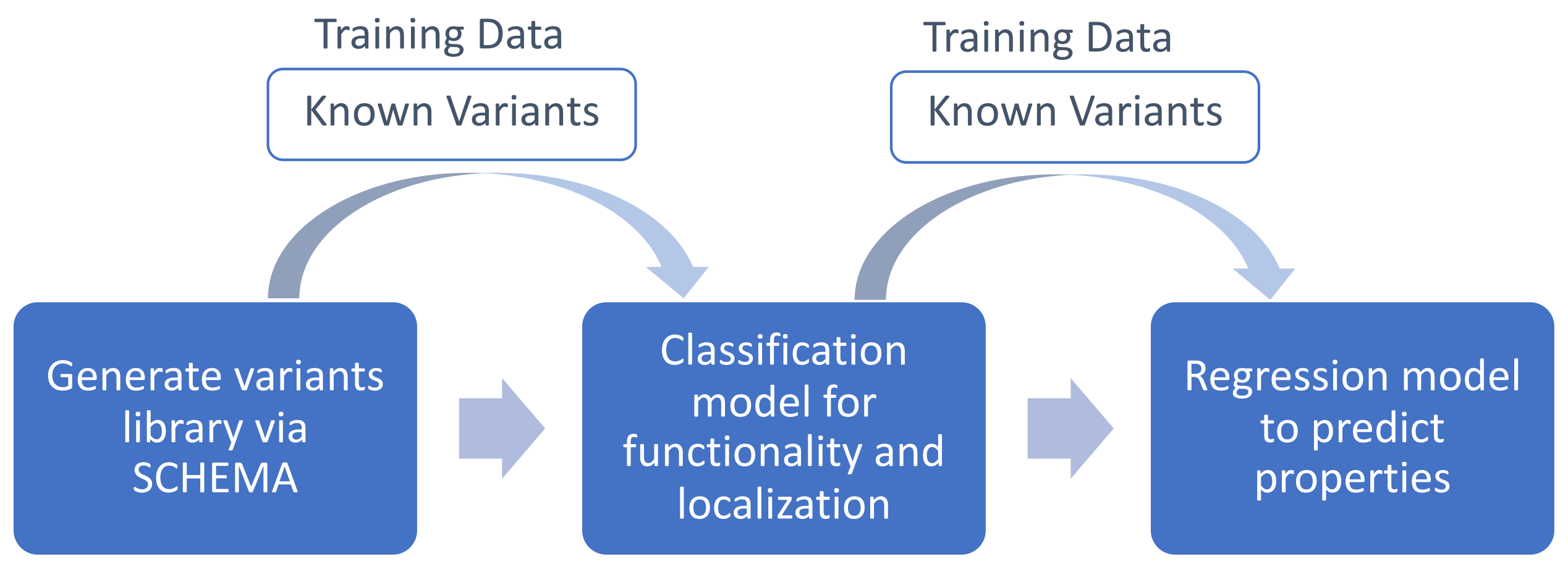
Machine Learning Guided Rhodopsin Engineering for Voltage Imaging

Mentee: Yijia Li (College of Arts and Sciences 2026, liyijia@sas.upenn.edu)
Mentor: Professor Lu Lu (Department of Chemical and Biomolecular Engineering)

Introduction

As the only currently known family of ion channels that are directly gated by light, Channelrhodopsins (ChRs) allows for direct fluorescence imaging of voltage changes, an attractive method for studying neuronal circuits and heart muscle, and a useful complement to traditional electrode-based methods or calcium imaging. In this research project, we aim to engineer high performing rhodopsin proteins with the tools of machine learning. Using Archaerhodopsin-3 as our parent protein, along with some existing variants including archon1, QuasAr3, and QuasAr6a , our goal is to generate a library of new variants and predict their functionality via classification and regression models.

Machine Learning Models



Training Data
Known Variants
Training Data
Known Variants
Generate variants library via SCHEMA
Classification model for functionality and localization
Regression model to predict properties

SCHEMA recombination shuffles sequence elements/blocks defined by a set of crossover locations in homologous proteins to generate novel chimeric proteins.

Classification Model
Use existing ChR variants to train binary classification models, then use the trained model to predict whether uncharacterized ChR sequence variants were functional. Threshold out non-functioning ChRs from the library based on predicted probability of functioning and localizing.

Regression Model
Use GP regression model to predict max-peak, voltage sensitivity, and kinetics

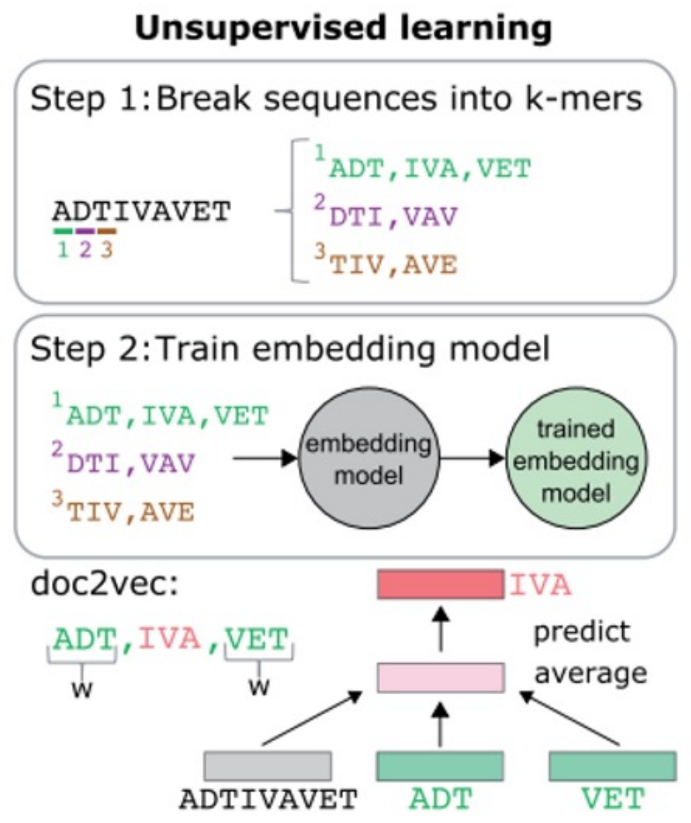
Model Input

The machine learning models infer predictive values for uncharacterized sequences from training examples by assuming that similar inputs (ChR variants) will have similar outputs (photocurrent properties). To quantify the relatedness ChR variants, we consider both sequence and structure. ChR sequence information is encoded in the amino acid sequence. For structural comparisons, we convert three-dimensional crystal-structural information into a 'contact map' that is convenient for modeling.

Doc2Vec Model for sequences

Often, one-hot encoding is used for input into ML models, but it is costly and a memory waste. We replace it with doc2vec model which reduces the input size of the sequence while generating the same results.

- *Each protein sequence is broken up into appropriate k-mer groups (1, 2, and 3).*
- *Each k-mer is like a word, and the embedding model will convert these "documents" into vector representations.*



Contact Map for 3D structure
Two residues are considered to be in contact and potentially important for structural and functional integrity if they have any nonhydrogen atoms within 4.5 Å.

We generate contact map for protein from its PDB (Protein Data Bank) file using 3D plotting and protein structure visualization libraries in Python.

1. Map three-letter amino acid codes to one-letter codes. This is useful for simplifying and standardizing residue names.
2. Parse a PDB file to extract the positions of atoms for each amino acid residue.
3. Calculate pairs of residues that are in contact, based on a specified contact distance

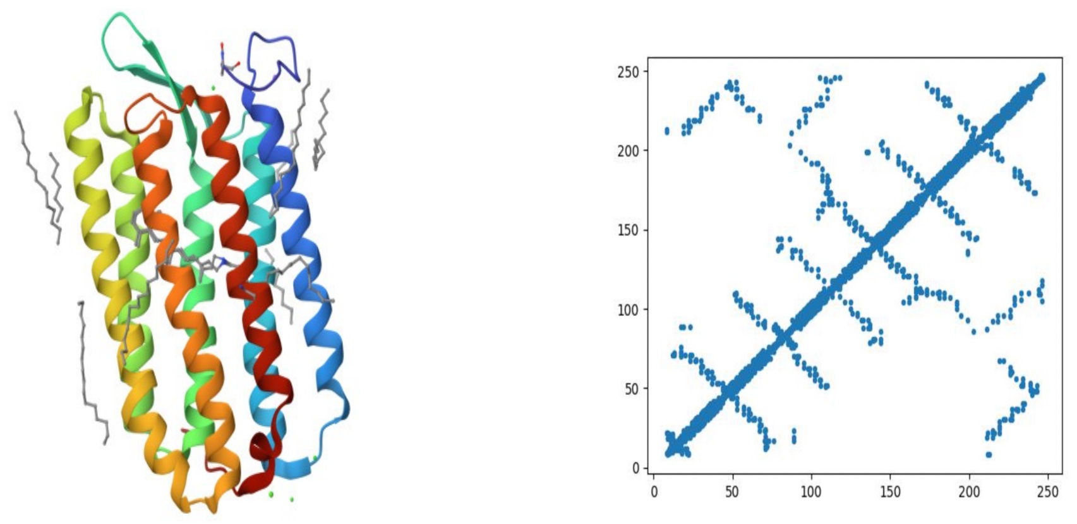


Figure: 3D Structure and Contact Map of 6GUY

Current Work

Using the ChR sequence and structure as well as functional data as inputs, our next step is to train GP classification and regression models with the following data that we have collected:

- 218 + 35 mKate (red) and GFP (green) florescence datapoints
- 75 red, green, and cyan photocurrent peak (pA) datapoints
- 154 red, green, and cyan peaks (nA), off-kinetics datapoints
- 249 datapoints on localization

Acknowledgment

I am working on this project under Lu Group in collaboration with Mitchell Daneker (Graduate Student Researcher for Lu Group), Wendy Deng (Undergraduate Student Researcher for Lu Group), Dr. Linlin Fan (Postdoctoral Fellow at Stanford University), and Dr. Lu Lu (Penn Institute for Computational Science). Special appreciation for the amazing work and support from my research group.

Reference

Bedbrook, C.N., Yang, K.K., Robinson, J.E. *et al.* Machine learning-guided channelrhodopsin engineering enables minimally invasive optogenetics. *Nature Methods* 16, 1176–1184 (2019). https://doi.org/10.1038/s41592-019-0583-8

Bedbrook, C.N., Rice, A.J., Yang, K.K. *et al*. Structure-guided SCHEMA recombination generates diverse chimeric channelrhodopsins. *National Academy of Sciences*, 2624-2633 (2017). https://doi.org/10.1073/pnas.1700269114

Yang, K.K., Wu, Z., Bedbrook, C.N., Arnold, F.H. *Bioinformatics*, 2642–2648
(2018). https://doi.org/10.1093/bioinformatics/bty178

Bedbrook, C.N., Yang, K.K., Rice, A.J., Gradinaru, V., & Arnold, F.H. Machine learning to design integral membrane channelrhodopsins for efficient eukaryotic expression and plasma membrane localization. *PLoS computational biology*, *13*(10), e1005786. https://doi.org/10.1371/journal.pcbi.1005786