptamer length.

The oligonucleotide is then optimized with the genetic optimization algorithm to find the most stable structure that minimizes the application specific constraints. The most stable sequence becomes the initial sequence for the *Warm Start* module. Simultaneous random perturbations are introduced to the initial sequence to generate number of mutated sequences. The potential interaction between the sequences and both the target and counter-target molecule are evaluated using a molecular docker. The best sequence, according to the scoring function, is selected and undergoes further perturbations. This iterative process explores the interaction landscape between the sequence and target molecules, selecting the predicted best overall sequence.

This process not only eliminates potential initial oligonucleotide library challenges, but the computational time is faster than *in vitro* selection time. The time from initial target to sequence validation can be further reduced with greater computational capacity. There is no cost associated with executing the digitalSELEX platform. Unlike other SELEX methods, the only cost incurred stems from the chemical synthesis and validation of the oligonucleotides. In the end, the platform provides only a handful (1 – 10 sequences) of optimized sequences for validation; not $10^3 – 10^4$ possible sequences.

Regarding high affinity and specificity, each validation problem presented in Chapter 5 achieved its end-state and in several cases, it surpassed its initial goal. All 19 synthesized oligonucleotides met the criteria for high affinity. Additionally, nine of the 16 oligonucleotides tested for specificity met that criterion as well. Of the two problems specifically directed towards high affinity and specificity (problems one and five), eight of the nine oligonucleotides met both criteria.

6.3 OPPORTUNITIES FOR IMPROVEMENT

Even though the digitalSELEX platform exceeded many goals of the validation process, there are opportunities for improvement. The first improvement is the validation of the Clustering Algorithm 2 in the *Cold Start* module. The algorithm has shown interesting preliminary results, but further testing followed by a complete validation problem is required.

The second opportunity for improvement is direct towards the molecular docker and scoring function. The data presented in Chapter 2 demonstrated that there is a relationship between the docking scores and K_d-values. This work however did not explore all the scoring functions that were developed later. Specific work regarding the scoring function should enable better sequences to be designed in fewer iterations.

The third improvement opportunity is the complete validation of AlphaFold structures. This validation requires comparing the clusters, oligonucleotides sequences, and K_d-values between aptamers generated with both AlphaFold and pdb files for the

same molecules. The incorporation of AlphaFold would enable digitalSELEX platform to develop biosensor probes faster than other existing methods.

The fourth improvement opportunity is not solely constrained to the digitalSELEX platform. This improvement is to examine the ideal oligonucleotide length versus target molecule size and corresponding application. As highlighted in Chapter 2, there are application as well as *steric hindrance* considerations for designing oligonucleotides. The length of oligonucleotides for the same target and application varies with no clear underlying rationale. A first order approximation of oligonucleotide length would limit the number of iterations required and sequences per round. This would ultimately reduce computational time.

6.4 LESSONS LEARNED

There are three key lessons learned regarding *in silico* oligonucleotide design that were identified during platform validation process. First, selecting the correct cluster for initial sequence generation can promote specificity. As previously mentioned, the degenerative relationship between the combinations of amino acids to nucleotides means that a single nucleotide string can potentially interact with multiple amino acids strings. Selecting unique clusters on the target versus counter-target should provide an innate level of specificity prior to the counter-selection module. This lesson learned drove the development of Algorithm 2 in the *Cold Start* clustering step, however it still needs to be validated.

The second lesson learned, which also drove an opportunity for improvement, is to consider the oligonucleotide length. As discussed in detail in Chapter 2, there is a relationship between the size of the oligonucleotide and its ability to interact with targets. With larger oligonucleotides there is potential *steric hindrance* which prevents binding between the aptamer and target. On the other hand, an aptamer that is too short could reduce functionality in its application. This lesson learned is evident in several of the published aptamers such as 1C, 4C for Spike and the PD4S for PD1 where the authors reduced the length of the aptamers by removing the primers and the K_d-value improved. Employment of the digitalSELEX platform requires consideration of the application and the overall oligonucleotide length as well as the core sequence length.

The third lesson learned is computational capacity and stability. The digitalSELEX platform can generate high affinity and specific oligonucleotides with greater efficiency than other selection methods. The number and quality of CPUs reduce the computational time required. When identifying *Warm Start* options, it is necessary to know the computational limits when selected the number of sequences to be generated and the number of counter-targets.

6.5 FUTURE APPLICATIONS

The goal in developing the digitalSELEX platform was to address lingering challenges of previous selection methods. The digitalSELEX platform designed nine oligonucleotides with high affinity and specificity for their target versus counter-target, as well as several high affinity oligonucleotides. The design constraints and validation steps

were focused on employing these oligonucleotides as biosensor probes, however, other applications of the platform / process are possible. The first application would be to design oligonucleotides for therapeutic purposes.

Another application is to modify the platform to design small peptides. This application requires adding an amino acid assignment table and modifying the simultaneous random perturbation step. AlphaFold can predict the three-dimensional structure of peptides which would enable molecular docking. An additional application is the platform could be used to design oligonucleotides against specific chemical compounds. The program ChemDraw can generate a three-dimensional structure. This would enable the digitalSELEX platform to be employed to develop probes for sensors to detect both biological and chemical weapon agents.

Problem	Aptamer ID	Target Molecule	Target K_d-value (nM)	High Affinity	Counter-Target Molecule	Counter-Target K_d-value (nM)	Specific
1	Aptamer 1C (pre-published)	Spike	0.553	Yes	HA (not selected)	0.699	No
	Aptamer 4C (pre-published)		2.3	Yes	HA (not selected)	1.47	No
	dnSpike – 1		1.06***	Yes	HA	6.59	Yes
	dnSpike – 1		1.72	Yes		7.67	Yes
	dnSpike – 2		0.425***	Yes		8.1	Yes
	dnSpike – 3		2.86	Yes		1.08	No
2	Aptamer 6 (pre-published)	ACE2	3.48	Yes	Spike (not selected)	2.47	No
	dnACE2-I		2.32	Yes	Spike (positive selection only)	0.424	No
	dnACE2-7		3.47***	Yes		16.8	Yes
	dnACE2-9		3.37	Yes		7.5	No
	dnACE2-21		7.74	Yes		14.3	No
	dnACE2-22		6.43	Yes		11.9	No
3	Aptamer 1 (pre-published)	HA	3.18	Yes	Spike (not selected)	6.67	No
	dnHA-I		5.49	Yes	Spike (positive selection only)	5.74	No
	dnHA-1		7.38	Yes		8.99	No
	dnHA-4		6.47	Yes		6.83	No
4	PD1-Initial	PD1	8.99	Yes	N/A	N/A	N/A
	PD1-AF-1-9		12.9	Yes			
	PD1-AF-3-1		9.61	Yes			
5	Aptamer 1C (pre-published)	Spike	0.553	Yes	ACE2	1.38	No
	Aptamer 4C (pre-published)		2.3	Yes		3.96	No
	dnSpike – 1		1.06***	Yes		13.9	Yes
	dnSpike-AUC-1		2.18***	Yes		11.2	Yes
	dnSpike-AUC-4		3.74	Yes		16.1	Yes
	dnSpike-AUC-6		3.12***	Yes		15.4	Yes
	dnSpike-AUC-22		2.84***	Yes		13.9	Yes

Table 6. Summary of Oligonucleotides form Validation Problems.
The oligonucleotides generated during the validation problems along with previously published aptamers for the target molecules are listed. The aptamers with significant differences in K_d-value for target versus counter-target are denoted. The significant marker (***) indicates $P < 0.001$.

7.0 REFERENCES

1. Khanna V, Ranganathan S. Physiochemical property space distribution among human metabolites, drugs and toxins. BMC Bioinformatics. 2009 Dec 3;10 Suppl 15(Suppl 15):S10. doi: 10.1186/1471-2105-10-S15-S10. PMID: 19958509; PMCID: PMC2788350.

2. Wouters OJ, McKee M, Luyten J. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. JAMA. 2020 Mar 3;323(9):844-853. doi: 10.1001/jama.2020.1166. Erratum in: JAMA. 2022 Sep 20;328(11):1110. Erratum in: JAMA. 2022 Sep 20;328(11):1111. PMID: 32125404; PMCID: PMC7054832.

3. Voskuil JL. The challenges with the validation of research antibodies. F1000Res. 2017 Feb 17;6:161. doi: 10.12688/f1000research.10851.1. PMID: 28357047; PMCID: PMC5333605.

4. Ibid.

5. https://www.roche.com/stories/antibodies-in-manufacturing-production/

6. Sumedha D Jayasena, Aptamers: An Emerging Class of Molecules That Rival Antibodies in Diagnostics, *Clinical Chemistry*, Volume 45, Issue 9, 1 September 1999, Pages 1628–1650, https://doi.org/10.1093/clinchem/45.9.1628

7. Kaiming Chen, Jie Zhou, Zhentao Shao, Jia Liu, Jia Song, Ruowen Wang, Juan Li, and Weihong Tan Journal of the American Chemical Society **2020** 142 (28), 12079-12086, DOI: 10.1021/jacs.9b13370

8. Lakhin AV, Tarantul VZ, Gening LV. Aptamers: problems, solutions and prospects. Acta Naturae. 2013 Oct;5(4):34-43. PMID: 24455181; PMCID: PMC3890987.

9. Xu Yixin, Jiang Xin, Zhou Yanhong, Ma Ming, Wang Minjin, Ying Binwu, Systematic Evolution of Ligands by Exponential Enrichment Technologies and Aptamer-Based Applications: Recent Progress and Challenges in Precision Medicine of Infectious Diseases, Frontiers in Bioengineering and Biotechnology, 9, 2021, 10.3389/fbioe.2021.704077

10. Ellington A.D., Szostak J.W. In vitro selection of RNA molecules that bind specific ligands. *Nature.* 1990;346:818–822. doi: 10.1038/346818a0.

11. Adachi T, Nakamura Y. Aptamers: A Review of Their Chemical Properties and Modifications for Therapeutic Application. *Molecules*. 2019;24(23):4229. Published 2019 Nov 21. doi:10.3390/molecules24234229

12. Tuerk C., Gold L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science.* 1990;249:505–510. doi: 10.1126/science.2200121.

13. Ellington A.D., Szostak J.W. In vitro selection of RNA molecules that bind specific ligands. *Nature.* 1990;346:818–822. doi: 10.1038/346818a0.

14. Alberts B, Johnson A, Lewis J, et al. Molecular Biology of the Cell. 4th edition. New York: Garland Science; 2002. Analyzing Protein Structure and Function. Available from: https://www.ncbi.nlm.nih.gov/books/NBK26820/

15. Chandola, C., Kalme, S., Casteleijn, M.G. *et al.* Application of aptamers in diagnostics, drug-delivery and imaging. *J Biosci* **41,** 535–561 (2016). https://doi.org/10.1007/s12038-016-9632-y

16. Song K-M, Lee S, Ban C. Aptamers and Their Biological Applications. *Sensors*. 2012; 12(1):612-631. https://doi.org/10.3390/s120100612

17. Ni, S., Zhuo, Z., Pan, Y., Yu, Y., Li, F., Liu, J., Wang, L., Wu, X., Li, D., Wan, Y., Zhang, L., Yang, Z., Zhang, B., Lu, A., and Zhang, G., Recent Progress in Aptamer Discoveries and Modifications for Therapeutic Applications, ACS Applied Materials & Interfaces, 13:8, 2020, https://doi.org/10.1021/acsami.0c05750

18. AD Ellington, JW Szostak, In vitro selection of RNA molecules that bind specific ligands, Nature, 346 (1990), pp. 818-822

19. L Gold, C Tuerk, Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase, Science, 249 (1990), pp. 505-510

20. Hays EM, Duan W, Shigdar S. Aptamers and Glioblastoma: Their Potential Use for Imaging and Therapeutic Applications. Int J Mol Sci. 2017;18(12):2576. Published 2017 Nov 30. doi:10.3390/ijms18122576

21. Liu, Q., Zhang, W., Chen, S. et al. SELEX tool: a novel and convenient gel-based diffusion method for monitoring of aptamer-target binding. J Biol Eng 14, 1 (2020). https://doi.org/10.1186/s13036-019-0223-y

22. Sefah, K., Shangguan, D., Xiong, X. et al. Development of DNA aptamers using Cell-SELEX. Nat Protoc 5, 1169–1185 (2010). https://doi.org/10.1038/nprot.2010.66

23. Miriam Jauset-Rubio, Mary Luz Botero, Vasso Skouridou, Gülsen Betül Aktas, Marketa Svobodova, Abdulaziz S. Bashammakh, Mohammad S. El-Shahawi, Abdulrahman O. Alyoubi, and Ciara K. O'Sullivan, ACS Omega 2019 4 (23), 20188-20196, DOI: 10.1021/acsomega.9b02412

24. Regina Stoltenburg, Nadia Nikolaus, Beate Strehlitz, "Capture-SELEX: Selection of DNA Aptamers for Aminoglycoside Antibiotics", Journal of Analytical Methods in Chemistry, vol. 2012, Article ID 415697, 14 pages, 2012. https://doi.org/10.1155/2012/415697

25. Subash C.B. Gopinath, Thangavel Lakshmipriya, M.K. Md Arshad, C.H. Voon, Tijjani Adam, Uda Hashim, Harbant Singh & Suresh V. Chinni (2017) Shortening full-length aptamer by crawling base deletion – Assisted by Mfold web server application, Journal of the Association of Arab Universities for Basic and Applied Sciences, 23:1, 37-42, DOI: 10.1016/j.jaubas.2016.07.001

26. J. Yang, M.T. Bowser, Capillary Electrophoresis-SELEX Selection of Catalytic DNA Aptamers for a Small-Molecule Porphyrin Target, Anal. Chem., 85 (3) (2013), pp. 1525-1530

27. M. Jing, M.T. Bowser, Isolation of DNA aptamers using micro free flow electrophoresis, Lab Chip, 11 (21), (2011), pp. 3703-3709.

28. N.P.G. Istamboulie, A. Triki, C. Lozano, L. Barthelmebs, T. Noguer, Selection of DNA aptamers against penicillin G using Capture-SELEX for the development of an impedimetric sensor, Talanta, 162 (2017), pp. 232-240

29. Bashir, A., Yang, Q., Wang, J. et al. Machine learning guided aptamer refinement and discovery. Nat Commun 12, 2366 (2021). https://doi.org/10.1038/s41467-021-22555-9

30. Emami, N., Ferdousi, R. AptaNet as a deep learning approach for aptamer–protein interaction prediction. Sci Rep 11, 6074 (2021). https://doi.org/10.1038/s41598-021-85629-0

31. Heredia FL, Roche-Lima A, Parés-Matos EI (2021) A novel artificial intelligence-based approach for identification of deoxynucleotide aptamers. PLoS Comput Biol 17(8): e1009247. https://doi.org/10.1371/journal.pcbi.1009247

32. Iwano, N., Adachi, T., Aoki, K. et al. Generative aptamer discovery using RaptGen. Nat Comput Sci 2, 378–386 (2022). https://doi.org/10.1038/s43588-022-00249-6

33. Thodima, Venkata, Mehdi Pirooznia, and Youping Deng. "RiboaptDB: a comprehensive database of ribozymes and aptamers." In *BMC bioinformatics*, vol. 7, no. 2, pp. 1-6. BioMed Central, 2006.

34. Torkamanian-Afshar, Mahsa, Hossein Lanjanian, Sajjad Nematzadeh, Maryam Tabarzad, Ali Najafi, Farzad Kiani, and Ali Masoudi-Nejad. "RPINBASE: an online toolbox to extract features for predicting RNA-protein interactions." *Genomics* 112, no. 3 (2020): 2623-2632.

35. "Apta-Index™ (Aptamer Database)" https://www.aptagen.com/apta-index/. Accessed 12 Apr. 2022.

36. Liu, Z.; Li, Y.; Han, L.; Li, J.; Liu, J.; Zhao, Z.; Nie, W.; Liu, Y.; Wang, R. PDB-wide Collection of Binding Data: Current Status of the PDBbind Database. Bioinformatics 2015, 31, 405– 412 DOI: 10.1093/bioinformatics/btu626

37. Emami, Neda, and Reza Ferdousi. "AptaNet as a deep learning approach for aptamer–protein interaction prediction." *Scientific Reports* 11, no. 1 (2021): 1-19.

38. Liu, Z.; Li, Y.; Han, L.; Li, J.; Liu, J.; Zhao, Z.; Nie, W.; Liu, Y.; Wang, R. PDB-wide Collection of Binding Data: Current Status of the PDBbind Database. Bioinformatics 2015, 31, 405– 412 DOI: 10.1093/bioinformatics/btu626

39. N. R. Markham & M. Zuker. UNAFold: Software for Nucleic Acid Folding and Hybridization. In Data, Sequence Analysis, and Evolution, J. Keith, ed., Bioinformatics: Volume 2, Chapter 1, pp 3-31, Humana Press Inc., 2008.

40. Antczak, M., Popenda, M., Zok, T., Sarzynska, J., Ratajczak, T., Tomczyk, K., Adamiak, R.W., Szachniuk, M. New functionality of RNAComposer: an application to shape the axis of miR160 precursor structure, Acta Biochimica Polonica, 2016, 63(4):737-744

41. Buglak AA, Samokhvalov AV, Zherdev AV, Dzantiev BB. Methods and Applications of In Silico Aptamer Design and Modeling. Int J Mol Sci. 2020 Nov 10;21(22):8420. doi: 10.3390/ijms21228420. PMID: 33182550; PMCID: PMC7698023.

42. Xiao J, Salsbury FR. Molecular dynamics simulations of aptamer-binding reveal generalized allostery in thrombin. J Biomol Struct Dyn. 2017 Nov;35(15):3354-3369. doi: 10.1080/07391102.2016.1254682. Epub 2016 Nov 29. PMID: 27794633; PMCID: PMC6876308.

43. Ya-chen Xie, Leif A Eriksson, Ru-bo Zhang, Molecular dynamics study of the recognition of ATP by nucleic acid aptamers, Nucleic Acids Research, Volume 48, Issue 12, 09 July 2020, Pages 6471–6480, https://doi.org/10.1093/nar/gkaa428

44. Song, J.; Zheng, Y.; Huang, M.; Wu, L.; Wang, W.; Zhu, Z.; Song, Y.; Yang, C. A Sequential Multidimensional Analysis Algorithm for Aptamer Identification Based on Structure Analysis and Machine Learning. Anal. Chem. 2020, 92, 3307–3314, doi:10.1021/acs.analchem.9b05203.

45. Ishida, R.; Adachi, T.; Yokota, A.; Yoshihara, H.; Aoki, K.; Nakamura, Y.; Hamada, M. RaptRanker: In Silico RNA Aptamer Selection from HT-SELEX Experiment Based on Local Sequence and Structure Information. Nucleic Acids Res. 2020, 48, doi:10.1093/nar/gkaa484.

46. Buglak AA, Samokhvalov AV, Zherdev AV, Dzantiev BB. Methods and Applications of In Silico Aptamer Design and Modeling. Int J Mol Sci. 2020 Nov 10;21(22):8420. doi: 10.3390/ijms21228420. PMID: 33182550; PMCID: PMC7698023.

47. Buglak AA, Samokhvalov AV, Zherdev AV, Dzantiev BB. Methods and Applications of In Silico Aptamer Design and Modeling. Int J Mol Sci. 2020 Nov 10;21(22):8420. doi: 10.3390/ijms21228420. PMID: 33182550; PMCID: PMC7698023.

48. Ibid.

49. Bell, D.R.; Weber, J.K.; Yin, W.; Huynh, T.; Duan, W.; Zhou, R. In silico design and validation of high-affinity RNA aptamers targeting epithelial cellular adhesion molecule dimers. Proc. Natl. Acad. Sci. USA 2020, 117, 8486–8493.

50. Song, Y.; Song, J.; Wei, X.; Huang, M.; Sun, M.; Zhu, L.; Lin, B.; Shen, H.; Zhu, Z.; Yang, C. Discovery of aptamers targeting the receptor-binding domain of the SARS-CoV-2 spike glycoprotein. Anal. Chem. 2020, 92, 9895–9900.

51. Sabri, M.Z.; Abdul Hamid, A.A.; Sayed Hitam, S.M.; Abdul Rahim, M.Z. In silico screening of aptamers configuration against hepatitis B surface antigen. Adv. Bioinform. 2019, 2019, 6912914

52. Soon, S.; Nordin, N.A. In silico predictions and optimization of aptamers against Streptococcus agalactiae surface protein using computational docking. Mater. Today Proc. 2019, 16, 2096–2100

53. Rockey, W.M.; Hernandez, F.J.; Huang, S.Y.; Cao, S.; Howell, C.A.; Thomas, G.S.; Liu, X.Y.; Lapteva, N.; Spencer, D.M.; McNamara, J.O.; et al. Rational truncation of an RNA aptamer to prostate-specific membrane antigen using computational structural modeling. Nucleic Acid Ther. 2011, 21, 299–314

54. Bavi, R.; Liu, Z.; Han, Z.; Zhang, H.; Gu, Y. In silico designed RNA aptamer against epithelial cell adhesion molecule for cancer cell imaging. Biochem. Biophys. Res. Commun. 2019, 509, 937–942.

55. Bell, D.R.; Weber, J.K.; Yin, W.; Huynh, T.; Duan, W.; Zhou, R. In silico design and validation of high-affinity RNA aptamers targeting epithelial cellular adhesion molecule dimers. Proc. Natl. Acad. Sci. USA 2020, 117, 8486–8493.

56. H. Yu, O. Alkhamis, J. Canoura, Y. Liu, Y. Xiao, Angew. Chem. Int. Ed. 2021, 60, 16800.

57. Lakhin AV, Tarantul VZ, Gening LV. Aptamers: problems, solutions and prospects. Acta Naturae. 2013 Oct;5(4):34-43. PMID: 24455181; PMCID: PMC3890987.

58. Ibid.

59. Famulok M, Mayer G. Aptamers and SELEX in chemistry & biology. Chem Biol. 2014;21:1055–8.

60. Shuaijian Ni, Zhenjian Zhuo, Yufei Pan, Yuanyuan Yu, Fangfei Li, Jin Liu, Luyao Wang, Xiaoqiu Wu, Dijie Li, Youyang Wan, Lihe Zhang, Zhenjun Yang, Bao-Ting Zhang, Aiping Lu, and Ge Zhang, ACS Applied Materials & Interfaces 2021 13 (8), 9500-9519, DOI: 10.1021/acsami.0c05750

61. Zhuo, Zhenjian, et al. "Recent advances in SELEX technology and aptamer applications in biomedicine." International journal of molecular sciences 18.10 (2017): 2142.

62. Wuchty, S., Fontana, W., Hofacker I., Schuster P., Biopolymers (1999), 9:145-165.

63. Mathews, D., Sabina, J., Zuker, M., Turner, D., J. Mol. Biol. (1999), 288:911-940.

64. Antczak, M., Popenda, M., Zok, T., Sarzynska, J., Ratajczak, T., Tomczyk, K., Adamiak, R.W., Szachniuk, M. New functionality of RNAComposer: an application to shape the axis of miR160 precursor structure, Acta Biochimica Polonica, 2016, 63(4):737-744

65. Hays EM, Duan W, Shigdar S. Aptamers and Glioblastoma: Their Potential Use for Imaging and Therapeutic Applications. Int J Mol Sci. 2017;18(12):2576. Published 2017 Nov 30. doi:10.3390/ijms18122576

66. Proteins: Structures and molecular properties (2nd edition). by Thomas E. Creighton, W. H. Freeman, New York, 1992, p142.

67. Ibid, p144.

68. Ibid, p145.

69. Ibid, p146.

70. Ibid, p146.

71. Ibid, p147.

72. Luscombe NM, Laskowski RA, Thornton JM. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. Nucleic Acids Res. 2001 Jul 1;29(13):2860-74. doi: 10.1093/nar/29.13.2860. PMID: 11433033; PMCID: PMC55782.

73. Ibid.

74. Sheu SY, Yang DY, Selzle HL, Schlag EW. Energetics of hydrogen bonds in peptides. Proc Natl Acad Sci U S A. 2003 Oct 28;100(22):12683-7. doi: 10.1073/pnas.2133366100. Epub 2003 Oct 14. PMID: 14559970; PMCID: PMC240678.

75. Elizabeth Barratt, Richard J. Bingham, Daniel J. Warner, Charles A. Laughton, Simon E. V. Phillips, and Steve W. Homans, Journal of the American Chemical Society 2005 127 (33), 11827-11834, DOI: 10.1021/ja0527525

76. Glaser, R., Chapter 2: Molecular Structure of Biological Systems, Biophysics, Springer Press, p26, 2012.

77. Ibid, p27.

78. Hollingsworth SA, Dror RO. Molecular Dynamics Simulation for All. Neuron. 2018;99(6):1129-1143. doi:10.1016/j.neuron.2018.08.011

79. Huizenga DE, Szostak JW. A DNA aptamer that binds adenosine and ATP. Biochemistry. 1995;34:656–65.

80. Zimmermann GR, Jenison RD, Wick CL, Simorre JP, Pardi A. Interlocking structural motifs mediate molecular discrimination by a theophylline-binding RNA. Nat Struct Biol. 1997;4:644–9.

81. Chen ZB, Chen L,Ma H, Zhou T, Li XX. Aptamer biosensor for label-free impedance spectroscopy detection of potassium ion based on DNA G-quadruplex conformation. Biosens Bioelectron. 2013;48:108–12.

82. Wilson DS, Szostak JW. In vitro selection of functional nucleic acids. Annu Rev Biochem. 1999;68:611–47.

83. Lin CH, Patel DJ. Structural basis of DNA folding and recognition in an AMP-DNA aptamer complex: distinct architectures but common recognition motifs for DNA and RNA aptamers complexed to AMP. Chem Biol. 1997;4:817–32.

84. Yang Y, Kochoyan M, Burgstaller P, Westhof E, Famulok M. Structural basis of ligand discrimination by two related RNA aptamers resolved by NMR spectroscopy. Science. 1996;272:1343–7.

85. Jenison RD, Gill SC, Pardi A, Polisky B. High-resolution molecular discrimination by RNA. Science. 1994;263:1425–9.

86. Alexandra Wrist, Wanqi Sun, and Ryan M. Summers, ACS Synthetic Biology 2020 9 (4), 682-697, DOI: 10.1021/acssynbio.9b00475

87. Kohlberger, M, Gadermaier, G. SELEX: Critical factors and optimization strategies for successful aptamer selection. *Biotechnology and Applied Biochemistry*. 2021; 1– 22. https://doi.org/10.1002/bab.2244

88. Kohlberger, M, Gadermaier, G. SELEX: Critical factors and optimization strategies for successful aptamer selection. *Biotechnology and Applied Biochemistry*. 2021; 1– 22. https://doi.org/10.1002/bab.2244

89. Huang D.-B., Vu D., Cassiday L.A., Zimmerman J.M., Maher L.J., Ghosh G. Crystal structure of NF-kappaB (p50)2 complexed to a high-affinity RNA aptamer. *Proc. Natl. Acad. Sci. USA.* 2003;100:9268–9273. doi: 10.1073/pnas.1632011100.

90. Ibid.

91. Long S.B., Long M.B., White R.R., Sullenger B.A. Crystal structure of an RNA aptamer bound to thrombin. *RNA.* 2008;14:2504–2512. doi: 10.1261/rna.1239308.

92. Adachi T, Nakamura Y. Aptamers: A Review of Their Chemical Properties and Modifications for Therapeutic Application. *Molecules*. 2019;24(23):4229. Published 2019 Nov 21. doi:10.3390/molecules24234229

93. Song, Y.; Song, J.; Wei, X.; Huang, M.; Sun, M.; Zhu, L.; Lin, B.; Shen, H.; Zhu, Z.; Yang, C. Discovery of aptamers targeting the receptor-binding domain of the SARS-CoV-2 spike glycoprotein. Anal. Chem. 2020,
92, 9895–9900.

94. Singh R, Mozzarelli A. Cofactor chemogenomics. Methods Mol Biol. 2009;575:93-122. doi: 10.1007/978-1-60761-274-2_4. PMID: 19727612.

95. Mulero MC, Shahabi S, Ko MS, Schiffer JM, Huang DB, Wang VY, Amaro RE, Huxford T, Ghosh G. Protein Cofactors Are Essential for High-Affinity DNA Binding by the Nuclear Factor κB RelA Subunit. Biochemistry. 2018 May 22;57(20):2943-2957. doi: 10.1021/acs.biochem.8b00158. Epub 2018 May 10. PMID: 29708732; PMCID: PMC5993198.

96. Ormerod, Michael, Flow Cytometry: A Basic Introduction, February 2019, page 1.

97. Yanling Song, Jia Song, Xinyu Wei, Mengjiao Huang, Miao Sun, Lin Zhu, Bingqian Lin, Haicong Shen, Zhi Zhu, and Chaoyong Yang, Analytical Chemistry 2020 92 (14), 9895-9900, DOI: 10.1021/acs.analchem.0c01394

98. Yan Y, Tao H, He J, Huang S-Y.* The HDOCK server for integrated protein-protein docking. Nature Protocols, 2020; doi: https://doi.org/10.1038/s41596-020-0312-x.

99. Plotly Technologies Inc. Title: Collaborative data science Publisher: Plotly Technologies Inc. Place of publication: Montréal, QC Date of publication: 2022 URL: https://plot.ly

100. Plotly Technologies Inc. Title: Collaborative data science Publisher: Plotly Technologies Inc. Place of publication: Montréal, QC Date of publication: 2022 URL: https://plot.ly

101. Ahirwar, R., Bariar, S., Balakrishnan, A., and Nahar, P., BSA Blocking in Enzyme-linked immunosorbent assays in non-mandatory step: a perspective study on mechanism of BSA Blocking in Common ELISA protocols, RSC Adv., 2015,5, 100077-100083

102. Zhou J, Soontornworajit B, Snipes MP, Wang Y. Structural prediction and binding analysis of hybridized aptamers. J Mol Recognit. 2011 Jan-Feb;24(1):119-26. doi: 10.1002/jmr.1034. PMID: 21194122.

103. Zhang P, Zhao N, Zeng Z, Feng Y, Tung CH, Chang CC, Zu Y. Using an RNA aptamer probe for flow cytometry detection of CD30-expressing lymphoma cells. Lab Invest. 2009 Dec;89(12):1423-32. doi: 10.1038/labinvest.2009.113. Epub 2009 Oct 12. PMID: 19823169; PMCID: PMC3763839.

104. Henri, Justin Liam; Bayat, Narges; Macdonald, joanna; Shigdar, Sarah (2019): A guide to using nucleic acid aptamers in cell based assays. Deakin University. Journal contribution. https://hdl.handle.net/10536/DRO/DU:30132731

105. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL, https://www.R-project.org/

106. Ritz C, Baty F, Streibig JC, Gerhard D (2015) Dose-Response Analysis Using R. PLoS ONE 10(12): e0146021. https://doi.org/10.1371/journal.pone.0146021

107. Seber, G. A. F. and Wild, C. J (1989) Nonlinear Regression, New York: Wiley & Sons (pp. 330–331).

108. Ritz, C (2009) Towards a unified approach to dose-response modeling in ecotoxicology, Environ Toxicol Chem.

109. Luscombe NM, Laskowski RA, Thornton JM. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res*. 2001;29(13):2860-2874. doi:10.1093/nar/29.13.2860

110. MATLAB. (2021). Version 9.10.0.1710957 (R2021a) Update 4. Natick, Massachusetts: The MathWorks Inc.

111. PDB-101: Educational resources supporting molecular explorations through biology and medicine. Christine Zardecki, Shuchismita Dutta, David S. Goodsell, Robert Lowe, Maria Voigt, Stephen K. Burley. (2022) Protein Science 31: 129-140, doi:10.1002/pro.4200

112. https://zhanggroup.org/SSIPe/pdb_atom_format.html

113. https://www.umass.edu/microbio/rasmol/pdb.htm

114. https://iupac.org/what-we-do/nomenclature/

115. Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, Philip E. Bourne, The Protein Data Bank, Nucleic Acids Research, Volume 28, Issue 1, 1 January 2000, Pages 235–242,

116. Arthur, David, and Sergi Vassilvitskii. "K-means++: The Advantages of Careful Seeding." SODA '07: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms. 2007, pp. 1027–1035.

117. Spath, H. Cluster Dissection and Analysis: Theory, FORTRAN Programs, Examples. Translated by J. Goldschmidt. New York: Halsted Press, 1985.

118. Luscombe NM, Laskowski RA, Thornton JM. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res*. 2001;29(13):2860-2874. doi:10.1093/nar/29.13.2860

119. Hoffman MM, Khrapov MA, Cox JC, Yao J, Tong L, Ellington AD. AANT: the Amino Acid-Nucleotide Interaction Database. *Nucleic Acids Res*. 2004;32(Database issue):D174-D181. doi:10.1093/nar/gkh128

120. Park, B., Kim, H. & Han, K. DBBP: database of binding pairs in protein-nucleic acid interactions. *BMC Bioinformatics* **15,** S5 (2014). https://doi.org/10.1186/1471-2105-15-S15-S5

121. Katoch, S., Chauhan, S.S. & Kumar, V. A review on genetic algorithm: past, present, and future. Multimed Tools Appl 80, 8091–8126 (2021). https://doi.org/10.1007/s11042-020-10139-6

122. Kumar V, Chhabra JK, Kumar D (2014) Parameter adaptive harmony search algorithm for unimodal and multimodal optimization problems. J Comput Sci 5(2):144–155

123. Michalewicz Z, Schoenauer M (1996) Evolutionary algorithms for constrained parameter optimization problems. Evol Comput 4(1):1–32

124. Katoch, S., Chauhan, S.S. & Kumar, V. A review on genetic algorithm: past, present, and future. Multimed Tools Appl 80, 8091–8126 (2021). https://doi.org/10.1007/s11042-020-10139-6

125. Michalewicz Z, Schoenauer M (1996) Evolutionary algorithms for constrained parameter optimization problems. Evol Comput 4(1):1–32

126. Michalewicz Z (1992) Genetic algorithms + data structures = evolution programs. Springer-Verlag, New York

127. Holland JH (1975) Adaptation in natural and artificial systems. The U. of Michigan Press

128. Mathworks. (2022). Global Optimization Toolbox: User's Guide (r2022a). Retrieved May 19, 2022 from https://www.mathworks.com/help/pdf_doc/gads/gads.pdf

129. Mathworks. (2022). Global Optimization Toolbox: User's Guide (r2022a). Retrieved May 19, 2022 from https://www.mathworks.com/help/pdf_doc/gads/gads.pdf

130. Mathworks. (2022). Global Optimization Toolbox: User's Guide (r2022a). Retrieved May 19, 2022 from https://www.mathworks.com/help/pdf_doc/gads/gads.pdf

131. Thompson, I.A.P., Zheng, L., Eisenstein, M. et al. Rational design of aptamer switches with programmable pH response. Nat Commun 11, 2946 (2020).

132. Hasegawa H, Savory N, Abe K, Ikebukuro K. Methods for Improving Aptamer Binding Affinity. Molecules. 2016;21(4):421. Published 2016 Mar 28. doi:10.3390/molecules21040421

133. Afanasyeva A, Nagao C, Mizuguchi K. Prediction of the secondary structure of short DNA aptamers. *Biophys Physicobiol*. 2019;16:287-294. Published 2019 Nov 29. doi:10.2142/biophysico.16.0_287

134. Mathews DH, Sabina J, Zuker M, Turner DH. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol. 1999 May 21;288(5):911-40. doi: 10.1006/jmbi.1999.2700. PMID: 10329189.

135. Mathews DH, Sabina J, Zuker M, Turner DH. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol. 1999 May 21;288(5):911-40. doi: 10.1006/jmbi.1999.2700. PMID: 10329189.

136. Wuchty S, Fontana W, Hofacker IL, Schuster P. Complete suboptimal folding of RNA and the stability of secondary structures. Biopolymers. 1999 Feb;49(2):145-65. doi: 10.1002/(SICI)1097-0282(199902)49:2<145::AID-BIP4>3.0.CO;2-G. PMID: 10070264.

137. Lorenz, R., Bernhart, S.H., Höner zu Siederdissen, C. et al. ViennaRNA Package 2.0. Algorithms Mol Biol 6, 26 (2011). https://doi.org/10.1186/1748-7188-6-26

138. Mathworks. (2022). Rnafold (r2021a). Retrieved August 04, 2022.

139. Mathworks. (2022). Rnaplot (r2021a). Retrieved August 04, 2022.

140. Antczak, M., Popenda, M., Zok, T., Sarzynska, J., Ratajczak, T., Tomczyk, K., Adamiak, R.W., Szachniuk, M. New functionality of RNAComposer: an application to shape the axis of miR160 precursor structure, Acta Biochimica Polonica, 2016, 63(4):737-744 (doi:10.18388/abp.2016_1329).

141. Popenda, M., Szachniuk, M., Antczak, M., Purzycka, K.J., Lukasiak, P., Bartol, N., Blazewicz, J., Adamiak, R.W. Automated 3D structure composition for large RNAs, Nucleic Acids Research, 2012, 40(14):e112 (doi:10.1093/nar/gks339).

142. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE. A geometric approach to macromolecule-ligand interactions. J Mol Biol. 1982 Oct 25;161(2):269-88. doi: 10.1016/0022-2836(82)90153-x. PMID: 7154081.

143. Jones G., Willett P., Glen R.C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. J. Mol. Biol. 1995; 245:43–53. doi: 10.1016/S0022-2836(95)80037-9.

144. Jain A.N. Surflex: Fully Automatic Flexible Molecular Docking Using a Molecular Similarity-Based Search Engine. J. Med. Chem. 2003; 46:499–511. doi: 10.1021/jm020406h.

145. Trott O., Olson A.J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J. Comput. Chem. 2010; 31:455–461. doi: 10.1002/jcc.21334.

146. Friesner R.A., Banks J.L., Murphy R.B., Halgren T.A., Klicic J.J., Mainz D.T., Repasky M.P., Knoll E.H., Shaw D.E., Shelley M., et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. J. Med. Chem. 2004; 47:1739–1749. doi: 10.1021/jm0306430.

147. Pantsar T, Poso A. Binding Affinity via Docking: Fact and Fiction. Molecules. 2018 Jul 30;23(8):1899. doi: 10.3390/molecules23081899. PMID: 30061498; PMCID: PMC6222344.

148. Brian Jiménez-García, Jorge Roel-Touris, Miguel Romero-Durana, Miquel Vidal, Daniel Jiménez-González, Juan Fernández-Recio, LightDock: a new multi-scale approach to protein–protein docking, Bioinformatics, Volume 34, Issue 1, 01 January 2018, Pages 49–55, https://doi.org/10.1093/bioinformatics/btx555

149. Krishnanand, K.N., Ghose, D. Glowworm swarm optimization for simultaneous capture of multiple local optima of multimodal functions. Swarm Intell 3, 87–124 (2009). https://doi.org/10.1007/s11721-008-0021-5

150. https://www.ebi.ac.uk/msd-srv/capri/round47/participants.html

151. Yan Y, Tao H, He J, Huang S-Y.* The HDOCK server for integrated protein-protein docking. Nature Protocols, 2020; doi: https://doi.org/10.1038/s41596-020-0312-x.

152. Yan Y, Zhang D, Zhou P, Li B, Huang S-Y. HDOCK: a web server for protein-protein and protein-DNA/RNA docking based on a hybrid strategy. Nucleic Acids Res. 2017;45(W1):W365-W373.

153. Yan Y, Zhang D, Zhou P, Li B, Huang S-Y. HDOCK: a web server for protein-protein and protein-DNA/RNA docking based on a hybrid strategy. Nucleic Acids Res. 2017;45(W1):W365-W373.

154. Yan Y, Tao H, He J, Huang S-Y.* The HDOCK server for integrated protein-protein docking. Nature Protocols, 2020; doi: https://doi.org/10.1038/s41596-020-0312-x.

155. Yan Y, Zhang D, Zhou P, Li B, Huang S-Y. HDOCK: a web server for protein-protein and protein-DNA/RNA docking based on a hybrid strategy. Nucleic Acids Res. 2017;45(W1): W365-W373.

156. Yan Y, Wen Z, Wang X, Huang S-Y. Addressing recent docking challenges: A hybrid strategy to integrate template-based and free protein-protein docking. Proteins 2017; 85:497-512.

157. Huang S-Y, Zou X. A knowledge-based scoring function for protein-RNA interactions derived from a statistical mechanics-based iterative method. Nucleic Acids Res. 2014;42:e55.

158. Huang S-Y, Zou X. An iterative knowledge-based scoring function for protein-protein recognition. Proteins 2008; 72:557-579.

159. https://www.mathworks.com/help/matlab/ref/histcounts.html

160. Ali, A., Vijayan, R. Dynamics of the ACE2–SARS-CoV-2/SARS-CoV spike protein interface reveal unique mechanisms. Sci Rep 10, 14214 (2020). https://doi.org/10.1038/s41598-020-71188-3

161. Song Y, Song J, Wei X, Huang M, Sun M, Zhu L, Lin B, Shen H, Zhu Z, Yang C. Discovery of Aptamers Targeting the Receptor-Binding Domain of the SARS-CoV-2 Spike Glycoprotein. Anal Chem. 2020 Jul 21;92(14):9895-9900. doi: 10.1021/acs.analchem.0c01394. Epub 2020 Jul 2. PMID: 32551560; PMCID: PMC7336720.

162. Song Y, Song J, Wei X, Huang M, Sun M, Zhu L, Lin B, Shen H, Zhu Z, Yang C. Discovery of Aptamers Targeting the Receptor-Binding Domain of the SARS-CoV-2 Spike Glycoprotein. Anal Chem. 2020 Jul 21;92(14):9895-9900. doi: 10.1021/acs.analchem.0c01394. Epub 2020 Jul 2. PMID: 32551560; PMCID: PMC7336720.

163. Song Y, Song J, Wei X, Huang M, Sun M, Zhu L, Lin B, Shen H, Zhu Z, Yang C. Discovery of Aptamers Targeting the Receptor-Binding Domain of the SARS-CoV-2 Spike Glycoprotein. Anal Chem. 2020 Jul 21;92(14):9895-9900. doi: 10.1021/acs.analchem.0c01394. Epub 2020 Jul 2. PMID: 32551560; PMCID: PMC7336720.

164. CDC, “What is the Difference between Influenza (flu) and COVID-19?”, https://www.cdc.gov/flu/symptoms/flu-vs-covid19.htm

165. Villa A, Brunialti E, Dellavedova J, Meda C, Rebecchi M, Conti M, Donnici L, De Francesco R, Reggiani A, Lionetti V, Ciana P. DNA aptamers masking angiotensin converting enzyme 2 as an innovative way to treat SARS-CoV-2 pandemic. Pharmacol Res. 2022 Jan;175:105982. doi: 10.1016/j.phrs.2021.105982. Epub 2021 Nov 16. Erratum in: Pharmacol Res. 2021 Dec 22;:106042. PMID: 34798263; PMCID: PMC8594078.

166. Ibid.

167. Ibid.

168. Shang, J., Ye, G., Shi, K. et al. Structural basis of receptor recognition by SARS-CoV-2. Nature 581, 221–224 (2020). https://doi.org/10.1038/s41586-020-2179-y

169. Villa A, Brunialti E, Dellavedova J, Meda C, Rebecchi M, Conti M, Donnici L, De Francesco R, Reggiani A, Lionetti V, Ciana P. DNA aptamers masking angiotensin converting enzyme 2 as an innovative way to treat SARS-CoV-2 pandemic. Pharmacol Res. 2022 Jan;175:105982. doi: 10.1016/j.phrs.2021.105982. Epub 2021 Nov 16. Erratum in: Pharmacol Res. 2021 Dec 22;:106042. PMID: 34798263; PMCID: PMC8594078.

170. CDC, “What is the Difference between Influenza (flu) and COVID-19?”, https://www.cdc.gov/flu/symptoms/flu-vs-covid19.htm

171. Bhardwaj J, Chaudhary N, Kim H, Jang J. Subtyping of influenza A H1N1 virus using a label-free electrochemical biosensor based on the DNA aptamer targeting the stem region of HA protein. Anal Chim Acta. 2019 Aug 8;1064:94-103. doi: 10.1016/j.aca.2019.03.005. Epub 2019 Mar 9. PMID: 30982523.

172. Wenkai Li, Xinru Feng, Xing Yan, Keyi Liu, and Le Deng. A DNA Aptamer Against Influenza A Virus: An Effective Inhibitor to the Hemagglutinin–Glycan Interactions. Nucleic Acid Therapeutics.Jun 2016.166-172.

173. Wenkai Li, Xinru Feng, Xing Yan, Keyi Liu, and Le Deng. A DNA Aptamer Against Influenza A Virus: An Effective Inhibitor to the Hemagglutinin–Glycan Interactions. Nucleic Acid Therapeutics.Jun 2016.166-172.

174. Tian Gao and Renjun Pei, Isolation of DNA Aptamer Targeting PD-1 with an Antitumor Immunotherapy Effect, ACS Applied Bio Materials 2020 3 (10), 7080-7086, DOI: 10.1021/acsabm.0c00919

175. Tian Gao and Renjun Pei, Isolation of DNA Aptamer Targeting PD-1 with an Antitumor Immunotherapy Effect, ACS Applied Bio Materials 2020 3 (10), 7080-7086, DOI: 10.1021/acsabm.0c00919

176. Ibid.

177. Jumper, J et al. Highly accurate protein structure prediction with AlphaFold. Nature (2021)

178. Varadi, M et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Research (2022)

179. Song Y, Song J, Wei X, Huang M, Sun M, Zhu L, Lin B, Shen H, Zhu Z, Yang C. Discovery of Aptamers Targeting the Receptor-Binding Domain of the SARS-CoV-2 Spike Glycoprotein. Anal Chem. 2020 Jul 21;92(14):9895-9900. doi: 10.1021/acs.analchem.0c01394. Epub 2020 Jul 2. PMID: 32551560; PMCID: PMC7336720.

180. Song Y, Song J, Wei X, Huang M, Sun M, Zhu L, Lin B, Shen H, Zhu Z, Yang C. Discovery of Aptamers Targeting the Receptor-Binding Domain of the SARS-CoV-2 Spike Glycoprotein. Anal Chem. 2020 Jul 21;92(14):9895-9900. doi: 10.1021/acs.analchem.0c01394. Epub 2020 Jul 2. PMID: 32551560; PMCID: PMC7336720.

181. P. Towler, B. Staker, S.G. Prasad, S. Menon, J. Tang, T. Parsons, D. Ryan, M. Fisher, D. Williams, N.A. Dales, M.A. Patane, M.W. Pantoliano, ACE2 X-ray structures reveal a large hinge-bending motion important for inhibitor binding and catalysis, J. Biol. Chem., 279 (17) (2004), pp. 17996-18007

182. Daniel, W., & Wang, N. (2020). Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science*. 10.1126/science.abb2507

183. Daniel, W., & Wang, N. (2020). Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science. 10.1126/science.abb2507

184. R. Xu, D. C. Ekiert, J. C. Krause, R. Hai, J. E. Crowe Jr.., I. A. Wilson, Structural basis of preexisting immunity to the 2009 H1N1 pandemic influenza virus. Science 328, 357–360 (2010).

185. P. Towler, B. Staker, S.G. Prasad, S. Menon, J. Tang, T. Parsons, D. Ryan, M. Fisher, D. Williams, N.A. Dales, M.A. Patane, M.W. Pantoliano, ACE2 X-ray structures reveal a large hinge-bending motion important for inhibitor binding and catalysis, J. Biol. Chem., 279 (17) (2004), pp. 17996-18007

186. R. Xu, D. C. Ekiert, J. C. Krause, R. Hai, J. E. Crowe Jr.., I. A. Wilson, Structural basis of preexisting immunity to the 2009 H1N1 pandemic influenza virus. Science 328, 357–360 (2010).

187. Tzarum N, de Vries RP, Zhu X, Yu W, McBride R, Paulson JC, Wilson IA. Structure and receptor binding of the hemagglutinin from a human H6N1 influenza virus. Cell Host Microbe. 2015 Mar 11;17(3):369-376. doi: 10.1016/j.chom.2015.02.005. PMID: 25766295; PMCID: PMC4374348.

188. Towler P, Staker B, Prasad SG, Menon S, Tang J, Parsons T, Ryan D, Fisher M, Williams D, Dales NA, Patane MA, Pantoliano MW. ACE2 X-ray structures reveal a large hinge-bending motion important for inhibitor binding and catalysis. J Biol Chem.

2004 Apr 23;279(17):17996-8007. doi: 10.1074/jbc.M311191200. Epub 2004 Jan 30. PMID: 14754895; PMCID: PMC7980034.

189. Villa A, Brunialti E, Dellavedova J, Meda C, Rebecchi M, Conti M, Donnici L, De Francesco R, Reggiani A, Lionetti V, Ciana P. DNA aptamers masking angiotensin converting enzyme 2 as an innovative way to treat SARS-CoV-2 pandemic. Pharmacol Res. 2022 Jan;175:105982. doi: 10.1016/j.phrs.2021.105982. Epub 2021 Nov 16. Erratum in: Pharmacol Res. 2021 Dec 22;:106042. PMID: 34798263; PMCID: PMC8594078.