Design a screening method for the *de novo* designed RSV immunogens based on machine learning

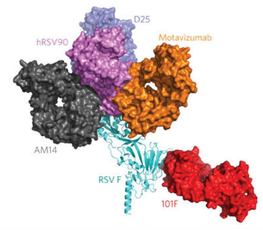
Shi-Lin Wang[1, 3], Luciano Abriata[3], Michele Ceriotti[1,2], Matteo Dal Peraro[3]

[1]School of Engineering, Materials Science and Engineering Department, École polytechnique fédérale de Lausanne (EPFL)
[2]Laboratory of Computational Science and Modeling. Materials Science and Engineering Department, École polytechnique fédérale de Lausanne (EPFL)
[3]Laboratory of Biomolecular Modelling, Life Science Department, École polytechnique fédérale de Lausanne (EPFL)

※ Data (sequence and structure) used in this project is obtained with the help from Prof. Bruno Correia and Dr. Yang Che et al.

Introduction & Motivation

Respiratory syncytial virus (RSV) is the primary cause for the lower respiratory infection. To design the vaccine, template-free de novo methods is powerful to deal with structurally complex epitope. However, not all designed sequences are necessarily good candidates. Consequently, another preliminary screening process must be performed. Normally, the screening process requires intensive experiments (e.g., Next-Generation Sequencing experiment) [1-2]. This motivates one to design a computational screening process. Thus, this project aims to search for a set of feature and model to approximate a screening function that can determine the quality of a de novo designed sequence.

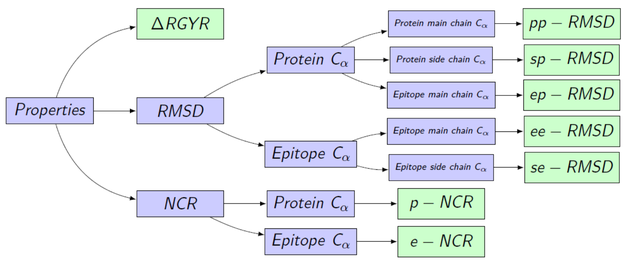


Trajectory Method

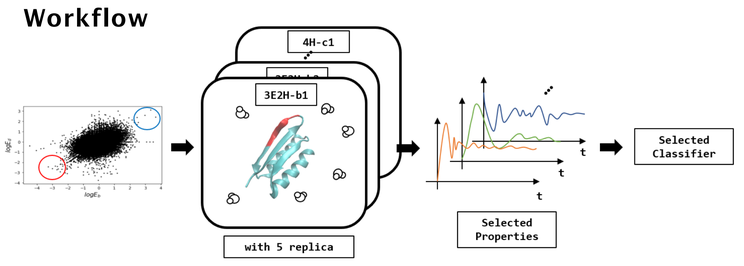
Methodology

The goal of the trajectory method is to fit the classifier using shape-related ensemble averaged properties obtained from simulation trajectories. Specifically, molecular dynamics (MD) is applied to simulate a de novo designed sequence under physiological conditions. A small number of de novo designed sequences were first selected, then corresponding structures were generated and simulated by MD. Afterward, shape-related ensemble averaged properties were calculated based on MD trajectory. Finally, the classifiers (Random forest, support vector machine (SVM) and logistic regression classifiers) were trained by these properties.

Target Antibody	101F			D25
Design	3E2H	4E1H	4E2H	4H
# Best (g)	4	3	4	3
# Worst (b)	4	3	2	3
# Experiment (c)	0	1	0	1

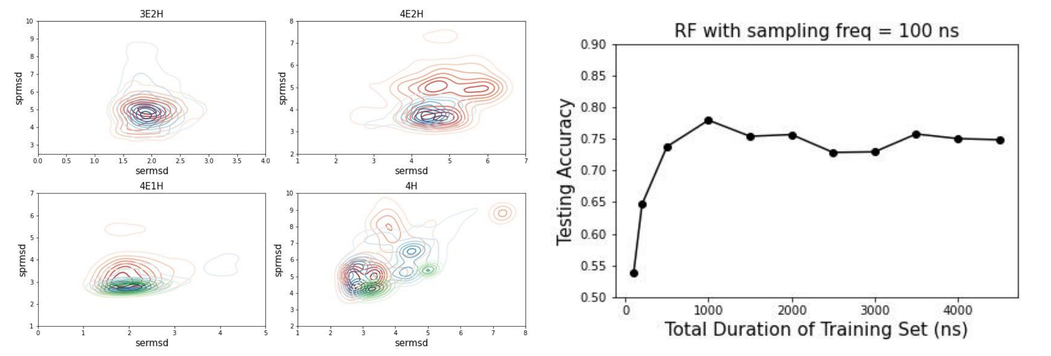


Workflow



Replica-1	Replica-2	Replica-3	Replica-4	Replica-5
3E2H-g-1	3E2H-g-1	3E2H-g-1	3E2H-g-1	3E2H-g-1
3E2H-g-2	3E2H-g-2	3E2H-g-2	3E2H-g-2	3E2H-g-2
⋮	⋮	⋮	⋮	⋮
4E2H-b-2	4E2H-b-2	4E2H-b-2	4E2H-b-2	4E2H-b-2
3E2H-b-1	3E2H-b-1	3E2H-b-1	3E2H-b-1	3E2H-b-1

Results

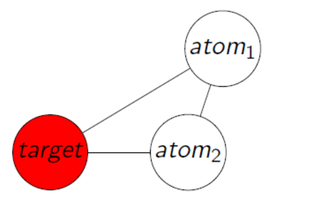


Structure Method

Methodology

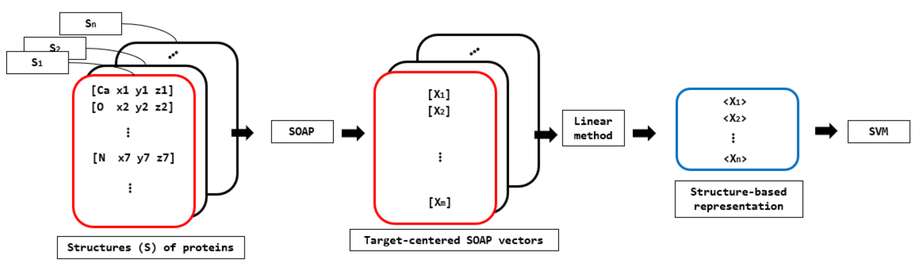
Since shape-related ensemble averaged properties are insufficient to make a satisfying prediction, additional factors should be considered. Therefore, the structure method used SOAP vectors as features. First, all structures (in pdb file format) were originally in Cartesian coordinates. These structures were transformed into SOAP vectors. Subsequently, a linear method was applied to these vectors to obtain a structure-based representation. Finally, these structure-based representations were used as features to train an SVM.

Design	3E2H	4E2H
Length	65/68	84
# Structures	1121/3820	1952
# Good-Bad	1794-3756	910-1042



$$\rho_{\mathcal{X}_i} = \sum_{\alpha n \tilde{\alpha} \tilde{n} l} \langle \alpha n \tilde{\alpha} \tilde{n} l | \mathcal{X}_i \rangle$$

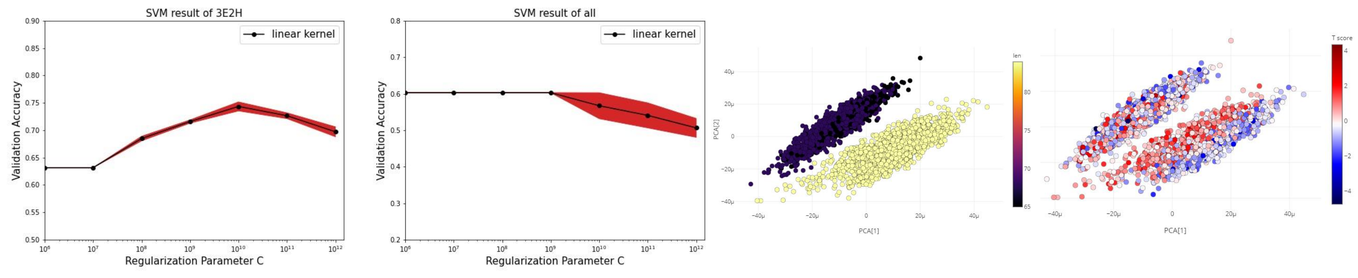
Workflow



Case 1 - train on all	
Train set	Test set
80% of (3E2H and 4E2H)	20% of (3E2H and 4E2H)
Good - Bad 3252 - 2262	Good - Bad 932 - 447

Case 2 - train on 3E2H		
Train set	Internal test set	External test set
80% of 3E2H	20% of 3E2H	All 4E2H
Good - Bad 2496 - 1456	Good - Bad 646 - 343	Good - Bad 1042 - 910

Results



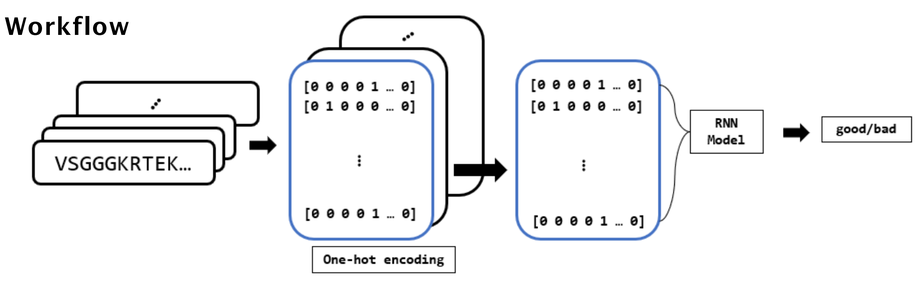
Sequence Method

Methodology

The structure method was only valid when all structures have a single design. In addition, it is itself computationally costly. To find a better model, proteins' sequences were adopted as feature for the sequence method. First, all the sequences were one-hot encoded as input. Next, according to a cross-entropy loss function, the RNN model was trained. Finally, the best model was obtained based on a grid search. The predicted class of given sequences was chosen to be the class that had a larger probability.

Design	#Sequence	Target Ab	Length-#Sequences	Good-Bad
4E1H	82798	101F	62/63/64/65 – 11741/15812/38930/16315	34662 – 48136
3E2H	5005	101F	65/68 – 1121/3884	1826 – 3179
4E2H	1952	101F	84 – 1952	910 – 1042
4H	282	D25	85 – 282	104 – 158

Workflow

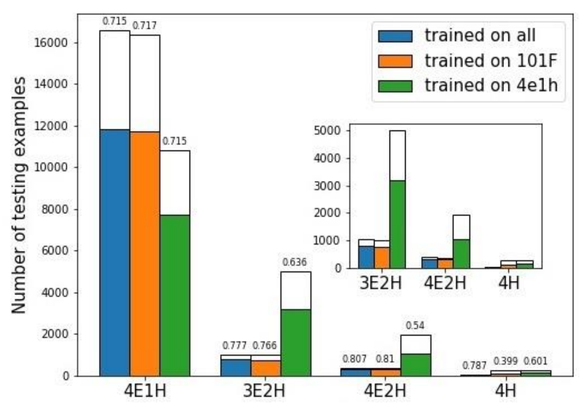


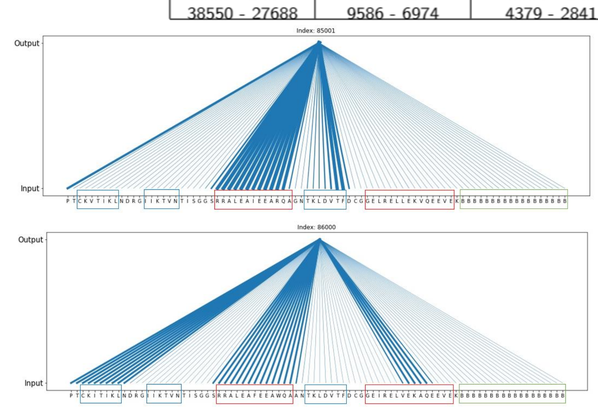
Case 1 - train on all	
Train set	Test set
80% of all	20% of all
Good - Bad 42062 - 29952	Good - Bad 10453 - 7551

Case 2 - train on 101F		
Train set	Internal test set	External test set
80% of 101F	20% of 101F	All D25
Good - Bad 41932 - 29872	Good - Bad 10425 - 7526	Good - Bad 105 - 158

Case 3 - train on 4E1H		
Train set	Internal test set	External test set
80% of 4E1H	20% of 4E1H	All other
Good - Bad 38550 - 27688	Good - Bad 9586 - 6974	Good - Bad 4379 - 2841

Results





Conclusion

The trajectory method failed to classify any given sequence. Although the structure method could be useful to some extent, such method was still unrealistic. Finally, the sequence method was able to process 18 times more sequences and twice faster then the structure method. Moreover, the testing accuracy was 0.72 when several designs were included. The classification results showed that the quality of the designed sequence was determined by the sequence's structure regardless of epitope mimicking domain. This disproved previous assumption that sequences with the same epitope-mimicking domain are in the same distribution to the model. Nevertheless, the sequence method could still be helpful because the motivation of this project is to obtain good sequences through computation.

[1] F. Sesterhenn, C. Yang, et al. "De novo protein design enables precise induction of functional antibodies in vivo". In: bioRxiv (2020), p. 685867.

[2] F. Sesterhenn, M. Galloux, et al. "Boosting subdominant neutralizing antibody responses with a computationally designed epitope-focused immunogen". In: PLoS biology 17.2 (2019)