# DATA-DRIVEN ENGINEERING OF CRISPR-CAS12A FOR PAM RECOGNITION

by

Apoorv Saraogee

A thesis submitted in partial fulfillment
of the requirements for the degree of

Masters of Science

(Chemical and Biological Engineering)

at the

UNIVERSITY OF WISCONSIN-MADISON
2020

# DATA-DRIVEN ENGINEERING OF CRISPR-CAS12A FOR PAM RECOGNITION

Approved by:


Dr. Philip Romero, Advisor
College of Agricultural and Life Sciences
*University of Wisconsin-Madison*

Dr. Sean Palecek
College of Engineering
*University of Wisconsin-Madison*

Dr. Eric Shusta
College of Engineering
*University of Wisconsin-Madison*

Dr. John Yin
College of Engineering
*University of Wisconsin-Madison*


Date Approved: September 1, 2020

# ACKNOWLEDGEMENTS

I have loved my time as a graduate student at UW-Madison. It has truly been a rewarding and enriching experience. I would like to especially thank Professor Philip Romero for this exciting project and creating a stimulating lab environment and people of diverse training backgrounds. I would like to thank the members of the Romero lab for their constant support in research - Jonathan Greenhalgh, Dr. Job Grant, Ben Bremer, Hridindu Rowchowdhury, Dr. Mark Politz, Dr. Aaron Lin and Juan Diaz for their constant support. I would also like to thank Haiyang (Ocean) Zheng for all of his help in analyzing data using the CHTC. Mark Politz served as an excellent mentor the first couple of years, showing me the way of molecular biology.

I am grateful to my friends and colleagues in the Chemical and Biological Engineering department – Mike Jindra, Alec Linot, Jonathan Sheavly, Francesca Gambacorta, Koji Foreman, Leida Vazquez and Pratyush Kumar. I also thank my committee members Eric Shusta, John Yin and Sean Palecek for their flexibility in scheduling of my thesis exam.

Finally, I would like to thank my family for their support and perspective.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# ABSTRACT

We know that amino acids are combined in sequence to constitute proteins for an undefined number of biological functions. Proteins thus evolved for millions of years before being repurposed for human applications in the medical field, food and chemicals. CRISPR enzymes are emerging as a highly versatile workhorse for targeting of specific DNA sequences, useful in biomedicine and biotechnology. Exploring the vast space of possible protein sequences is intractable using traditional protein engineering approaches of rational design and directed evolution. Data-driven methods can greatly accelerate protein engineering strategies and aid in CRISPR enzyme engineering. Data-driven methods also leverage the vast and exponentially growing volume of biological data. Here we design an experimental and computational pipeline to investigate the binding function of CRISPR-Cas12a. CRISPR-Cas12a works as a pair of molecular scissors that are programmed using an RNA molecule to a site with matching genetic material in DNA. An important limitation for human applications is that before they bind to their target DNA site, they must also bind to a protospacer adjacent motif (PAM). We design a library of mutant CRISPR-Cas12a proteins with chimeric sequences made by DNA recombination. To investigate PAM binding function, we develop an assay based on a Green Fluorescent Protein (GFP) reporter system presented by collaborators in the Beisel lab. We generate data on the order of millions of sequences by using long-read DNA sequencing or nanopore sequencing after we performed fluorescence activated cell sorting (FACS) using our assay on our chimeric library. Our assay is reproducible, shown by enrichment analysis on chimeric sequences, which yielded a consensus protein sequence between three sorting replicates. We further

demonstrate machine learning methods to investigate a generalized model for CRISPR-Cas12a-PAM binding.

# CHAPTER 1. DATA-DRIVEN PROTEIN ENGINEERING

Jonathan Greenhalgh[1*], Apoorv Saraogee[1*], and Philip A. Romero[1,2]

1. Department of Chemical and Biological Engineering, University of Wisconsin--Madison

2. Department of Biochemistry, University of Wisconsin--Madison

*these authors contributed equally to this work

*A version of this chapter is currently in press.*

## 1.1 Introduction

A protein's sequence of amino acids encodes its function. This "function" could refer to a protein's natural biological function, or it could also be any other property including binding affinity toward a particular ligand, thermodynamic stability, or catalytic activity. A detailed understanding of how these functions are encoded would allow us to more accurately reconstruct the tree of life and possibly predict future evolutionary events, diagnose genetic diseases before they manifest symptoms, and design new proteins with useful properties. We know that a protein sequence folds into a three-dimensional structure, and this structure positions specific chemical groups to perform a function; however, we're missing the quantitative details of this sequence-structure-function mapping. This mapping is extraordinarily complex because it involves thousands of molecular interactions that are dynamically coupled across multiple length and time scales.

Computational methods can be used to model the mapping from sequence to structure to function. Tools such as molecular dynamics simulations or Rosetta use atomic

representations of protein structures and physics-based energy functions to model structures and functions [1–3] . While these models are based on well-founded physical principles, they often fail to capture a protein's overall global behavior and properties. There are numerous challenges associated with physics-based models including consideration of conformational dynamics, the requirement to make energy function approximations for the sake of computational efficiency, and the fact that, for many complex properties such as enzyme catalysis, the molecular basis is simply unknown [4] . In systems composed of thousands of atoms, the propagation of small errors quickly overwhelms any predictive accuracy. Despite tremendous breakthroughs and research progress over the last century, we still lack the key details to reliably predict, simulate, and design protein function.

Machine learning and artificial intelligence are transforming marketing, finance, healthcare, security, internet search, transportation, and nearly every aspect of our daily lives. These approaches leverage vast amounts of data to find patterns and quickly make optimal decisions. In this chapter, we present how these ideas are starting to impact the field of protein engineering. Instead of physically modeling the relationships between protein sequence, structure, and function, data-driven methods use ideas from statistics and machine learning to infer these complex relationships from data. This top-down modeling approach implicitly captures the numerous and possibly unknown factors that shape the mapping from sequence to function. These statistical models complement physical models and can even be used to improve physics-based models. Statistical models have been used to understand the molecular basis of protein function and provide exceptional predictive accuracy for protein design. We present three key stages in data-driven protein

engineering—(1) representation: how to encode protein sequence/structure/function data, (2) learning: automatic detection of patterns and relationships in data, and (3) prediction: applying the learned models to design new proteins.

## 1.2 The data revolution in biology

The volume of biological data has exploded over the last decade. This is being driven by advances in our ability to read and write DNA, which are progressing faster than Moore's law [5] . Simultaneously, we have also gained unprecedented ability to characterize biological systems with advances in automation, miniaturization, multiplex assays, and genome engineering. It is now routine to perform experiments on thousands to millions of molecules, genes, proteins, and/or cells. The resulting data provides a unique opportunity to study biological systems in a comprehensive and less biased manner.

Protein sequence and structure databases have been growing exponentially for decades (**Fig 1bc**). Currently, the UniProt database [6] contains over 100 million unique protein sequences and the Protein Data Bank [7] contains over 100,000 experimentally determined protein structures. While there is an abundance of protein sequence and structure data, there is still relatively little data mapping sequence to function. ProtaBank is a new effort to build a protein function database [8] . Function data is challenging to standardize because it is highly dependent on experimental conditions and even the particular researcher that performed the experiments. Therefore, statistical modeling approaches are most useful on data that is generated by an individual researcher/research group. This allows for a consistent definition of "function" that is not influenced by uncontrolled experimental factors.

Many sequence-function data sets are generated by protein engineering experiments that involve screening libraries of sequence variants for improved function. These variants may include natural homologs, random mutants, targeted mutants, chimeric proteins generated by homologous recombination, and computationally designed sequences. Each of these sequence diversification methods explores different features of the sequence-function mapping and varies in their information content. Important factors include the sequence diversity of a library, the likelihood of functional vs nonfunctional sequences, and the difficulty/cost of building the desired gene sequences.

Recent advances in high-throughput experimentation have enabled researchers to map sequence-function relationships for thousands to millions of protein variants [9, 10] . These "deep mutational scanning" experiments start with a large library of protein variants, and this library is passed through a high-throughput screen/selection to separate variants based on their functional properties (**Fig 1e**). The genes from these variant pools are then extracted and analyzed using next-generation DNA sequencing. Deep mutational scanning experiments generate data containing millions of sequences and how those sequences map to different functional classes (e.g. active/inactive, binds ligand 1/binds ligand and 2). The resulting data have been used to study the structure of the protein fitness landscape, discover new functional sites, improve molecular energy functions, and identify beneficial combinations of mutations for protein engineering [9, 11–13].

### 1.3 Statistical representations of protein sequence, structure, and function

The growing trove of biological data can be mined to understand the relationships between protein sequence, structure, and function. This complex and heterogenous protein data needs to be represented in simple, machine-readable formats to leverage advanced tools in

pattern recognition and machine learning. There are many possible ways of representing proteins mathematically including simple sequence-based representations or more advanced structure/ physics-based representations. In general, a good representation is low dimensional but still captures the system's relevant degrees of freedom.

### *1.3.1 Representing protein sequences*

A protein's amino acid sequence contains all the information necessary to specify its structure and function. Each position in this sequence can be modeled as a categorical variable that can take on one of twenty amino acid values. Categorical data can be represented using a one-hot encoding strategy that assigns one bit to each possible category. If a particular observation falls into one of these categories, it is assigned a "1" at that category's bit, otherwise it is assigned a "0." A protein sequence of length $l$ can be represented with a vector of $20l$ bits; 20 bits for each sequence position (**Fig 2**). For example, assuming the amino acid bits are arranged in alphabetical order (A, C, D, E ... W, Y), if a protein has alanine (A) at the first position, the first bit would be 1 and the next 19 bits would be 0. If a protein has aspartic acid (D) at the first position, the first two bits would be 0, the third bit 1, and the next 17 bits 0. This encoding strategy can be applied to all amino acid positions in a protein and represent any sequence of length $l$. One-hot encoding sequence representations are widely used in machine learning because they are simple and flexible. However, they are also very high dimensional ($20l \approx$ thousands of variables for most proteins) and therefore require large quantities of data for learning.

Machine learning is widely used in the fields of text mining and natural language processing to understand sequences of characters and words. The tools word2vec and

doc2vec use neural networks to learn vector representations that encode the linguistic context of words and documents [14, 15]. These embeddings attempt to capture word/document "meaning" and are much lower dimensional than the original input space. Similar concepts have recently been applied to learn embedded representations of amino acid sequences [16]. This approach breaks amino acid sequences into all possible subsequences of length *k.* These subsequences are referred to as *k*-mers. As an example, the sequence PRFYLA contains the four 3-mers: PRF, RFY, FYL, and YLA. An amino acid sequence's k-mers are treated as "words" and a neural network is used to learn other words that are found before/after a given word (i.e. a word's context). Importantly, words that are found in similar contexts tend to have similar meanings. This concept can be used to build low-dimensional vector spaces that place similar words close together. For an amino acid sequence, this might mean that one amino acid triplet is comparable to another, and therefore, we only need one variable to represent both. This produces a low-dimensional representation or "protein embedding" that captures the entire protein sequence. These protein embeddings can then be used to model specific properties such as thermostability.

### *1.3.2 Representing protein structures*

The properties of proteins depend on sequence through their structure, therefore structure-based representations provide a more direct link to function. Experimentally determining a protein's three-dimensional structure (via crystallography, NMR, CryoEM) is significantly more challenging and time consuming than determining sequence or function. Therefore, most sequence-function data sets do not contain experimentally determined protein structures. Instead, this missing structural information can be approximated by

taking advantage of the extreme conservation of structures within a family. Homologous proteins with as low as 20% sequence identity still have practically identical three-dimensional structures [17].

A protein's overall fold can be represented by specifying which residues are "contacting" in the three-dimensional structure. These contacting residues could be defined as any pair of residues that has an atom within five angstroms. Other contact definitions could include different distance cutoffs, Cα-Cα distances, or Cβ-Cβ distances. A protein's contact map specifies all pairs of contacting residues and provides a coarse-grained description of the protein's overall fold. Importantly, contact maps are highly conserved within a protein family, and therefore any two evolutionarily related proteins have practically identical contact maps. If we assume a fixed contact map for a protein family, structural information can be represented using a one-hot encoding scheme similar to sequence encoding described above. Each pair of contacting residues can take on one of 400 ($20^2$) possible amino acid combinations, which can be one-hot encoded using 400 bits. Therefore, the structure of a protein with $c$ contacts can be represented with $400c$ bits. In contrast to sequence-based representations, this contact-based representation can capture pairwise interactions between residues. However, this increased flexibility comes at the cost of significantly higher dimensionality.

Three-dimensional protein structures can also be predicted using molecular modeling and simulation software. Most protein sequence-function data sets can take advantage of homology modeling approaches that start with a closely related template structure, mutate differing residues to the target sequence, and run minimization methods to relax the structure into a local energy minimum. State-of-the-art homology modeling

methods can reliably predict protein structures with less than 2 angstrom atomic RMSD [18]. These predicted structures can be analyzed to extract key physiochemical properties such as surface areas, solvent exposure, and physical interactions (**Fig 2**). This approach was recently applied to model the kinetic properties of β-glucosidase point mutants [19]. The substrate was docked into β-glucosidase homology models, and this enzyme-substrate interaction was used to extract 59 physical features such as interface energy, number of intermolecular hydrogen bonds, and change in solvent accessible surface area. A simple linear regression model could relate these physical features to β-glucosidase turnover number, Michaelis constant, and catalytic efficiency. Physics-based representations tend to be lower dimensional than the sequence and contact encodings described above. They may also have good generalization within a protein family or even across protein families because they are based on fundamental biophysical principles.

### 1.4 Learning the sequence-function mapping from data

Advanced pattern recognition and machine learning techniques can be used to automatically identify key relationships between protein sequence, structure, and function. These tools are used for two primary tasks: supervised learning and unsupervised learning. Supervised methods, such as regression and classification, attempt to learn the mapping between a set of input variables and output variables. The term "supervised learning" arises because the algorithms are given examples of input-output mapping to guide the learning process. In contrast, unsupervised methods are not given information about the output variable, but instead try to learn relationships between the various input variables. Similar concepts have been used extensively in quantitative structure-activity relationship (QSAR) models, which are typically used to predict the chemical and biological properties of small

molecules [20]. QSAR models have also been applied to peptide and DNA sequences [21, 22].

### *1.4.1 Supervised Learning (Regression/Classification)*

Regression is a supervised learning technique that is used to model and predict continuous properties. Continuous protein properties could include thermostability, binding affinity, or catalytic efficiency. Regression methods span from simple linear models to advanced, nonlinear models such as neural networks.

Linear regression is the simplest regression technique and applies fixed weights to each input variable. A linear model is described by the following equation:

$$y = X\beta + \epsilon,$$

where $y$ is a vector of continuous output variables, $X$ is a matrix of sequence/structure features (one protein variant per row), $\beta$ is the weight vector, and $\epsilon$ is the model error. The model parameters ($\beta$) can be estimated by minimizing the sum of the squared error. This least-squares parameter estimate has an analytical solution:

$$\hat{\beta} = (X^T X)^{-1}(X^T y)$$

Here, $\hat{\beta}$ corresponds to an estimate of the true $\beta$. $\hat{\beta}$ can then be applied to new proteins to predict their properties:

$$\hat{y} = X_{new}\hat{\beta}$$

Linear regression provides a simple framework for relating sequence/structure to function, and predicting the properties of previously uncharacterized proteins.

Linear regression has been used to model chimeric cytochrome P450 thermostability [23]. A library of chimeric P450s was generated by shuffling sequence elements from three related bacterial P450s [24] . The thermostability of 184 randomly chosen chimeric P450s was determined, and a linear regression model was used to relate sequence to thermostability. Each chimeric protein's sequence was one-hot encoded by specifying which sequence elements were present. This encoding scheme is similar to the sequence-based one-hot encoding described above, but sequence "blocks" are used rather than individual amino acids. This simple regression model revealed a strong correlation between the predicted and observed thermostability (**Fig 3**). The model was applied to predict the thermostabilities of all 6,351 possible sequences in the chimeric P450 library, and the most stable predicted sequences were validated experimentally.

Supervised learning methods, including linear regression, are highly susceptible to overfitting data. A linear model must have at least as many data points as model parameters to avoid overfitting. More complex nonlinear models require even more data. Overfitting occurs when there is not sufficient data and the model fits spurious correlations or noise, rather than the true underlying signal. An overfit model will display very small error on the training data, but large prediction error on new data points.

All statistical models must be evaluated for overfitting and their ability to generalize to new, unseen data points. One method for model validation involves training the model on some fraction of the data and using the remainder to evaluate the model's predictive ability. For example, one could train a model on 60% of the data and test the model on the remaining 40%. This holdout method is simple to implement, but also throws out valuable information because the model is not learning from the entire data set. Cross-

validation is another method for model evaluation that more effectively utilizes the available data. Cross-validation is similar to the holdout method, but rotates through multiple training set-test set combinations. For example, ten-fold cross-validation breaks the data into ten subsets; a model is trained on nine of these subsets and used to predict the tenth subset. This process is repeated over all ten data folds (i.e. testing on all ten subsets) and the results are averaged. Cross-validation allows all data points to be used in model training and evaluation.

Overfitting can be reduced using regularization methods that favor simpler models. Regularized parameter estimation involves minimizing the model's squared error in addition to the magnitude of the model parameters. This can be achieved by including a penalty term on the norm of the parameter vector:

$$\min_{\beta}(X\beta - y)^2 + \lambda\|\beta\|_n$$

Here, the first term corresponds to the model's squared error, the second term is the magnitude of the model parameters, and $\lambda$ tunes the relative influence of these two terms. $n$ determines the type of vector norm and is typically equal to 0, 1, or 2. L0 regularization (n=0) penalizes the total number of non-zero parameters in the model, L1 regularization (n=1) penalizes the sum of the parameter absolute values, and L2 regularization (n=2) penalizes the sum of the squared parameters. This minimization problem can be solved analytically if n=2 or using convex optimization if n=1. The hyperparameter $\lambda$ can be determined using cross-validation. Combinations of these penalties can also be used, such as elastic net regression, which utilizes both L1 and L2 norms.

While regression methods model continuous properties, classification methods are used to model discrete protein properties such as folded/unfolded or active/inactive.

Classifiers are especially important for modeling data generated by high-throughput methods such as deep mutational scanning because these methods often bin proteins into broad functional classes. Classification methods try to relate input feature vectors to functional classes (e.g. active/inactive or folded/unfolded). Like the regression models discussed above, classification models can be evaluated using cross-validation, and regularization can be used to prevent overfitting.

Logistic regression is simple classification method that transforms a linear model through the logistic (sigmoid) function to produce binary outputs. The name "logistic regression" is a misnomer because it actually performs classification rather than regression. Logistic regression parameters can be identified using iterative methods or convex optimization. Logistic regression was recently used to refine molecular energy functions for designing de novo miniproteins [25] . Thousands of miniproteins were designed using Rosetta protein design software, and these designs were screened for folding using a high-throughput yeast display assay. Each protein's structure was modeled and used to generate physical input features such as number of H-bonds, Lennard-Jones energies, and net charge. Logistic regression was then used to map these physical features to whether a design was successful or unsuccessful. The statistical model revealed that a protein's buried nonpolar surface area was a dominant factor in determining design success. The logistic regression model was used to rank designs and drastically improved the rate of successful designs.

Kernel methods are another modeling approach that is widely used in machine learning and bioinformatics. In contrast to the parametric regression/classification methods described above, kernel methods do not require input feature vectors, but instead a user

defined similarity function (or kernel function) is used to compute the "implicit features" by comparing pairs of data points. Kernel methods are more effective at dealing with high dimensional problems than parametric models because they do not have to store large parameter matrices. The similarity function could be as simple as an inner product between feature vectors, or it can represent more complex, potentially infinite dimensional, relationships between data points [26] . This flexibility allows them to learn from unstructured objects such as biological systems. Popular kernel methods include Support Vector Machines (SVMs) and Gaussian Process (GP) regression/classification.

Gaussian processes use kernel functions to define a prior probability distribution over a function space. This allows predictions of both the function mean and its confidence intervals. Gaussian processes have been used to model stability and activity of cytochrome P450s [27] . A structure-based kernel function was developed to define structural similarity between pairs of proteins. GP regression using this kernel function explained 30% more of the variation in P450 thermostability in comparison to linear regression and sequence-based kernels. The structure-based kernel was also used to model enzyme activity and binding affinity for several P450 substrates.

### *1.4.2 Unsupervised learning/semi-supervised learning*

Unlike supervised learning, where the data is labeled or categorized, in unsupervised learning there are no labels associated with each data point. Unsupervised learning can be used to find patterns such as clusters or correlations within data. The main drawback of unsupervised techniques is that the outputs are unknown, i.e. there is no mapping to protein function. However, these techniques still provide valuable information about proteins because of the massive amount of protein sequence data that is currently available.

Examples of unsupervised methods include clustering, where data points are grouped based on similarity, and principal component analysis (PCA). PCA is a projection of data onto lower dimensional space in a way that maximizes the variance of the projection. This converts high dimensional input variables into a set of uncorrelated principle components that are ranked based on their variance. These principle components can be used to reduce the dimensionality of a problem and identify important relationships among variables [28]

Unsupervised methods can be used to identify patterns in multiple sequence alignments (MSAs) of evolutionarily related proteins. Statistical coupling analysis (SCA) analyzes residue coevolution by performing principal component analysis on a protein family's MSA [29] . The dominant principle components consist of positions that coevolve and can reveal networks of spatially connected amino acids called protein sectors (**Fig 4**). Protein sectors have been demonstrated to play roles in protein dynamics and allostery and may represent functional modules [30, 31] . EVmutation is another unsupervised method that models natural sequence variation and simultaneously considers epistasis (non-independence of mutational effects) [32] . Although EVmutation is only parameterized on an MSA (i.e. it is unsupervised), it is capable of predicting the functional effects of amino acid substitutions and residue interdependencies.

Semisupervised methods learn from data sets that contain both unlabeled and labeled data points. Semisupervised approaches can be used in protein engineering to transfer knowledge across protein families. A semisupervised approach was recently developed that trained an unsupervised embedding model (doc2vec) on a large protein sequence database [16] . These embeddings were then used as the inputs for supervised

Gaussian process regression. This approach was used to model channelrhodopsin membrane localization, P450 thermostability, and epoxide hydrolase enantioselectivity.

## 1.5 Applying statistical models to engineer proteins

Statistical modeling approaches provide unprecedented predictive accuracy for a wide variety of complex protein functions/properties. These models can be used to understand protein function and design new proteins. In addition, many classes of statistical models can provide confidence intervals for their predictions. These confidence intervals can be used to gauge whether a prediction is valid or if it contains too much uncertainty to be useful. We discuss several protein engineering strategies that leverage the predictive power of statistical models.

The most straightforward data-driven protein engineering approach involves training a model on a data set and then extrapolating that model to design best predicted sequences. This method was applied to engineer thermostable fungal cellobiohydrolase class II (CBHII) cellulases [33] . A panel of 33 chimeric CBHIIs was characterized for their thermal inactivation half-lives at elevated temperatures. This data was used to train a linear regression model that related sequence blocks to thermal tolerance. This model was then used to design 18 chimeras that were predicted to have enhanced stability relative to the parent enzymes. Most of these designed CBHII chimeras could hydrolyze cellulose at higher temperatures than most stable parent. A key feature of this extrapolation-based design approach is a relatively small training set (<1% of possible chimeras) can be used to make predictions over a massive combinatorial sequence space. The CBHII regression model also pointed to a single sequence block that contributed over 8 °C of thermostability [34] . Further analysis revealed that a single amino acid substitution in that block (C313S)

was responsible for the elevated thermostability. This example highlights how statistical models can be used to uncover molecular mechanisms contributing to protein function.

It is important to consider the space of sequences that a statistical model can make valid predictions on. This prediction domain is highly dependent on the model's sequence/structure representation. For example, consider a model that uses one-hot encoding to represent protein sequences. This model can only learn the effect of amino acids that are observed in the training set, and therefore can only make predictions about sequences composed of combinations of these observed amino acids. Representations that include information about amino acid properties and/or protein structure can broaden a model's prediction domain. Representations that use three-dimensional structural models to extract key physiochemical properties have potential to generalize well within a protein family and even across protein families.

Statistical models can be incorporated into an iterative directed evolution framework. ProSAR uses a statistical model to guide the search for beneficial mutations [35] . This model consists of a one-hot encoded sequence representation and a partial least squares linear regression model to relate sequence to function. A mutational library is screened, and the model classifies each amino acid substitution as deleterious, neutral, beneficial, or underdetermined (i.e. needing more information). Substitutions that are beneficial or underdetermined are combined with new substitutions in the next round, and this screen-and-learn process is repeated over multiple rounds. The ProSAR method was used to engineer bacterial halohydrin dehalogenases (HHDH) to perform a cyanation reaction important for the synthesis of the cholesterol-lowering drug Lipitor [35] . 18 rounds of ProSAR yielded HHDH variants with over 35 mutations and increased the

volumetric productivity of target reaction by ~4,000-fold. More recently, ProSAR-driven evolution was used to evolve ultra-stable carbonic anhydrase variants (107 ˚C thermostability at pH 10 in 4.2 M solvent) that enhanced the rate of $CO_2$ capture by 25-fold over the natural enzyme [36] .

Statistical models can also be used in an active learning setting that very efficiently explores protein sequence space. Active learning involves sequentially designing an informative experiment, performing that experiment, learning from the resulting data, and repeating the process over multiple cycles (**Fig 5a**). For protein engineering, the active learning algorithm must first learn the sequence-function mapping and then apply this knowledge to design optimized sequences. The primary challenge is how to allocate experimental resources toward understanding the sequence-function mapping versus designing optimized sequences. This trade-off is referred to as the "exploration-exploitation dilemma", and the objective is to minimize the amount of exploration that is needed to predict optimized sequences. Upper confidence bound (UCB) algorithms provide a principled framework for trading off between exploration and exploitation modes [37] . The UCB algorithm iteratively selects the point with the largest upper confidence bound (predicted mean plus confidence interval) and therefore encourages sampling of points that are simultaneously optimized and uncertain (**Fig 5b**). A UCB search algorithm was combined with a Gaussian process regression model to optimize cytochrome P450 thermostability [27] . Eight rounds of UCB optimization identified thermostable P450s that were more stable than variants made by rational design, recombination or directed evolution.

## 1.6 Conclusions and future outlook

The protein sequence-structure-function mapping involves thousands of interacting atoms, a practically infinite number of dynamic conformational states, and physical processes that span multiple length and time scales. This mapping is extremely difficult to model from a physical perspective. In contrast, statistical methods are able to learn complex interrelationships directly from experimental data. This top-down understanding of complex systems allows discovery of new functional mechanisms and provides exceptional predictive accuracy.

This chapter provides an overview of emerging data-driven approaches to model and engineer proteins. We have described statistical representations of proteins, how these representations can be used to learn from data, and practical protein engineering applications of these models. As a relatively new field, there is still significant room for improving these methods, especially in the area of sequence and structure representations. Ideal representations would be sparse, but still have a broad prediction domain. These representations may integrate different sources of information (evolutionary, biochemical, and physical) into a single unified model. Advanced machine learning methods such as dictionary learning and deep learning attempt to learn new representations directly from data and could play an important role in protein modeling. Another key challenge for the field is data access and sharing. While there are many interesting sequence-structure-function data sets, they are often buried in a publication's supplemental information and very difficult to parse/organize. Efforts to share data on public repositories and databases such as ProtaBank will greatly accelerate progress in the field.

In addition to proteins, statistical approaches can be used to model genotype-phenotype relationships across all levels of biological organization. For example, linear regression was used to model product titers in a multi-enzyme biosynthetic pathway; this model was then used to optimize enzyme expression levels to maximize overall product production [38] . Another example used compressed sensing methods to model a protein's DNA-binding specificity [39] . Statistical methods have been widely used in genetics relate phenotypes to genetic loci using quantitative trait locus (QTL) mapping [40] .

Data-driven approaches are transforming every field of science and engineering. This revolution has been triggered by the confluence of advances in data generation, data access, and data analysis/interpretation. Advanced experimental technologies are allowing us to analyze biological systems on an unprecedented scale and resolution. The resulting data is also becoming readily accessible through large, public biological databases and repositories. At the same time, there have been tremendous advances in artificial intelligence and pattern recognition. Widespread interest in machine learning has also driven improvements in software packages such as the Scikit-learn and Keras deep learning Python libraries. Data-driven approaches leverage the continuously expanding sea of data and will play an increasingly important role in biological discovery and engineering.

## 1.7 Figures

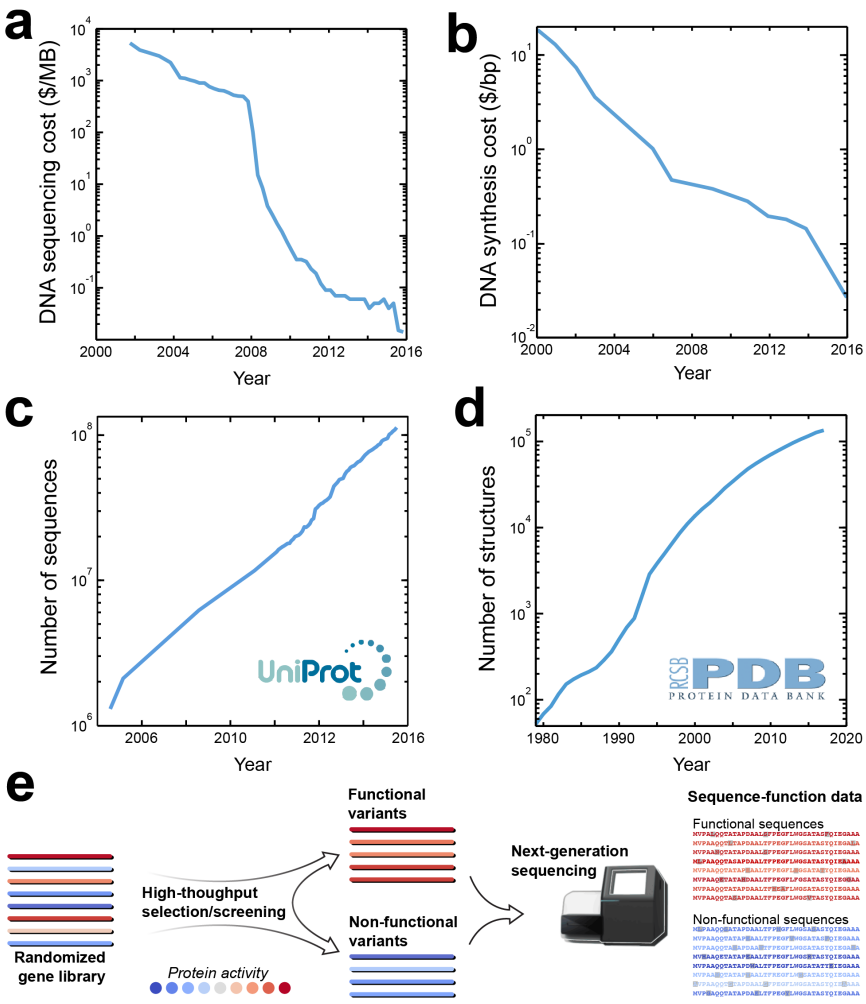


**Figure 1.1:** The growth of biological data. (a,b) DNA sequencing and synthesis technologies are advancing faster than Moore's law. As a result, costs have decreased exponentially over the last two decades. (c,d) Large-scale genomics, metagenomics, and structural genomics initiatives have resulted in exponential growth of protein sequence and structure databases. (e) Deep mutational scanning experiments combine high-throughput screens/selections with next-generation DNA sequencing to map sequence-function relationships for thousands to millions of protein variants.

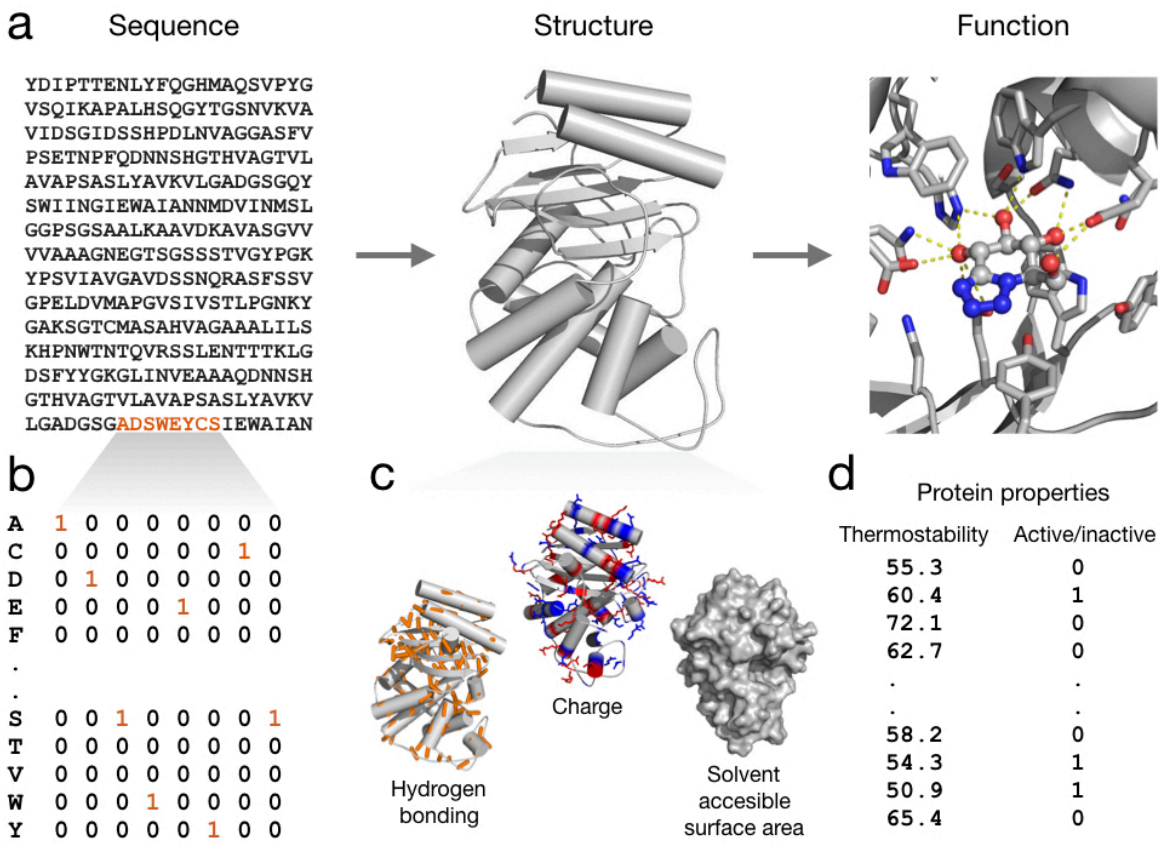


**Figure 1.2:** Sequence, structure, and function representations. (a) A protein's sequence folds into a three-dimensional structure, and this structure determines its function and properties. (b) Protein sequences can be represented using a one-hot encoding scheme that assigns 20 amino acid bits to each residue position. A bit is assigned a value of "1" if the protein has the corresponding amino acid at a particular residue position. (c) Structure-based representations use modeled protein structures to extract key physiochemical properties such as hydrogen bonds, total charge, or molecular surface areas. (d) Protein functions can be continuous properties such as thermostability or catalytic efficiency, or discrete properties such as active/inactive. Discrete properties can be represented using a binary (0 or 1) encoding.

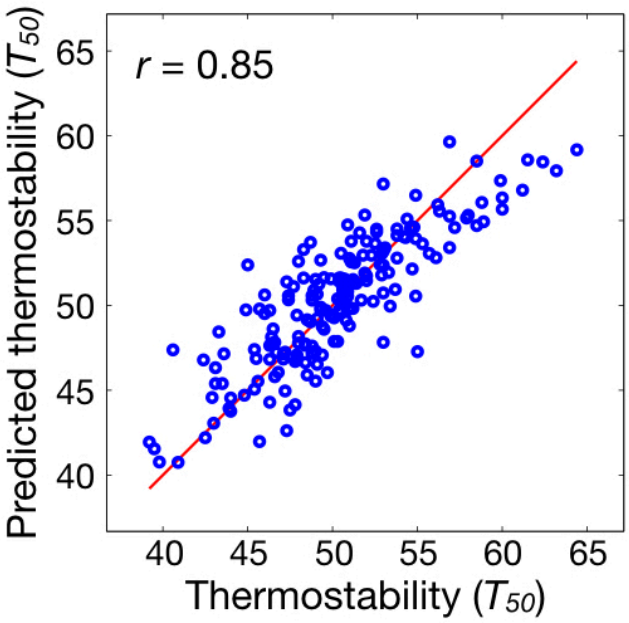


**Figure 1.3:** A linear regression model for cytochrome P450 thermostability. This model relates sequence blocks of chimeric P450s to their thermostability values. The plot shows the model's cross-validated predictions for 184 chimeric P450s.

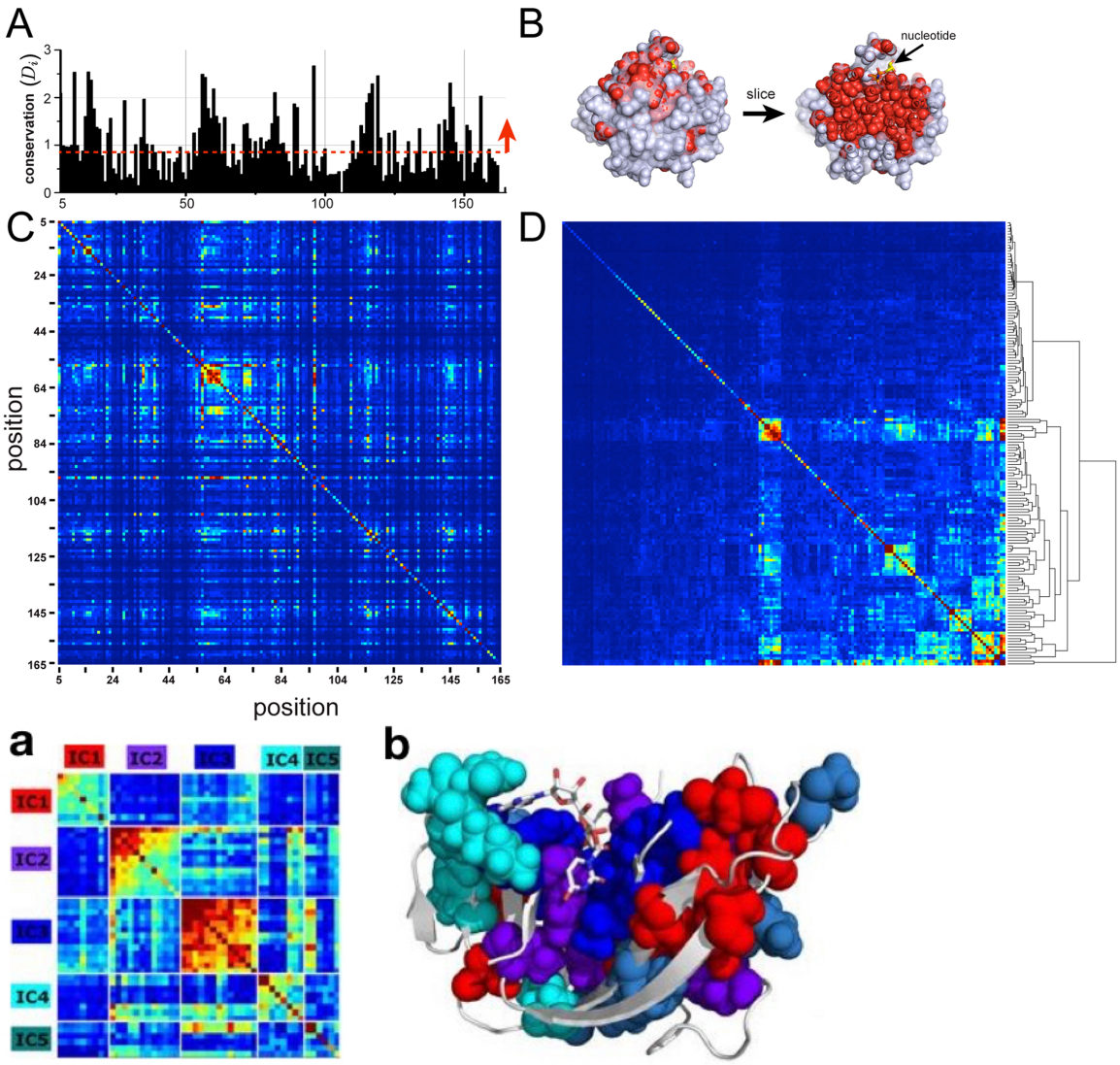


**Figure 1.4:** Unsupervised learning from protein sequences. (A) Statistical coupling analysis of the RNase superfamily reveals five independent components (ICs) that correspond to groups of coevolving residues (B) These five ICs form contiguous "sectors" in the three-dimensional protein structure. Figure was adapted from [31] .

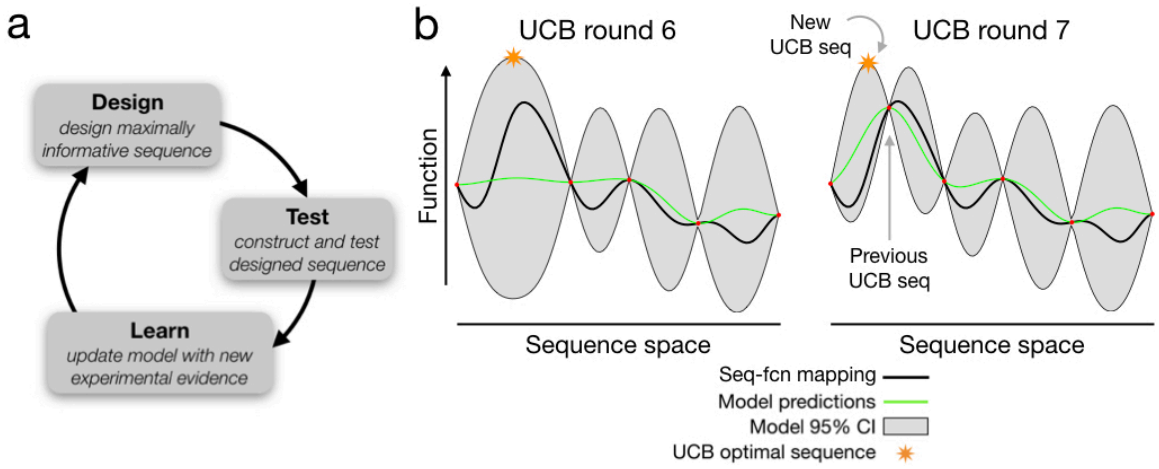


**Figure 1.5.** Active machine learning. (a) Active learning involves designing maximally informative sequences, experimentally characterizing these sequences, learning from the resulting data, and repeating this process over multiple iterations. (b) Upper-confidence bound (UCB) optimization involves iteratively selecting the sequence with the largest upper confidence bound (mean + confidence interval). The schematic illustrates sequence space in one dimension and the true mapping from sequence to function as a black line. Characterized sequences (small red dots) have accurate model predictions and small confidence intervals. The first panel shows five characterized sequences, which cause the model to propose one UCB optimal sequence (marked with a star). The second panel shows the results after this UCB optimal sequence is characterized—this causes a new UCB sequence to be proposed. This iterative process is guaranteed to efficiently converge to the optimal point.

# CHAPTER 2. ENGINEERING CRISPR-CAS12A WITH ALTERED PAMs

## 2.1 Introduction

CRISPR nucleases are leading a major upheaval with widespread use for many important applications in biotechnology and biomedical research [41-42]. CRISPR enzymes are highly versatile and work as a pair of molecular scissors to target user-specified DNA sequences. These enzymes have been used to develop gene editing technologies as well as biodiagnostics and medical therapies. For example, therapies are being developed using CRISPR-Cas9 for genetic diseases such as sickle cell anemia [43]. However, a different CRISPR locus found involving Cas12a has possible advantages. Notably, CRISPR-Cas12a has both increased specificity and reduced off-target effects in both mammalian and plant genomes [44]. Additionally, Cas12a has been shown to be more suitable for multiplexed applications [45] because it can process a single RNA strand into multiple different gRNAs by recognizing repeat sequences.

CRISPR associated (Cas) proteins such as Cas9 and Cas12a have been derived from prokaryotic immune systems. The microbes target the genetic material or DNA of invading virus using Cas9 and Cas12a by using a guide-RNA transcribed separately. However, to prevent self-targeting of the guide-RNA transcripts, microbes have evolved to additionally require a protospacer adjacent motif or PAM sequence (2-6 bp) for Cas-function by forming an R-loop (See Figure 2.1). For example, the wild type version of Cas12a derived from *Acidaminococcus sp. BV3L6* (AsCas12a) can be used to only target DNA that has a

correct corresponding PAM motif (TTTV). The PAM motif is highly conserved with most Cas12a homologs requiring thymidine-rich PAMs 18 to 23 base pairs before the cleavage site. CRISPR-Cas proteins have varying PAM profiles and thus PAMs are attractive targets for protein engineering.

Millions of years of evolution have made PAMs challenging to alter. Protein design is non-trivial [46]; the sequence space is vast and on the order of $20^N$ for proteins with N amino acids [47]. Traditional methods of rational design and directed evolution are slow. While rational design approaches have high function diversity, they require prior knowledge. Moreover, directed evolution approaches yield low functional diversity. As Cas12a is a multi-domain protein and ~1300 amino acids long with unknown conformations, CRISPR engineering can be accelerated using data-driven strategies [48].

Here, we develop a data-driven strategy to engineer Cas12a with a high throughput screening assay, a large library of variants and statistical analysis. Fluorescence activated cell sorting provided a platform for a high throughput screening assay. Homologous recombination of structurally conserved regions or 'blocks' Cas12a yielded a large library of variants. With six homologs and 8 blocks, $6^8$ variants are generated (See Figure 2.2). In addition, to reconstruct whole chimeric protein sequences, sorted population sequences ($10^6$) were run using nanopore sequencing methods and processed on high-throughput computing (CHTC) pools. Experimental data analysis using log enrichment showed a strong shift after sorting. Together, these data show that our data-driven strategy can be used to engineer Cas12a with novel molecular functions.

## 2.2 Results

### *2.2.1 Development of a high-throughput assay for Cas12a-PAM binding*

Fluorescence activated cell sorting (FACS) is a commercialized application of flow cytometry. In a FACS instrument, fluorescent parameters of cells can be analyzed by a focused laser beam at high rates of $10^7$/hour [49]. A charge is applied to the cells, which are deflected into a collection tube by a charged plate. We report a FACS-based high-throughput screening method for screening dCas12a enzyme libraries. The screening is based on an assay called PAM-SCANR developed in the Beisel lab. As shown in Figure 2.3A and 2.3B, the assay links dCas12a binding and GFP (Green Fluorescent Protein) production [50]. Functional dCas12a binds to their target site to repress the *lacI* gene that is repressing the expression a green fluorescent protein (GFP) reporter. With high repression activity, higher GFP is expressed and fluorescence can be quantified. To benchmark this method for protein engineering, an additional control with a non-interacting dCas12a was made by early truncation using a stop codon and addition of a unique NotI restriction site. Flow cytometry was performed using cells transformed with Beisel labs 3-plasmid PAM-SCANR system with (1) functional PAM (TTTA) and non-functional PAM (AAAA) reporter genes; (2) plasmid encoding guide-RNA to *lacI* promoter; (3) plasmid encoding protein dFnCas12a/NotI-dFnCas12a. Flow cytometry results are shown in Figure 2.3C show an expected high fluorescence visible to the eye with the functional PAM reporter gene and functional dFnCas12a.

To further adapt PAM-SCANR to protein engineering, protein expression was optimized using a bicistronic design (BCD) for the ribosome binding site. The initial

ribosome binding site used in PAM-SCANR wasn't producing fluorescent activity with the chimeras. In BCD, a leader peptide is produced followed by your target gene under control of the same ribosome promoter. Here, the stop codon of the leader peptide (TAA) and start codon (ATG) of your gene overlap so that ribosomes are 'pointed' to the target gene of interest [51].

### *2.2.2 Chimeric Cas12a library design and construction*

To perform chimeragenesis for large multi-domain proteins such as Cas12a, protein domains can serve as logical components to shuffle. As larger proteins with multiple domains can have more conformational diversity for function, mutations distal from the active site can have substantial effects on enzyme activity [52]. Conserved structural domains were identified based on crystal structures of FnCas12a, AsCas12a and LbCas12a [53]. By shuffling conserved structural protein domains, we infer that there should be low disruption of structural contacts and a library with high functional activity. Six parents were chosen to create a family of chimeric proteins. All known parent enzymes commonly used as mammallian genome editing tools were chosen as parent homologs – *Acidaminococcus sp. BV3L6* (AsCas12a), *Lachnospiraceae bacterium ND2006* (LbCas12a), *Francisella tularensis subsp. novicida* (FnCas12a), and *Moraxella bovoculli* (MbCas12a). Additionally, two other homologous proteins were selected with low activities to add functional diversity – Lachnospiraceae bacterium COE1 (Lb6Cas12a) and Porphyromonas crevioricanis (PcCas12a). Moderately low pairwise sequence identity was found (36-49%) with a multiple sequence alignment of the parent proteins.

Based on the sequence alignment of domain boundaries, 8 blocks were made as

shown in Figure 2.4A – (1) WED-I, REC-I; (2) REC-II; (3) WED-II; (4) PI; (5) WED-III; (6) RuvC-I, II; (7) Nuc; (8) RuvC-III. Catalytic residues in the RuvC domain (Block 6) were inactivated to make a dCas12a library. This block could be switched to having active residues for use in a screening assay with active Cas12a. Unique sequence motifs that could be recognized by type IIS restriction enzymes were used as block breakpoints to enable use of Golden Gate assembly [54]. These sites were carefully selected to result in overhang sequences that are assembled in a specific orientation without scarring. After Golden gate assembly in two steps and linearization, the plasmid library can be sequenced for analysis. As shown in Figure 2.4B, we constructed 2 mini-libraries with half of the domains each to produce $6^4$ or 1296 combinations. The resulting theoretical library size is then $1296^2$ or 1,679,616 chimeras. We showed moderate coverage of this library after construction with 3.5 million transformants checked by counting colony forming units (cfu). The constructed library for this protein engineering strategy has close to 3% fraction functional based on colony counts and flow cytometry data. This translates to ~50,000 active Cas12a mutants binding to the functional PAM sequence TTTA.

### *2.2.3 Sequence-function mapping with FACS and nanopore sequencing*

For testing reproducibility of our results, FACS was done three times on the constructed chimeric library binding to the canonical PAM – TTTA. The library was transformed into CB414 cells (a strain of *E. coli* from Beisel lab which is ΔCRISPR-Cas and Δ*lacI-lacZ*) containing the guide-RNA plasmid and TTTA PAM-SCANR reporter gene plasmids. The library was grown overnight at 37°C, 250 rpm in triplicate before sorting on BDFACSAria (UW Flow Core). We controlled for random growth bias in the experiment by sequencing

libraries both before and after sorting. For each sorting experiment, roughly 15,000,000 cells were sorted with ~20,000 cells sorted as positive. After cells were sorted, they were recovered by growth overnight at 37°C, 250 rpm and on agar plates. See sort statistics in Table 1.

A sequencing read was labelled as functional if identified in the sorted library. To identify sequences, we can employ short-read or long-read sequencing. Short-read or Ilumina sequencing can be problematic with large chimeric proteins because the sequence homology between the small sequencing fragments of different variants can be hard to discern. We thus performed long-read nanopore sequencing on our chimeric libraries. For an unsorted library, we show that nanopore sequencing is able to capture long plasmid reads (8kb fragments) and could capture 4 million reads with a median length of 7kb. However, nanopore sequencing data is error prone with error rates as high as 15% [55]. Fortunately, this is not an experimental limitation for use on chimeric protein sequences which are made up of blocks of DNA. These blocks of DNA can be identified with the errors using computational methods by computing the sequence identity between the read and a block. Correct block identifications had >90% sequence identity compared to <80% sequence identity for incorrect block identifications.

Each read was aligned with every possible block sequence (6 parents and 8 blocks) and both top and bottom strands of DNA using Basic Local Alignment Search Tool (BLAST). Due to the number of reads in ever library being on the order of $10^6$ and 96 BLAST alignments per read, processing on a local drive, assuming 1 second per alignment, would take >100 days of computational time. We developed a computational pipeline in Python3 to shorten the computational time using high-throughput computing ‘pools’ at the

Center for High-Throughput Computing (CHTC) at UW-Madison. We validated our method by performing a single alignment for predicted sequences as shown in Figure 2.5. Using this computational pipeline, we identified close to 260,000 unique chimeric proteins in a library after Golden gate assembly.

### *2.2.4 Enrichment analysis of the blocks*

After nanopore sequencing on the FACS runs, the sorted and unsorted datasets for each sorting experiment were analyzed using site-wise or block-wise enrichment and sequence based enrichment. Analysis was done by normalizing number of observations of a block or a sequence by the total number of observations in a dataset. All data sets have 1,000,000 to 2,000,000 sequences as shown in Table 2.

Site-wise enrichment was calculated for every parent at each block position as shown in Figure 2.6. This initial bias we see in the unsorted libraries is likely due to growth bias from growth in *E. coli*. The unsorted *(u*) and sorted (*s*) libraries were both sequenced for each sorting replicate. The number of observations or counts for each sorting replicate was then analyzed using log enrichment scores defined as:

$$ES_{ij} = \log\frac{s_{ij}}{s_{Tj}} - \log\frac{u_{ij}}{u_{Tj}}, i = 1,2,3 \dots 6, j = 1,2,3 \dots 8$$

where *i* and *j* correspond to the parent number and block number respectively; *s* and *u* correspond to sorted and unsorted counts, T refers to total for that block number. We also demonstrate that our screens are reproducible. The enrichment scores are highly correlated between sorting replicates as shown by the goodness of fit or R-squared values (>0.85). This is further demonstrated with calculating the best chimeric sequence for each replicate.

The best chimeric sequence was constructed piece-wise using enrichment scores: 45216635, 45616645, 45616635 for the first, second and third sorting replicate respectively. They are shown to be similar with a consensus sequence of 45616635 (*Fn-Lb6-Pc-As-As-Lb-Lb6)*.

Sequence-wise enrichment was calculated using a similar metric for enrichment scores defined as:

$$ES_i = \log\frac{s_i}{s_T} - \log\frac{u_i}{u_T}, i = 1,2,3 \ldots > 10^6$$

where *i* instead corresponds to the sequence number in the set; *s* and *u* correspond to sorted and unsorted counts, T refers to total for that sequencing set. We did observe 65,443 unique sequences across replicates. Due to the large theoretical library size (~1,600,000), there is some more variability in sequences observed between replicates that limited our analysis.

### *2.2.5 Classification analysis for FnCas12a-PAM binding*

CRISPR-Cas12a-PAM interactions have been characterized by collaborators in the Beisel lab for all 256 PAMs (4 nucleotide) based on enrichment [50]. Using this, we modeled FnCas12a-PAM binding to further our understanding of how many experiments to run and how much data we would need to make a generalized model for Cas12a-PAM binding. Future studies could be done with data from FACS screens using different subsets of PAM reporter plasmids to build a generalized model of Cas12a's binding properties to all 256 PAMs. We performed a classification analysis on this data to test different models for analysis with future experiments.

For classification analysis, we looked at clustering the top PAMs of the dataset. The average enrichment score of the top 16 PAMs was 10-fold higher than the average of the remaining PAMs. Thus, we set the top 16 PAMs and classified them as 'binders' for our model: TTTG, TTTC, TTTA, GTTA, GTTC, CTTA, GTTG, CTTC, CTTG, AAAA, ATTG, ATTA, AAAT, ATTC, AATT, TCTC. We compared different machine learning methods in Python3's `sklearn` package to predict binding – linear regression, logistic regression and neural networks. Models were trained using data from 2 additional PAMs (TTCA, TGTA) and then used to predict the binding (yes/no) for the remaining 253 PAMs. These PAMs were chosen as they captured major deviations from the canonical sequence of TTTA at the second and third nucleotide positions. The ROC curves shown in Figure 2.7 show that all of our models perform well with high AUC values. We constructed multi-layer perceptrons in `sklearn` (a form of neural networks) comprised of three layers – 100 nodes on the input layer, 10 nodes on the other layers. Logistic regression has a high AUC and is the most suitable for future analysis as the model is less complex.

## 2.3 Discussion

Protein sequence space is high-dimensional. For Cas12a, which is 1300 amino acids long, it would take billions of years to test every possible sequence– $20^{1300}$ mutants ~ infinity. This combinatorial explosion continues with interaction terms between residues; Cas12a uses WED, REC1 and PI domains to interact with the PAM site [56]. Proteins are thus highly versatile and amenable to high-throughput methods and data-driven strategies. Our chimeric protein libraries, FACS-based screening assay and nanopore sequencing methods demonstrate a data-driven strategy to explore this vast space. We have identified sequences for further testing by looking at the best sequence from block-based enrichment scores.

The experiments and statistical analysis we present can be used to perform informed or augmented rounds of directed evolution.

We demonstrate the construction of a large chimeric library, made using domain swapping, for analysis. Chimeric proteins maintain high functionality as sequences selected through evolution and mimics iterative homologous recombination seen in nature. Random mutagenesis can lead to many unstable proteins that are less functional; chimeric proteins occupy a functionally 'enriched ridge' in protein sequence space as they are combinations of functional folds [57]. While orthologous proteins maintain conserved active sites, chimeric proteins have been shown to change substrate specificity in cytochrome p450s [58-60]. Indeed, for Cas12a, domain swapping has been used to change PAM specificity [61]. Thus, there is reason to infer that our library would contain mutants with novel PAM recognition profiles. A previous study has been conducted to alter PAM specificity in Cas12a effector proteins by altering residues in close proximity to the PAM DNA duplex using random mutagenesis [62].

Our developed assay's level of sequencing throughput ($10^6$) and sequence diversity ($10^6$) is what makes our strategy data-driven. We showed that sequence-function mapping using FACS is robust and can screen millions of cells containing variants of Cas12a with >90% sorting efficiency. FACS provides high quality sorting using commercial equipment. We see that long-read sequencing or nanopore sequencing can identify millions of variants using a computational pipeline to identify chimeric sequences. MinION nanopore sequencing provides high sequencing throughput for larger proteins such as Cas12a.

We demonstrate a generalized model for Cas12a-PAM binding using machine learning methods on FnCas12a-PAM enrichment data from the Beisel lab. Using just 2 additional PAMs, we can capture most of the functional diversity for all of the possible PAM sequences. This is promising in that for a given chimera, we may only need to test only a few PAM sequences to predict activity on others. However, since this data analysis was limited to FnCas12a, generalizing our analysis would first need to be limited to proteins that are close to the homolog FnCas12a and other parent proteins.

You can only obtain proteins for which you select the properties while performing your assay. Our dCas12a assay screens for binding activity to a particular PAM which may not necessarily translate to active Cas12a PAM profile. A kinetic study using stopped flow fluorescence found cleavage to be favorable and fast after R-loop formation [44]. Thus, it is theorized that assaying for site specific binding using dCas12a should correlate well with subsequent site specific cleavage using Cas12a. However, our assay does not select for PAM specificity or off-target effects – both desirable targets for protein engineering.

Future directions on this work would include testing for other Cas12a PAMs or other screening conditions such as pH, temperature. For example, in the application for genome editing in plants, Cas12a is temperature sensitive at temperatures lower than ~28°C [63]. It would be simple to assay for PAM binding at different temperature by just growing the bacteria at different temperatures overnight. Library growth conditions can be modulated to the desired screening outcome.

## 2.4 Figures and Tables

**Table 2.1:** Sort statistics for each sorting replicate. Each sort was done after controls were run – GFP reporter gene with both positive and negative PAM and dFnCas12a plasmids. Roughly 15,000,000 cells were run with 20,000 events sorted positively. Function after sort is checked by colony counts for GFP from post-sort plating on agar plates.

| | Sort replicate 1 | Sort replicate 2 | Sort replicate 3 |
|---|---|---|---|
| Sorted cells | 19,178 | 19,725 | 18,208 |
| Total events | 16,156,994 | 15,527,086 | 15,715,748 |
| Time to sort | 23:15 | 25:02 | 22:20 |
| Function after sort | 88% | 91% | 91% |

**Table 2.2:** Nanopore sequencing data – chimeras with complete sequences. All data sets contain 1,000,000 to 2,000,000 complete sequences after analysis on the CHTC.

| | Unsorted (counts) | Sorted (counts) |
|---|---|---|
| Sort replicate 1 | 1,699,697 | 1,373,371 |
| Sort replicate 2 | 1,763,761 | 1,870,868 |
| Sort Replicate 3 | 1,237,235 | 1,445,443 |


**Figure 2.1.** R-loop formation of Cas12a. An RNA molecule guides Cas12a to the target DNA site with matching genetic material. This does first require binding to the PAM site preceding the target DNA site (TTTA for *Acidaminococcus sp. BV3L6)*. Cleavage of DNA occurs 18 to 23 bases downstream of the PAM site on the non-target DNA strand. The double-stranded break is staggered.

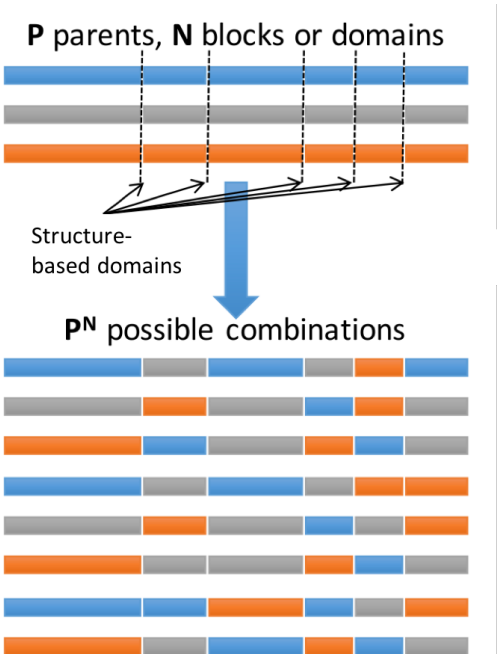


**Figure 2.2.** Chimeragenesis of proteins. Parent proteins are chosen based on sequence homology. Blocks or domains of parent proteins are recombined to make chimeras with novel molecular folds.

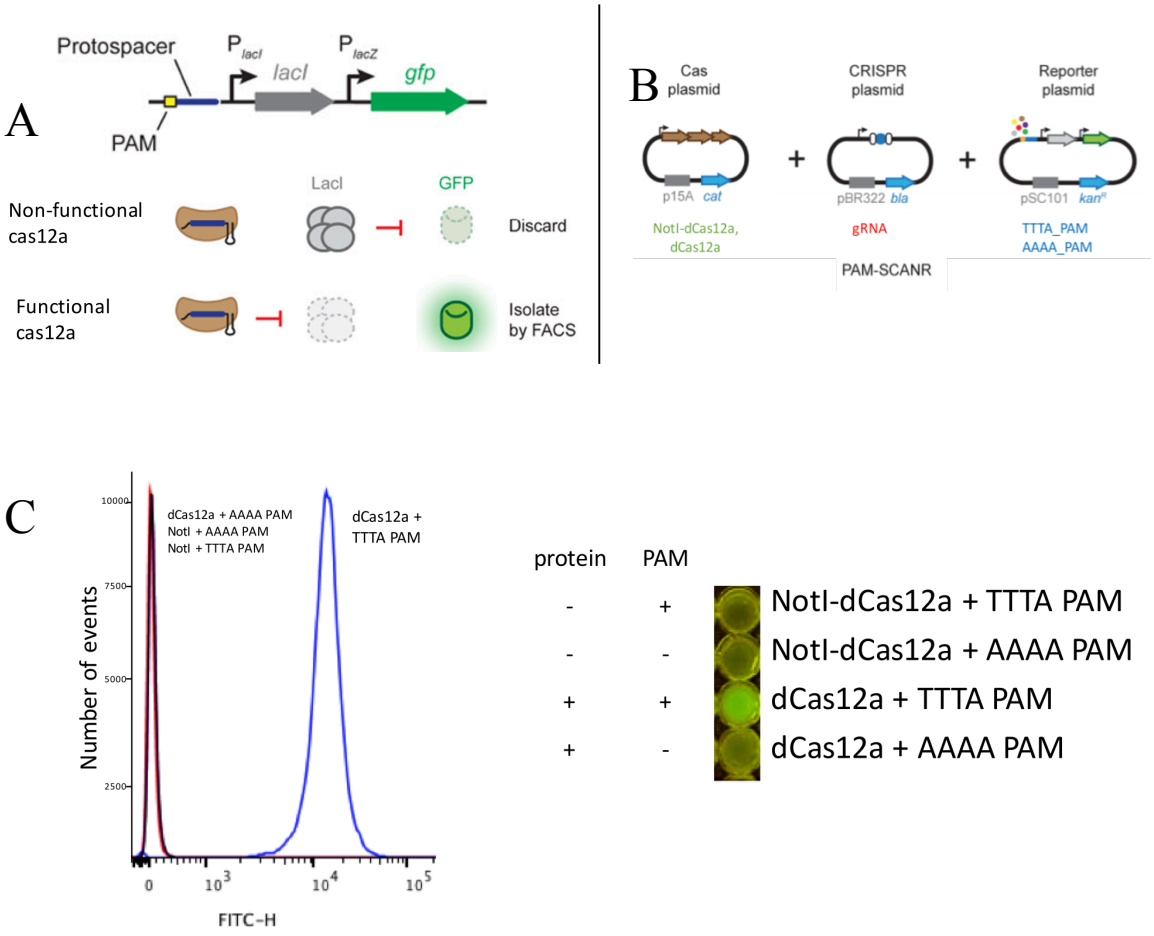


**Figure 2.3.** Assay for Cas12a-PAM binding**.** (A) Fluorescent assay for binding activity with dCas12a NOT-gate repression system from Beisel lab. dCas12a will bind to a PAM site adjacent to a lacI promoter. GFP is under control of *lacI* with a *lacZ* promoter site. Functional dCas12a thus turns fluorescent by inhibiting *lacI* expression and thus stopping *lacI* from blocking GFP expression (B) Assay from Beisel lab called PAM-SCANR has three plasmid system. PAM-SCANR consists of a Cas plasmid expressing dCas12a, a CRISPR plasmid expressing guideRNA and a reporter plasmid expressing GFP with different PAM sites. To test the assay, a Cas plasmid was made with an early stop codon NotI in the middle of the protein to form an inactivated version as a control. (C) The three-plasmid system was tested using two different PAM sites – positive (TTTA) and negative (AAAA) by transformation into *E. coli*. Flow cytometry shows a clear separation for the positive sample with functional dCas12a and positive PAM. Results are also visible under a blue-light transilluminator (right).

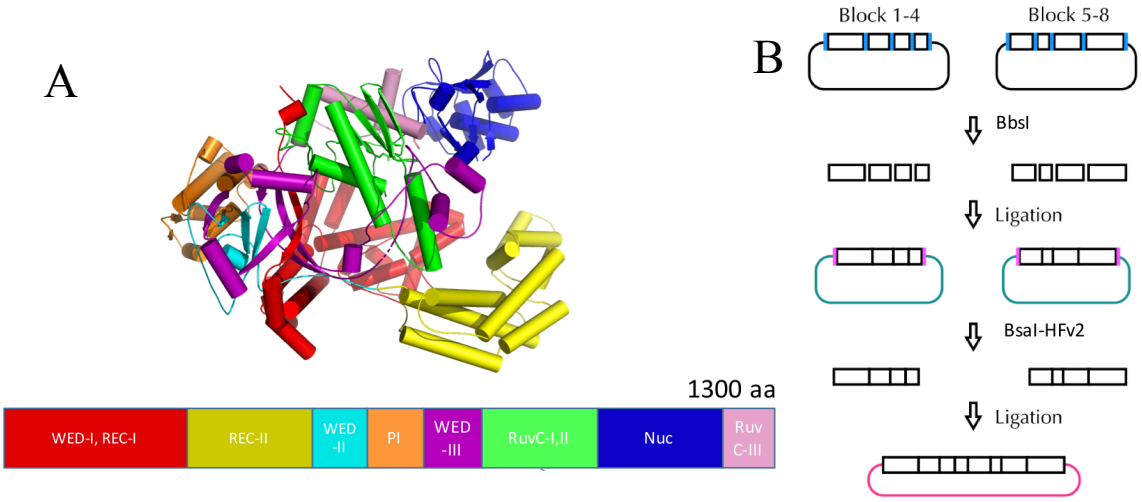


**Figure 2.4.** CRISPR-Cas12a blocks and cloning. (A) Pymol rendering for *Acidaminococcus sp. BV3L6* (AsCas12a) with different structural domains highlighted. Domains are highlighted by block number as shown on the bottom starting with WED-I, REC-I. (B) Domains shown are recombined in two halves using Golden gate cloning [54]. Golden gate cloning is a method using type IIs restriction enzymes (BbsI, BsaI-HFv2) to recombine DNA fragments in a specific orientation.

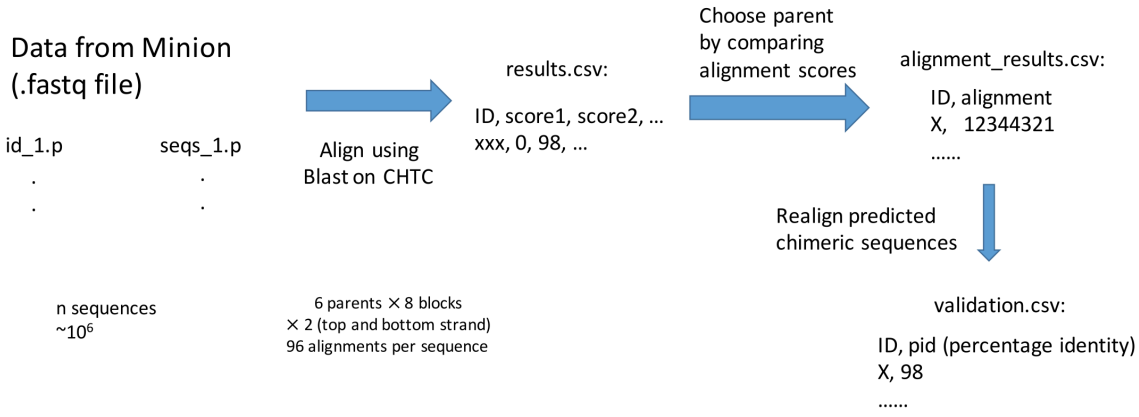


**Figure 2.5**. MinION nanopore sequencing data processing workflow. Sequences are aligned using BLAST on the CHTC to get alignment scores for each parent-block combination for the top and bottom strands. The scores are compared to identify the chimeric sequence. These sequences are validated by realignment with the sequencing read to ensure a high percentage identity.

| | b1 | b2 | b3 | b4 | b5 | b6 | b7 | b8 |
|---|---|---|---|---|---|---|---|---|
| p1 | 0.35 | -0.56 | 0.10 | 0.91 | -0.77 | 0.37 | -0.09 | -0.82 |
| p2 | -1.27 | -0.13 | 1.15 | -0.05 | -0.71 | -0.69 | -0.39 | -0.60 |
| p3 | -1.38 | -0.46 | -0.61 | -0.44 | -1.30 | -1.34 | 0.25 | -0.97 |
| p4 | 1.08 | -0.07 | -0.55 | 0.07 | 0.27 | -1.24 | 0.19 | 0.02 |
| p5 | -0.38 | 0.87 | -1.15 | -0.86 | 0.43 | 0.50 | 0.15 | 0.69 |
| p6 | -0.60 | -0.82 | 0.81 | -0.61 | 0.80 | 0.89 | -0.23 | -0.09 |

| | b1 | b2 | b3 | b4 | b5 | b6 | b7 | b8 |
|---|---|---|---|---|---|---|---|---|
| p1 | 0.39 | -0.74 | 0.34 | 0.90 | -0.77 | 0.51 | 0.00 | -0.51 |
| p2 | -1.16 | -0.12 | 0.86 | -0.65 | -0.77 | -0.61 | -0.29 | -0.49 |
| p3 | -1.37 | -0.51 | -0.57 | -0.40 | -1.10 | -1.19 | 0.21 | -1.02 |
| p4 | 1.03 | 0.06 | -0.61 | 0.13 | 0.35 | -1.34 | 0.25 | -0.05 |
| p5 | -0.85 | 0.89 | -1.45 | -0.97 | 0.21 | 0.42 | 0.12 | 0.68 |
| p6 | -0.69 | -0.76 | 0.95 | -0.61 | 0.93 | 0.79 | -0.31 | -0.17 |

| | b1 | b2 | b3 | b4 | b5 | b6 | b7 | b8 |
|---|---|---|---|---|---|---|---|---|
| p1 | 0.38 | -0.82 | 0.19 | 0.91 | -0.84 | 0.46 | -0.06 | -0.73 |
| p2 | -1.34 | -0.12 | 0.80 | -2.54 | -0.59 | -0.70 | -0.46 | -0.71 |
| p3 | -1.50 | -0.37 | -0.57 | -0.33 | -1.32 | -1.34 | 0.35 | -0.84 |
| p4 | 1.08 | -0.18 | -0.63 | 0.05 | 0.30 | -1.33 | 0.08 | -0.01 |
| p5 | -0.58 | 0.92 | -1.19 | -0.99 | 0.24 | 0.48 | 0.12 | 0.73 |
| p6 | -0.56 | -0.88 | 0.98 | -0.66 | 0.99 | 0.87 | -0.23 | -0.25 |

**Figure 2.6**. Enrichment scores on Cas12a blocks. The results are shown from sorting replicate 1 (left), sorting replicate 2 (middle), sorting replicate 3 (right) We performed enrichment analysis using log scores on each sort. The site-wise or block-wise enrichment scores look reproducible across sorts with a similar pattern seen across replicates here. Here, p1 refers to AsCas12a, p2 refers to MbCas12a, p3 refers to LbCas12a, p4 refers to FnCas12a, p5 refers to Lb6Cas12a and p6 refers to PcCas12a.

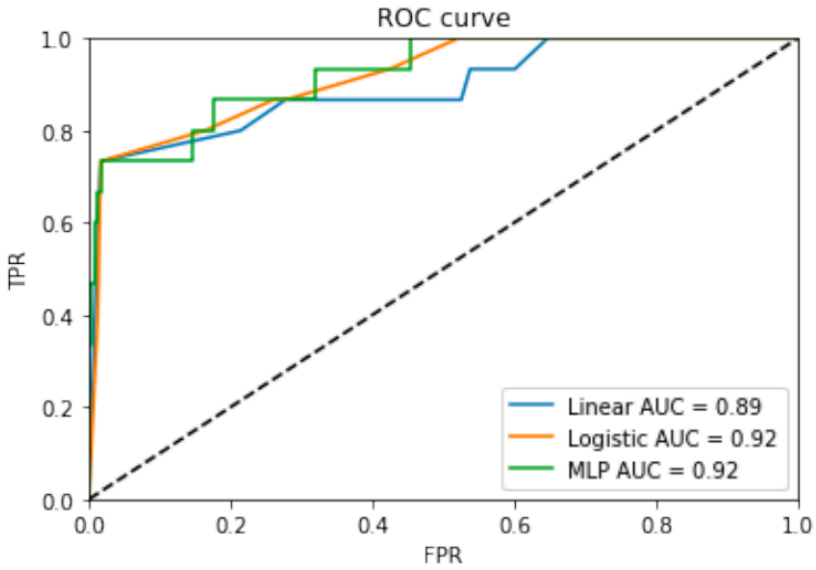


**Figure 2.7.** Comparison of machine learning models for PAM classification. Classification is performed on 253 PAMs after training a model made using linear regression, logisitic regression and multi-layer perceptrons (MLPs) on data for 3 PAMs (TTTA, TTCA, TGTA). ROC curves are drawn and AUCs are calculated to compare these models. The AUCs are comparable and indicate that our models perform well at classification of binder PAMs.

# REFERENCES

**[1]** Lazaridis T, Karplus M (2000) Effective energy functions for protein structure prediction. Curr Opin Struct Biol 10(2):139–145.

[2] Li Z, Yang Y, Zhan J, Dai L, Zhou Y (2013) Energy functions in de novo protein design: current challenges and future prospects. Annu Rev Biophys 42:315–35.

[3] Bradley P, Misura KM, Baker D (2005) Toward high-resolution de novo structure prediction for small proteins. Science (80- ) 309(5742):1868–1871.

[4] Baker D (2010) An exciting but challenging road ahead for computational enzyme design. Protein Sci 19(10):1817–1819.

[5] Kosuri S, Church GM (2014) Large-scale de novo DNA synthesis: Technologies and applications. Nat Methods 11(5):499–507.

[6] Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Natale D a, O'Donovan C, Redaschi N, Yeh L-SL (2017) UniProt: the universal protein knowledgebase. Nucleic Acids Res 45:D158–D169.

[7] Rose PW, Prlic A, Altunkaya A, Bi C, Bradley AR, Christie CH, Di Costanzo L, Duarte JM, Dutta S, Feng Z, Green RK, Goodsell DS, Hudson B, Kalro T, Lowe R, Peisach E, Randle C, Rose AS, Shao C, Tao Y-P, Valasatava Y, Voigt M, Westbrook JD, Woo J, Yang H, Young JY, Zardecki C, Berman HB, Burley SK (2017) The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic Acids Res 45:D271–D281.

[8] Wang CY, Chang PM, Ary ML, Allen BD, Chica RA, Mayo SL, Olafson BD (2018) ProtaBank: A repository for protein design and engineering data. Protein Sci. doi:10.1002/pro.3406.

[9] Hietpas RT, Jensen JD, Bolon DNA (2011) Experimental illumination of a fitness landscape. Proc Natl Acad Sci U S A 108(19):7896–7901.

[10] Araya CL, Fowler DM (2011) Deep mutational scanning: Assessing protein function on a massive scale. Trends Biotechnol 29(9):435–442.

[11] Romero PA, Tran TM, Abate AR (2015) Dissecting enzyme function with microfluidic-based deep mutational scanning. Proc Natl Acad Sci U S A 112(23):7159–7164.

[12] Whitehead TA, Chevalier A, Song Y, Dreyfus C, Fleishman SJ, De Mattos C, Myers CA, Kamisetty H, Blair P, Wilson IA, Baker D (2012) Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. Nat Biotechnol 30(May):1–9.

[13] Bloom JD (2014) An experimentally determined evolutionary model dramatically improves phylogenetic fit. Mol Biol Evol 31(8):1956–1978.

[14] Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient Estimation of Word Representations in Vector Space. CoRR abs/1301.3. doi:10.1162/153244303322533223.

[15] Le Q V, Mikolov T (2014) Distributed Representations of Sentences and Documents. CoRR abs/1405.4. doi:10.1145/2740908.2742760.

[16] Yang KK, Wu Z, Bedbrook CN, Arnold FH (2018) Learned protein embeddings for machine learning. Bioinformatics 34(15):2642–2648.

[17] Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5(4):823–826.

[18] Misura KM, Chivian D, Rohl CA, Kim DE, Baker D (2006) Physically realistic homology models built with ROSETTA can be more accurate than their templates. Proc Natl Acad Sci 103(14):5361–5366.

[19] Carlin DA, Caster RW, Wang X, Betzenderfer SA, Chen CX, Duong VM, Ryklansky C V., Alpekin A, Beaumont N, Kapoor H, Kim N, Mohabbot H, Pang B, Teel R, Whithaus L, Tagkopoulos I, Siegel JB (2016) Kinetic characterization of 100 glycoside hydrolase mutants enables the discovery of structural features correlated with kinetic constants. PLoS One. doi:10.1371/journal.pone.0147596.

[20] Dudek A, Arodz T, Galvez J (2006) Computational Methods in Developing Quantitative Structure-Activity Relationships (QSAR): A Review. Comb Chem High Throughput Screen 9(3):213–228.

[21] Hellberg S, Sjöström M, Skagerberg B, Wold S (1987) Peptide Quantitative Structure-Activity Relationships, a Multivariate Approach. J Med Chem 30(7):1126–1135.

[22] Jonsson J, Norberg T, Carlsson L, Gustafsson C, Wold S (1993) Quantitative sequence-activity models (QSAM) - tools for sequence design. Nucleic Acids Res 21(3):733–739.

[23] Li Y, Drummond DA, Sawayama AM, Snow CD, Bloom JD, Arnold FH (2007) A diverse family of thermostable cytochrome P450s created by recombination of stabilizing fragments. Nat Biotechnol 25(9):1051–1056.

[24] Otey CR, Landwehr M, Endelman JB, Hiraga K, Bloom JD, Arnold FH (2006) Structure-Guided Recombination Creates an Artificial Family of Cytochromes P450. PLoS Biol 4(5):e112.

[25] Rocklin GJ, Chidyausiku TM, Goreshnik I, Ford A, Houliston S, Lemak A, Carter L, Ravichandran R, Mulligan VK, Chevalier A, Arrowsmith CH, Baker D (2017) Global analysis of protein folding using massively parallel design, synthesis, and testing. Science (80- ) 357(6347):168–175.

[26] Rasmussen CE, Williams C (2006) Gaussian Processes for Machine Learning (MIT Press, Cambridge, MA).

[27] Romero PA, Krause A, Arnold FH (2013) Navigating the protein fitness landscape with Gaussian processes. Proc Natl Acad Sci U S A 110(3):E193–E201.

[28] Bishop CM (2006) Pattern Recognition and Machine Learning (Springer, New York). 1st Ed.

[29] Lockless SW, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. Science (80- ) 286(5438):295–299.

[30] Halabi N, Rivoire O, Leibler S, Ranganathan R (2009) Protein sectors: evolutionary units of three-dimensional structure. Cell 138(4):774–786.

[31] Narayanan C, Gagné D, Reynolds KA, Doucet N (2017) Conserved amino acid networks modulate discrete functional properties in an enzyme superfamily. Sci Rep 7(1). doi:10.1038/s41598-017-03298-4.

[32] Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CPI, Springer M, Sander C, Marks DS (2017) Mutation effects predicted from sequence co-variation. Nat Biotechnol 35(2). doi:10.1038/nbt.3769.

[33] Heinzelman P, Snow CD, Wu I, Nguyen C, Villalobos A, Govindarajan S, Minshull J, Arnold FH (2009) A family of thermostable fungal cellulases created by structure-guided recombination. Proc Natl Acad Sci U S A 106(14):5610–5615.

[34] Heinzelman P, Snow CD, Smith MA, Yu X, Kannan A, Boulware K, Villalobos A, Govindarajan S, Minshull J, Arnold FH (2009) SCHEMA recombination of a fungal cellulase uncovers a single mutation that contributes markedly to stability. J Biol Chem 284(39):26229–26233.

[35] Fox RJ, Davis SC, Mundorff EC, Newman LM, Gavrilovic V, Ma SK, Chung LM, Ching C, Tam S, Muley S, Grate J, Gruber J, Whitman JC, Sheldon RA, Huisman GW (2007) Improving catalytic function by ProSAR-driven enzyme evolution. Nat Biotechnol 25(3):338–344.

[36] Alvizo O, Nguyen LJ, Savile CK, Bresson JA, Lakhapatri SL, Solis EOP, Fox RJ, Broering JM, Benoit MR, Zimmerman SA, Novick SJ, Liang J, Lalonde JJ (2014) Directed evolution of an ultrastable carbonic anhydrase for highly efficient carbon capture from flue gas. Proc Natl Acad Sci 111(46):16436–16441.

[37] Auer P (2002) Using Confidence Bounds for Exploitation-Exploration Trade-offs. J Mach Learn Res 3(3):397–422.

[38] Lee ME, Aswani A, Han AS, Tomlin CJ, Dueber JE (2013) Expression-level optimization of a multi-enzyme pathway in the absence of a high-throughput assay. Nucleic Acids Res 41(22):10668–10678.

[39] AlQuraishi M, McAdams HH (2011) Direct inference of protein-DNA interactions using compressed sensing methods. Proc Natl Acad Sci 108(36):14819–14824.

[40] Doerge RW (2002) Mapping and analysis of quantitative trait loci in experimental populations. Nat Rev Genet 3(1):43–52.

[41] H. Ledford, “CRISPR, the disruptor,” *Nature*, vol. 522, no. 7554, pp. 20–24, Jun. 2015, doi: 10.1038/522020a.

[42] R. Barrangou and J. A. Doudna, “Applications of CRISPR technologies in research and beyond,” *Nat. Biotechnol.*, vol. 34, no. 9, pp. 933–941, Sep. 2016, doi: 10.1038/nbt.3659.

[43] S. Demirci, A. Leonard, J. J. Haro-Mora, N. Uchida, and J. F. Tisdale, “CRISPR/Cas9 for Sickle Cell Disease: Applications, Future Possibilities, and Challenges,” in *Advances in Experimental Medicine and Biology*, vol. 1144, Springer New York LLC, 2019, pp. 37–52.

[44] I. Strohkendl, F. A. Saifuddin, J. R. Rybarski, I. J. Finkelstein, and R. Russell, “Kinetic Basis for DNA Target Specificity of CRISPR-Cas12a,” *Mol. Cell*, vol. 71, no. 5, pp. 816-824.e3, Sep. 2018, doi: 10.1016/J.MOLCEL.2018.06.043.

[45] M. Wang, Y. Mao, Y. Lu, X. Tao, and J.-K. Zhu, “Multiplex Gene Editing in Rice Using the CRISPR-Cpf1 System.,” *Mol. Plant*, vol. 10, no. 7, pp. 1011–1013, Jul. 2017, doi: 10.1016/j.molp.2017.03.001.

[46] N. A. Pierce and E. Winfree, “Protein Design is NP-hard,” Oxford Academic, Oct. 2002. doi: 10.1093/PROTEIN/15.10.779.

[47] W. Mandecki, “The game of chess and searches in protein sequence space,” *Trends Biotechnol.*, vol. 16, no. 5, pp. 200–202, Dec. 1998, doi: 10.1016/S0167-7799(98)01188-3.

[48] K. K. Yang, Z. Wu, and F. H. Arnold, “Machine-learning-guided directed evolution for protein engineering,” *Nature Methods*, vol. 16, no. 8. Nature Publishing Group, pp. 687–694, Aug. 01, 2019, doi: 10.1038/s41592-019-0496-6.

[49] E. Farinas, “Fluorescence Activated Cell Sorting for Enzymatic Activity,” *Comb. Chem. High Throughput Screen.*, vol. 9, no. 4, pp. 321–328, Apr. 2006, doi: 10.2174/138620706776843200.

[50] R. T. Leenay *et al.*, “Identifying and visualizing functional PAM diversity across CRISPR-Cas systems,” *Mol. Cell*, vol. 62, no. 1, p. 137, Apr. 2016, doi: 10.1016/J.MOLCEL.2016.02.031.

[51] V. K. Mutalik *et al.*, “Precise and reliable gene expression via standard transcription and translation initiation elements,” *Nat. Methods*, vol. 10, no. 4, pp. 354–360, Apr. 2013, doi: 10.1038/nmeth.2404.

[52] X. Gao *et al.*, “Directed evolution and structural characterization of a simvastatin synthase.,” *Chem. Biol.*, vol. 16, no. 10, pp. 1064–74, Oct. 2009, doi: 10.1016/j.chembiol.2009.09.017.

[53] S. Stella, P. Alcón, and G. Montoya, “Structure of the Cpf1 endonuclease R-loop complex after target DNA cleavage,” *Nature*, vol. 546, no. 7659, pp. 559–563, Jun. 2017, doi: 10.1038/nature22398.

[54] C. Engler, R. Gruetzner, R. Kandzia, and S. Marillonnet, “Golden Gate Shuffling: A One-Pot DNA Shuffling Method Based on Type IIs Restriction Enzymes,” *PLoS One*, vol. 4, no. 5, p. e5553, May 2009, doi: 10.1371/journal.pone.0005553.

[55] M. Jain *et al.*, “Nanopore sequencing and assembly of a human genome with ultra-long reads,” *Nat. Biotechnol.*, vol. 36, no. 4, pp. 338–345, Jan. 2018, doi: 10.1038/nbt.4060.

[56] T. Yamano *et al.*, “Crystal Structure of Cpf1 in Complex with Guide RNA and Target DNA.,” *Cell*, vol. 165, no. 4, pp. 949–62, May 2016, doi: 10.1016/j.cell.2016.04.003.

[57] D. A. Drummond, J. J. Silberg, M. M. Meyer, C. O. Wilke, and F. H. Arnold, “On the conservative nature of intragenic recombination.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 15, pp. 5380–5, Apr. 2005, doi: 10.1073/pnas.0500729102.

[58] M. Landwehr, M. Carbone, C. R. Otey, Y. Li, and F. H. Arnold, “Diversification of Catalytic Function in a Synthetic Family of Chimeric Cytochrome P450s,” *Chem. Biol.*, vol. 14, no. 3, pp. 269–278, Mar. 2007, doi: 10.1016/j.chembiol.2007.01.009.

[59] Y. Li, D. A. Drummond, A. M. Sawayama, C. D. Snow, J. D. Bloom, and F. H. Arnold, “A diverse family of thermostable cytochrome P450s created by recombination of stabilizing fragments,” *Nat. Biotechnol.*, vol. 25, no. 9, pp. 1051–1056, Sep. 2007, doi: 10.1038/nbt1333.

[60] C. R. Otey, M. Landwehr, J. B. Endelman, K. Hiraga, J. D. Bloom, and F. H. Arnold, “Structure-guided recombination creates an artificial family of cytochromes P450.,” *PLoS Biol.*, vol. 4, no. 5, p. e112, May 2006, doi: 10.1371/journal.pbio.0040112.

[61] R. M. Liu *et al.*, “Synthetic chimeric nucleases function for efficient genome editing,” *Nat. Commun.*, vol. 10, no. 1, pp. 1–11, Dec. 2019, doi: 10.1038/s41467-019-13500-y.

[62] L. Gao *et al.*, “Engineered Cpf1 variants with altered PAM specificities.,” *Nat. Biotechnol.*, vol. 35, no. 8, pp. 789–792, Aug. 2017, doi: 10.1038/nbt.3900.

[63] M. A. Moreno-Mateos *et al.*, “CRISPR-Cpf1 mediates efficient homology-directed repair and temperature-controlled genome editing,” *Nat. Commun.*, vol. 8, no. 1, Dec. 2017, doi: 10.1038/s41467-017-01836-2.