NORTHWESTERN UNIVERSITY

# Finding the Needle in the Haystack: Applying Data Science to Address Biological Questions

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Chemical and Biological Engineering

By

Albert Y. Xue

EVANSTON, ILLINOIS

September 2018

# ABSTRACT

Finding the Needle in the Haystack:

Applying Data Science to Address Biological Questions

Albert Y. Xue

Biology is entering the exciting world of big data. Modern high-throughput experimental techniques often produce large datasets that aim to capture complex relationships often found in biological systems. While these larger data sets contain vast amounts of useful information, the answers are often locked behind a wall of numbers. As a result, the big data revolution has spawned the field of data science composed of scientific methods, algorithms, and systems to unlock useful information using modern data science tools that blend various tools such as statistical methods, machine learning models, and data visualization pipelines. When applied to new scientific fields, these tools accelerate the discovery and understanding of novel scientific insights.

In my thesis, I apply modern data science tools to various biological datasets to investigate the complex relationships and produce actionable insights that inform future experiments. The investigated datasets are united by the common theme of

big data and require data science tools to extract useful scientific results. In the first project, I investigate the signal quality of peptide arrays and call attention to the under-studied complexities of peptide behavior in mass spectrometers. For the second project, I extract useful synthesis designs of a potential nanoparticle cancer-immunotherapy, and I expand the capabilities of the synthesis pipeline using supervised machine learning models. The third project creates an improved and automated methodology to systematically label and visualize RNA folding events in SHAPE-Seq datasets. I conclude this thesis by discussing an issue present with many supervised learning models: how do we interpret models? I focus on deep learning interpretation techniques as applied to medical tasks and how these current techniques fall short of emulating clinical practices.

## Acknowledgements

My mentor, Professor Neda Bagheri has been receptive to my ideas and directed me towards the correct paths. She is integral in developing me not just as a student, but as a person, a professional, and beyond. Thank you for your guidance.

I thank my collaborators who have helped with all this work: Lindsey Szymczak, Jennifer Grant, Weston Kightlinger, Gokay Yamakurt, Professor Eric Berns, Angela Yu, Erin Garfield, and Victor Quan. The work in this dissertation is impossible without them.

My committee and other faculty gave me valuable advice, discussed research, and provided feedback. Thank you Professor Luis Amaral, Professor Doug Downey, Professor Pedram Gerami, Professor Julius Lucks, Professor Milan Mrksich, and Professor Fengqi You.

I also owe this work to the individuals in the Bagheri Lab past and present. I thank them for their creativity, diverse perspectives, scientific curiosity, and generally positive attitude. Their feedback has been helpful in shaping this work for the better.

I would especially like to thank my friends and family who have helped me enjoy my experiences in graduate school and guiding me through life.

# Table of Contents

# List of Tables

# List of Figures

CHAPTER 1

# Introduction

## 1.1. The new age of big data in biology

Understanding and unraveling the complex world of biological systems has always been challenging. Fortunately, modern experimental techniques generate large datasets that capture the complicated relationships that often underlie biological systems. Whereas previous biology experiments were limited with restricted impact, newer high-throughput experiments allow a more comprehensive perspective to better understand the complex world of biology. These big data experiments are becoming bountiful, and a simple Google Scholar search for "big data biology" currently produces over 1.6 million results. Recently, big data applications have led to successes across various biological fields including metabolism[1,2], genomics[3,4], medicine[5,6]. I define big data not in terms of the absolute size of data[7] but in terms of the relative data size increase that enables application of modern data science tools. Modern big-data generating experiments include SAMDI, a technique that immobilizes various peptide sequences to a surface[8]. SAMDI allows simultaneous probing of hundreds or thousands of peptides in a single experiment, which represents a significant increase over the status quo. Previous peptide profiling techniques analyze few peptide samples[9,10], but SAMDI simultaneously queries hundreds to thousands of samples[8] and represents one of the many high-throughput techniques generating big data. SAMDI

and similar jumps in data-generation capabilities allow academic research to leapfrog from limited discoveries to broader impacts to accelerate scientific advances within biological domains.

## 1.2. Finding the needle in the haystack

Big data leads to a new challenge: if big data can be represented as the haystack, it is increasingly challenging to find the needle, or key biological insights, buried in a sea of experimental data. These challenges span multiple issues including visualizing data, understanding complex relationships, or extracting actionable insights. In terms of visualization, it is difficult to construct an effective visual for a high-dimensional dataset.[11] Though humans have highly developed visual processing centers, it is limited to three dimensions and is easily leveraged for high-dimensional dataset. Similarly, big data means that simple 2-dimensional plots often bring an overwhelming amount of information that is difficult to digest and understand. As a result, traditional visualization techniques are often insufficient, and creating understandable visuals requires significant forethought and integration with newer tools. This principle is generalizable across big data analysis and is especially relevant for identifying those important nuggets of information within a dataset. The following introductory sections explore how modern data science tools are able to tackle the challenges present in big data. Chapters 2-4 focuses on finding the needle within the haystack of various biological datasets.

## 1.3. Solving big data with data science tools: data visualization and machine learning

I apply tools from the field of data science, a new discipline that mixes and combines data visualization, statistics, machine learning, and other data-related domains, to tackle the challenges of understanding big data. Unlike traditional approaches that concentrate on understanding one of these disciplines, data science requires familiarity and usage of these tools in an integrative manner. In addition, data science intimately combines the quantitative tools with the studied domain. For example, my knowledge of organic chemistry improved the effectiveness of data science tools when applied to a chemical dataset in chapter 2. As a result, a data scientist should be able to match domain-specific hypotheses to the specific computational tool yielding specific insights for subsequent experimental validation.

Of these tools, newer visualization implementations enable both flexible and powerful methods to construct unique visualizations that effectively represent larger data sets better than older implementations. These newer implementations include both R and python open-sourced languages as well as the visual-focused tableau and D3 frameworks. Where visual tools offer a qualitative view into big data, machine learning tools provide a quantitative method to understand specific relationships. Machine learning blends older statistical tools with modern optimization techniques and have undergone revolutionary innovations in terms of flexibility, accessibility, and power. The hallmark of machine learning is that the computer *learns* (optimizes itself) to better perform a task. Due to this flexibility, these tasks are broad and

include predictive modeling, classification, data clustering, natural-language processing, decision-making systems, and more. This combination of flexibility and power allows researchers to train a model to perform unique tasks and further understanding in emerging big data experiments.

### 1.3.1. Refined data visualizations convey more compelling messages

I believe that data visualization, when done properly, is a good starting point for understanding big data. Larger datasets contain far too many numbers to intuit with a table, and identifying the correct visualization is non-trivial. For example, slight improvements in formatting lead to dramatic improvements in visual effectiveness. Changing the arrangement of a published SAMDI dataset visualization enables easier identification of trends that correlate with the variable of interest. This dataset explores the relationship between peptide sequence and deacetylation activity.[8] Peptides are composed of a sequence of amino acids, represented as letters. Here, each peptide has a unique pair of 19 amino acids in two variable positions (X or Z) and are exposed to cellular lysate to measure deacetylation activity. Deacetylase is an enzyme that leads to a chemical reaction called deacetyation and is important in various biological process, such as stem cell differentiation.[12] When arranged in a 2-dimensional heatmap (Fig. 1.1 left), color represents the degree of deacetylation and each row/column is an amino acid. However, this implementation does not easily convey the effect that each amino acid has upon deacetylation. Which X-position amino acid is associated with the highest deacetylation, or the second highest? Are

there outlier peptides? In addition, the SAMDI experiment includes replicates, but the heatmap representation collapses the replicates into a mean value. As a result, it is unclear if any peptide has consistently high deacetylation activity or not. This representation does not easily untangle the needle in the proverbial haystack.

In contrast, the right bubble chart conveys a more clear message about the data. This representation sorts the rows/columns by mean deacetylation and scales the bubble size by the variance across replicates. The sorted amino acids convey an ordering of how amino acids relate to deacetylation activity and informs how to design future experiments around peptide sequences. In addition, the bubble size enables the user to intelligently judge whether or not to trust specific measurements; selecting a peptide for high deacetylation is worthless if the measurement is called into question based on noise. This formatting also displays the few peptides that buck the trend and generates hypotheses for future experimental testing. This example demonstrates that refined designs lead to integrated visuals that deliver more effective messages. Well-constructed visual representations leads to a better view when finding that needle in the haystack.

### 1.3.2. Interactive data visualization promotes rapid testing of effective visual strategies

Older visualizations are often static, resulting in visuals in the form of pdfs or passive figures. These simpler visual tools tend to explore few dimensions, which results in

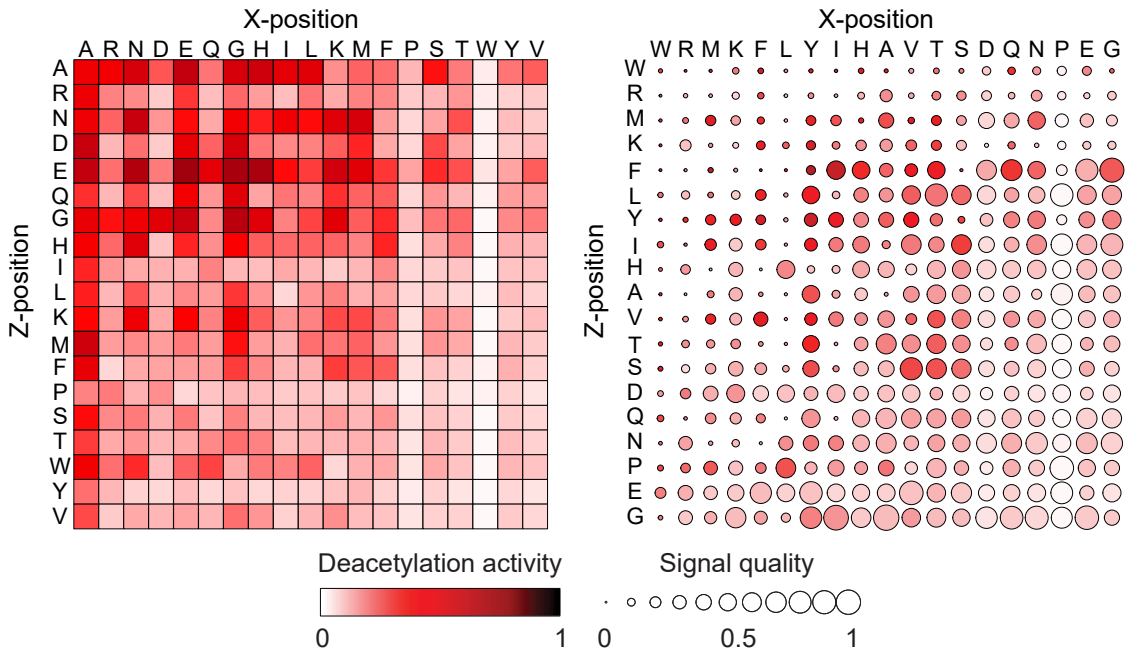



Figure 1.1. **Effective visual implementation improves understanding of complex datasets** Both left and right figures display the same dataset. This data examines deacetylation activity of 361 peptides with variable X- and Z- positions. Color and bubble size represent activity and signal quality, respectively. The right visual sorts the amino acids by mean activity, giving a clear ordering and effect on activity. The right visual also integrates signal quality, informing the audience of which peptides to focus upon, without compromising the clarity of deacetylation activity. As a result, the right design delivers more relevant information and is representative of careful design to understand big datasets.

a naive perspective of a large-dimensional dataset. When a dataset contains hundreds of variables, there are many thousands of combinations of 2- or 3- dimensions and selecting the correct dimensions for basic visualization seems an impossible task. Newer visualization tools, such as shiny or D3, are designed with user-interactivity in mind, fostering more flexible visuals to adapt for the particular task. For example, the right visual in Figure 1.1 contains a different message when the rows or columns are sorted differently. However, the most effective sorting method is not obvious. This process is not easily automated because visual effectiveness depends on our human visual processing, which is not easily expressed in a computer. Fortunately, exploring a different ordering is realizable with a small programmatic change, as evident with a simple toggle switch in a user-interface (UI) found at github.com/bagherilab/bubble_chart_app. By adding a simple toggle or other flexible control devices, interactive visualization methods lead to figures that are better tailored for understanding.

Interactive visual tools allow manual tuning when no good automatic optimization method exists, such as in the dimensional stacking visual[13] in Figure 1.2. Here, the different explanatory variables are "stacked" upon one another in columns/rows and the bubble color represents the response variable (for a detailed explanation, go to Chapter 3). The variable stacking order has a large effect on visual clarity and message. When comparing Figure 1.2 to Figure 1.3, the poorly-ordered first visual is incongruous and almost random. There is a regular pattern in the bubbles but the relationship between explanatory variables and response is not well illustrated.

In contrast, the bottom visual has been fine-tuned and the strongest relationships become obvious. This particular example contains 80,640 possible combinations of variable ordering; displaying all orderings and visually selecting the most effective one is impractical. Similarly, automatic optimization of variable ordering is non-obvious, suggesting that human-directed tuning is required. By creating an interactive UI (github.com/bagherilab/dimensional_stacking), the intuitive human visual processing abilities are leveraged to explore and optimize the possible orderings in a rapid testing manner. Interactive visual tools lower the barrier for exploring and determining approaches to visualize big datasets.

### 1.3.3. Supervised machine learning tools perform a variety of tasks and enable flexible hypothesis testing

Visualization tools by themselves cannot unlock the information with big data. Much of the following work relies on supervised machine learning tools to understand various biological datasets. This reliance is due to supervised learning's core ability to convert the explanatory variables, $X$, into a response vector, $Y$, using a learning functions (also called a model), $f$, and is represented as a general equation, $f(X) = Y$. $X$ is a matrix and contains rows and columns of data that represent samples and variables (also called features), respectively. $Y$ is typically a vector of the predicted variable (also called the response). $f$ is the learning equation and is adaptable to different instances of $X$ and $Y$ because of internal optimization methods that attempt to better convert $X$ into $Y$. To gauge how well a supervised learning model has fit

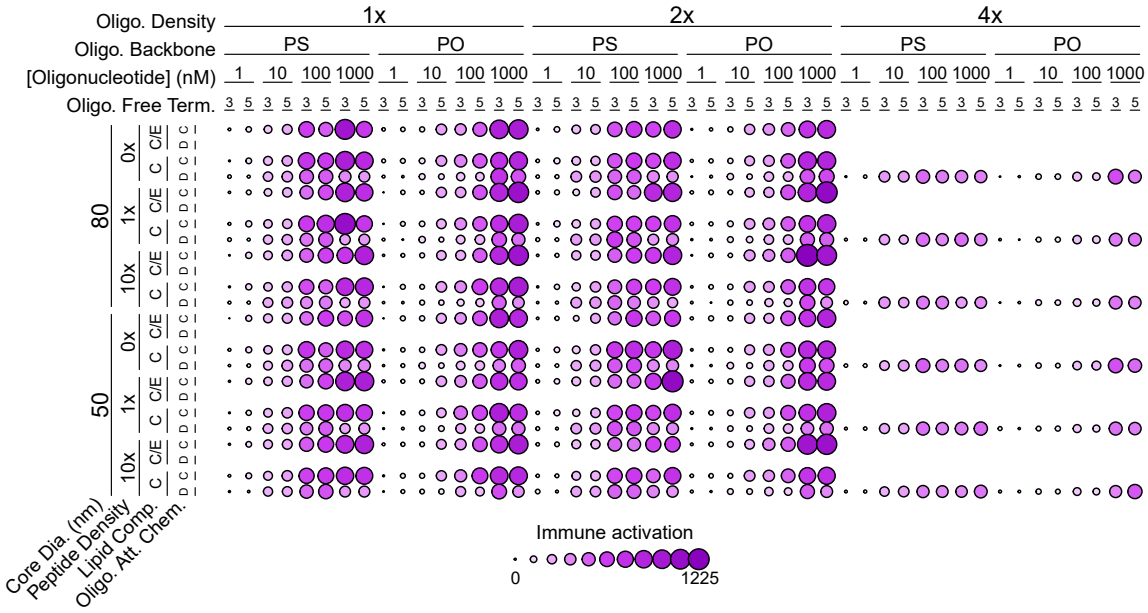



Figure 1.2. **A non-optimized visual does not readily display relationships within data** This dimensional stacking visual shows the relationship between eight nanoparticle design properties (shown in rows/columns) and the response, immune activation, shown in bubble size/color. The variables are randomly ordered in the row/column levels and in contrast with Figure 1.3, the importance of variables and their effect on immune activation is unclear.

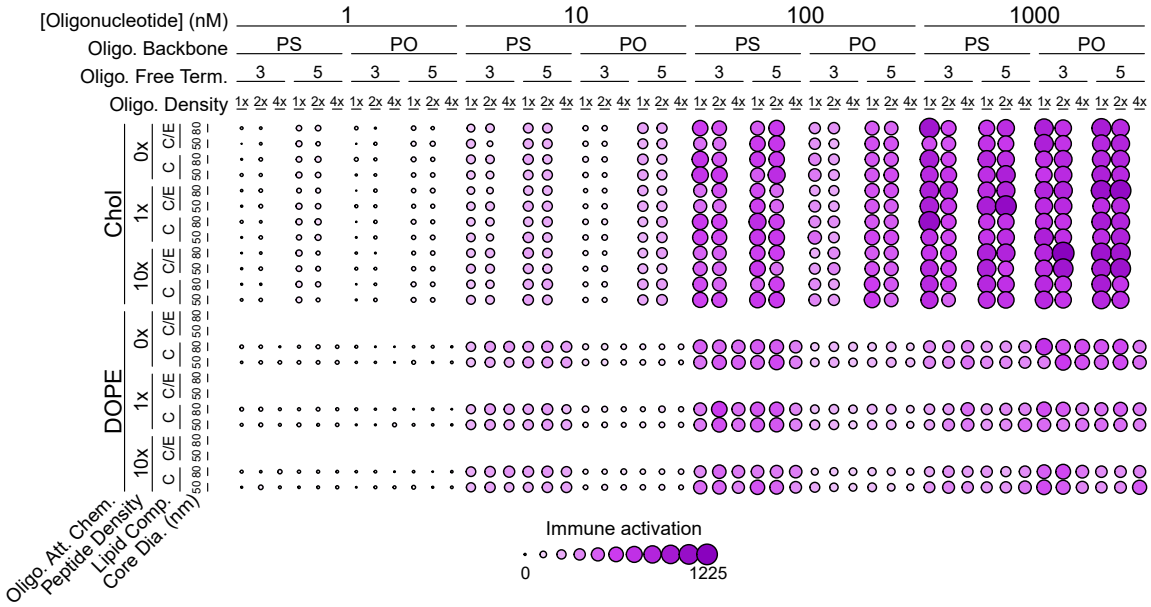



Figure 1.3. **A visual optimized with interactive visual tools clearly conveys important relationships within a dataset** In contrast with Figure 1.3, this visual has been optimized by experts in this nanoparticle design space. This interactive process results in a clean visual that displays the important/unimportant relationships. For example, the variable [Oligonucleotide] has the strongest effect on immune activation and has been placed at the top level of the columns, resulting in the general trend of immune activation increase from left to right.

to the data, cross-validation is used where a training data subset is used to train (optimize) the model, and a separate test dataset (randomly selected rows) is left out for model testing.[14] Performance is measured from how well the model predicts the response within the test dataset.

Often, predicting $Y$ is the central goal of supervised learning, but I apply supervised models to extract deeper meaning and understanding within a dataset. Modifications in $f$, $X$, or $Y$, allow for flexible hypothesis testing to reveal unseen relationships. Put broadly, if a modification leads to increases/decreases in predictive performance, then the modification is relevant to understanding the data. Changing the learning function is a good starting point; they come in many different flavors, ranging from linear to highly non-linear methods, and their performance belies the underlying relationship between the $X$ and $Y$. For example, if a non-linear model outperforms a linear model when predicting $Y$, then I can conclude that there exist non-linear relationships between variables within $X$ and with $Y$. Other than $f$, modifying $X$ also lead to changes in predictive performance. By selecting different columns of $X$, in a process called feature selection, model performance can be increased/decreased[15], suggesting that some features are more predictive or relevant than others. For example, removing an irrelevant feature is expected to improve model performance because the model is not led astray. Similarly, adding in redundant features is not expected to improve performance. This style of hypothesis testing is highly flexible because all combinations of features are potentially explorable, and there are not necessarily assumptions about the features that is present in statistical

testing (normality, independent distribution, etc.). As a result, this analysis focuses attention towards the most relevant features for deeper investigations.

Modifying $X$ is not limited to the columns/features, but includes modifying the number of rows or samples. Because biological experiments are usually costly and resource-intensive, future experiments benefit greatly if a smaller experiment yields the same amount of information. In other words, if a supervised learning model accurately predicts $Y$ from a few samples, then future experiments need only synthesize those few samples to train a model to predict the remaining unsynthesized samples. Depending on how few samples are needed, this analysis expands the capabilities of a typical high-throughput experiment several times over, which improves and accelerates future discoveries.

## 1.4. Organization of dissertation

In this work, I apply modern data science tools to three domains. In Chapter 2, I investigate how to improve peptide array experiments by exploring how peptide sequence affects signal-to-noise ratio in mass spectrometry. This research focuses on an under-explored aspect of many peptide experiments involving mass-spectrometry: how does the peptide sequence or composition affect the quality of its signal in a mass spectrometer? I uncover results demonstrating a strong relationship between sequence and signal quality and show that supervised learning models lead to predictive models for improving signal quality. The second project in Chapter 3 involves discovering optimal designs of immune-activating nanoparticle with implications in

cancer therapeutics. This work examines the relationship between nanoparticle design properties and immune activation to inform future experiments. In addition, the supervised learning models demonstrate that relatively few synthesized nanoparticles are able to predict immune activity of a larger set of nanoparticles, which greatly expands experimental capabilities. The final research project in Chapter 4 focuses on detecting interesting RNA folding events in SHAPE-Seq data and integrates interactive visual UIs with a custom-made event detector. Here, there are few data points, prohibiting application of supervised learning models and forcing me to carefully engineer an event detector from simple assumptions. The final event detector enables a systematic and quantitative method to detect events that is less prone to error in human judgment. In the concluding remarks of Chapter 5, I address a common concern of supervised learning models: their (lack of) interpretability. I specifically highlight how the machine learning and medical communities are isolated, leading to model interpretation techniques that perform well in one community, but not the other. I state that future supervised model development and implementation should become increasingly integrated with the end application, especially when it includes model interpretation.

CHAPTER 2

# Machine learning on signal to noise ratios improves peptide array design in SAMDI mass spectrometry

This work was published with Lindsey M. Szymczak and Professor Milan Mrksich in ACS Analytical Chemistry in 2017.

## 2.1. Abstract

Emerging peptide array technologies are able to profile molecular activities within cell lysates. However, the structural diversity of peptides leads to inherent differences in peptide signal to noise ratios (S/N). These complex effects can lead to potentially unrepresentative signal intensities and can bias subsequent analyses. Within mass spectrometry-based peptide technologies, the relation between a peptide's amino acid sequence and S/N remains largely non-quantitative. To address this challenge, we quantify and analyze mass spectrometry S/N of two peptide arrays, and we use this analysis to portray quality of data and to design future arrays for SAMDI mass spectrometry. Our study demonstrates that S/N varies significantly across peptides within peptide arrays, and variation in S/N is attributable to differences of single amino acids. We apply supervised machine learning to predict peptide S/N based on amino acid sequence, and identify specific physical properties of the amino acids that govern variation of this metric. This study illustrates how machine learning can

accurately predict the S/N of a peptide given its sequence, allowing for the efficient design of arrays through selection of high S/N peptides.

## 2.2. Introduction

Peptide arrays have emerged as an enabling tool for identifying biologically relevant peptide substrates and molecular recognition sites, and hold great promise as a new analytical method for basic and translational research in the biomedical sciences[16,17]. Uses of peptide arrays include measuring changes in enzymatic activity specifically enzymes that add or remove post-translational modifications to gain insight into different cellular pathways and processes[18,19,20]. Other applications include diagnostic or detection-focused arrays such as differential peptide arrays to detect specific analytes in complex mixtures[21,22], or diagnose diseases[23,24]. Many existing methods are based on either radioisotopic or fluorescent labels to detect reaction products[25,26]. These methods introduce additional protocol steps, and for the latter, can alter natural biological activity leading to false interpretations, as when resveratrol was erroneously found to enhance deacetylation on a peptide with an attached fluorophore[27].

We recently introduced the SAMDI mass spectrometry method, which uses MALDI mass spectrometry to analyze peptides that are immobilized to a self-assembled monolayer of alkanethiolates on gold (Fig. 2.1), and we have demonstrated the use of this method for profiling enzyme specificities[28], for discovering new enzymes[29], and for profiling activities in a lysate[8]. This method provides many benefits, including the use of surface chemistries that are intrinsically inert to the non-specific

adsorption of protein, the availability of a broad range of chemistries for immobilization of peptides, and, most significantly, the compatibility with matrix assisted laser desorption ionization mass spectrometry to analyze the masses of the peptide-alkanethiolate conjugates. This ability to directly measure peptide masses[30] allows a straightforward analysis of peptide modifications by identifying the corresponding mass shifts. This method has also been demonstrated to provide a semi-quantitative measure of the peptides' substrate activity[8]. However, the S/N of a mass peak for a peptide often depends on its amino acid sequence, resulting in both well-suited and poorly-suited peptides for inclusion in an array.

In practice, the signal to noise ratio (S/N) of a peptide in mass spectrometry can vary, making certain sequences poorly compatible with the detection method[31,32]. Hence, some fraction of peptides serves no useful purpose in an experiment. To identify peptide array designs that maximize S/N, we predicted S/N from amino acid sequences measured by SAMDI mass spectrometry. We identified amino acids associated with high S/N peptides in two peptide arrays and used machine learning to highlight properties that predict the relationship between amino acids and S/N. While SAMDI-specific results are not generalizable, the method we describe can be adapted and applied to diverse peptide array technologies.

Previous work has explored S/N relationships involving peptide charge (as with arginine residues)[33,34], or hydrophilicity, where hydrophilic proteins can be preferentially detected in MALDI-MS due to easier co-crystallization with MALDI matrix[35,36]. In addition to hydrophilicity, many specific and complex peptide-matrix

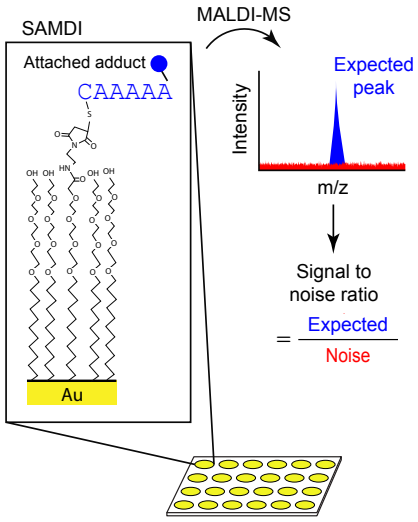



Figure 2.1. **Measuring S/N on peptide arrays using SAMDI MS.** SAMDI MS uses MALDI mass spectrometry to analyze peptides that are immobilized to a self-assembled monolayer of alkanethiolates on gold. Depending on the enzyme of study, the peptides may contain a chemical adduct, such as an acetyl group if deacetylases are the enzymes of interest. The expected peak before enzyme treatment includes the peptide immobilized to the alkanethiolate with the attached chemical adduct of interest. We quantify the expected mass peak and noise using their area under the curve to calculate peptide S/N.

interactions can explain MALDI peptide S/N[34 37 38]. Single amino acids have been reported to improve signal strength. For example, Krause and co-workers reported that peptides containing arginine or phenylalanine typically contributed to higher MALDI signal strength[39]. Additionally, the relationship between S/N and amino acid sequence gains complexity with the addition of chemical adducts. For example, Kolarich and coworkers reported peptides with attached N-glycans have altered signal strengths depending on MS instrument types or subtle changes to peptides from glycosylation[40]. Many studies use peptides that may have undergone oxidation[39 41 42 43] which likely also affects peptide signal strength. These peptide modifications introduce difficulties in signal detection and emphasize the need to integrate computational strategies to better understand the relationship between the amino acid sequence of a peptide and the quality of its signal. We select peptide libraries that are unbiased in their composition to evaluate differences in S/N due to differing amino acid sequences, and we offer a complete empirical analysis relating amino acid composition and S/N of the peptides.

Using statistical and machine learning strategies, we investigated how amino acid composition affects S/N in SAMDI mass spectrometry and how subtle amino acid differences can give rise to different S/N. We focus on two peptide arrays, each containing two consecutive variable positions (represented by all 19 amino acids except for cysteine). The amino acids surrounding the variable positions however are different. The two peptide arrays are Ac-GRKacXZC (K-array) and Ac-GXZHGC

(H-array). We collected peptide spectra by SAMDI mass spectrometry and calculated the S/N of each peptide. Statistical analysis identified amino acids associated with low or high S/N peptides. We trained machine learning models to identify factors that predict S/N from the physical properties of the peptide's amino acids. We then predicted the S/N of peptides and experimentally screened for high S/N peptides in SAMDI high-throughput data. Accurate prediction of peptide S/N from machine learning models allows for the selection of peptides that are better suited for inclusion in the array without costly screening.

## 2.3. Methods

### 2.3.1. Solid phase peptide synthesis.

Data was collected from K- and H-peptide array experiments. The K-peptide array synthesis and methods have been previously published[8] and contains peptides of the form Ac-GRK$^{ac}$XZC, where X and Z represent all combinations of 19 amino acids (cysteine omitted) for a total of 361 peptides. We synthesized another 361 membered unmodified histidine peptide array with the sequence Ac-GXZHGC, referred to as the H-array. The constant amino acids (everything except X or Z) are referred to as the outside amino acids. Peptides were synthesized using standard solid phase peptide synthesis on Fmoc-Rink Amide MBHA resin purchased from Anaspec. Fmoc-protected amino acids were purchased from either Anaspec or Sigma-Aldrich. The Fmoc-Rink Amide resin was swelled in dimethylformamide (DMF) for 30 min and treated with 20% piperidine in DMF for 20 minutes to remove the Fmoc protecting

group. The first Fmoc-protected amino acid was coupled to the resin with pybop and N-methylmorpholine at a 4:4:8 ratio, which was repeated until all the amino acids were coupled to the resin. Once the Fmoc protecting group was removed from the final amino acid, the resin was treated with 10% acetic anhydride in DMF for 30 minutes to acetylate the N-terminus. The peptide was cleaved from the resin with a solution of 95% tri-floroacetic acid (TFA), 2.5% triethylsilane, and 2.5% milli-q water for 2 hours. To remove the resin, the solution was filtered and precipitated with peptides with ethyl ether. The peptides were re-suspended in 0.1% TFA, lyophilized and re-suspended in 0.1% TFA again. The peptides are neutralized by dilution into 50 mM Tris buffer pH 7.5 before im-mobilization.

### 2.3.2. Preparing peptide arrays.

Peptide arrays were prepared as described previously.[28 30] Briefly, steel plates were evaporated with 384 gold spots. The plates were soaked in an alkanethiolate solution that self-assembles onto the gold surfaces. The alkanethiolate monolayers presented a functional maleimide group against a background of tri(ethylene glycol). Peptides were transferred onto the gold spots using Tecan robotics and incubated at room temperature for 1 hour for immobilization. Peptide immobilization occurs through conjugate addition of the thiol on the terminal cysteine residue to the maleimide.

### 2.3.3. SAMDI Mass Spectrometry.

The SAMDI peptide array plates were coated with a 10 mg/mL 2',4',6'- Trihydroxy-acetophenone MALDI matrix in acetonitrile. Each immobilized peptide was analyzed in the reflector positive mode with 900 shots on an AB Sciex TOF/TOF 5800 MALDI mass spectrometer.

### 2.3.4. Statistical testing to identify amino acids associated with high or low S/N.

The S/N of all peptides were calculated by dividing the integrated product (area under the curve) of the expected peptide peaks (the signal) by the integrated product of a region in the spectrum devoid of peaks (the noise). The S/N for each peptide were sorted and ranked from lowest to highest. The S/N increase for consecutive peptides was calculated, and the low region boundary was defined as when a large change in S/N increase occurs. Similarly, a high S/N region was identified with the same process. This method allows different sizes for low and high regions. Amino acid enrichment in either region was determined using the Fischer exact test, which calculates the probability to observe at least as many amino acids in the region. Since there were 19 amino acids, the significance threshold was determined by a Bonferroni corrected p-value cutoff of $10^{-4}$; all reported p-values define the likelihood that the observed number of amino acids is within the low or high regions by random chance.

## 2.4. Results and Discussion

### 2.4.1. Experimental design.

We calculated peptide S/N using SAMDI mass spectrometry in two peptide arrays: Ac-GRK$^{ac}$XZC (K-array) and Ac-GXZHGC (H-array) where X and Z represent all combinations of 19 amino acids (cysteine omitted) for a total of 384 peptides in each array. To investigate the relationships between specific amino acids and S/N, we conducted statistical tests and machine learning. We applied the corresponding results to a published peptide array data set to reveal how S/N information can inform and serve as a guide for experimental design and analysis. In doing so, we discovered specific amino acid interactions that can explain observed S/N-amino acid relations. Through subsequent machine learning analysis, we identified physical properties and amino acid positions that predict the peptide's observed S/N. Finally, we used our machine learning model to predict the S/N of an unknown array and a partially synthesized array.

### 2.4.2. Preparation of peptide arrays.

We used solid-phase peptide synthesis to synthesize two peptide libraries containing terminal-cysteine residues, Ac-GRK$^{ac}$XZC (K-array) and Ac-GXZHGC (H-array), where X and Z represent all amino acids except cysteine for a total of 361 peptides in each array. Steel plates with 384 gold spots were soaked in a solution of disulfides as described earlier.[24] The monolayers self-assembled onto the gold surfaces and

presented a functional maleimide group allowing for the immobilization of thiol-containing molecules. We treated each monolayer surface with a unique peptide, which was immobilized to the surface through the side-chain thiol of the terminal cysteine residues. Eleven identical arrays were printed for the experiments that follow.

### 2.4.3. S/N is attributable to single amino acids in the K-array.

Comprehensive analysis of the K-array revealed general trends of single amino acids in a peptide on the observed S/N for that peptide. We collected spectra for each immobilized peptide on an AB Sciex 5800 MALDI mass spectrometer using reflector positive mode. Noise was quantified as the area under the curve (AUC) of the mass spectrum in a region devoid of signals, and the peptide signal was quantified as AUC of the expected peptide-terminated alkanethiol mass minus the noise AUC. Finally, we calculated S/N as the peptide's signal AUC divided by the noise AUC and calculated the mean for each peptide over the eleven plates in each array.

We used the Fischer exact test (Bonferroni corrected $p<10^{-4}$) to determine whether peptides with low or high S/N were enriched with specific amino acids. The corresponding p-values reflect the probability that the observed number of amino acids is within either the low or high S/N regions by random chance (gray regions in Figure 2.2A). All p-values are reported in Supplementary Figs. A.1 and A.2. We found enrichment of peptides with X-position tryptophan and leucine in the low S/N region and enrichment of peptides with Z-position glycine in the high S/N region

(Fig. 2.2B). This result suggests that single amino acids can have a strong effect on a peptide's detectability in MALDI-MS. The exceptionally low S/N of tryptophan and leucine-containing peptides suggest that their S/N-lowering effect is particularly strong, further suggesting that future K-arrays can disregard tryptophan and leucine while favoring glycine.

Peptides in the K-array display a wide range of S/N-from 3.8 to 313.7 (S/N is unitless)-demonstrating a wide range of poorly-detectable to detectable peptides (Fig. 2.2A). Combined with the statistical tests, this result suggests that poorly-detectable peptides can be predicted by their sequence. This observation may explain differences in MS-detectable peptide fragments after protein digestion.[44,45] In MS-based proteomics experiments, proteins are commonly digested and the fragments are detected using mass spectrometry. It is rare for complete detection of all peptide sequences digestion[46,47], and incorporation of known poorly-detectable peptide information could increase confidence of protein observation. As we demonstrate, characterization of a MALDI-MS experimental pipeline with known peptide sequences can inform subsequent protein quantification experiments.

### 2.4.4. A machine learning model predicts SAMDI-MS S/N as a function of amino acid sequence.

We developed a machine learning model to predict the S/N of peptide-terminated alkanethiolates in the SAMDI spectrum based on amino acid sequence with high

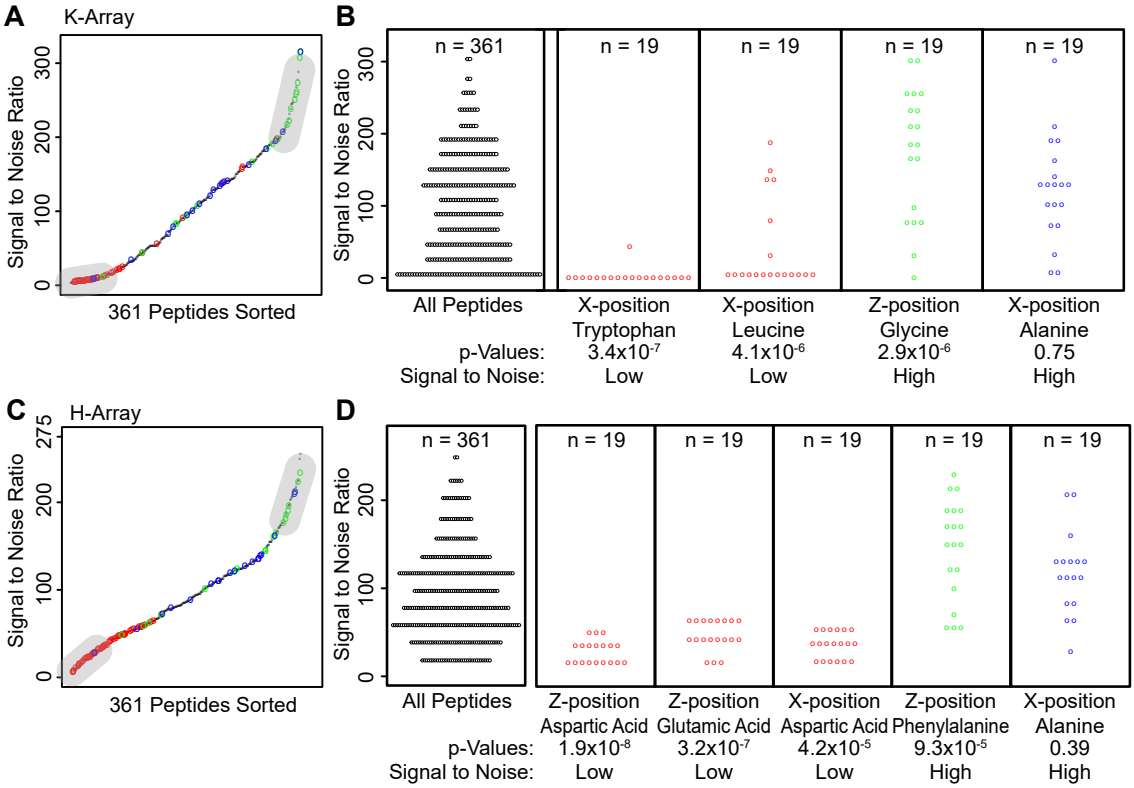



Figure 2.2. **Low Peptide S/N is observed in peptides containing tryptophan & leucine, and aspartic acid & glutamic acid, in K- and H-arrays, respectively.** Peptide S/N was averaged over 11 control plates. (A) Peptides in the K-array were sorted according to S/N. Low/high S/N regions are identified (See Methods). S/N ranges from 3.8 to 313.7, demonstrating that peptides vary greatly in S/N. (B) Amino acids found in the low/high regions were found to be statistically significant (Bonferroni corrected $p<10^{-4}$) using a Fischer exact test. The reported p-value is the chance the observed number of amino acids is within the low or high region by random chance. Peptides with X-position tryptophan and leucine have statistically low S/N, and peptides with Z-position glycine have statistically high S/N. Peptides that have X-position alanine are not statistically significant and are representative of other amino acids. (C) and (D) describe the same methods for the H-array. S/N has a similarly large range for both arrays, but the differences in amino acids observations suggest that dissimilar mechanisms are responsible for S/N.

accuracy suggesting that amino acid composition drives S/N observations in a predictive manner. We trained a random forest[48] machine learning model to predict S/N based on the hydrophilic, steric, and electronic physical properties of amino acids.[49] The training data contained 361 peptides (rows) and 39 associated physical properties for each of the X- or Z-position amino acids (resulting in 78 columns). The response vector, or predicted variable, defines the mean S/N from the 11 control plates. We used cross-validation, where a data sample (randomly selected rows) is left out for model testing, to calculate the predictive power $Q^2$ statistic[14]:

$$Q^2 = 1 - \frac{\sum_i^n (y_i - \hat{y}_i)^2}{\sum_i^n (y_i - \bar{y}_{train})^2} \tag{2.1}$$

In this formulation, $y_i$ is the true S/N for the left-out test peptide $i$, $\hat{y}_i$ is the predicted S/N of the test peptide, $\bar{y}_{train}$ is the sample mean of S/N in the training set, and $n$ is the number of cross-validated test peptides. The $Q^2$ statistic can take on values from $-\infty$ to 1, where 1 represents perfect prediction and 0 is equivalent to random performance. We create an explicit null model for each case by randomizing the data values prior to model training; the average $Q^2$ value of the null case was about 0.

The K-array analysis resulted in a $Q^2$ of 0.59 using both X- and Z-positions and all 39 amino acid physical properties. The high $Q^2$ value confirms our hypothesis that S/N values can be reliably predicted from amino acid sequences. This performance further suggests that S/N of new amino acids, such as non-natural amino

acids, can be predicted using their known physical properties. Together these results strongly indicate that amino acid sequence influences S/N in MALDI-MS. However, the inability to acquire a $Q^2$ value closer to 1 suggests that hidden variables-such as chemical interactions with amino acids outside the X- and Z-positions-play an important role in the overall response. These interactions are challenging to take into account, as they cannot be characterized with physical properties alone.

### 2.4.5. A bubble chart illustrates the S/N as an experimental design parameter.

As a measure of data quality, the S/N becomes another experimental design parameter. When studying enzyme activity on SAMDI peptide arrays, we measure the extent of peptide conversion with the enzyme.[8] Enzyme-treated peptides can be sorted into four categories: (i) high enzyme activity and high S/N, (ii) high enzyme activity and low S/N, (iii) low enzyme activity and high S/N, and (iv) low enzyme activity and low S/N. In the past, the SAMDI peptide array data was compiled into heat-maps that portrayed only enzymatic activity. We wanted to incorporate a metric into SAMDI array data output to differentiate between peptides that offer reliable and valuable information (category i) from those of lesser importance.

To this end, we include S/N information to complement a previously published experiment.[8] We construct a bubble chart where each peptide is represented by a circle, whose color represents the extent of peptide conversion to the product, and

whose size represents normalized S/N of the peptide before enzyme treatment. Previous approaches that use a color-only heatmap give the impression that each data point is equally valid in an analysis of the array data. However, some of the peptides contribute information that is more reliable because they have smaller errors. Observed enzyme activity on a peptide does not always correlate to significance. By incorporating S/N in bubble size, we rule out low performance signals and focus the analysis on high S/N ones. We illustrate this approach by replotting the heatmaps from Kuo *et al.*[8] to include S/N (Fig. 2.3).

In Kuo *et al.*[8], the K-array was exposed to cell lysates, and endogenous deacetylase activity was quantified by measuring the fraction of deacetylated peptides with MALDI mass spectrometry.[8] Deacetylation activity was quantified as the AUC of the modified (deacetylated) peptide divided by the AUC of both modified and unmodified peptides. AUC of each peptide is the sum of the three background-subtracted ion peaks in MALDI-MS: H+, Na+, and K+.

This new analysis revealed additional insights into the previous data. Peptides containing amino acids tryptophan, leucine, arginine, methionine, and lysine reflect low S/N, suggesting that their activity profiles are less useful. Conversely, peptides containing proline, glutamic acid, and glycine reflect high S/N, suggesting that their activity pro-files are more useful. Peptides containing leucine exhibit low S/N exclusively in the X-position, demonstrating that certain amino acids can have positional effects on S/N. Though amino acid presence can largely explain a peptide's S/N, we also find that some peptides have inexplicably low S/N—such as KAA, KIT,

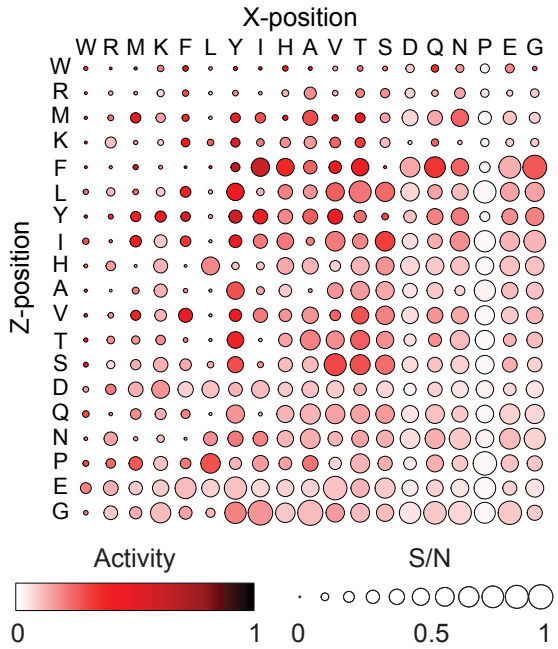



Figure 2.3. **Heatmap of cell lysate deacetylation activity and S/N highlights trustworthy peptides.** Bubble color is based on deacetylase activities from Kuo *et al.*[8] for lysate treated ac-GRK$^{ac}$XZKC peptide arrays. Bubble area represents peptide S/N before lysate treatment, normalized by max S/N. Amino acids are sorted by their general trend in peptide S/N when in either X- or Z-position. Peptides containing tryptophan (W), arginine (R), methionine (M), and lysine (K) have consistently low S/N, regardless of position. This illustration emphasizes peptides that are both active in terms of enzymatic activity and reliable in terms of S/N. In contrast, the highest activity peptides (darkest in color) do not necessarily give the highest S/N (largest bubble).

and KIQ—despite general trends suggesting that these peptides should have high S/N. This peculiarity highlights the complexities in S/N and reinforces the utility of machine learning strategies to predict S/N, which can be a critical design factor for future arrays.

Testing for S/N does not supplant tightly controlled and validated peptide array experiments. Instead, we suggest that accounting for unknown influences that lower confidence of a signal's true value—such as peptide synthesis inefficiencies, side reactions, peptide loss from washing, or ionization efficiencies—can better guide experimental design and data analysis. These influences are especially complicated with peptide species, where it is not clear how different amino acid sequences affect S/N. Machine learning can easily account for such effects.

### 2.4.6. Low S/N peptides offer unrepresentative signals.

The experiments by Kuo *et al.* demonstrated that low S/N peptides have higher variance across replicates. The same K-array measurements were carried out on two time points and across three different cellular conditions. We compared the variance in replicates of peptides in the top 20% of S/N to those in the bottom 20%. A one-sided F-test verified that the top 20% peptides have lower replicate variance than the lower 20% across all three conditions and across both time points ($p<10^{-10}$ for all cases). This finding suggests that peptides with low S/N have unrepresentative (or possibly random) signals, and they should be weighted less during analysis to avoid misleading conclusions. To investigate further, we calculated the standard deviation

of deacetylase activity on each peptide and plotted it against S/N (Supplementary Fig. A.3A); peptides with lower S/N have a higher variance in deacetylase activity. This trend was consistent for all days and for all experimental conditions, with a high anti-correlation coefficient (ranging from -0.814 to -0.975, Supplementary Fig. A.3B), demonstrating that peptides with low S/N can give unrepresentative measurements.

### 2.4.7. S/N is attributable to single amino acids in the H-array.

We investigate the H-array, Ac-GXZHGC, to analyze the generalizability of our findings; that is, do S/N characteristics of peptides from the K-array also apply to other peptide arrays? Similar to the K-array, the H-array has a wide range of S/N values, ranging from 5.5 to 255 (Fig. 2.2C and 2.2D), reinforcing the fact that peptides span a wide range of non-detectable to detectable signals in MALDI. The statistically low S/N peptides contain aspartic acid and/or glutamic acid, suggesting that their synthesis may be unnecessary in future experiments. Peptides with phenylalanine have statistically high S/N values suggesting that additional phenylalanine may improve peptide signals in the H-array.

### 2.4.8. Context matters: S/N characteristics are inconsistent between the K-array and H-array.

In addition to the variable composition of amino acids, the surrounding amino acids (those not in the X- or Z-position) play a role. Within the same array, S/N appears consistent between positions (Supplementary Fig. A.4), suggesting that S/N remains

largely unchanged when amino acid substitutions are made in the X- or Z-positions. However, peptides that had the lowest S/N in the K-array contained tryptophan, arginine and methionine; those with the lowest S/N in the H array had aspartic acid and glutamic acid. This disparity demonstrates that S/N characteristics in one array can be contextual and are not always consistent with a different array (Fig. 2.4). This observation suggests that the outside amino acids—arginine and lysine in the K-array and histidine in the H-array—strongly influence SAMDI peptide detection. In other words, the amino acid context around X- and Z-positions influences overall peptide detection, and partial knowledge of amino acid sequence is insufficient in understanding S/N values.

### 2.4.9. Physical interactions help inform S/N differences.

S/N differences can arise from a variety of sources, including synthesis inefficiencies, side reactions, and poor MALDI-MS ionization. Peptides with both methionine and tryptophan have low S/N in the K-array, and both have shown sensitivity to oxidation[41,42], sequestering the relevant peaks and lowering signal strength. However, Lee and coworkers demonstrated greater oxidation of histidine than either methionine or tryptophan[50], and Stafford and coworkers reported similar findings in oxidation of histidine in peptide arrays[41]. Their results are contrary to our lack of observed histidine oxidation (or low S/N) in either array, which remains unexplained.

In contrast to the K-array where methionine and tryptophan associate with the lowest S/N, glutamic acid and aspartic acid have the lowest S/N in the H-array

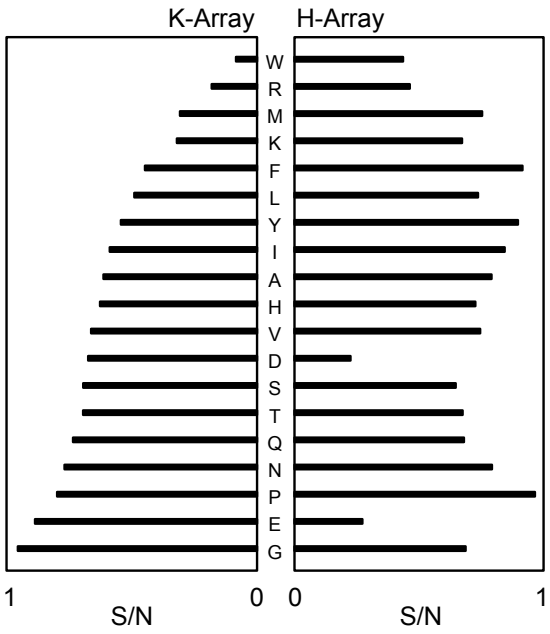



Figure 2.4. **Amino acid influence is context dependent.** Amino acids are sorted by their correlated mean peptide S/N in the K-array when in either X- or Z-positions. Bars represent S/N normalized by the highest value within each array. The H-array shows little agreement, suggesting that the surrounding amino acids strongly influence S/N. The two negatively charged amino acids, aspartic and glutamic acid, have the largest difference between the two arrays, suggesting a relation between charge and S/N, but only within the H-array.

(Fig. 2.4). Lysine has strongly favorable hydrogen bonding energies[51], and when in close vicinity of methionine and tryptophan, hydrogen bonding could catalyze oxidation.[51 52] Tryptophan-containing peptides have statistically low S/N specifically when in the X-position (Fig. 2.2), which is directly adjacent to the lysine and further supports this hypothesis. If hydrogen bonding stabilization is required for methionine or tryptophan oxidation, then the presence of carboxylic acid groups on acidic amino acids may be unfavorable for oxidation. To explore this concept further, we compared methionine or tryptophan containing K-array peptides with either glutamic or aspartic acid to those without glutamic or aspartic acid. We applied a Mann-Whitney U test and found that peptides with one of methionine or tryptophan and one of glutamic or aspartic acid had higher S/N values ($p=0.0050$) than peptides with methionine or tryptophan without either glutamic or aspartic acid, maybe indicating that the two acidic amino acids protect against methionine and tryptophan oxidation.

High S/N peptides in the K-array commonly contain hydrophilic amino acids such as glutamic acid, asparagine, and glutamine, potentially due to more efficient crystallization within the matrix. This finding is in agreement with a report by Fenselau and coworkers, where hydrophilic proteins were preferentially detected in MALDI-MS due to differences in the co-crystallization[35]. However, the H-array has high S/N associated with hydrophobic amino acids: proline, tyrosine, phenylalanine, and isoleucine. The divergence in S/N of hydrophobic and hydrophilic amino acids suggests that mechanisms leading to high S/N are different between the two arrays.

A two sided Mann-Whitney U test (Bonferroni corrected $p<2.6x10^{-3}$) reveals peptides that contain eight amino acids in the X- or Z-position that have statistically different S/N values between the two arrays: glutamic acid, tryptophan, aspartic acid, methionine, arginine, lysine, glycine, and phenylalanine. This test directly compares differences between the two arrays rather than within the array, which has resulted in more amino acids than the Fischer exact test in Figure 2.2. Only arginine and phenylalanine differ from Figure 2.2, and both amino acids have lower S/N in the K-array. This result contrasts with those of Krause and coworkers where peptides with higher numbers of arginine or phenylalanine typically contributed to higher MALDI signal strength[39] (the K-array has an additional arginine). The unusual observation may be due to an unknown interaction with other outside amino acids, indicating that peptide S/N should be tested for each peptide array.

### 2.4.10. Machine learning performance across positions and physical properties help explain S/N observations.

We trained random forest models with individual physical properties to assess the impact each property has on S/N. Highly predictive properties (namely those with highest $Q^2$) suggest that the associated physical property is highly relevant and predictive of SAMDI-MS S/N. In addition, we independently evaluated the X- and Z-positions to see if one position reflected more predictive power. Positional differences suggest that the amino acid position, and not merely composition, influences the predictive power of our machine learning model.

Using both X- and Z-positions and all physical properties, the K-array and H-array had a $Q^2$ of 0.59 and 0.61, respectively (Fig. 2.5). The similar $Q^2$ values suggest that the models reached an upper limit to predictive performance from amino acid sequence. Predictions based on the amino acids in both X- and Z- positions consistently performed better than predictions based solely on one position: $Q^2$=0.22 and 0.20 for the X- and Z-positions in the K-array, respectively, and $Q^2$=0.16 and 0.28 in the X- and Z-positions of the H-array, respectively. As expected, more complete amino acid information results in better prediction. However, the higher $Q^2$ for the Z-position in the H-array suggests that positions can have varied influence on S/N. This observation suggests that the Z-position interacts with the histidine to change S/N detection in MALDI-MS more strongly than the X-position. In addition, the highest single property $Q^2$ values—0.57 and 0.54 in the K- and H-array, respectively—are close to the $Q^2$ value of all properties. This observation indicates that few properties are necessary to predict S/N and that many physical properties are redundant or uninformative.

In terms of physical properties (39 total), we do find both consistent and inconsistent trends for the two arrays. Electronic properties (15 total) tend to be less predictive for both arrays than steric or hydrophilic properties (Supplementary Fig. A.5). Steric properties (16 total) and hydrophilic properties (8 total) are especially highly predictive in the K- and H-arrays, respectively. Hydrophilic properties are highly predictive in the H-array potentially due to the hydrophilicity of glutamic

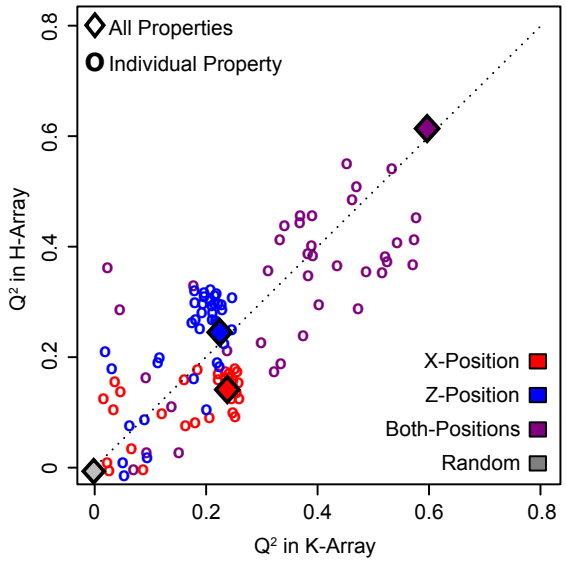



Figure 2.5. **Peptide S/N is predicted as a function of amino acid properties.** Peptide S/N was predicted using a random forest machine learning model based on 39 amino acid physical properties, shown in diamonds, of the amino acid in either the X-position (red), Z-position (blue), or both (purple). Also, models were fit on individual properties to identify their predictive power, shown in circles. Random forest models contained 1000 trees and the predictive power, quantified by the $Q^2$ metric, was calculated based on 5-fold cross-validation. All $Q^2$ values are listed in Supplementary Tables 1 and 2. Consistent for both peptide ar-rays, the highest $Q^2$ values were attained when using both positions with all physical properties (purple diamond). Z-position $Q^2$ values (blue) are higher in the H-array, which suggests that positions have varied predictive power on S/N. In addition, most properties lie near the diagonal indicating that they have similar predictive power between peptide arrays; the amino acid disagreement in Figure 2.4, however, suggests that those properties are predictive for different reasons.

and aspartic acid and their association with low S/N exclusively in the H-array. Similarly, hydrophobic amino acids like proline, tyrosine, phenylalanine, and isoleucine tend to have high S/N. This alignment explains why hydrophilic properties are predictive in the H-array. However, it is unclear why electronic properties are relatively less predictive while steric properties are more predictive in the K-array.

Despite these differences, physical properties are similarly predictive between the K-array and H-array (Fig. 2.5), as evident in their closeness to the diagonal. That is, a predictive or non-predictive property remains largely the same between arrays, but there still exist small differences between the performance of steric and hydrophilic properties between arrays. This relation demonstrates that the same properties govern S/N observations, but because single amino acids differ in S/N characteristics (Fig. 2.4), these results altogether suggest that S/N values manifest from different mechanisms between the arrays. These different mechanisms are likely a direct result of context differences, specifically relating to the outside amino acids.

Machine learning cannot predict S/N on completely unknown peptide arrays. We trained various machine learning models on the K-peptide array and tested them on the H-array, and vice versa, to assess the feasibility of predicting S/N on a *de novo* peptide array. We trained models for every positional combination to interrogate exhaustively the entire space. For example, we trained a model on X-position data in the K-array, then tested on the Z-position in the H-array, and we continued with all combinations of positions. We also trained several types of models based on random forest[48], deep learning[53], nearest neighbor regression[54], and partial least squares

regression.[55] The models had the following model-specific parameters: random forest had 1000 trees, deep learning consisted of two layers of 200 nodes with feed-forward connections; nearest neighbors regression used 10 neighbors; and partial least squares regression used one component, or loading vector.

All models trained on the K-array failed to predict S/N in the H-array, and vice versa ($Q^2$<0.1). This failure is attributable to S/N disagreement between peptide arrays for each amino acid (Fig. 2.4), which arises from the unique outside amino acids in the two arrays (GRK$^{ac}$XZC and GXZHGC). This finding reinforces the idea that context matters: interactions with outside amino acids influences S/N.

### 2.4.11. Only 1/3 of peptides in an array are required for machine learning model prediction of peptide S/N.

We investigated the minimum number of peptides in an array needed to train a model that could accurately predict S/N of the full array. We simulated a partially synthesized array by randomly selecting training peptides to predict S/N of the non-selected peptides. The number of training peptides ranged from 5 to 350, and each training size contained 200 repetitions of selecting random peptides. We trained a random forest model with all 39 physical properties in both amino acid positions. We identified the point of diminishing returns, which balances minimum training size with maximum predictive power, by normalizing the number of training peptides and finding the sample size closest to training size 1 and $Q^2$=1. The point

of diminishing returns was found to be 87 and 111 peptides for the K-array and H-array, respectively, both of which had a $Q^2$=0.48 (shown with arrows in Figure 2.6). This result shows that we can partially screen future peptide arrays by synthesizing about 100 of the planned 361-sized array, or roughly one-third, reducing the use of resources and time. Though we cannot generalize this specific ratio to larger array sizes, these results suggest that only a fraction requires synthesis. This machine learning technique can prevent costly experimental screens and allow researchers to focus on predicted *a priori* high S/N peptides. The null model with randomized data performed consistently around $Q^2$=0.

There are significant variations in the intensities of peaks in SAMDI mass spectrometry that can arise from different peptide sequences. SAMDI analysis of peptide arrays demonstrates that peptide signals can have a wide range of S/N, where many of the peptides are nearly undetectable. We find that S/N is attributable to single amino acids, offering design choices to increase information content. However, the underlying basis of S/N is unclear and may be due to complex interactions among amino acids, matrix, crystallization, or ionization efficiencies. Additionally, we find that the two arrays used in this work exhibited different S/N values for different amino acids, demonstrating that the whole amino acid sequence can affect S/N values in MALDI-MS. Machine learning identified physical properties that predict S/N with high accuracy. Machine learning models can be trained on a fraction of the peptide sequences and still describe the full set of sequences, allowing early selection of high S/N peptides. Such computational models allow peptide experiments

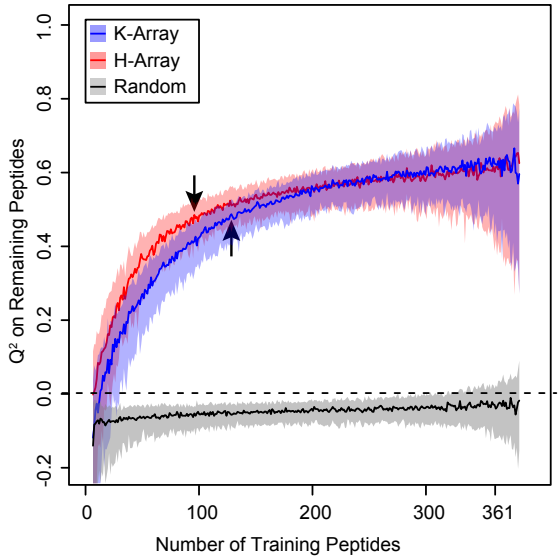



Figure 2.6. **Peptide array S/N can be predicted from a minimal peptide subsample.** A specified number of peptides were randomly selected for training to predict S/N of the remaining peptides using all physical properties of both X- and Z-position amino acids. Due to computational constraints, random forest was used with 100 trees for training set sizes from 5 to 350. The mean $Q^2$ and 80% confidence intervals are shown for 200 random training sets. For both peptide arrays, predictive power increases with training size and levels out around 100 peptides. The optimal tradeoff was identified by normalizing the number of training peptides and finding the sample size closest to training size 1 and $Q^2$=1. The tradeoff is shown with arrows: 87 training peptides for K-array and 111 for H-array, which demonstrates that machine learning can predict S/N for future peptide arrays, avoiding costly experiments that screen for high S/N peptides. A randomized dataset performed consistently around $Q^2$=0.

with high S/N arrays without costly screens or ineffective peptide library syntheses. Additionally, accounting for S/N as a design choice can prevent inaccurate results drawn from poor peptide measurements.

This work significantly improves and simplifies high-throughput data analysis by factoring in data quality. The statistical and machine learning methods presented here allow us to discover the most valuable information from peptide arrays and plan future experiments with more confidence. As demonstrated, these methods can inform the design of new peptide arrays using a small set of peptides. The presented methodology and applications of S/N are adapted to maximize the information learned from peptide array experiments and can improve peptide design across a wide range of technologies.

CHAPTER 3

# Addressing Nanomedicine Complexity with Novel High-Throughput Screening and Machine Learning

This work is in collaboration with Gokay Yamankurt, Eric Berns, Milan Mrksich, and Chad Mrkin and is currently under review.

## 3.1. Abstract

A tiny fraction of the nanomedicine design space has been explored, largely due to the complexities of such structures and the lack of high-throughput methods to make and analyze them. To address this challenge, we studied spherical nucleic acids (SNAs) as a first example because they have at least 11 unique parameters that can be systematically and independently varied to optimize performance in the context of immune cell activation. We defined reasonable ranges of these parameters and identified approximately 1000 therapeutic candidate structures that are qualitatively similar but could have significant differences in activity, thereby creating both a synthesis and an analysis challenge. To address this challenge, we developed a high-throughput method for making such structures at the picomole scale in a 384-well format and utilized a self-assembled monolayers for matrix-assisted laser desorption ionization (SAMDI) mass spectrometry assay to rapidly measure innate

immune system activation by quantitatively determining NF-$\kappa$B activity. Traditionally, cell-based optical assays are used for measuring these activities, but they suffer from artifacts due to the absorption and scattering of light associated with nanostructures. Using this novel methodology, we identified structure-activity relationships for immune activation of SNAs, and new design rules for SNA-based cancer vaccine candidates. Finally, we utilized machine learning to quantitatively model the immune activation of SNAs, and applied it to identify the minimum number of SNAs needed to capture optimum structure-activity relationships for a given library. By doing so, one can reduce the number of nanoparticles that need to be tested by an order of magnitude, and still obtain the same information as from screening the entire library. Importantly, these insights and techniques can be generalized to include many other types of nanomedicines and provide a next generation screening tool for therapeutic development.

## 3.2. Introduction

Nanotechnology is beginning to play a major role in developing new therapeutic modalities. Currently, over 100 drugs based upon nanomaterials are in clinical trials or approved for therapeutic use.[56] These structures are promising because of their multifunctionality, which directly relates to their relatively large size and often complex architectures when compared with conventional small molecules or biologics. However, due to this complexity, little attention has been paid as to how structural changes inform biological activity. Consider, for example, spherical nucleic acids (SNAs), which are made by chemically arranging short sequences of DNA or RNA

around a nanoparticle core (Fig. 3.1a).[57,58] SNAs exhibit properties that are substantively different from the short, linear oligonucleotides that comprise them, including the ability to actively cross mammalian cell membranes without the need for transfection reagents, a resistance to nuclease degradation, and the ability to carry large and complex cargo (such as oligonucleotides and peptides) into many cell types.[59–62]

These properties make SNAs an attractive candidate in cancer immunotherapy, as structures with dual functionality can be rapidly prepared from lipids, oligonucleotide adjuvants, and peptide antigens. When delivered to antigen presenting cells (APCs), SNAs activate the immune system and, in a lymphoma model, show superior activity compared to the same free antigen and linear oligonucleotides.[60] However, the modularity of an SNA allows for a large number of possible designs, and identifying the nanoparticle architectures best for inducing multiple aspects of cellular immune responses, such as potency, selectivity, and efficacy, remains a challenge. Herein, we describe a new approach for synthesizing a library of SNAs that are qualitatively similar but structurally distinct, in conjunction with a mass spectrometry-based screening protocol that can rapidly and quantitatively determine the ability of an SNA structure to activate the TLR9 pathway. In this first example, we show how this methodology can be used to make and screen 1000 (800 of which are unique) SNA architectures. In addition, we describe how machine learning models can be trained with this data, and subsequently used to accurately predict the TLR9 stimulatory activity of SNAs based on structural features. Significantly, these models provide a ranking of the order of importance of eleven structural parameters as well

as SNA drug concentration. Collectively, these insights have important implications in designing SNA-based therapeutics. Additionally, since this methodology can be extended to other nanotherapeutics, this work points towards a new way of designing and optimizing nanomedicines for a wide variety of uses.

## 3.3. Methods

### 3.3.1. The Modular Design of Spherical Nucleic Acids

Immunostimulatory SNAs consist of three modular components: the nanoparticle core, the oligonucleotide shell, and the peptide antigen, each of which can be arranged in a variety of configurations.[60] To establish an appropriate library for high-throughput evaluation, we focused on eleven properties across these components (Fig 3.1b). We used 1,2-dioleoyl-*sn*-glycero-3-phosphocholine (DOPC) and 1,2-dioleoyl-*sn*-glycero-3-phosphoethanolamine (DOPE) to form liposomes that are biocompatible, straightforward to synthesize, and capable of encapsulating the antigen.[63] We focused on two liposome core sizes with average diameters of 70 and 100 nm that were made from DOPC or a mixture of 80% DOPC and 20% DOPE. The size of the SNA can influence its rate of cellular uptake, and inclusion of DOPE in the liposomes is believed to affect the peptide release rate and their endosomal escape, which is important for peptide processing.[64,65]

The oligonucleotide shell serves two roles. It facilitates cellular uptake and serves as the adjuvant, which activates the innate immune system in a sequence-specific manner.[60] The oligonucleotides used in the design of SNAs in the library varied in

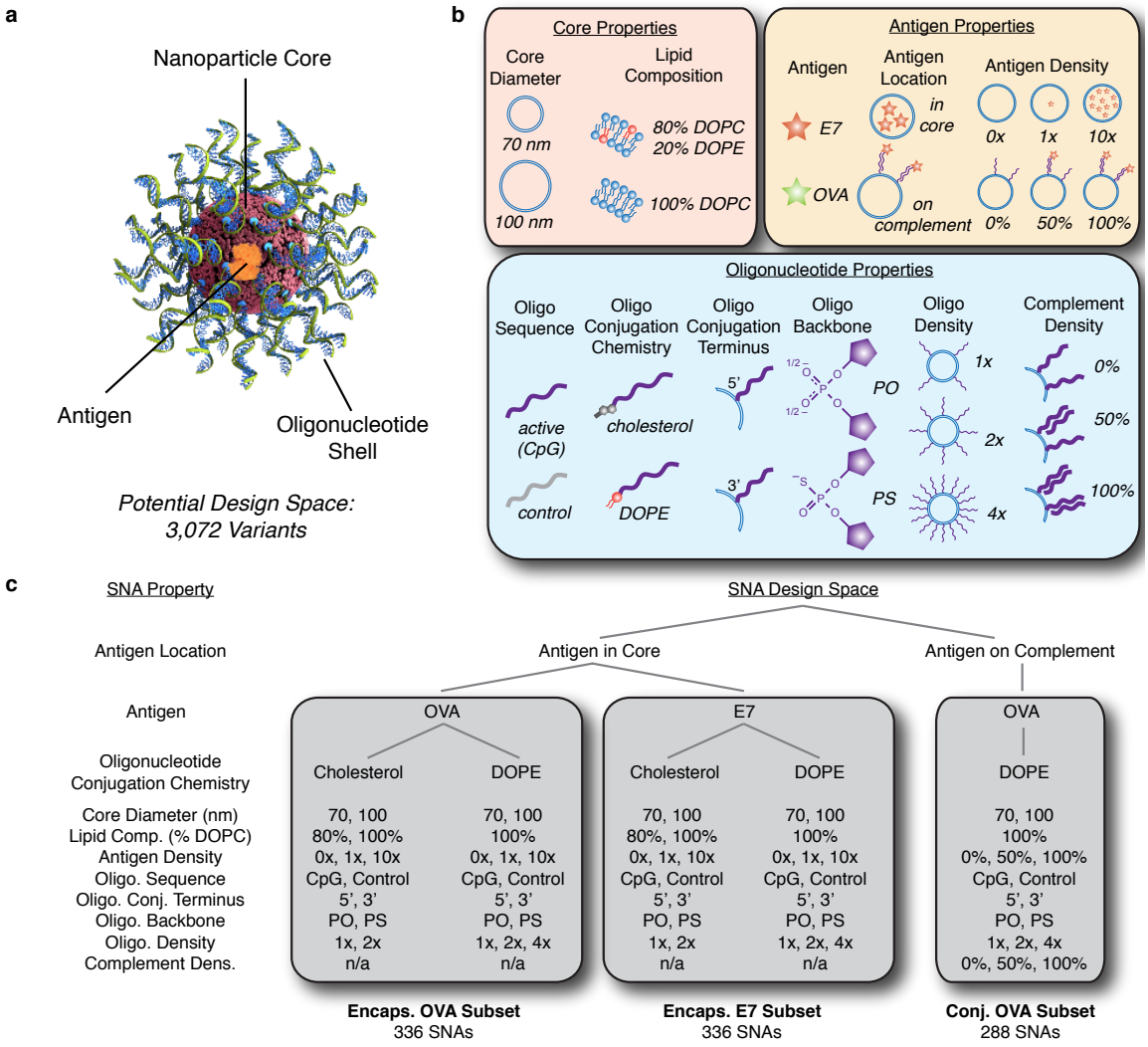



Figure 3.1. **Overview of synthesized SNAs.** **a**, The three components of immunostimulatory spherical nucleic acids (SNAs). **b**, The parameters investigated for each of the SNA design properties, organized by core, antigen, and oligonucleotide property categories. **c**, The total design space investigated in this study, divided into three subsets.

five ways: sequence, backbone chemistry, conjugation chemistry to the liposome, site of lipid functionalization, and surface density of presentation by SNAs (albeit over a narrow range). We chose a CpG DNA oligonucleotide (ODN1826) known to activate mouse Toll-like receptor 9 (TLR9), as well as an inactive control where the CpG motif is inverted to GpC.[66,67] TLR9 is an endosomal protein that recognizes unmethylated CpG oligonucleotides associated with bacteria and viruses.[68] To explore the importance of backbone composition, we synthesized linear oligonucleotides with either phosphodiester (PO) or phosphorothioate (PS) backbones since PS oligonucleotides are known to induce higher immune activation, but SNAs comprised of PO backbones present activities comparable to PS structures.[60,69] We evaluated distinct strategies for conjugating oligonucleotides to the nanoparticles by preparing structures with cholesterol or DOPE, both of which insert into the liposomal cores and can be chemically attached to the 3'- or 5' ends of the oligonucleotides. Finally, since oligonucleotide density is known to influence cellular uptake and protein binding of SNAs, we evaluated the oligonucleotide surface density at 0.5, 1, and 2 pmol/cm$^2$ (referred to as 1x, 2x and 4x, respectively).[70,71] The 4x structure represents the upper-limit of what is synthetically viable via our high-throughput procedures, at present.

As our test case, we chose the $OVA_{257\text{-}264}$ peptide from ovalbumin, a well-studied model antigen. Since peptide properties can vary dramatically with amino acid

composition, we also tested a peptide antigen from the E7 protein of human papillomavirus.[72] To study how the release rate of the antigen influences NF-$\kappa$B activation, we evaluated SNAs wherein the antigen was either encapsulated within the SNA architecture or hybridized to the oligonucleotide shell through a complementary oligonucleotide. As a control, we investigated how the addition of a complement affects TLR9 stimulation.

We synthesized and tested three subsets of SNAs (OVA encapsulated SNAs, E7 encapsulated SNAs, and surface-presented OVA SNAs) representing the key possible combinations of the parameters, with a few synthesis-limited exceptions noted below regarding lipid composition, oligonucleotide surface density, and surface conjugated peptide antigen (Fig. 3.1c). Variation across the eleven structural features—spanning the nanoparticle core, the oligonucleotide chemistry, the surface presentation of oligonucleotides, and the incorporation of antigen—led to the design of a library with 960 total SNAs, 800 of which are unique.

### 3.3.2. High-throughput screening of SNA libraries

To enable the screening of SNA libraries, we developed a high-throughput assay for the rapid and quantitative measurement of cellular responses to the SNAs (Fig. 3.2a). We cultured RAW-Blue macrophages in 384-well plates and treated each well with a distinct SNA at four oligonucleotide concentrations between 1 nM and 1 $\mu$M (each separated by a factor of 10). RAW-Blue cells are engineered to secrete embryonic alkaline phosphatase (SEAP) upon activation of NF-$\kappa$B, a major transcription factor

that is activated by TLR9 signaling, as well as other signals, to regulate the immune response. We collected the culture media and determined the concentration of SEAP using SAMDI (Self-Assembled Monolayers for MALDI) mass spectrometry (MS), a label-free assay for high throughput, quantitative analysis of enzymatic activity.[73–76] SAMDI uses monolayers presenting a selective capture chemistry against a background of non-binding tri(ethylene glycol) groups to isolate substrates and products from a complex mixture.[76,77] Subsequent analysis of the monolayers by MALDI-MS quantitates the amount of substrate and product, which is a direct measure of the enzyme concentration (Fig. 3.2b and c). Here, we mixed the media containing SEAP with a phosphorylated peptide substrate, captured the substrate and dephosphorylated product on monolayers, and then analyzed the samples by SAMDI. We chose this platform for its ability to quantify enzyme activities at high throughput, without dependence with the common optical methods, which can be negatively affected by the light scattering and absorbance of the nanoparticles. These artifacts are difficult to correct because of their dependence on nanoparticle properties such as size, concentration, and aggregation. Furthermore, SAMDI is compatible with small sample volumes for analysis, thereby reducing the amount of SNAs, cells, and reagents necessary for evaluation, by 6-fold compared to the amounts used in optical assays.

With this assay, we measured the responses to 960 SNAs at four concentrations and with two biological replicates and acquired two SAMDI spectra for each sample. Along with standards and controls, more than 8,500 cell culture wells were used and more than 17,000 SAMDI spectra were analyzed. These data revealed many insights

into the importance of each structural feature, and how the combinations of features impact immune activation. Below, we highlight some of the most prominent trends.

## 3.4. Results and Discussion

### 3.4.1. SNAs induce higher immune activation than linear oligonucleotides

Varying the design parameters of SNAs induced a broad range of immune activation (Fig. 3.3a shows the encapsulated OVA subset with the active CpG oligonucleotide sequence, Supplementary Figure B.1 shows the encapsulated E7 subset). Almost all SNAs with the active oligonucleotide sequence outperformed the linear PO oligonucleotide. Additionally, many SNAs, including those with a PO backbone, were more potent than the linear oligonucleotide with the PS backbone.

### 3.4.2. Conjugation chemistry of oligonucleotide-liposome association significantly affects immune activation by SNAs

With eleven design parameters under investigation, we sought to identify the relative importance of design choices on immune activation. Multifactor analysis of variance (ANOVA) (Supplementary Table B.1) revealed that, unsurprisingly, oligonucleotide concentration and oligonucleotide sequence (i.e., active or control) heavily influenced activation. After sequence, the feature that had the greatest impact on immune activation was the lipid moiety conjugated to the oligonucleotide for liposome attachment. Cholesterol conjugation resulted in higher levels of immune activation than DOPE conjugation ($P=4.8\text{x}10^{-16}$). However, SNAs with cholesterol-conjugated

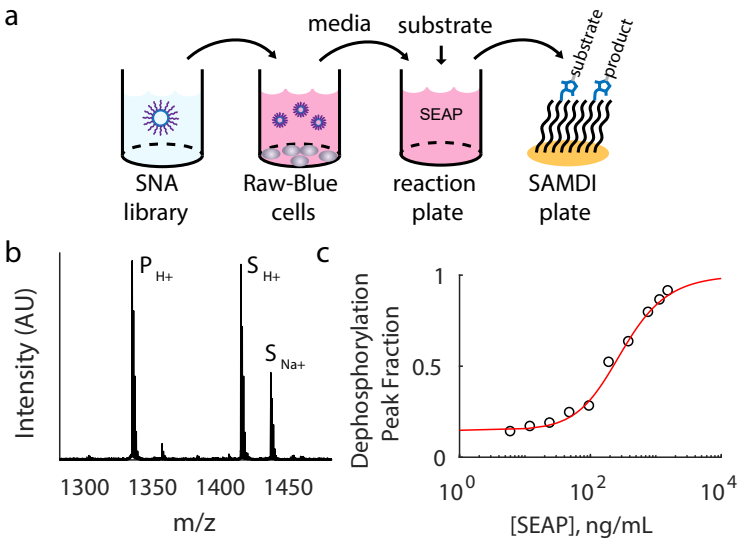



Figure 3.2. **Description of assay used to measure immune activation.** **a**, The assay used to evaluate the structure-activity relationships between SNA properties and TLR9 activation of APCs. Libraries of SNAs are incubated with RAW-Blue macrophages, engineered to secrete embryonic alkaline phosphatase (SEAP) into the media, in 384-well plates. After 16 hours, the media is transferred, processed, and mixed with a phosphorylated substrate. The solution is transferred to SAMDI plates with 1536-spot arrays of monolayers presenting maleimides to selectively capture the substrate and product by a maleimide-thiol reaction. **b**, An example SAMDI spectrum showing the immobilized substrate and product. Performing MALDI-MS on the self-assembled monolayers (ie. SAMDI) results in mass spectra containing quantitative information on the relative amounts of substrate and product (i.e. extent of dephosphorylation). **c**, An example standard curve used to convert the SAMDI spectral data for the library into SEAP concentration.

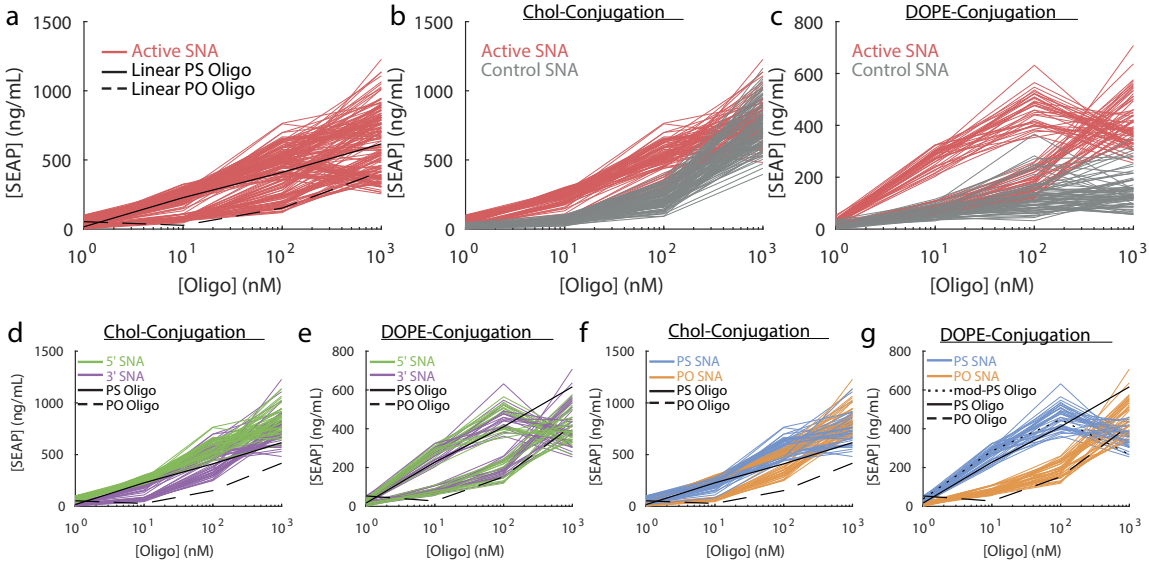



Figure 3.3. **Visualizing the relationship between SNA design and immune activation in the encapsulated OVA subset.** **a**, The SEAP concentrations observed for all active-sequence SNAs in the encapsulated OVA subset (all data in this figure is from this subset), compared to the PO and PS versions of linear oligonucleotides with the same active sequence. Comparison of SNAs with the active and control sequences, broken into the groups of SNAs with **b**, cholesterol-conjugated oligonucleotides and **c**, DOPE-conjugated oligonucleotides. **d & e**, Comparison of 5' and 3' conjugation termini of SNAs with active sequence, grouped by conjugation chemistry. **f & g**, Comparison of PO and PS backbones of SNAs with active sequence, grouped by conjugation chemistry.

oligonucleotides without CpG motifs also induced similarly high levels of activation at 1 $\mu$M oligonucleotide concentration (798 and 747 ng/mL SEAP for active and inactive, respectively—Fig. 3.3b), indicating a sequence-independent activation of TLR9. The linear oligonucleotide does not activate TLR9; therefore, these results indicate that these SNAs may activate NF-$\kappa$B through another mechanism. One possible explanation is that cholesterol groups delivered to cells on the SNA induces additional activation. Our cholesterol conjugation chemistry utilizes carbamates, which can be cleaved by esterases, including sterol *O*-acyltransferases.[78] Any potentially released cholesterol, which is known to activate the UPR pathway in macrophages, may also induce NF-$\kappa$B activation.[79]

In contrast, SNAs without CpG-containing oligonucleotides conjugated to DOPE (instead of cholesterol) lead to dramatically lower secretion of SEAP compared to their cholesterol-conjugated counterparts ($P<1x10^{-16}$, Fig. 3.3c). We conclude that DOPE conjugation provides a way to synthesize SNAs that trigger an innate immune response exclusively through activation of TLR9. However, the combination of TLR9 stimulation and non-specific activation by SNAs with cholesterol-conjugated oligonucleotides may be advantageous for inducing a greater overall immune response.

### 3.4.3. Conjugation terminus of the oligonucleotide influences the immune activation in a conjugation chemistry dependent manner

Because of the dominant effects of conjugation chemistry, we analyzed the remaining SNA properties separately for SNAs with cholesterol- and DOPE-conjugated

oligonucleotides. Interestingly, we observed differences in the preferred conjugation terminus when different conjugation chemistries were used (Fig. 3.3e and f). With cholesterol conjugation, 5' conjugated SNAs showed significantly higher activity than 3' conjugated SNAs (OVA subset: $P<2.2\text{x}10^{-16}$ for all concentrations; 566 and 439 ng/mL mean SEAP at 100 nM for 5' and 3' conjugation, respectively), however, DOPE-conjugated SNAs did not show a difference with conjugation terminus (OVA subset: P=1 for all concentrations; 324 and 330 ng/mL mean SEAP at 100 nM for 5' and 3' conjugation, respectively). Furthermore, conjugation from the 5' terminus did not lead to loss of immune activation for either conjugation chemistry, which contradicts reports that modifications at the 5' end inactivate the TLR9 activity of linear CpG oligonucleotides.[80,81]

### 3.4.4. Phosphorothioate oligonucleotide backbone increases immune activation compared to phosphodiester backbone

Similar to well-known trends with linear oligonucleotides, the oligonucleotide backbone also influenced the immunostimulatory activity of the SNAs (Table B.1 and Fig. 3.3g, h, and i).[69] SNAs with PS backbones generally outperformed their PO counterparts ($P=5\text{x}10^{-9}$ for DOPE and $P=2.7\text{x}10^{-4}$ for cholesterol-conjugated SNAs). However, a more pronounced dependence on oligonucleotide backbone was observed with DOPE-conjugated SNAs than with cholesterol-conjugated SNAs. For DOPE-conjugated SNAs, the mean SEAP concentrations were 191 and 463 ng/mL for PO

and PS backbones, respectively, whereas for cholesterol-conjugated SNAs, they were 431 and 573 ng/mL (all at 100 nM).

In contrast, at the highest concentration of 1 $\mu$M, SNAs with PO oligonucleotides outperformed their PS counterparts,. Notably, the activity induced by DOPE-conjugated SNAs with PS oligonucleotides consistently decreased when the oligonucleotide concentration increased from 100 nM to 1 $\mu$M. The DOPE-conjugated PS linear oligonucleotide, but not the PO backbone, showed a similar reduction in activity at 1 $\mu$M (Fig. 3.3h), suggesting that this behavior is due to the specific stimulatory properties of the DOPE-conjugated oligonucleotide.

These results lead us to conclude that DOPE-conjugated oligonucleotides with PS backbones provide an advantage if greater potency is desired. PS backbones have the added benefit of resistance to nuclease degradation *in vivo*.[82] However, these results also show that SNAs with oligonucleotides composed of PO backbones can achieve similar levels of activation when present at higher concentrations. While class B CpG oligonucleotides are less effective with PO backbones, using SNAs with PO oligonucleotides may be worth the loss in potency because of the reduction in toxicity and cost, since the SNA structure may provide sufficient resistance to nuclease activity.[83–85]

### 3.4.5. Oligonucleotide density on the surface of the nanoparticle has a small and variable impact on immune activation

Surprisingly, there was not a strong or consistent trend in how oligonucleotide density affected activity, with neither the highest or lowest densities showing the best activity. In previous studies, SNAs with higher oligonucleotide densities led to higher biological activity in cellular uptake and RNAse H mediated degradation of mRNA; however, the nanoparticle designs in those studies were limited to gold cores, and used different core sizes and oligonucleotide densities compared to this study.[70,71] From these observations, we conclude that the choice of oligonucleotide density for these constructs over this narrow density range should be based on other considerations, such as stability in vivo, which is inextricably linked to potency.

### 3.4.6. Core diameter and lipid composition influences the immune activation of SNAs in an encapsulated peptide-specific manner

In both encapsulated SNA subsets, the lipid composition generally did not have a significant impact on activity as determined by ANOVA (Table B.1), except for one particular context discussed below. Additionally, core diameter was not a significant parameter in the encapsulated OVA subset, whereas it had a significant impact with encapsulated E7 group.

Since all combinations of parameters evaluated were both with and without peptide, we were able to isolate the effects of peptide encapsulation by comparing pairs of SNAs with identical properties except for the amount of peptide encapsulation.

We subtracted the SEAP concentration of the SNA without peptide from the SNA with highest peptide concentration (Fig. 3.4a). This analysis revealed that core diameter and lipid composition were influential when E7, but not OVA, was encapsulated. Specifically, for the E7 subset, SNAs with 100 nm cores containing peptide induced higher levels of NF-$\kappa$B activation (P=5.7x10$^{-5}$), and the magnitude of this effect also depended on lipid composition (Fig. 3.4b). Within the subset of SNAs with cholesterol-conjugated oligonucleotides on 100 nm cores, the SNAs with 100% DOPC cores showed higher immune activation than 80% DOPC, 20% DOPE cores (P=0.0011) (Fig. 3.4b). We observed no dependence between the presence of antigen and immune activation when the antigen was OVA. (Fig. 3.4c).

These results clearly illustrate that peptide encapsulation can impact the ability of SNAs to activate TLR9 and reveal crosstalk between the molecular components of SNAs intended to induce innate or adaptive immunity. Unlike oligonucleotides, the physicochemical properties of peptides vary dramatically with sequence, which can affect their interaction with the rest of the SNA structure. For example, the differences in isoelectric points of the peptides, which are 5.7 and 8.8 for the E7 and OVA peptides, respectively, result in different net charges for the peptides, which could affect their interaction with the positively charged liposome core. We conclude that the interactions between liposomes and peptides must be taken into account when designing and evaluating nanomedicines, as they can lead to large shifts in the immune activation of SNAs, especially at high levels of peptide encapsulation.

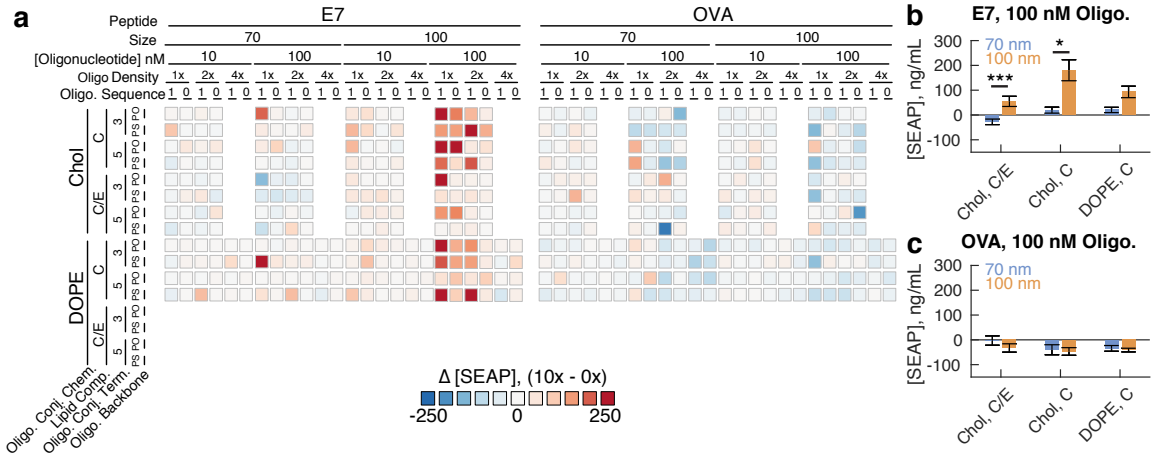



Figure 3.4. **Peptide encapsulation has a selective effect on immune activation. a**, The difference in SEAP concentration between each SNA with the highest peptide concentration (10x) and the corresponding SNA without any peptide (0x), for the two encapsulated peptide subsets. **b & c**, The average difference in SEAP concentration between SNAs with 10x and 0x E7 peptide concentration, grouped by core diameter, the combination of conjugation chemistry and lipid composition, for the E7 and OVA subsets, respectively. Chol, C/E indicates cholesterol conjugation and 80% DOPC/20% DOPE lipid composition. Chol, C indicates cholesterol conjugation and 100% DOPE lipid composition. DOPE, C indicates DOPE conjugation and 100% DOPE lipid composition (Chol, C/E and Chol, C: n = 16, DOPE, C: n = 24; *: $P < 0.05$, ***: $P < 0.001$)

### 3.4.7. Effects of hybridization of complementary strands onto the SNAs

The versatility of the SNA architecture allows for alternative methods of incorporating the antigen into the structure, apart from loading in the lipid core. We investigated one such alternative – conjugation of the antigen to a complementary oligonucleotide, which is then hybridized to a lipid-anchored oligonucleotide. As a control, we also synthesized SNAs with the complementary oligonucleotide but without peptide conjugation. In these SNAs, the CpG containing oligonucleotide is double-stranded, and thus is differentiated from SNAs with only single-stranded oligonucleotides. In this conjugated OVA subset, we used DOPE-conjugated oligonucleotides to prevent the non-specific NF-$\kappa$B activation by cholesterol-conjugated SNAs described above.

Our results show that SNAs synthesized with this strategy shared some trends with their single-stranded counterparts. After oligonucleotide sequence, the most influential property on immune activation was backbone chemistry, with PS backbones outperforming PO versions (Fig. 3.5a). Again, we found that the core properties of lipid composition and core diameter were not significant.

Interestingly, for the SNAs with PS oligonucleotides, the addition of the complement oligonucleotide, either to half or to all of the anchored oligonucleotides, did not change immune activation at concentrations of 100 nM or 1 $\mu$M, respectively (b). Furthermore, there was no difference between SNAs composed of the complement, with and without conjugated peptide. However, at low concentrations (10 nM), higher complement densities led to higher immune activation. This effect may be

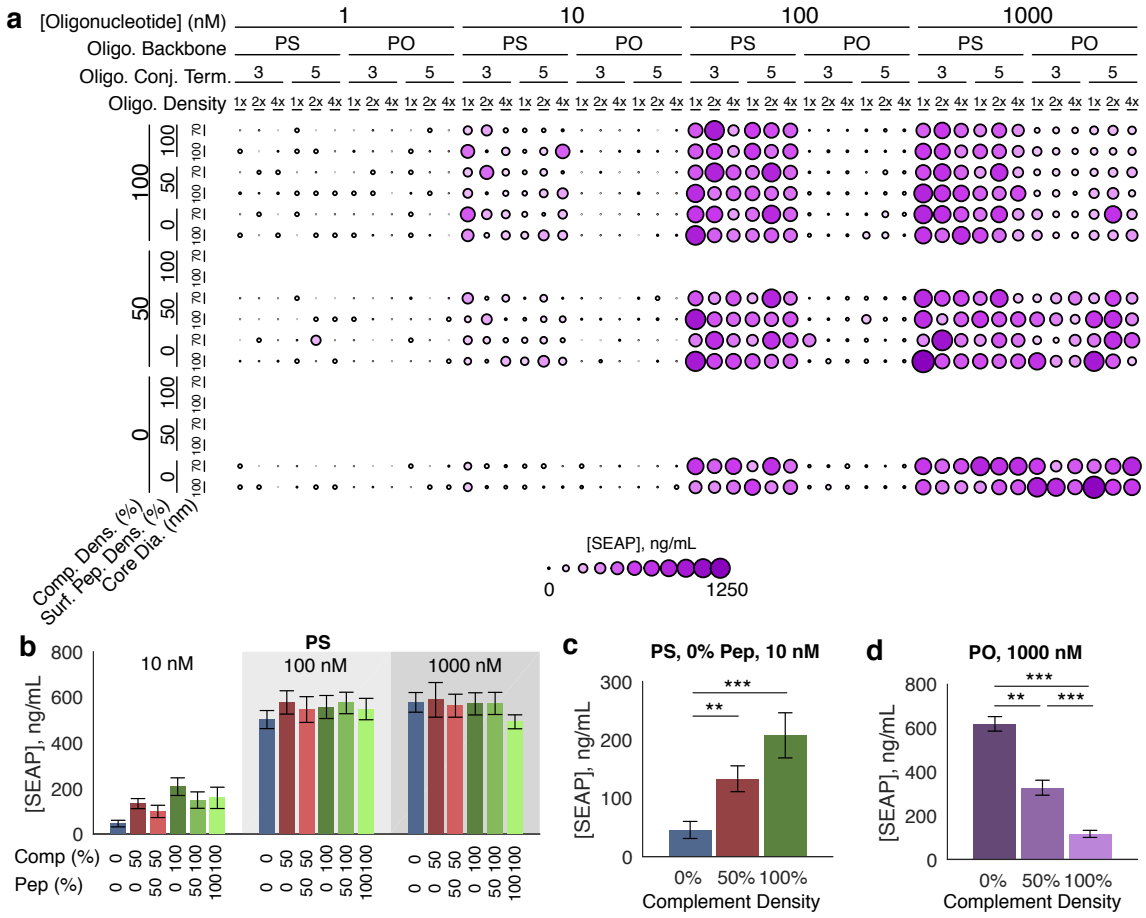



Figure 3.5. **Visualizing immune activation in the surface-presented OVA subset.** **a**, The SEAP concentrations for all active-sequence SNAs in the surface-presented OVA subset. **b**, The mean SEAP concentration of PS-backbone, active-sequence SNAs, grouped by the combinations of complement density and surface antigen density (n = 12). **c**, The mean SEAP concentration of SNAs with PS-backbone, 0% peptide and active-sequence, as a function of complement density, at 10 nM oligonucleotide concentration (n = 12). **d**, The mean SEAP concentration of SNAs with PO-backbone and active-sequence, as a function of complement density, at 1 $\mu$M oligonucleotide concentration. (100%: n = 36, 50%: n = 24, 0%: n = 12; **: $P < 0.01$, ***: $P < 0.001$))

a function of SNA uptake, where higher complement densities create higher charge densities on the surface and increase the uptake of SNAs, which in turn lead to higher immune activation. In contrast, complementation strongly reduced the activity of PO-backbone SNAs at the highest concentration tested (1 $\mu$M). A possible explanation for the decreased activity in duplexed SNAs is that the duplexing interferes with the oligonucleotide interaction with TLR9, however, it is not clear why the interaction with TLR9 would be different with PO and PS backbones. These results suggest that the strategy of including antigens by duplexing antigen-conjugated complementary oligonucleotides is effective with PS SNAs without concern for losing activation of TLR9.

### 3.4.8. Supervised machine learning captures non-linearity of property interactions and confirms trends in biological importance of properties

Because many of the parameters studied were interdependent in defining the activity of the SNAs, we trained linear and non-linear supervised learning to predict immune activation and to evaluate the relationships between SNA properties and confirm their relative impact.[86,87] Specifically, multiple linear regression, logistic regression, and non-linear xgboost were employed to fit training data and cross-validation of test data was conducted using the $Q^2$ statistic. $Q^2$ quantifies the accuracy of the predicted SEAP concentrations against measured values, and ranges from -$\infty$ to 1,

where 0 indicates no predictive power (equivalent to predicting the mean) and 1 indicates perfect prediction.[14]

We trained each model with all combinations of properties (i.e., 2 properties at a time, 3 properties at a time, and so on) and analyzed their $Q^2$ performance. As additional properties were added to the models, the $Q^2$ performance increased, plateauing for most models and decreasing in the xgboost model for the surface-presented OVA subset (Fig. 3.6a and b). Since clear non-linear trends were observed in the data as described above, the model performance increased with the non-linearity of the model in both subsets (mean increase from 0.53 for the linear model to 0.83 for xgboost). Analysis of the most predictive SNA property combinations demonstrate that highly predictive properties remain significant and informative as more properties are introduced into the model (Supplementary Fig. B.2a and b). In addition, the order of importance of the properties was largely consistent between the encapsulated OVA and the surface-conjugated OVA subsets, suggesting that the ordering is robust regardless of peptide localization.

For the encapsulated OVA and surface-presented OVA subsets, the $Q^2$ value stopped increasing beyond five and four properties, respectively (Fig. 3.6a). At first glance, one might conclude that only these highly predictive properties are relevant; however, when repeating this analysis with fixed values for sequence and concentration (the two features with the greatest impact),the $Q^2$ values stopped increasing

after another five properties were added (Fig. 3.6a), indicating that formerly seemingly non-predictive properties do, in fact, influence immune activation (Supplementary Fig. B.2c). Taken together, these properties, which appear non-influential in a global context, become impactful in a restricted design space.

### 3.4.9. Capturing maximal structure-activity relationship with minimal SNA synthesis and evaluation

We next investigated if a similar $Q^2$ level is attainable with fewer, randomly selected SNA designs. This question is particularly relevant when synthesis and evaluation of full libraries are impractical, but where exploration of a large design space is desired. In that case, one could synthesize a random subset that would capture the most important trends and then suggest additional candidates to evaluate. To this end, we simulated this process by training an xgboost model on a random selection of SNAs and testing predictions on the remaining, unselected SNAs within the three subsets (Fig. 3.6c). We identified the point of diminishing returns, which balances the minimum number of SNAs with maximum $Q^2$, by calculating the sample size closest to training size 1 and $Q^2$=1. This point is 90, 20, and 31 SNAs (out of 336, 336, and 288 SNAs) with $Q^2$= 0.67, 0.88, and 0.66 for encapsulated E7, encapsulated OVA, and surface-presented OVA subsets, respectively. These points represent a mean of 16% of the total number of SNAs, suggesting that a small number of randomly selected SNAs can predict SAR of a relatively large SNA library. In practice, this *external* $Q^2$ (prediction of non-synthesized SNAs) cannot be measured with

a randomized sub-sample, but an *internal* $Q^2$ can be measured by cross-validating within the randomized sub-sample. We show that the internal and external $Q^2$ are highly correlated (Fig. 3.6d and Supplementary Fig. B.3), suggesting that we can identify the point of diminishing returns as we continually synthesize random SNAs from an arbitrary library size. Combined with the high-throughput SNA synthesis and characterization approach described above, the machine learning analysis shows that a combined experimental/computational method can probe and predict the SAR of tens of thousands of SNAs with a much smaller subset (order of thousands) of structures.

### 3.4.10. Conclusion

In conclusion, this work—as well as other approaches pioneered by Anderson and Langer[88,89]—makes clear the need to consider the full range of structure-activity relationships when designing nanomedicines by high-throughput processes. Although high-throughput techniques are industry-standards in the combinatorial screening of small molecule drugs, such approaches are just beginning to be implemented to define structure-activity relationships for therapeutic nanoconstructs. The data presented herein show that such properties can be strongly interrelated in non-obvious ways, and emphasize the risks in using limited data to make global conclusions about one structural consideration being more critical than others. This interdependence and non-linearity are underscored when applying the non-linear machine-learning models, as opposed to linear ones, in predicting the biological response of SNAs.

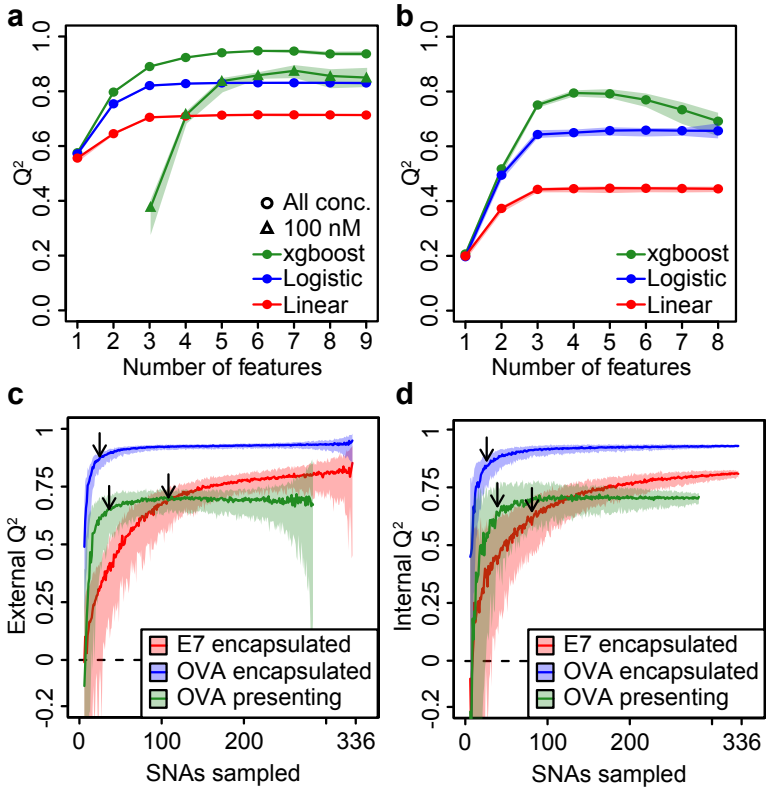



Figure 3.6. **Machine learning identifies relevant SNA properties and expands exploration capabilities. a**, The $Q^2$ of the highest performing SNA property combinations are shown across different numbers of properties for encapsulated OVA and **b**, surface-presented OVA subsets. For the encapsulated OVA subset and xgboost model, the active sequence and 100 nM subset is shown ($\Delta$) in addition to both active/inactive sequences with all concentrations (o). **c**, Xgboost $Q^2$ performance when selecting and training on a random SNA subsample and testing predictions on the unselected SNAs or **d**, cross-validating within the selected subsample. All plots have 90% confidence intervals. The arrows point to the point of diminishing returns.

Indeed, to realize rational approaches to vaccinology, this work makes a strong case for the combination of high-throughput experimentation and computational analysis, in determining the structure-activity relationships of nanomedicines in general and SNAs in particular.

CHAPTER 4

# DUETT quantitatively identifies unknown events in nascent RNA structural dynamics from chemical probing data

This work is in collaboration with Angela Yu and Julius Lucks and is in preparation for submission.

## 4.1. Abstract

RNA molecules undergo complex structural dynamics, especially during transcription, which influence their biological functions. Recent high-throughput chemical probing experiments study RNA cotranscriptional folding to generate nucleotide-resolution reactivities for each length of growing nascent RNA. However, the manual annotation and qualitative interpretation of reactivity across these large datasets can be nuanced, laborious, and difficult for new practitioners. We therefore developed a quantitative and systematic approach to automatically detect RNA folding events from these datasets to reduce human bias/error and standardize event discovery. Newly discovered events generate hypotheses about RNA folding trajectories for further analysis and experimental validation. Detection of Unknown Events with Tunable Thresholds (DUETT) identifies two types of RNA structural transitions in cotranscriptional RNA chemical probing datasets: swing events denote large structural changes at defined points in transcription, and ramp events denote gradual

structural refinements or relaxations that occur across multiple RNA lengths. We employ separate methods to detect each: a method inspired by feedback control identifies swing events, and a linear regression method identifies ramp events. Each method relies on interpretable and independently tunable parameter thresholds that are adjusted to match qualitative user expectations with quantitatively identified folding events. We validate the approach by identifying known RNA structural transitions of the *E. coli.* signal recognition particle (SRP) RNA and the *Bacillus cereus crcB* fluoride riboswitch. We identify previously overlooked RNA behavior such as consistently heightened reactivities in the SRP about 12 nucleotide lengths before base pair rearrangement. We then apply a sensitivity analysis to identify trade-offs when choosing parameter thresholds. Finally, DUETT is tunable across a wide range of contexts, enabling flexible application to study broad classes of RNA folding mechanisms.

## 4.2. Introduction

RNA molecules play diverse functional roles ranging from catalysis of splicing and peptide bond formation, regulation of mRNA processing and gene expression, and molecular scaffolding and localization among many others.[90,91] The particular RNA structure in the cellular environment is central to these functional roles. These structures are diverse and can precisely orient nucleotides for ligand recognition or catalysis in specific mechanisms, and they can generally prohibit/promote interactions with other cellular RNAs, proteins, and metabolites to enable the broad range of RNA function. For example, bacteria can have simple RNA hairpin structures to

inhibit transcription elongation[92], translation initiation[93], and RNA degradation.[94] Bacterial RNA can even be engineered as regulatory switches for each of these processes.[95] In eukaryotes, regulatory RNAs such as microRNAs and small interfering RNAs form hairpin structures before they are processed within the RNA interference machinery[96], and there is growing evidence that RNA structure may generally impact many gene expression processes.[97–100] Despite their importance, our ability to interrogate RNA structures in folding regimes that match the complexities of the cellular environment is relatively new[99], and there is a wealth of knowledge yet to be uncovered surrounding the principles of the RNA structure-function relationship.

Within the cellular RNA folding regimes, we know perhaps the least about how nascent RNAs fold during transcription[101,102]. Due to timescales of RNA folding and transcription, RNA molecules can begin to fold immediately as they emerge from RNA polymerase.[103] Local folds can form almost instantly, and as more stable folds become possible due to an increase in available sequence, RNAs can transition between states in a cotranscriptional folding pathway that dictate RNA function. For example, natural RNA biosensors called riboswitches dynamically alter structure during transcription in response to ligand binding, leading ligand-dependent structural, and ultimately regulatory, switching.[104] In addition, there is emerging evidence that cotranscriptionally-formed RNA structures influence a range of processes in eukaryotes such as splicing[105] and 3' end processing of histone mRNAs.[106]

There has thus been a great interest in developing both computational and experimental approaches to uncover RNA cotranscriptional folding pathways and their implications for cellular RNA function.

Recently, new experimental techniques characterize cotranscriptional RNA folding at nucleotide resolution[107,108] by utilizing high-throughput chemical probing of RNA structure.[109] Chemical probing experiments use the structure-dependent reactions of probes with an RNA[110,111]: once folded, RNAs are treated with probes that covalently modify RNAs preferentially at positions that are unstructured. For example, SHAPE (selective 2'-hydroxyl acylation analyzed with primer extension) probes attach as adducts at the 2'-OH backbone position of each nucleotide.[112] Adduct positions are then mapped by reverse transcribing the RNA, sequencing the resulting cDNAs, and analyzing the resulting reads for mutations[113,114] or cDNA truncations[112,115] that are indicative of adduct position. When coupled with high-throughput sequencing, these experiments reveal detailed reactivity patterns that uncover RNA structural properties—highly reactive positions indicate regions of unstructured RNA and lowly reactive positions indicate constrained regions due to structure or interaction with other RNAs, ligands, or proteins.[116–118] Recently, an experimental variant called SHAPE-Seq probes the structure of each intermediate length RNA during active RNA polymerase transcription.[107,108] This experiment results in a matrix of reactivities, where the rows in the matrix correspond to reactivity at each length of a growing nascent RNA chain, and the columns represent reactivity changes within specific nucleotides that reflect structural state changes during

transcription. For example, a decrease in reactivity down a column indicates the presence of a folding event, while an increase indicates an unfolding event during transcription.

Cotranscriptional SHAPE-Seq has uncovered important mechanistic insights into cotranscriptional folding pathways that underlie long-range interactions in non-coding RNAs as well as the nature of the ligand-dependent folding bifurcations of riboswitches. However, analysis of the cotranscriptional reactivity matrices has so far been mostly qualitative, relying on manual identification of reactivity trends to identify key regions and changes that have biological significance. As the number and complexity of these datasets grow, quantitative and automated techniques are needed to robustly identify patterns in datasets and corroborate functional significance through additional experimental data. This automated quantitative approach is challenging, as SHAPE datasets that contain complete annotations with validated structures through other methods such as crystallography are not bountiful; we lack a vast amount of "ground truth" examples. The scarcity of training data presents difficulties when defining statistical models[119] and prohibits application of machine learning models[120], such as visually identifying single nucleotide variants in RNA structures labeled by experimental experts.[121] These limitations suggest that we require a systematic method that identifies RNA structural dynamics from interpretable rules.

To overcome this challenge, we sought to develop a quantitative and automated approach to identify trends in cotranscriptional SHAPE-Seq datasets. Combining

approaches from control theory and linear regression, we developed two complementary methods that are suited for identifying two major patterns of reactivity changes that are common in these datasets and are indicative of biologically relevant folding transitions. Specifically, we sought a systematic detection method that remains human-tunable across diverse contexts, and with an interpretable set of parameters. Due to RNA structural complexities and the flexibility of SHAPE-Seq applications/implementations, we opted for a systematic approach to detect generic events. This philosophy is common in domains with poorly defined events such as detecting surprising instances in videos[122] or identifying unknown genomic deletions and insertions.[123,124]

We therefore implemented Detection of Unknown Events with Tunable Thresholds (DUETT). DUETT detects swing events using a strategy inspired by proportional-integral feedback control[125] and detects ramp events using a linear regression approach. Swing events represent rapid structural changes that occur over few transcript lengths. In contrast, ramp events represent slower events that span many transcript lengths. DUETT provides automated threshold parameter optimization, but because SHAPE-Seq data and RNA structural dynamics vary between experiments, DUETT also allows human-defined parameter-tuning to match a wide range of experimental contexts. We first establish these methods and find parameters that robustly identify known folding events within the *E. coli* SRP RNA. We extend the same methodology to the *B. cereus* fluoride riboswitch and corroborate previous manually identified transitions. In both datasets, our analysis revealed unexpected

behavior such as subtle reactivity increases that consistently occur roughly 12 nucleotide lengths before a reactivity decrease, suggesting a highly-reactive transient structure. Finally, we provide parameter sensitivity analyses to explore the relationship between parameter values and observed RNA structural dynamics. Due to the flexible approach and interpretable tuning parameters, we expect DUETT to be applicable to many high-throughput experimental systems that require event detection.

## 4.3. Methods

### 4.3.1. Event detection

Structural events are characterized by significant changes in reactivity across sequential transcript lengths. We consider two common yet distinct phenomena in cotranscriptional SHAPE-Seq datasets: swing and ramp events. Swing events correspond to rapid step-changes in reactivity across a limited number of transcript lengths (matrix rows). Upswings and downswings reflect unfolding and folding transitions, respectively. Ramp events correspond to gradual changes in reactivity that persist over many sequential transcript lengths or matrix rows, and may similarly occur in either direction. These two qualitatively different classes of dynamic behavior motivate separate detection methods for each event type. Assumptions are explicitly listed alongside their design consequences in STable C.1.

### 4.3.2. Swing event detection is motivated by PI control

We detect swing events using a technique inspired by control theory. Feedback control is widely deployed throughout the process industries to mitigate fluctuations of key process variables about a desired system state[126]. One common application is to maintain steady state behavior by taking corrective action based on the measured deviation of controlled variables from their nominal steady state values. This strategy is premised on the notion that major sustained deviations necessitate more aggressive intervention than minor brief fluctuations. PI controllers scale the strength of their corrective action both proportional to ($P$) and with the integral of ($I$) the measured deviation from steady state. The $P$ and $I$ terms thus provide a convenient framework for quantifying the magnitude and duration by which the system has deviated from steady state.

Swing events are characterized by abrupt changes in SHAPE-Seq reactivity during transcript elongation. These events may also be thought of as deviations ($D$) about a constant reactivity value:

$$D_i = z_{i+1} - \bar{z}_{i:(i-n)} \tag{4.1}$$

where $i$ indexes transcript length, $z$ is the SHAPE-Seq reactivity, and $\bar{z}_{i:(i-n)}$ is the mean reactivity within a sliding window of length $n$. DUETT quantifies the absolute magnitude and duration of these deviations at each transcript length by adopting

both $P$ and $I$ terms:

$$P_i = D_i \tag{4.2}$$

$$I_i = \frac{1}{n} \int_{i+1}^{i+1+n} D_i \tag{4.3}$$

where the $I$ term is normalized by the window size. A third quantity captures the relative ($R$) magnitude of deviations:

$$R_i = \frac{D_i}{\bar{z}_{i:(i-n)}} \tag{4.4}$$

An upswing event is detected when each of the $P$, $I$, and $R$ ($PIR$) values exceed user-defined thresholds. The $P$ and $R$ thresholds ensure that changes in reactivity are sufficiently large and distinct from the local steady state to represent true RNA structural dynamics (Figure 1). The $I$ threshold ensures that deviations reflect sustained changes in reactivity rather than brief noise-driven fluctuations (Figure 1). All $PIR$ thresholds are specified with positive values, with values of zero denoting constant reactivity.

Downswing events are similarly detected using the additive inverse of the PIR thresholds. The downswing R threshold requires a slight modification to retain an equivalent magnitude to its upswing counterpart:

$$R_{down} = \frac{-R_{up}}{1 + R_{up}} \tag{4.5}$$

For example, an upswing with 50% increase relative to the local steady state reactivity ($R_{up} = 0.5$) followed by a downswing of equivalent magnitude ($R_{down} = 0.33$, 33% decrease) ultimately results in no net change in reactivity.

We introduced two additional threshold parameters to further mitigate the impact of minor fluctuations driven by measurement noise. We remove swing events that are shorter than a specified duration threshold as we assume that structurally informative events generally persist for longer durations than noise-driven fluctuations (STable C.1). Conversely, we merge swing events separated by a gap less than a swing event gap threshold to reject noise-driven fluctuations that intersperse real swing events.

### 4.3.3. Automated parameter selection for swing event detection

DUETT provides a method to automatically select *PIR* thresholds for any given dataset. Due to the subjective nature of event detection, the automated method relies on a heuristic similar to the elbow method in cluster analysis[127]. The heuristic identifies a threshold combination that balances lenient with stringent thresholds. The automated search starts with low *PIR* thresholds where both noise and real events are detected. DUETT scans over combinations of increasing *PIR* values (with a user-defined window size) and records the number of detected events. We expect that increasing thresholds removes noise and sharply decreases the number of detected events until leveling off, forming an elbow that corresponds to detection of true events. DUETT identifies this elbow by finding the threshold set closest to

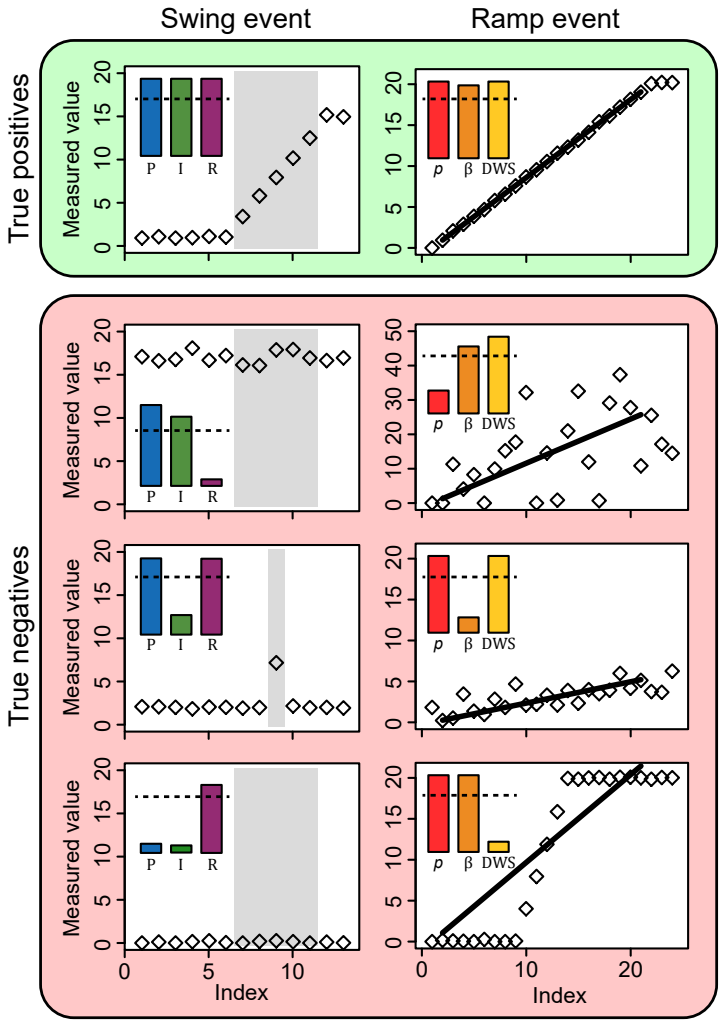



Figure 4.1. **Three thresholds filter out true positives from true negatives for each swing and ramp events** Swing and ramp events differ in terms of event length and have their own set of thresholds. For swing events, $P$ and $R$ filter out noise at low and high magnitudes, respectively, and $I$ filters out anomalously high values that persist for short durations. We show toy examples and the $PIR$ values in the region of interest (grey). For ramp events, $p$, $\beta$ and $DWS$ filter out high noise, low slope ramps, and swing events that may be erroneously classified as ramps, respectively. We show $p$, $\beta$ and $DWS$ values for the fitted line (solid line). $p$ is shown as -$\log_{10}(p)$ so larger values pass thresholds (dotted line).

the origin. This elbow represents the appropriate balance between lenient and stringent thresholds. If needed, the automatically identified thresholds serve as a starting point for manual tuning.

### 4.3.4. Ramp event detection using linear regression

Ramp events correspond to gradual changes in reactivity over broad stretches of sequential transcript lengths. The swing event detection method overlooks this class of events because it emphasizes rapid changes in reactivity constrained to brief sequences of transcript lengths. Instead, DUETT detects ramp events using a strategy based on linear regression. Given a user-specified window size corresponding to the minimum expected ramp length, lines of the form $y = \beta x$ are fit within windows sliding down each column of the SHAPE-Seq data matrix. Here, $x$ is a vector of sequential integers with length equal to the window size, $y$ is the measured reactivities within the corresponding window, and $\beta$ is a regression coefficient.

A ramp is detected when the fitted line passes three user-specified thresholds: a maximum $p$-value calculated from a one sample t-test for the regression coefficient, a minimum regression coefficient ($\beta$), and a minimum Durbin-Watson Statistic ($DWS$). Manual tuning of these thresholds is required to confidently detect ramp events. Fortunately, these parameters are readily interpretable (Figure 4.1). The $p$-value threshold controls the robustness of event detection against measurement noise; low values improve specificity at the expense of sensitivity. The $\beta$ threshold constrains effect size; high values exclude relatively flat ramp events with low average

change in reactivity. Finally, the $DWS$ threshold tunes the selectivity of ramp versus swing event detection. Swing events yield strong positive autocorrelation among residuals about the regression line (Figure 4.1, right panel of row 4), while residuals associated with true ramp events are uncorrelated. A $DWS$ threshold of unity precludes misclassification of swing events as ramps by filtering events whose sequential residuals are positively correlated[128].

### 4.3.5. Identifying concurrent events

Multiple events detected at similar transcript lengths likely reflect a common structural change. For example, a pair of upswings independently detected at two different RNA positions are involved in the same structural rearrangement if they occur at approximately the same transcript length. We label such instances as concurrent events by applying a transcript length proximity threshold.

### 4.3.6. Computational development and graphical user interface

DUETT was programmed in the freely available statistical software environment R and RStudio. We provide the source code and a graphical user interface as an R Shiny app located at github.com/bagherilab/DUETT. The app facilitates parameter tuning by continually updating figures as parameter values are varied by the user. The app also provides a suite of formatting tools for generating appropriate figures and tables.

### 4.3.7. Application to Cotranscriptional SHAPE-Seq Datasets

We apply DUETT to two RNA sequences characterized by previous SHAPE-Seq experiments[108]: the *E. coli* SRP RNA and the *B. cereus crcB* fluoride riboswitch. These published datasets were obtained from the Small Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/sra) with the BioProject accession code PRJNA342175. The data was processed with Spats v1.01 (https://github.com/LucksLab/spats/releases/) and the scripts are located at https://github.com/LucksLab/Cotrans_SHAPE-Seq_Tools/releases/.

## 4.4. Results and Discussion

We validate DUETT by identifying known cotranscriptional structural events in two RNA molecules, the signal recognition particle (SRP) RNA from *E. coli* and the fluoride riboswitch from *B. cereus*..[108,129,130] We use the automated approach to select *PIR* threshold parameters and manually select the same linear ramp thresholds across all datasets. During the automated search, the increase in *PIR* thresholds causes the number of detected events to rapidly decrease until reaching an elbow (SFig. C.1). As expected, the point closest to the origin lies at the start of the elbow and corresponds to the automatically selected threshold values. We apply DUETT on each of the three replicates and retain events conserved across all replicates to decrease the likelihood of spurious events. This approach creates similar results as averaging all replicates (SFigure C.2) but avoids scenarios where a single replicate has anomalous values, biasing detection. We identify both known and unknown

structural events and propose novel hypotheses for further study. We highlight patterns and events that DUETT found but are difficult for a human to identify. We conclude with sensitivity analyses to explore the relation-ship between user-defined threshold parameters and observed events.

### 4.4.1. DUETT is validated on the *E. coli* SRP RNA and identifies unknown structural dynamics

Previous studies showed that the *E. coli* SRP RNA forms an intermediate 5' hairpin (H1) that reforms into a long helical structure with a hairpin loop and multiple inner loops[108,129,130], which we label H2-H5. Many upramps begin at or close to the nucleotide's (nt) transcription by RNAP. This association suggests SHAPE attachment begins almost immediately after RNAP transcription. Due to experimental limitations, these short RNA fragments are difficult to detect leading to reduced signal. As the RNA elongates, SHAPE adducts become increasingly detectable and create an upwards linear ramp. As a result, we infer that bases with upramps close to the initial transcription site are identifying SHAPE adducts that manifest as experimental artifacts due to their position near the 3' end of the RNA.

**4.4.1.1. DUETT identifies expected H1 hairpin formation and rearrangement.** Bases 14-15 and 18 have upswings around length 45 nt or upramps that conclude around length 50 nt (Fig. 4.2). These positions remain unpaired in the intermediate H1 hairpin, validating the upswings/upramps. This hairpin rearranges

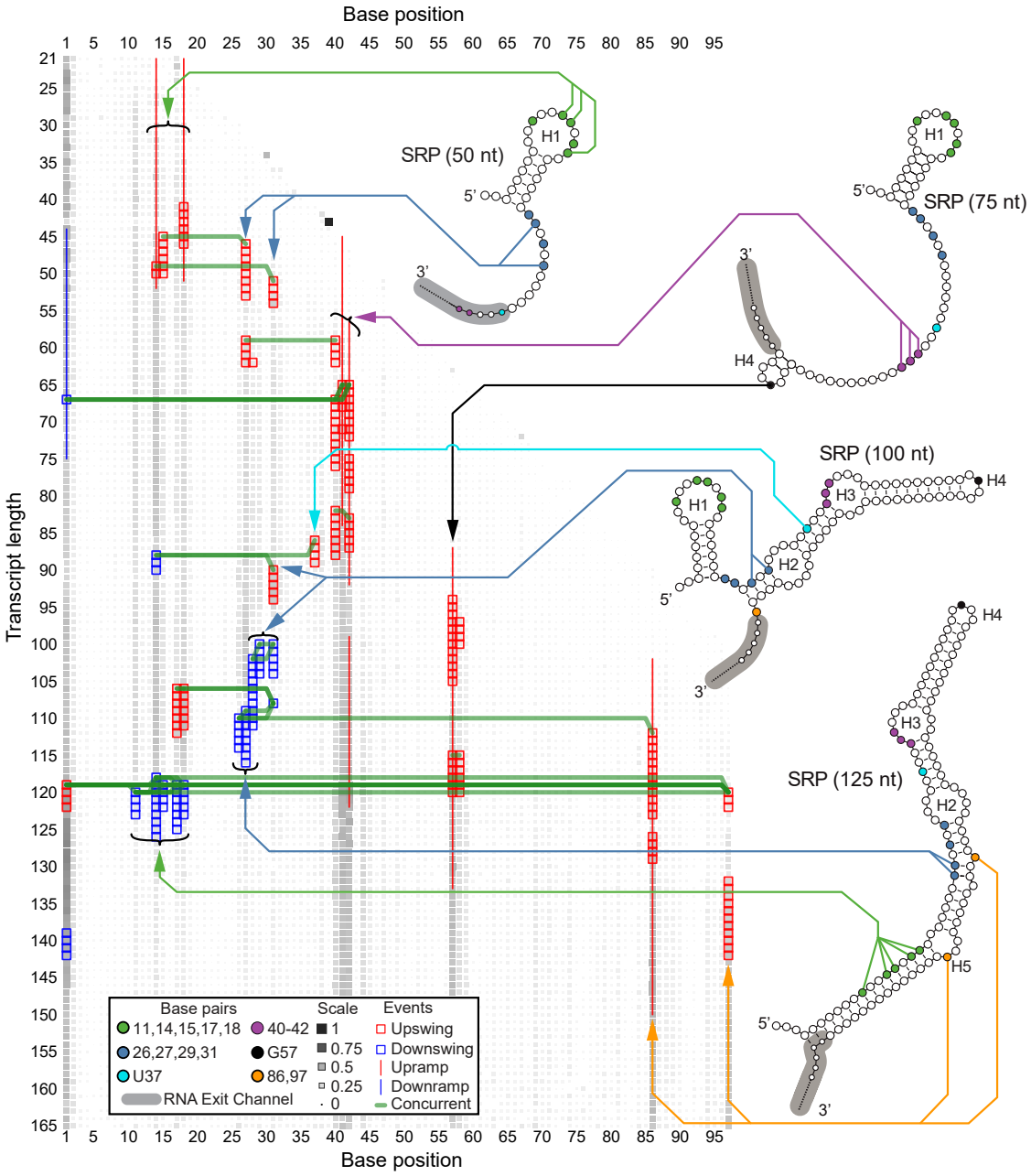



Figure 4.2. **SHAPE-Seq event detector identifies known RNA structural events in *E. coli* SRP RNA** Four structural conformations of SRP RNA are shown with arrows linking specific bases to identified structural changes. The event detector identifies multiple instances of hairpin formation/reformation and we propose structural explanations for novel events.

into the long helical structure, which DUETT identifies as downswing events between lengths 120 and 125 nt at bases 11, 14-15, and 17-18. Though observations largely agree with expectations, the curious upswing in bases 17-18 at 106 nt differs from the behavior of the other positions and is suggestive of increased flexibility that expose these 3' positions to SHAPE reactivity. We hypothesize that increased flexibility is a precursor to the H1 hairpin rearrangement and occurs roughly 12 positions beforehand.

**4.4.1.2. Identification of multiple expected pairings validates DUETT.** DUETT identifies the expected dynamics of bases 26-27, 29, and 31. Upon initial transcription, bases 27 and 31 have upswings corresponding to their unpaired state, and bases 28-29 and 31 have downswings at 100 nt that agree with the previously proposed 100 nt structure. As expected, bases 26-27 remain reactive and lack downswings until length 110 nt, when they presumably base pair before rearranging into the final helical structure. Like bases 17-18, base 31 exhibits an upswing roughly 12 nt lengths before its downswing event. In addition, base 27 has a similar non-detected upswing in the same location (Appendix C File 1) and correlation analysis suggests that base 27 is similar to base 26, 29, and 31 (Appendix C File 4). Coincidentally, when bases 26-27 or bases 29-31 pair with their respective partners, the consecutive upswing/downswing is in the most 3' base. This observation agrees with bases 17-18 and supports the hypothesis that increased flexibility in the 3' side occurs before pairing up. We note the difficulty in manually detecting this pattern, justifying our systematic approach.

We provide additional validation by identifying the formation of other unpaired positions. A cluster of upramps/upswings in bases 40-42 between lengths 55 and 85 nts corresponds to the open region in hairpin H3. Base 40 was reported to be paired by length 100 nt[108], corresponding to an undetected downswing at 94 nt (Appendix C File 1). In contrast, bases 41-42 continue increasing reactivity. Though base 40 has lower reactivity than bases 41-42 (Appendix C File 1), its reactivity is higher than expected from a canonical base pair and is correlated with bases 41-42 (Appendix C File 4). Our results suggest that base 40 is more labile than previously reported.[108,129] We attribute this accessibility to relatively few canonical pairs in the vicinity around bases 40 and 72. It is also attributable to the generally less stable nature of GU pairs, helix ends, and near inner loops.[131,132] Finally, bases 86 and 97 have upramps/upswings immediately after transcription that corroborates their expected unpaired status.

**4.4.1.3. Unexpected events highlight previously overlooked structural dynamics.** We identify two novel and unexpected events in bases 14 (downswing), and 37 (upswing) at lengths 88 and 90 nts, respectively. These observations are discordant with the previously proposed folding model of the SRP RNA: base 14 remains unpaired in hairpin H1 and base 37 pairs up between lengths 75-100 nts. Qualitatively, the downswing in base 14 is concurrent with other non-detected downswings in neighboring bases 11 and 15 (Appendix C File 1). Similarly, base 37's upswing is concurrent with non-detected upswings in bases 32-33, 36, and 38 (Appendix C File 1). These observations lead us to believe that the detected events are not spurious

but lack an explanation by previously published studies and highlight the discoveries enabled by our systematic method.

### 4.4.2. SHAPE-Seq event detect identifies known and novel structural differences in a fluoride riboswitch

We apply DUETT to published SHAPE-Seq data from the *B. cereus* fluoride riboswitch.[108] The riboswitch was exposed to either fluoride-positive (10 mM NaF) or fluoride-negative (0 mM NaF) conditions, which causes structural changes in RNA folding.[2] We apply DUETT to both conditions and compare results. We tuned threshold parameters (STable C.2) to match expected structural dynamics such as the series of swing events around bases 52-55 between lengths 80 and 95 nts. We chose thresholds such that lower thresholds detect spurious events.

**4.4.2.1. Detected events before 69 nt agree with expectations that structures are identical in both conditions.** Before the structural divergence at 69 nt, our detected events agree with the proposed model that RNA structures are identical in both Fluoride conditions. These shared events occur between 22 nt and 55 nt in bases 13-16, 25, 29-30, and 34 (Fig. 4.3 and Fig. 4.4). Bases 13-16, 25, and 34 have upswings/upramps that confirm their unpaired status in both conditions. Additionally, bases 12-13 and 16 have downswings around 60 nt that agree with their pairing off prior to the 69 nt structure. Though bases 29-30 are consistent in both conditions, the detected upswings around 48 and 53 nt disagree with the riboswitch model; bases 29-30 are paired off within a hairpin stem, which should not

manifest as an upswing. These results suggest that bases 29-30 undergo increased SHAPE reactivity prior to pairing off and represents a nuanced structural dynamic. Otherwise, the detected events before 69 nt agree between conditions, reflective of identical RNA structures.

**4.4.2.2. Identification of delayed terminator nucleation agrees with riboswitch model.** We validate DUETT results by identifying events that agree with the delayed terminator nucleation. Exclusively in the 0 mM NaF condition, bases 12-16 are expected to unpair before the 77 nt structure, allowing bases 52-55 to pair up into a hairpin stem earlier than the 10 mM NaF condition.[108] Both conditions exhibit upswings for bases 52-55 around 73 nt due to different reasons: increased reactivity prior to hairpin formation in the fluoride-negative condition and unpaired bases in the fluoride-positive condition. The fluoride-negative bases immediately decrease in reactivity corresponding to pairing off while the fluoride-positive bases continue to increase in reactivity. The delayed terminator nucleation manifests with a series of downswings around 90 nt exclusively in the 10 mM NaF condition, which corresponds to forming the hairpin stem. In addition, bases 56 and 59 exhibit upswings/upramps in both conditions, corroborating their unpaired nature.

We analyze events that occur after terminator formation, when RNAP transcription is expected to halt exclusively in the fluoride-negative condition.[108,133] After about 90 nt, bases 69-71 and 74 remain unpaired in the 10 mM NaF condition and contain upswings shortly after transcription as expected by their reactive nature (Fig. 4.4). These upswings are expectedly missing in the 0 mM NaF condition except for

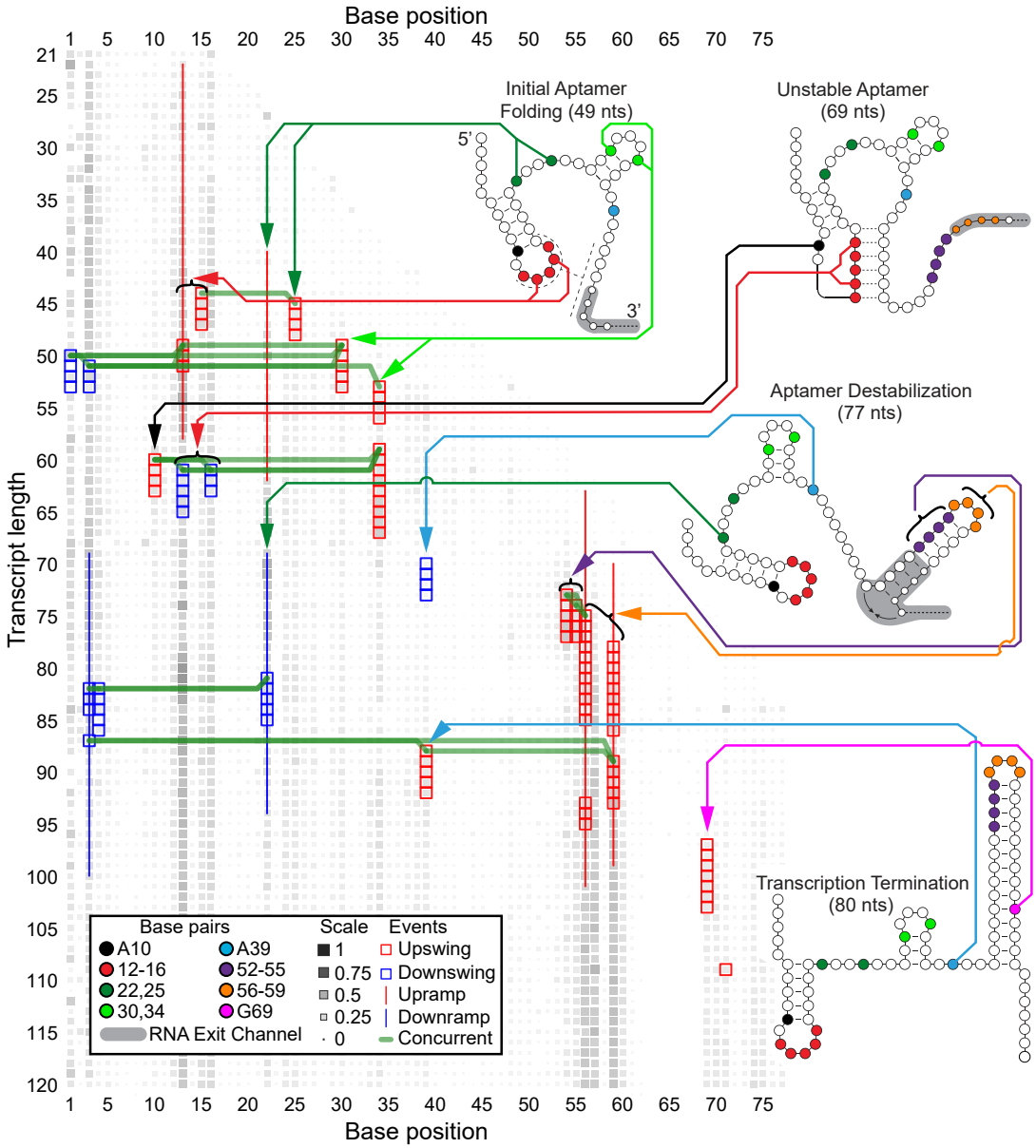



Figure 4.3. **Event detector identifies RNA structural dynamics in a *B. cereus* fluoride riboswitch, fluoride-negative condition.** The riboswitch was exposed to fluoride-negative (0 mM NaF) or fluoride-positive (10 mM NaF, Fig. 4.4) conditions. The SHAPE-Seq event detector identifies multiple known and novel structural events between in the fluoride-negative condition, which is shown with arrows linking nucleotides to previously proposed structural conformations.

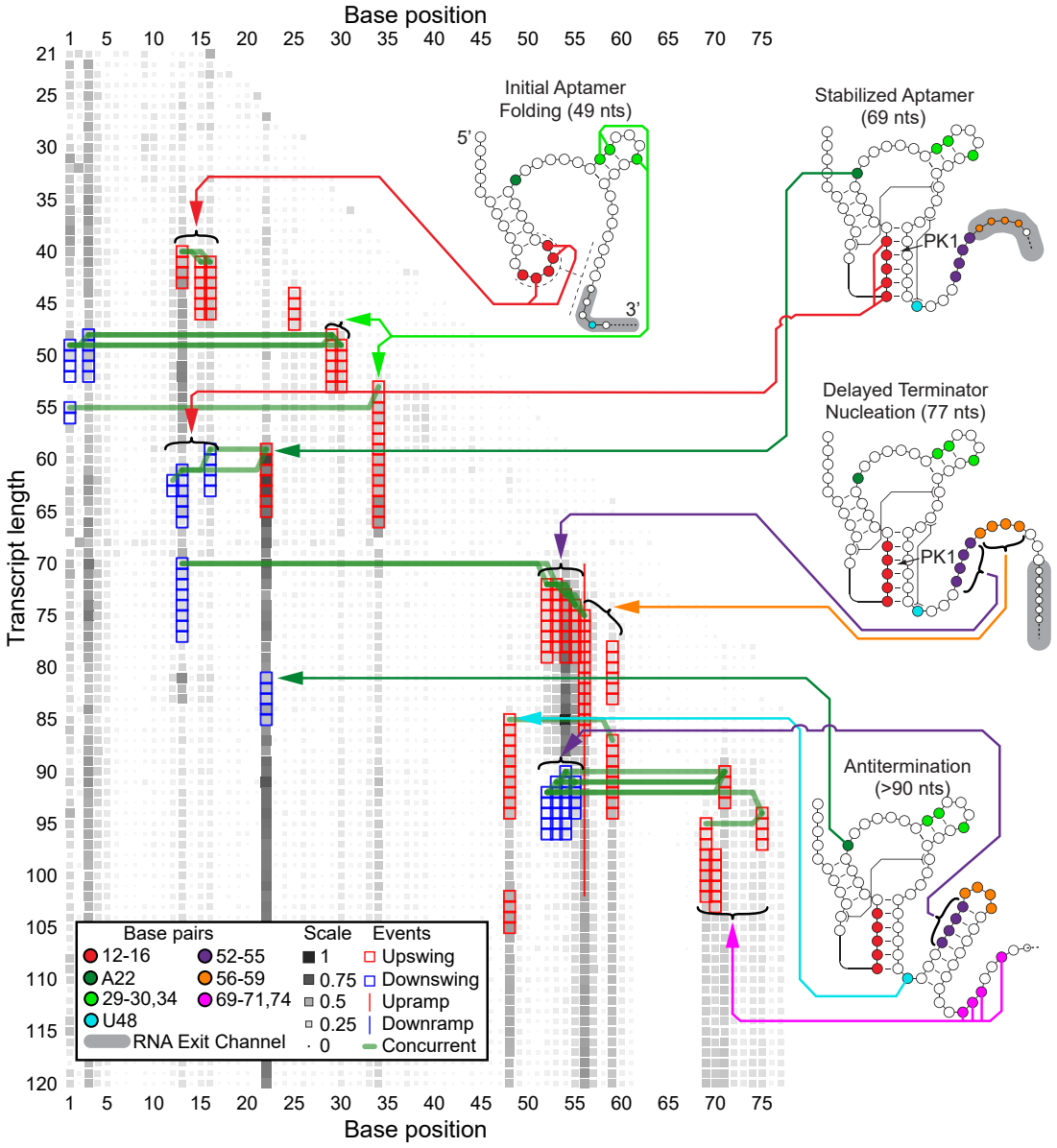



Figure 4.4. **RNA structural dynamics in the fluoride-positive riboswitch.** The riboswitch was exposed to the fluoride-positive (10 mM NaF). These results are compared to Fig. 4.3 results to examine structural divergences between the two fluoride conditions.

G69 (Fig. 4.3). G69 exhibits an unexpected upswing at 97 nt, which is after the fully formed terminator stem and should not be observed. However, it was shown that mutations that remove G69's pairing also prevent formation of the terminator stem, meaning that G69 pairing is one prerequisite of terminator formation.[108] A separate study, using a slightly different riboswitch sequence, found that a single base pair in the same terminator stem area plays a pivotal role in functional terminator stem formation.[134] These findings coupled with the G69 upswing suggest that a subpopulation of SRP RNA lost G69 pairing, leading to SHAPE reactivity and lost terminator function. However, the mechanism behind G69 unpairing is unclear and requires further work. This previously overlooked finding demonstrates DUETT's ability to flag interesting events for follow-up analysis.

**4.4.2.3. Identified events in bases 10 and 48 corroborate long-range interactions.** We inspect and corroborate two previously reported long-range interactions: A10-U38 and A40-U48.[108,133,134] These interactions are hypothesized to increase stability of the aptamer region (69 nts structure) and persist through transcription of the riboswitch only when fluoride binds.[108] In the 0 mM NaF condition, we observe an upswing in A10 at length 60 nt, which corresponds to aptamer formation and increased reactivity (Fig. 2.3). Conversely, this upswing is absent in the 10 mM NaF condition because the A10-U38 interaction prohibits SHAPE reactivity (Fig. 2.4). The other long-range interaction, A40-U48, is proposed to unpair between the 77 and 88 nt structures[108], which we corroborate with an upswing in U48 at length 85 nt.

Additionally, A39 is situated between the two long range interactions and exhibits a downswing and upswing at 70 and 88 nt, respectively, exclusively in the fluoride-negative condition. The downswing's existence and exclusivity are unexpected. Upon closer inspection, both conditions have a small undetected upswing around 55 nt followed by a downswing (Appendix C Files 2 and 3). Qualitatively, both conditions exhibit similar behaviors, but the events went undetected due to our rigorous approach to include only events detected across all three replicates. We conclude that A39 has similar SHAPE reactivity in both conditions before the structural divergence at 77 nt, as expected. After 77 nt, A39 exhibits structural divergence with an unexpected upswing at 88 nt that corresponds to increased reactivity. Previous NMR research shows that A39 (A35 in their numbering) undergoes local structural dynamics when no fluoride is bound and is stabilized when fluoride is bound.[134] The upswing may reflect those local structural changes. Conversely, the fluoride-positive condition lacks this upswing due to the neighboring A10-U38 long-range interaction and continued aptamer formation that prohibit SHAPE reactivity. Altogether, the detected swing events in A10, A39, and U48 support the proposed aptamer stabilization via long-range interactions.

**4.4.2.4. Novel A22 dynamics identified by DUETT.** We observe SHAPE hyperactivity in A22 via an upswing at length 59 nt in the 10 mM NaF condition; A22 hyperactivity is associated with aptamer stabilization.[108] This upswing is followed by a sharp downswing at length 81 nt and another undetected upswing shortly afterwards. The second upswing went undetected due to the short duration of the

previous downswing, which causes high and low reactivity positions to lump together during the sliding window averaging. Afterwards, the reactivity plateaus at a high value comparable to the 69 nt structure levels. The 0 mM NaF condition has similar dynamics but are less extreme and show up as ramps (Appendix C Files 2 and 3) demonstrating that swing and ramp events differentiate small from large changes as intended. We conclude that base 22 has similar dynamics (except in magnitude) across both conditions until about length 90 nt where only the fluoride-positive condition exhibits the rebound upswing. While the fluoride-negative downswing is justified (aptamer destabilization), the analogous fluoride-positive downswing disagrees with the stable aptamer and lacks a mechanistic explanation. A22's complex behavior was overlooked earlier due to the visual upper limit (4) set in the original figure.[108] While the upper limit simplifies data analysis/visualization, DUETT accounts for all magnitudes, and the detector is partially insulated from disadvantages in human visualizations. Altogether, DUETT identified several expected structural differences between the fluoride conditions, and we generate multiple hypotheses on unknown or unexpected events.

### 4.4.3. Tuning threshold parameters requires a tradeoff between true positive events and false positive/negative events

When choosing threshold parameters, a user balances identifying true positive events with accidentally identifying false positive/negative events. Slight differences between detected and non-detected events highlight user preferences; if identifying

small magnitude events (true positives) is prioritized, then thresholds can be relaxed to avoid undetected events (false negatives).

We perform sensitivity analysis to explore the true positive/false positive tradeoff, and we demo-strate that large events are retained despite drastic threshold parameter adjustments. We highlight two scenarios in the *E. coli* SRP RNA dataset using a stringent and a lenient set of *PIR* thresholds (additional sensitivity tests in C.1). The stringent *PIR* thresholds (50% increase in each threshold) yield fewer overall events (Fig. 4.5 right) relative to the original baseline (Fig. 4.5 center). As expected, qualitatively small changes are removed: the downswing at length 88 nt in base 14 and the downswings around 100 nt in bases 28-29. We observed that base 14's downswing is likely non-spurious and marks a new discovery. Similarly, the downswings in bases 28-29 are attributed to their pairing off before the final structure. These removals in the stringent scenario underscore the tradeoff that while higher thresholds lower false positives, it can also turn true positives into false negatives. We chose a large parameter increase but retained many of the originally detected events, suggesting that large events have a wide acceptable range of threshold values.

On the flip side, lenient thresholds (Fig. 4.5 left) generally allow more true and false positives. The downswing in base 11 at length 88 nt and the upswings in bases 27, 29, and 31-32 around length 90 nt are detected with lenient thresholds. By inspection, these events seem non-spurious and occur concurrently with other similar events (Appendix C File 1) leading to the conclusion that these are true positives that were original non-detected. Conversely, the lenient scenario creates potential

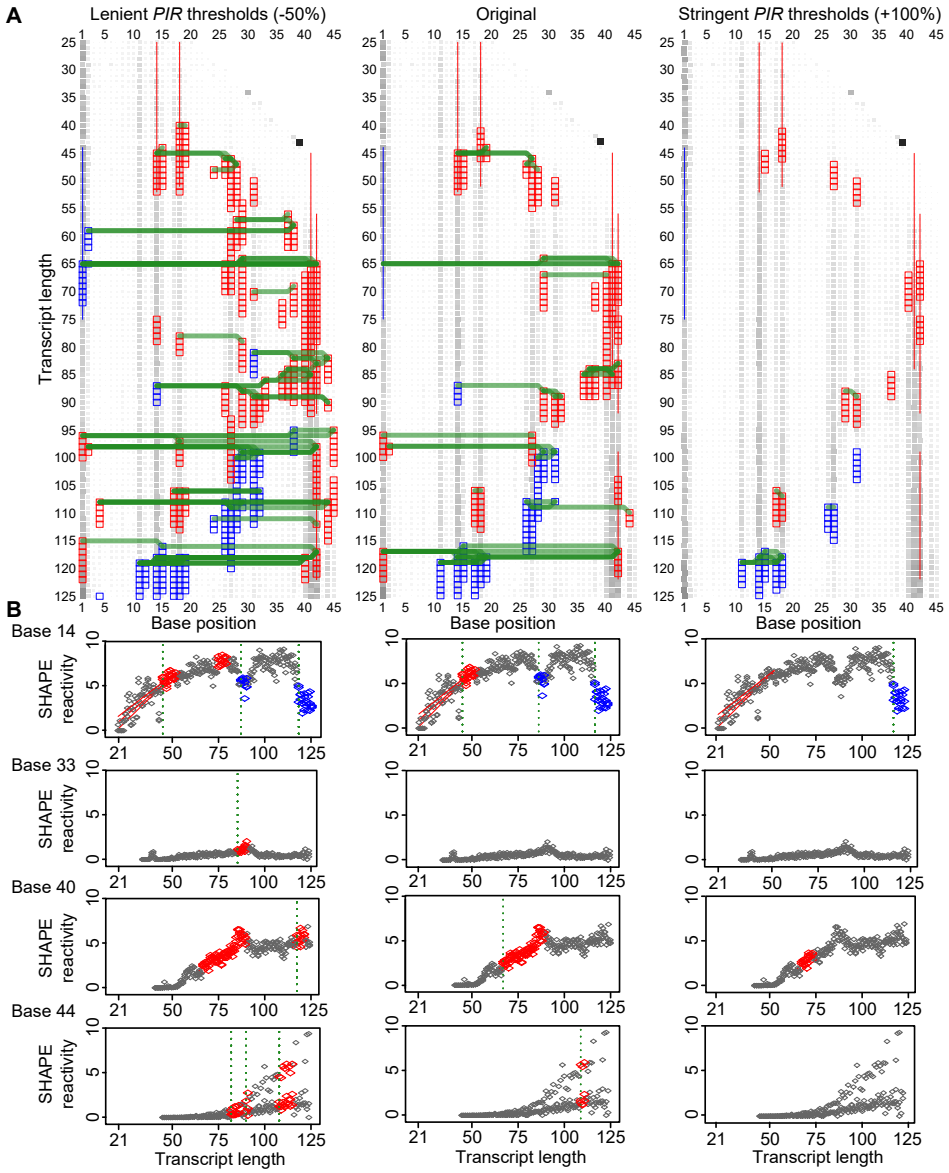



Figure 4.5. **Sensitivity analysis of user-defined thresholds illustrates the tradeoff between true positives and false positives/negatives. A**) The original *PIR* thresholds (center) are compared to a lenient scenario (left, -50%) and a stringent scenario (right, +50%). In general, lenient thresholds increase sensitivity towards small magnitude events, but false positives are introduced. Conversely, stringent thresholds have fewer false positives, but multiple false negatives are created instead. Large magnitude events tend to be non-sensitive towards thresholds. **B**) Individual examples are highlighted.

false positives. For example, the upswings in base 40 and 44 at lengths 121 and 112 nt, respectively, do not appear qualitatively like upswings and they lack a structural explanation. The upswings in base 44 are especially confounding; one replicate has increased reactivity with a potential upswing while the others remain flat. Instead, we conclude that this detected upswing in the lenient scenario is spurious and arose by chance due to the lenient thresholds. These findings reinforce the tradeoff between true and false positives and that selection of parameters is ultimately subjective and should be tailored according to user expectations. We chose drastic changes to parameters thresholds to illustrate their effect. Fortunately, many originally detected events remained even in the stringent scenario and few spurious events arose in the lenient scenario, suggesting that our methodology creates concordant results across a wide acceptable range of parameters. We provide additional sensitivity analysis on window length and linear ramp threshold parameters in C.1.

DUETT was designed to emulate human visual inspection of cotranscriptional SHAPE-Seq data in an efficient and systematic manner to both reduce potential biases and discover not easily identifiable events. Cotranscriptional SHAPE-Seq creates a wealth of data as many RNAs and RNA lengths are probed. This leads to increased complexity when interpreting the data as its cotranscriptional nature requires consideration of both the structures at each length and structural transitions between lengths. Study of riboswitches adds an additional layer of complexity in that ligand-dependent structural changes are studied as well. When interpreting cotranscriptional SHAPE-Seq data, it is also important to keep in mind that halted

nascent RNA structures are probed and fleeting structural changes are difficult to detect with this method. However, we detected some of these multi-state positions that were previously reported as events. Cotranscriptional SHAPE-Seq complements aptamer structure and dynamics information known from crystallography and NMR with the ability to probe nascent RNA structures through transcription. Thus, DUETT quickly establishes transcription lengths and nucleotides of interest from reactivities to be further interpreted and developed into a structural model. We hope this method will be adopted to provide a tunable baseline across RNA folding experiments and to identify structural events that elude visual identification in cotranscriptional data.

CHAPTER 5

# Interpreting supervised learning models

The work presented thus far has focused on applying data science tools to understand a dataset. These tools, especially supervised machine learning, are often criticized for being uninterpretable leading to models that do their job, such as a predictive task, without necessarily understanding *how* the model works. Because medical diagnostics require justification, how well do we need to understand the model's inner workings? Can the cutting edge research of machine learning address these concerns today? And what should the future of interpreting supervised models look like? The following opinion piece attempts to bridge the gap between machine learning and medical communities and present dialogue at a critical junction where data science tools are becoming a common answer for big data challenges.

## 5.1. Can an AI think like a doctor?

Data-driven models, such as deep learning models, have scored multiple successes in medical imaging tasks including identifying dermoscopic patterns[135,136] or classifying melanoma[137–139] and provide promising methods for learning health systems.[140,141] Deep learning models create predictions by processing data via layers of artificial neurons that mimic neuronal activity, leading to models that create predictions on traditionally difficult tasks, such as image analysis. Appropriately, hesitation to

adopt these models is partially due to their black box nature[142,143]; deep learning models offer accurate but inscrutable predictions. Given the need to justify medical treatment, models require interpretation[144] but models that are developed alongside interpretation techniques are not common. A review of prognostic cancer models found that two out of 47 models included a published explanation for patients.[145] Fortunately, deep learning inscrutability is a common concern leading to rapid developments in interpretation techniques.[142,146] While this research area is rapidly growing, it is often siloed within the machine learning community, and this separation has created techniques that succeed on common machine learning tasks but do not emulate human-level interpretations in a medical imaging context. In this opinion piece, I focus on how interpretation techniques do not yet emulate how clinical experts interpret images, and we motivate the issue with dermoscopic examples. I discuss the goal of interpretation research in medical imaging contexts; do interpretations need to emulate current human analysis or is it sufficient to offer a different but still convincing explanation? I offer an outlook that establishes a future organization that facilitates cross-community dialogue while hosting clinically-relevant challenges to accelerate research into clinical applications of data-driven models.

## 5.2. Current deep learning interpretation techniques are inconsistent with human image analysis

I discuss how current interpretation techniques fall short of emulating human-level medical image analysis. I divide interpretation approaches into two broad categories:

image-specific interpretations and model-based reasonings. Image-specific interpretations typically result in highlighted image regions that are relevant for a given prediction. However, image-specific interpretations often do not identify context-dependent visual features, such as heterogeneity or texture, that are commonly relevant in medical imaging tasks. In contrast, model-based reasoning methods probe a model's internal logic with example images and have demonstrated that models can identify complex visual features, such as texture. However, model-based reasoning does not produce interpretations for specific images resulting in inscrutable predictions.

Image-specific interpretation techniques often focus on identifying borders or areas within an image that are relevant to a model's prediction. These techniques result in an outline of an object in a diverse environment such as highlighting faces in a room[147,148] or outlining a car on a busy road[149] (Fig. 5.1A). The outline is interpreted as the region that is relevant for the deep learning model's prediction.[146] Input modification is one of the main deep learning interpretation techniques[146] and is based on modifying pixels; if a modified pixel changes the model's prediction, then the pixel is relevant and highlighted. The focus on outlining or highlighting relevant image segments is natural within the machine learning community because image recognition tasks often contain multiple objects with unique shapes within a single image. Well-studied objects include cars or pedestrians, and it is useful when a model/interpretation technique outlines the unique shape (Fig. 5.1A).

Image-specific interpretation techniques do not correspond well to dermoscopic analysis. Dermatologists examine lesions by viewing dermoscopic images and diagnosing benign vs. melanoma-specific visual features (morphology) that range from context-independent to context-dependent. Context-independent features are identifiable without inspecting the surroundings, but more complex features require a larger image context. Context-independent visual features include comma-shaped vessels that are generally benign in nature as well as serpentine or corkscrew-shaped blood vessels that associate with malignancies.[150] Context-dependent features include variation, symmetry, homogeneity, and other complicated patterns such as pigment networks with empty patches.[150] Similarly, subtle variations and non-uniformity in lesion color also correspond to malignancy.[150] These context-dependent visual features often require identification within the framework of the larger image, and an outline of relevant pixels does not convey that the model understands complex visual features (Fig. 5.2). For example, multiple distinct independent features arranged asymmetrically can signal malignancy (Fig. 5.2). While each feature does not signal malignancy by itself, the combination of features indicates malignancy due to asymmetric/heterogeneous organization. In Fig. 5.2 left, a border does not differentiate why the region is important. The first two examples have multiple visual features that when examined individually, each feature seems benign. When combined, these features are asymmetric and associated with malignancy. Similarly, the third example is a pigmented network, but empty patches become apparent only in context of the whole network. In all examples, a standard interpretation technique

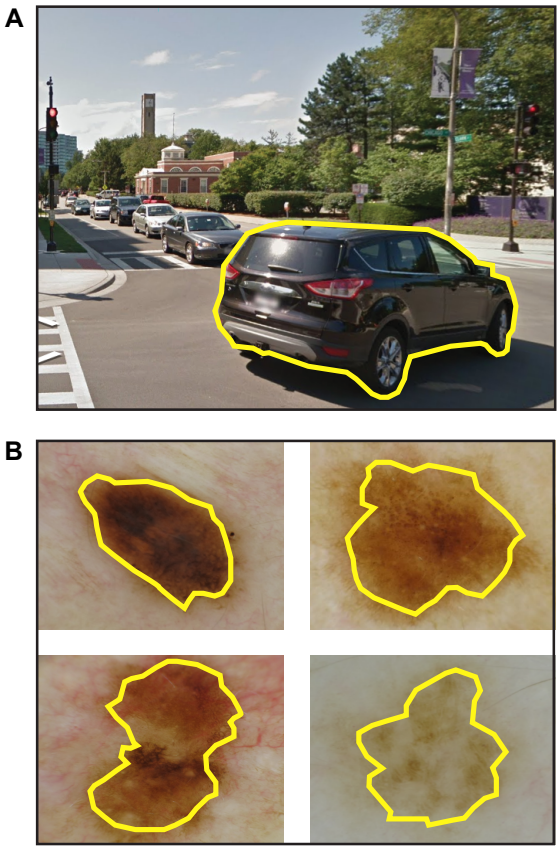



Figure 5.1. **Interpretation techniques identify the border of an object rather the combination or context of multiple visual features** Deep learning interpretation techniques often highlight relevant pixels resulting in outlines of areas that are relevant for a prediction. (A) In common machine learning tasks, the correct border is often useful to understand that the model correctly identified a uniquely shaped object within a complex background, such as the nearest car on a road. (B) In contrast, medical image analysis requires identification of context-dependent visual features that are not easily summarized by an outline. For example, a simple border does not convey how a model distinguishes between malignant (left) and benign (right) lesions in dermoscopic image analysis.

should highlight the whole lesion as relevant (left). This outline is not wrong; the whole lesion is relevant. However, this process does not emulate how a dermatologist might interpret these examples (Fig. 5.2 right): individual features are segmented and identified but their aggregation forms a complex feature such as asymmetry.

Unlike image-specific interpretation, model-based reasoning methods directly probe a model's internal logic by selecting or synthesizing example images and examining model behavior. One method includes optimizing an arbitrary image to cause a specific model prediction, resulting in an image with visual features relevant for that prediction.[149,151] In one respect, model-based reasoning approaches invert the normal image classifier by producing images that correspond to specific classes.[152] The resulting images are often unnatural-looking[153], but unlike image-specific interpretations, they may contain complex visual features such as context-dependent asymmetry or heterogeneity. Another approach selects images that cause a specific prediction and identifies the visual features conserved across the selection.[154] These conserved features can be complex[153], suggesting that this technique holds promise when assessing whether a model has understood medically relevant and context-dependent visual features. Despite these benefits, model-based reasoning techniques do not explicitly create explanations for a specific image-diagnosis pair. This approach does not allow a clinician to scrutinize how a model arrived at a specific diagnosis, unlike image-specific interpretations. While the machine learning expert can analyze and verify a model's internal rationale, the ultimate end-user, the clinician and patients, will have no explanations for specific diagnoses. As a result, we

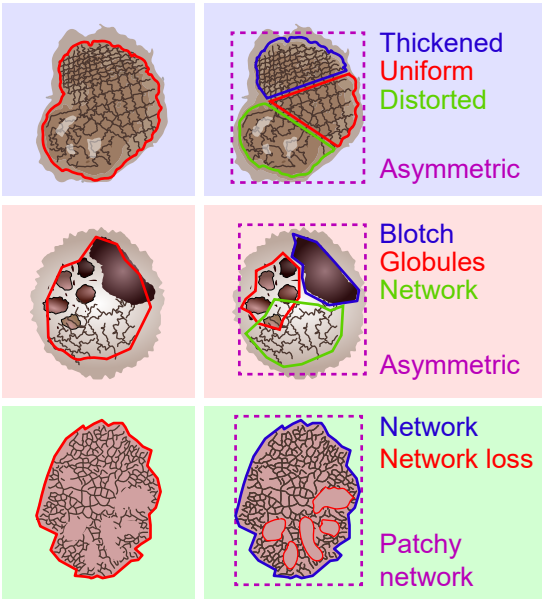



Figure 5.2. **A dermatologist's analysis differs from a model's interpretation output** An outline of relevant pixels does not emulate how clinicians analyze and make decisions from medical images. A standard machine learning interpretation (left) creates simple outlines around three illustrations of malignant lesions that contain context-dependent visual features. In contrast, a possible dermatologist interpretation highlights individual visual features and contextualizes the whole image before making a diagnosis (right). Ideally, interpretations that em-late a clinician's reasoning enables critical analysis of internal model rationale, engenders trust, and improves clinical adoption of high accuracy predictions.

are left with interpretation techniques that create diagnosis-specific interpretations or identify medically-relevant visual features, but not both, suggesting that current methods are insufficient for clinical adoption.

### 5.3. The goal of model interpretation research should be explored alongside medical standards

Model interpretation research is currently siloed within the machine learning community, leading to methods that may not be convincing for the target audience, medical professionals and patients. While research has shown that explanations increase human confidence in model predictions[142], it is unknown what type or format of explanations are required before clinical adoption becomes widespread. Is it enough to mimic how a clinician conducts image analysis? Is it acceptable if model interpretations become indistinguishable from a clinician's? Or would the medical community require interpretations that go beyond the capabilities of a standard clinician and reveal otherwise unattainable insights? In general, what is needed for non-technical experts to "trust" the predictions of a machine learning model? The answers to these questions are unclear but would have significant impact in shaping future research directions towards a well-defined end goal. As evident in the inability of interpretation techniques to both create image-specific interpretations and identify medically-relevant visual features, ignoring the end goal might result in methods that do not fully convince medical experts. Though emulating the image analysis process of a clinician is a promising direction, this path requires constant dialogue to understand what is needed for clinical adoption.

### 5.4. Continual dialogue between medical and machine learning communities accelerates improvement of data-driven models

The model interpretability issue is applicable across visual-based medical tasks involving deep learning such as radiological diagnosis[155] or converting X-ray images from 2D to 3D.[156] Because the issue spans across medical fields, community-wide dialogue should drive continuous assessment of data-driven models. These assessments require centralized efforts and are not easily conducted at the individual researcher level. For example, expertly segmented and standardized datasets provides a solution towards improving image-specific interpretations. The labeled segments allow adaptation of current interpretation techniques to better emulate current dermatology. For instance, input modification methods swap a labeled segment with a different visual feature; if a benign feature is switched with a malignant feature, then the model should predict a higher likelihood of malignancy. Similarly, if a single labeled segment is split into multiple distinct features, then the asymmetry and malignancy has increased. If the model's prediction behaves unexpectedly, then the model is less trustworthy. These sophisticated questions become explorable and allow experts to query the inner model workings with counterfactual scenarios, which begins to emulate dermatology training. However, there are few standardized and widespread dermoscopy reporting formats that are both practical and amenable to model interpretation techniques. To the authors' knowledge, the International Skin Imaging Collaboration (ISIC) contains the largest publicly-available segmented dermoscopic dataset with 1142 images with labeled globules and streaks[137], only two

visual features. Dermoscopic images encompass many more features; comprehensively segmented datasets are not abundant. Clearly, lone researchers cannot solely fix this challenge, but this relatively minor change in data reporting represents potentially major improvements that are realizable through centralized efforts to explore and communicate effective directions.

A centralized organization facilitates dialogue and interdisciplinary research to improve model interpretation for clinical applications. Such an organization oversees how to standardize datasets, create a platform to share data, and host community-driven challenges to develop new methods. We take inspiration from the Dialogue for Reverse Engineering Assessments and Methods (DREAM) consortium and International Skin Imaging Collaboration (ISIC). DREAM broadly connects researchers across biomedical areas[157] while ISIC focuses on developing image analysis tools for melanoma applications.[137] Both organizations regularly host crowd-sourced challenges where researchers compete by creating models from public datasets, and winners are awarded cash prizes, invited talks, and other rewards. DREAM challenges has advanced network inference methods[158] and the prediction of human olfactory perception[159]; the 2016 ISIC challenge created a baseline for assessing various computational methods and validated the value of crowd-sourced challenges in dermoscopic imaging.[160] DREAM further incentivizes innovation by awarding prizes for high-performance models that borrow code from other researchers, which spurred cross-collaboration and model sharing. Within 24 hours, a competitor created the

top-ranked model that combined their clinical experienced with other modeling methods.[157] Overall, both DREAM and ISIC have created integrated and interdisciplinary communities that a-celerate model development.

I believe that crowd-sourced DREAM or ISIC challenges are appropriate platforms for tackling specific issues such as emulating human-level interpretations in clinical applications. Specific to dermoscopic analysis, we envision an initial challenge that explores data-reporting methods amenable to both machine learning and medical standards. It begins with multiple disparate dermoscopic datasets for competitors to merge and test for optimal standardization methods. Such a challenge requires both medical and machine learning experts to collaborate to form effective pipelines. Once an appropriate standard is established, a second challenge leverages the new protocol to identify effective model interpretation techniques. A panel of both machine learning and medical experts judge model interpretations and rate them based on correctness, clarity, and relevance to clinical practices. These tiered challenges build upon previous successes and has been similarly applied to develop network inference methods.[158,161,162]

Solving the overall dialogue gap between machine learning and clinical experts requires a greater investment than a series of crowd-sourced challenges. This dialogue gap has been prevalent for some time and is inherent in sharing and using medical data [33–36].[163–166] One potential solution is creating an organization to tackle model interpretation in clinical tasks. As with DREAM, this organization provides a platform for the exchange of data and is especially cognizant of data issues specific

in medical domains such as legality, ethics, and confidentiality.[163] This idea is not a unique solution and has been proposed for issues such as sharing functional MRI data[165], trauma support networks[166], and clinical trial data sharing.[163,164] These varied proposals suggest that the dialogue gap stems from a common problem: few centralized medical data organizations exist. The model interpretability challenge has become one of a growing list of reasons to establish such an organization.

## 5.5. Summarizing the dialogue gap between machine learning and medicine

Data-driven models, especially deep learning models, offer exciting opportunities across medical imaging applications. Though research into model interpretation techniques is growing, their applications remain limited by an inability to emulate human-level interpretations on medical imaging tasks. Improved dialogue between the medical and machine learning communities should help align interpretation techniques with the end goal of clinical adoption. Standardization of protocols, such as in data-reporting, has the potential to expand machine learning capabilities and make current model interpretation techniques more useful. These solutions require centralized efforts to explore and implement effective methods.

## 5.6. Big data and data science is the future of scientific advancement

Big data offers exciting opportunities in finally unlocking the complexities in biology systems and engineering solutions to many of life's problems. Though large datasets themselves present few answers and often create larger haystacks when searching for

the needle, I am fully confident that the development of future scientists will further modern data science tools to efficiently tackle this problem. My work here is predicated upon decades of focused work from seemingly disparate areas that only recently became connected by the common theme of big data. If recent developments are any indication, big data generation techniques and data science tools are accelerating in popularity and I expect this dissertation to become rapidly eclipsed. I believe I have only scratched the surface of learning and applying data science tools, and I have a bright future ahead of me.

# References


1. Dante A Pertusi, Andrew E Stine, Linda J Broadbelt, and Keith EJ Tyo. Efficient searching and annotation of metabolic networks using chemical similarity. *Bioinformatics*, 31(7):1016–1024, 2014.

2. Monya Baker. Metabolomics: from small molecules to big ideas, 2011.

3. Aisling O'Driscoll, Jurate Daugelaite, and Roy D Sleator. 'big data', hadoop and cloud computing in genomics. *Journal of biomedical informatics*, 46(5):774–781, 2013.

4. Doug Howe, Maria Costanzo, Petra Fey, Takashi Gojobori, Linda Hannick, Winston Hide, David P Hill, Renate Kania, Mary Schaeffer, Susan St Pierre, et al. Big data: The future of biocuration. *Nature*, 455(7209):47, 2008.

5. Travis B Murdoch and Allan S Detsky. The inevitable application of big data to health care. *Jama*, 309(13):1351–1352, 2013.

6. Muin J Khoury and John PA Ioannidis. Big data meets public health. *Science*, 346(6213):1054–1055, 2014.

7. Vivien Marx. Biology: The big challenges of big data, 2013.

8. Hsin-Yu Kuo, Teresa A DeLuca, William M Miller, and Milan Mrksich. Profiling deacetylase activities in cell lysates with peptide arrays and samdi mass spectrometry. *Analytical chemistry*, 85(22):10635–10642, 2013.

9. Josep Villanueva, John Philip, David Entenberg, Carlos A Chaparro, Meena K Tanwar, Eric C Holland, and Paul Tempst. Serum peptide profiling by magnetic particle-assisted, automated sample processing and maldi-tof mass spectrometry. *Analytical chemistry*, 76(6):1560–1570, 2004.

10. CR Jimenez, KW Li, K Dreisewerd, S Spijker, R Kingston, RH Bateman, AL Burlingame, AB Smit, J Van Minnen, and WPM Geraerts. Direct mass

spectrometric peptide profiling and sequencing of single neurons reveals differential peptide patterns in a small neuronal network. *Biochemistry*, 37(7):2070–2076, 1998.

11. Ekaterina Olshannikova, Aleksandr Ometov, Yevgeni Koucheryavy, and Thomas Olsson. Visualizing big data. In *Big Data Technologies and Applications*, pages 101–131. Springer, 2016.

12. Jeong-Heon Lee, Suzanne RL Hart, and David G Skalnik. Histone deacetylase activity is required for embryonic stem cell differentiation. *Genesis*, 38(1):32–38, 2004.

13. Jeffrey LeBlanc, Matthew O Ward, and Norman Wittels. Exploring n-dimensional databases. In *Visualization, 1990. Visualization'90., Proceedings of the First IEEE Conference on*, pages 230–237. IEEE, 1990.

14. Gerrit Schüürmann, Ralf-Uwe Ebert, Jingwen Chen, Bin Wang, and Ralph Kühne. External validation and prediction employing the predictive squared correlation coefficient test set activity mean vs training set activity mean. *Journal of Chemical Information and Modeling*, 48(11):2140–2145, 2008.

15. Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

16. Jac MMJG Aarts, Si Wang, René Houtman, Rinie MGJ van Beuningen, Walter MA Westerink, Beppy J Van De Waart, Ivonne MCM Rietjens, and Toine FH Bovee. Robust array-based coregulator binding assay predicting er$\alpha$-agonist potency and generating binding profiles reflecting ligand structure. *Chemical research in toxicology*, 26(3):336–346, 2013.

17. Norie Murayama, Rinie van Beuningen, Hiroshi Suemizu, Christiane Guguen-Guillouzo, Norio Shibata, Kanako Yajima, Masahiro Utoh, Makiko Shimizu, Christophe Chesné, Masato Nakamura, et al. Thalidomide increases human hepatic cytochrome p450 3a enzymes by direct activation of the pregnane x receptor. *Chemical research in toxicology*, 27(2):304–308, 2014.

18. Burcu Ayoglu, Nicholas Mitsios, Ingrid Kockum, Mohsen Khademi, Arash Zandian, Ronald Sjöberg, Björn Forsström, Johan Bredenberg, Izaura Lima Bomfim, Erik Holmgren, et al. Anoctamin 2 identified as an autoimmune target in multiple sclerosis. *Proceedings of the National Academy of Sciences*, 113(8):2188–2193, 2016.

19. Chenggang Wu, Mike Haiting Ma, Kevin R Brown, Matt Geisler, Lei Li, Eve Tzeng, Christina YH Jia, Igor Jurisica, and Shawn S-C Li. Systematic identification of sh3 domain-mediated human protein–protein interactions by peptide array target screening. *Proteomics*, 7(11):1775–1785, 2007.

20. Ulrich Reineke, Claudia Ivascu, Marén Schlief, Christiane Landgraf, Seike Gericke, Grit Zahn, Hanspeter Herzel, Rudolf Volkmer-Engert, and Jens Schneider-Mergener. Identification of distinct antibody epitopes and mimotopes from a peptide array of 5520 randomly generated sequences. *Journal of immunological methods*, 267(1):37–51, 2002.

21. Alona P Umali, Sarah E LeBoeuf, Robert W Newberry, Siwon Kim, Lee Tran, Whitney A Rome, Tian Tian, David Taing, Jane Hong, Melissa Kwan, et al. Discrimination of flavonoids and red wine varietals by arrays of differential peptidic sensors. *Chemical Science*, 2(3):439–445, 2011.

22. Eman Ghanem, Saazina Afsah, Parisa N Fallah, Alexandria Lawrence, Elise LeBovidge, Sneha Raghunathan, Dominic Rago, Michelle A Ramirez, Mitchell Telles, Michelle Winkler, et al. Differentiation and identification of cachaça wood extracts using peptide-based receptors and multivariate data analysis. *ACS sensors*, 2(5):641–647, 2017.

23. Blake Farrow, Sung A Hong, Errika C Romero, Bert Lai, Matthew B Coppock, Kaycie M Deyle, Amethist S Finch, Dimitra N Stratis-Cullum, Heather D Agnew, Sung Yang, et al. A chemically synthesized capture agent enables the selective, sensitive, and robust electrochemical detection of anthrax protective antigen. *ACS nano*, 7(10):9452–9460, 2013.

24. Jessica A Pfeilsticker, Aiko Umeda, Blake Farrow, Connie L Hsueh, Kaycie M Deyle, Jocelyn T Kim, Bert T Lai, and James R Heath. A cocktail of thermally stable, chemically synthesized capture agents for the efficient detection of anti-gp41 antibodies from human sera. *PLoS One*, 8(10):e76224, 2013.

25. Marie-Laure Lesaicherre, Mahesh Uttamchandani, Grace YJ Chen, and Shao Q Yao. Developing site-specific immobilization strategies of peptides in a microarray. *Bioorganic & Medicinal Chemistry Letters*, 12(16):2079–2083, 2002.

26. Mizuki Takahashi, Kiyoshi Nokihara, and Hisakazu Mihara. Construction of a protein-detection system using a loop peptide library with a fluorescence label. *Chemistry & biology*, 10(1):53–60, 2003.

27. Matt Kaeberlein, Thomas McDonagh, Birgit Heltweg, Jeffrey Hixon, Eric A Westman, Seth D Caldwell, Andrew Napper, Rory Curtis, Peter S DiStefano, Stanley Fields, et al. Substrate-specific activation of sirtuins by resveratrol. *Journal of Biological Chemistry*, 280(17):17038–17045, 2005.

28. Zachary A Gurard-Levin, Kristopher A Kilian, Joohoon Kim, Katinka Bähr, and Milan Mrksich. Peptide arrays identify isoform-selective substrates for profiling endogenous lysine deacetylase activity. *ACS chemical biology*, 5(9):863–873, 2010.

29. Lan Ban, Nicholas Pettit, Lei Li, Andreea D Stuparu, Li Cai, Wenlan Chen, Wanyi Guan, Weiqing Han, Peng George Wang, and Milan Mrksich. Discovery of glycosyltransferases using carbohydrate arrays and mass spectrometry. *Nature chemical biology*, 8(9):769, 2012.

30. Zachary A Gurard-Levin, Michael D Scholle, Adam H Eisenberg, and Milan Mrksich. High-throughput screening of small molecule libraries using samdi mass spectrometry. *ACS combinatorial science*, 13(4):347–350, 2011.

31. Katalin F Medzihradszky, Jennifer M Campbell, Michael A Baldwin, Arnold M Falick, Peter Juhasz, Marvin L Vestal, and Alma L Burlingame. The characteristics of peptide collision-induced dissociation using a high-performance maldi-tof/tof tandem mass spectrometer. *Analytical chemistry*, 72(3):552–558, 2000.

32. Renato Zenobi and Richard Knochenmuss. Ion formation in maldi mass spectrometry. *Mass spectrometry reviews*, 17(5):337–366, 1998.

33. Eberhard Krause, Holger Wenschuh, and Peter R Jungblut. The dominance of arginine-containing peptides in maldi-derived tryptic mass fingerprints of proteins. *Analytical chemistry*, 71(19):4160–4165, 1999.

34. Mari-Luz Valero, Ernest Giralt, and David Andreu. An investigation of residue-specific contributions to peptide desorption in maldi-tof mass spectrometry. *Letters in Peptide Science*, 6(2-3):109–115, 1999.

35. Victor Ryzhov and Catherine Fenselau. Characterization of the protein subset desorbed by maldi from whole bacterial cells. *Analytical chemistry*, 73(4):746–750, 2001.

36. Anna Pashkova, Eugene Moskovets, and Barry L Karger. Coumarin tags for improved analysis of peptides by maldi-tof ms and ms/ms. 1. enhancement in maldi ms signal intensities. *Analytical chemistry*, 76(15):4550–4557, 2004.

37. Ronald C Beavis and John N Bridson. Epitaxial protein inclusion in sinapic acid crystals. *Journal of Physics D: Applied Physics*, 26(3):442, 1993.

38. Francisco ML Amado, P Domingues, M Graça Santana-Marques, AJ Ferrer-Correia, and KB Tomer. Discrimination effects and sensitivity variations in matrix-assisted laser desorption/ionization. *Rapid communications in mass spectrometry*, 11(12):1347–1352, 1997.

39. Sabine Baumgart, Yvonne Lindner, Ronald Kühne, Axel Oberemm, Holger Wenschuh, and Eberhard Krause. The contributions of specific amino acid side chains to signal intensities of peptides in matrix-assisted laser desorption/ionization mass spectrometry. *Rapid communications in mass spectrometry*, 18(8):863–868, 2004.

40. Kathrin Stavenhagen, Hannes Hinneburg, Morten Thaysen-Andersen, Laura Hartmann, Daniel Varón Silva, Jens Fuchser, Stephanie Kaspar, Erdmann Rapp, Peter H Seeberger, and Daniel Kolarich. Quantitative mapping of glycoprotein micro-heterogeneity and macro-heterogeneity: an evaluation of mass spectrometry signal strengths using synthetic peptides and glycopeptides. *Journal of Mass Spectrometry*, 48(6):627–639, 2013.

41. Joseph Barten Legutki, Zhan-Gong Zhao, Matt Greving, Neal Woodbury, Stephen Albert Johnston, and Phillip Stafford. Scalable high-density peptide arrays for comprehensive health monitoring. *Nature communications*, 5:4785, 2014.

42. ER Stadtman and RL Levine. Free radical-mediated oxidation of free amino acids and amino acid residues in proteins. *Amino acids*, 25(3-4):207–218, 2003.

43. Steven Gay, Pierre-Alain Binz, Denis F Hochstrasser, and Ron D Appel. Peptide mass fingerprinting peak intensity prediction: extracting knowledge from spectra. *Proteomics*, 2(10):1374–1391, 2002.

44. Vincent A Fusaro, DR Mani, Jill P Mesirov, and Steven A Carr. Prediction of high-responding peptides for targeted protein assays by mass spectrometry. *Nature biotechnology*, 27(2):190, 2009.

45. William S Sanders, Susan M Bridges, Fiona M McCarthy, Bindu Nanduri, and Shane C Burgess. Prediction of peptides observable by mass spectrometry applied at the experimental set level. In *BMC bioinformatics*, volume 8, page S23. BioMed Central, 2007.

46. Christian Scheler, Stephanie Lamer, Zaoming Pan, Xin-Ping Li, Johannes Salnikow, and Peter Jungblut. Peptide mass fingerprint sequence coverage from differently stained proteins on two-dimensional electrophoresis patterns by matrix assisted laser desorption/ionization-mass spectrometry (maldi-ms). *Electrophoresis*, 19(6):918–927, 1998.

47. Jeffrey C Silva, Marc V Gorenstein, Guo-Zhong Li, Johannes PC Vissers, and Scott J Geromanos. Absolute quantification of proteins by lcmse a virtue of parallel ms acquisition. *Molecular & Cellular Proteomics*, 5(1):144–156, 2006.

48. Vladimir Svetnik, Andy Liaw, Christopher Tong, J Christopher Culberson, Robert P Sheridan, and Bradley P Feuston. Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences*, 43(6):1947–1958, 2003.

49. HU Mei, Zhi H Liao, Yuan Zhou, and Shengshi Z Li. A new set of amino acid descriptors and its application in peptide qsars. *Peptide Science*, 80(6):775–786, 2005.

50. IBC Matheson and John Lee. Chemical reaction rates of amino acids with singlet oxygen. *Photochemistry and photobiology*, 29(5):879–881, 1979.

51. Christian Schöneich. Methionine oxidation by reactive oxygen species: reaction mechanisms and relevance to alzheimer's disease. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1703(2):111–119, 2005.

52. Krzysztof Bobrowski and Christian Schdneich. Hydroxyl radical adduct at sulfur in substituted organic sulfides stabilized by internal hydrogen bond. *Journal of the Chemical Society, Chemical Communications*, pages 795–797, 1993.

53. Alessandro Lusci, Gianluca Pollastri, and Pierre Baldi. Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. *Journal of chemical information and modeling*, 53(7):1563–1575, 2013.

54. Walter Cedeño and Dimitris K Agrafiotis. Using particle swarms for the development of qsar models based on k-nearest neighbor and kernel regression. *Journal of computer-aided molecular design*, 17(2-4):255–263, 2003.

55. Kevin A Janes, Jason R Kelly, Suzanne Gaudet, John G Albeck, Peter K Sorger, and Douglas A Lauffenburger. Cue-signal-response analysis of tnf-induced apoptosis by partial least squares regression of dynamic multivariate data. *Journal of Computational Biology*, 11(4):544–561, 2004.

56. Daniel Bobo, Kye J Robinson, Jiaul Islam, Kristofer J Thurecht, and Simon R Corrie. Nanoparticle-based medicines: a review of fda-approved materials and clinical trials to date. *Pharmaceutical research*, 33(10):2373–2387, 2016.

57. Chad A Mirkin, Robert L Letsinger, Robert C Mucic, and James J Storhoff. A dna-based method for rationally assembling nanoparticles into macroscopic materials. *Nature*, 382(6592):607, 1996.

58. Joshua I Cutler, Evelyn Auyeung, and Chad A Mirkin. Spherical nucleic acids. *Journal of the American Chemical Society*, 134(3):1376–1391, 2012.

59. Chung Hang J Choi, Liangliang Hao, Suguna P Narayan, Evelyn Auyeung, and Chad A Mirkin. Mechanism for the endocytosis of spherical nucleic acid nanoparticle conjugates. *Proceedings of the National Academy of Sciences*, 110(19):7625–7630, 2013.

60. Aleksandar F Radovic-Moreno, Natalia Chernyak, Christopher C Mader, Subbarao Nallagatla, Richard S Kang, Liangliang Hao, David A Walker, Tiffany L Halo, Timothy J Merkel, Clayton H Rische, et al. Immunomodulatory spherical nucleic acids. *Proceedings of the National Academy of Sciences*, 112(13):3892–3897, 2015.

61. Nathaniel L Rosi, David A Giljohann, C Shad Thaxton, Abigail KR Lytton-Jean, Min Su Han, and Chad A Mirkin. Oligonucleotide-modified gold nanoparticles for intracellular gene regulation. *Science*, 312(5776):1027–1030, 2006.

62. Dwight S Seferos, Andrew E Prigodich, David A Giljohann, Pinal C Patel, and Chad A Mirkin. Polyvalent dna nanoparticle conjugates stabilize nucleic acids. *Nano letters*, 9(1):308–311, 2008.

63. Jing Li, Xuling Wang, Ting Zhang, Chunling Wang, Zhenjun Huang, Xiang Luo, and Yihui Deng. A review on phospholipids and their main applications in

drug delivery systems. *Asian journal of pharmaceutical sciences*, 10(2):81–98, 2015.

64. Alan J Schroit, J Madsen, and R Nayar. Liposome-cell interactions: in vitro discrimination of uptake mechanism and in vivo targeting strategies to mononuclear phagocytes. *Chemistry and physics of lipids*, 40(2-4):373–393, 1986.

65. Sergio Simoes, Vladimir Slepushkin, Nejat Düzgünes, and Maria C Pedroso de Lima. On the mechanisms of internalization and intracellular delivery mediated by ph-sensitive liposomes. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1515(1):23–37, 2001.

66. Michael J McCluskie and Heather L Davis. Cpg dna as mucosal adjuvant. *Vaccine*, 18(3-4):231–237, 1999.

67. Arthur M Krieg, Ae-Kyung Yi, Sara Matson, Thomas J Waldschmidt, Gail A Bishop, Rebecca Teasdale, Gary A Koretzky, and Dennis M Klinman. Cpg motifs in bacterial dna trigger direct b-cell activation. *Nature*, 374(6522):546, 1995.

68. H Hemmi, O Takeuchi, T Kawai, T Kaisho, S Sato, H Sanjo, M Matsumoto, K Hoshino, H Wagner, K Takeda, et al. A toll-like receptor recognizes bacterial dna. *Nature*, 408(6813):740–745, 2000.

69. Qiuyan Zhao, Jamal Temsamani, Patricia L Iadarola, Zhiwei Jiang, and Sudhir Agrawal. Effect of different chemically modified oligodeoxynucleotides on immune stimulation. *Biochemical pharmacology*, 51(2):173–182, 1996.

70. David A Giljohann, Dwight S Seferos, Pinal C Patel, Jill E Millstone, Nathaniel L Rosi, and Chad A Mirkin. Oligonucleotide loading determines cellular uptake of dna-modified gold nanoparticles. *Nano letters*, 7(12):3818–3821, 2007.

71. Andrew E Prigodich, Ali H Alhasan, and Chad A Mirkin. Selective enhancement of nucleases by polyvalent dna-functionalized gold nanoparticles. *Journal of the American Chemical Society*, 133(7):2120–2123, 2011.

72. Kristin B Gendron, Alex Rodriguez, and Duane A Sewell. Vaccination with human papillomavirus type 16 e7 peptide with cpg oligonucleotides for prevention of tumor growth in mice. *Archives of Otolaryngology–Head & Neck Surgery*, 132(3):327–332, 2006.

73. Eric J Berns, Maria D Cabezas, and Milan Mrksich. Cellular assays with a molecular endpoint measured by samdi mass spectrometry. *Small*, 12(28):3811–3818, 2016.

74. Dal-Hee Min, Wei-Jen Tang, and Milan Mrksich. Chemical screening by mass spectrometry to identify inhibitors of anthrax lethal factor. *Nature biotechnology*, 22(6):717, 2004.

75. Milan Mrksich. Mass spectrometry of self-assembled monolayers: a new tool for molecular surface science. *Acs Nano*, 2(1):7–18, 2008.

76. Jing Su and Milan Mrksich. Using mass spectrometry to characterize self-assembled monolayers presenting peptides, proteins, and carbohydrates. *Angewandte Chemie International Edition*, 41(24):4715–4718, 2002.

77. Jing Su, Tharinda W Rajapaksha, Marcus E Peter, and Milan Mrksich. Assays of endogenous caspase activities: A comparison of mass spectrometry and fluorescence formats. *Analytical chemistry*, 78(14):4945–4951, 2006.

78. Rod Humerickhouse, Karen Lohrbach, Lin Li, William F Bosron, and M Eileen Dolan. Characterization of cpt-11 hydrolysis by human liver carboxylesterase isoforms hce-1 and hce-2. *Cancer Research*, 60(5):1189–1192, 2000.

79. Yankun Li, Robert F Schwabe, Tracie DeVries-Seimon, Pin Mei Yao, Marie-Christine Gerbod-Giannone, Alan R Tall, Roger J Davis, Richard Flavell, David A Brenner, and Ira Tabas. Free cholesterol-loaded macrophages are an abundant source of tumor necrosis factor-$\alpha$ and interleukin-6 model of nf-$\kappa$b- and map kinase-dependent inflammation in advanced atherosclerosis. *Journal of Biological Chemistry*, 280(23):21763–21772, 2005.

80. Dong Yu, Qiuyan Zhao, Ekambar R Kandimalla, and Sudhir Agrawal. Accessible 5 -end of cpg-containing phosphorothioate oligodeoxynucleotides is essential for immunostimulatory activity. *Bioorganic & medicinal chemistry letters*, 10(23):2585–2588, 2000.

81. Ekambar R Kandimalla, Lakshmi Bhagat, Dong Yu, Yanping Cong, Jimmy Tang, and Sudhir Agrawal. Conjugation of ligands at the 5 '-end of cpg dna affects immunostimulatory activity. *Bioconjugate chemistry*, 13(5):966–974, 2002.

82. Erik De Clercq, F Eckstein, and TC Merigan. Interferon induction increased through chemical modification of a synthetic polyribonucleotide. *Science*,

165(3898):1137–1139, 1969.

83. Tara L Roberts, Matthew J Sweet, David A Hume, and Katryn J Stacey. Cutting edge: species-specific tlr9-mediated recognition of cpg and non-cpg phosphorothioate-modified oligonucleotides. *The Journal of Immunology*, 174(2):605–608, 2005.

84. Ulrike Flierl, Tracy L Nero, Bock Lim, Jane F Arthur, Yu Yao, Stephanie M Jung, Eelo Gitz, Alice Y Pollitt, Maria TK Zaldivia, Martine Jandrot-Perrus, et al. Phosphorothioate backbone modifications of nucleotide-based drugs are potent platelet activators. *Journal of Experimental Medicine*, 212(2):129–137, 2015.

85. Scott P Henry, Greg Beattie, Grace Yeh, Alfred Chappel, Patricia Giclas, Angela Mortari, Mark A Jagels, Douglas J Kornbrust, and Arthur A Levin. Complement activation is responsible for acute toxicities in rhesus monkeys treated with a phosphorothioate oligodeoxynucleotide. *International immunopharmacology*, 2(12):1657–1666, 2002.

86. Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

87. Scott Menard. *Applied logistic regression analysis*, volume 106. Sage, 2002.

88. Akin Akinc, Andreas Zumbuehl, Michael Goldberg, Elizaveta S Leshchiner, Valentina Busini, Naushad Hossain, Sergio A Bacallado, David N Nguyen, Jason Fuller, Rene Alvarez, et al. A combinatorial library of lipid-like materials for delivery of rnai therapeutics. *Nature biotechnology*, 26(5):561, 2008.

89. Daniel G Anderson, David M Lynn, and Robert Langer. Semi-automated synthesis and screening of a large library of degradable cationic polymers for gene delivery. *Angewandte Chemie*, 115(27):3261–3266, 2003.

90. Phillip A Sharp. The centrality of rna. *Cell*, 136(4):577–580, 2009.

91. Thomas R Cech and Joan A Steitz. The noncoding rna revolution—trashing old rules to forge new ones. *Cell*, 157(1):77–94, 2014.

92. Guru Jagadeeswaran, Yun Zheng, Niranji Sumathipala, Haobo Jiang, Estela L Arrese, Jose L Soulages, Weixiong Zhang, and Ramanjulu Sunkar. Deep sequencing of small rna libraries reveals dynamic regulation of conserved and novel microrna and microrna-stars during silkworm development. *BMC genomics*, 11(1):52, 2010.

93. Kirill A Afonin, Mathias Viard, Philip Tedbury, Eckart Bindewald, Lorena Parlea, Marshall Howington, Melissa Valdman, Alizah Johns-Boehme, Cara Brainerd, Eric O Freed, et al. The use of minimal rna toeholds to trigger the activation of multiple functionalities. *Nano letters*, 16(3):1746–1753, 2016.

94. Monica P Hui, Patricia L Foley, and Joel G Belasco. Messenger rna degradation in bacterial cells. *Annual review of genetics*, 48:537–559, 2014.

95. Allison Hoynes-O'Connor, Kristina Hinman, Lukas Kirchner, and Tae Seok Moon. De novo design of heat-repressible rna thermosensors in e. coli. *Nucleic acids research*, 43(12):6166–6179, 2015.

96. V Narry Kim, Jinju Han, and Mikiko C Siomi. Biogenesis of small rnas in animals. *Nature reviews Molecular cell biology*, 10(2):126, 2009.

97. Silvi Rouskin, Meghan Zubradt, Stefan Washietl, Manolis Kellis, and Jonathan S Weissman. Genome-wide probing of rna structure reveals active unfolding of mrna structures in vivo. *Nature*, 505(7485):701, 2014.

98. Yiliang Ding, Yin Tang, Chun Kit Kwok, Yu Zhang, Philip C Bevilacqua, and Sarah M Assmann. In vivo genome-wide profiling of rna secondary structure reveals novel regulatory features. *Nature*, 505(7485):696, 2014.

99. Robert C Spitale, Ryan A Flynn, Qiangfeng Cliff Zhang, Pete Crisalli, Byron Lee, Jong-Wha Jung, Hannes Y Kuchelmeister, Pedro J Batista, Eduardo A Torre, Eric T Kool, et al. Structural imprints in vivo decode rna regulatory mechanisms. *Nature*, 519(7544):486, 2015.

100. Jason Talkish, Gemma May, Yizhu Lin, John L Woolford, and C Joel McManus. Mod-seq: high-throughput sequencing for chemical probing of rna structure. *Rna*, 20(5):713–720, 2014.

101. Sarah A Woodson. Compact intermediates in rna folding. *Annual review of biophysics*, 39:61–77, 2010.

102. Tao Pan and Tobin Sosnick. Rna folding during transcription. *Annu. Rev. Biophys. Biomol. Struct.*, 35:161–175, 2006.

103. Elizabeth A Dethoff, Jeetender Chugh, Anthony M Mustoe, and Hashim M Al-Hashimi. Functional complexity and regulation through rna dynamics. *Nature*, 482(7385):322, 2012.

104. Kathryn D Smith, Sarah V Lipchock, Tyler D Ames, Jimin Wang, Ronald R Breaker, and Scott A Strobel. Structural basis of ligand binding by a c-di-gmp riboswitch. *Nature Structural and Molecular Biology*, 16(12):1218, 2009.

105. Sanjeev Shukla and Shalini Oberdoerffer. Co-transcriptional regulation of alternative pre-mrna splicing. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1819(7):673–683, 2012.

106. Tassa Saldi, Michael A Cortazar, Ryan M Sheridan, and David L Bentley. Coupling of rna polymerase ii transcription elongation with pre-mrna splicing. *Journal of molecular biology*, 428(12):2623–2635, 2016.

107. Eric J Strobel, Kyle E Watters, Yuri Nedialkov, Irina Artsimovitch, and Julius B Lucks. Distributed biotin–streptavidin transcription roadblocks for mapping cotranscriptional rna folding. *Nucleic acids research*, 45(12):e109–e109, 2017.

108. Kyle E Watters, Eric J Strobel, M Yu Angela, John T Lis, and Julius B Lucks. Cotranscriptional folding of a riboswitch at nucleotide resolution. *Nature Structural and Molecular Biology*, 23(12):1124, 2016.

109. Julius B Lucks, Stefanie A Mortimer, Cole Trapnell, Shujun Luo, Sharon Aviran, Gary P Schroth, Lior Pachter, Jennifer A Doudna, and Adam P Arkin. Multiplexed rna structure characterization with selective 2 -hydroxyl acylation analyzed by primer extension sequencing (shape-seq). *Proceedings of the National Academy of Sciences*, 108(27):11063–11068, 2011.

110. Richard Lavery and Alberte Pullman. A new theoretical index of biochemical reactivity combining steric and electrostatic factors: An application to yeast trnaphe. *Biophysical chemistry*, 19(2):171–181, 1984.

111. Jennifer L McGinnis, Jack A Dunkle, Jamie HD Cate, and Kevin M Weeks. The mechanisms of rna shape chemistry. *Journal of the American Chemical Society*, 134(15):6617–6624, 2012.

112. Edward J Merino, Kevin A Wilkinson, Jennifer L Coughlan, and Kevin M Weeks. Rna structure analysis at single nucleotide resolution by selective 2 '-hydroxyl acylation and primer extension (shape). *Journal of the American Chemical Society*, 127(12):4223–4231, 2005.

113. Nathan A Siegfried, Steven Busan, Greggory M Rice, Julie AE Nelson, and Kevin M Weeks. Rna motif discovery by shape and mutational profiling (shape-map). *Nature methods*, 11(9):959, 2014.

114. Meghan Zubradt, Paromita Gupta, Sitara Persad, Alan M Lambowitz, Jonathan S Weissman, and Silvi Rouskin. Dms-mapseq for genome-wide or targeted rna structure probing in vivo. *Nature methods*, 14(1):75, 2017.

115. Christine Brunel and Pascale Romby. [1] probing rna structure and rna-ligand complexes with chemical probes. 2000.

116. Eckart Bindewald, Michaela Wendeler, Michal Legiewicz, Marion K Bona, Yi Wang, Mark J Pritt, Stuart FJ Le Grice, and Bruce A Shapiro. Correlating shape signatures with three-dimensional rna structures. *RNA*, 17(9):1688–1696, 2011.

117. Kady-Ann Steen, Arun Malhotra, and Kevin M Weeks. Selective 2 -hydroxyl acylation analyzed by protection from exoribonuclease. *Journal of the American Chemical Society*, 132(29):9940–9943, 2010.

118. Kyle E Watters, M Yu Angela, Eric J Strobel, Alex H Settle, and Julius B Lucks. Characterizing rna structures in vitro and in vivo with selective 2 -hydroxyl acylation analyzed by primer extension sequencing (shape-seq). *Methods*, 103:34–48, 2016.

119. Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, Franciska de Jong, and Emiel Caron. A survey of event extraction methods from text for decision support systems. *Decis. Support Syst.*, 85(C):12–22, May 2016.

120. Dragos Margineantu, Weng-Keen Wong, and Denver Dash. Machine learning algorithms for event detection. *Machine Learning*, 79(3):257–259, 2010.

121. Chanin Tolson Woods and Alain Laederach. Classification of rna structure change by 'gazing'at experimental data. *Bioinformatics*, 33(11):1647–1655, 2017.

122. Laurent Itti and Pierre Baldi. A principled approach to detecting surprising events in video. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 631–637. IEEE, 2005.

123. Kai Ye, Marcel H Schulz, Quan Long, Rolf Apweiler, and Zemin Ning. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, 25(21):2865–2871, 2009.

124. Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Jianyong Wang, Helen Pinto, Qiming Chen, Umeshwar Dayal, and Mei-Chun Hsu. Mining sequential patterns by pattern-growth: The prefixspan approach. *IEEE Transactions on knowledge and data engineering*, 16(11):1424–1440, 2004.

125. Daniel E Rivera, Manfred Morari, and Sigurd Skogestad. Internal model control: Pid controller design. *Industrial & engineering chemistry process design and development*, 25(1):252–265, 1986.

126. Nicolas Minorsky. Directional stability of automatically steered bodies. *Journal of ASNE*, 42(2):280–309, 1922.

127. David J Ketchen and Christopher L Shook. The application of cluster analysis in strategic management research: an analysis and critique. *Strategic management journal*, 17(6):441–458, 1996.

128. Marc Nerlove and Kenneth F Wallis. Use of the durbin-watson statistic in inappropriate situations. *Econometrica: Journal of the Econometric Society*, pages 235–238, 1966.

129. Robert T Batey, Robert P Rambo, Louise Lucast, Brian Rha, and Jennifer A Doudna. Crystal structure of the ribonucleoprotein core of the signal recognition particle. *Science*, 287(5456):1232–1239, 2000.

130. Terrence N Wong, Tobin R Sosnick, and Tao Pan. Folding of noncoding rnas during transcription facilitated by pausing-induced nonnative structures. *Proceedings of the National Academy of Sciences*, 104(46):17995–18000, 2007.

131. John A Jaeger, Douglas H Turner, and Michael Zuker. Improved predictions of secondary structures for rna. *Proceedings of the National Academy of Sciences*, 86(20):7706–7710, 1989.

132. Catherine Papanicolaou, Manolo Gouy, and Jacques Ninio. An energy model that predicts the correct folding of both the trna and the 5s rna molecules. 1984.

133. Aiming Ren, Kanagalaghatta R Rajashankar, and Dinshaw J Patel. Fluoride ion encapsulation by mg 2+ ions and phosphates in a fluoride riboswitch. *Nature*, 486(7401):85, 2012.

134. Bo Zhao, Sharon L Guffy, Benfeard Williams, and Qi Zhang. An excited state underlies gene regulation of a transcriptional riboswitch. *Nature chemical biology*, 13(9):968, 2017.

135. Sergey Demyanov, Rajib Chakravorty, Mani Abedini, Alan Halpern, and Rahil Garnavi. Classification of dermoscopy patterns using deep convolutional neural networks. In *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*, pages 364–368. IEEE, 2016.

136. Christoph Sinz, Philipp Tschandl, Cliff Rosendahl, Bengu Nisa Akay, Giuseppe Argenziano, Andreas Blum, Ralph P Braun, Horacio Cabo, Jean-Yves Gourhant, Juergen Kreusch, et al. Accuracy of dermatoscopy for the diagnosis of nonpigmented cancers of the skin. *Journal of the American Academy of Dermatology*, 77(6):1100–1109, 2017.

137. David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1605.01397*, 2016.

138. Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

139. Lequan Yu, Hao Chen, Qi Dou, Jing Qin, and Pheng-Ann Heng. Automated melanoma recognition in dermoscopy images via very deep residual networks. *IEEE transactions on medical imaging*, 36(4):994–1004, 2017.

140. Bright I Nwaru, Charles Friedman, John Halamka, and Aziz Sheikh. Can learning health systems help organisations deliver personalised care? *BMC medicine*, 15(1):177, 2017.

141. Paul Wicks, Matthew Hotopf, Vaibhav A Narayan, Ethan Basch, James Weatherall, and Muir Gray. It'sa long shot, but it just might work! perspectives on the future of medicine. *BMC medicine*, 14(1):176, 2016.

142. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.

143. Seonwoo Min, Byunghan Lee, and Sungroh Yoon. Deep learning in bioinformatics. *Briefings in bioinformatics*, 18(5):851–869, 2017.

144. Lisa M McShane, Margaret M Cavenagh, Tracy G Lively, David A Eberhard, William L Bigbee, P Mickey Williams, Jill P Mesirov, Mei-Yin C Polley, Kelly Y Kim, James V Tricoli, et al. Criteria for the use of omics-based predictors in clinical trials. *Nature*, 502(7471):317, 2013.

145. Susan Mallett, Patrick Royston, Rachel Waters, Susan Dutton, and Douglas G Altman. Reporting performance of prognostic models in cancer: a review. *BMC medicine*, 8(1):21, 2010.

146. Felix Grün, Christian Rupprecht, Nassir Navab, and Federico Tombari. A taxonomy and library for visualizing learned features in convolutional neural networks. *arXiv preprint arXiv:1606.07757*, 2016.

147. Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

148. Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.

149. Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

150. Ashfaq A Marghoob, Josep Malvehy, and Ralph P Braun. *An atlas of dermoscopy*. CRC Press, 2012.

151. Jonathan L Long, Ning Zhang, and Trevor Darrell. Do convnets learn correspondence? In *Advances in Neural Information Processing Systems*, pages 1601–1609, 2014.

152. Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. 2015.

153. Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017.

154. Donglai Wei, Bolei Zhou, Antonio Torrabla, and William Freeman. Understanding intra-class knowledge inside cnn. *arXiv preprint arXiv:1507.02379*, 2015.

155. Shijun Wang and Ronald M Summers. Machine learning and radiology. *Medical image analysis*, 16(5):933–951, 2012.

156. Shun Miao, Z Jane Wang, and Rui Liao. A cnn regression approach for real-time 2d/3d registration. *IEEE transactions on medical imaging*, 35(5):1352–1363, 2016.

157. Adam A Margolin, Erhan Bilal, Erich Huang, Thea C Norman, Lars Ottestad, Brigham H Mecham, Ben Sauerwine, Michael R Kellen, Lara M Mangravite, Matthew D Furia, et al. Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. *Science translational medicine*, 5(181):181re1–181re1, 2013.

158. Daniel Marbach, James C Costello, Robert Küffner, Nicole M Vega, Robert J Prill, Diogo M Camacho, Kyle R Allison, Andrej Aderhold, Richard Bonneau, Yukun Chen, et al. Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796, 2012.

159. Andreas Keller, Richard C Gerkin, Yuanfang Guan, Amit Dhurandhar, Gabor Turu, Bence Szalai, Joel D Mainland, Yusuke Ihara, Chung Wen Yu, Russ Wolfinger, et al. Predicting human olfactory perception from chemical features of odor molecules. *Science*, page eaal2014, 2017.

160. Michael A Marchetti, Noel CF Codella, Stephen W Dusza, David A Gutman, Brian Helba, Aadi Kalloo, Nabin Mishra, Cristina Carrera, M Emre Celebi, Jennifer L DeFazio, et al. Results of the 2016 international skin imaging collaboration international symposium on biomedical imaging challenge: Comparison

of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. *Journal of the American Academy of Dermatology*, 78(2):270–277, 2018.

161. Gustavo Stolovitzky, Robert J Prill, and Andrea Califano. Lessons from the dream2 challenges. *Annals of the New York Academy of Sciences*, 1158(1):159–195, 2009.

162. Pablo Meyer, Thomas Cokelaer, Deepak Chandran, Kyung Hyuk Kim, Po-Ru Loh, George Tucker, Mark Lipson, Bonnie Berger, Clemens Kreutz, Andreas Raue, et al. Network topology and parameter estimation: from experimental design methods to gene regulatory network kinetics using a community based approach. *BMC systems biology*, 8(1):13, 2014.

163. Corrado Barbui. Sharing all types of clinical data and harmonizing journal standards. *BMC medicine*, 14(1):63, 2016.

164. Nophar Geifman, Jennifer Bollyky, Sanchita Bhattacharya, and Atul J Butte. Opening clinical trial data: are the voluntary data-sharing portals enough? *BMC medicine*, 13(1):280, 2015.

165. Kristina M Visscher and Daniel H Weissman. Would the field of cognitive neuroscience be advanced by sharing functional mri data? *BMC medicine*, 9(1):34, 2011.

166. Nikolaos K Kanakaris and Peter V Giannoudis. Trauma networks: present and future challenges. *BMC medicine*, 9(1):121, 2011.

# APPENDIX A

Supporting information for Chapter 2

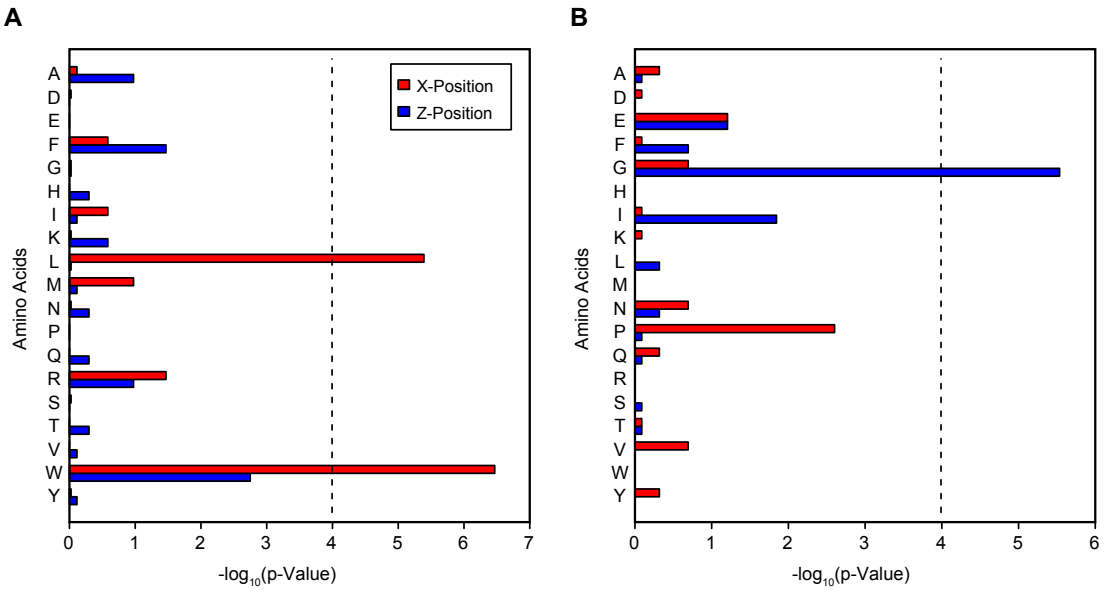



Figure A.1. **Peptide S/N is affected most by tryptophan, leucine, and glycine in K-array peptides.** The peptide's S/N were calculated from the AUC of the peptide peaks in each mass spectrum for 11 control plates containing Ac-GRK$^{ac}$XZC peptides. Peptides were identified to have low or high S/N as shown in shaded region in Figure 2.2 with 50 low and 30 high S/N peptides (See Methods). A Fischer exact test (Bonferroni corrected $p<10^{-4}$) was performed to determine general trends in peptides containing specific amino acids. The reported p-value is the chance the observed number of amino acids is within the bottom 50 or top 30 by random chance. (A) The low S/N region was enriched with peptides having X-position tryptophan, W, and leucine, L, and (B) the high S/N region was enriched with peptides having Z-position glycine, G.

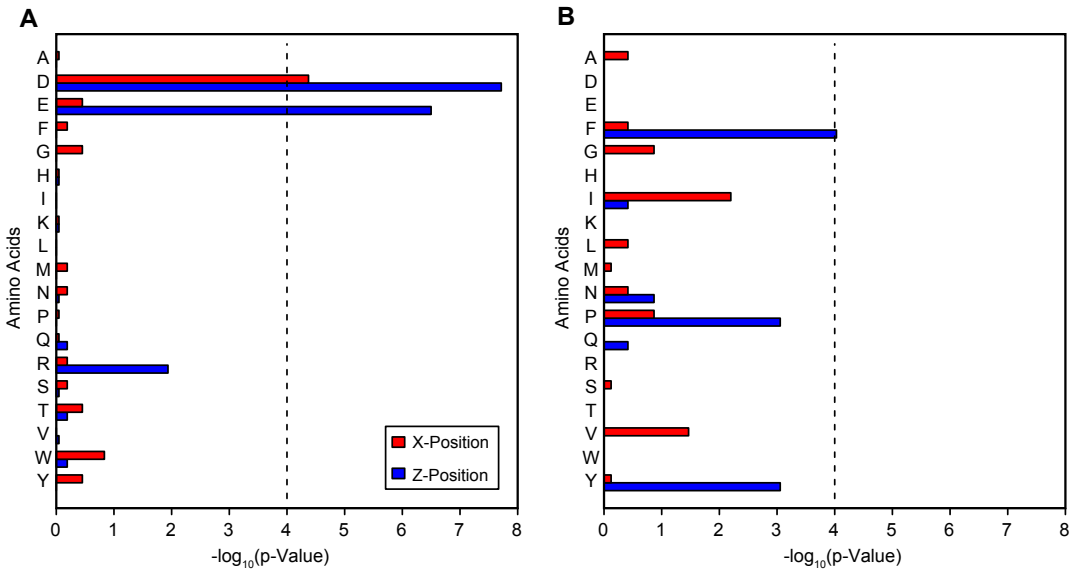
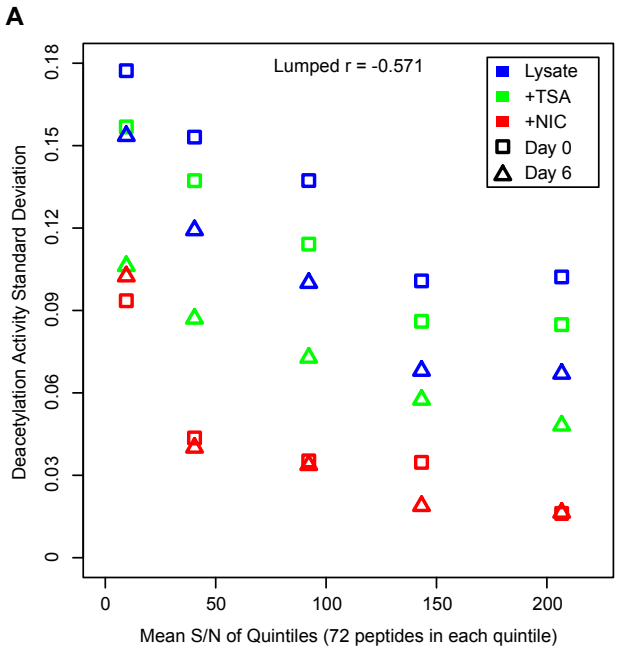



Figure A.2. **Peptide S/N is affected most by aspartic acid, glutamic acid, and phenylalanine in H-array peptides.** The peptide's S/N were calculated from the AUC of the peptide peaks in each mass spectrum for 11 control plates containing Ac-GXZHGC peptides. Peptides were identified to have low or high S/N as shown in shaded region in Figure 2.2 with 40 low and 25 high S/N peptides. A Fischer exact test (Bonferroni corrected $p<10^{-4}$) was performed to determine general trends in peptides containing specific amino acids. The reported p-value is the chance the observed number of amino acids is within the bottom 40 or top 25 by random chance. (A) The low S/N region was enriched with peptides containing X- or Z-position aspartic acid, D, and Z-position glutamic acid, E, and (B) the high S/N region was enriched with peptides having Z-position phenylalanine, F.




B

| Lysate Condition | Day | r |
|---|---|---|
| Lysate | Day 0 | -0.948 |
| Lysate | Day 6 | -0.957 |
| +TSA | Day 0 | -0.843 |
| +TSA | Day 6 | -0.942 |
| +NIC | Day 0 | -0.975 |
| +NIC | Day 6 | -0.814 |

Figure A.3. **Peptide S/N is anti-correlated with deacetylation activity standard deviation.** Data is drawn from Kuo *et al.*[8] Peptides are grouped into five quintiles (72 peptides each) based on S/N, and the standard deviation across replicates for each quintile is calculated. (A) The standard deviation of deacetylase activity versus mean S/N of each quintile is shown for the three cellular conditions of untreated lysate (blue), +TSA (green), and +NIC (red), and for day 0 (square) and day 6 (triangle). This con-sistent trend suggests that low S/N peptides have higher variations in deacetylase activity, which indicate that those peptides give unreliable data. (B) The lumped correlation (r) is the Pearson correlation over all conditions and days, and data with randomized signal S/N gives zero correlation. Within each condition and day, signal standard deviation and mean S/N is strongly anti-correlated.

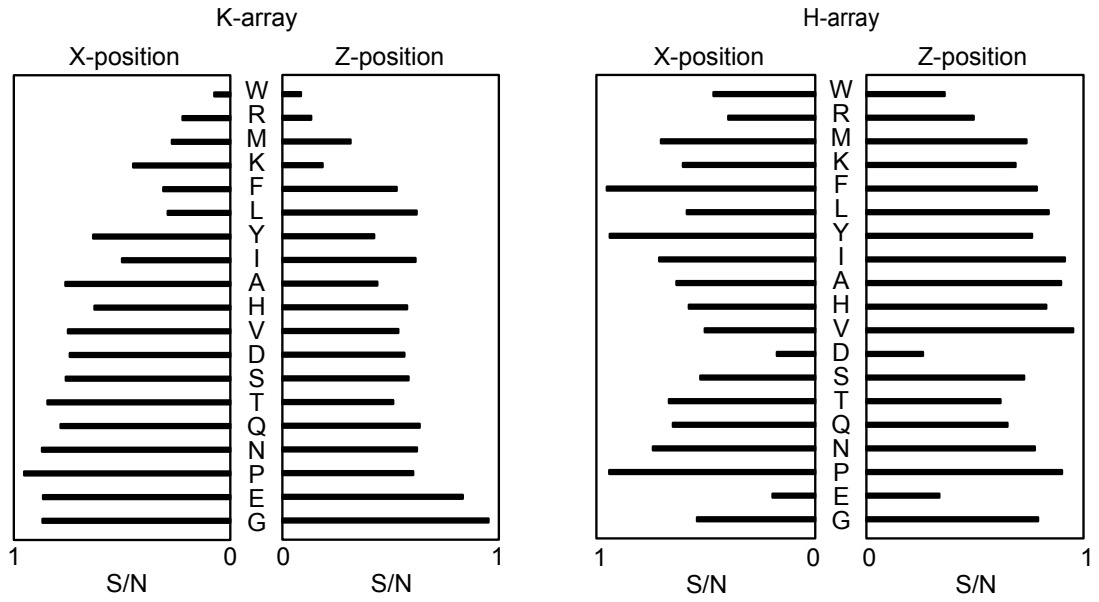



Figure A.4. **Peptide S/N stays consistent between positions, but not between peptide arrays.** Peptides with specific amino acids are sorted by their mean S/N when in either X- or Z-positions in the K-array. The two rows of bar plots are the K-array and H-array, and both have the same amino acid order. Within the K-array, S/N is correlated between the X- and Z-position amino acids. However, the correlation disappears between the H- and K-arrays, demonstrating that the surrounding amino acids have an influence on peptide S/N. Conversely, peptide with certain amino acids, such as proline, have similar S/N values on both peptide arrays which suggests that some amino acids have a consistent effect regardless of surrounding amino acids.

| | | $Q^2$ of individual properties on K-array | | |
|---|---|---|---|---|
| Property type | Physical properties | X-position | Z-position | Both positions |
| Steric | Substituent van der Waals volume | 0.224 | 0.204 | 0.568 |
| Steric | Average volume of buried residue | 0.251 | 0.189 | 0.565 |
| Steric | STERIMOL length of the side chain | 0.234 | 0.204 | 0.563 |
| Steric | Radius of gyration of side chain | 0.238 | 0.214 | 0.535 |
| Hydrophobic | Solvation free energy | 0.228 | 0.173 | 0.525 |
| Steric | Refractivity | 0.234 | 0.209 | 0.517 |
| Steric | Distance between C-alpha and centroid of side chain | 0.239 | 0.203 | 0.513 |
| Steric | Normalized van der Waals volume | 0.238 | 0.199 | 0.508 |
| Steric | STERIMOL maximum width of the side chain | 0.252 | 0.239 | 0.479 |
| Steric | van der Waals parameter epsilon | 0.244 | 0.171 | 0.465 |
| Steric | Average accessible surface area | 0.248 | 0.208 | 0.462 |
| Electronic | Isoelectric point | 0.242 | 0.209 | 0.454 |
| Hydrophobic | Retention coefficient in HPLC pH 2.1 | 0.227 | 0.221 | 0.445 |
| Steric | Residue accessible surface area in tripeptide | 0.233 | 0.211 | 0.427 |
| Steric | side chain torsion angle phi | 0.237 | 0.174 | 0.395 |
| Electronic | Electron-ion interaction potential values | 0.236 | 0.187 | 0.383 |
| Hydrophobic | Hydration number | 0.227 | 0.225 | 0.383 |
| Hydrophobic | Partition coefficient in thin-layer chromatography | 0.215 | 0.172 | 0.381 |
| Electronic | aNH chemical shifts | 0.244 | 0.221 | 0.375 |
| Hydrophobic | Melting point | 0.233 | 0.213 | 0.375 |
| Steric | Graph shape index | 0.240 | 0.212 | 0.366 |
| Electronic | pKCOOH(COOH on C_alpha) | 0.212 | 0.162 | 0.363 |
| Hydrophobic | Free energy of solution in water | 0.214 | 0.189 | 0.361 |
| Electronic | Polarity | 0.177 | 0.167 | 0.360 |
| Hydrophobic | Retention coefficient in HPLC pH 7.4 | 0.153 | 0.185 | 0.333 |
| Steric | Side-chain angle theta | 0.173 | 0.217 | 0.326 |
| Electronic | Nuclear magnetic resonance (NMR) chemical shift of αcarbon | 0.241 | 0.201 | 0.325 |
| Electronic | Amphiphilicity index | 0.114 | 0.194 | 0.314 |
| Electronic | aCH chemical shifts | 0.247 | 0.239 | 0.304 |
| Steric | van der Waals parameter Ro | 0.156 | 0.107 | 0.291 |
| Electronic | pKNH2(NH2 on C_alpha) | 0.198 | 0.181 | 0.230 |
| Electronic | Net charge | 0.040 | 0.110 | 0.170 |
| Electronic | A parameter of charge transfer donor capability | 0.080 | 0.088 | 0.144 |
| Electronic | A parameter of charge transfer capability | 0.059 | 0.083 | 0.131 |
| Steric | STERIMOL minimum width of the side chain | 0.016 | 0.046 | 0.086 |
| Hydrophobic | Number of hydrogen-bond donors | 0.027 | 0.056 | 0.085 |
| Electronic | Positive charge | 0.019 | 0.044 | 0.063 |
| Electronic | Negative charge | 0.009 | 0.024 | 0.038 |
| Electronic | Localized electrical effect | 0.029 | 0.013 | 0.016 |

Figure A.5. **Predictive power of amino acid physical properties on K-array.**

$Q^2$ of individual properties on H-array

| Property type | Physical properties | X-position | Z-position | Both positions |
|---|---|---|---|---|
| Hydrophobic | Retention coefficient in HPLC pH 2.1 | 0.156 | 0.280 | 0.544 |
| Hydrophobic | Solvation free energy | 0.167 | 0.292 | 0.535 |
| Steric | Average accessible surface area | 0.168 | 0.291 | 0.501 |
| Electronic | Isoelectric point | 0.139 | 0.304 | 0.479 |
| Hydrophobic | Hydration number | 0.132 | 0.219 | 0.450 |
| Hydrophobic | Free energy of solution in water | 0.154 | 0.303 | 0.450 |
| Steric | Substituent van der Waals volume | 0.139 | 0.274 | 0.446 |
| Electronic | Polarity | 0.172 | 0.257 | 0.436 |
| Hydrophobic | Retention coefficient in HPLC pH 7.4 | 0.153 | 0.275 | 0.431 |
| Steric | Average volume of buried residue | 0.147 | 0.310 | 0.407 |
| Electronic | Nuclear magnetic resonance (NMR) chemical shift of αcarbon | 0.151 | 0.315 | 0.406 |
| Steric | Radius of gyration of side chain | 0.119 | 0.290 | 0.400 |
| Hydrophobic | Partition coefficient in thin-layer chromatography | 0.164 | 0.313 | 0.396 |
| Hydrophobic | Melting point | 0.146 | 0.256 | 0.381 |
| Electronic | Electron-ion interaction potential values | 0.134 | 0.290 | 0.377 |
| Steric | Distance between C-alpha and centroid of side chain | 0.136 | 0.287 | 0.376 |
| Electronic | pKCOOH(COOH on C_alpha) | 0.144 | 0.287 | 0.370 |
| Steric | Refractivity | 0.143 | 0.260 | 0.367 |
| Steric | STERIMOL length of the side chain | 0.164 | 0.262 | 0.361 |
| Steric | Residue accessible surface area in tripeptide | 0.157 | 0.309 | 0.359 |
| Electronic | Localized electrical effect | 0.149 | 0.204 | 0.356 |
| Electronic | αCH chemical shifts | 0.129 | 0.244 | 0.350 |
| Steric | STERIMOL maximum width of the side chain | 0.119 | 0.302 | 0.348 |
| Steric | Normalized van der Waals volume | 0.148 | 0.297 | 0.346 |
| Electronic | aNH chemical shifts | 0.173 | 0.289 | 0.342 |
| Electronic | Net charge | 0.131 | 0.193 | 0.323 |
| Steric | side chain torsion angle phi | 0.135 | 0.262 | 0.288 |
| Steric | van der Waals parameter epsilon | 0.086 | 0.155 | 0.282 |
| Electronic | Negative charge | 0.119 | 0.173 | 0.279 |
| Steric | Graph shape index | 0.093 | 0.183 | 0.232 |
| Steric | van der Waals parameter Ro | 0.070 | 0.184 | 0.220 |
| Electronic | pKNH2(NH2 on C_alpha) | 0.085 | 0.245 | 0.206 |
| Steric | Side-chain angle theta | 0.075 | 0.176 | 0.182 |
| Electronic | Amphiphilicity index | 0.092 | 0.099 | 0.168 |
| Hydrophobic | Number of hydrogen-bond donors | 0.099 | 0.069 | 0.157 |
| Electronic | A parameter of charge transfer capability | 0.029 | 0.081 | 0.104 |
| Electronic | A parameter of charge transfer donor capability | -0.010 | 0.011 | 0.022 |
| Steric | STERIMOL minimum width of the side chain | 0.003 | -0.021 | 0.021 |
| Electronic | Positive charge | -0.011 | 0.003 | -0.009 |

Figure A.6. **Predictive power of amino acid physical properties on H-array.**

# APPENDIX B

Supporting information for Chapter 3

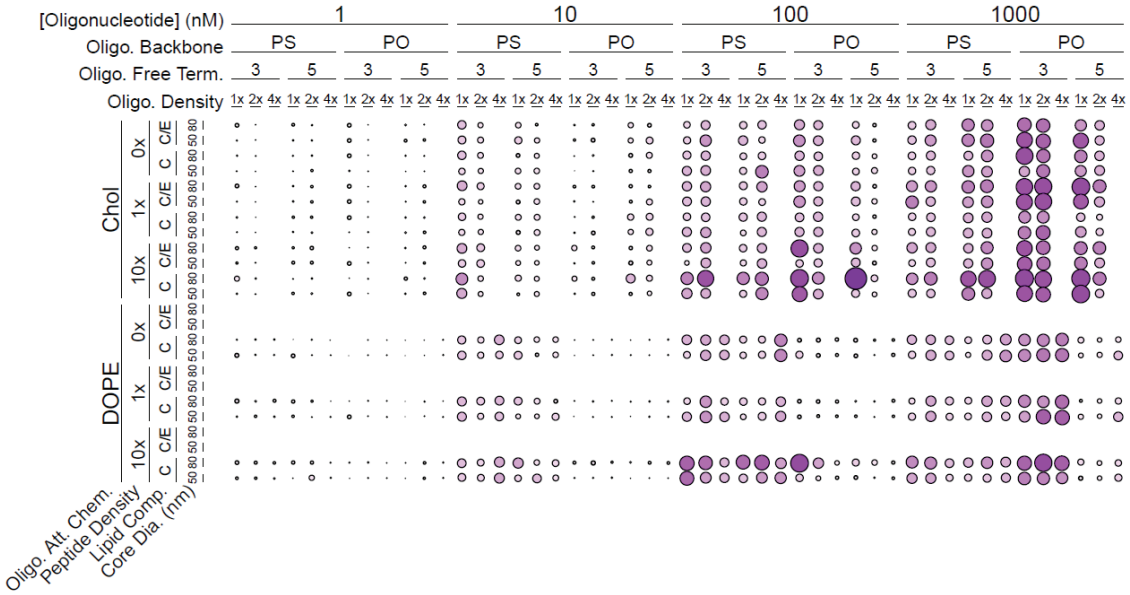



Figure B.1. **Dimensional stacking visual for encapsulated E7 subset.** A dimension-stacking plot of the active-sequence SNAs in the encapsulated E7 subset, showing the SEAP concentration for each combination of design properties. Larger and darker circles indicate greater SEAP concentration.

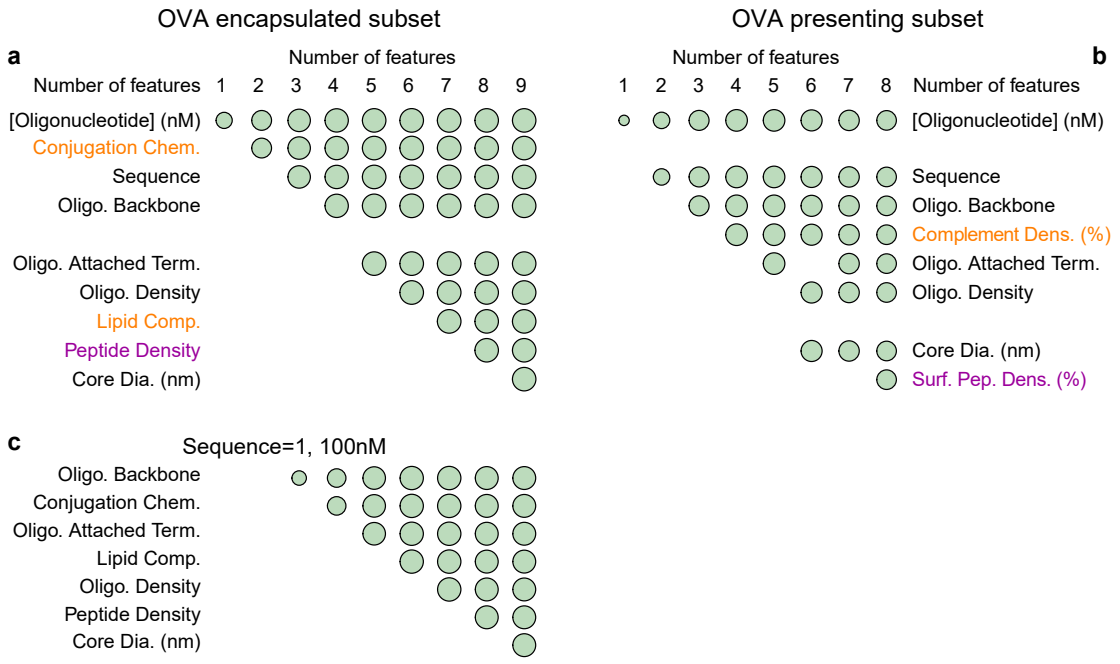



Figure B.2. **Machine learning identifies order of importance for SNA design properties.** Highest $Q^2$ scoring property combinations are shown across different number of properties for the **a**, encapsulated OVA subset, **b** surface-presenting OVA subset, and **c** encapsulated OVA subset with active sequence and 100 nM. Bubble areas correspond to $Q^2$ values from Fig. 3.6. Orange and purple properties denote exclusive and shared properties between the two subsets, respectively.

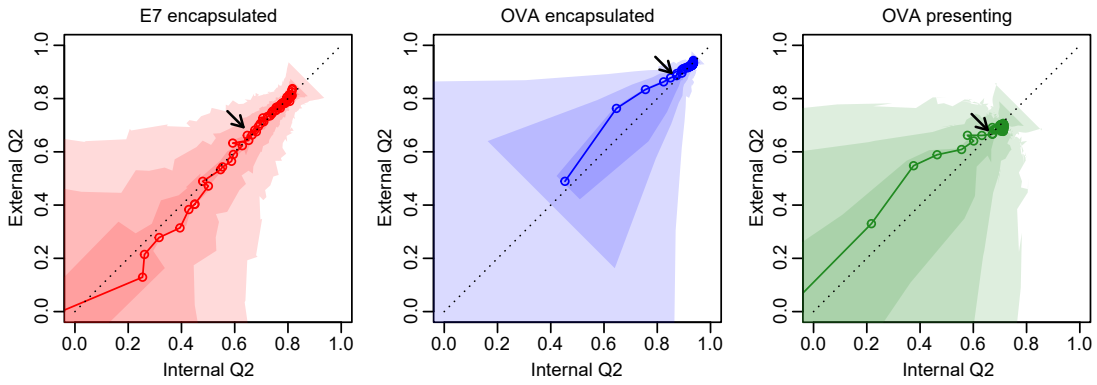



Figure B.3. **External $Q^2$ is highly correlated with internal $Q^2$.** The non-observable external $Q^2$ (predicting immune activity of non-synthesized SNAs from a synthesized subsample) is plotted against the observable internal $Q^2$ (cross-validating within the synthesized subsample) for all three subsets. The median line and 90%, 50% and 20% confidence intervals are shown.

| | Encapsulated OVA | | | Encapsulated E7 | | | Surface-Presented OVA | | |
|---|---|---|---|---|---|---|---|---|---|
| Factor | d.f. | F | P | d.f. | F | P | d.f. | F | P |
| Concentration | 3 | 1240 | <E-220 | 3 | 412 | 3E-220 | 3 | 183 | 5E-106 |
| Sequence | 1 | 381 | 2.0E-79 | 1 | 261 | 3.6E-56 | 1 | 246 | 1.2E-52 |
| Conj. Chem. | 1 | 338 | 4.3E-71 | 1 | 103 | 6.0E-24 | N/A | N/A | N/A |
| Backbone | 1 | 22.6 | 2.1E-06 | 1 | 3.64 | 0.056 | 1 | 241 | 8.5E-52 |
| Conj. Term. | 1 | 32.6 | 1.3E-08 | 1 | 3.34 | 0.068 | 1 | 2.73 | 0.099 |
| Oligo. Dens. | 2 | 5.59 | 0.0038 | 2 | 11.5 | 1.0E-05 | 2 | 2.23 | 0.11 |
| Antigen Dens. | 2 | 0.945 | 0.39 | 2 | 33.2 | 5.6E-15 | 2 | 0.673 | 0.51 |
| Lipid Comp. | 1 | 2.17 | 0.14 | 1 | 0.0839 | 0.77 | N/A | N/A | N/A |
| Core Diameter | 1 | 0.0248 | 0.87 | 1 | 20.4 | 6.6E-06 | 1 | 0.0218 | 0.88 |
| Comp. Dens. | N/A | N/A | N/A | N/A | N/A | N/A | 2 | 1.34 | 0.26 |

Table B.1. Multi-factor ANOVA of 3 SNA subsets

# APPENDIX C

Supporting information for Chapter 4

## C.1. Additional sensitivity analysis

### C.1.1. Window length

We explore how window lengths affect event detection. Generally, longer window lengths lead to higher sensitivity (more true positives) because reactivities from shorter/earlier transcript lengths affect the average window used in *PIR* equations. With a 100% longer window (SFig. C.3 right), this effect is shown with base 14, resulting in detected upswings. As a tradeoff, spurious events are included such as the downswing in base 37 (SFig. C.3 right). On the other hand, a shorter window generally decreases sensitivity (SFig C.3 left). The upswings in base 40 after 75 nt are no longer steep enough to be included and the overall number of swing events has decreased.

A longer window does not globally increase event detection sensitivity as evident in false negatives for downswings in base 14 at length 87 nt (SFig C.3). This observation is due to two effects. First, the transcript lengths before 87 nt have an upward trend, and the longer window length averages over lower reactivity transcript lengths and causes the downswing to appear less significant. Second, the integral length for $I$ defaults to the same value as the window length, and longer integral lengths lower sensitivity. Instead, specifying a shorter window length (5) for the $I$ length generally increases sensitivity and recapitulates the downswings in bases 11 and 14 (SFig C.4).

This analysis demonstrates that longer windows and shorter windows generally enhance and decrease sensitivity, respectively, but scenarios exist that go against

this rule. Our detailed analysis serves as a cautionary tale that there are tradeoffs between identifying false and true events.

### C.1.2. Durbin-Watson Statistic

We provide sensitivity analysis for the Durbin-Watson statistic ($DWS$). We recommend a default setting of one as lower values traditionally correspond to a positive correlation, and we test a lenient (0.1) and a stringent threshold (1.5). As expected, the lenient and stringent thresholds have more and fewer detected ramps, respectively (Supplementary Fig. C.5). The lenient threshold allows linear ramps where the residuals are not uniformly distributed down the length of the ramp. For example, the downramp in base 14 has points clustered above and below the ramp, and we argue that this pattern does not conform to our expectations of a ramp and appears more like a downswing (it is detected as a downswing). In contrast, the upramp in base 41 appears genuine but is removed with the stringent $DWS$. Though minor, the flat region around length 50 nt creates a sequence of negative residuals that correspond to a positive autocorrelation, which fails the stringent $DWS$. However, we observe that the upramp in base 42 is preserved because it lacks a flat region as large as in base 41. These three examples showcase how the $DWS$ threshold parameter removes ramps that resemble swing events

Overall, we apply large threshold changes and most qualitatively large events are still identified in all scenarios. This demonstrates that large and clean events can be insensitive to threshold changes, and the true positive-false positive tradeoff

is applicable to qualitatively small or noisy events. Consequently, it is likely that the SHAPE-Seq event detector identifies large RNA structural events with similar accuracy as a human, but the detector can also systematically identify small events whereas a human might not. Similarly, RNA structural events are not clearly defined and encompass a degree of subjectivity meaning that tradeoffs between identifying true positives and avoiding false positives/negatives must be considered. Altogether, these complexities justify the need for our quantitative and systematic approach.

### C.2. Supplementary file descriptions

Supplementary Files are found at github.com/bagherilab/DUETT. Supplementary Files 1-3 contain the profiles for each nucleotide in SRP, fluoride-negative riboswitch, and fluoride-positive riboswitch datasets. Upswings, downswings, upramps, and downramps are shown with red diamonds, blue diamonds, red lines, and blue lines, respectively. Concurrent events are denoted with a dotted green line.

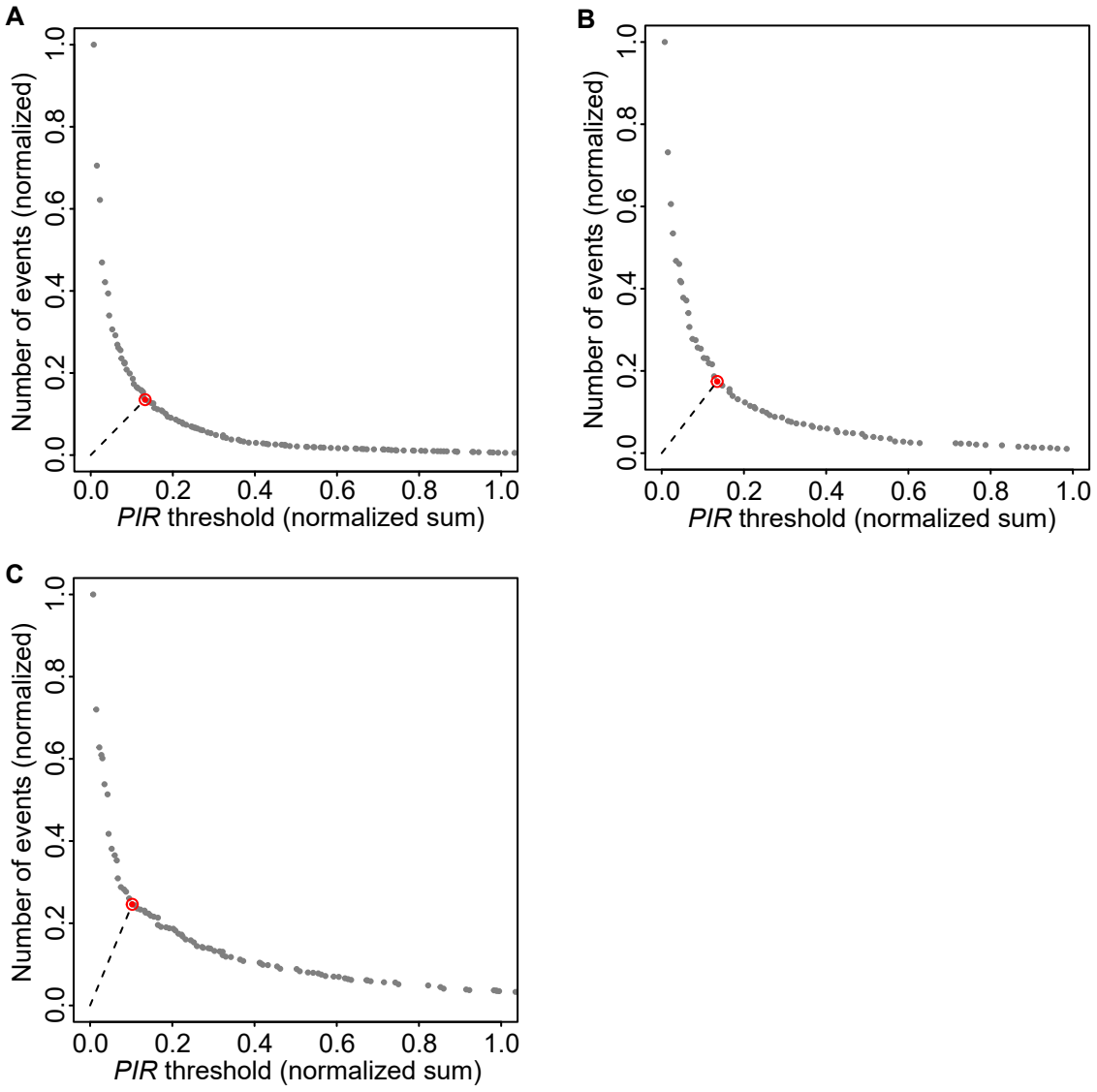



Figure C.1. **Automated *PIR* threshold selection identifies the balance between too lenient and too stringent.** DUETT provides an automated method to select appropriate *PIR* thresholds for A) the SRP, B) the 0 mM NaF riboswitch, and C) the 10 mM NaF riboswitch datasets. After scanning over potential combinations of *PIR* thresholds, DUETT identifies the threshold set closest to the origin (red point with dotted line from origin). This point corresponds to the elbow where noise is removed but real events are retained. The horizontal axis is the sum of all three *PIR* thresholds and both axes are normalized. This calculation is done in 4-dimensions ($P$, $I$, $R$, and number of events) and the selected thresholds may not appear closest to the origin when plotted in 2D.

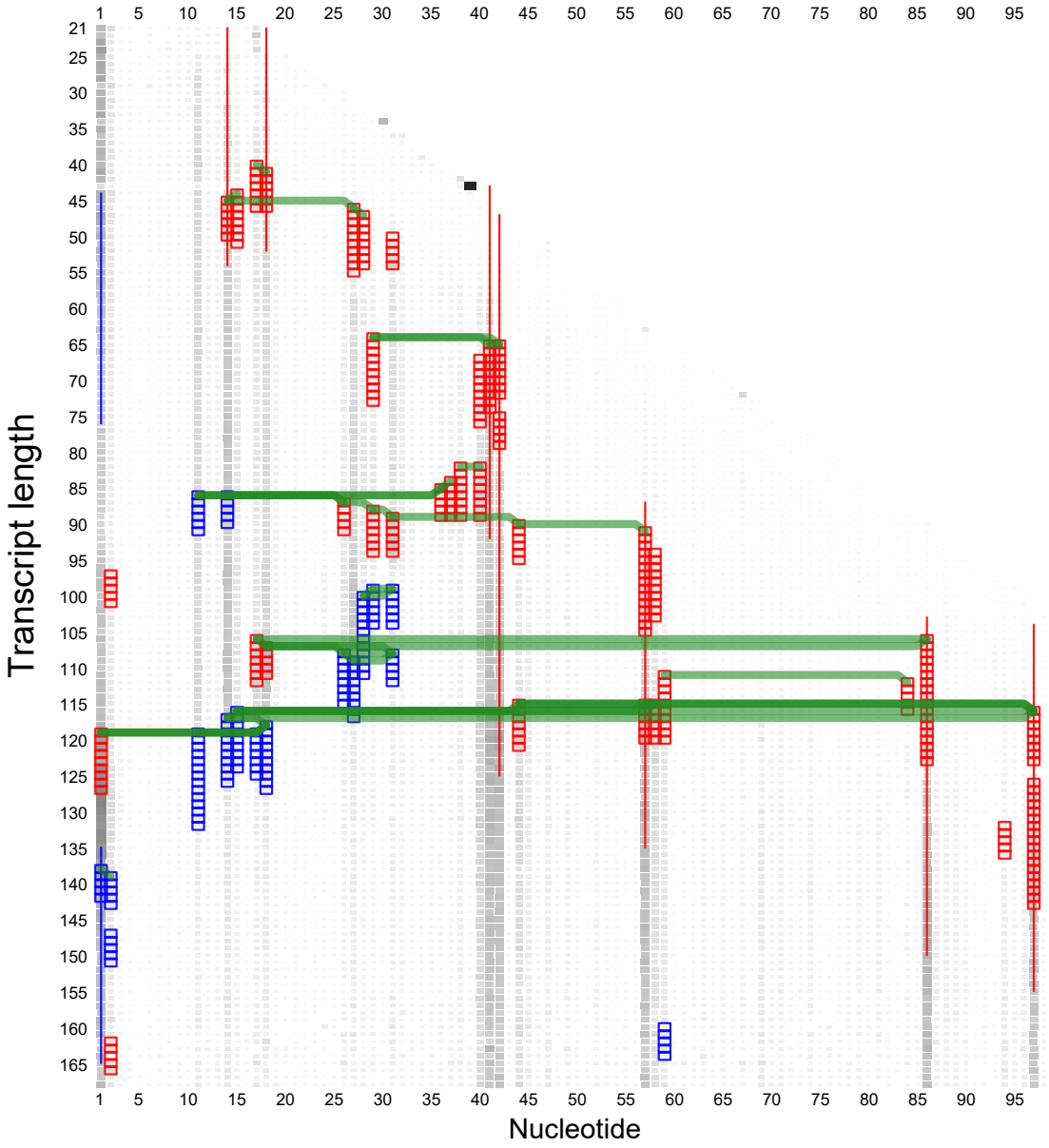



Figure C.2. **Similar results are created when applying SHAPE-Seq event detector to the average of replicates.** Figure 4.2 shows results identified events that are shared in each of the three replicates. Here, all three replicates are averaged then event detection is conducted. The *PIR* thresholds are slightly more stringent than in Figure 4.2 because few events pass all thresholds and are conserved in all replicates. Each *PIR* threshold was increased by 0.1.

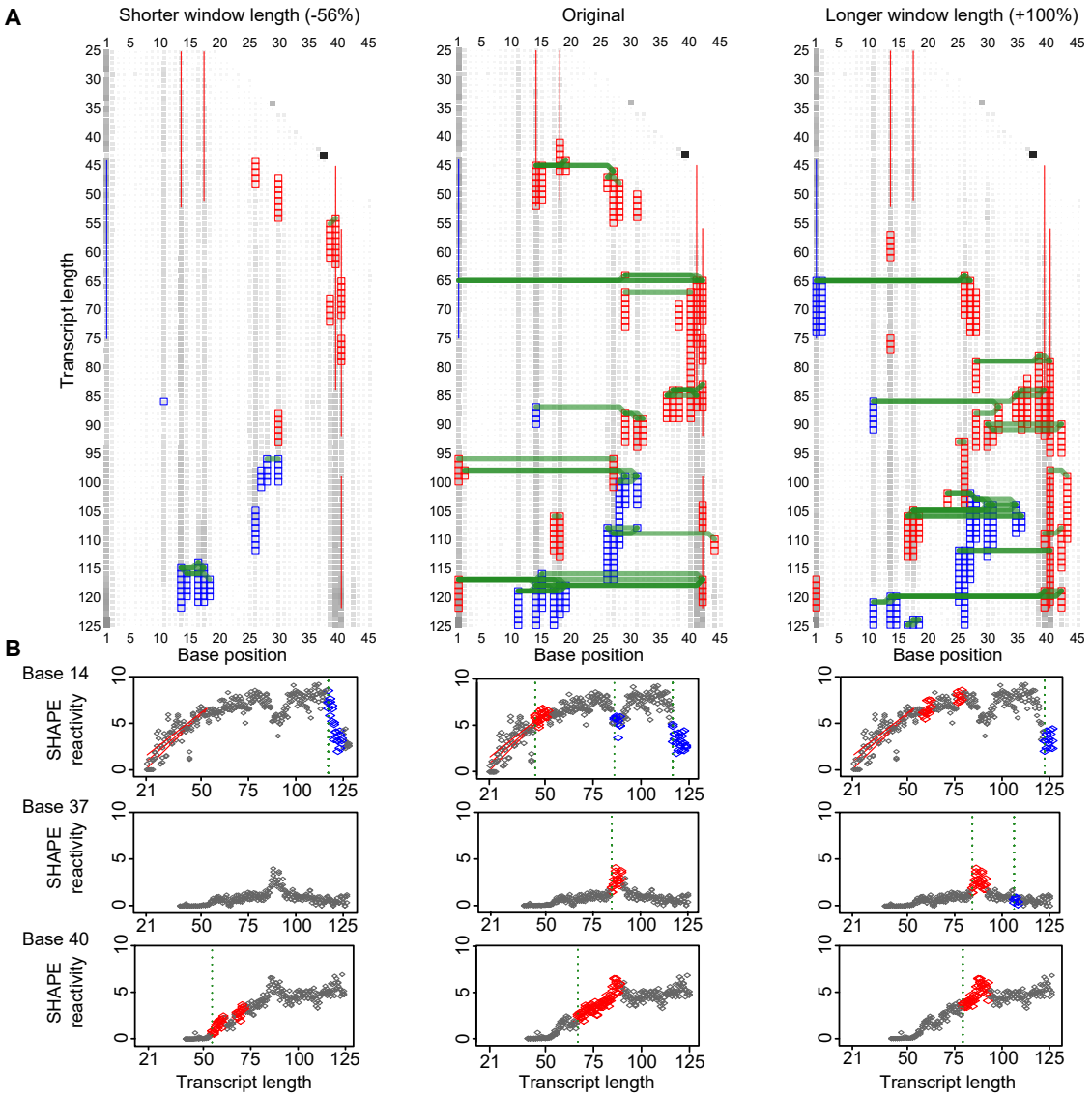



Figure C.3. **Sensitivity analysis of window length.** The window length determines the number of positions that are averaged together before *PIR* calculation. Longer lengths generally lead to longer memory and higher sensitivity to true positives. In contrast, shorter lengths generally leads to fewer events and false positives/negatives.

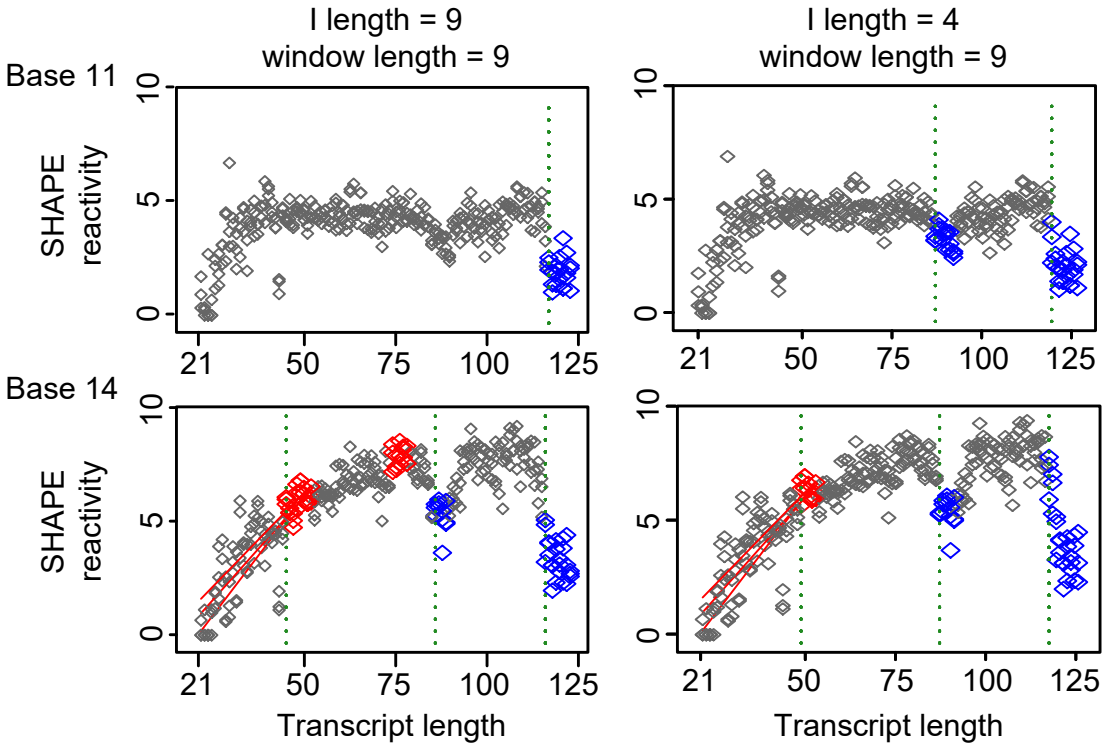



Figure C.4. **Higher $I$ length generally lowers sensitivity and longer window length does not always raise sensitivity.** By default, $I$ length is the same as the window length (left) but can be specified separately (right). Lower and higher values of $I$ length generally increase and lower sensitivity, respectively.

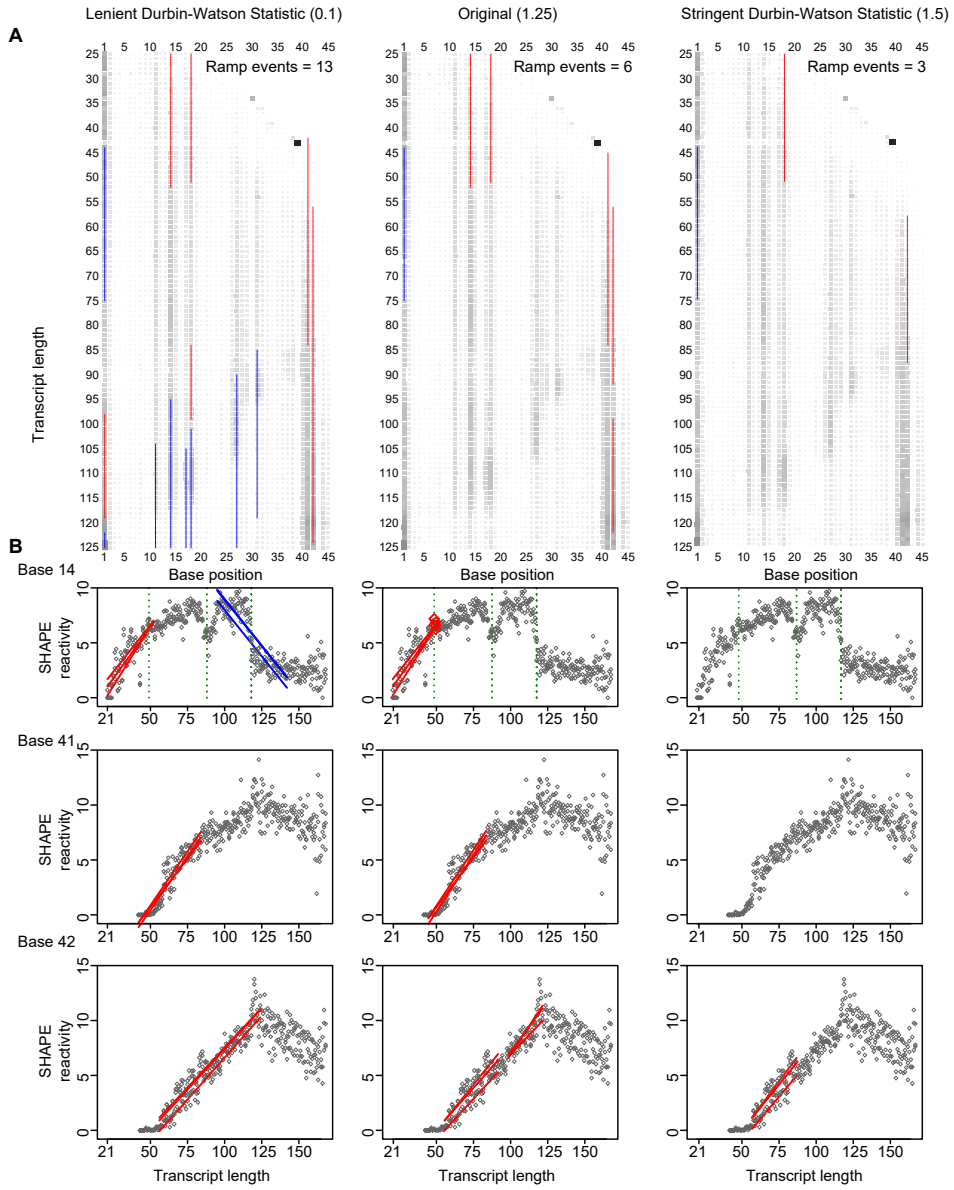



Figure C.5. **Sensitivity analysis of Durbin-Watson statistic.** $DWS$ ranges from 0-4 where 2 represents our ideal scenario, no autocorrelation in the residuals. We present effects of lower and higher $DWS$ thresholds and compare the quality of fitted linear ramps. Generally, lower $DWS$ thresholds are more lenient where lines are fitted on less line-like segments. To simplify visual analysis, all swing events were removed.

| # | Assumption | Design Implication |
|---|---|---|
| 1 | Low magnitude measurements have noise that appear large in a proportional sense | $P$ thresholds filter out low magnitude noise |
| 2 | High magnitude measurements have high magnitude noise that appear large in an absolute sense | $R$ thresholds filter out high magnitude noise |
| 3 | Short-lived swing events are likely due to noise | $I$ and event length thresholds filter out short events |
| 4 | Real swing events may become fragmented due to noise | Merge together short but adjacent swing events via event gap parameter |
| 5 | True linear ramps have a clean and low noise ramp | Ramp p-value threshold filters out noisy ramps |
| 6 | True linear ramps have a non-trivial slope | $\beta$ threshold (slope) filters out shallow ramps |
| 7 | Ramp residuals should be uniform down the length of the ramp; Ramps should be longer than swing events | Durbin-Watson Statistic threshold filters out non-uniform residuals; avoids ramps fitted on swing events |
| 8 | Events that occur nearby in terms of transcript length are likely to be related | Concurrent events are identified within a specified length |

Table C.1. Explicit assumptions and design implications in the SHAPE-Seq event detector

| **SRP RNA settings** | | **F riboswitch settings** | |
|---|---|---|---|
| PIR | | | |
| Window length | 9 | Window length | 9 |
| $P$ | 0.3 | $P$ | 0.35 |
| $I$ | 0.025 | $I$ | 0.05 |
| $R$ | 0.1 | $R$ | 0.025 |
| $I$ length | default | $I$ length | default |
| Noise length | 4 | Noise length | 4 |
| Event gap | 1 | Event gap | 1 |
| Ramp | | | |
| Ramp length | 30 | Ramp length | 30 |
| p-value | 1.00E-04 | p-value | 1.00E-04 |
| $\beta$ | 0.15 | $\beta$ | 0.15 |
| $DWS$ | 1.25 | DWS | 1.25 |
| Concurrent distance | 2 | Concurrent distance | |

Table C.2. Automated *PIR* and user-defined linear ramp threshold parameters for the SRP and riboswitch examples.