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

Shi-Lin Wang

École Polytechnique Fédérale de Lausanne (EPFL)

School of Engineering, Material Science and Engineering Department

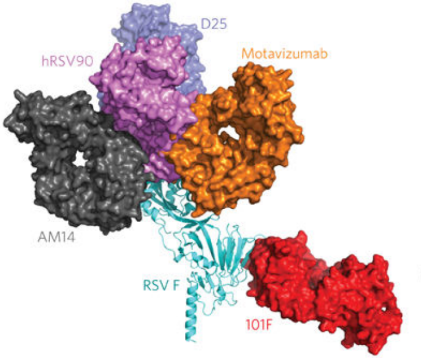
Feb. 2021

Advisor: Prof. Matteo Dal Peraro, Prof. Michele Ceriotti, Dr. Luciano Abriata

Outlines

Introduction

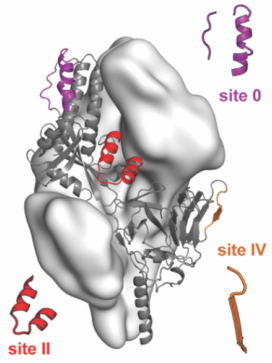
Background: respiratory syncytial virus (RSV)

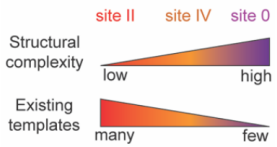


1. A primary cause of lower respiratory infections.
2. Treatments:
   - Vaccine (?)
   - **Antibody** (e.g. D25, 101F, ... etc.)
3. **Epitope**: distinct antigenic sites that, are part of RSV's glycoproteins sequences that specific antibodies target.
4. Therefore, the specificity between antibodies and their target epitopes sheds light on the design of an RSV vaccine.

Background: computational protein design

Goal: design an immunogen that accommodates certain epitope of RSV (i.e. epitope-mimicking immunogens).



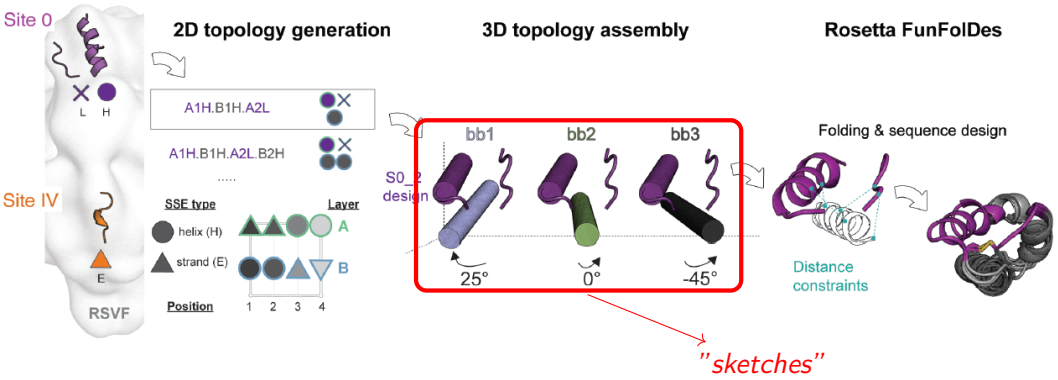


Template-based de novo method:

- Exhaustively search for structures that match the target epitope in a database.
- Limitations:
  1. Availability of existing templates.
  2. Optimization through in-vitro evolution.

Background: template-free de novo method

- Mimics the structure of epitopes by directly matching the epitopes' structural motifs.



EPFL

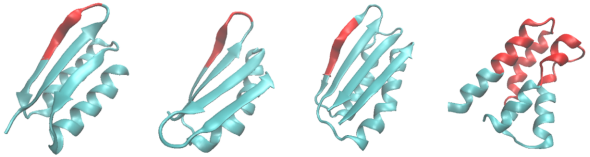
Motivation & Objective

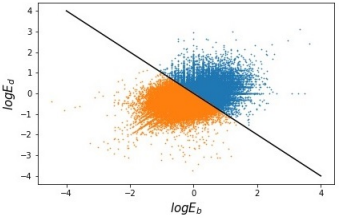
- After applying the de novo method, one obtains several best sketches, and each sketch comes with several immunogens designed by *Rosetta FunFolDes*.
- However, not all immunogens are necessarily "good" candidates. As a consequence, another preliminary **screening process** should be performed in advance of any actual implementation.
- Normally, the screening process (based on NGS[1]) requires intensive experiments. This motivates one to design **a computational screening process**.
- Thus, this project aims to search for **feature and model** to approximate a screening function that can determines the quality of a given sequence.

EPFL

Data

1. **Sequence data**[2]: selected sketches (3E2H/4E1H/4E2H/4H) → mutations.
2. **Structure data**[3]: selected sketches (3E2H/4E2H) → *Rosetta*.
3. **Label**: without subjective criteria, the screening problem is arbitrarily simplified to a binary classification problem.





Generic threshold function for classification

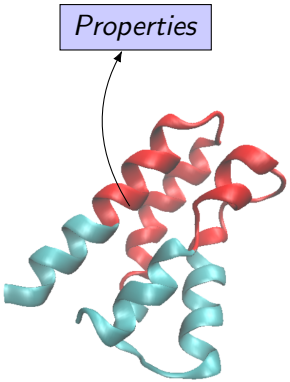
$$T(E_d, E_b) = \begin{cases} 0 & \text{if } log(E_d) + log(E_b) > 0 \\ 1 & \text{if } log(E_d) + log(E_b) \leq 0 \end{cases} \quad (1)$$

Trajectory Method

EPFL

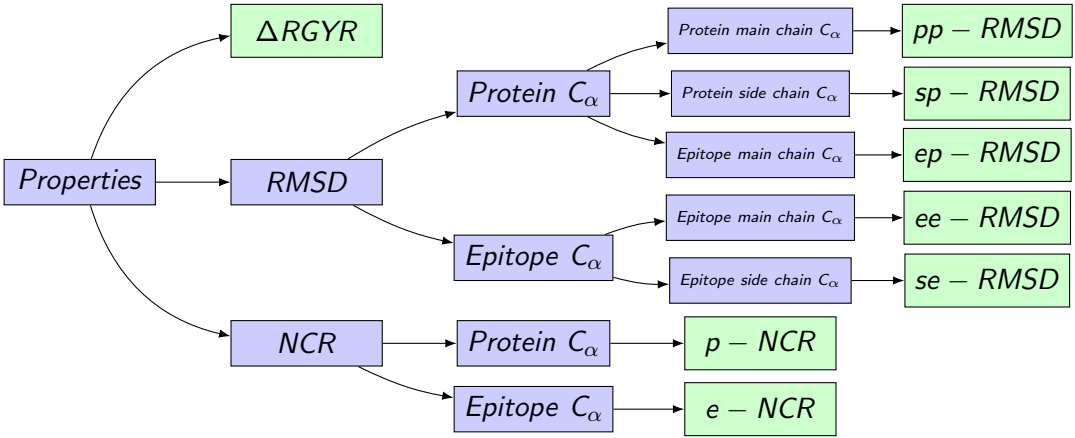
Introduction: prerequisite

1. The quality of a de novo designed sequence is determined by E_b and E_d.
2. Naive approach: simulate the entire process (i.e. the binding process between the antibody and designed sequence) $\rightarrow$ Intractable.
3. Recall that de novo designed sequences have
   - specific domains that mimic the epitope and,
   - the overall structure that is designed to support epitope-mimicking domain.

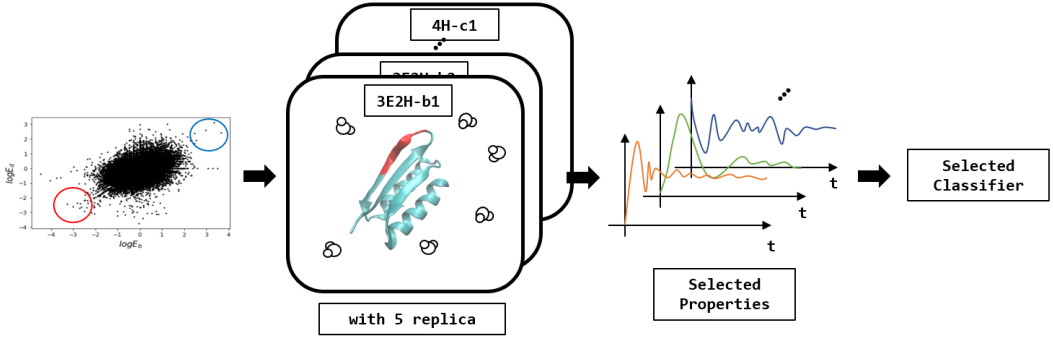


Assumption: There exists a set of **shape-related ensemble averaged properties** that can determine a sequence's quality.

Introduction: selected properties[4]



Methodology: workflow[5]

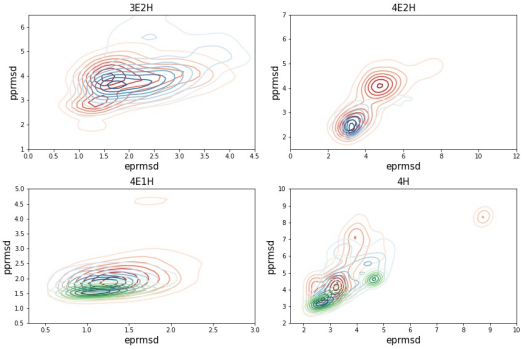


- Selected classifiers: logistic regression classifier (LRC), random forest classifier (RFC), and support vector machine (SVM).

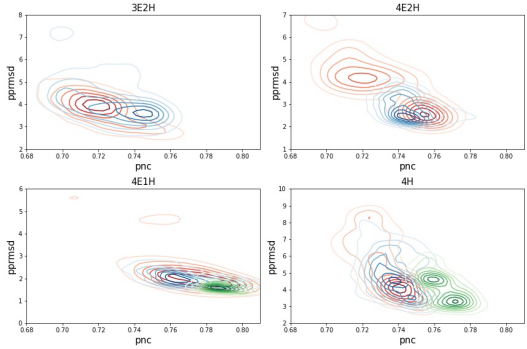
Results: MD simulation (1)

Best, Worst, and Experiment.

ep − RMSD vs. *pp − RMSD*

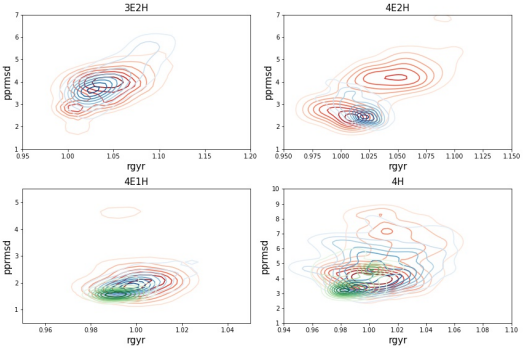


p − NRC vs. *pp − RMSD*

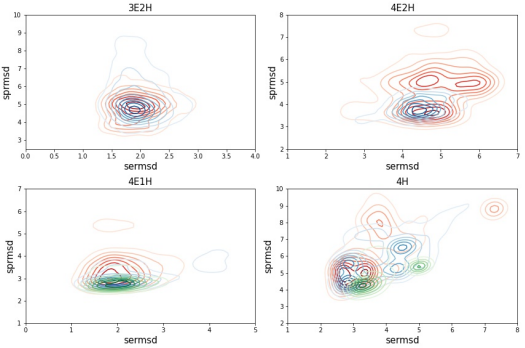


Results: MD simulation (2)

$\Delta RGYR$ vs. $pp - RMSD$

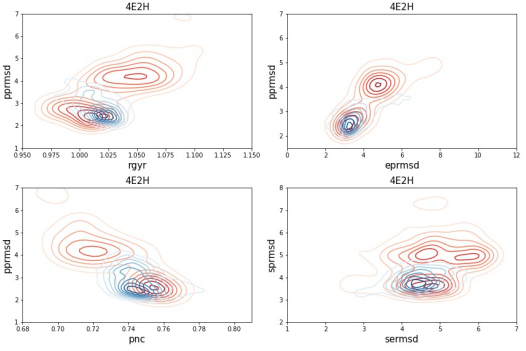


$se - RMSD$ vs. $sp - RMSD$

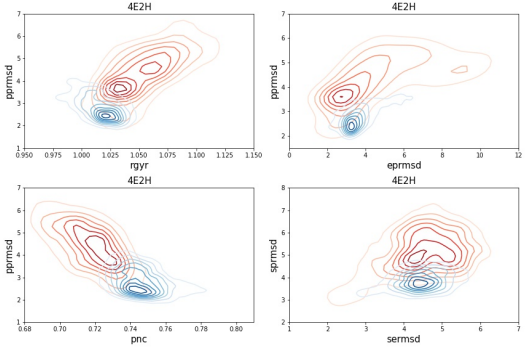


Results: MD simulation[8] (3)

original 4E2H (worst-1)



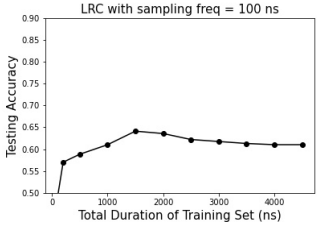
mutated 4E2H (worst-1)



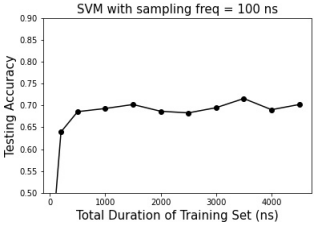
Results: Classification[9]

- External test set: 3E2H-worst-1
- Train & test sets: All but 3E2H-worst-1

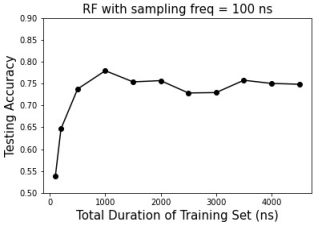
LRC



SVM



RFC



- However, RFC only reached an external testing accuracy of **0.39 (86/217)**.

EPFL

Discussion

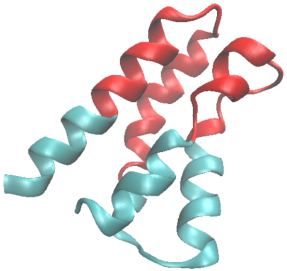
1. Good sequences had more confined distributions when compared with the bad structures.
2. There is a gap between theoretical and actual structures even though their structures are similar.
3. Resulting distribution changed significantly after a single shift of amino acid was made.
4. The trajectory method failed to classify any unseen sequence.

Structure Method

EPFL

Introduction: prerequisite

1. Although some interesting results were obtained from MD simulation, the trajectory method failed to predict the quality of given sequence.
2. Apparently, shape-related ensemble averaged properties are insufficient and additional factors should be considered.
3. Recall that the ensemble averaged properties (E_b and E_d) are presumably determined by the sequence's structure if the environment remains the same.

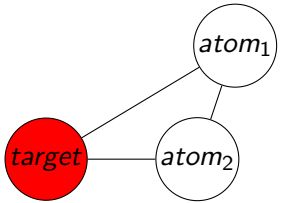


Assumption: The structure of a given sequence already carries enough information to determine the quality of that sequence.

EPFL

Introduction: structural descriptor SOAP

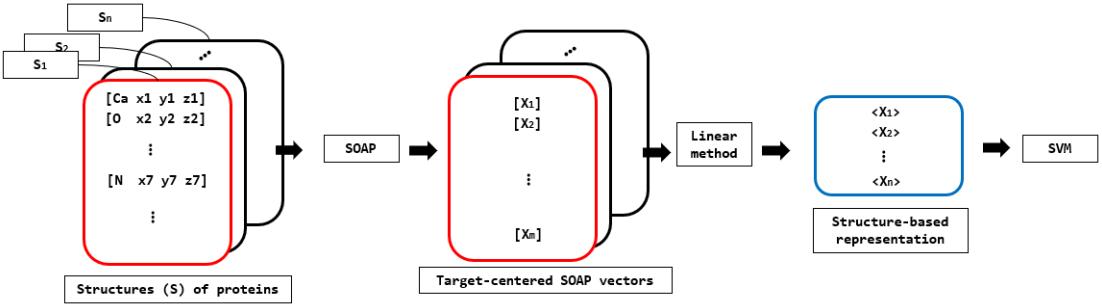
- The SOAP representation is a complete, translation, permutation, and rotation invariant descriptor.
- It is similar to a three-body correlation function. A SOAP vector builds a description of a **hypothetical triplet** of atoms.
- If the hypothetical triplet matches a real triplet in the structure, the SOAP feature measures the **strength** of this match.



The SOAP representation[10]

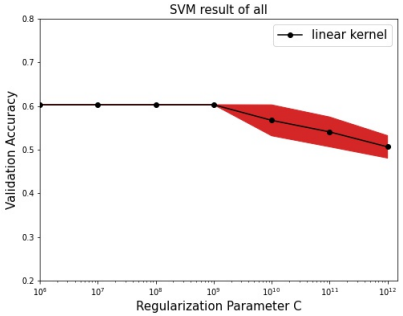
$$\rho_{\mathcal{X}_i} = \sum_{\alpha n \tilde{\alpha} \tilde{n} l} \langle \alpha n \tilde{\alpha} \tilde{n} l | \mathcal{X}_i \rangle \qquad (2)$$

Methodology: workflow[12]

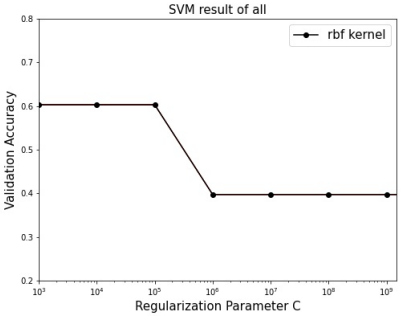


Results: Case.1- train on all (grid search)[14]

with linear kernel



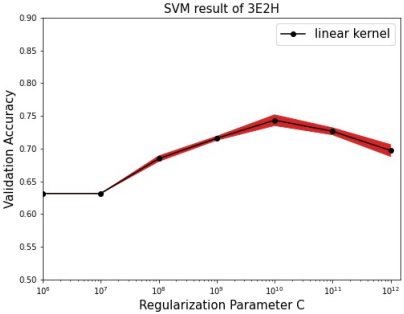
with rbf kernel



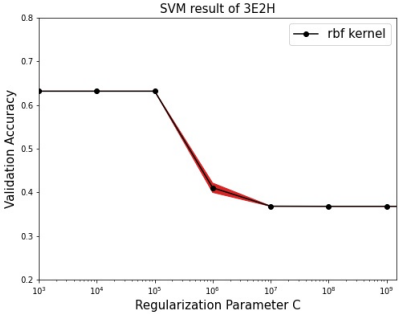
The best model reached a **testing accuracy of 0.62**.

Results: Case.2- train on 3E2H (grid search)[14]

with linear kernel



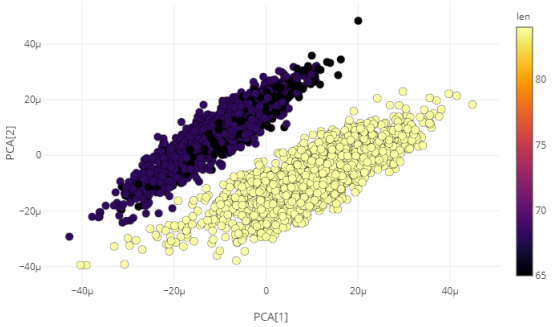
with rbf kernel



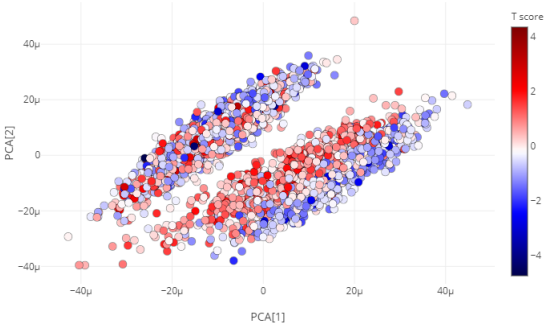
When only 3E2H structures constituted the train set, the best model reached an **internal testing accuracy of 0.77**. However, the **external testing accuracy was only 0.53**. The same conclusion holds for model trained on 4E2H.

Results: PCA analysis

colored by sequence length



colored by $logE_b + logE_d$



EPFL

Discussion

1. The structure method is capable of classifying designed sequences for the same sketch, it performs poorly when multiple sketches exist simultaneously.
2. The failure was because structures with different sketches belong to different distributions. This was confirmed by PCA result.
3. Although the structure method is powerful, it is computationally expensive.
   - Sequence transformation
   - SOAP transformation
   - Structure-based representation

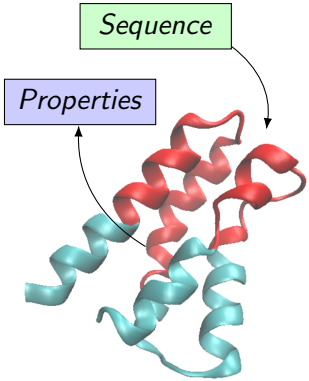
Sequence Method

EPFL

Introduction: prerequisite

1. Although the structure method successfully classified structures with a testing accuracy of 0.7, it was only valid for a data set limited to the same sketch. In addition, it is computationally costly.

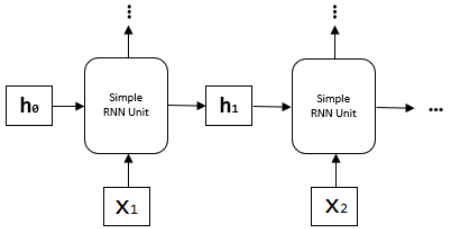
2. All proteins' properties should originate from their structures, that are presumably determined by their corresponding sequences given the same environment.



Assumption: There exists an underlying rule governing the combination of amino acids. Such a rule dictates the quality of de novo designed sequences.

Introduction: recurrent neural network (RNN)

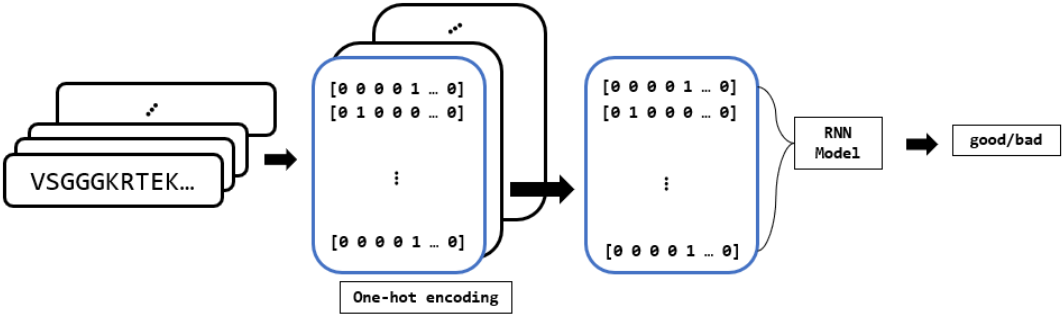
1. To model the sequential data, RNN **recursively** takes the **previous hidden state** as an additional input. However, traditional RNN suffes from vanishing gradients or exploding gradients[17].
2. To solve this issue, long short-term memory unit (LSTM)[18] was tested. Other improvements were also tested such as bidirectional RNN and attention mechanism[19].



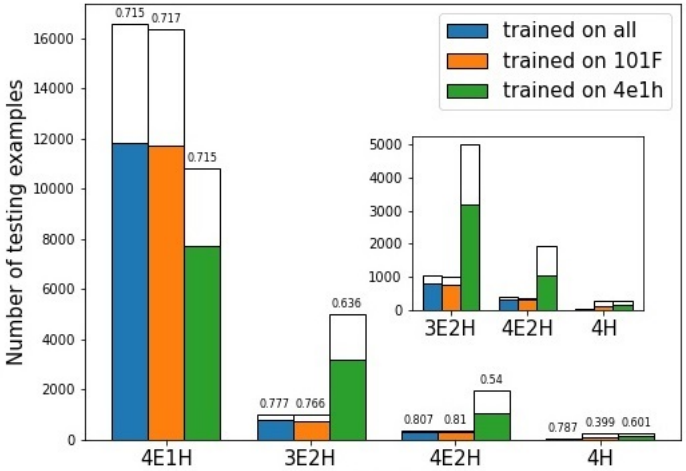
RNN's feedforward pass

$$h^{(t)} = \sigma(s^t) = \sigma\left(W^{hx}x^{(t)} + W^{hh}h^{(t-1)} + b_h\right) \quad (3)$$

Methodology: workflow[20]

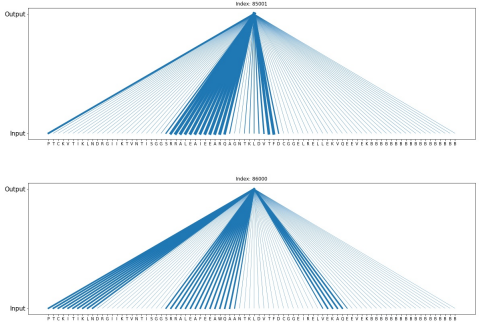


Results: classification[22]

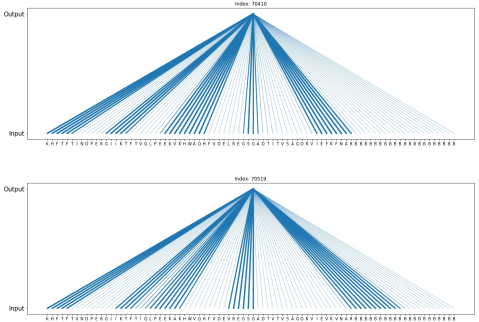


Results: attention weights (1)

3E2H

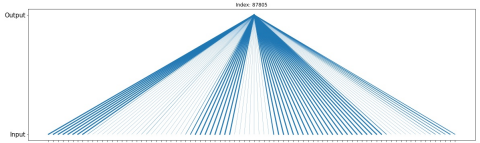


4E1H

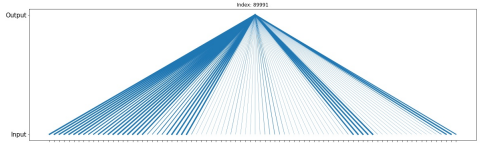


Results: attention weights (2)

4E2H

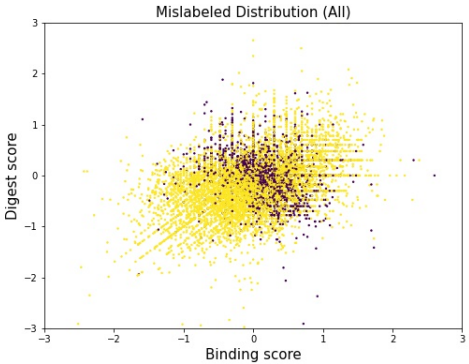


4H

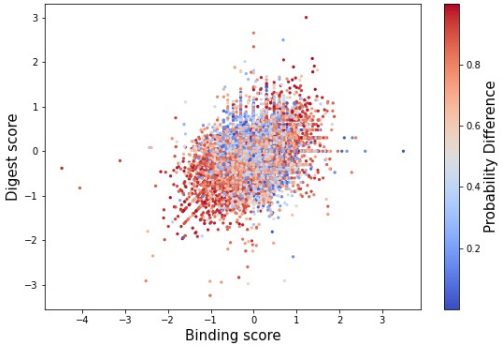


Results: error analysis (train on all)

mis-labeled



absolute probabilities difference



EPFL

Discussion

1. Sequence method was able to process more sequences and faster then the structure method. Moreover, the testing accuracy was 0.72 when several sketches were included.
2. Sequences with different sketches are in the same distribution regardless of the same epitope mimicking domain. This can be indirectly visualized by their specific patterns of attention weights.
3. The error was introduced when the screening problem was simplified into a binary classification problem.
4. Nevertheless, the sequence method could still be helpful because the motivation of this project is simply to obtain good sequences through computation.

Thank you!

Appendix

Appendix: NGS screening process

back

Definition of enrichment score

$$E_{\text{pressure}} = \frac{\#\text{Sequence Count(high selective pressure)}}{\#\text{Sequence Count(low selective pressure)}} \tag{4}$$

The screening process is based on two enrichment scores obtained from next generation sequencing (NGS).
Specifically,

1. binding score (E_b) represents the enrichment score subjected to the target antibody and,
2. digest score (E_d) represents the enrichment score subjected to chymotrypsin protease.

Generally, a **higher** enrichment value corresponds to a **higher** affinity (or resistance) to a selective pressure.

EPFL

Appendix: Sequence data

back

1. At the final step of the template-free de novo method, *Rosetta FunFolDes* generates 1000 to 10000 (parent) sequences.
2. From the subset of sequences in the selected sketches, several new sequences are made based on mutations sequences based on the selected sketch.
3. These mutations locate 8 to 14 different critical hydrophobic core positions and surface positions close to the binding interface.

Sketch	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

Appendix: Structure data

back

1. On top of the sequence data, *Rosetta FunFolDes* builds structures in pdb file format according to their unmutated sequences.
2. Unconstrained *Rosetta FastRelax* optimizes output structures.

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

EPFL

Appendix: Selected properties

back

Definition of *RGYR*

$$\Delta R_g = \frac{R_g(t) - R_g(0)}{R_g(0)} \tag{5}$$

Definition of *NC*

$$NC^j = \sum_{i \notin n_j}^{N} \Phi(|r_{ij} - r_0|) \tag{6}$$

Definition of *NCR*

$$NCR(t) = \frac{\sum_{i=1} NC^i(t)}{\sum_{i=1} NC^i(t_0)} \tag{7}$$

EPFL

Appendix: data & supervised classification

back

1. Although, in principle, one can simply take all of the structures as inputs to run a simulation, this method would take too much time and resource. To practically apply the trajectory method, several **"best", "worst", and experimental** structures[6] were selected from sequence data as inputs for the MD simulation[7].

2. The mean of each selected properties was calculated every 100 *ns* along each trajectory as data point.

3. Due to the amount of data, only classifiers that required low amount of data were considered: Logistic Regression Classifier (LRC), Random Forest Classifier (RFC), and support vector machine (SVM).

Appendix: data for trj-method

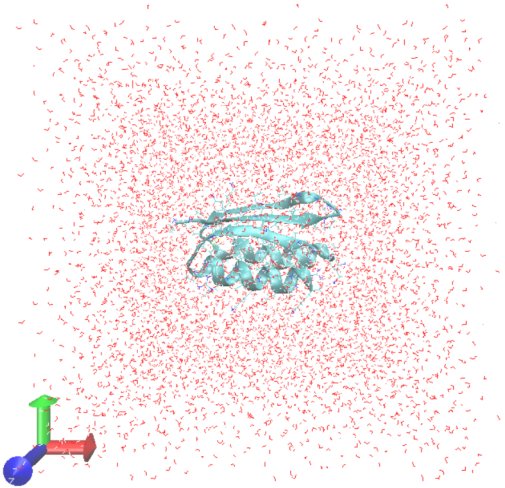
back

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

PS. two X-ray structures of good proteins obtained from the X-ray experiment were also included and named as "experiment".

Appendix: MD simulation detail

back



- *GROMACS version of 2019.2-mpi*.
- NPT emsemble: 303.15K and 1 *atm*.
- Box: the dimension of the protein $\pm$ 1.2 *nm*.
- Solvated by water.
- Neutralized by KCl.

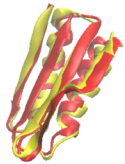
Appendix: 4E2H mutation

back

- Original 4E2H (worst): yellow
- Mutated 4E2H (worst): red
- In the sequence alignment, the yellow highlight indicates the mutation sites.

```
4E2H_b1_o    PNTISCTAPTRDEVEQCVQEQWKELVKKLDGSKKITFSVNAGGERRGIIKTSSGLPEEDFQALEKAVKEGASGF-VTVTVSVK   82
4E2H_b2_o    PNTISCTAPTRDEVAQCFQEHWKELVKKLDGS-KVTFSANVPKLSRGIIKTSDGNIEEVLQALAKGLKEGASGFVVTVTISVK   82
4E2H_b3_o    PNTFSCTAPTRDELEQCLQEHWKELAKKLDGSKKITFSVNAGGERRGIIKTSSGLPEEDFQALEKALKEGASGF-VTITVSVK   82
4E2H_b4_o    PNTISCTAPTRDELEQCVQEQWKELVKKLDGSKKVTFSVNAGGERRGIIKTSSGLPEEDFQALEKAVKEGASGF-VTVTVSVK   82

4E2H_b1_m    PNTISCTAPTRDELAQCVQEQWKELVKKLDGSKKITFSVNAGGERGIIKTTSSGLPEEDFQALEKALKEGASGF-VTVTVSVK   82
4E2H_b2_m    PNTFSCTAPTRDEVAQCVQEHWKELAKKLDGSKKITFSVNAGGERGIIKTTSSGLPEEDFQALEKALKEGASGF-VTVTVSIK   82
4E2H_b4_m    PNTFSCTAPTRDELAQCVQEHWKELVKKLDGSKKITFSVNAGGERGIIKTTSSGLPEEDFQALEKALKEGASGF-VTVTVSIK   82
4E2H_b3_m    PNTISCTAPTRDELEQCLQEHWKEWAKKLDGSKKVTFSVNAGGERGIIKTTSSGLPEEDFQALEKGVKEGASGF-VTVTVSVK   82
               eee    hhhhhhhhhhhhhhhhh     eeeeee     eeeeee      hhhhhhhhhhhh    e  eeeeeee
```



Appendix: date set for supervised classification

back

1. All data points from the 3E2H (worst-1) were extracted as an **external test set**.
2. All data points from the fifth trajectory were extracted as an **internal test set**.
3. Remaining data points were used as a training set.
4. Data was shuffled before splitting.
5. The classifier was optimized by 4-fold cross-validation.

EPFL

Appendix: SOAP transformation

In the formulation of SOAP, the chemical environment was originally represented by the atomic probability density function, which is composed of Gaussian functions.

Projecting on radial and spherical harmonic basis

$$\rho^{\alpha}_{\mathcal{X}_j}(\vec{r}) = \sum_{i\in\alpha} \Phi(\vec{r}_{ij})\; e^{-\frac{\left|\vec{r}-\vec{r}_{ij}\right|^2}{2\sigma^2}} = \langle \alpha\vec{r}|\mathcal{X}_j\rangle \tag{8}$$

$$= \int d\vec{r}\, \mathcal{R}_n \mathcal{Y}^{\ell}_m \langle \alpha\vec{r}|\mathcal{X}_j\rangle = \langle \alpha nlm|\mathcal{X}_j\rangle \tag{9}$$

EPFL

Appendix: SOAP transformation

back

To make the representation rotation invariant, two probability density functions' overlap are integrated over all possible rotations, which yields the similarity kernel.

Similarity kernel

$$S^{(\nu)}\left(\rho^{\alpha}_{\mathcal{X}_i}, \rho^{\widetilde{\alpha}}_{\mathcal{X}_j}\right) = \int d\widehat{R} \left\langle \rho^{\alpha}_{\mathcal{X}_i} \middle| \widehat{R} \middle| \rho^{\widetilde{\alpha}}_{\mathcal{X}_j} \right\rangle^{\nu} \rightarrow \sum_{\alpha n \widetilde{\alpha} \widetilde{n} l} \langle \mathcal{X}_i | \alpha n \widetilde{\alpha} \widetilde{n} l \rangle \langle \alpha n \widetilde{\alpha} \widetilde{n} l | \mathcal{X}_j \rangle \qquad (10)$$

EPFL

Appendix: data & supervised classification

back

1. All structures (3E2H and 4E2H) were first built and optimized by *Rosetta*.
2. All structures (in pdb file format) are transformed into SOAP representations. To reduce memory usage, only C_α were selected as targets.
3. The structure method concerns, not individual targets, but the overall structure. Thus, the target-centered SOAP vectors were further transformed into the structure-based representations by **linear method**[13].
4. Due to the inherent characteristics of the formulation of a SOAP vector, only the SVM was selected as the classifier. Two kernel methods are applied (linear and rbf).

EPFL

Appendix: structure-based representation

back

Assume that the local information is additive.

Structure-based representation (f) based on linear method

$$f = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix} = \begin{bmatrix} \langle X_1 \rangle \\ \langle X_2 \rangle \\ \vdots \\ \langle X_n \rangle \end{bmatrix} \tag{11}$$

where

$$\langle X_i \rangle = \frac{1}{m_i} \sum_{j \in S_i}^{m_i} X_i^{(j)} \tag{12}$$

Appendix: data set for str-method

Case 1 - train on all back	
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 back		
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

- Data was shuffled before splitting.
- The classifier was optimized by 5-fold cross-validation.

Appendix: data set for str-method

back

Case 3 - train on 4E2H	
Train set	test set
80% of 4E2H	20% of 4E2H
Good - Bad 833 - 728	Good - Bad 209 - 182

- Data was shuffled before splitting.
- The classifier was optimized by 5-fold cross-validation.

Appendix: structure-based representation

back

Structure-based representation (f) based on rbf kernel method

$$f = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix} = \begin{pmatrix} K(S_1, S_1) & K(S_1, S_2) & \cdots & K(S_1, S_n) \\ K(S_2, S_1) & K(S_2, S_2) & \cdots & K(S_2, S_n) \\ \vdots & \vdots & \ddots & \vdots \\ K(S_n, S_1) & K(S_n, S_2) & \cdots & K(S_n, S_n) \end{pmatrix} \tag{13}$$

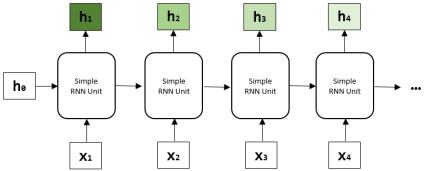
where

$$K(S_a, S_b) = \sum_{j \in S_a} \sum_{k \in S_b} k(X_j, X_k) = k(X_a^{(1)}, X_b^{(1)}) + \cdots + k(X_a^{(m)}, X_b^{(m)}) \tag{14}$$

EPFL

Appendix: vanishing and exploding gradients

back



Back-propagation of RNN

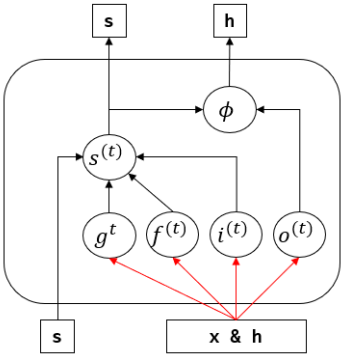
$$\frac{\partial \mathcal{L}^t}{\partial W^{hh}} = \frac{\partial \mathcal{L}^t}{\partial h^t} \times \frac{\partial h^t}{\partial h^{t-1}} \times \frac{\partial h^{t-1}}{\partial h^{t-2}} \times \cdots \times \frac{\partial h^1}{\partial W^{hh}} \tag{15}$$

$$= \frac{\partial \mathcal{L}^t}{\partial h^t} \times \left(\frac{\partial h^t}{\partial s^t} \times \frac{\partial s^t}{\partial h^{t-1}} \right) \times \left(\frac{\partial h^{t-1}}{\partial s^{t-1}} \times \frac{\partial s^{t-1}}{\partial h^{t-2}} \right) \times \cdots \times \left(\frac{\partial h^1}{\partial s^1} \times \frac{\partial s^1}{\partial W^{hh}} \right) \tag{16}$$

$$= \frac{\partial \mathcal{L}^t}{\partial h^t} \times \left(\sigma'(s^t) \times W^{hh} \right) \times \left(\sigma'(s^{t-1}) \times W^{hh} \right) \times \cdots \times \left(\sigma'(s^1) \times h^0 \right) \tag{17}$$

Appendix: LSTM

back



LSTM back-propagation

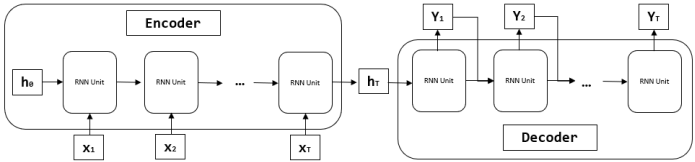
$$\frac{\partial s^{(t)}}{\partial p} = \frac{\partial}{\partial p}\left(g^{(t)} \odot i^{(t)} + s^{(t-1)} \odot f^{(t)}\right) \quad (18)$$

$$= \frac{\partial}{\partial p}\left(g^{(t)} \odot i^{(t)}\right) + \frac{\partial}{\partial p}\left(s^{(t-1)} \odot f^{(t)}\right) \quad (19)$$

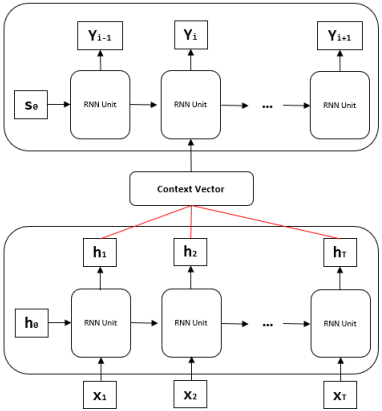
Appendix: attention mechanism

back

encoder-decoder architecture



attention mechanism



Appendix: data & supervised classification

back

1. To approximate the probabilities for both classes (good and bad) of a given sequence, all the sequences were one-hot encoded[21] as input.
2. Based on sparse categorical cross-entropy loss, several RNN models were trained and a grid search was conducted to search for the best RNN model.
3. During the inference stage, the prediction of a given sequence was chosen to be the class that had a larger probability.

Appendix: one-hot encoding map

back

Name	Three-Letter Code	One-Letter Code
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic Acid	Asp	D
Cysteine	Cys	C
Glutamic Acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V
Blank	Blk	B

Appendix: data set fro seq-method

back

All

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

Different epitope-mimicking domains: 101F and D25 (4H)

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

Different sketches: 3E2H, 4E1H, 4E2H and 4H

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