Protein Variational Autoencoders for Antibody-Based Drug Design

Technical Report:
Year 1, FY 2020

Scott A. Walper (PI), Research Biologist
Scott N. Dean (Co-PI), Research Biologist
George P. Anderson, Research Biologist
Jerome E. Alvarez, SSEP

March 15th, 2021

Prepared For
Dr. Sweta Batni
Science and Technology Manager
Research and Development Directorate
Chemical and Biological Technologies Department
Defense Threat Reduction Agency
8725 John J. Kingman Road
Ft. Belvoir, VA 22060

Organization
Scott A. Walper
Research Biologist
U.S. Naval Research Laboratory
Center for Bio/Molecular Science and Engineering
4555 Overlook Ave SW
Washington, DC 20375
202.404.6070
Scott.walper@nrl.navy.mil

Table of Contents

1. Executive Summary

This report details the *Year 1* efforts of the Naval Research Laboratory (NRL) toward the development of Artificial Intelligence/ Machine Learning (AI/ML) algorithms to enable directed modification of antibody and antibody-like proteins to enhance their pharmokinetic properties. Over the course of this program, NRL researchers will establish and build capabilities in AI/ ML technologies identifying best practices in algorithm design and application, dataset generation and curation, and methods for validation of biomolecule properties. It is anticipated that this foundational research will benefit current and future Department of Defense (DoD) efforts in biomolecule development and utilization for a wide array of potential applications including the development of next generation therapeutics, diagnostics, and sensors.

1.1. Overview

The harnessing of biological systems and processes has driven the development of human civilization. Recognized at the historical level as the development of reliable agricultural systems, mankind has evolved its capabilities in biological utilization. Today commercial entities can produce biomolecules such as enzymes, antibodies, and others at incredible scale for commercial and medical applications. In order to achieve these incredible successes, however, biomolecules must be selected and/or engineered to exhibit biophysical properties that make them amenable to both production and application. While some biomolecules exhibit a hardiness that allows them to function under extremes of heat in the presence of organic compounds, most bio-derived materials are less stable when removed from the confines of the cell or host system and quickly lose biological function or activity. This significantly complicates the identification and production of suitable biomolecules for commercial and medical applications which often relies on large-scale sequence mining via bioinformatics, screening of mutant libraries, and cumbersome downstream testing protocols. With growing capabilities in Artificial Intelligence and Machine Learning (AI/ML) as well as available datasets of biomolecules for algorithm training, the potential to circumvent many of the aforementioned complications in biomolecule evolution has emerged as a realistic possibility.

Machine learning is considered a sub set of AI and is typically comprised of fewer algorithms that are typically “trained” rather than programmed. Training occurs through an iterative or looped process that proceeds until a defined criteria is met or a specified number of cycles has elapsed. By design, ML systems can process and require large data systems on which to be trained and generate assumptions. Deep learning (DL) expands upon the foundations of ML and relies on a series of stacked algorithmic layers which closely resemble the architecture and functionality of the brain. The stacked layers of a Deep learning system are artificial neural networks that rely on many layers of nonlinear processing units for learning data representations. This architecture has proven useful for automation in chemical design following algorithm training on the large datasets comprised of hundreds of thousands of chemical structures. While still in the early stages of testing and validation, it is theorized that DL systems will have significant impact on the design and evolution of biomolecules.

Antibody and recombinant antibody-like molecules are routinely used as therapeutics and in diagnostic assays and sensors. Antibodies are often characterized by their antigen specificity, however, not all antibodies are created equal. Within a polyclonal assortment antibodies can show dramatic variation in binding specificity and affinity, protein/thermal stability, aggregation potential, and numerous other biophysical properties. Additionally, while polyclonal antibodies have utility in a number of applications, monoclonal antibodies are necessary to reduce lot-to-lot variability and are required for many therapeutic applications. Therefore, researchers and developers must balance benefits and flaws when attempting to select for a monoclonal antibody for a given application. Further engineering through sequence mutation and evolution is often undertaken to improve the antibody leading to significant investments of time and money. Advancements in AI/ML techniques combined with large data collections of antibody sequences offers a path toward antibody by design and are the goals of this program

1.2. Updates

In the first year of this effort, focused efforts on developing AI/ML capabilities for the selection of recombinant antibody-like molecules with enhanced thermal stability. While numerous recombinant antibody-like structures have been described in the literature, the small (100-110 amino acids, 300-330 base pairs) minimal structure of the *Camelid*-derived, single domain antibody (sdAb) was selected as a starting platform (**Figure 1**). Single domain antibodies were chosen as a starting platform for several reasons 1) readily accessible expertise and datasets within NRL, 2) small set of fully characterized sdAb sequences to assist in algorithm training and controls for comparison, 3) ease of recombinant production and purification. In combination, these advantages allowed for rapid *Go / No Go* analysis of AI/ML algorithms and the associated sequence output as sdAb proteins could be evaluated *in silico* and biochemically to verify properties such as binding affinity, thermal stability, and aggregation potential. Unlike conventional antibodies that must be produced in mammalian cell culture, the sdAb chassis is more conducive to higher-throughput methods of evaluation.

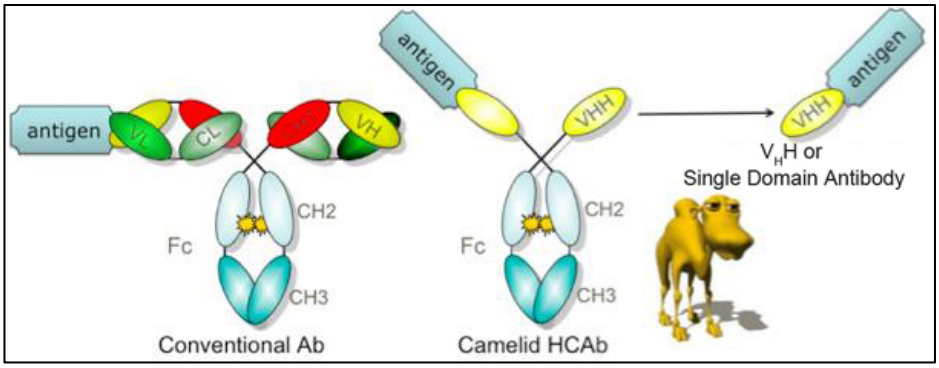


Figure 1. Structure of a single domain antibody.
Conventional antibodies are tetrameric proteins comprised of identical heavy and light chain subunits that partner to form an antigen binding domain consisting of amino acids from each subunit. Camelids and sharks possess unique antibody structures that are dimers of a single, identical heavy chain subunits. As with conventional antibodies the antigen binding domain is formed from a short terminal amino acid sequence or motif. This protein motif can be isolated and produced recombinantly maintaining its antigen binding capability. Image reproduced from https://www.creative-biolabs.com/blog/index.php/single-domain-antibody-made-it-possible-for-curing-diseases-with-antibody-drugs/

1.2.1 Dataset Assembly

The initial sdAb database was assembled from several online repositories containing the amino acid sequence of *Camelid*-derived sdAb. Sequences from the Single Domain Antibody Database (Wilton, Opyr et al. 2018), Institute Collection and Analysis of Nanobodies (Zuo, Li et al. 2017), and The Structural Antibody Database (Dunbar, Krawczyk et al. 2014) were combined as to form the initial dataset for AI/ML training. While each of these databases are associated with some level of metadata, none contain melting temperature (T_m) or other physiochemical information. This combined dataset was later appended to a dataset provided by Dr. George Anderson and Dr. Ellen Goldman of NRL which included several hundred sdAb sequences, many with extensive biophysical characterization including binding affinity, thermal stability/ melting temperature (T_m), and aggregation potential. In combination, this starting dataset consisted of ~3000 unique sequences. To further expand the number of sdAb sequences available for training ML and DL models, we leveraged a raw sdAb sequence library that had been obtained through a previous NRL program (Anderson, Goldman). This genetic sequence resource contained ~15 million antibody-like gene sequences. Refining this database to remove non-productive and redundant sequences yielded a useable dataset of nearly ~6 million unique sdAb sequences (**Figure 2**).

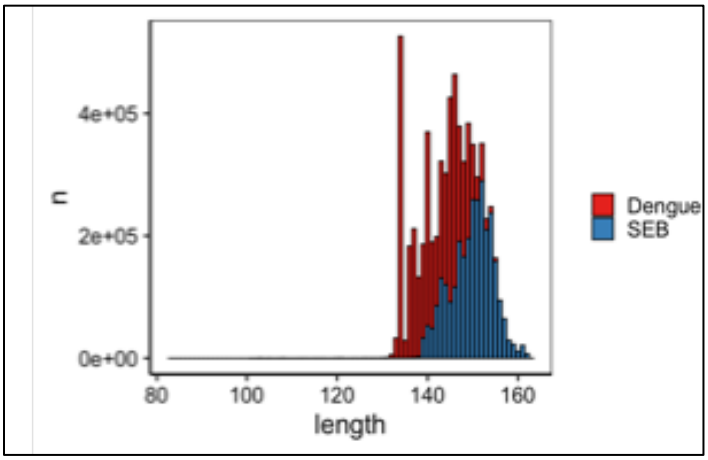


Figure 2. Length distributions of sequences in the NRL sdAb database.
Following data cleaning and removal of redundant sequences, approximately 6 million sequences remained from the two Dengue and Staphylococcal Enterotoxin B (SEB) datasets. Length distributions were plotted. Median dengue sequences were found to be slightly shorter than SEB: 146 vs. 152 amino acids. Nearly all sequences were between 130 and 165 amino acids in length.

To assist in algorithm training and analysis of generated sequences, sdAbs known binding specificities, affinities, and verified T_m were chosen as control sdAbs. Here we targeted sdAbs for Staphylococcal Enterotoxin B (SEB) reference sequences from extant literature and internal NRL data, totaling ~150 unique sequences. In addition to these SEB-binding sdAbs, additional sdAb sequences with experimental T_ms were obtained from Patrick Kunz (Coriolis Pharma, Munich).

1.2.2 AI/ ML algorithms

The number of tools or systems to enable AI/ML technologies is constantly rising as shown in **Table 1** allowing users to partner tool selection with experimental requirements. The applicability of these systems to biomolecule selection and design have not been adequately validated. Over the course of this program we intend to evaluate several of these systems validating AI/ML predictions with experimental testing. Year one efforts focused on the use of a *Recursive Neural Network* (RNN) for dataset evaluation, however, additional methods including a *Variational Autoencoder* (VAE) and the *Local Interpretable Model-Agnostic Explanations* (LIME) system were also examined.

Category	Model	Reference
VAE	LSTMVAE	(Bowman et al., 2016)
	CNNVAE	(Yang et al., 2017)
	HybridVAE	(Semeniuta et al., 2017)
GAN	SeqGAN	(Yu et al., 2017)
	TextGAN	(Zhang et al., 2017)
	RankGAN	(Lin et al., 2017)
	MaliGAN	(Che et al., 2017)
	LeakGAN	(Guo et al., 2018)
	MaskGAN	(Fedus et al., 2018)
Seq2Seq	RNN	(Sutskever et al., 2014)
	Transformer	(Vaswani et al., 2017b)
	GPT-2	(Radford et al.)
	XLNet	(Yang et al., 2019)
	BERT2BERT	(Rothe et al., 2020)
	BART	(Lewis et al., 2020)

Table 1. Common AI / ML algorithms

1.2.2.1 *Recursive Neural Networks*

A generative RNN, a type of generative DLtechnique, can make use of a variety of sequence or text-like data. RNNs are very effective text generators and have previously been shown to produce functional proteins when trained on a dataset of known sequences (Karpathy 2015, Müller, Hiss et al. 2018). The standard approach developed for using RNNs is to train on a large corpus of known sequences, such that the model can learn a generalizable vocabulary and grammar of the subject – in this case antibodies – then refine or tune the output by subsequently feeding the model a narrow type of sequence.

The RNN was initially trained on the dataset of ~3000 sdAb amino acids sequences, followed by tuning the training to generate SEB-binders by feeding the model the ~150 SEB-binding sdAbs. The RNN embeddings developed by the model (dimensions = 128) were reduced to two dimensions for visualization (**Figure 3**). In lieu of T_m values for all of the sdAbs trained on, our initial plan was to utilize the aliphatic index, calculatable solely from sequence information, and thus immediately available for all sequences. Known to correlate to thermostability of proteins (with a R^2 of ~0.4), we overlaid aliphatic index quartiles as part of the selection criteria.

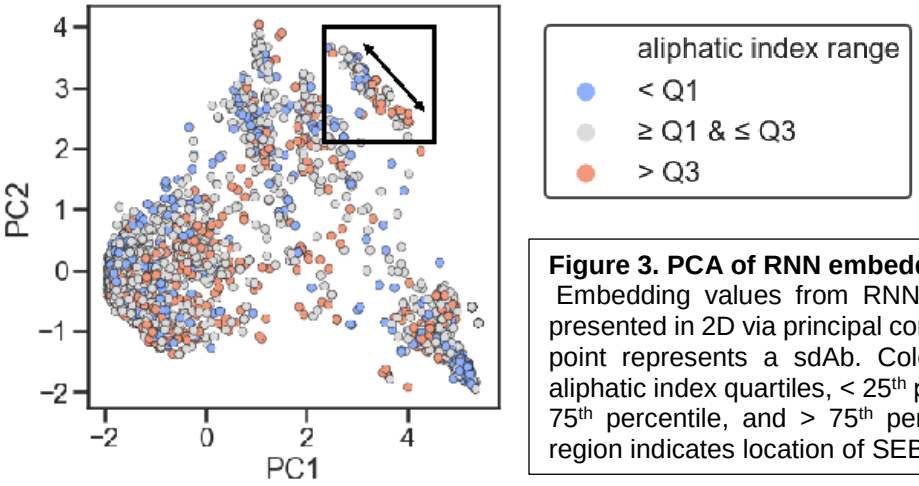


Figure 3. PCA of RNN embeddings.
Embedding values from RNN trained on sdAbs are presented in 2D via principal component analysis. Each point represents a sdAb. Coloring applied indicates aliphatic index quartiles, < 25th percentile (Q1), ≥ 25th to 75th percentile, and > 75th percentile (Q3). Enclosed region indicates location of SEB-binding sequences.

From the output amino acid sequences, twelve putative SEB-binding sdAb genes were selected and ordered for synthesis from ATUM Biosciences (**Supplemental Table 1**). Four sequences were selected from each of the aliphatic index groups, low (or < Q1), medium (≥ Q1 and ≤ Q3), and high (> Q3), hypothesized to provide us with a range of T_ms and allow for improve prediction and selection of sdAb amino acid sequences from the RNN output.

Since the aliphatic index was thought to be an inadequate stand-in for experimentally determined T_m values, we next created an in-house T_m prediction model which is described in detail below (sections **1.2.3** and **1.2.5.1**).

1.2.2.2 Local Interpretable Model-Agnostic Explanations

While the RNN generative model allows for use of very large sdAb datasets, the vast majority of these lack experimentally defined biophysical properties which complicates initial predictions. To counter this issue, we developed a low-sample size method for generating new antibody sequences. The novel optimization algorithm we developed makes use of the recently introduced explainable AI technique Local Interpretable Model-Agnostic Explanations (LIME) for directing mutations at sites selected by the associated machine or deep learning model (Ribeiro, Singh et al. 2016). The method takes an initial input sequence, then using LIME paired with a chosen regression model – here XGBoost regressor – the LIME explainer selects locations within the sequence that most positively or negatively impact the regression model prediction. In this case, increased T_m values was the chosen goal, so mutations that negatively impacted the regression output were selected. After finding the best mutation location, each amino acid is assessed via the same regression model, and after identifying the best replacement amino acid, the process repeats N times (see schematic in **Figure 4**). This process is generalizable and could theoretically be applied to any dataset of polymers that are paired with a numeric property descriptor, such as T_m, binding affinity, or serum half-life, while not requiring thousands or millions of examples for functionality. In the case of sdAbs, this LIME-based generative model allows for production of new sdAb sequences using a greatly reduced dataset of ~150 sdAb, each of which had defined T_m values arrived at experimentally.

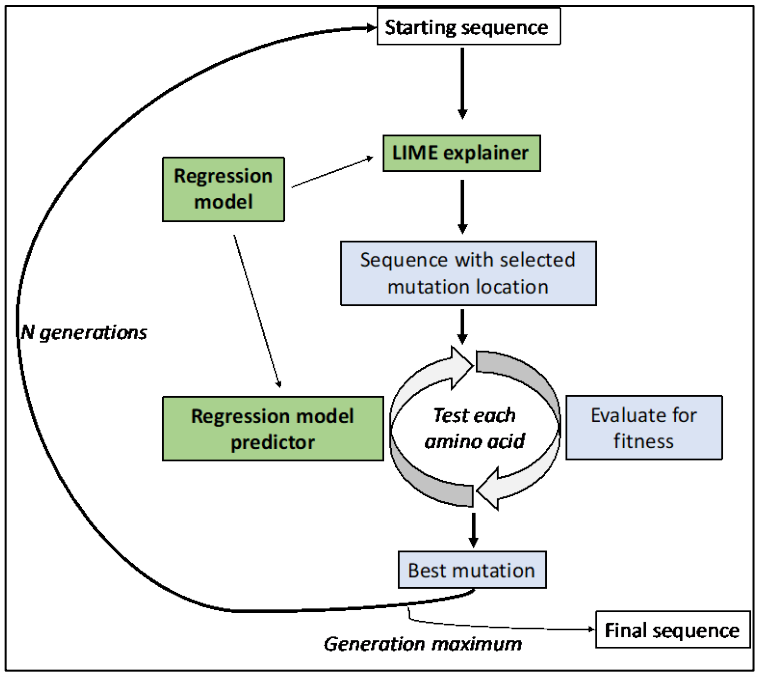


Figure 4. Schematic of LIME-based algorithm for protein sequence optimization.
With an input sequence, using LIME paired with a chosen regression model (green boxes), the LIME explainer can select locations within the sequence that most positively or negatively impact the regression output. After selecting the best mutation location, each amino acid is tested via the same regression model. After identifying the best mutation, the process repeats N times.

Using the described LIME-based method, the previously identified XGB model with optimized parameters was used (**Figure 5A**). This model was then subjected to the LIME explainer method of **Figure 4** and iterated over the input sdAb sequence, identified the most problematic location, mutated it to a different amino acid, and then repeated the process N times. In the case of the ordered sequences N was set to 30. Applying this process to low Tm sdAbs, each predicted Tm increases (**Figure 5B**). The sdAb highlighted in red is synthetic construct referred to as Ca immunoglobulin heavy chain variable region gene (GenBank: KU508542.1) from Turner *et al.* (Turner, Naciri et al. 2016).

LIME-generated sdAb amino acid sequences were translated to DNA sequences which in turn were synthesized by ATUM. Parental sdAb sequences and mutated progeny are reported in **Table 2** (also **Supplemental Table 2**). Other parental sequences selected were synthetic construct Cd immunoglobulin heavy chain variable region gene (GenBank: KU508545.1), clone S222-A2 camelid anti-SEB recombinant antibody, and clone H9-B11 camelid anti-SEB recombinant antibody. Both Ca and Cd were obtained from NRL-lead publications (Turner, Naciri et al. 2016) as were S222-A2 and H9-B11 (Turner, Zabetakis et al. 2014) along with their thermal stability and binding affinity characteristics. Each Ca, Cd, S222-A2, and H9-B11 was selected due to their low starting T_ms: between 45 and 55 °C.

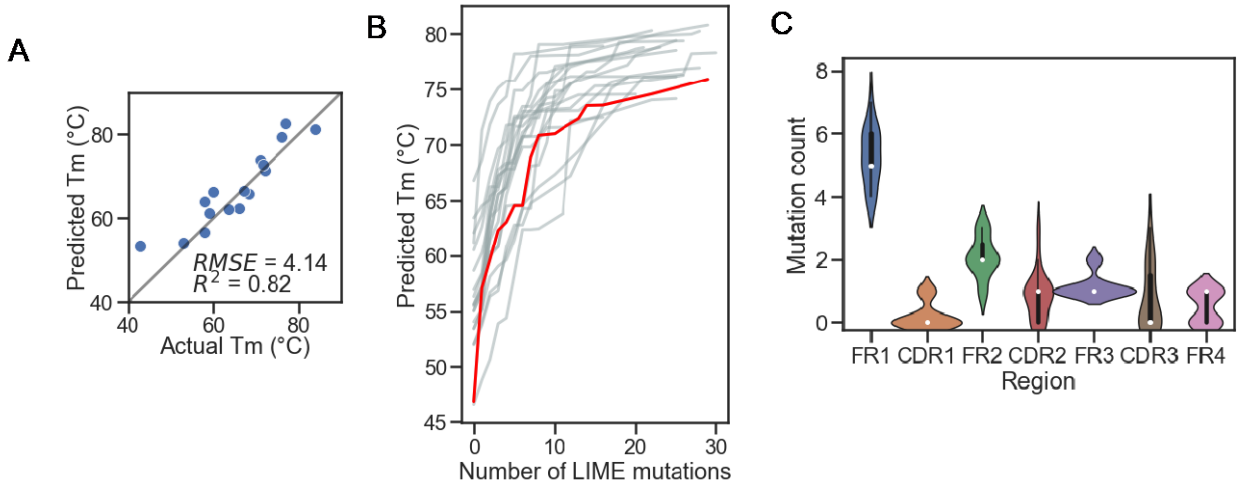


Figure 5. Generation of new sdAb sequences using LIME-based optimization algorithm. A) The XGBoost was selected previously as the best Tm-prediction algorithm. R^2 was found to be ~0.8 on the validation set. B) LIME (using XGB) was then used to iterate over the sdAb sequence, find the most problematic location, mutate it to a better amino acid, and then repeat process N times. In the case of the ordered sequences N was set to 20. Applying this process to low Tm sdAbs, each predicted Tm increases. C) Visualization of the region-level frequency of mutations selected by the model for mutation.

By visualizing the frequency of mutations categorized by region suggested by LIME after 30 iterations (**Figure 5B**), most mutations are not in the antigen binding domain, increasing the probability of maintaining antigen SEB-binding activity (**Figure 5C**). As can be seen in **Figure 6** (and **Supplemental Figure 1**), LIME produces sdAbs that are very highly divergent in sequence from both their starting sdAb sequences and the previously discussed RNN-generated sequences, likely to due to the high number of mutated residues, particularly after a large number of LIME generations. Production of LIME sdAbs for subsequent experimental characterization are in-progress.

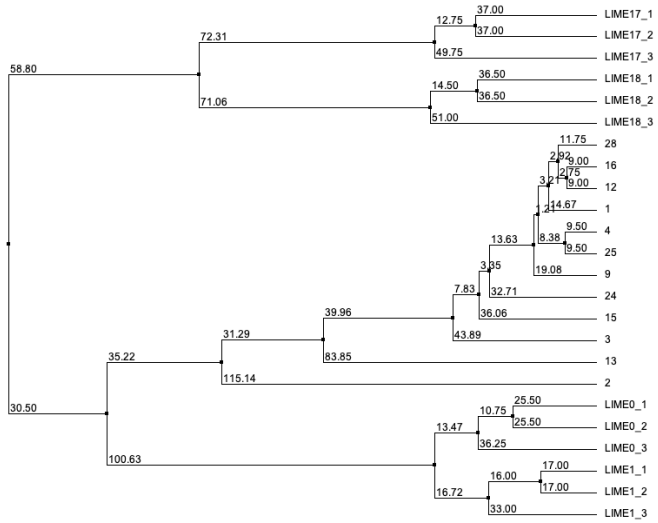


Figure 6. Sequence comparison. Sequence distance tree, according to BLOSUM algorithm. LIME-produced sdAbs are very highly divergent from RNN sdAbs.

1.2.3 Melting temperature prediction model

In order to develop our own sdAb T_m prediction model, we made use of a handful of tree-based machine learning algorithms: EXtreme Gradient Boosting (XGB), Gradient Boosting (GB), Light Gradient Boosting (LGB) and Random Forest (RF). Tree-based models were utilized with the forethought that explainable AI techniques, or techniques that assist in removing the black box from machine and deep learning and helping explaining model output, are particularly well-suited for use with decision-tree based models. Each implemented in Python, we trained the models using the small, characterized dataset of SEB-binding sdAbs. Representative starting point results are shown in **Figure 7A**. Following repeated shuffle split cross-validation, XGB with a train-test split of 90:10 was found to be the most accurate model by R^2 with a median of ~0.6 (**Figure 7B**). The root-mean square error (RMSE) for the best performing model was approximately 4.0. Given the size of the dataset, this performance was expected and sufficient for distinguishing between low, medium, and high T_m sdAbs with high confidence.

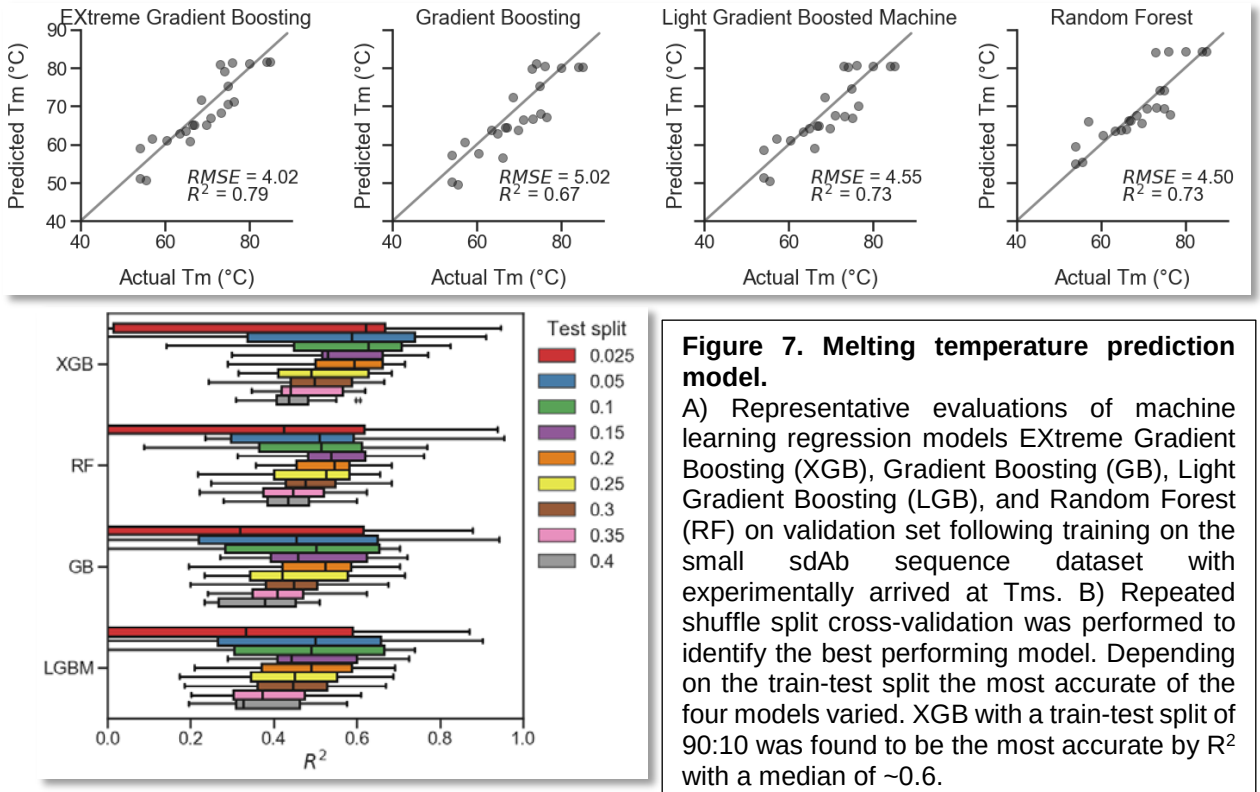


Figure 7. Melting temperature prediction model.
A) Representative evaluations of machine learning regression models EXtreme Gradient Boosting (XGB), Gradient Boosting (GB), Light Gradient Boosting (LGB), and Random Forest (RF) on validation set following training on the small sdAb sequence dataset with experimentally arrived at Tms. B) Repeated shuffle split cross-validation was performed to identify the best performing model. Depending on the train-test split the most accurate of the four models varied. XGB with a train-test split of 90:10 was found to be the most accurate by R^2 with a median of ~0.6.

1.2.4 Recombinant protein production

Genetic sequences encoding both the RNN and LIME-generated were synthesized by ATUM Biosciences. RNN sdAbs were supplied in storage vectors and required mobilization into bacterial expression plasmids. These added steps of cloning and sequence confirmation slowed bacterial expression. LIME-generated sequences were therefore constructed and provided in bacterial expression plasmids. This should speed evaluation of AI/ML-generated sdAb sequences using higher-throughput methods such as cell-free protein synthesis.

Recombinant sdAb production was carried out in the bacterial expression strain *Escherichia coli* BL21(DE3) using a plasmid vector that directed recombinant proteins to the periplasmic space. As a reducing environment, the periplasm allows for the formation of disulfide bonds though at the cost of overall protein expression levels. Each RNN-generated sdAb construct was studied to determine optimal expression conditions. For each construct expression conditions including temperature, inducer concentration, and duration of expression was examined.

Recombinant proteins were purified from whole cell lysates using a combination of chromatography protocols and dialysis. Both RNN- and LIME-generated sdAbs sequences possess a C-terminal hexahistidine sequence that allows for purification via immobilized metal affinity chromatography (IMAC). AS seen in **Figure 8**, IMAC purification removed the majority of the cellular proteins, however, higher molecular weight (MW) products were also captured which may be the result of aggregated or multimeric forms of the sdAbs. Therefore a secondary purification scheme using size exclusion chromatography was also employed to separate monomeric sdAb proteins from aggregates and other cellular contaminants.

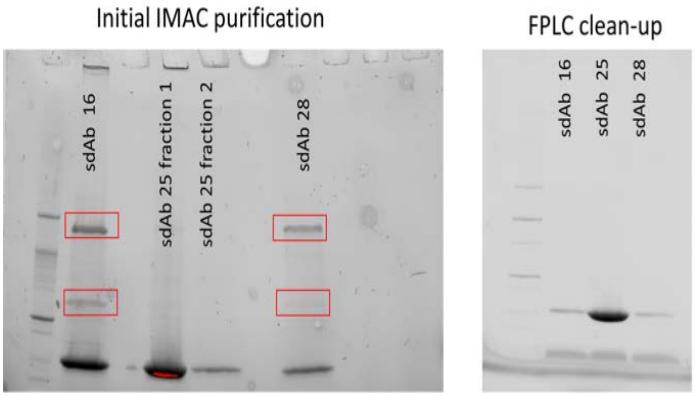


Figure 8. Purification of RNN sdAbs Coomassie stained SDS-PAGE gel following initial IMAC purification and FPLC clean-up of RNN sdAbs #16, #25, and #28.

1.2.5 Analysis of Biophysical Properties

Similar to conventional antibodies and other biomolecules, the amino acid sequence of conserved regions of sdAbs have evolved to convey biophysical properties such as *in vivo* stability, host tolerance (pharmacokinetics), and other attributes. Subtle variations in amino acid sequence can lead to significant changes in sdAb properties as expected. This program targets AI/ML training to avoid such alterations for antibody/biomolecule design but requires a dataset be established that encompasses not only successes but observations that produce non-viable/non-functional products. As an example, expression and purification of our first three sdAbs (sdAb 16, 25, and 28) showed interesting changes in protein folding and multimerization. While it is known that some sdAbs will multimerize, these proteins are primarily monomers under normal conditions. As shown in **Figure 8,** both sdAb 16 and sdAb 28 showed increased multimerization during purification which led to aggregation of purified protein. In the subsequent section we detail methods and results in determine some of the biophysical properties of the RNN-generated sdAb sequences. This data is incorporated directly into the sdAb dataset generated in this program to improve AI/ML predictions in the next rounds of evaluation.

1.2.5.1 *Thermal stability*

Melting temperatures were assessed using a SYPRO orange dye melt assay. In the SYPRO orange assay, the first derivative that was taken for each sdAb melting curve indicated a range of T_m from 57.7 to 93.4 °C, for #12 and #16 respectively, with a mean of 71.3 °C (**Figure 9A**). In these initial studies, protein thermostability did not coincide with predictions using aliphatic index for all of the RNN-sdAbs produced. This result was anticipated as there was a minimal physicochemical information data available for the majority sdAb sequence used to train the deep learning algorithm, and the relationship between T_m and aliphatic index, while present, is weak. The data generated in these initial studies will be used to iteratively improve all models used to further improve subsequent rounds of predictions and sequences generated. The in-house developed XGB-based Tm predictor, however, performed better than correlation with aliphatic

index. Although each of the predicted T_m for the RNN sdAbs ranged between a relatively narrow 64 and 80 °C (**Figure 9A**), a smaller range than what was found experimentally, the results from our T_m prediction model had a useful positive correlation with the experimental results (**Figure 9C**). These promising results suggest our T_m predictor could be used in the LIME-based model construction.

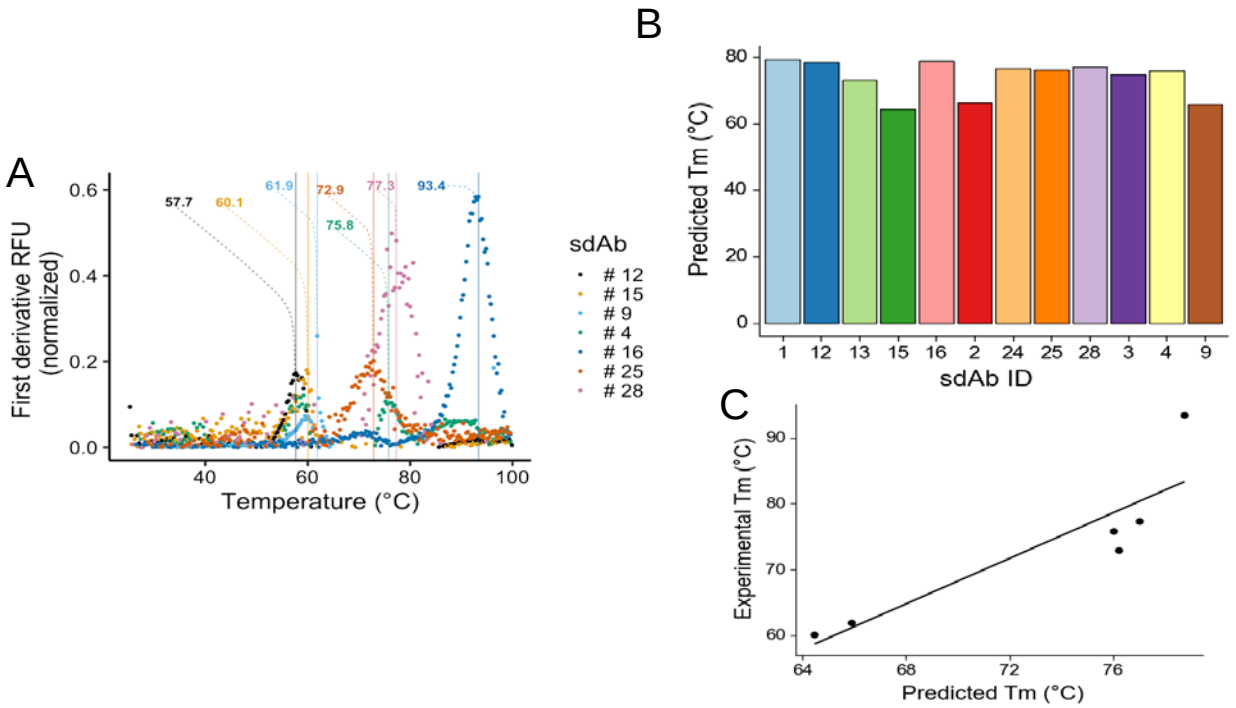


Figure 9. Melting temperature determinations and prediction evaluation.
A) Melting temperatures of RNN sdAbs were assessed using a SYPRO orange dye melt assay. The first derivative taken for each sdAb indicated the Tm. B) Prediction results from sequence information alone using our developed XGB-based Tm predictor. C) Correlation between experimental and predicted Tm results of the RNN-generated antibodies examined thus far.

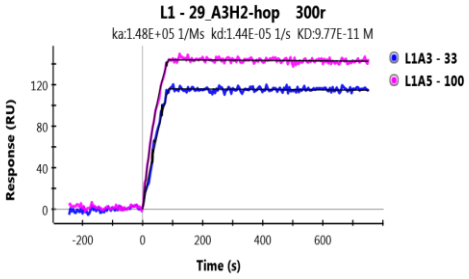
1.2.5.2 *Binding affinity*

As has been seen in other studies, changes in the primary amino acid sequence of sdAb and other recombinant antibody-like platforms can directly impact antigen specificity and other biophysical properties. Here we used Surface Plasmon Resonance (SPR) to determine antigen binding capabilities. Direct binding assays in which SEB antigen in immobilized to gold surfaces and RNN antigens were flowed over the surface were performed. This assay format allows for determination of not only the sdAbs affinity for the target antigen but also the rate of dissociation which provides insight as the strength of protein – protein interactions. As seen in **Figure 10**, comparing RNN-generated sdAbs to the known SEB-binder, A3H2, all save one – RNN #15 – bound to SEB with varying degrees of affinity (**Figure 10A**). Importantly, RNN #25 showed a similar level of affinity to the control with a kd (s^{-1}) of 4.7 x 10^{-5}, supporting the hypothesis that SEB-binding sdAbs can be produced by our RNN-based technique without any manual intervention. Response curves for both control A3H2 and RNN #25 are shown in representative plots in **Figure 10B and 10C**, respectively.

A

	ka ($M^{-1}s^{-1}$)	kd (s^{-1})	KD (M)
sdAb A3	$1.48E^{+05}$	$1.44E^{-05}$	$9.77E^{-11}$
RNN 4	$1.14E^{+05}$	$2.43E^{-04}$	$2.13E^{-09}$
RNN 9	$8.49E^{-01}$	$3.82E^{-04}$	$4.50E^{-04}$
RNN12	$3.49E^{+04}$	$4.01E^{-05}$	$1.15E^{-09}$
RNN15	-	-	-
RNN16	$3.53E^{+03}$	$7.58E^{-06}$	$2.15E^{-09}$
RNN 25	$6.29E^{+04}$	$4.67E^{-05}$	$7.43E^{-10}$
RNN 28	$9.25E^{-01}$	$6.57E^{-09}$	$7.10E^{-09}$

B



C

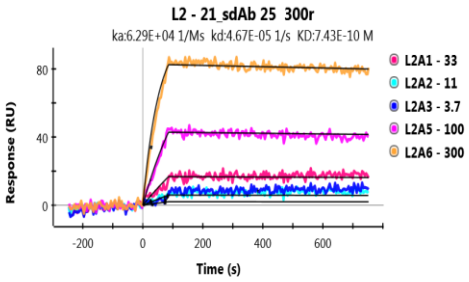


Figure 10. Surface plasmon resonance (SPR) experiments on RNN sdAbs.
Binding affinity of the RNN-generated sdAbs to SEB was confirmed via SPR. A) Obtained binding constants for each of the sdAbs, with A3 as a control known binder. B) and C) Representative response curves of A3 and RNN 25, respectively.

1.2.5.3 Aggregation potential

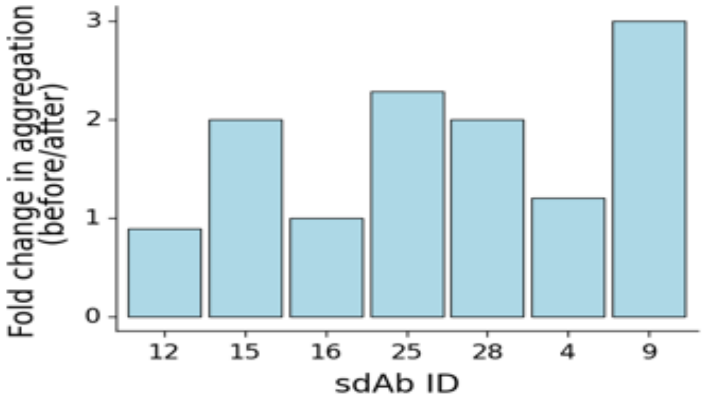


Figure 11. Aggregation of sdAbs following thermocycling.
Fold change in aggregation measured by protein concentration before and after thermocycling between room temperature and 70 °C. Each sdAb ranged between ~1.0 (no aggregation) and ~3.

In addition to T_m, aggregation potential is another attribute of focus for antibody design and optimization. In order to perform an initial investigation into the aggregation potential of the RNN-designed antibodies, proteins were denatured through thermal cycling and precipitation/ aggregation monitored through measurement of the optical density at 280 nm (OD_{280}). Total protein concentration was measured prior to and following temperature oscillations between 25 and 70 °C. After this, fold change was calculated where ~1.0 indicates no increase or decrease in measured protein concentration, and therefore no introduced aggregation. Results showed that each of the examined sdAb ranged between ~1 and ~3, with the measurable protein concentration of RNN #9 decreasing significantly following

thermocycling indicating a large increase in aggregation (**Figure 11**). In future when optimizing for aggregation is targeted, experimental controls such as A3H2 used above will be used for comparison.

1.2.6 In silico Modeling

Given the diversity of the amino acid sequences compared to parental and RNN-generated sdAbs (see **Figure 6**) i*n silico* characterization of LIME antibodies was performed. Tertiary structures of LIME-generated sdAbs were predicted via a deep learning model by I-TASSER (Yang and Zhang 2015) which provides protein structure prediction and structure-based function annotation. Sequence identities and mutations compared to their respective parent sdAbs were obtained using multi-sequence alignments with protein BLAST. The ligand binding site residues were predicted by COACH This server generates its predictions using two comparative methods, TM-SITE and S-SITE, which recognize ligand-binding templates from the BioLiP protein function database by binding-specific substructure and sequence profile comparisons. Based on these results, **Table 2** shows consistency of the LIME-based method in selecting mutations for potential novel antibodies (e.g., the SDT mutation site in each of the progeny sequences) and maintaining high sequence identities with each parent. Visualizing the predicted of PDBs (in 3D where the

Reference sdAb	LIME-sdAb	Sequence Identity	Mutated Residues	Total Residues Mutated	Ligand Binding Site Residues
Ca (GenBank: KU508542.1)	LIME_0_1	100%	clone	0	31,32,33,35,50,51,52,53, 56,57,58,98
	LIME_0_2	95%	6-7 (DT), 13 (R), 17 (R), 21 (T), 112 (W)	6	31,33,35,47,50,51, 52,53,56,57,58,98, 111
	LIME_0_3	90%	5-7 (SDT), 13 (R), 16-17 (FR), 21 (T), 36 (R), 57 (T), 104 (V), 109 (D), 112 (W)	12	31,32,33,50,52,56, 58,98, 99,102,111
Cd (GenBank: KU508545.1)	LIME_1_1	100%	clone	0	31,32,33,35,47,50, 51,52,56,57,58,98, 111,112
	LIME_1_2	96%	5-7 (SDT), 13 (R), 21 (T)	5	31,33,35,47,50,51,52,53,56,57,58,98,111
	LIME_1_3	91%	5-7 (SDT), 13 (R), 17 (R), 21 (T), 36 (T), 57 (V), 102 (A), 104 (V), 112 (W)	11	31,32,33,35,50,51,52,53,56,57,58,98,104
S222-A2	LIME_17_1	100%	clone	0	33,35,37,47,51,96,97,98, 104,106
	LIME_17_2	96%	5-7 (SDT), 13 (R), 21 (T)	5	31,32,33,35,50,51,52,53,56,57,58,98
	LIME_17_3	91%	5-7 (SDT), 13-14 (RA), 17 (R), 21 (T), 36 (H), 54 (S), 92 (M), 94 (W)	11	33,35,37,50,96,98,106
H9-B11	LIME_18_1	100%	clone	0	26,27,31,32,33,50,52,98,99,100
	LIME_18_2	95%	5-7 (SDT), 13-14 (RA), 21 (T)	6	33,35,37,47,50,51,98,104
	LIME_18_3	90%	5-7 (SDT), 13-14 (RA), 17 (R), 21 (T), 29 (A), 36 (H), 67 (V), 83 (A), 104 (V)	12	33,50,52,56, 58,98

Table 2. Control sdAb and LIME-generated sequences and their mutations

LIME-sdAbs depicted as lime green color and mutated residues are highlighted purple; see **Figure 12**) suggests that there may be some significant changes to tertiary structure. Looking at a representative Ca antibody-based series, the mutations (LIME_0_1 through LIME_0_3) the 12 mutations do result in some disturbance to the clearly defined beta sheets of the parent. Protein production and experimental evaluation of both SEB-binding and thermostability will need to be performed in order to determine whether these changes are favorable.

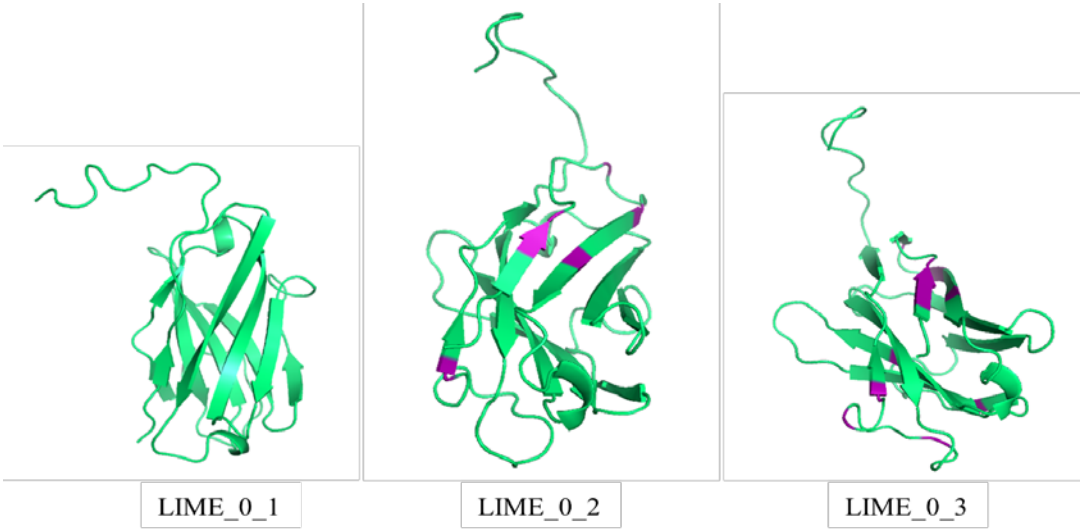


Figure 12. *In silico* modeling of sdAb sequences
LIME_0_1 through LIME_0_3 are secondary structure predictions in 3D where the LIME-sdAbs depicted as lime green color and mutated residues are highlighted purple. These structures are visualized through PyMOL software.

1.3. Conclusion and Future Development

This effort focused on building capabilities and examining a number of machine learning systems for their potential in biomolecule design. This is represented by the significant investment in building the IT capabilities within NRL and its subsequent utilization in several machine learning-based approaches to the single domain antibody design. Currently, there are a considerable number of ML/AI algorithms that are applicable to biomolecule design but few have been empirically tested in the laboratory. Initial efforts focused on a *Recursive Neural Network* (RNN) and less well-known *Local Interpretable Model-Agnostic Explanations* (LIME) for dataset analysis and prediction of sdAb sequences with improved thermal stability.

In silico methods of evaluating protein structure as a tool for predicting thermostability and antigen recognition were also examined in this study. Initial results suggest that partnering predictive and design tools such as *iTasser* and *Rosetta* could prove invaluable in down-selecting those sequences output from AI/ML algorithms based on known protein folding patterns and molecular assembly. Additionally, such capabilities could provide another avenue of further improving biophysical properties through modeling and direct protein design and mutagenesis.

In conclusion, the team was able to assemble an extensive sdAb dataset which was served as foundational work in AI /ML-based biomolecule design. While biomolecules such as conventional antibodies, cellular ligands, and others are targeted biomolecules for future studies, the complexity of these molecules and their interactions will require greater understanding of AI/ML methodologies and how they can be adapted to specific applications in biological systems. The development of these tools and methods will be a focus of future efforts.

While efforts will continue with the sdAb platform to develop viable methods, future efforts will include the production and purification of conventional antibodies which will require specialized equipment and protocols. The team examined a number of commercial systems for recombinant biomolecule production in mammalian cell lines. Leveraging existing research programs within NRL, the research group has been granted access to much of the necessary equipment and is currently seeking NRL approval for the relevant expression cell lines. Additionally, the group has connected with researchers from FiberCell Systems, a small business that manufactures and distributes a hollow fiber bioreactor that will be used for the scalable production of each antibody sequence generated through the ML/AI method and partnered *in silico* modeling efforts.

Progress

- **IT infrastructure** including the purchase Lambda Quad system with 4x RTX5000 GPUs and two desktop computer with RTX2080 Tis to facilitate ML/AI systems.
- **Assembly of sdAb dataset** comprised of over 6 million sequences from both published and NRL-derived amino acid sequence libraries
- **Algorithm training** and sdAb sequence predictions using *RNN* and *LIME* systems.
- **Testing and evaluation** of ML/AI-determined sdAb sequences assessing thermal stability, aggregation potential, and binding kinetics.

Next steps

- **Testing and evaluation** of ML/AI-determined sdAb sequences assessing thermal stability, aggregation potential, and binding kinetics.
- **Establish conventional antibody dataset** including sequence clustering and organization to enable ML/AI training.
- **Establish conventional antibody production platform.**

1.4 Acknowledgements

This research was supported by core funds from the Naval Research Laboratory (WU 1V33) and the Defense Threat Reduction Agency (DTRA #19-143).

1.5 References

Dunbar, J., et al. (2014). "SAbDab: the structural antibody database." Nucleic acids research **42**(D1): D1140-D1146.

Karpathy, A. (2015). "The Unreasonable Effectiveness of Recurrent Neural Networks."

Müller, A. T., et al. (2018). "Recurrent neural network model for constructive peptide design." Journal of chemical information and modeling **58**(2): 472-479.

Ribeiro, M. T., et al. (2016). " Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.

Turner, K. B., et al. (2016). "Next-generation sequencing of a single domain antibody repertoire reveals quality of phage display selected candidates." PLoS One **11**(2): e0149393.

Turner, K. B., et al. (2014). "Isolation and epitope mapping of staphylococcal enterotoxin B single-domain antibodies." Sensors **14**(6): 10846-10863.

Wilton, E. E., et al. (2018). sdAb-DB: the single domain antibody database, ACS Publications.

Yang, J. and Y. Zhang (2015). "I-TASSER server: new development for protein structure and function predictions." Nucleic acids research **43**(W1): W174-W181.

Zuo, J., et al. (2017). "Institute collection and analysis of Nanobodies (iCAN): a comprehensive database and analysis platform for nanobodies." BMC genomics **18**(1): 1-5.