2023-12-01

Computation-Assisted Molecular Discovery For Biomedical Applications: Seeking Small Molecules And Dna Sequences With High Affinity Target Binding

Payam Kelich
University of Texas at El Paso

COMPUTATION-ASSISTED MOLECULAR DISCOVERY FOR BIOMEDICAL APPLICATIONS: SEEKING SMALL MOLECULES AND DNA SEQUENCES WITH HIGH AFFINITY TARGET BINDING

PAYAM KELICH

Doctoral Program in Chemistry

APPROVED:

Lela Vuković, Ph.D., Chair

Chu-Young Kim, Ph.D.

Wen-Yee Lee, Ph.D.

Ming-Ying Leung, Ph.D.

Stephen L. Crites, Jr., Ph.D.
Dean of the Graduate School

COMPUTATION-ASSISTED MOLECULAR DISCOVERY FOR BIOMEDICAL APPLICATIONS: SEEKING SMALL MOLECULES AND DNA SEQUENCES WITH HIGH AFFINITY TARGET BINDING

by

PAYAM KELICH, MS

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

Department of Chemistry and Biochemistry
THE UNIVERSITY OF TEXAS AT EL PASO
December 2023

Acknowledgements

I'm grateful to Dr. Lela Vukovic for her continuous encouragement and support during my research journey. I am equally thankful to my committee members, Dr. Kim, Dr. Lee, and Dr. Leung, for their guidance and insightful contributions throughout this process. I would also like to extend my gratitude to Huanhuan Zhao, Anju Yadav, Sayantani Chakraborty, Peter Girnt, Gabriela Molina Aguirre, and Dr. Parth Chaturvedi, our former lab colleagues, for their invaluable feedback and discussions. Thanks to the Texas Advanced Computing Center (TACC) for the computational facilities. Additionally, I would like to recognize the financial support from the National Institutes of Health (grant number: 1 R03 AI142553-01), the National Science Foundation (grant number: CBET-2106587), and the Department of Chemistry and Biochemistry at the College of Science, University of Texas at El Paso. I am also thankful to the College of Science for kindly providing travel funding for my visit to the ACS conference in Fall 2023.

Abstract

Binding affinity between two molecules is an essential property in drug and sensor discovery. Several computational and experimental methods exist to find molecules with high binding affinities to desired target molecules. These methods are often complementary, where fast computational methods can be used for the initial screening of molecules, and experimental methods can then screen and determine the molecules of interest and sometimes define the structures of bound complexes. After these steps, computational methods, like molecular dynamics (MD) simulations, can provide detailed insights into atomic interactions and binding, and machine learning approaches can analyze experiment-derived data to discern patterns and trends. The above computational methods were employed to tackle several research questions in this dissertation. In the first project, lipid-wrapped single-walled carbon nanotube (SWNT) conjugates and their interactions were examined with several membrane-disrupting molecules. The results of our simulations with the experimental optical emission spectra of these conjugates were compared, and the magnitude of the optical signal from the magnitude of the observed structural disruption was predicted. In the subsequent project, machine learning approaches were used to predict new DNA sequences in DNA-SWNT conjugates that can sense serotonin molecules. In the last project, *BinderSpace*, an open-source Python package was coded and developed for motif analysis, sequence visualization, and clustering. This tool was instrumental in analyzing datasets of oligonucleotides binding to single-wall carbon nanotubes and cyclic peptidomimetics interacting with bovine carbonic anhydrase protein.

Overall, this dissertation demonstrates the effective combination of computational methodologies in molecular science and contributes valuable tools and knowledge that can significantly impact sensor technology.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Binding affinity between two molecules is a pivotal property in drug discovery and sensor development. There exists a complementary utilization of computational and experimental methods to identify molecules with high binding affinities to specific target molecules. Initial screenings often employ rapid computational techniques, followed by experimental methods to refine, and sometimes elucidate the structures of bound complexes. Subsequently, computational methods, particularly molecular dynamics (MD) simulations, provide intricate details of interactions and binding, while the machine learning methods offer predictive analytics and pattern recognition from the accumulated data, further enhancing the accuracy of prediction of new candidates proposed to have high affinity to target molecules.

In the second chapter, the computational methodologies were used to perform the research described in the subsequent chapters.

Chapter three focuses on lipid-functionalized SWNTs, examining their interactions with various membrane-disrupting molecules using MD simulations. The study revealed that these SWNTs favor asymmetrical positioning within the phosphatidylcholine (POPC) corona phase. By juxtaposing our simulation outcomes with experimental optical emission spectra of these conjugates (obtained by our collaborators), we aimed to discern if the magnitude of the optical signal could be predicted from the observed structural perturbations. The findings of this chapter were published in the journal *ACS Applied Materials & Interfaces* under the title "Characterizing the Interactions of Cell-Membrane-Disrupting Peptides with Lipid-Functionalized Single-Walled Carbon Nanotubes."[1]

The fourth chapter, DNA-wrapped single-walled carbon nanotube (SWNT) conjugates, known for their unique optical properties in biosensing and imaging, were explored. The challenge lay in predicting DNA sequences that yield strong analyte-specific optical responses. Using near-infrared (nIR) fluorescence datasets, machine learning (ML) models were trained to predict DNA sequences with pronounced optical responses to the neurotransmitter serotonin. This approach led to the discovery of five novel DNA–SWNT sensors with enhanced fluorescence response to serotonin. The insights from this chapter were published in the journal *ACS Nano* in an article entitled "Discovery of DNA–Carbon Nanotube Sensors for Serotonin through Machine Learning and Near-infrared Fluorescence Spectroscopy."[2]

In the last chapter, the discovery of target-binding molecules, such as oligonucleotides and peptides, was enhanced using *BinderSpace*, an open-source Python tool was coded and developed. This tool proved useful for analyzing datasets of oligonucleotides binding to single-wall carbon nanotubes and cyclic peptidomimetics interacting with bovine carbonic anhydrase protein, emphasizing the importance of bioinformatics in understanding large datasets and identifying high-affinity binders. The findings from this chapter were featured in the *Journal of Computational Chemistry* under the title "BinderSpace: A Tool for Analyzing Sequence Spaces in Datasets of Affinity-Selected Oligonucleotides and Peptide-Based Molecules."[3]

Chapter 2: Methods

2.1. Molecular dynamics simulation

The systems in chapter three were investigated using classical atomistic molecular dynamics simulations. This section covers the various aspects of MD simulations, including their theoretical foundations, force fields, and interactions between atoms. The integration method for solving the equations of motion in each thermodynamic ensemble is also discussed.

Classical Molecular Dynamics

The simplest way to perform MD simulations is to solve the classical equations of motion using Newton's equations for all atoms present in the system of interest. These equations are used for systems that have constant energy as a constraint[4]. For a system containing N particles, the Cartesian coordinates r_i and velocities v_i of each particle i present in the system are obtained by solving Newton's equations of motion. For the system which contains N particles, the force on particle i, $\vec{f_i}$, can be obtained from Newton's equations of motion as follows:

$$m_i \frac{\partial^2 \vec{r_i}}{\partial t^2} = \vec{f_i} \tag{2-1}$$

$$\vec{f_i} = -\frac{\partial}{\partial \vec{r_i}} U(\vec{r_1}, \vec{r_2}, \dots, \vec{r_N}) \qquad , \tag{2-2}$$

where m_i is the mass of particle i, t is time, and U is the total potential energy of the system, a function of coordinates of all the N atoms, which is defined in the section below.

To solve the equations, we need to calculate the forces $\vec{f_i}$ exerted on particle i by all the other particles present in a system. The main computational task for MD software, such as NAnoscale Molecular Dynamics (NAMD)[5] which is used in this dissertation research, is the efficient calculation of the potential energy U and the above equations. The potential function U, which defines the interaction between particles, is also called a force field, which will be introduced below.

The MD simulations performed in this dissertation used Langevin equation of motion, which are slightly different from Newton's equations of motion. These equations are used to maintain constant temperature T and constant pressure p inside the system, which usually corresponds to the conditions in the laboratory.

Atomistic Force Field

As mentioned above, the potential function U is defined through the interactions between all the atoms present in a simulated system. The definitions of these interactions and specific parameters assigned to atoms is overall called a force field. Every atom present in the system is assigned its type, which has its defined parameters. All the parameters associated with each atom type have been previously determined and validated by other researchers and are available for use by those who perform MD simulations. The force field used in this thesis is called Chemistry at HArvard Macromolecular Mechanics (CHARMM) force field, whose parameters were determined by quantum mechanical calculations[4]. The force field parameters are validated by comparison of simulated and experimental properties, such as solvation free energy and vaporization heats, for molecules in different thermodynamics conditions.[6,7]

The force field parameters are divided into two main groups: non-bonded and bonded parameters. The non-bonded parameters are comprised of Coulombic and van der Waals interactions, while the bonded parameters include the intramolecular interactions, namely, bonds, angles, dihedrals, and improper angles. The potential energy of the system is defined as an additive potential. Namely, we define a full system potential $V_N = U(\overrightarrow{r_1}, \overrightarrow{r_2}, \dots, \overrightarrow{r_N})$ as a sum of bonded and non-bonded interaction energies:

$$V_N = \sum V_{N,bonded} + \sum V_{N,non-bonded} \tag{2-3}$$

Where $V_{N,bonded}$ is the potential energy arising from intramolecular interactions, which contains stretching, bending and torsions of chemical bonds. Chemical bonds and angles are considered as springs in CHARMM forcefield. In Figure 2-1 and Equations 2-4 to 2-7, $V_{N,bonded}$ of CHARMM force field[8] are depicted and defined for all angles, bonds, improper dihedrals, and dihedral angles defined in the system:

$$V_{bond} = \sum_{bond\ i} k_i^{bond} (r_i - r_{0i})^2 \tag{2-4}$$

$$V_{angle} = \sum_{angle\ i} k_i^{angle} (\theta_i - \theta_{0i})^2 \tag{2-5}$$

$$V_{dihedral} = \sum_{dihedral\ i} k_i^{dihedral} [1 + cos(n_i \phi_i - \gamma_i)] \quad n_i \neq 0 \tag{2-6}$$

$$V_{improper} = \sum_{improper\ i} k_i^{improper} (\phi_i - \gamma_i)^2 \quad n_i = 0 \tag{2-7}$$

Above, the total energies arising from bonds, angles and improper dihedrals, labeled as V_{bond} , V_{angle} and $V_{improper}$, are described by weak harmonic potential forms with spring constants (k_i^{bond}, k_i^{angle} , $k_i^{improper}$) and associated instantaneous internal coordinates r_i, angles θ_i, dihedrals and improper dihedrals ϕ_i, which differ from their equilibrium values (r_{0i}, θ_{0i}, γ_i). The equilibrium values (r_{i0}, θ_{i0}, γ_{i0}) represent stable equilibrium state with minimum stretching, bending and torsional energies. Cosine function is used to define $V_{dihedral}$, which is dependent on force constant $k_i^{dihedral}$, periodicity n_i and dihedrals (ϕ_i) varying from γ_i.

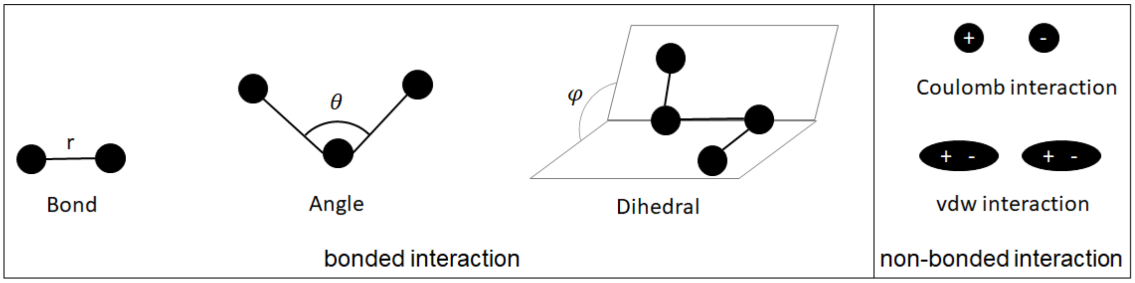


Figure 2-1: Illustration of bonded and non-bonded interactions in CHARMM force field.

Non-bonded potential energies, due to Coulomb and van der Waals (vdW) interactions, are defined as pairwise interactions. They are calculated using the Coulomb law. The Coulomb potential between two atoms i and j, which have partial charges q_i and q_j (defined as force field parameters), is given as follows:

$$V_{coul}\left(r_{ij}\right) = \sum_i \sum_{j>i} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \tag{2-8}$$

where, ϵ_0 and r_{ij} denote vacuum permittivity and the distance between the centers of atoms i and j, respectively. This term is evaluated explicitly within a predetermined cutoff distance, beyond which the Coulomb interactions are either ignored or estimated using different approaches due to the prohibitive computational cost of calculating every pairwise interaction in large systems. One commonly used method for accounting for long-range electrostatic interactions beyond this cutoff is the Particle Mesh Ewald (PME) method[9]. The PME method divides the computation of long-range electrostatics into real space and reciprocal space interactions, thereby allowing for an accurate and efficient treatment of the Coulomb potential over the entire system, despite the presence of a cutoff. This technique is particularly important in periodic systems where long-range interactions play a significant role in the physical behavior of the system.

The short-ranged Lennard-Jones (LJ) potential is used to describe vdW interactions between a pair of atoms *i* and *j*. The LJ potential is composed of two terms. The Pauli repulsion term is important at short range distances between atoms due to overlapping electron orbitals. The attractive long-range term describes the attraction between atoms at long range distances. LJ potential is defined as:

$$V_{LJ}(r_{ij}) = \sum_i \sum_{j>i} \varepsilon_{ij} \left[\left(\frac{R_{min,ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{min,ij}}{r_{ij}} \right)^{6} \right] \quad (2\text{-}9)$$

where r_{ij} is the distance between atoms i and j, and $R_{min,ij}$ and ε_{ij} are calculated from the Lorentz-Berthelot rules:

$$R_{min,ij} = \frac{R_{min,i}+R_{min,j}}{2} \quad \varepsilon_{ij} = \sqrt{\varepsilon_i \varepsilon_j} \quad (2\text{-}10)$$

where the particles i and j's respective radii and potential well depths are represented by $R_{min,i}$, $R_{min,j}$ and ε_i, ε_j.

Integration method for MD

By operation of integration, Newton's equations of motion can provide coordinates and velocities of the particles present in the system. To compute the integrals, the velocity-Verlet algorithm is used. The algorithm considers the position of the particle, its velocity, and its force at the time of its calculation. The velocity-Verlet algorithm procedures are defined as follows:

"half-kick" → $v_{n+1/2} = v_n + m^{-1} f_n . \Delta t/2$ (2-11)

"drift" → $R_{n+1} = R_n + v_{n+1/2} . \Delta t$ (2-12)

"Compute force" → $f_{n+1} = f(R_{n+1})$ (2-13)

"half-kick" → $v_{n+1} = v_{n+1/2} + m^{-1} f_{n+1} . \Delta t/2$ (2-14)

The time reversibility and simplistic properties of the velocity-Verlet method contribute to momentum conservation, an important theoretical feature for researchers carrying out the MD simulations.

Periodic boundary condition (PBC)

Experimental systems typically contain many atoms (> Avogadro's number). However, MD simulations are not capable of efficient simulations of so many atoms. The systems that are chosen to form a simulation unit cell are therefore typically much smaller, containing from several atoms to about a billion atoms. To make simulated systems resemble a much larger experimental system, a periodic boundary condition (PBC) is introduced in the simulations. This condition propagates the simulation unit cell box in x, y, and z dimension an infinite number of times, so that the original simulation boxes interact with propagated adjacent unit cells.

Ensembles

Experimental systems usually exist in conditions with well-defined thermodynamic parameters. We can say that these systems can be described as thermodynamic ensembles. We can perform MD simulations with a different choice of such thermodynamics ensembles. A first common ensemble is NVE ensemble, where the number of particles (N), volume (V) and total energy (E) of the system are held constant. Other common ensembles are NVT (N, V, and temperature (T) are held constant), and NPT (N, V, and pressure (P) are held constant).

The simulations performed using the NPT ensemble in this thesis were conducted with Langevin dynamics equations of motion. In simulations performed with Langevin equations, temperature and pressure of the systems are maintained at constant values.

$$m\frac{\partial^2 \vec{r}}{\partial t^2} = m\dot{v} = F(r) - \gamma_{Lang} m v + \sqrt{2\gamma_{Lang} k_b T m}\, G(t) \tag{2-15}$$

where, respectively, r, t, m, v, and F stand for coordinates, time, mass, velocity, and force. γ_{Lang}, k_b, and T stand for the Boltzmann constant, temperature, and friction factor, respectively, depending on the system and user definition. *G(t)* stands for a Gaussian process with a single variable. Damping and fluctuation terms are the second and third terms in the equation, respectively. γ_{Lang}, controls the magnitudes of damping and fluctuating terms.

2.2. Machine Learning

This section presents a thorough overview of machine learning (ML) techniques, particularly highlighting convolutional neural networks (CNN) and support vector machines (SVM), as used in Chapter 4. Furthermore, it introduces dimensionality reduction techniques, specifically the t-distributed stochastic neighbor embedding (t-SNE) and principal component analysis (PCA), along with clustering methods, all of which are utilized in Chapter 5.

Classification and Regression

Machine learning problems can be broadly categorized into classification and regression. Classification deals with discrete variables, wherein the output variable belongs to a specific category or group, such as determining whether a DNA sequence corresponds to high or low responses to serotonin. On the other hand, regression concerns continuous variables, where the output represents a continuous value like the dissociation constant value. Both classification and regression are foundational to various ML algorithmic methods.

Perceptron Networks

All neural networks consist of input layer, hidden layers, and output layer as main sections. A layer contains neurons. Figure 2-2 shows these three sections. All different kinds of neural networks like convolutional neural networks contain perceptron at the end of their models. According to the dataset the perceptron networks can be used alone without adding any other layers.

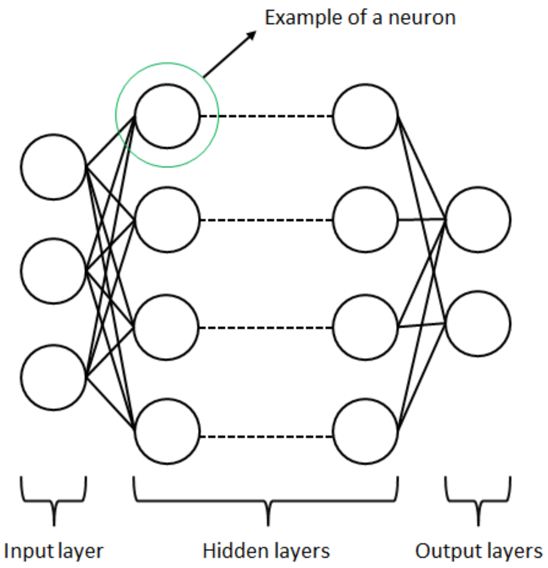


Figure 2-2: Input layer, hidden layers, and output layers. Each circle indicates a neuron in layers.

The perceptron neural network processes input data, each instance of which is represented by a vector of attributes. For each input instance *I*, the perceptron computes an output by multiplying this input vector with a weight matrix *W*, often followed by adding a bias term *b*. The weight matrix *W* consists of individual weights w_i, each corresponding to the importance or contribution of an input attribute towards the resulting output. The perceptron output is thus a weighted sum given by $W^T \cdot I + b$, as shown in equation 2-16 and figure 2.3:

$$y = fuction(\sum_{i=1}^{n} w_i I_i + b) = fuction(W^T I + b) \qquad (2\text{-}16)$$

where W^T is the transpose of the weight matrix *W*, and *b* is a scalar bias that offsets the input to the activation function. During the training phase, the network adjusts the weights *W* and the bias *b* based on a set of training data, with the goal of minimizing the error in prediction or classification tasks. For classification tasks involving multiple classes, the perceptron's output—before serving as the final predicted class—can be converted into a probability distribution using the SoftMax function. This function, given by equation 2-17, is applied to the raw output scores, or logits, *z*, which can be obtained from the weighted sum before the activation function is applied:

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \qquad (2\text{-}17)$$

In this context, the SoftMax function, denoted as $\sigma(\vec{z})_i$, calculates the probability that a given input belongs to class *i*. The function employs the exponential *e* to ensure that each raw score z_i from the input vector $\vec{z}$ is positive and scales these values such that the sum across all *K* classes is 1, forming a valid probability distribution. Here, *K* signifies the total number of classes in the classifier.

During the training phase, the network fine-tunes the weights *W* and the bias *b* with the aim of minimizing the discrepancy between the network's predictions and the actual outcomes of the training data. For a perceptron that is part of a multi-layer network handling multiple categories, the SoftMax function is typically applied at the final layer. This function translates the perceptron's outputs into a probability distribution, which aids in identifying the most probable class for a given input as illustrated in equation 2-17.

Separately, the network utilizes an activation or transfer function to transform the weighted input sum before classification. One commonly used activation function in neural network models is the step function, which outputs a value of 1 if the processed input is above a threshold and -1 if below. This binary output is pivotal during learning, as it is contrasted with the expected output to compute the error. The error is then used to adjust the weights proportionally to the input magnitude and a predetermined learning rate, thereby incrementally improving the model's accuracy.

The perceptron neural network is essentially a linear classifier, which partitions input data into two distinct groups with a decision boundary. Throughout training, this boundary, represented by the weights, is iteratively optimized. Initially, weights are assigned randomly, and the model is progressively refined following equation 2-16. Upon completing the training cycle, the model can make predictions.

It should be noted that the description of the SoftMax function generating a binary value of 1 or -1 is incorrect; SoftMax outputs probabilities, not binary classifications. Binary results are typical of the sign or step function, which is not the same as the SoftMax function.

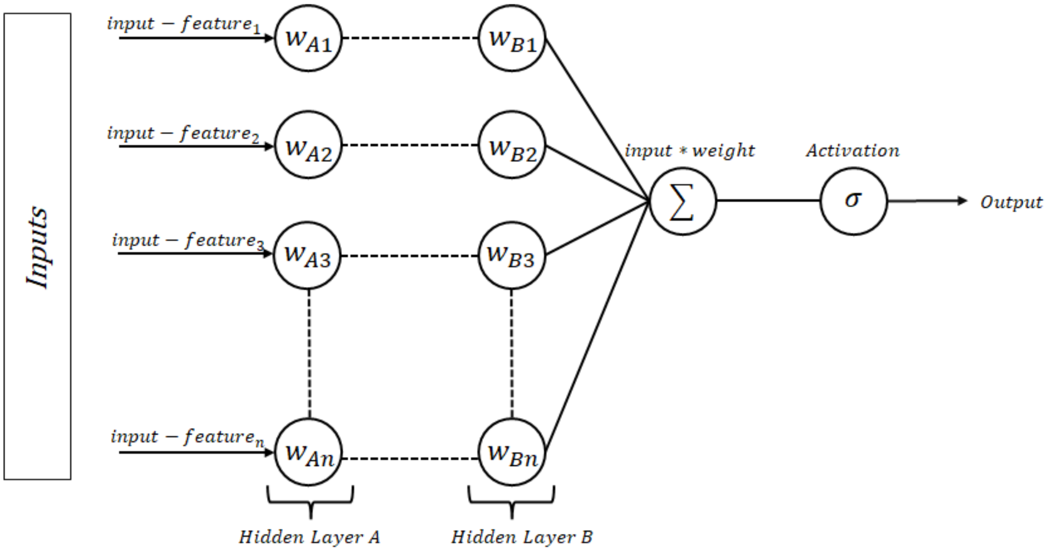


Figure 2-3: Input layer, Hidden layers, and output layers. Each circle indicates a neuron in layers.

Convolutional Neural Networks

Convolutional neural network (CNN) is a deep learning algorithm that can receive as input images and text data which contains DNA sequences data. A CNN model is divided into feature extraction and classifier parts, and the last part is a perceptron network. In the feature extraction part, CNN assigns importance (learnable and bias weights) to each of the objects/aspects and extracts features in the images and text and can distinguish them from each other. Feature extraction is the critical difference between ML and CNN deep learning models. Usually, a DNA sequence is a sequence of four or fewer nucleoids, adenine (A), cytosine (C), guanine (G), and thymine (T) which can be treated as a letter and the DNA sequence as text. Since all ML and DL methods work with numerical data, these letters need to be converted to numbers, which is known as encoding. Two types of encodings are typically used to encode the text data, one hot encoding, which changes the sequence to 0 and 1, and a label encoding. Figure 2.4 shows an example of how the encoding works for the DNA sequences.

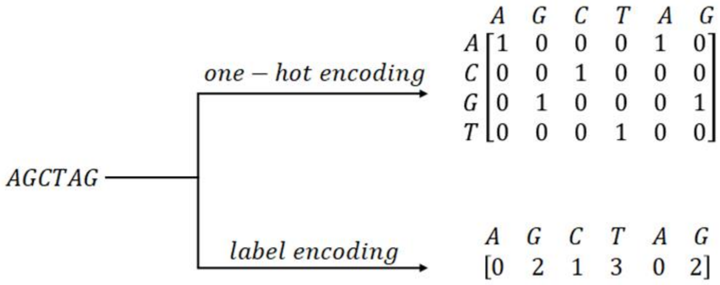


Figure 2-4: Illustration of one-hot encoding (also referred to as position-specific vector encoding) and label encoding methods.

The convolutional layer is a hidden layer, as shown in Figure 2.5, to extract features that will add more dimensions. A max-pooling layer is added after each convolutional layer to reduce the dimensions of extracted features. Kernel size and the number of filters is the most critical hyperparameter in the convolutional layer that can affect the learning result. Flattening is used after the max pooling layer to convert all the resultant 2-dimensional arrays from pooled feature maps into a single long continuous linear vector. Adding a fully connected layer is an inexpensive way to learn high-level nonlinear combinations of features as provided by the feature extraction layers (convolutional layers + max pooling + flatten layer). The fully connected layer in that space retains a possibly nonlinear function. With all the layers, the input image or text is transformed into a suitable output to be fed to the classifier section to assign classes[10–13].

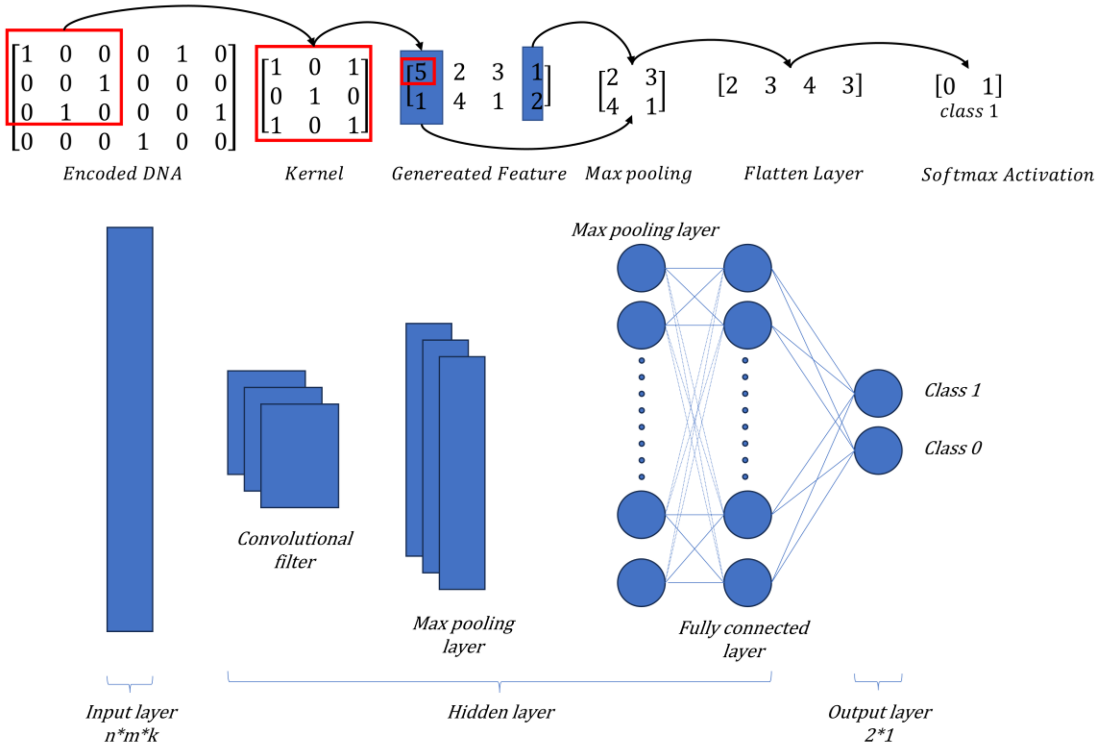
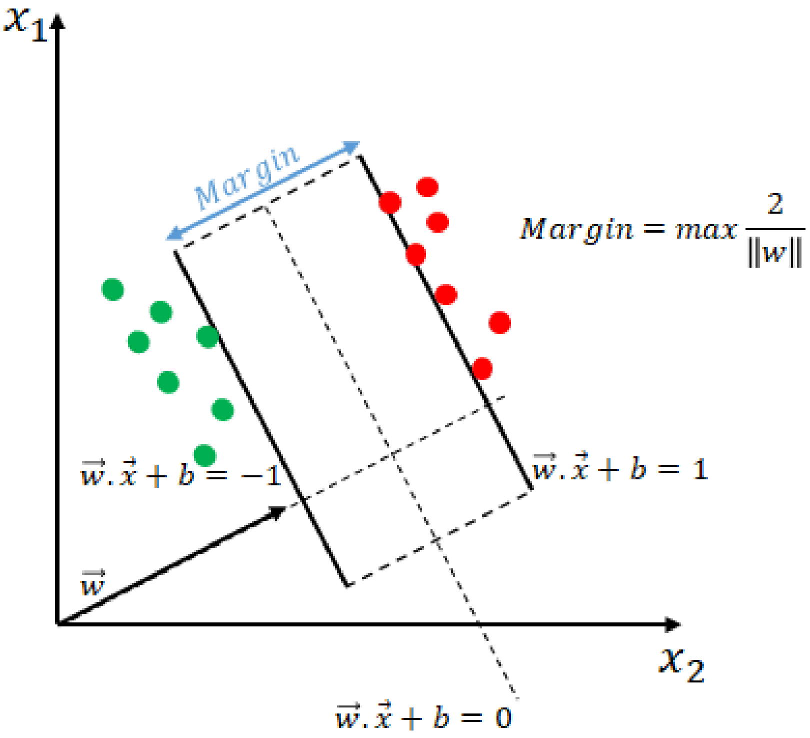


Figure 2-5: Workflow of a convolutional neural network (CNN) for DNA sequence classification. The process begins with binary-encoded DNA sequences, followed by feature extraction using convolutional filters (kernels). The resultant feature maps are then down sampled through max pooling and flattened to form a one-dimensional vector. This vector is subsequently passed through a fully connected neural layer, culminating in a SoftMax activation for probabilistic classification into two distinct classes: Class 0 and Class 1.

Support Vector Machines

Support vector machines (SVMs) Support vector machines (SVMs) are a set of supervised learning algorithms used for classification and regression tasks. In the context of classification, SVMs aim to find the optimal separating hyperplane that divides classes of data with the maximum margin. Figure 2.6 provides a visual representation of how SVMs accomplish this task (in two dimensions rather than the usual multiple dimensions that typically are present in real application problems). The figure illustrates a two-dimensional space where the x-axis is $X_{2,}$ and the y-axis is X_1. In this space, data points from two different classes are plotted: datapoints from one class are represented by green dots, and the datapoints from the other class by red dots. The SVM algorithm seeks to find the line (in two dimensions) or a hyperplane (in higher dimensions) that best separates these two classes. This line is the decision boundary, and the equations **w**·**x**+b=−1 and **w**·**x**+b=1 represent the borders of the margin on either side of it.[14]



x_1
Margin
$Margin = max\frac{2}{\|w\|}$
$\vec{w}.\vec{x} + b = -1$
$\vec{w}.\vec{x} + b = 1$
$\vec{w}$
x_2
$\vec{w}.\vec{x} + b = 0$

Figure 2-6: SVM Decision Boundary Visualization: The two classes, represented by green and red dots. The parallel dashed lines denote the support vectors, given by equations w.x + b = -1 and w.x + b = 1. The distance between these support vectors is the margin, calculated as max(2/||w||), which SVM aims to maximize for optimal class separation.

In Figure 2.6, the margin is indicated by the distance between the two dashed lines parallel to the decision boundary, which is the line **w**·**x**+b=0. The margin is the region that encompasses the closest points of both classes, which are equidistant from the decision boundary. The SVM classifier works by maximizing this margin, and the maximum width of the margin is calculated as 2/‖**w**‖, where **w** is the weight vector perpendicular to the separating hyperplane. The points that lie on the dashed lines are known as the support vectors, as they 'support' the margin.

The b in the equations represents the bias, which adjusts the position of the hyperplane along the direction of **w**. Together, **w** and b define the decision boundary: **w** determines its orientation, and b its displacement from the origin. The goal of the SVM training process is to determine the optimal values of **w** and b such that the margin is maximized, subject to the condition that all data points are correctly classified, meaning that all points from one class fall on one side of the margin and all points from the other class fall on the opposite side.

For cases where the data is not linearly separable in the original input space (the space where X_1 and X_2 are defined), SVMs employ a mathematical function known as the kernel function to project the data into a higher-dimensional space where it is possible to find a separating hyperplane. This is where the phi function (ϕ) mentioned comes into play. It is a transformative feature mapping that takes the original input data and projects it into a higher-dimensional space. This process is essential when the data is not linearly separable in its original space. The phi function is implicitly represented through the kernel function, which computes the dot product of data points in this new, larger space without having to compute the high-dimensional space explicitly. This is known as the kernel trick, which allows the SVM to efficiently create a decision boundary for complex datasets.

In summary, by adjusting **w** and b based on the input data, SVMs find the widest possible margin between classes, as depicted in Figure 2.6. Once these parameters are learned, the SVM model can be used to predict the class of new data points by determining which side of the decision boundary they fall on, using the inequalities $(\mathbf{w}\cdot\mathbf{x}+b)\geq 1$ for class 1 and $(\mathbf{w}\cdot\mathbf{x}+b)\leq -1$ for class 0.[15]

$$(w.x + b) \geq 1, x \in \ class\ 1 \tag{2-18}$$

$$(w.x + b) \leq -1, x \in \ class\ 0 \tag{2-19}$$

Machine learning evaluation metrics

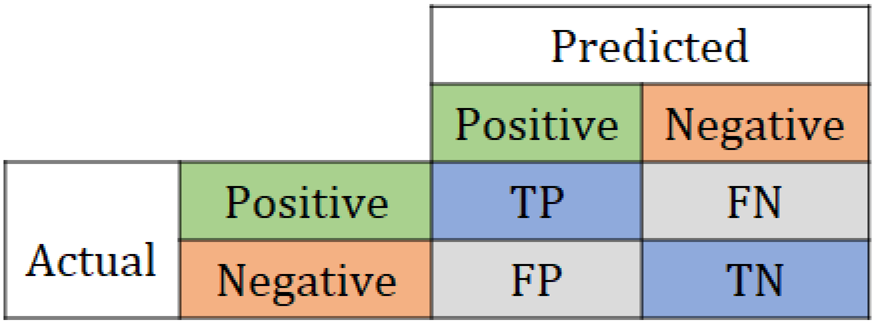
In machine learning, datasets are typically divided into two portions: training and testing. The training dataset, which is typically the larger subset, is used to train the model, while the testing dataset is employed to evaluate the trained model's performance. Performance metrics such as accuracy, recall, precision, and f1 score play a crucial role in evaluating the model's quality. Accuracy is determined by the proportion of correct predictions out of all the test datapoints and is given by equation 2-20. Terms like True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) are fundamental in understanding these metrics and are illustrated in Figure 2-7. The recall, which represents the model's ability to correctly identify positive instances, is calculated as shown in equation 2-21. Precision, on the other hand, measures the proportion of positive identifications that were correct and is given by equation 2-22. The f1 score harmonizes precision and recall into a single metric and is defined as below.[16]

$$accuracy = \frac{number\ of\ the\ correct\ prediction}{test\ size} * 100 \tag{2-20}$$

$$Recall = \frac{TP}{FN+TP} \tag{2-11}$$

$$Precision = \frac{TP}{FP+TP} \tag{2-22}$$

$$f1\ score = \frac{2*precision*recall}{precision+recall} \tag{2-23}$$



		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

Figure 2-7: A confusion matrix illustrating the four primary outcomes in binary classification: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). It provides a comprehensive view of a model's performance by comparing predicted results against actual values.

Receiver Operating Characteristic (ROC) curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. It plots the true positive rate (sensitivity) against the false positive rate (1-specificity) at various threshold settings.[17]

Area Under the Curve (AUC), refers to the area under the ROC curve and is a measure of the overall ability of the test to discriminate between positive and negative class values. A larger AUC indicates better model performance, with a value of 1 representing perfect discrimination and a value of 0.5 suggesting no discriminative power, equivalent to random guessing.[17]

Dimensionality Reduction

In this dissertation, dimensionality reduction is leveraged to condense the complex, high-dimensional information present in DNA or peptide sequences into a more manageable form, aiding in data visualization and analysis. The two techniques utilized are principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE).

PCA serves to simplify high-dimensional data by finding the directions, or axes, where the data shows the most significant variance. Variance in this context is a statistical measure of how much the data points deviate from their average value. In essence, PCA finds where the data is most spread out and projects it onto new axes that capture this spread most effectively. The construction of a covariance matrix is a preliminary step in PCA. This matrix is a systematic arrangement of values that quantifies the extent to which each dimension varies in concert with the others. When two dimensions increase or decrease together, they have a high covariance, indicating a relationship between them. Principal components are extracted from this matrix. They are the eigenvectors that point in the direction of the highest variance, with each direction's significance quantified by its corresponding eigenvalue. Eigenvalues measure the magnitude of variance along their respective eigenvectors, revealing the relative importance of each direction.

The PCA transformation is captured mathematically by the equation $Y=XW$. Here, Y is the transformed dataset expressed in terms of new axes, X is the original dataset, and W is the matrix of eigenvectors that define the new coordinate system derived from the covariance matrix.[18]

While PCA streamlines data by emphasizing its most spread-out features, t-SNE focuses on preserving the local structure of the data as it reduces dimensionality. It places each data point into a lower-dimensional space, a process known as embedding, which is akin to creating a detailed map that maintains the proximity of neighborhoods in a complex city layout. t-SNE compares two distributions: one representing similarities between data points in the original space, and the other representing similarities in the new, lower-dimensional space. The goal of t-SNE is to make the lower-dimensional distribution reflect the high-dimensional one as closely as possible. The divergence, specifically the Kullback-Leibler divergence, is a mathematical criterion that t-SNE aims to minimize. It is a measure of how one probability distribution diverges from a second, expected probability distribution. In the context of t-SNE, minimizing this divergence is the objective of a cost function—a mathematical function that the algorithm tries to minimize. Because t-SNE is non-convex, the optimization process can arrive at different solutions with each run, as it does not settle at a single best answer but rather at one of the many local minima in the solution landscape. t-SNE's core formula is:

$$q_{ij} = \frac{(1+||y_i - y_j||^2)^{-1}}{\sum_{k\neq l}(1+||y_k - y_l||^2)^{-1}} \quad , \qquad (2\text{-}2)$$

calculates the probabilities that define similarity in the lower-dimensional space. Here, q_{ij} represents the likelihood that points *i* and *j* are neighbors in the embedded space. The embedded space is the lower-dimensional representation where each data point's coordinates, y_i and y_j, maintain the essential relationships seen in the high-dimensional original dataset.

Both PCA and t-SNE offer insightful methods for visualizing high-dimensional data, each through a distinct lens that caters to different aspects of the data's innate structure. PCA highlights the global structure by pinpointing the directions with the greatest variance, effectively revealing the overarching spread and range of the dataset. Conversely, t-SNE illuminates the local patterns by preserving the neighborhood relationships, ensuring that data points close to each other in the high-dimensional space remain close in the reduced-dimensional representation. When selecting the appropriate method for data analysis, researchers must consider the unique characteristics of the data along with the specific investigative questions they aim to address. This decision is crucial as it determines whether the emphasis should be on the data's broad trends or on the finer, subtle interactions within it.[18]

Clustering

Clustering is a technique in unsupervised learning where data points are grouped into distinct clusters based on their similarities, without any prior labeling. This method is especially useful for discovering inherent patterns and structures in datasets. Among the various clustering algorithms, balanced iterative reducing and clustering using hierarchies (BIRCH) algorithm is adept at handling large datasets by constructing a tree structure, thus reducing the complexity. K-means, on the other hand, partitions the dataset into 'K' number of centroids and is one of the most widely used clustering methods. Density-based spatial clustering of applications with noise (DBSCAN) identifies clusters based on the density of data points, effectively distinguishing between high-density regions and areas of noise or outliers. Lastly, Gaussian mixture model (GMM) assumes that data is generated from a mixture of several Gaussian distributions. It seeks to identify these individual distributions, making it a probabilistic approach to clustering. Each of these methods offers unique advantages, and the choice of algorithm often depends on the nature of the data and the specific goals of the clustering process. The process of dimensionality reduction and clustering was employed to construct the BinderSpace, as detailed in Chapter 5.

Chapter 3: Characterizing the Interactions of Cell-Membrane-Disrupting Peptides with Lipid-Functionalized Single-Walled Carbon Nanotubes

3.1. Introduction

Antimicrobial molecules often function by permeating and disrupting bacterial membranes, leading to bacterial cell death either by dissolving their membranes or creating pores that result in cytoplasmic leakage and membrane potential destruction.[19–21] However, bacteria naturally develop resistance to antimicrobials over time, which leads to a continuing need for the development of new antimicrobial molecules. Many organisms produce antimicrobial peptides (AMPs) as a part of their innate immune responses.[22] Drawing inspiration from these natural peptides, numerous antimicrobial screening efforts aim to discover novel membrane-disrupting molecules either through the chemical design and synthesis of new peptides[23] or by creating extensive combinatorial peptide libraries[24] and evaluating their antimicrobial activity.

While modern advancements in automated peptide synthesis technologies have expedited the synthesis of new peptide-based compounds,[24–26] the evaluation of their antimicrobial properties remains a significant challenge. Traditional screening methods involve assessing the effects of potential antimicrobials on live target cells using growth inhibition and cell death assays.[27] However, these methods are often time-consuming, costly, and come with contamination risks. As a result, alternative antimicrobial screening techniques have been developed. One such approach involves the use of droplet-based microfluidic platforms, where individual cells and potential compounds are encapsulated in picolitre-volume emulsion droplets for analysis through imaging or fluorescence reporters.[28,29] Another approach employs abiotic systems that screen for molecules specifically designed to disrupt cell membranes. These systems utilize simplified cell membrane models like artificial planar lipid bilayers and liposomes.[30,31]

Despite the potential of existing antimicrobial discovery methods, each comes with its own set of advantages and limitations. These limitations often pertain to the speed, throughput, detail level, and the applicability of results to live cells. Recognizing these challenges, recent research has proposed an abiotic system based on lipid-wrapped single-walled carbon nanotube (SWNT) conjugates as a novel optical sensing platform for antimicrobial compound screening.[32] This platform leverages SWNTs as near-infrared (nIR) fluorescent transducers to report lipid interactions with antimicrobial compounds occurring on the SWNT surface.

SWNTs, when wrapped by various polymers, have found applications in diverse fields such as biological catalysis, bio separations, gene therapy, photothermal cancer therapy, and bio analyte detection. [2,33–48] The role of these polymers is twofold: solubilizing hydrophobic SWNTs in aqueous media and binding to potential analytes in sensing applications. SWNTs, known for their remarkable electronic and optical properties,[49] are pivotal in polymer-wrapped SWNT conjugates for optical sensing.

While previous studies have linked changes in SWNT emission to significant alterations in dielectric constants and local electric fields, the impact of small analyte concentrations on the SWNT environment remains largely unexplored. In the context of POPC-wrapped SWNTs interacting with antimicrobial peptides, the addition of minimal peptide concentrations primarily led to a decrease in SWNT emission intensity.[32] The structural changes in the lipid corona that induce this decrease in SWNT fluorescence emission intensity remain unidentified.

In this study, we focus on characterizing phospholipid-coated SWNTs, exploring the effects of membrane-disrupting peptides on these sensors and phospholipid bilayers, and comparing the effects across both phospholipid systems. We employ atomistic molecular dynamics (MD) simulations to detail peptide binding and the subsequent perturbations in these phospholipid systems. Our findings, in conjunction with previous experimental results,[32] are analyzed for potential correlations and mechanistic insights. We specifically chose POPC-coated SWNTs for our investigations due to the availability of experimental data for these systems.[32].

3.2. METHOD

Antimicrobial Building Atomistic Models of POPC-SWNT Systems

A segment of a (6,5) single-walled carbon nanotube (SWNT) measuring 8 nm was constructed using the Carbon Nanostructure Builder tool in VMD software.[50] A single POPC molecule's structure (1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine) was sourced from a POPC membrane segment, which was crafted using the Membrane Builder tool in VMD software and the CHARMM36 topology.[51] Using a custom bash-TCL script which I developed specifically for this project, copies of this POPC molecule were systematically arranged around the SWNT segment. This script positions a specified number of POPC molecules radially around the SWNT. The script's input includes the nanotube's structure, the single POPC molecule's structure, the angle between neighboring radially oriented POPC molecules, and the number of times the POPC molecules are replicated along the SWNT's length. This script can be accessed on GitHub (https://github.com/vukoviclab/POPCsensor). Using this approach, four POPC-SWNT systems were generated with mass density ratios of 9:1, 15:1, 20:1, and 30:1. Each system was then solvated with TIP3P water molecules. Details of the constructed systems are provided in Ref[1].

Constructing the Atomistic Model of the POPC Bilayer

A POPC bilayer was assembled by Ms. Yadav using the CHARMM-GUI membrane builder[52] and CHARMM36 topology,[51] set in a tetragonal box containing 160 lipid molecules (80 per leaflet) and 31,838 TIP3P water molecules. The molecule counts were chosen to maintain a similar POPC concentration as in the 15:1 POPC-SWNT system. Details are provided in Ref[1].

Building Models of POPC-SWNT Systems and POPC Bilayers with Membrane-Disrupting Peptides

Three peptides known for membrane disruption, namely colistin (Col), TAT peptide was constructed by Ms. Yadav, and a crotamine-derived peptide (Cro), were explored in the simulations. The Col structure was prepared using the ChemSpider MOL structure (id = 65877) and supplemented with hydrogen atoms. Parameters for Col, based on the CHARMM general force field (CGenFF), were sourced from the CGenFF web interface.[53,54] The TAT structure was derived from the HIV-derived TAT peptide's crystal structure (pdb ID: 1K5K).[55] The Cro structure was adapted from the crotamine's crystal structure (pdb ID: 4GV5)[56] with specific mutations to achieve the desired sequence.[32] Additionally, a poly-R peptide made of three arginines (RRR) was crafted using the TAT peptide's structure (pdb ID: 1K5K).[55] These peptides' parameters were based on the CHARMM36 protein force field.[6] Subsequently, six POPC-SWNT systems were developed with Col, TAT, and Cro molecules to study their impact on POPC-SWNT conjugates. Each system contained a POPC-SWNT conjugate with a 9:1 or 15:1 mass density ratio and ten molecules of either Col, TAT, or Cro, strategically placed within 20 Å of the POPC corona (detailed are prepared in Ref[1]). Structures of the POPC-SWNT conjugates used were pre-equilibrated in 1 μs MD simulations. All the new systems were solvated in TIP3P water and charge-neutralized by adding Cl^- ions. Additionally, a system was created with ten RRR peptides and a 9:1 POPC-SWNT mass density ratio system (Detailed in Ref[1]). This system was constructed similarly to the six previously described POPC-SWNT systems. Three separate POPC bilayers were also developed with Col, TAT, and Cro molecules. All these bilayer systems had ten molecules of either Col, TAT, or Cro, placed within 20 Å of the POPC bilayer headgroups. Each system was solvated in TIP3P water and charge-neutralized by adding Cl^- ions.

Molecular Dynamics Simulations

All systems were simulated using the NAMD2.13[57] software package and CHARMM36 force-field parameters.[6,51,58] The simulations employed Langevin dynamics in the NPT ensemble, with set conditions of 310 K temperature and 1 bar pressure. The integration time step was 2 fs, and interactions were evaluated within a 12 Å cutoff distance. Long-range interactions were assessed using the particle-mesh Ewald (PME) method with periodic boundary conditions applied.[59] The pure POPC-SWNT systems underwent an initial minimization of 20,000 steps. Following this, solvent molecules were equilibrated for 1 ns, with SWNTs and POPC restrained using harmonic forces. The systems were then subjected to production runs of varying lengths. Six POPC-SWNT systems with added peptides (Detailed in Ref[1]) underwent a similar minimization and equilibration process. During these stages, SWNT and POPC molecules were restrained. The systems were then subjected to 1 μs long production runs. Simulations of the 9:1 POPC-SWNT system with RRR peptides followed a similar protocol, but with a 200 ns long production run. Separate simulations were conducted to study single disruptor molecules' interactions with equilibrated 15:1 POPC-SWNT systems. These systems were minimized, pre-equilibrated, and then subjected to 1 μs long production runs. Simulations of four POPC bilayer systems were also conducted. These systems underwent a 20,000-step minimization, followed by a 1 ns pre-equilibration and 1 μs long production runs.

Analysis of POPC Corona Thickness

The POPC distribution around the SWNT was analyzed by calculating the POPC corona's thickness, d_{POPC}, based on the angle θ. The value of dPOPC(θ) was determined by averaging over bins along the z-axis of the system.

$$d_{POPC}(\theta) = \frac{1}{N_l} \sum_{l_{min}}^{l_{max}} \left(r_{max,\theta}(l) - r_{SWNT} \right) \tag{3-1}$$

Where l_{min} refers to the smallest z coordinate in the bin, while l_{max} represents the largest z coordinate in the bin. The term $r_{max,\theta}(l)$ denotes the furthest radial distance from a single-wall carbon nanotube (SWNT) surface to a POPC atom, specifically the POPC atom that is at the greatest radial distance within the bin and r_{SWNT} signifies the radius of SWNT.

Calculating Distances between Disruptor Molecules and SWNT Surfaces

To understand the binding nature of disruptor molecules to POPC-SWNT conjugates, the distances of all bound disruptor molecules from SWNT surfaces were calculated.

$$\langle d_i(t)\rangle = \langle r_i(t) - r_{SWNT}\rangle \qquad (3\text{-}2)$$

Contact Area Calculations

The contact areas between SWNTs and different species in the solution were calculated to understand the SWNT surface's exposure to various molecules and functional groups.

$$\mathrm{A(t)} = \frac{a_{SWNT}(\mathrm{t}) + a_M(\mathrm{t}) - a_{SWNT,M}(\mathrm{t})}{2} \qquad (3\text{-}3)$$

Calculation of Distances between Disruptor Molecules and Centers of POPC Bilayers

To understand the binding nature of disruptor molecules to pure POPC bilayers, the z-axis component of distances of all bound disruptor molecules from the center of the POPC bilayer was calculated.

$$\langle d_i(t)\rangle = \langle z_i(t) - z_{COM,bilayer}(t)\rangle \qquad (3\text{-}4)$$

3.3. Results

Asymmetric Distribution of POPC Corona Around SWNTs

Our preliminary MD simulations explored the interaction of POPC lipids with (6,5) SWNTs. We constructed four systems where POPC lipids were cylindrically arranged around SWNTs (detailed in the ref paper[1]). These systems had POPC to SWNT mass density ratios of 9:1, 15:1, 20:1, and 30:1. Notably, the 30:1 ratio mirrors the mass density ratio from prior experimental setups.[32] Such setups are detailed in Ref.[32], involving a 2 mL mixture of 0.5 mg/mL SWNT and 20 mg/mL POPC, which underwent sonication and centrifugation. Our simulations aim to replicate these experimentally derived POPC-SWNTs in water-based solutions. We focused on individual SWNTs because only they, not bundled SWNTs, are believed to emit light in POPC-SWNT conjugates.[60–62]

After 175 ns to 1 μs of MD simulation equilibration, the systems settled into the configurations depicted in Figure 1a. The 9:1 POPC-SWNT system took a cylindrical form, while others resembled bilayers, with this likeness intensifying with the mass density ratio. In all systems, the central position of the SWNT shifted, leading to an increasingly asymmetrical POPC corona over time. This evolving asymmetry for the 15:1 POPC-SWNT system (Detailed is illustrated in ref paper[1]). The SWNTs ended up being enveloped by the hydrophobic tails of the POPC, but one side always remained in touch with the zwitterionic heads of the POPC molecules. This asymmetry is evident in the POPC corona thickness measurements for the 9:1 and 15:1 system, as shown in Figure 3-1b.

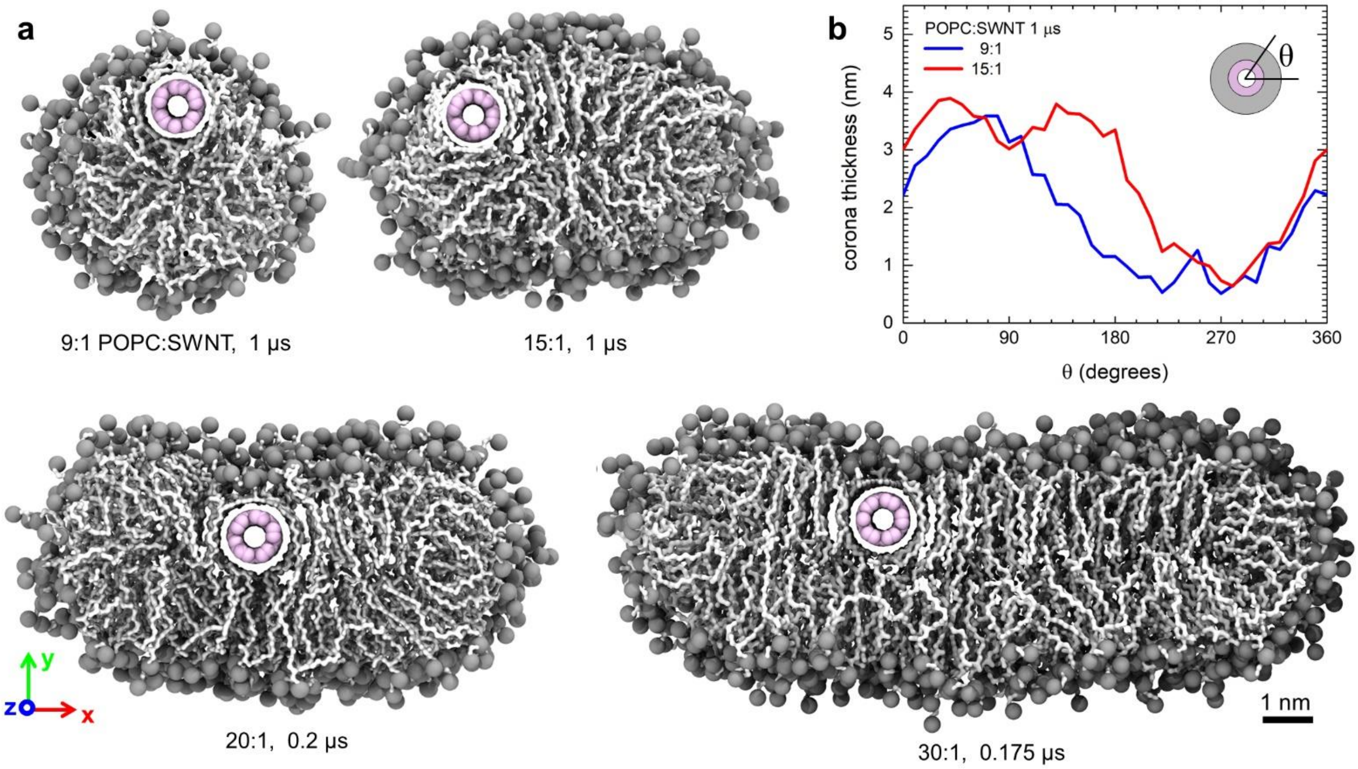


Figure 3-1. Equilibrated configurations of POPC-SWNT assemblies from MD simulations. a) Four representative images of POPC-SWNT combinations at mass density ratios of 9:1, 15:1, 20:1, and 30:1, post-simulation equilibration. Duration of equilibration for each setup is mentioned below their respective images. SWNT atoms are depicted in pink, while POPC tails appear as white links and the P and N atoms of POPC as grey orbs. Water molecules are omitted for visual simplicity. b) Measurement of the POPC layer thickness surrounding SWNTs in 9:1 and 15:1 mass density ratio setup after 1 μs of equilibration.

Interactions of POPC Membrane Disruptors with POPC-SWNTs

We then studied the interactions between POPC-SWNT systems with 9:1 and 15:1 ratio and three peptides known for their disruptive effects on POPC structures. These peptides were Col, TAT, and Cro. We focused on the 9:1 and 15:1 system for better simulation sampling. The 15:1 system already resembles a bilayer, akin to the 20:1 and 30:1 system. After introducing ten molecules to the equilibrated POPC-SWNT systems, they were equilibrated in water for 1 μs, resulting in the structures shown in Figure 3-2a. The disruptor molecules attached to various parts of the POPC corona, with some binding directly to the nanotube. Figure 3-2's results allow for a comparison of the three peptides based on their disruptive effects on POPC-SWNT systems. Col exhibited the most significant disruption, followed by TAT and then Cro. I performed Col and Cro simulations, while Ms. Anju Yadav performed TAT simulations.

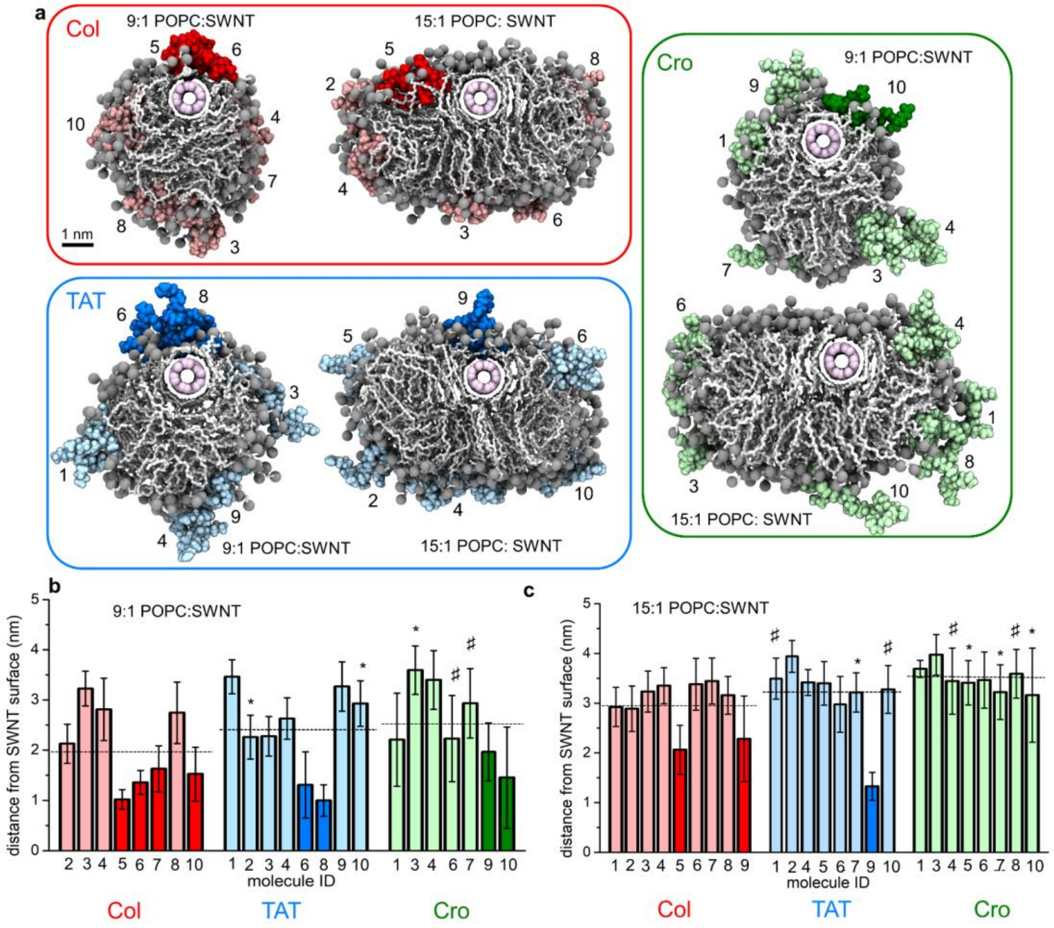


Figure 3-2. Interaction of colistin, TAT peptide, and crotamine-derived peptide with POPC-SWNT assemblies. a) Visual depictions of stabilized POPC-SWNT assemblies in the presence of Col (red), TAT (blue), and Cro (green) at 9:1 and 15:1 POPC:SWNT mass density ratios. Darker shades signify molecules directly touching SWNT in the latter half of their paths, while paler shades represent molecules attached to the thicker POPC layers away from the SWNT core. b) Average distances between the molecules' central points and the SWNT surface while they are linked to 9:1 POPC-SWNTs, detailed further in Table S4. Disruptor molecules have unique identifiers on the x-axis, denoting each of the ten examined molecules in each scenario. Histogram bar shades differentiate molecules directly bound to SWNT (darker shade) and those attached to the POPC layer (lighter shade). Bars marked with (*) represent molecules that intermittently attach and detach from POPC-SWNT, while bars with (♯) indicate molecules with two distinct binding occurrences with POPC-SWNT. Broken lines represent the mean distance of all linked molecules from the SWNT exterior in each setup. c) Central points of molecules' distances from the SWNT surfaces, averaged for the periods they are linked to 15:1 POPC-SWNTs. Color distinctions and markings remain consistent with panel b's description.

Interactions of Peptides with POPC Bilayers

We also investigated the interactions of Col, TAT, and Cro with pure POPC bilayers, maintaining the same concentration ratios as in the POPC-SWNT systems. After 1 μs of MD simulation equilibration, the peptide-bilayer interaction structures are shown in Figure 3-3a. Compared to POPC-SWNT systems, fewer molecules are bound to the pure bilayers. However, the binding modes to both systems were similar. Figure 3-3b quantifies the depth and probability of insertion into bilayers for each molecule. All three peptides reduced the order of POPC chains in the bilayer, as shown in Figure 3-3c.

In summary, the results from Figure 3-3 allow us to rank the peptides based on their disruptive activity on POPC bilayers: Col is the most disruptive, followed by TAT, then Cro. This ranking aligns with their behavior in POPC-SWNT systems, suggesting that the disruption experienced by POPC-SWNT conjugates by these molecules is like that in lipid bilayers. These membrane simulations were performed and analyzed by Ms. Anju Yadav.

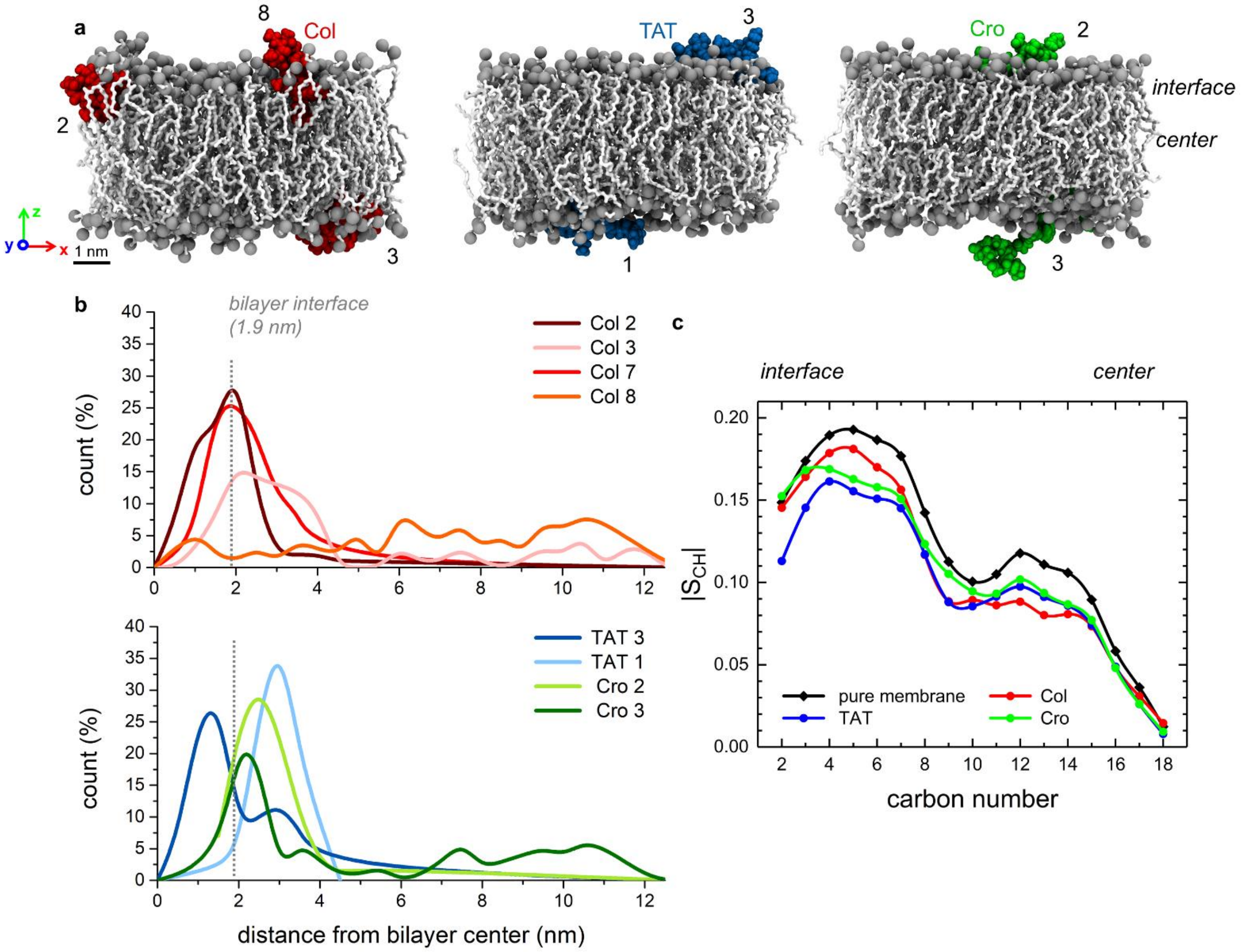


Figure 3-3. Interaction of colistin, TAT peptide, and crotamine-derived peptide with standalone POPC bilayers. a) Visual representations of stabilized POPC bilayers in the company of Col (red), TAT (blue), and Cro (green). The coloring and depiction methods mirror those used in preceding figures. b) Dispersal of central point distances (along the z-axis) for adhered Col, TAT, and Cro molecules relative to the central point of the bilayer. c) Structuredness measure of POPC bilayers (both unadulterated and in conjunction with Col, TAT, or Cro molecules).

3.4. Conclusion

We studied the behavior of POPC-coated (6,5) SWNT conjugates through MD simulations. Contrary to our initial expectations based on prior studies with (18,0) SWNTs, the SWNTs in our simulations were asymmetrically positioned within the POPC coronas, as depicted in Figure 3-1. This deviation suggests that the SWNT's position in lipid assemblies might be influenced by the type of lipid and possibly the SWNT size. The intricacy of our findings and those from previous studies highlight the need for more in-depth research.

Subsequently, we explored the interactions between POPC-SWNT conjugates and three known cell membrane disruptors: colistin, TAT peptide, and crotamine-derived peptide. We also analyzed these peptides' interactions with POPC bilayers. Our findings revealed that Col and TAT deeply penetrate both POPC bilayers and the POPC corona on SWNTs, while Cro primarily adheres to the POPC surface in both scenarios. These interactions align with previously documented interaction mechanisms between antimicrobial cationic peptides and lipid bilayers.

Another unique aspect of POPC-SWNT systems is the presence of a distinct interface where peptides can bind, which is absent in POPC bilayers. This interface is the thinnest part of the POPC corona, where the SWNT surface is partially exposed. Interestingly, peptides like Col and TAT, which cause the most disruption in bilayers, also induce significant perturbations at this POPC-SWNT interface.

In essence, our simulations allowed us to rank the peptides based on the disruption they cause in POPC-SWNT conjugates: Col being the most disruptive, followed by TAT, and then Cro. This ranking is consistent with the behavior of these peptides when interacting with POPC bilayers. The degree of structural disruptions aligns with the experimentally observed nIR optical signal changes in lipid functionalized SWNTs upon peptide addition. Our findings suggest that lipid-functionalized SWNTs can serve as simplified cell membrane models, with their optical signals reflecting peptide-induced structural changes in the lipid phase at the SWNT surface. This indicates the potential of lipid functionalized SWNTs as promising platforms for preliminary screening of molecules that can disrupt cell membranes. Further testing on larger compound libraries could offer more insights into the efficacy of these systems for antimicrobial screening through optical emission responses. Additionally, exploring the optical responses of other nanomaterials, like quantum dots, to antimicrobial peptides could lead to exciting new areas of research.

Chapter 4: Discovery of DNA–Carbon Nanotube Sensors for Serotonin with Machine Learning and Near-infrared Fluorescence Spectroscopy

4.1. Introduction

Single walled carbon nanotubes (SWNT) are integral components of numerous hybrid materials crafted for nanotechnological uses, spanning areas like sensing, biological visualization, electronics, and gene transport.[38,40,63–70] A common technique to modify SWNTs and make them soluble in water is noncovalent polymer adsorption, which creates a "corona phase" on the SWNT exterior. Various polymers, such as nucleic acids, peptides, surfactants, lipids, and peptoids, have been employed for this purpose.[41,43,71–78] Notably, SWNT conjugates functionalized with nucleic acids are particularly prevalent and hold significant potential for applications like optical sensing of vital biological compounds,[38,40] delivering polynucleotides (DNA/RNA) for genetic modification,[67,79] and segregating multi-chirality SWNT samples into pure chiral groups.[33,80–83]

The DNA sequence is pivotal in DNA–SWNT conjugates used for optical sensing, as it's the primary factor determining specific molecular recognition of analytes. An effective sequence should not only bind strongly to the analyte but also to the SWNT surface, leading to a noticeable change in the SWNT's near-infrared (nIR) fluorescence, ΔF/F, when the target analyte is present. Previous studies have shown that even a minor alteration, like changing a single nucleotide in the DNA sequence, can negate the sensor's response to its target.[40]

Historically, DNA–SWNT sensors have been created either by using pre-established molecular recognition elements[84,85] or by screening a limited number (less than 100) of DNA sequences for their fluorescence changes in the presence of target analytes.[40,86] The latter method, which often results in random successful outcomes, depends on the accidental identification of potential sensors. While this method has its merits in pioneering new research field, it's not ideal for refining existing sensing technologies or creating sensors for hard-to-detect analytes. To address this, we recently introduced a technique (SELEC) that "evolves" ssDNA–SWNT-based molecular recognition towards a specific analyte, enhancing selectivity with each evolution cycle.[87] This method allows for the evolution of around 10^{10} unique ssDNA strands for target analyte recognition while still being attached to a nanomaterial's surface.

The SELEC method produces data sets rich in details about DNA sequences that offer selectivity and SWNT binding affinity. In this study, we utilize these data sets to curate a collection of approximately 100 DNA–SWNT conjugates, obtained by the lab led by Dr. Markita Landry at the University of California Berkeley. We then assess their ΔF/F nIR fluorescence response to a selected analyte. We employ this data to construct machine learning (ML) models that can predict and identify useful ssDNA sequences, which might have been overlooked in earlier experiments, that can bind to and optically detect the chosen analyte on SWNT surfaces. The Landry Lab then experimentally validated these model predictions experimentally, use the outcomes to refine our models, and predict DNA sequences that yield a higher ΔF/F response to the target. While our methodology can be adapted for various analytes, in this instance, we focus on serotonin (5-hydroxytryptamine, 5-HT), a neurotransmitter with significant functions both within[88] and outside the brain.[89] Given the critical nature of serotonin biosensing, numerous recent endeavors have been directed towards its sensor development.[84,87,90,91] Figures and tables labeled with an 'S' can be referenced in the published paper associated with this chapter.

4.2. METHOD

Data Preparation and Machine Learning (ML) Implementation

We used datasets comprising ssDNA sequences and their corresponding ΔF/F values for our preliminary ML model training and testing. These data were sourced from experiments conducted by the Landry Lab at UC Berkeley, as detailed in Ref.[87]. This data can be found in Table 4-1.

Table 4-1. presents the original collection of DNA sequences attached to single-walled carbon nanotubes (SWNT) and their change in fluorescence intensity (ΔF/F) in reaction to serotonin. All listed sequences contribute to both the training and evaluation of model M1. However, for model M1B, sequences highlighted in yellow, orange, and green are excluded as they represent repeated measurements. Each DNA sequence is bookended by C6-mer PCR primers. Most of the sequences and their ΔF/F metrics were initially documented in Ref[87], with their designated names correlating to datasets produced in that study (with E/C denoting experimental/control groups, followed by the selection round number and the individual ID within the SELEC dataset). Sequences marked as N/A were selected at random for testing purposes.

Seq ID	Sequence	ΔF/F (1195 nm)	Name
1	ACACACACAACGACGCGG	0.70223	E4#2755, **E5#10**, E2#180224
2	AGCACAACACGGCAACCT	1.63693	E4#8701, E5#24, C6#105380, **E6#1**
3	AACACACCACAGACTCTG	0.74523	E4#39060, E5#83, **E6#2**
4	ACACACCATCAGACGCCG	0.61943	E4#2479, E5#21, **E6#3**
5	AGCAGCACACGACACACT	0.96503	E5#29, **E6#4**
6	ACGCCAACACATTCCGCT	1.68843	E4#12871, E5#23, **E6#5**
7	AACACACACAGCCGTCCG	0.78507	E4#467850, E5#18, **E6#6**
8	AACACACACAGACGCACG	1.0623	E4#734956, E5#1703, **E6#7**
9	AGCACCAGACAGCACACT	1.9069	E5#137, **E6#8**
10	ACCACGATCCTCACTCCG	0.59733	E4#184836, E5#239, **E6#9**
11	ACGCACCGACAGCACACT	0.54127	E4#233419, **E5#1**, E6#1208
12	ACACCACACCACACCGAT	0.47627	E4#184339, **E5#2**, E6#43
13	ACGACAACCAACACTGTG	1.33259	**E5#3**, E6#56
14	AGCACACTACACACGGCG	0.8454	**E5#4**
15	ACACCACCTCACGACGTG	0.77627	E4#34446, **E5#5,** E6#665
16	ACACCACCAGACACTGCG	0.80127	**E5#6**, E6#1534
17	ACCAACACCAGCCGTGCG	0.63761	E4#111696, **E5#7,** E2#300713, E6#326
18	ACACACACCACACGTGCT	0.65277	E4#207701, **E5#8**, **E6#10**
19	ACACAACACCCGACGCGG	0.66363	E4#681284, **E5#9**, E6#25
20	ACACACACAACGACGCGG	0.77176	E4#2755, **E5#10**, E2#180224
21	GATCCAACCGCTGCCACA	1.3514	E3#7742, **E5#10**, E6#3137
22	ACGACGTACACTCCTCCT	1.27193	**E4#1**, E5#565, E6#1609
23	AACCGCATGTACTCTCCG	1.02657	**E4#2**, E6#445196
24	AACATGCACAGACGTCCG	1.11015	**E4#3,** E5#1033, E6#482
25	AACCATGCACAACGCGTG	1.04517	**E4#4,** E5#2084, E6#1925
26	ACACAACCTGCTCCTCCT	1.15197	**E4#5**, E5#193, E6#36
27	CCCCCCCCCCCCCCCCCC	0.81727	**E4#6,** E5#95, C6#17553, E3#14, C2#773, C3#8098, C5#1640, C4#125089, E6#6067
28	ACGCACAATCCGGCACTT	1.01673	**E4#7,** E6#441458
29	ACAGACTGCAGTCATGTG	0.76213	**E4#8**
30	ACACCAGCCACACGTGCG	0.54147	**E4#9**, E5#32, E6#103
31	ACCTGACACGATCCTATG	0.20597	N/A
32	GGCACAACGCTCGATGCT	0.62686	**E3#1**
33	ATTACAGCGGACAAGTGT	0.39338	**E3#2**

34	TAAGGCCGATCCCACTAT	0.21325	**E3#3**
35	TGACTCCATAACAGTGTG	1.15087	**E3#4**
36	GACACCCTGGACCCGTCG	0.73723	**E3#5**
37	TGGCGTACAAACCGTCTG	0.65847	**E3#6**
38	ACACACTCTACTCTTCCA	0.26397	**E3#7**
39	GACGTTGTGCCCAAGTTG	0.98223	**E3#8**
40	AAGGGACTGAAAGCAATG	1.45753	**E3#9**
41	ACCGCATCGACATGTGCT	1.01393	**E4#3165, E2#642689, C3#1153, C5#8791, C4#124653, E6#173**
42	ACCGCACGAGCCAGTGTG	0.54137	**E4#128294, C6#1, E2#855501, C2#556426, E6#4999**
43	TCACCACATTCCGCTGTG	0.65097	**E4#592353, C6#2, E2#141998, C5#10221, C4#6551, E6#21314**
44	ACCGAGAGCAGACGATGT	1.00867	**E4#299356, C6#3, C2#14685, C4#26020, E6#473021**
45	AACACCACACACGGCGCT	1.1451	**E4#3934, C6#4, E2#35033, C2#49973, C5#1, C4#13, E6#325887**
46	GCAGCGTGACTTGACGTG	1.13463	**E4#665722, C6#5, E2#1833, E3#10611, E6#2458**
47	AACACGGCCCTCATGTCG	0.58513	**C6#6, C2#331535, C5#6179, C4#40744, E6#123552**
48	AGCCGTATGCACACCTCA	0.52833	**C6#7, E2#767751, E3#480, C5#2434, C4#2983**
49	ACACACCGTTCATCCGCG	0.805	**C6#8, E3#38449, C3#252187, E6#219759**
50	GCTGATCGACGACACGTG	0.78593	**C6#9**
51	AACACCACACACGGCGCT	0.64953	**E4#3934, C6#4, E2#35033, C2#49973, C5#1, C4#13, E6#325887**
52	AGCACACTCCACTCCGCT	0.95393	**C6#320, C3#75273, C5#2**
53	GCACACACCAGCCGTCTG	0.77587	**C6#2551, C3#19037, C5#3, C4#12**
54	AACCACACACCGTCCGCT	0.9622	**E5#5415, C6#294, E3#18495, C3#244055, C5#4, C4#3825, E6#5869**
55	ACCACACCATCGACGCGT	0.97047	**C6#1396, C3#110889, C5#5, C4#4762**
56	AGCCACACGACGCGCTCT	0.39057	**E4#29995, C6#984, E2#121542, C5#6**
57	ACGGCACACACCATCGCT	0.66563	**C6#5556, C5#7**
58	ACGACACTGCACGACGCG	0.56955	**C6#195757, C5#8**
59	ACGGCAACTCCCATTCCG	0.8091	**C6#11460, C2#614315, C5#9**
60	ACGACACCACACTGCTCT	0.50943	**C6#105782, C5#10**
61	ACACAGCATCATTCCGCT	0.44783	**C3#346295, C5#12**
62	GCACCAACCAGCCGTCTG	0.874	**E4#28645, C6#321, C5#47, C4#1**
63	TCACCACATTCGACGGCG	0.4382	**C5#2646, C4#2**
64	ACCACAAGTGACTGTCCT	0.47915	**C4#3**
65	GCCGACATGACTCCTCCT	0.42083	**C6#421, C3#80010, C5#1693, C4#4, E6#410151**
66	ACACACCAATGACCTGTG	0.61373	**C5#135, C4#5**
67	TACCCACACCACACACTG	0.70733	**C6#2111, C3#398357, C5#152, C4#6**
68	ACTGCACATCGACGCGCG	0.46227	**C6#29012, C5#41, C4#7**
69	ATTGCCGCCATCCTCATG	0.6527	**C4#8**
70	AGGCCACCGTCGCACGTG	0.41914	**C4#9**
71	ACAGACCGACGTGTGCTG	0.28237	**N/A**
72	TGGGAGCCATCTTGTGCG	0.30723	**C3#1**

73	GTTCAGCCTTTTCGTTCG	0.42323	**C3#2**
74	GGAATCTCCGGCGTCTAT	0.33207	**C3#3**
75	TAGCACAGGTCGTCTATT	0.38593	**C3#4**
76	GCCAATATAGCCCTTCCG	0.79868	**C3#5**
77	AATCACTGCAATGGTCGT	0.31937	**C3#6**
78	AACACATTGACGTGCACT	0.2447	**C3#7**
79	GGGCTGTGCCGTCATGCG	0.55663	**C3#8**
80	GATGGGGAATCATGCGTG	0.27437	**C3#9**
81	GCACAATCCAGCGCACAA	0.73998	**N/A**
82	ACGACGGAACTACACACC	0.91085	**N/A**
83	GAGACTCAACCGAACACC	1.3473	**N/A**
84	ACCACACAACCGACTGTG	1.02557	**N/A**
85	AACCCCAACCACGGTTGG	0.82087	**N/A**
86	AGGACAACCCCGCGTGTG	1.2009	**N/A**
87	ACACACCGACACGGTGTG	0.3311	**N/A**
88	ACCACGACGACGACTGTG	0.53127	**E6#3500**
89	ACCAACACACACTCCGCT	0.85663	**N/A**
90	AACACACCAACACCCGCT	1.0687	**N/A**
91	ACACACACACACTCCGCT	0.49517	**N/A**
92	ACACCACCACACTCCGCT	0.32298	**N/A**
93	ACCACACAACGCTCCGCT	0.334	**N/A**
94	ACACACCGCTCTCCCTCT	0.49743	**E4#393, E5#538, E6#20**
95	ACGACATGGCACACCGAT	0.87663	**E4#26, E5#578, C3#108525, E6#28**
96	ACACAACCTGCTCCTCCT	1.41713	**E4#5, E5#193, E6#36**
97	ACACCAATCGCACTTCCG	1.4976	**E4#385, E5#232, E6#59**
98	ACACGATCCAACACTCCG	0.95543	**E4#404, E5#74, E6#94**
99	AGCACCAGACAGCACACT	0.767	**E5#137, E6#8**

The ssDNAs had an 18-nt variable segment bordered by two C6-mers on each end. The 18-nt variable segments of the ssDNAs served as the input for our models. Multiple data encodings were explored, such as position specific vectors (1-gram, denoted as psv1 and illustrated in figure 2-4) The data from Table 4-1 was categorized into two groups: class 0 for DNA sequences with low serotonin response and class 1 for those with high serotonin response. We adjusted the threshold value for class 0 (t_0) across various parameters, while maintaining the threshold for class 1 (t_1) at a constant value of 0.9. This thresholding ensured a balanced distribution of classes, essential for effective model training and testing.

We evaluated the efficacy of multiple ML classifiers, such as AdaBoost, logistic regression, linear support vector classification, and random forest. For these models, sequences were represented as 1 × 72 binary arrays, derived from the sequential arrangement of psv1 matrix columns into a linear array. Additionally, we assessed the convolutional neural network (CNN) models' performance on psv1 and term frequency vector inputs, which had shown promise in prior DNA and RNA sequence specificity predictions.[92] All models aimed to predict the likelihood of an input sequence exhibiting a high or low serotonin response. We employed the Scikit-learn library for ML model training, while Keras[93] and TensorFlow 2[94] served as the foundation for the CNN models. The primary data set used for training, as referenced in the paper[2], was instrumental in this study. The CNN models with psv_1 encoding exhibited the best performance, leading to the exclusive use of the CNN methodology for subsequent models applied to extended data sets.

All coding for ML classification and regression model training is accessible on GitHub at https://github.com/vukoviclab/DNAsensor.

Evaluation Criteria

Our CNN models were designed to forecast the serotonin response of 18-nt DNA sequences, or more precisely, to predict the likelihood of these sequences belonging to either class 0 or class 1. Both class probabilities were assessed independently. Sequences predicted to have a high response were identified based on normalized class 1 probabilities exceeding 0.5.

For every model, we computed various metrics like accuracy, precision, recall, f1 score, ROC curves, and AUC, while also tracking TP, TN, FP, and FN values. We used standard definitions to calculate accuracy, precision, recall, and f1 score. For ML models, singular values of precision, recall, and f1 score were determined. In contrast, CNN models provided two sets of these metrics, enabling separate evaluations for sequences with both low and high serotonin responses. All model performances were further assessed using ROC curves and AUC values. For each CNN model, two ROC curves and corresponding AUC values were derived, one for low-response sequences and another for high-response sequences. For most data sets, 200 models were crafted with varying random states. For certain model groups, we provided a range of statistical evaluations related to model quality metrics.

PCA Assessment

The top 200 sequences from R6E and R6C SELEC data sets underwent principal component analysis (PCA) using the Scikit-learn library. We then analyzed the positions of some experimentally tested sequences within the established PCA framework.

4.3. RESULTS AND DISCUSSION

Categorizing DNA Sequences in DNA–SWNT Conjugates by Their Optical Reaction to Serotonin

Our initial endeavor was to develop and evaluate classifier models that could predict 18-nucleotide (nt) ssDNA sequences' relative nIR fluorescence response to serotonin after being conjugated to SWNT. We began by training models on a foundational data set of 96 distinct ssDNA sequences, which were pinpointed from prior SELEC studies. This foundational data set was curated from a comprehensive library of roughly 1010 18-nt ssDNA sequences that either bind competitively to SWNT (control) or to SWNT in serotonin's presence (experimental).[87] The SELEC method, depicted in Figure 4-1a, underwent multiple selection rounds, yielding data sets of chosen DNA sequences and their prevalence. The primary 96 sequences, mainly the most prevalent ones from SELEC rounds 3 to 6, were selected for subsequent serotonin response spectroscopic evaluations, thereby creating the foundational data set for model training. We measured the nIR fluorescence emission for these 96 distinct ssDNA–SWNT conjugates before and after introducing 100 μM serotonin. This data has been previously documented in ref (31) and is displayed in Figure 4-1b. The conjugates' response to serotonin was derived from the fluorescence emission spectra for the dominant (8,6) chirality peak (centered around ~1195 nm) using the formula $\Delta F/F = (Fa - F)/F$, where F represents the fluorescence signal before serotonin's addition, and Fa represents the signal post-addition (Figure 5-1a). The $\Delta F/F$ values for sequences in the foundational data set span from 0.2 to 1.9. This foundational data set, color-coded based on their SELEC group affiliation (experimental, control, or neither), is also depicted in Detailed in Ref[2].

Using the methodology illustrated in Figure 4-1c, we trained and evaluated convolutional neural network (CNN) classifier models on the acquired data set of 96 DNA sequences and their associated $\Delta F/F$ values (Figure 4-1b). This data set was bifurcated into two categories based on serotonin response: class 1 sequences exhibited a pronounced serotonin response ($\Delta F/F$ threshold $t_1 > 0.9$), while class 0 sequences had a subdued response (with variable $\Delta F/F$ thresholds of $t_0 <$ 0.85, 0.8, 0.7, 0.6, and 0.5). These thresholds were chosen to discern sequences resulting in DNA–SWNT conjugates with either an exceptionally high or notably low response to the target analyte, information crucial for the practical application of DNA–SWNT sensors. The $t_1 > 0.9$ threshold meant that class 1 encompassed 32% of sequences from the entire data set (31 out of 96 sequences). The fluctuating t0 threshold allowed us to study the impact of data set dimensions and equilibrium on ML model efficacy and consistency. Larger thresholds would result in more sequences in class 0, imbalanced classes, more robust models, and the learning of sequences with a median serotonin response. Conversely, smaller thresholds would yield fewer entries in class 0, balanced classes, less consistent models, and the learning of sequences with a minimal serotonin response. The CNN models' input was ssDNA sequences, transformed into position-specific vector (psv_1) format with binary values (an example is provided in Figure 2-4). The trained CNN models' outputs are the likelihoods of the input sequences belonging to either class 0 or class 1 (evaluated independently).

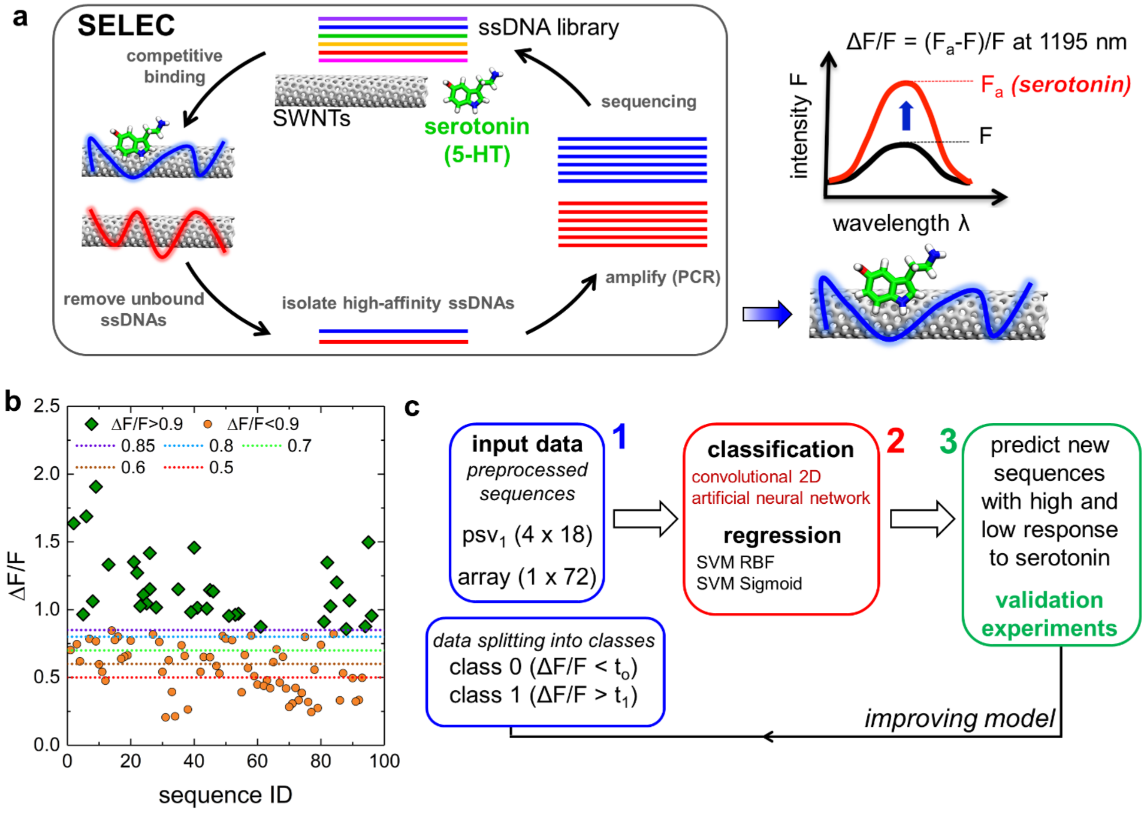


Figure 4-1. Methodology for identifying DNA sequences in DNA-SWNT conjugates that exhibit a significant reaction to serotonin. a) A SELEC protocol, conducted for up to 6 cycles, isolates ssDNAs with a pronounced affinity for SWNTs and, distinctively, SWNTs when serotonin is present. A subset of these high affinity ssDNAs are then chosen for detailed fluorescence emission spectroscopic studies of their conjugates with SWNTs both pre and post the introduction of 100 µM serotonin. b) The optical shift, represented as ΔF/F, of 96 distinct ssDNA-SWNT combinations when exposed to 100 µM serotonin. This dataset also encompasses repeat readings for 4 of these sequences. c) The primary computational strategy involves converting DNA sequences into either the binary psv1 format or a basic binary sequence. These sequences are then categorized based on their optical reactions, using specific ΔF/F thresholds. Both the sequences and their corresponding ΔF/F values serve to educate both classification and predictive models. The top-performing models then forecast sequences with either heightened or diminished reactions to serotonin. These predictions undergo experimental verification. The results from these tests then guide the creation of subsequent models.

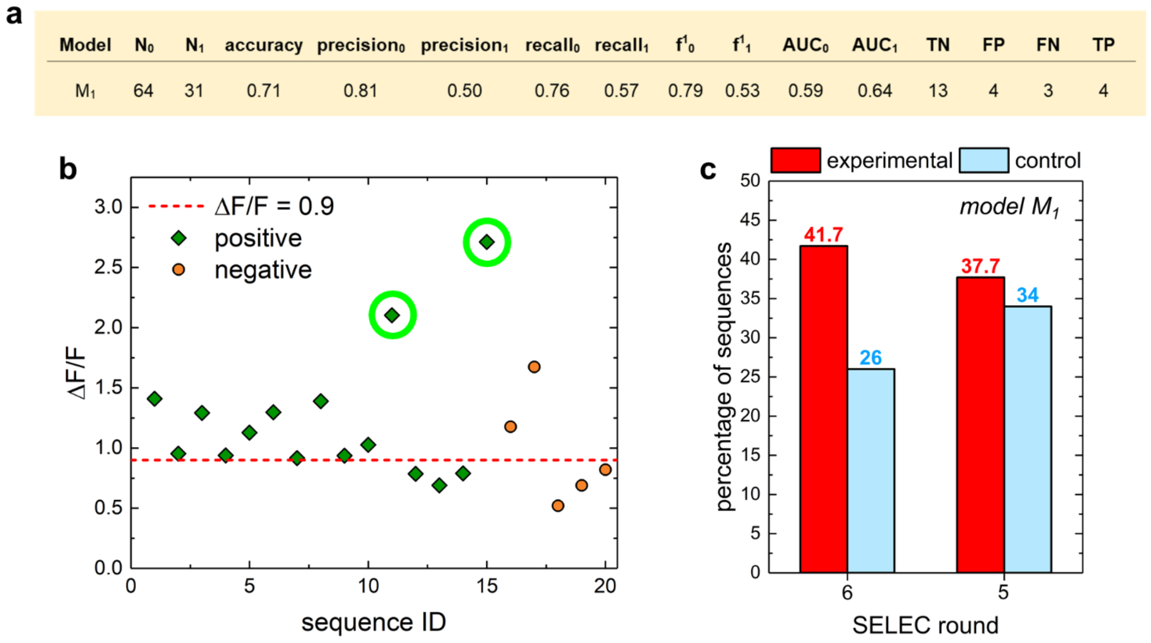
One of the most effective CNN models trained on the foundational data set, labeled M1, has its quality metrics displayed in Figure 4-2a. For M1, the area under the receiver operating curve (AUC) values were 0.59 for predicting class 0 sequences and 0.64 for class 1 sequences. Meanwhile, precision/recall values were 0.81/0.5 and 0.76/0.57 for predicting class 0 and class 1 ssDNA sequences, respectively. The models were sensitive to the removal of several sequences from the input, especially those from class 1. This sensitivity was evident when attempting to develop a high-quality CNN model trained on a reduced foundational data set containing only 93 data points. Quality metrics for a representative model, M1B, crafted with 93 data points from the foundational data set, are detailed in Ref[2]. In general, while the CNN methodology yielded decent, albeit not outstanding, quality metrics, we opted to employ it for DNA sequence classification. This decision was made because several other tested ML techniques, including AdaBoost, logistic regression, support vector classification, and random forest, consistently produced unsatisfactory models. These alternative methods invariably resulted in class 0 and class 1 probabilities of 0.5 ± 0.2, signifying a lackluster distinction between sequences with high and low serotonin responses when trained using these methods (are detailed in Ref[2]). Separately, we favored the psv1 sequence encoding over other encoding types reported by others[95] that we also evaluated.

Subsequently, we analyzed model M1's predictions for the most prevalent DNAs from the control and experimental SELEC data sets. We employed model M1 to categorize the 300 most prevalent DNA sequences from rounds 6 and 5 of the experimental (R6E/R5E) and control (R6C/R5C) SELEC data sets, excluding any overlapping sequences present in both experimental and control data sets from identical rounds. Model M1 anticipates that 41.7%/37.7% of R6E/R5E sequences and 26%/34% of R6C/R5C sequences will exhibit a high $\Delta F/F$ serotonin response (Figure 4-2c). Given that sequences from the experimental data set were chosen based on their strong affinity for SWNTs in serotonin's presence, in contrast to the sequences from the control data set, we anticipated that the experimental data sets would contain a higher number of serotonin-responsive sequences. This assumption aligns with the predictions presented in Figure 4-2c.

To verify the predictive accuracy of model M1, we chose 20 DNA sequences from the top 300 prevalent sequences in the R6E SELEX dataset. These sequences underwent experimental validation by the Landry Lab at UC Berkeley. Based on model M1's probability predictions for these sequences to belong to classes 0 or 1, 15 of the sequences are anticipated to have a high serotonin response, while the remaining 5 are expected to have a low response (SI, Table S4). Figure 4-2b and SI, Table S4 showcase the experimentally measured ΔF/F values for the chosen DNA sequences. Notably, 12 out of the 15 predicted high-response sequences exhibited ΔF/F values surpassing the class 1 threshold, $t_1 > 0.9$ (80%, derived from 12 out of 15 sequences). Moreover, the validation tests identified two sequences with ΔF/F values of 2.1 and 2.7, indicating a stronger serotonin response than any sequence from the foundational data set. These sequences correspond to ID#90 (with 8 reads, ΔF/F = 2.1) and ID#115 (with 7 reads, ΔF/F = 2.7), based on their read counts in the R6E data set. On the other hand, 3 out of the 5 predicted low-response sequences recorded ΔF/F values below the class 0 threshold, t0 < 0.85. Overall, there wasn't a statistically significant correlation between the predicted probabilities and the experimental ΔF/F values (SI, Figure S3).

Performance of Individual CNN Models Trained on Our Extended Data Set

To determine if incorporating more experimental data points could yield more predictive models, we trained a second representative CNN model, M2, on an expanded data set of 113 sequences. This data set combined the foundational data set (singly measured sequences in Figure 4-1b with additional experimental data from the first set of validation tests Figure 4-2b more detail are available in the ref paper[2]). A representative model, M2, achieved an accuracy of 0.64, AUC values of 0.71 for predicting class 0 sequences, and 0.75 for class 1 sequences. Additionally, precision/recall values were 0.77/0.59 and 0.53/0.73 for predicting class 0 and class 1 ssDNA sequences, respectively. Besides the enhanced AUC values, the M2 model exhibited significantly improved ROC curves compared to models M1 and M1B (Figure 4-3b, and are detailed in Ref[2]).

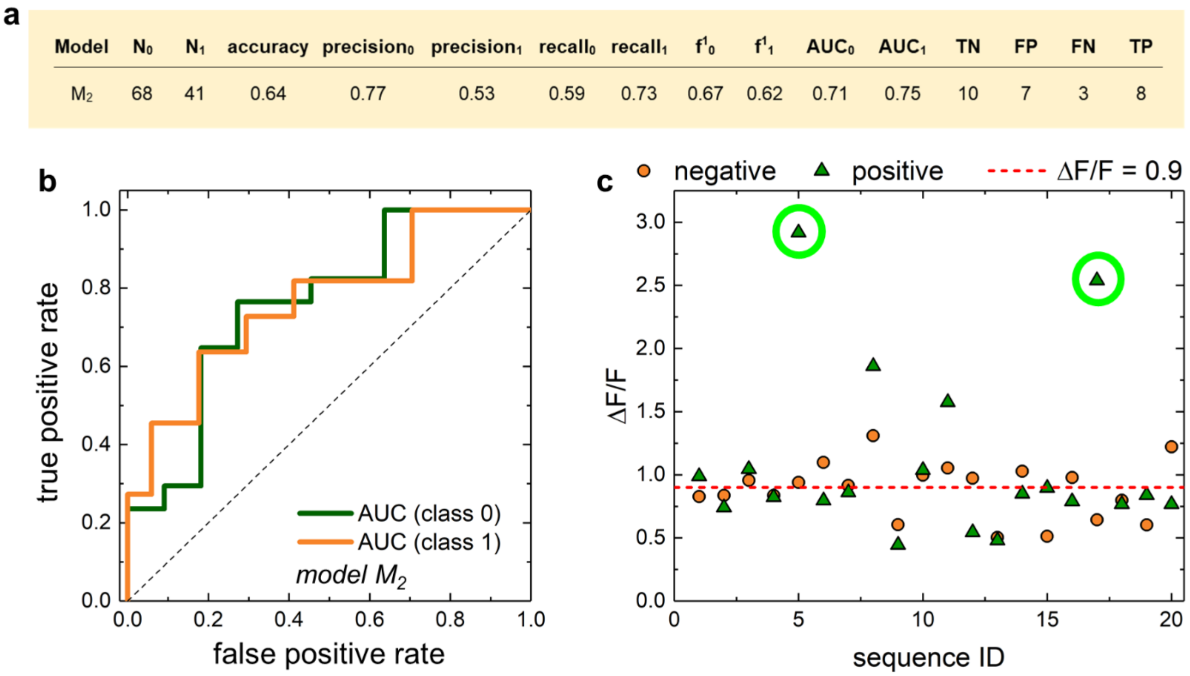


a

Model	N_0	N_1	accuracy	$precision_0$	$precision_1$	$recall_0$	$recall_1$	f^1_0	f^1_1	AUC_0	AUC_1	TN	FP	FN	TP
M_1	64	31	0.71	0.81	0.50	0.76	0.57	0.79	0.53	0.59	0.64	13	4	3	4

Figure 4-2. Assessment of a typical CNN model based on the primary data collection. a) Review of a standard CNN M1 model educated using the preliminary dataset, with criteria set at $t_1 > 0.9$ and $t_0 < 0.85$. b) The optical shift, symbolized as ΔF/F, of ssDNA-SWNT combinations when introduced to 100 μM serotonin. This data is derived from 20 novel ssDNA sequences that were forecasted by model M1 to exhibit either a strong (marked as positive) or weak (indicated as negative) reaction to serotonin. Sequences that present ΔF/F measurements surpassing 1.9, the peak values from the initial data set, are highlighted with green circles. c) Ratio of sequences from R6E, R5E, R6C, and R5C SELEC collections that model M1 anticipates being highly reactive to serotonin. This percentage is deduced from the inaugural 300 SELEC data sequences in the designated experiment/control set that aren't found in the associated control/experiment set.

The Landry Lab at UC Berkeley experimentally validated these sequences. To validate the accuracy of model M2's predictions, we selected 40 DNA sequences from the 280 untested most prevalent sequences in the R6E dataset. Among these 40 DNA sequences, half were predicted by M2 to have a low response (termed as negative), and the other half were anticipated to have a high response (termed as positive) (are detailed in Ref[2]). Figure 4-3c display the experimentally measured ΔF/F values for the chosen DNA sequences. Model M2 tends to overestimate false positive sequences since only 7 out of the 20 sequences (35%) predicted to have a high serotonin response recorded ΔF/F values exceeding the class 1 threshold of $t_1 = 0.9$. The remaining 13 out of 20 sequences (65%) registered ΔF/F < 0.9. Single models like M2 might be predicting sequences with responses akin to those of randomly selected sequences. Notably, model M2 identified two previously unknown sequences from the R6E evolution group with a robust serotonin response, recording ΔF/F values of 2.5 and 2.9. These sequences are associated with ID#264 (with 6 reads, ΔF/F = 2.5) and ID#156 (with 7 reads, ΔF/F = 2.9), based on their read counts in the R6E evolution group. Separately, while all sequences predicted to have a low serotonin response recorded ΔF/F < 1.3, only 9 out of the 20 sequences (45%) had ΔF/F values below the class 0 threshold of $t_0 = 0.85$.



Model	N_0	N_1	accuracy	precision$_0$	precision$_1$	recall$_0$	recall$_1$	f^1_0	f^1_1	AUC$_0$	AUC$_1$	TN	FP	FN	TP
M_2	68	41	0.64	0.77	0.53	0.59	0.73	0.67	0.62	0.71	0.75	10	7	3	8

Figure 4-3. Assessment of CNN models based on the augmented data collection. a) Analysis of a typical CNN M2 model developed using the enlarged dataset, with criteria set at t1 > 0.9 and t0 < 0.85. b) ROC trajectories for the M2 model when forecasting sequences for class 0 and class 1. c) Optical shift, denoted as ΔF/F, for ssDNA-SWNT combinations when exposed to 100 µM serotonin, derived from 40 novel ssDNA sequences. Sequences that present ΔF/F measurements that surpass 1.9, the apex value from the preliminary data collection, are highlighted with green circles.

Predicting High-Response DNAs Using Combined Classification and Regression Models

During the training of models M1 and M2, we observed their unpredictable behavior and their reliance on the random state variable (resulting in different training/testing data set divisions) chosen during the training process. To understand the unpredictability of these models trained on our limited data sets of approximately 100 sequences, we subsequently analyzed their accuracy and f1 scores. This analysis was conducted on 200 CNN models trained on the expanded data set using various random state variables and several t_0 threshold values (0.5, 0.6, 0.7, 0.8, 0.85). The accuracy and f1 score distributions of these models, depicted in Figure 4-4a, b, vary between 0.4 to 0.93 and 0.2 to 0.9, respectively. Although these distributions cover a broad spectrum, most models have accuracy and f1 scores above 0.5, indicating their predictive potential. Furthermore, over half of the models in Figure 4-4b have accuracy and f1 scores exceeding 0.6. Intriguingly, predictions of high-response sequences are of superior quality for smaller t_0 thresholds (are detailed in Ref[2]). Model unpredictability diminishes, and model consistency enhances when data sets contain 500 or more sequences for each class (Figure 4-4c, input sequences sourced from SELEC data sets).

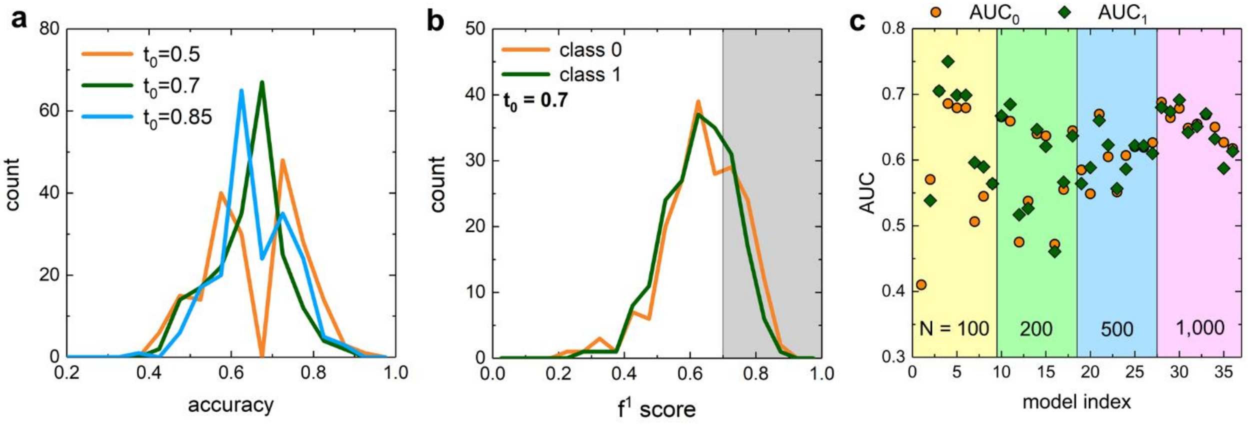


Figure 4-4. Variability in CNN model outcomes. a) Display of accuracy metrics across 200 CNN models using psv1 input, determined by varying random state settings for multiple t0 values. b) Spread of f1 score metrics across 200 CNN models, achieved with varying random states at a fixed t0 value of 0.7. c) Relationship between dataset volume and model consistency. AUC metrics for nine distinct CNN models, each developed using 100, 200, 500, and 1,000 sequence sets from two classes, sourced from R6C (class 0) and R6E (class 1) SELEC data collections. Every model is characterized by a unique random state setting.

With the goal of predicting DNA sequences with the highest ΔF/F values, we subsequently trained regression models that predict ΔF/F values based on DNA sequence input. These regression models were developed using the support vector machine (SVM) regression algorithm with radial basis function (RBF) and sigmoid kernels, inspired by the successful application of these algorithms for sequence input.[96] One of the most effective SVM RBF regression models, trained on the expanded data set with sequences having ΔF/F > 0.9 and ΔF/F < 0.6, is illustrated in Figure 5a. There's a strong correlation between the ΔF/F values of test sequences obtained experimentally and those predicted by this SVM model, with r^2 = 0.448 and a Pearson coefficient of rPearson = 0.67 (p-value = 0.001). However, like classification models, the quality of regression models also hinges on the random state variable. This dependency is evident in the r^2 value distributions for 200 models crafted with SVM RBF and SVM sigmoid methods, various random state variables, and different t_0 values (Figure 5-5b, c). For both SVM RBF and SVM sigmoid methods, r^2 values range from negative values to 0.5, with SVM RBF models generally being of superior quality compared to SVM sigmoid models.

Considering the substantial number of effective classification and regression models we developed, our next step was to evaluate if merging these model predictions could help identify DNA sequences with either high or low serotonin responses. To achieve this, we trained CNN models using input data with set thresholds: t_1 = 0.9 and either t_0 = 0.5 or 0.6. This resulted in values of $f1_0$ and $f1_1$ exceeding 0.6. These models were then employed to predict the responses of 3000 untested, abundant R6E sequences. In parallel, the top-performing regression models, with an r2 value greater than 0.45 and trained on an expanded dataset with sequences within the thresholds t1 > 0.9 and t0 < 0.5, were used to estimate the ΔF/F values for these 3000 sequences.

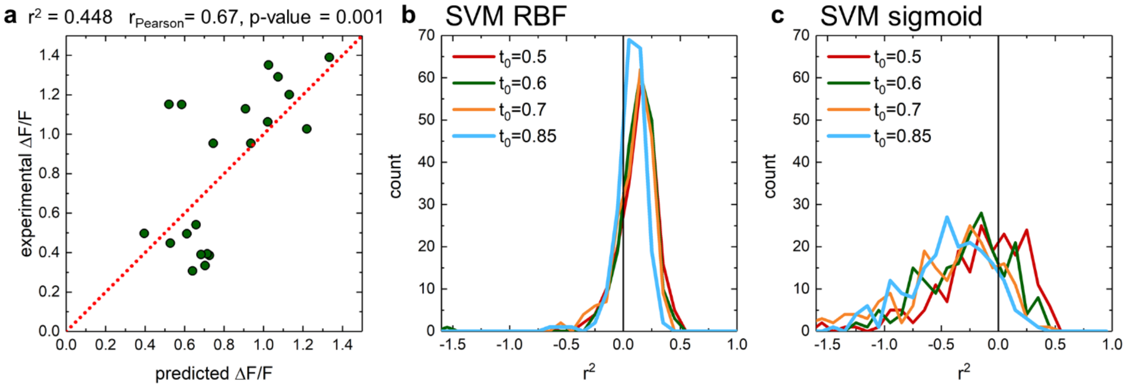


Figure 4-5. Utilizing SVM regression models to forecast ΔF/F values in ssDNA-SWNT combinations. a) A juxtaposition of experimentally derived ΔF/F values with those projected by a standout SVM RBF regression model, which was trained on the augmented dataset using criteria of $t_1 > 0.9$ and $t_0 < 0.6$, and a specific random state variable. b) Spread of r^2 metrics across 200 SVM RBF models, generated using varied random state settings. c) Dispersion of r^2 metrics for 200 SVM sigmoid models, acquired through different random state configurations.

After organizing the sequences based on their predicted ΔF/F values from regression, we selected the top 10 sequences that the CNN models also identified as having a high response, due to their consistently high or low likelihood of belonging to class 1 or class 0, as shown in Figure 6a. For a comparative analysis, we also picked the 10 sequences ranked at the bottom, which the CNN models identified as having a low serotonin response, based on their consistent probabilities, as depicted in Figure 4-6a. The likelihood of these 20 sequences being in class 1 or class 0 was further confirmed by an ensemble of multilayer perceptron artificial neural network (MLP-ANN) models, as illustrated in ref paper[2]. This ensemble incorporated models trained with thresholds t_1 = 0.9 and t_0 = 0.7, achieving $f1_0$ and $f1_1$ scores values above 0.6. The probabilities derived from the MLP-ANN models (refer to paper[2]) aligned closely with the trends observed from the CNN models (as seen in Figure 4-6a).

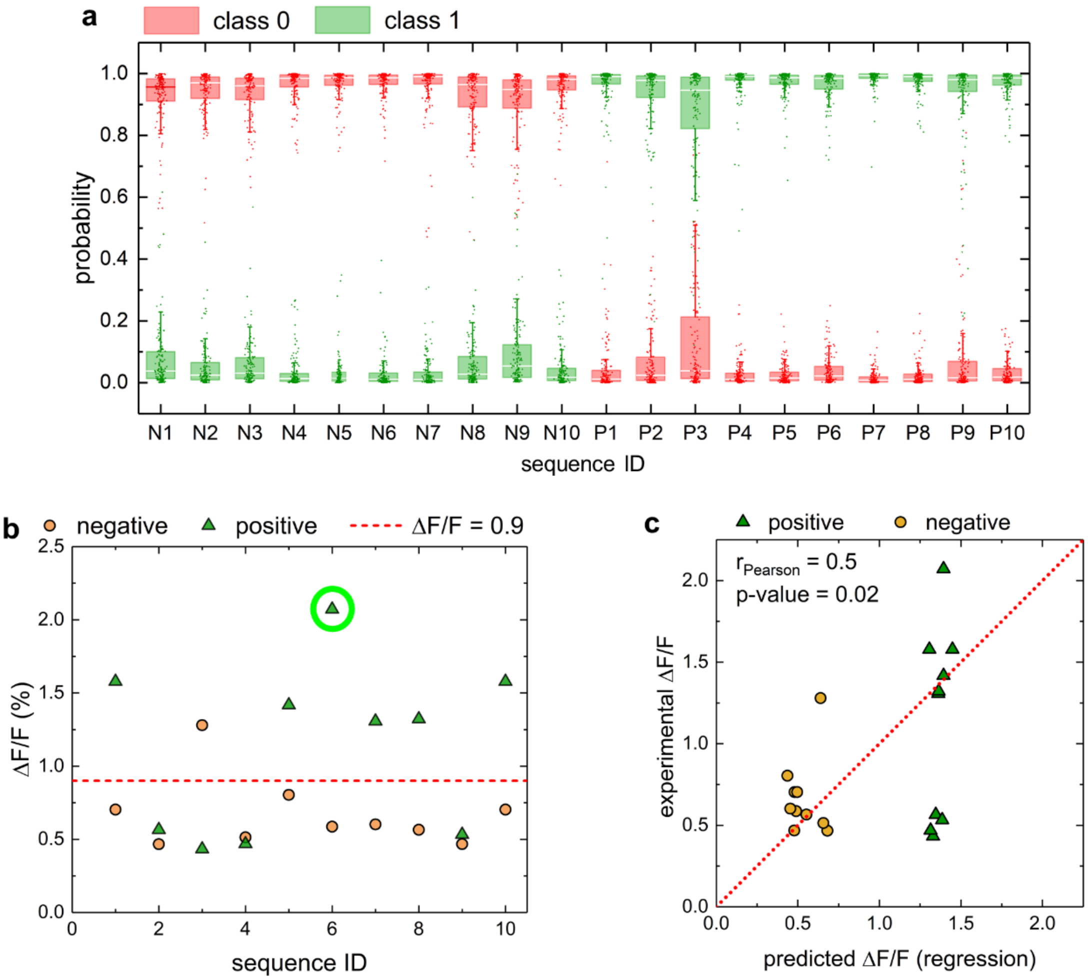


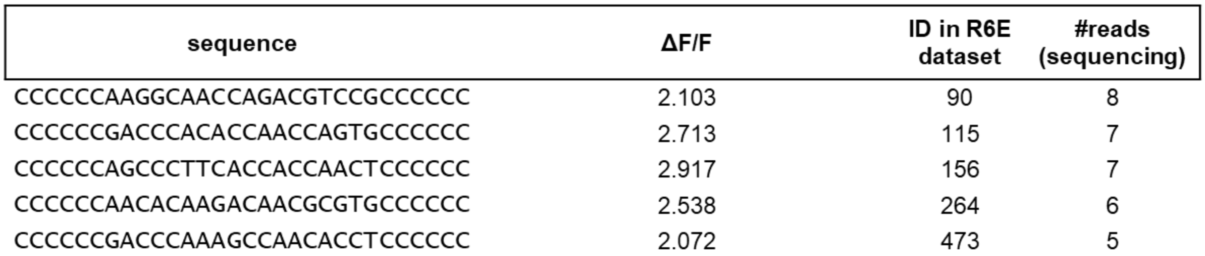
Figure 4-6. Anticipating the reaction of DNA sequences to serotonin utilizing numerous top-tier classification and regression models. a) Predicted likelihoods for 20 DNA sequences to exhibit a strong (class 1) or weak (class 0) reaction to serotonin; these sequences were handpicked based on the outputs of several elite classification and regression models, as elaborated in the main content. b) The optical shift, depicted as ΔF/F, for ssDNA-SWNT combinations when introduced to 100 μM serotonin, derived from those 20 ssDNA sequences. Any sequence that displays a ΔF/F measurement surpassing 1.9, the peak value from the initial data collection, is accentuated with a green circle. c) A side-by-side comparison of both experimentally obtained and projected ΔF/F values for those specific 20 ssDNA sequences.

Subsequently, we conducted experimental evaluations on the top and bottom 10 sequences, labeled as positive and negative respectively (refer to Figure 4-6b). Remarkably, 60% (or 6 out of 10) of the positive sequences displayed ΔF/F responses surpassing the class 1 threshold of $t_1 = 0.9$, with one sequence even achieving a ΔF/F value of 2.1 (as are detailed in Ref[2]), which is higher than any value in our initial dataset (1.9). On the other hand, 90% (or 9 out of 10) of the negative sequences had ΔF/F responses below the class 1 threshold of $t_1 = 0.9$. Notably, there was a significant correlation between the experimentally measured and predicted ΔF/F values (as seen in Figure 6c), with a Pearson correlation coefficient (r_{Pearson}) of 0.5 and a p-value of 0.02.

4.4. Conclusions

In this study, we utilized machine learning (ML) techniques to identify DNA–SWNT sensors that exhibit a strong nIR fluorescence response to serotonin. Previously, we selected 96 DNA–SWNT sensors based on the frequency of each DNA sequence from SELEC experiments, assuming that sequence abundance was the primary indicator of sensor performance. This method overlooked less common DNA sequences. Our findings indicate that ML can enhance this abundance-driven approach. ML models can independently discern the relationship between DNA sequences and their fluorescence reactions to substances, aiding in the selection of superior sensor candidates. Using the most frequent sequences from SELEC experiments ensures that the initial dataset for ML training is rich in high-response sequences, a scenario less probable with entirely random selections.

Table 4-2. Newly developed DNA-SWNT detectors specific to serotonin and their ΔF/F values, as determined by the Landry lab. The identification numbers and count of reads are sourced from the R6E SELEC data collection.



sequence	ΔF/F	ID in R6E dataset	#reads (sequencing)
CCCCCCAAGGCAACCAGACGTCCGCCCCCC	2.103	90	8
CCCCCCGACCCACACCAACCAGTGCCCCCC	2.713	115	7
CCCCCCAGCCCTTCACCACCAACTCCCCCC	2.917	156	7
CCCCCCAACACAAGACAACGCGTGCCCCCC	2.538	264	6
CCCCCCGACCCAAAGCCAACACCTCCCCCC	2.072	473	5

ML models can interpret patterns from previously tested datasets and forecast potential DNA sequences. Our evaluation of various ML techniques revealed that convolutional neural network (CNN) classifier models are most effective for datasets of around 100 DNA sequences, a common size in sensor research. Initially, we trained two individual CNN classifier models to predict DNA sequences with high serotonin responses. While the prediction accuracies from the test data varied from experimental results, these models still identified several DNA sequences with enhanced serotonin responses than earlier experimental findings. Additionally, we developed regression models to predict the relative nIR fluorescence response based on DNA sequences.

Our analysis of various models, considering different data splits, revealed that both classification and regression models based on our limited datasets are somewhat unpredictable. Still, most models are predictive, with the majority achieving over 50% accuracy. Our investigation into the unpredictability based on dataset size suggests that consistent models can be realized with datasets containing 500 samples per category. Given the challenge of obtaining such extensive experimental datasets, we combined predictions from top-performing CNN classifiers and SVM regression models, drawing inspiration from model ensembling strategies. This integrated method achieved 60% accuracy for high-response sequences and 90% for low-response sequences, underscoring its potential in predicting effective DNA sequences and hastening sensor development. A simpler principal component analysis (PCA) method seemed promising in preliminary analyses but lacked strong correlation in validation tests, unlike the successful CNN classifiers and SVM regression models. Many sequence patterns were observed in high-response sequences, but most were infrequent, emphasizing the advantage of ML in predicting high-response sequences from existing datasets.

In summary, our ML strategies led to the discovery of five serotonin DNA–SWNT sensors, as detailed in Table 4-2. Notably, these sensors outperformed those identified through traditional manual screening based on sequence abundance in the R6E SELEC library. The capability of our models to predict non-responsive DNA sequences is also crucial for sensor design. Collectively, our findings indicate that ML can efficiently pinpoint high-performing DNA sequences, potentially accelerating advancements in fields reliant on DNA–SWNT combinations, such as biosensors, bioelectronics, and SWNT chirality separation.

Chapter 5: BinderSpace: A package for sequence space analyses for datasets of affinity-selected oligonucleotides and peptide-based molecules

5.1. INTRODUCTION

The discovery of molecules that strongly bind to target molecules is a crucial step in creating new therapeutic treatments and research tools.[97,98] However, finding such molecules is a complex task, especially when they need to bind selectively to the target and possibly perform other specific functions, like emitting light signals or triggering targeted protein degradation.[99] A common method for finding these target-binding molecules involves creating and testing combinatorial libraries of molecules, which can either stand alone or be attached to entities like oligonucleotides or phages. In display tests, the sequence of the oligonucleotide or phage genome can be decoded using next-generation sequencing, giving each molecule in the library a unique identifier. These encoded molecule libraries can be screened in one go for target binding, often resulting in the selection of high-affinity binders.

The libraries can consist of various molecules, including oligonucleotides, modified nucleic acid polymers,[100] small molecules,[101,102] and different types of peptides.[103–108] One technique for selecting target binders from diverse oligonucleotide sequences[109,110] is known as systematic evolution of ligands by exponential enrichment (SELEX). This method involves multiple rounds where, in each cycle, high-affinity binders become increasingly prevalent. SELEX typically produces a dataset of oligonucleotide sequences that have been enriched for target binding. Targets in SELEX can vary from small molecules,[111] to cancer cells,[112] to carbon nanotubes.[87] Once high-affinity binders, or aptamers, are identified, they can be utilized in various ways.

Over the past decades, methods for high-throughput screening of peptide-based molecule libraries have also been developed. As noted, experimental display technologies[96,113,114] like phage or mRNA display are commonly used for peptide ligand discovery. These techniques can create and test vast libraries of peptide molecules (up to 10^{10}). Each library usually has a single structural base with peptide segments that have both fixed and variable amino acid positions, sometimes with added synthetic fragments.[103–105,115] Like SELEX, peptide selection from phage or mRNA display libraries typically results in large datasets of target-binding sequences. While these datasets likely contain high-affinity binders, identifying the best binders requires more detailed, low-throughput experiments that measure the equilibrium dissociation constant K_D between each molecule and its target. As these experiments can only be done on a limited scale, further methods or experiments are needed to identify the highest affinity molecules.

With the rise of artificial intelligence (AI), there's growing interest in using experimental selection datasets to train machine learning (ML) models. These models can predict molecules, whether oligonucleotides or peptides, with a high affinity for targets[101,116] or those that can produce a specific functional response, like fluorescence.[2,117] Bioinformatics can help understand the sequence makeup of experimental datasets and assess patterns in molecules predicted by ML models to have high target affinity. For instance, sequence motif analyses can identify motifs prevalent in experimental datasets, which is valuable since motifs often play a role in target affinity.[118] Comparative analyses of control and selection datasets can also guide the selection of candidate molecules for more detailed experiments. Various bioinformatics tools have been developed for motif discovery in DNA, protein, and peptide datasets, with applications ranging from gene regulation to the discovery of therapeutic peptides.[119] However, many of these tools, like MEME[120] and MERCI[121], might not be optimized for analyzing selection datasets.

Another valuable bioinformatics tool is the visual analysis of reduced-dimensionality sequence spaces, using methods like principal component analyses (PCA) and t-distributed stochastic neighbor embedding (t-SNE). These methods can reveal structures within large datasets and have been applied in various biological contexts, including DNA methylation[122] and single-cell transcriptomics.[123] Clustering methods can further categorize sequences in these reduced spaces, helping identify sequences with similar properties.

To our knowledge, no single toolkit currently combines motif discovery with reduced-dimensionality sequence space visualizations for oligonucleotide and peptide-based selection datasets. In response, we've created BinderSpace, a Python package designed for efficient analysis of sequence compositions from selection processes. BinderSpace can analyze motifs in DNA, RNA, or peptide sequences, visualize sequences in reduced spaces, and cluster sequences. We showcase BinderSpace's capabilities using datasets of oligonucleotides selected for binding to carbon nanotubes in the presence of serotonin[2,87]and cyclic peptidomimetics chosen for binding to bovine carbonic anhydrase protein.[104]

5.2. Methods

In this section, we introduce the BinderSpace package, designed to analyze datasets of related DNA and peptide-based sequences derived from selection experiments. As illustrated in Figure 1, BinderSpace offers functionalities such as motif analysis, visualization of sequence space, cluster analysis, and extraction of sequences from specific clusters.

BinderSpace is a Python3-based open-source tool. Its source code is accessible on GitHub and can be installed or downloaded from the Python Package Index (PyPI) repository. The repository includes a code folder and a tutorial-style example analysis folder.

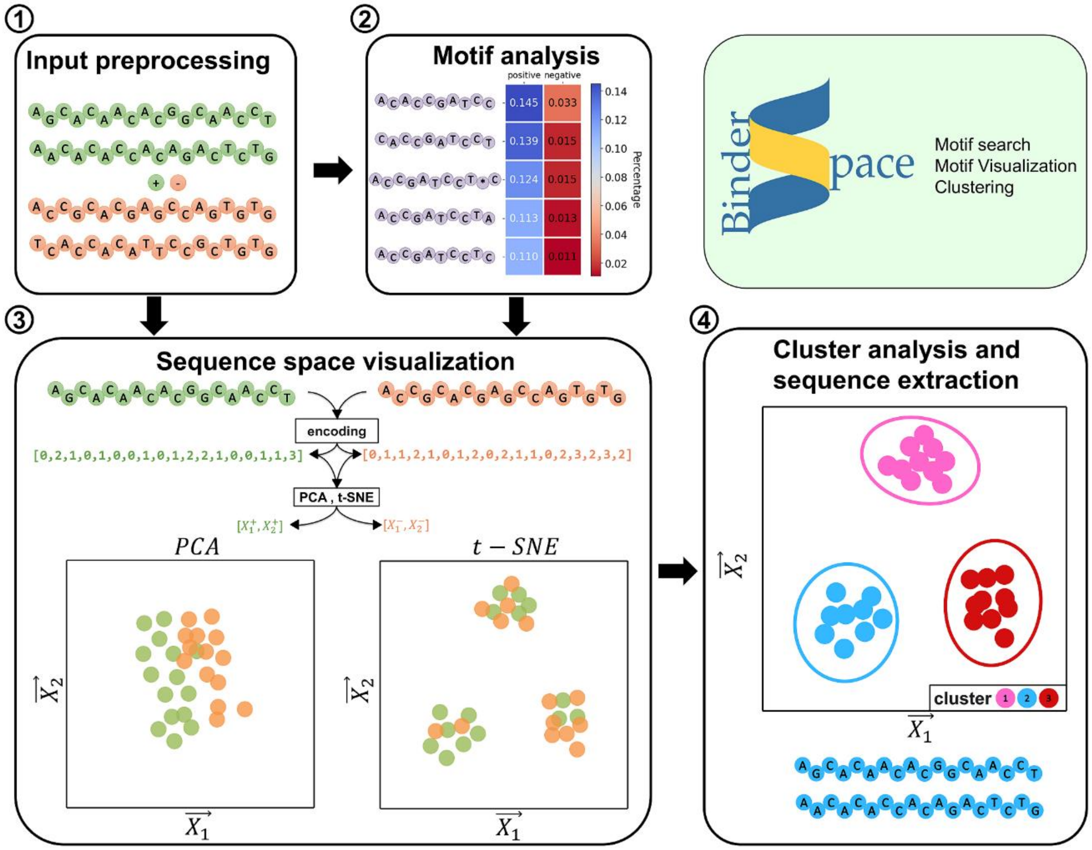


Figure 5-1. A breakdown of the capabilities of the BinderSpace tool. Inputs for the package (box 1) consist of DNA or peptide molecular sequences. This includes molecules known for their strong affinity to a target (positive dataset) and those from a control set, which are presumed not to bind strongly to the target (negative dataset). The initial result (box 2) involves evaluating motifs within both positive and negative sets. The subsequent output (box 3) is a graphical representation examining the sequence distribution of both sets. The final output (box 4) analyzes the cluster evaluation of these sequence space representations, allowing for the extraction of molecular sequences from desired clusters.

Motif Search in Affinity-Selected Molecule Datasets

The initial task in BinderSpace when analyzing datasets of affinity-selected DNA or amino acid-based molecules is motif search. These datasets might comprise sequences with strong target binding, termed positive sequences. Alternatively, datasets might have both positive and negative sequences, with the latter assumed to lack strong target binding. Ms. Zhao employed the Apriori algorithm,[124] the motif search identifies the top K motifs most prevalent in the positive sequence set. The result is a csv file listing motifs and their frequency in the positive sequence set. If both positive and negative sequence sets are analyzed, the search yields top K motifs frequent in the positive set but rare in the negative set. The output ranks motifs based on the difference in their frequency between the positive and negative sets. As our tool is designed for genetically encoded affinity-selected molecules of uniform size but varying sequences, it assumes consistent sequence lengths across the dataset.

To execute the motif search task, users run *motif_search.py* from the command line, a parallelized Python 3 script. Table 1 lists and describes the required and optional flags for this command. The essential flag is (-i), specifying the input dataset of positive sequences. When incorporating a negative sequence dataset using the optional flag (-n), the output also includes the Chi-square and p-value statistics. In practical scenarios, the control dataset might contain sequences with low or non-specific target affinity. Alternatively, the random_sequence function in module.py can generate a negative dataset of random sequences absent in the positive dataset. Several other options for the motif search task are detailed in Table 5-1. The flag (-c) lets users specify the molecule type in the dataset, either DNA or peptide. The flag (-f) sets the minimum motif occurrence frequency in the positive dataset, influencing search speed. Users can define minimum (-m) and maximum (-l) motif lengths, allowable motif gaps (-g), and the maximum gap length (-a). Additionally, the motif search can run on multiple processors, with the number of processors specified using the (-p) flag. While our search yields a detailed list of individual motifs, broader motif families can be identified using other software, like GLAM2 in the MEME suite,[120] which visualizes motifs as sequence logos (detailed in Ref[3]).

Table 5-1. Overview of choices available within the *motif_search.py* script.

flag	description	default	format
-i	calls the input file containing positive sequences (csv)	-	file name, required
-n	calls the input file containing negative sequences (csv)	-	file name
-o	defines the name of the output files listing the found motifs to be different from default (motifs.csv)	motifs.csv	file name
-c	defines if used for protein or DNA sequences	amino_acids	dna or amino_acids
-f	the minimal occurrence frequency for the positive sequences	0.001	0<number < 1
-m	the minimal motif length	3	integer > 1
-l	the maximal motif length	the length of the longest positive sequence	integer > 1
-g	maximal number of gaps in motifs	0	integer ≥ 0
-a	maximal gap length	1	integer ≥ 1
-p	number of processors used for running motif_search.py	20	integer

PCA and t-SNE Analyses

BinderSpace's second task involves PCA and t-SNE analyses on DNA or peptide-based molecule datasets. These can be executed on entire datasets or subsets related to specific motifs. We recommend users interact with the binder_space.py code in a Jupyter notebook for flexibility. The most straightforward PCA and t-SNE analyses are in two- and three-dimensional spaces, enabling users to visually inspect the dataset in these condensed spaces. The PCA and t-SNE analyses in binder_space.py utilize scikit-learn modules, optimized based on our test DNA dataset.

However, the t-SNE section offers comprehensive guidance on hyperparameter testing. For both analyses, molecule sequences in datasets are represented as matrices, either (1 × N) for peptide-based molecules with N amino acids or (1 × M) for DNA molecules with M nucleotides (Table S1). Users can choose the dataset for analysis, either the entire dataset or subsets containing specific motifs. The code provides instructions for each analysis type and suggests hyperparameter optimization. The code can also export PCA and t-SNE coordinates for sequences across all dataset dimensions when using the standard scikit-learn library. These coordinates are saved as a csv output. For extensive datasets, like our test DNA dataset with over 750,000 sequences, t-SNE analysis might be slow. In such cases, we recommend the cuML library for t-SNE analysis, which currently supports only two-dimensional analysis and is compatible with Linux-like systems.[124]

Clustering in PCA and t-SNE Spaces

Our provided Jupyter notebooks allow users to cluster data on the two-dimensional maps obtained from PCA and t-SNE analyses. Depending on data distribution in the maps, users can opt for one of four clustering methods from the scikit-learn library: DBSCAN, Birch, Gaussian Mixture, and k-means. Each method offers various parameters for refining clustering results.

Sample Datasets

We utilized two datasets, summarized in Table 5-2, to showcase BinderSpace's capabilities. The SDNA dataset, sourced from Jeong, Landry, and colleagues and detailed in Ref.,[87] consists of 18-nt long DNA sequences, divided into positive (SDNA-pos) and negative (SDNA-neg) sequences. The SDNA-pos sequences were selected for binding to single-walled carbon nanotubes (SWNTs) in the presence of serotonin, while SDNA-neg sequences were chosen for binding only to SWNTs. Another dataset, BCA-pos and BCA-neg, initially sourced from experiments by Ekanayake, Derda, and colleagues and described in Ref.[104] contains cyclic peptides with 1,3-

diketone groups. These molecules were selected for binding to bovine carbonic anhydrase (BCA) and compared against a control protein, bovine serum albumin (BSA).

Table 5-2. Overview of datasets utilized to showcase the application of BinderSpace. "X" denotes the variable locations in the molecules.

Molecule type	Dataset	Length	Sequence template	Number of sequences
DNA	SDNA-pos	18-nt	C_6-X_{18}-C_6	570,926
	SDNA-neg	18-nt	C_6-X_{18}-C_6	219,382
peptidomimetic	BCA-pos	6-aa	SXCXXXC-DKMP	7,815
	BCA-neg	6-aa	SXCXXXC-DKMP	7,644

5.3. Results and Examples

Motif Analysis

Upon importing the positive and negative datasets, BinderSpace sifts through the sequences to identify recurring motifs. It then provides an output detailing these sequence motifs and their prevalence within the datasets. These motifs can be continuous or may contain gaps. The results are presented in csv files, each showcasing sequence motifs of a specific length. Using the SDNA and BCA datasets, we first showcased the motif search feature. Figure 5-2a, b displays the top 5 motifs of two distinct lengths, both with and without sequence gaps. The program motif_search.py produces a csv file for each motif length. All motifs exceeding the user-specified minimum length are identified in a single search. Alongside the detailed list of motifs in csv format, our tool can also visually display the top motifs and their corresponding percentages, color-coded based on their magnitude (Figure 5-2a, b). This allows for an easy visual assessment of motif representation in the dataset. Figure 5-2a further demonstrates our tool's ability to visually represent functional properties of sequences containing each motif. For the SDNA dataset, this functional property is the change in optical fluorescence emission, termed ΔF/F, following serotonin analyte addition to a sample containing DNA sequences wrapped around single-walled carbon nanotubes.

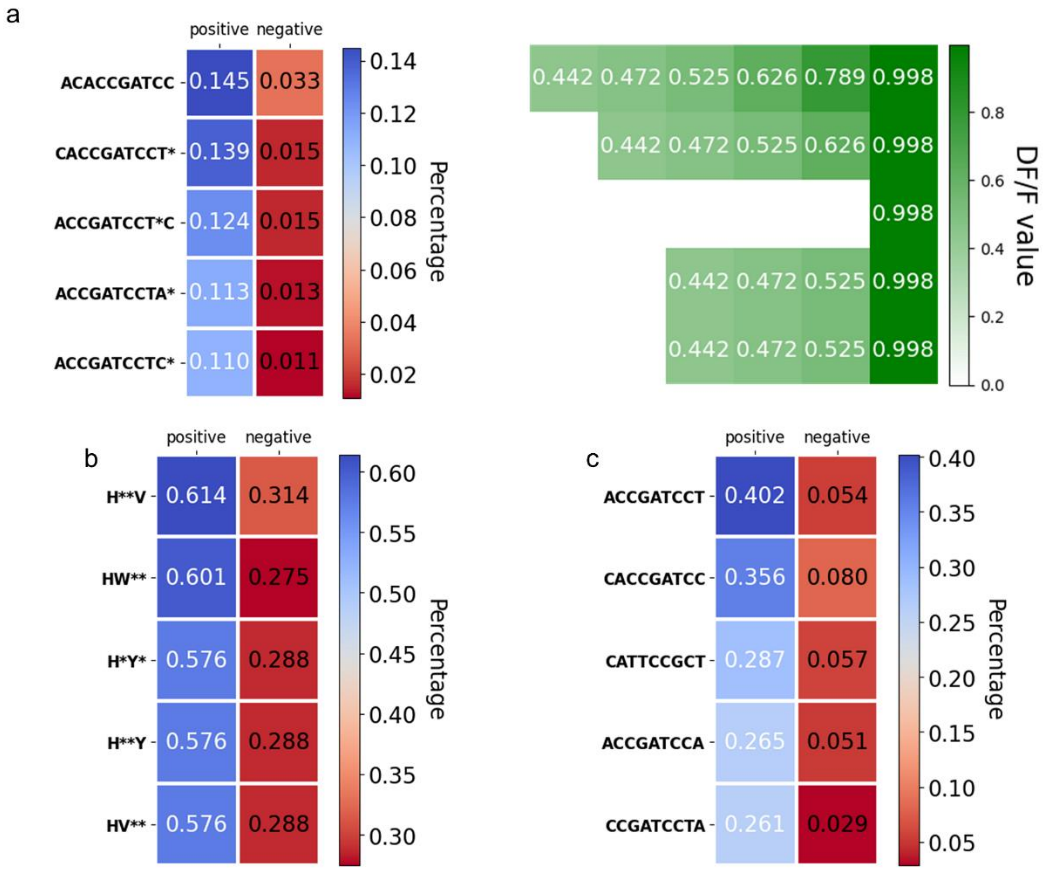


Figure 5-2. Illustrations of the graphical results of identified motifs through BinderSpace from two sample datasets. The motif analysis yields a csv file (comprehensive list) and a heatmap (prominent motif list) where motifs are arranged based on their frequency in the positive group. a) Heatmap showing leading 10-nt long motifs with either a single gap or none from the SDNA dataset, shaded according to ratios in positive and negative groups (left). Another heatmap represents ΔF/F values for DNA sequences encompassing these motifs, sourced from specific experiments24 (right). b) Heatmap spotlighting prominent 4-amino acid motifs with a pair of gaps from the BCA collection. c) Heatmap showcasing leading 9-nt long motifs devoid of gaps from the SDNA collection.

This visualization can be instrumental in pinpointing motifs linked to high functional property values. For the BCA dataset, our motif search can identify motifs with gaps as well as two and three amino acid motifs, as depicted in Figure 2b. We evaluated the runtime of BinderSpace's motif_search.py on a desktop workstation with an Intel Corporation Xeon E7 v3/Core i7 processor and 32 GB of memory, using the SDNA dataset, which contains 790,308 sequences. Execution time depends on the frequency option (-f) and the number of gaps (-g). Increasing the frequency and reducing gaps speeds up the process, as shown in Table S2. For instance, with the SDNA dataset, a default frequency of 0.001 and 1 gap resulted in a runtime of 1128 s (18.8 min). Adjusting the frequency to 0.01% reduced this to 85 s (1.42 min). Conversely, a laptop with an 11th Generation Intel Core i7-11800H Processor and 64 GB of CPU-allocated memory had a runtime of 43.19 s. For the BCA dataset, which contains 15,459 amino acid sequences, each four amino acids long, the runtime was consistently under 3 s (detailed in Ref[3]).

Visualizing Affinity-Selected Molecule Sequence Space

To visualize the sequence space of affinity-selected molecules in the SDNA dataset, we conducted PCA and t-SNE analyses on 18-nt DNAs, initially represented as 1 × 18 encoded arrays (Table S1). Figure 3A,B displays the positions of the top 500 SDNA-pos and SDNA-neg dataset sequences, overlaid with sequences containing the C*CATTCCGCT motif, previously identified as functionally significant.[87] The specific sequence and motif were chosen due to their notable 169% increase in optical emission in the presence of the serotonin analyte.[87] The PCA and t-SNE analyses were based on all 790,308 sequences from the SDNA dataset, but for clarity, only select sequences are visualized. Figure 5-3a, b reveals that four DNA sequence subsets occupy similar regions in both PCA and t-SNE spaces. The high overlap makes it challenging to discern regions associated with sequences binding to serotonin from those that don't. Figure 5-3c, d displays the positions of SDNA-pos and SDNA-neg sequences with the C*CATTCCGCT motif in PCA and t-SNE spaces, based on the subset of SDNA sequences with this motif. The plots reveal distinct sequence clusters, especially in t-SNE space, with varying compositions of SDNA-pos and SDNA-neg sequences with the C*CATTCCGCT motif. One potential application of datasets like SDNA is to train machine learning models to predict new high-affinity binding sequences.

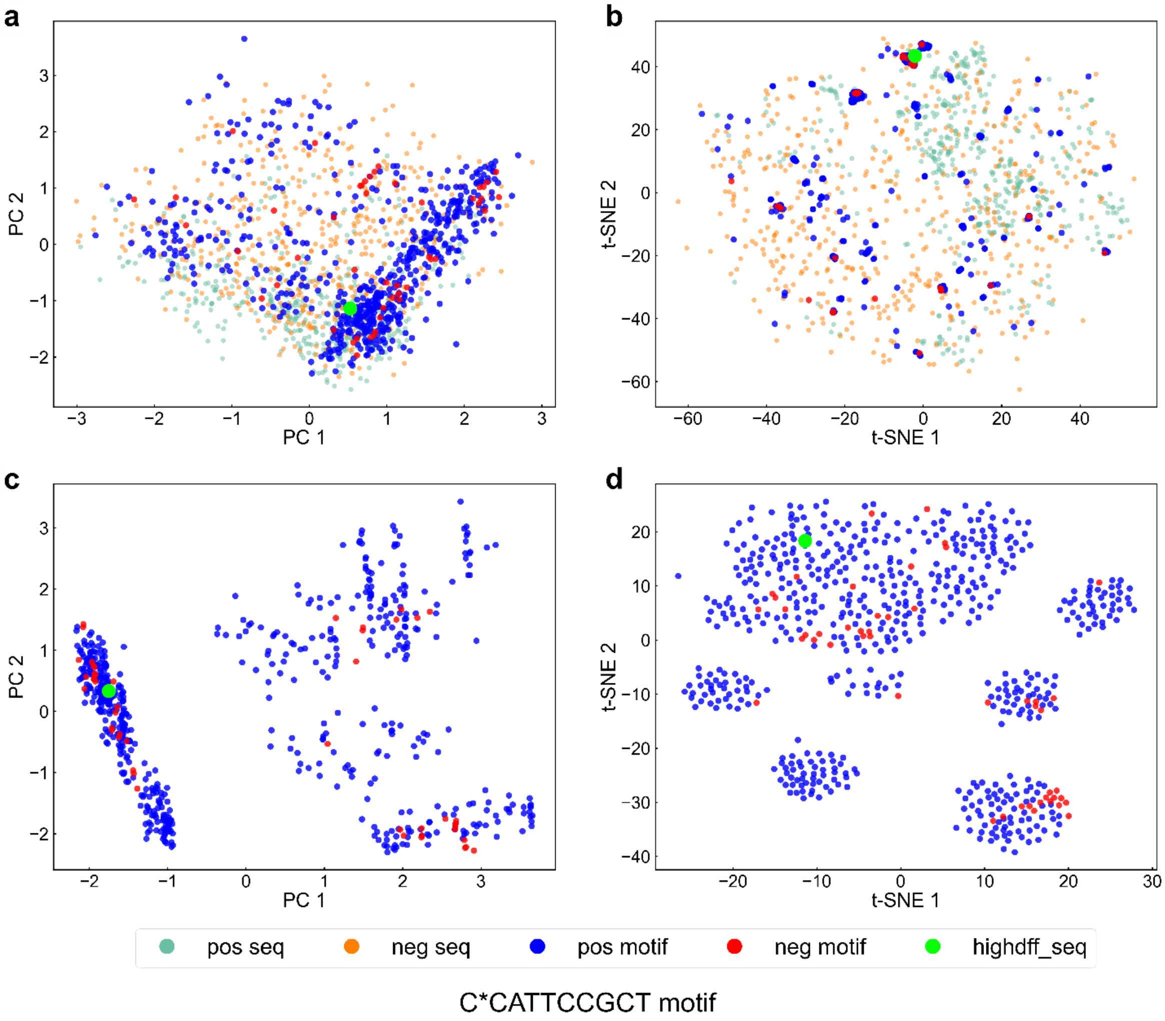


Figure 5-3. Exploration of the sequence landscape represented by the SDNA dataset. The visualizations were crafted by employing either the PCA or t-SNE techniques for shrinking the dimensionality of the original 18-nt long DNA sequences, initially showcased as 1 x 18 matrices. a) Contrast of the leading 500 sequences from SDNA-pos (depicted in orange) and SDNA-neg (in green) datasets, juxtaposed with sequences embracing the C*CATTCCGCT motif from both SDNA-pos (in blue) and SDNA-neg (in red) sets. This motif emerges in the ACGCCAACACATTCCGCT sequence (highlighted in fluorescent green), a sequence of noted functional relevance24. The principal component (PC) arena is drawn from the all-encompassing SDNA dataset. b) Contrast of sequences identical to panel a, yet analyzed using t-SNE. The t-SNE domain is derived from the entire SDNA dataset. c) A juxtaposition of sequences holding the C*CATTCCGCT motif from SDNA-pos (in blue) and SDNA-neg (in red) datasets. The PC realm is crafted from the SDNA sequences inclusive of the C*CATTCCGCT motif. d) Review of sequences identical to panel c, but through t-SNE. The t-SNE domain is shaped from SDNA sequences encompassing the C*CATTCCGCT motif.

The analyses in Figure 5-3 highlight the separation between SDNA-pos and SDNA-neg sequences, which can inform the potential accuracy of ML models. For instance, the better separation in Figure 5-3 c, d suggests that the dataset used here might yield higher-quality ML models than the datasets in Figure 5-3a, b, where sequences overlap more. However, since ML models would be based on specific data subsets, they might not generalize to sequences lacking the motif. Depending on project goals, users should examine maps from both full datasets and subsets to determine the viability of predictive ML models.

Cluster Analysis and Sequence Extraction

Our findings in Figure 5-3 c, d indicate that points in reduced dimensionality sequence space sometimes form clusters. These clusters, discernible visually, have diverse compositions of sequences from positive and negative dataset sections. Some clusters predominantly contain sequences from positive datasets, suggesting high target affinity. We theorize that such clusters might contain functionally significant sequences, warranting further experimental investigation. To identify sequences within clusters and extract them, I incorporated clustering analysis functionality into our Jupyter notebook codes.

Figure 5-4 presents example clustering analyses using the t-SNE map of sequences with the C*CATTCCGCT motif from SDNA-pos and SDNA-neg datasets. The same map displays cluster labels obtained by four different methods: DBSCAN, Birch, Gaussian Mixture, and k-means. All four methods required hyperparameter tuning to enhance cluster labeling. Among the tested methods, DBSCAN (Figure 5-4a) yielded the best cluster labeling. Birch and Gaussian Mixture methods excelled in small cluster labeling but struggled with larger clusters. The k-means method's labeling was also subpar compared to DBSCAN, as one distinct large cluster was labeled as multiple clusters. After clustering, the code offers an option to export sequences forming a specific cluster in csv format.

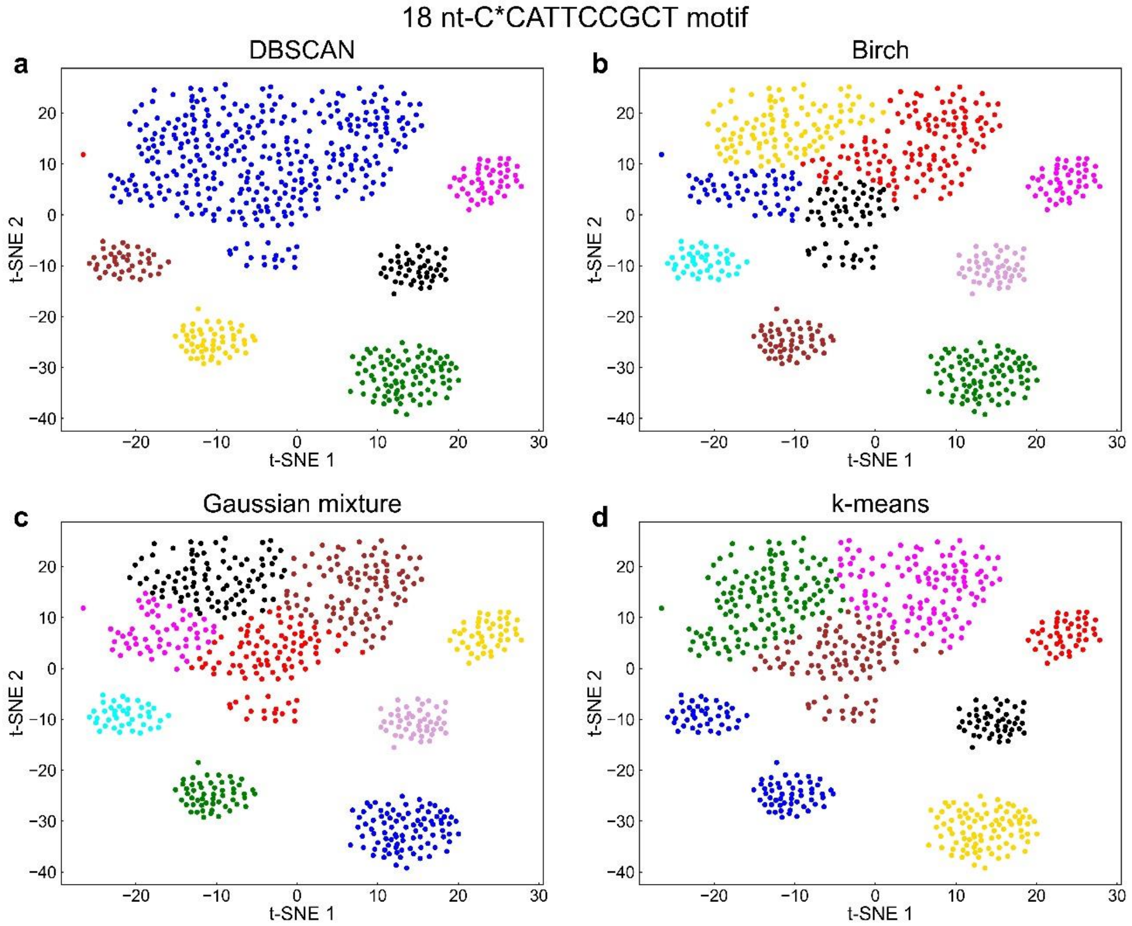


Figure 5-4. Cluster examination of the t-SNE visualization for sequences featuring the C*CATTCCGCT motif from the SDNA collection. This study is based on the two-dimensional t-SNE depiction presented in Figure 3c. a) Cluster evaluation employing the DBSCAN technique, set with ε = 6 and a threshold of 20 points to define a cluster. b) Cluster study utilizing the Gaussian Mixture approach, configured for 9 clusters, spherical covariance mode, and capped at 5000 cycles. c) Cluster evaluation through the BIRCH approach, calibrated for 9 clusters. d) Cluster analysis leveraging the k-means strategy, adjusted for 7 clusters.

5.4. Conclusion

BinderSpace serves as a comprehensive tool for examining datasets derived from DNA and peptide-based molecule sequences, which are acquired from selection experiments like SELEX13, 14 and phage display18, to name a few. This Python3-based open-source tool offers capabilities such as motif analysis, visualization of sequence space, cluster analysis, and extraction of sequences from specific clusters. A standout feature of BinderSpace is its motif search functionality, which leverages the Apriori algorithm and is adept at handling even short sequence datasets. Beyond text-based results, the motif analysis offers visual outputs and can be paired with visual representations of specific functional properties for molecules containing the identified motifs, given that these properties have been previously determined. Users have the flexibility to conduct PCA and t-SNE analyses on either the entire dataset or specific motif-related subsets. These analyses can spotlight sequences of functional significance within the PCA and t-SNE visualizations. If sequences form discernible clusters within the PCA or t-SNE two-dimensional maps, users can further looks into cluster analyses. In essence, BinderSpace is a valuable resource for exploring the interplay between molecular sequences and the underlying principles governing their target binding interactions.

Chapter 6: Conclusion and Future works

In my dissertation, I included the research that was performed for three publications where I was a first or a co-first author.

In the first study, I used molecular dynamics simulations to characterize POPC-coated (6,5) SWNT conjugates. Unexpectedly, the SWNTs showed asymmetric positioning within POPC coronas, contrasting with previous findings in different lipid types. This suggests that SWNT positioning in lipid assemblies may be influenced by lipid type and SWNT size. Furthermore, the interactions of cell membrane disruptors with these conjugates were explored, revealing varying interaction mechanisms. Colistin and TAT peptides are deeply inserted into POPC bilayers and coronas, whereas crotamine-derived peptides are primarily adsorbed to the POPC surface. A potential future direction could be to characterize the interactions of cell-membrane-disrupting peptides with lipopolysaccharide-coated single-walled carbon nanotubes, extending the understanding of these materials' interactions with complex lipid structures.

In the second study, I applied machine learning methods to discover DNA–SWNT sensors for serotonin with high near-infrared fluorescence response. Machine learning models, particularly convolutional neural network classifier models, effectively predicted promising sensor DNA sequences, significantly improving traditional selection methods based on sequence abundance. My study demonstrated the potential of machine learning in accelerating sensor discovery and optimization. Future work could focus on expanding the capability of the sensors to detect other neurotransmitters while ensuring specificity, such as developing sensors that specifically detect serotonin without interference from other neurotransmitters like dopamine.

In the third study, the BinderSpace package was developed to analyze datasets of DNA and peptide-based molecules obtained from selection experiments. Implemented in Python3, it performs motif analysis, sequence space visualization, cluster analysis, and sequence extraction from clusters of interest. Its ability to work with very short sequences and provide insights into the relationships between molecule sequences and their binding targets highlights its utility in molecular recognition research. Future development could involve enhancing BinderSpace to operate on GPUs for increased computational efficiency and scalability and adapting it to accept datasets with varying sequence lengths, facilitating more comprehensive analyses in molecular biology and bioinformatics.

References

1. Yadav, A., Kelich, P., Kallmyer, N., Reuel, N. F., & Vuković, L. (2023). Characterizing the Interactions of Cell-Membrane-Disrupting Peptides with Lipid-Functionalized Single-Walled Carbon Nanotubes. *ACS Applied Materials & Interfaces*, *15*(20), 24084–24096. https://doi.org/10.1021/acsami.3c01217

2. Kelich, P., Jeong, S., Navarro, N., Adams, J., Sun, X., Zhao, H., Landry, M. P., & Vuković, L. (2022). Discovery of DNA–Carbon Nanotube Sensors for Serotonin with Machine Learning and Near-infrared Fluorescence Spectroscopy. *ACS Nano*, *16*(1), 736–745. https://doi.org/10.1021/acsnano.1c08271

3. Kelich, P., Zhao, H., Orona, J. R., & Vuković, L. (2023). <scp> *BinderSpace* </scp> : A package for sequence space analyses for datasets of <scp>affinity-selected</scp> oligonucleotides and <scp>peptide-based</scp> molecules. *Journal of Computational Chemistry*, *44*(22), 1836–1844. https://doi.org/10.1002/jcc.27130

4. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., & Karplus, M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, *4*(2), 187–217. https://doi.org/10.1002/jcc.540040211

5. Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kalé, L., & Schulten, K. (2005). Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, *26*(16), 1781–1802. https://doi.org/10.1002/jcc.20289

6. Huang, J., & MacKerell, A. D. (2013). CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *Journal of Computational Chemistry*, *34*(25), 2135–2145. https://doi.org/10.1002/jcc.23354

7. Feller, S. E., & MacKerell, A. D. (2000). An Improved Empirical Potential Energy Function for Molecular Simulations of Phospholipids. *The Journal of Physical Chemistry B*, *104*(31), 7510–7515. https://doi.org/10.1021/jp0007843

8. Denning, E. J., Priyakumar, U. D., Nilsson, L., & Mackerell, A. D. (2011). Impact of 2′-hydroxyl sampling on the conformational properties of RNA: Update of the CHARMM all-atom additive force field for RNA. *Journal of Computational Chemistry*, *32*(9), 1929–1943. https://doi.org/10.1002/jcc.21777

9. Darden, T., York, D., & Pedersen, L. (1993). Particle mesh Ewald: An N ·log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics*, *98*(12), 10089–10092. https://doi.org/10.1063/1.464397

10. Jiang, D., Wu, Z., Hsieh, C.-Y., Chen, G., Liao, B., Wang, Z., Shen, C., Cao, D., Wu, J., & Hou, T. (2021). Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *Journal of Cheminformatics*, *13*(1), 12. https://doi.org/10.1186/s13321-020-00479-8

11. Wu, Y., Sutton, G. D., Halamicek, M. D. S., Xing, X., Bao, J., & Teets, T. S. (2022). Cyclometalated iridium-coumarin ratiometric oxygen sensors: improved signal resolution and tunable dynamic ranges. *Chemical Science*, *13*(30), 8804–8812. https://doi.org/10.1039/D2SC02909J

12. Valueva, M. V., Nagornov, N. N., Lyakhov, P. A., Valuev, G. V., & Chervyakov, N. I. (2020). Application of the residue number system to reduce hardware costs of the convolutional neural network implementation. *Mathematics and Computers in Simulation*, *177*, 232–243. https://doi.org/10.1016/j.matcom.2020.04.031

13. Gunasekaran, H., Ramalakshmi, K., Rex Macedo Arokiaraj, A., Deepa Kanmani, S.,

Venkatesan, C., & Suresh Gnana Dhas, C. (2021). Analysis of DNA Sequence Classification Using CNN and Hybrid Models. *Computational and Mathematical Methods in Medicine*, *2021*, 1–12. https://doi.org/10.1155/2021/1835056

14. Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology*, *24*(12), 1565–1567. https://doi.org/10.1038/nbt1206-1565

15. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

16. Mishra, A. (2021). *Metrics to Evaluate your Machine Learning Algorithm.* https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

17. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

18. Bro, R., & Smilde, A. K. (2014). Principal component analysis. *Anal. Methods*, *6*(9), 2812–2831. https://doi.org/10.1039/C3AY41907J

19. Lipkin, R., & Lazaridis, T. (2017). Computational studies of peptide-induced membrane pore formation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1726), 20160219. https://doi.org/10.1098/rstb.2016.0219

20. Shai, Y. (1999). Mechanism of the binding, insertion and destabilization of phospholipid bilayer membranes by α-helical antimicrobial and cell non-selective membrane-lytic peptides. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, *1462*(1–2), 55–70. https://doi.org/10.1016/S0005-2736(99)00200-X

21. Benfield, A. H., & Henriques, S. T. (2020). Mode-of-Action of Antimicrobial Peptides: Membrane Disruption vs. Intracellular Mechanisms. *Frontiers in Medical Technology*, *2*.

https://doi.org/10.3389/fmedt.2020.610997

22. Giuliani, A., & Rinaldi, A. C. (2011). Beyond natural antimicrobial peptides: multimeric peptides and other peptidomimetic approaches. *Cellular and Molecular Life Sciences, 68*(13), 2255–2266. https://doi.org/10.1007/s00018-011-0717-3

23. Tucker, A. T., Leonard, S. P., DuBois, C. D., Knauf, G. A., Cunningham, A. L., Wilke, C. O., Trent, M. S., & Davies, B. W. (2018). Discovery of Next-Generation Antimicrobials through Bacterial Self-Screening of Surface-Displayed Peptide Libraries. *Cell*, *172*(3), 618-628.e13. https://doi.org/10.1016/j.cell.2017.12.009

24. Mijalis, A. J., Thomas, D. A., Simon, M. D., Adamo, A., Beaumont, R., Jensen, K. F., & Pentelute, B. L. (2017). A fully automated flow-based approach for accelerated peptide synthesis. *Nature Chemical Biology*, *13*(5), 464–466. https://doi.org/10.1038/nchembio.2318

25. Zheng, H., Wang, W., Li, X., Wang, Z., Hood, L., Lausted, C., & Hu, Z. (2013). An automated Teflon microfluidic peptide synthesizer. *Lab on a Chip*, *13*(17), 3347. https://doi.org/10.1039/c3lc50632k

26. Dopp, J. L., Rothstein, S. M., Mansell, T. J., & Reuel, N. F. (2019). Rapid prototyping of proteins: Mail order gene fragments to assayable proteins within 24 hours. *Biotechnology and Bioengineering*, *116*(3), 667–676. https://doi.org/10.1002/bit.26912

27. Rudilla, H., Merlos, A., Sans-Serramitjana, E., Fuste, E., M. Sierra, J., Zalacain, A., Vinuesa, T., & Vinas, M. (2018). New and old tools to evaluate new antimicrobial peptides. *AIMS Microbiology*, *4*(3), 522–540. https://doi.org/10.3934/microbiol.2018.3.522

28. Guo, M. T., Rotem, A., Heyman, J. A., & Weitz, D. A. (2012). Droplet microfluidics for high-throughput biological assays. *Lab on a Chip*, *12*(12), 2146.

https://doi.org/10.1039/c2lc21147e

29. Puentes, P. R., Henao, M. C., Torres, C. E., Gómez, S. C., Gómez, L. A., Burgos, J. C., Arbeláez, P., Osma, J. F., Muñoz-Camargo, C., Reyes, L. H., & Cruz, J. C. (2020). Design, Screening, and Testing of Non-Rational Peptide Libraries with Antimicrobial Activity: In Silico and Experimental Approaches. *Antibiotics*, *9*(12), 854. https://doi.org/10.3390/antibiotics9120854

30. Chen, X., & Shen, J. (2017). Review of membranes in microfluidics. *Journal of Chemical Technology & Biotechnology*, *92*(2), 271–282. https://doi.org/10.1002/jctb.5105

31. Schaich, M., Cama, J., Al Nahas, K., Sobota, D., Sleath, H., Jahnke, K., Deshpande, S., Dekker, C., & Keyser, U. F. (2019). An Integrated Microfluidic Platform for Quantifying Drug Permeation across Biomimetic Vesicle Membranes. *Molecular Pharmaceutics*, *16*(6), 2494–2501. https://doi.org/10.1021/acs.molpharmaceut.9b00086

32. N Kallmyer, D Eeg, R Khor, N Roby, N. R. (2021). *Detection of Cell Membrane Interactions with Lipid-functionalized Single-walled Carbon Nanotubes*. https://doi.org/10.21203/rs.3.rs-588765/v1

33. Zheng, M., Jagota, A., Semke, E. D., Diner, B. A., Mclean, R. S., Lustig, S. R., Richardson, R. E., & Tassi, N. G. (2003). DNA-assisted dispersion and separation of carbon nanotubes. *Nature Materials*, *2*(5), 338–342. https://doi.org/10.1038/nmat877

34. Wu, Y., Phillips, J. A., Liu, H., Yang, R., & Tan, W. (2008). Carbon Nanotubes Protect DNA Strands during Cellular Delivery. *ACS Nano*, *2*(10), 2023–2028. https://doi.org/10.1021/nn800325a

35. Zhou, F., Wu, S., Wu, B., Chen, W. R., & Xing, D. (2011). Mitochondria-Targeting Single-Walled Carbon Nanotubes for Cancer Photothermal Therapy. *Small*, *7*(19), 2727–2735.

https://doi.org/10.1002/smll.201100669

36. Moon, H. K., Lee, S. H., & Choi, H. C. (2009). In Vivo Near-Infrared Mediated Tumor Destruction by Photothermal Effect of Carbon Nanotubes. *ACS Nano*, *3*(11), 3707–3713. https://doi.org/10.1021/nn900904h

37. Beyene, A. G., Delevich, K., Del Bonis-O'Donnell, J. T., Piekarski, D. J., Lin, W. C., Thomas, A. W., Yang, S. J., Kosillo, P., Yang, D., Prounis, G. S., Wilbrecht, L., & Landry, M. P. (2019). Imaging striatal dopamine release using a nongenetically encoded near infrared fluorescent catecholamine nanosensor. *Science Advances*, *5*(7). https://doi.org/10.1126/sciadv.aaw3108

38. Kruss, S., Salem, D. P., Vuković, L., Lima, B., Vander Ende, E., Boyden, E. S., & Strano, M. S. (2017). High-resolution imaging of cellular dopamine efflux using a fluorescent nanosensor array. *Proceedings of the National Academy of Sciences*, *114*(8), 1789–1794. https://doi.org/10.1073/pnas.1613541114

39. Kruss, S., Hilmer, A. J., Zhang, J., Reuel, N. F., Mu, B., & Strano, M. S. (2013). Carbon nanotubes as optical biomedical sensors. *Advanced Drug Delivery Reviews*, *65*(15), 1933–1950. https://doi.org/10.1016/j.addr.2013.07.015

40. Beyene, A. G., Alizadehmojarad, A. A., Dorlhiac, G., Goh, N., Streets, A. M., Král, P., Vuković, L., & Landry, M. P. (2018). Ultralarge Modulation of Fluorescence by Neuromodulators in Carbon Nanotubes Functionalized with Self-Assembled Oligonucleotide Rings. *Nano Letters*, *18*(11), 6995–7003. https://doi.org/10.1021/acs.nanolett.8b02937

41. Qiao, R., & Ke, P. C. (2006). Lipid-Carbon Nanotube Self-Assembly in Aqueous Solution. *Journal of the American Chemical Society*, *128*(42), 13656–13657.

https://doi.org/10.1021/ja063977y

42. Bisker, G., Dong, J., Park, H. D., Iverson, N. M., Ahn, J., Nelson, J. T., Landry, M. P., Kruss, S., & Strano, M. S. (2016). Protein-targeted corona phase molecular recognition. *Nature Communications*, *7*(1), 10241. https://doi.org/10.1038/ncomms10241

43. Antonucci, A., Kupis-Rozmysłowicz, J., & Boghossian, A. A. (2017). Noncovalent Protein and Peptide Functionalization of Single-Walled Carbon Nanotubes for Biodelivery and Optical Sensing Applications. *ACS Applied Materials & Interfaces*, *9*(13), 11321–11331. https://doi.org/10.1021/acsami.7b00810

44. Safaee, M. M., Gravely, M., Rocchio, C., Simmeth, M., & Roxbury, D. (2019). DNA Sequence Mediates Apparent Length Distribution in Single-Walled Carbon Nanotubes. *ACS Applied Materials & Interfaces*, *11*(2), 2225–2233. https://doi.org/10.1021/acsami.8b16478

45. Alizadehmojarad, A. A., Zhou, X., Beyene, A. G., Chacon, K. E., Sung, Y., Pinals, R. L., Landry, M. P., & Vuković, L. (2020). Binding Affinity and Conformational Preferences Influence Kinetic Stability of Short Oligonucleotides on Carbon Nanotubes. *Advanced Materials Interfaces*, *7*(15), 2000353. https://doi.org/10.1002/admi.202000353

46. Mitchell, D. T., Lee, S. B., Trofin, L., Li, N., Nevanen, T. K., Söderlund, H., & Martin, C. R. (2002). Smart Nanotubes for Bioseparations and Biocatalysis. *Journal of the American Chemical Society*, *124*(40), 11864–11865. https://doi.org/10.1021/ja027247b

47. Lee, S. B., Mitchell, D. T., Trofin, L., Nevanen, T. K., Söderlund, H., & Martin, C. R. (2002). Antibody-Based Bio-Nanotube Membranes for Enantiomeric Drug Separations. *Science*, *296*(5576), 2198–2200. https://doi.org/10.1126/science.1071396

48. Jirage, K. B., Hulteen, J. C., & Martin, C. R. (1997). Nanotubule-Based Molecular-Filtration Membranes. *Science*, *278*(5338), 655–658. https://doi.org/10.1126/science.278.5338.655

49. Boghossian, A. A., Zhang, J., Barone, P. W., Reuel, N. F., Kim, J., Heller, D. A., Ahn, J., Hilmer, A. J., Rwei, A., Arkalgud, J. R., Zhang, C. T., & Strano, M. S. (2011). Near-Infrared Fluorescent Sensors based on Single-Walled Carbon Nanotubes for Life Sciences Applications. *ChemSusChem*, *4*(7), 848–863. https://doi.org/10.1002/cssc.201100070

50. Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, *14*(1), 33–38. https://doi.org/10.1016/0263-7855(96)00018-5

51. Klauda, J. B., Venable, R. M., Freites, J. A., O’Connor, J. W., Tobias, D. J., Mondragon-Ramirez, C., Vorobyov, I., MacKerell, A. D., & Pastor, R. W. (2010). Update of the CHARMM All-Atom Additive Force Field for Lipids: Validation on Six Lipid Types. *The Journal of Physical Chemistry B*, *114*(23), 7830–7843. https://doi.org/10.1021/jp101759q

52. Wu, E. L., Cheng, X., Jo, S., Rui, H., Song, K. C., Dávila-Contreras, E. M., Qi, Y., Lee, J., Monje-Galvan, V., Venable, R. M., Klauda, J. B., & Im, W. (2014). CHARMM-GUI Membrane Builder toward realistic biological membrane simulations. *Journal of Computational Chemistry*, *35*(27), 1997–2004. https://doi.org/10.1002/jcc.23702

53. Vanommeslaeghe, K., & MacKerell, A. D. (2012). Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing. *Journal of Chemical Information and Modeling*, *52*(12), 3144–3154. https://doi.org/10.1021/ci300363c

54. Vanommeslaeghe, K., Raman, E. P., & MacKerell, A. D. (2012). Automation of the CHARMM General Force Field (CGenFF) II: Assignment of Bonded Parameters and Partial Atomic Charges. *Journal of Chemical Information and Modeling*, *52*(12), 3155–3168. https://doi.org/10.1021/ci3003649

55. Grégoire, C., Péloponèse, J.-M., Esquieu, D., Opi, S., Campbell, G., Solomiac, M., Lebrun,

E., Lebreton, J., & Loret, E. P. (2001). Homonuclear 1 H-NMR assignment and structural characterization of human immunodeficiency virus type 1 Tat Mal protein. *Biopolymers*, *62*(6), 324–335. https://doi.org/10.1002/bip.10000

56. Coronado, M. A., Gabdulkhakov, A., Georgieva, D., Sankaran, B., Murakami, M. T., Arni, R. K., & Betzel, C. (2013). Structure of the polypeptide crotamine from the Brazilian rattlesnake Crotalus durissus terrificus. *Acta Crystallographica Section D Biological Crystallography*, *69*(10), 1958–1964. https://doi.org/10.1107/S0907444913018003

57. Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kalé, L., & Schulten, K. (2005). Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, *26*(16), 1781–1802. https://doi.org/10.1002/jcc.20289

58. Vanommeslaeghe, K., Hatcher, E., Acharya, C., Kundu, S., Zhong, S., Shim, J., Darian, E., Guvench, O., Lopes, P., Vorobyov, I., & Mackerell, A. D. (2009). CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of Computational Chemistry*, NA-NA. https://doi.org/10.1002/jcc.21367

59. Darden, T., York, D., & Pedersen, L. (1993). Particle mesh Ewald: An N ·log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics*, *98*(12), 10089–10092. https://doi.org/10.1063/1.464397

60. Duque, J. G., Gupta, G., Cognet, L., Lounis, B., Doorn, S. K., & Dattelbaum, A. M. (2011). New Route to Fluorescent Single-Walled Carbon Nanotube/Silica Nanocomposites: Balancing Fluorescence Intensity and Environmental Sensitivity. *The Journal of Physical Chemistry C*, *115*(31), 15147–15153. https://doi.org/10.1021/jp2012107

61. Aliev, A. E., Lima, M. H., Silverman, E. M., & Baughman, R. H. (2010). Thermal

conductivity of multi-walled carbon nanotube sheets: radiation losses and quenching of phonon modes. *Nanotechnology*, *21*(3), 035709. https://doi.org/10.1088/0957-4484/21/3/035709

62. Lefebvre, J., & Finnie, P. (2009). Photoluminescence and Förster Resonance Energy Transfer in Elemental Bundles of Single-Walled Carbon Nanotubes. *The Journal of Physical Chemistry C*, *113*(18), 7536–7540. https://doi.org/10.1021/jp810892z

63. He, X., Htoon, H., Doorn, S. K., Pernice, W. H. P., Pyatkov, F., Krupke, R., Jeantet, A., Chassagneux, Y., & Voisin, C. (2018). Carbon nanotubes as emerging quantum-light sources. *Nature Materials*, *17*(8), 663–670. https://doi.org/10.1038/s41563-018-0109-2

64. Saha, A., Gifford, B. J., He, X., Ao, G., Zheng, M., Kataura, H., Htoon, H., Kilina, S., Tretiak, S., & Doorn, S. K. (2018). Narrow-band single-photon emission through selective aryl functionalization of zigzag carbon nanotubes. *Nature Chemistry*, *10*(11), 1089–1095. https://doi.org/10.1038/s41557-018-0126-4

65. Harvey, J. D., Williams, R. M., Tully, K. M., Baker, H. A., Shamay, Y., & Heller, D. A. (2019). An in Vivo Nanosensor Measures Compartmental Doxorubicin Exposure. *Nano Letters*, *19*(7), 4343–4354. https://doi.org/10.1021/acs.nanolett.9b00956

66. Lin, C.-W., Bachilo, S. M., Zheng, Y., Tsedev, U., Huang, S., Weisman, R. B., & Belcher, A. M. (2019). Creating fluorescent quantum defects in carbon nanotubes using hypochlorite and light. *Nature Communications*, *10*(1), 2874. https://doi.org/10.1038/s41467-019-10917-3

67. Demirer, G. S., Zhang, H., Goh, N. S., González-Grandío, E., & Landry, M. P. (2019). Carbon nanotube–mediated DNA delivery without transgene integration in intact plants. *Nature Protocols, 14*(10), 2954–2971. https://doi.org/10.1038/s41596-019-0208-9

68. Harvey, J. D., Jena, P. V., Baker, H. A., Zerze, G. H., Williams, R. M., Galassi, T. V., Roxbury, D., Mittal, J., & Heller, D. A. (2017). A carbon nanotube reporter of microRNA hybridization events in vivo. *Nature Biomedical Engineering*, *1*(4), 0041. https://doi.org/10.1038/s41551-017-0041

69. Horoszko, C. P., Jena, P. V., Roxbury, D., Rotkin, S. V., & Heller, D. A. (2019). Optical Voltammetry of Polymer-Encapsulated Single-Walled Carbon Nanotubes. *The Journal of Physical Chemistry C*, *123*(39), 24200–24208. https://doi.org/10.1021/acs.jpcc.9b07626

70. Zhao, M., Chen, Y., Wang, K., Zhang, Z., Streit, J. K., Fagan, J. A., Tang, J., Zheng, M., Yang, C., Zhu, Z., & Sun, W. (2020). DNA-directed nanofabrication of high-performance carbon nanotube field-effect transistors. *Science*, *368*(6493), 878–881. https://doi.org/10.1126/science.aaz7435

71. O'Connell, M. J., Bachilo, S. M., Huffman, C. B., Moore, V. C., Strano, M. S., Haroz, E. H., Rialon, K. L., Boul, P. J., Noon, W. H., Kittrell, C., Ma, J., Hauge, R. H., Weisman, R. B., & Smalley, R. E. (2002). Band Gap Fluorescence from Individual Single-Walled Carbon Nanotubes. *Science*, *297*(5581), 593–596. https://doi.org/10.1126/science.1072631

72. Richard, C., Balavoine, F., Schultz, P., Ebbesen, T. W., & Mioskowski, C. (2003). Supramolecular Self-Assembly of Lipid Derivatives on Carbon Nanotubes. *Science*, *300*(5620), 775–778. https://doi.org/10.1126/science.1080848

73. Alizadehmojarad, A. A., Zhou, X., Beyene, A. G., Chacon, K. E., Sung, Y., Pinals, R. L., Landry, M. P., & Vuković, L. (2020). Binding Affinity and Conformational Preferences Influence Kinetic Stability of Short Oligonucleotides on Carbon Nanotubes. *Advanced Materials Interfaces*, *7*(15). https://doi.org/10.1002/admi.202000353

74. Bakota, E. L., Aulisa, L., Tsyboulski, D. A., Weisman, R. B., & Hartgerink, J. D. (2009).

Multidomain Peptides as Single-Walled Carbon Nanotube Surfactants in Cell Culture. *Biomacromolecules*, *10*(8), 2201–2206. https://doi.org/10.1021/bm900382a

75. Bisker, G., Dong, J., Park, H. D., Iverson, N. M., Ahn, J., Nelson, J. T., Landry, M. P., Kruss, S., & Strano, M. S. (2016). Protein-targeted corona phase molecular recognition. *Nature Communications*, *7*(1), 10241. https://doi.org/10.1038/ncomms10241

76. Chio, L., Del Bonis-O'Donnell, J. T., Kline, M. A., Kim, J. H., McFarlane, I. R., Zuckermann, R. N., & Landry, M. P. (2019). Electrostatic Assemblies of Single-Walled Carbon Nanotubes and Sequence-Tunable Peptoid Polymers Detect a Lectin Protein and Its Target Sugars. *Nano Letters*, *19*(11), 7563–7572. https://doi.org/10.1021/acs.nanolett.8b04955

77. Harvey, J. D., Baker, H. A., Ortiz, M. V., Kentsis, A., & Heller, D. A. (2019). HIV Detection via a Carbon Nanotube RNA Sensor. *ACS Sensors*, *4*(5), 1236–1244. https://doi.org/10.1021/acssensors.9b00025

78. Zhang, J., Landry, M. P., Barone, P. W., Kim, J.-H., Lin, S., Ulissi, Z. W., Lin, D., Mu, B., Boghossian, A. A., Hilmer, A. J., Rwei, A., Hinckley, A. C., Kruss, S., Shandell, M. A., Nair, N., Blake, S., Şen, F., Şen, S., Croy, R. G., … Strano, M. S. (2013). Molecular recognition using corona phase complexes made of synthetic polymers adsorbed on carbon nanotubes. *Nature Nanotechnology*, *8*(12), 959–968. https://doi.org/10.1038/nnano.2013.236

79. Zhang, H., Demirer, G. S., Zhang, H., Ye, T., Goh, N. S., Aditham, A. J., Cunningham, F. J., Fan, C., & Landry, M. P. (2019). DNA nanostructures coordinate gene silencing in mature plants. *Proceedings of the National Academy of Sciences*, *116*(15), 7543–7548. https://doi.org/10.1073/pnas.1818290116

80. Tu, X., & Zheng, M. (2008). A DNA-based approach to the carbon nanotube sorting problem. *Nano Research*, *1*(3), 185–194. https://doi.org/10.1007/s12274-008-8022-7

81. Roxbury, D., Tu, X., Zheng, M., & Jagota, A. (2011). Recognition Ability of DNA for Carbon Nanotubes Correlates with Their Binding Affinity. *Langmuir*, *27*(13), 8282–8293. https://doi.org/10.1021/la2007793

82. Yang, F., Wang, M., Zhang, D., Yang, J., Zheng, M., & Li, Y. (2020). Chirality Pure Carbon Nanotubes: Growth, Sorting, and Characterization. *Chemical Reviews*, *120*(5), 2693–2758. https://doi.org/10.1021/acs.chemrev.9b00835

83. Lyu, M., Meany, B., Yang, J., Li, Y., & Zheng, M. (2019). Toward Complete Resolution of DNA/Carbon Nanotube Hybrids by Aqueous Two-Phase Systems. *Journal of the American Chemical Society*, *141*(51), 20177–20186. https://doi.org/10.1021/jacs.9b09953

84. Dinarvand, M., Neubert, E., Meyer, D., Selvaggio, G., Mann, F. A., Erpenbeck, L., & Kruss, S. (2019). Near-Infrared Imaging of Serotonin Release from Cells with Fluorescent Nanosensors. *Nano Letters*, *19*(9), 6604–6611. https://doi.org/10.1021/acs.nanolett.9b02865

85. Williams, R. M., Harvey, J. D., Budhathoki-Uprety, J., & Heller, D. A. (2020). Glutathione-S-transferase Fusion Protein Nanosensor. *Nano Letters*, *20*(10), 7287–7295. https://doi.org/10.1021/acs.nanolett.0c02691

86. Kruss, S., Landry, M. P., Vander Ende, E., Lima, B. M. A., Reuel, N. F., Zhang, J., Nelson, J., Mu, B., Hilmer, A., & Strano, M. (2014). Neurotransmitter Detection Using Corona Phase Molecular Recognition on Fluorescent Single-Walled Carbon Nanotube Sensors. *Journal of the American Chemical Society*, *136*(2), 713–724. https://doi.org/10.1021/ja410433b

87. Jeong, S., Yang, D., Beyene, A. G., Del Bonis-O'Donnell, J. T., Gest, A. M. M., Navarro, N.,

Sun, X., & Landry, M. P. (2019). High-throughput evolution of near-infrared serotonin nanosensors. *Science Advances*, *5*(12). https://doi.org/10.1126/sciadv.aay3771

88. Vaswani, M., Linda, F. K., & Ramesh, S. (2003). Role of selective serotonin reuptake inhibitors in psychiatric disorders: a comprehensive review. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *27*(1), 85–102. https://doi.org/10.1016/S0278-5846(02)00338-X

89. Berger, M., Gray, J. A., & Roth, B. L. (2009). The Expanded Biology of Serotonin. *Annual Review of Medicine*, *60*(1), 355–366. https://doi.org/10.1146/annurev.med.60.042307.110802

90. Unger, E., Keller, J. P., Altermatt, M., Liang, R., Yao, Z., Sun, J., Matsui, A., Dong, C., Jaffe, D. A., Hartanto, S., Mizuno, G., Borden, P., Shivange, A., Sinning, S., Underhill, S., Carlin, J., Banala, S., Cameron, L. P., Olson, D. E., … Tian, L. (2019). Directed Evolution of a Selective and Sensitive Serotonin Biosensor Via Machine Learning. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3498571

91. Han, H. S., You, J.-M., Jeong, H., & Jeon, S. (2013). Synthesis of graphene oxide grafted poly(lactic acid) with palladium nanoparticles and its application to serotonin sensing. *Applied Surface Science*, *284*, 438–445. https://doi.org/10.1016/j.apsusc.2013.07.116

92. Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, *33*(8), 831–838. https://doi.org/10.1038/nbt.3300

93. Chollet, F. & others. (2015). *Keras*.

94. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G.,

Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y. & Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. *Proc. 12th USENIX Symp. Oper. Syst. Des. Implementation, OSDI 2016*.

95. Yang, Y., Zheng, M., & Jagota, A. (2019). Learning to predict single-wall carbon nanotube-recognition DNA sequences. *Npj Computational Materials*, *5*(1), 3. https://doi.org/10.1038/s41524-018-0142-3

96. Jenson, J. M., Xue, V., Stretz, L., Mandal, T., Reich, L. “Luther,” & Keating, A. E. (2018). Peptide design by optimization on a data-parameterized protein interaction landscape. *Proceedings of the National Academy of Sciences*, *115*(44). https://doi.org/10.1073/pnas.1812939115

97. Hajduk, P. J., Meadows, R. P., & Fesik, S. W. (1997). Discovering High-Affinity Ligands for Proteins. *Science*, *278*(5337), 497–499. https://doi.org/10.1126/science.278.5337.497

98. Chaturvedi, P., Han, Y., Král, P., & Vuković, L. (2020). Adaptive Evolution of Peptide Inhibitors for Mutating SARS-CoV-2. *Advanced Theory and Simulations*, *3*(12). https://doi.org/10.1002/adts.202000156

99. Békés, M., Langley, D. R., & Crews, C. M. (2022). PROTAC targeted protein degraders: the past is prologue. *Nature Reviews Drug Discovery*, *21*(3), 181–200. https://doi.org/10.1038/s41573-021-00371-6

100. Chen, J. C., Chen, J. P., Shen, M. W., Wornow, M., Bae, M., Yeh, W.-H., Hsu, A., & Liu, D. R. (2022). Generating experimentally unrelated target molecule-binding highly functionalized nucleic-acid polymers using machine learning. *Nature Communications*, *13*(1), 4541. https://doi.org/10.1038/s41467-022-31955-4

101. Franzini, R. M., Neri, D., & Scheuermann, J. (2014). DNA-Encoded Chemical Libraries:

Advancing beyond Conventional Small-Molecule Libraries. *Accounts of Chemical Research*, *47*(4), 1247–1255. https://doi.org/10.1021/ar400284t

102. Kleiner, R. E., Dumelin, C. E., & Liu, D. R. (2011). Small-molecule discovery from DNA-encoded chemical libraries. *Chemical Society Reviews*, *40*(12), 5707. https://doi.org/10.1039/c1cs15076f

103. Derda, R., & Ng, S. (2019). Genetically encoded fragment-based discovery. *Current Opinion in Chemical Biology*, *50*, 128–137. https://doi.org/10.1016/j.cbpa.2019.03.014

104. Ekanayake, A. I., Sobze, L., Kelich, P., Youk, J., Bennett, N. J., Mukherjee, R., Bhardwaj, A., Wuest, F., Vukovic, L., & Derda, R. (2021). Genetically Encoded Fragment-Based Discovery from Phage-Displayed Macrocyclic Libraries with Genetically Encoded Unnatural Pharmacophores. *Journal of the American Chemical Society*, *143*(14), 5497–5507. https://doi.org/10.1021/jacs.1c01186

105. Owens, A. E., Iannuzzelli, J. A., Gu, Y., & Fasan, R. (2020). MOrPH-PhD: An Integrated Phage Display Platform for the Discovery of Functional Genetically Encoded Peptide Macrocycles. *ACS Central Science*, *6*(3), 368–381. https://doi.org/10.1021/acscentsci.9b00927

106. Ladner, R. C. (1995). Constrained peptides as binding entities. *Trends in Biotechnology*, *13*(10), 426–430. https://doi.org/10.1016/S0167-7799(00)88997-0

107. Sohrabi, C., Foster, A., & Tavassoli, A. (2020). Methods for generating and screening libraries of genetically encoded cyclic peptides in drug discovery. *Nature Reviews Chemistry*, *4*(2), 90–101. https://doi.org/10.1038/s41570-019-0159-2

108. Pomplun, S., Gates, Z. P., Zhang, G., Quartararo, A. J., & Pentelute, B. L. (2020). Discovery of Nucleic Acid Binding Molecules from Combinatorial Biohybrid Nucleobase

Peptide Libraries. *Journal of the American Chemical Society*, *142*(46), 19642–19651. https://doi.org/10.1021/jacs.0c08964

109. Tuerk, C., & Gold, L. (1990). Systematic Evolution of Ligands by Exponential Enrichment: RNA Ligands to Bacteriophage T4 DNA Polymerase. *Science*, *249*(4968), 505–510. https://doi.org/10.1126/science.2200121

110. Ellington, A. D., & Szostak, J. W. (1990). In vitro selection of RNA molecules that bind specific ligands. *Nature*, *346*(6287), 818–822. https://doi.org/10.1038/346818a0

111. Kim, Y. S., & Gu, M. B. (2013). *Advances in Aptamer Screening and Small Molecule Aptasensors* (pp. 29–67). https://doi.org/10.1007/10_2013_225

112. Sefah, K., Shangguan, D., Xiong, X., O'Donoghue, M. B., & Tan, W. (2010). Development of DNA aptamers using Cell-SELEX. *Nature Protocols*, *5*(6), 1169–1185. https://doi.org/10.1038/nprot.2010.66

113. Scott, J. K., & Smith, G. P. (1990). Searching for Peptide Ligands with an Epitope Library. *Science*, *249*(4967), 386–390. https://doi.org/10.1126/science.1696028

114. Wilson, D. S., Keefe, A. D., & Szostak, J. W. (2001). The use of mRNA display to select high-affinity protein-binding peptides. *Proceedings of the National Academy of Sciences*, *98*(7), 3750–3755. https://doi.org/10.1073/pnas.061028198

115. Iannuzzelli, J. A., & Fasan, R. (2020). Expanded toolbox for directing the biosynthesis of macrocyclic peptides in bacterial cells. *Chemical Science*, *11*(24), 6202–6208. https://doi.org/10.1039/D0SC01699C

116. Bashir, A., Yang, Q., Wang, J., Hoyer, S., Chou, W., McLean, C., Davis, G., Gong, Q., Armstrong, Z., Jang, J., Kang, H., Pawlosky, A., Scott, A., Dahl, G. E., Berndl, M., Dimon, M., & Ferguson, B. S. (2021). Machine learning guided aptamer refinement and discovery.

Nature Communications, *12*(1), 2366. https://doi.org/10.1038/s41467-021-22555-9

117. Copp, S. M., Bogdanov, P., Debord, M., Singh, A., & Gwinn, E. (2014). Base Motif Recognition and Design of DNA Templates for Fluorescent Silver Clusters by Machine Learning. *Advanced Materials*, *26*(33), 5839–5845. https://doi.org/10.1002/adma.201401402

118. Thiel, W. H., Bair, T., Peek, A. S., Liu, X., Dassie, J., Stockdale, K. R., Behlke, M. A., Miller, F. J., & Giangrande, P. H. (2012). Rapid Identification of Cell-Specific, Internalizing RNA Aptamers with Bioinformatics Analyses of a Cell-Based Aptamer Selection. *PLoS ONE*, *7*(9), e43836. https://doi.org/10.1371/journal.pone.0043836

119. Tyagi, A., Kapoor, P., Kumar, R., Chaudhary, K., Gautam, A., & Raghava, G. P. S. (2013). In Silico Models for Designing and Discovering Novel Anticancer Peptides. *Scientific Reports*, *3*(1), 2984. https://doi.org/10.1038/srep02984

120. Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., & Noble, W. S. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research*, *37*(Web Server), W202–W208. https://doi.org/10.1093/nar/gkp335

121. Vens, C., Rosso, M.-N., & Danchin, E. G. J. (2011). Identifying discriminative classification-based motifs in biological sequences. *Bioinformatics*, *27*(9), 1231–1238. https://doi.org/10.1093/bioinformatics/btr110

122. Wei, T., Nie, J., Larson, N. B., Ye, Z., Eckel-Passow, J. E., Robertson, K. D., Kocher, J.-P. A., & Wang, L. (2021). CpGtools: a python package for DNA methylation analysis. *Bioinformatics*, *37*(11), 1598–1599. https://doi.org/10.1093/bioinformatics/btz916

123. Davie, K., Janssens, J., Koldere, D., De Waegeneer, M., Pech, U., Kreft, Ł., Aibar, S., Makhzami, S., Christiaens, V., Bravo González-Blas, C., Poovathingal, S., Hulselmans, G.,

Spanier, K. I., Moerman, T., Vanspauwen, B., Geurs, S., Voet, T., Lammertyn, J., Thienpont, B., … Aerts, S. (2018). A Single-Cell Transcriptome Atlas of the Aging Drosophila Brain. *Cell*, *174*(4), 982-998.e20. https://doi.org/10.1016/j.cell.2018.05.057

124. Agrawal, R., & Srikant, R. (n.d.). Mining sequential patterns. *Proceedings of the Eleventh International Conference on Data Engineering*, 3–14. https://doi.org/10.1109/ICDE.1995.380415

Vita

Payam Kelich (https://payamkelich.github.io) completed his undergraduate studies in Chemical Engineering in 2012 and subsequently obtained his master's degree in Polymer Engineering in 2015, both from the Isfahan University of Technology in Isfahan, Iran. In 2020, he commenced his doctoral journey in Chemistry at UTEP. As a member of Lela Vuković's research team within the Department of Chemistry and Biochemistry, he did research in computational modeling of biological systems. His expertise lies in molecular dynamics simulations and machine learning techniques for his research. To date, his scholarly contributions at UTEP include six published papers: Genetically encoded discovery of perfluoroaryl macrocycles that bind to albumin and exhibit extended circulation in vivo (https://doi.org/10.1038/s41467-023-41427-y). BinderSpace: A package for sequence space analyses for datasets of affinity-selected oligonucleotides and peptide-based molecules (https://doi.org/10.1002/jcc.27130). Characterizing the Interactions of Cell-Membrane-Disrupting Peptides with Lipid-Functionalized Single-Walled Carbon Nanotubes (https://doi.org/10.1021/acsami.3c01217). Discovery of DNA–Carbon Nanotube Sensors for Serotonin with Machine Learning and Near-infrared Fluorescence Spectroscopy (https://doi.org/10.1021/acsnano.1c08271). Computational Modeling of the Virucidal Inhibition Mechanism for Broad-Spectrum Antiviral Nanoparticles and HPV16 Capsid Segments (https://doi.org/10.1021/acs.jpcb.1c07436). Genetically Encoded Fragment-Based Discovery from Phage-Displayed Macrocyclic Libraries with Genetically Encoded Unnatural Pharmacophores (https://doi.org/10.1021/jacs.1c01186).