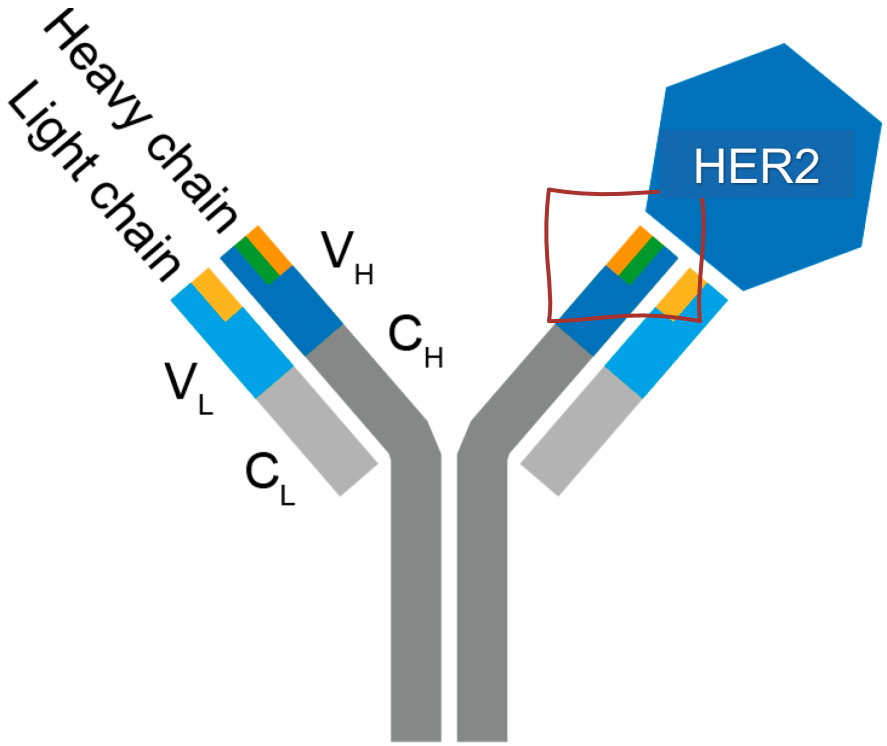
Case Study with deepCDR: Engineering Trastuzumab

Jinyan Tao
31. August 2021

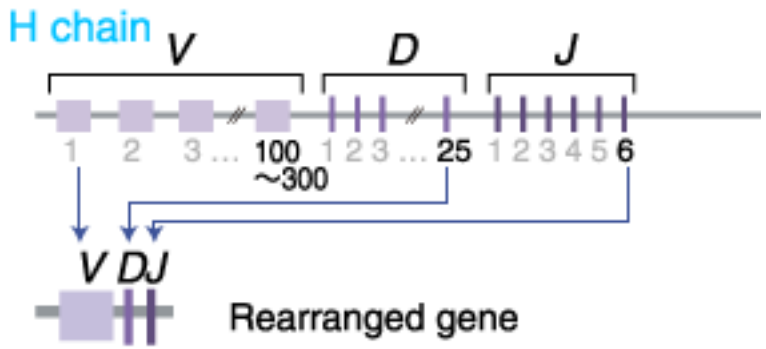
Agenda

- Introduction
- Task 1: Predicting the binding status of novel Trastuzumab Variants
- Task 2: Predicting the Affinity of Novel Trastuzumab Variants Towards HER2
- Take-away messages

Diversity in CDR3 regions determines antibody binding specificity



Trastuzumab



Genetic diversity

CSRYGVGGMFPYVYW
CSRFGQAALYAYCYW

.....

CSRWVASGFFVYLYW
CSRYNNGGLYVYRYW

CDR3 Sequence diversity

Strong binders?
Weak binders?
Non-binding?

Functional diversity

[1] Kreer, C et al. *Vaccines* **2020**, *8*, 13. https://doi.org/10.3390/vaccines8010013
[2] https://www.mblbio.com/bio/g/support/method/antibody-role.html

Machine/Deep learning optimizes antibody specificity by in-silico prediction

- Challenges:
  - Conventional bio-assays to measure binding affinity in a large-scale
- Modern solutions :
  - Machine learning models can predict binding affinities to pre-select variants from a large candidates pool

=> Acclerate antiboody discovery and optimization

CDR3 sequences

CSRYGVGGMFPYVYW
CSRFGQAALYAYCYW

.....
CSRWVASGFFVYLYW
CSRYNNGGLYVYRYW

X = sequence, Y = affinity

$$Y = f(X, \beta) + \varepsilon$$

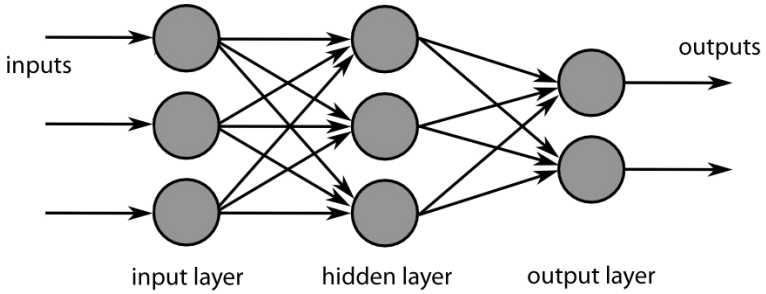
predict

Affinity

Strong binders?
Weak binders?
Non-binding?

Simple: linear regression

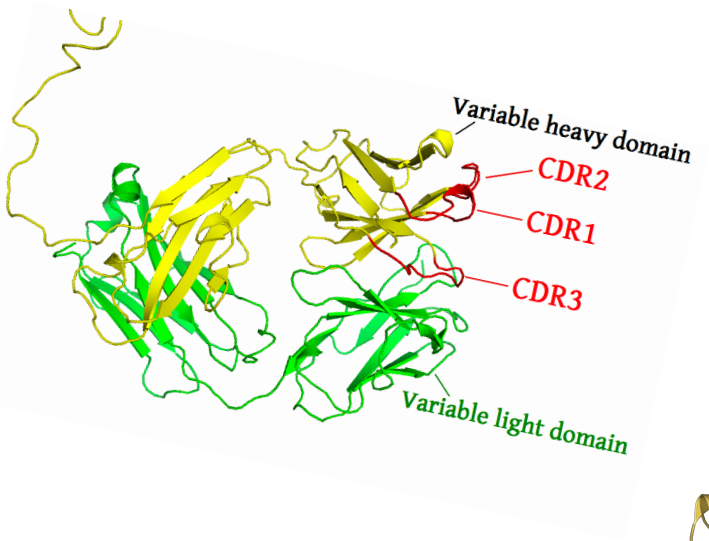
Complex:

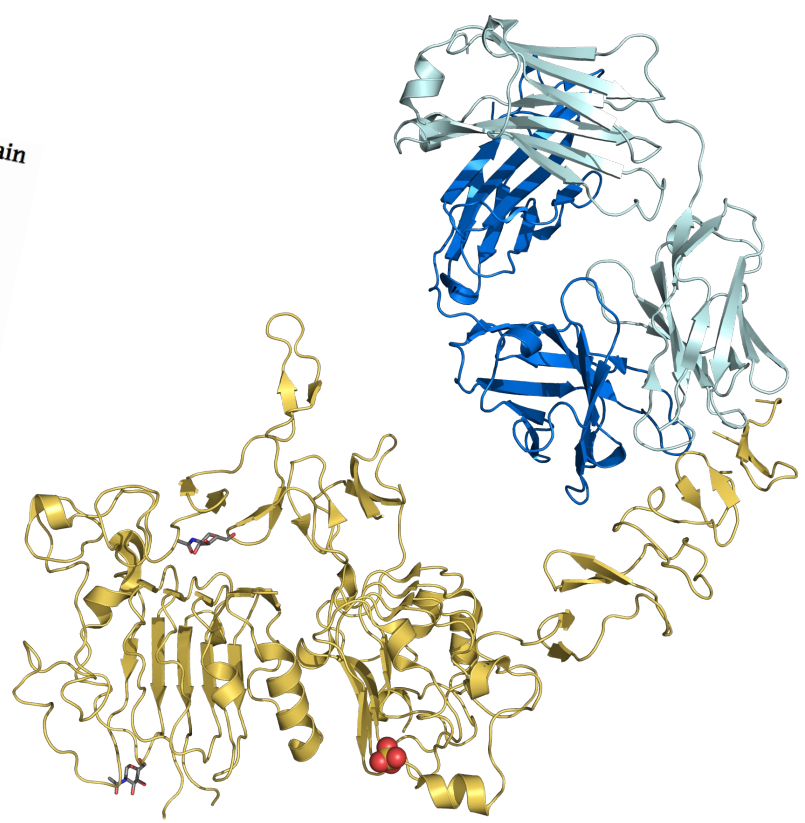


Deep neural network

https://de.wikipedia.org/wiki/Deep_Learning

Task 1: Predicting the binding status of novel Trastuzumab Variants





CDR3 sequences
of Trastuzumab variants

HER2

Binary classification

- Binding?
- Non-binding?

https://en.wikipedia.org/wiki/HER2/neu

Task 1: Predicting the binding status of novel Trastuzumab Variants

Key outlines:

1. Sequencing quality check
2. Sequence diversity analysis

1. Positional One-hot encoding
2. Feature-representation from Pre-trained language model

1. Convolutional neural networks
2. Random forest

Task 1: Predicting the binding status of novel Trastuzumab Variants

Quality control and exploratory data analysis

Data quality check

1. Applied Q score to ensure sequence quality control (QC):
    - Q score with thershold 25
        - Exclude any sequence with any base has a Q score less than 25
        - A rule of thumb cut-off is 20~30, apply a moderate thereshold 25 should guarantee high sequence quality

2. Deleted repetitive binders and non-binders with identical CDR3 amino acid sequence, however with altered CDR3 nucleotide sequence (due to codon degeneracy)

Number of Binder* entries		Number of Non-binder entries	
Before QC	After QC	Before QC	After QC
6146	**3725**	3888	**3620**

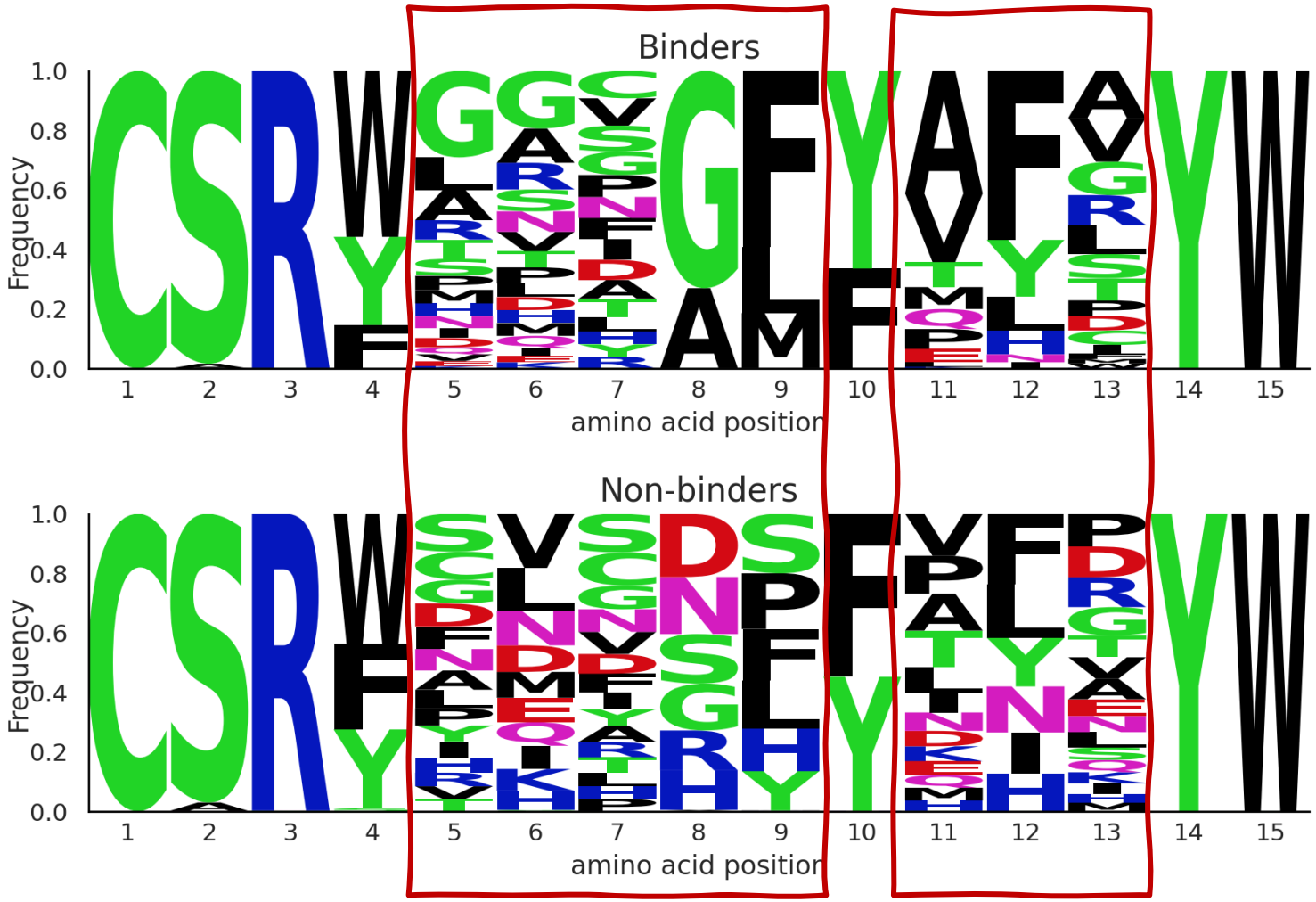
=> After two steps, data was used for downstream analysis

* note: for binders, two dataset was merged and performed quality check

[1] Q-score calculated referencing https://drive5.com/usearch/manual/quality_score.html

Positional amino acids (AAs) diversity analysis in CDR3 sequences

- Length of AA sequences : 15 residues
- **Binders and non-binders differ mainly from position 5-13**
  - Position 1-3 are moslty CSR, ending residues are YW
  - Frequency and compositions of AAs from pos. 5-13 vary



Task 1: Predicting the binding status of novel Trastuzumab Variants

Feature engineering

Feature engineering of CDR3 AA sequences representation

- **Key challenges:**
  1. To transform string-like AA sequences to machine-readable numeric values `CSRYW => 1,2,3,4,5,6`
  2. To reserve positional relationships of AAs
  3. To reserve functional properties of AAs

❖ Select AA sequences rather than nucleotide sequences, as:

- AA impacts protein interaction directly
- More diverse categories (20) to encode

Solutions:

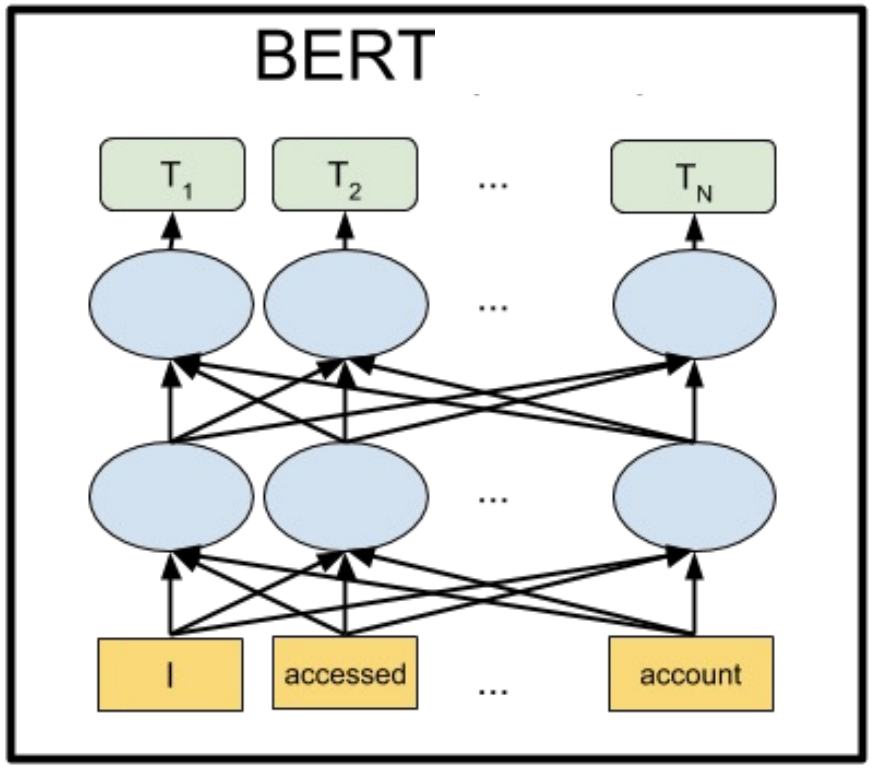
I. **Natural language processing (NLP) approach to learn 1024-vector numeric representation for each sequence**
II. **Positional one-hot encoding deriving 15 X 20 matrix to represent each sequence**

[1] Journal of Cheminformatics volume 5, Article number: 41 (2013)
[2] Mason, D.M., Friedensohn, S., Weber, C.R. *et al.* Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nat Biomed Eng* **5,** 600–612 (2021).

Feature engineering of CDR3 AA sequences

I. Natural language processing (NLP) approach to learn 1024-vector numeric representation for each sequence

- AA sequence is a language with unique syntax and semantics



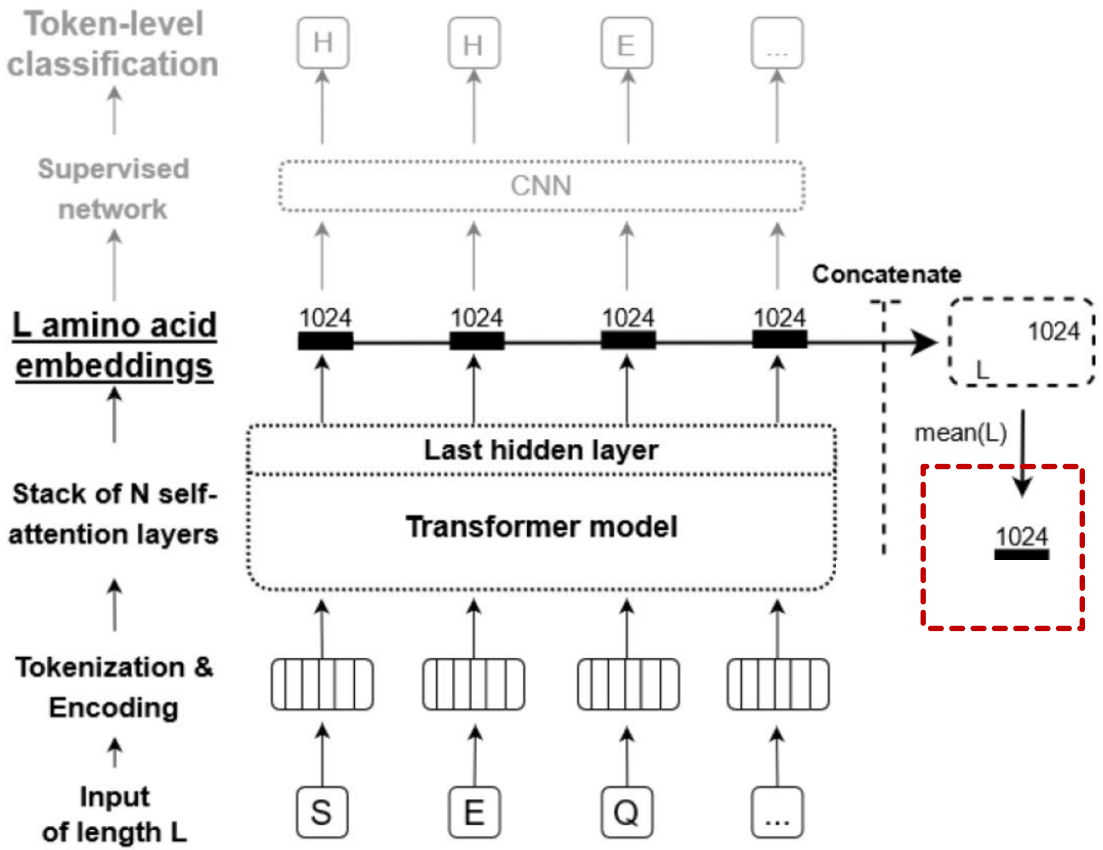
Key ideas:
Leverage BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers) model pretrained with large-scale protein sequences to generate informative representation

- bi-directional learning to gain interconnected context of CDR3 sequences from both left to right context and right to left context simultaneously

[1] A. Elnaggar *et al.*, "ProtTrans: Towards Cracking the Language of Lifes Code Through Self-Supervised Deep Learning and High Performance Computing," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, doi: 10.1109/TPAMI.2021.3095381.
[2] https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

Feature engineering of CDR3 AA sequences

I. Natural language processing (NLP) approach to learn a 1024-vector numeric representation for each sequence



Leverage BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers) model pretrained with large-scale protein sequences to generate informative representation

Key steps:

1. Tokenize CDR3 sequence and create positional encoding, feed forward to BERT
2. Generate 1024-long vector embedding for each AA
3. Concatenate embeddings and average them into one 1024-vector representation

[1] A. Elnaggar *et al.*, "ProtTrans: Towards Cracking the Language of Lifes Code Through Self-Supervised Deep Learning and High Performance Computing," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, doi: 10.1109/TPAMI.2021.3095381.

Feature engineering of CDR3 AA sequences

I. Natural language processing (NLP) approach to learn 1024-vector numeric representation for each sequence

II. Positional one-hot encoding deriving 2-dimensional 15 X 20 matrix to represent each sequence

Example sequence

C	S	R	Y	G	V	G	G	M	F	P	Y	V	T	W

=>

A	C	D	E	F	G	B	J	MLQRS	T	V	W	G
0	1	0	0	0	0	0	0	...	0	0	0	0
0	0	0	0	0	0	0	0	...	0	0	0	0
0	0	0	0	0	0	0	0	...	0	0	0	0
0	0	0	0	0	0	0	0	...	0	0	0	0
0	0	0	0	0	1	0	0	...	0	0	0	0
0	0	0	0	0	0	0	0	...	0	1	0	0
0	0	0	0	0	1	0	0	...	0	0	0	0
0	0	0	0	0	1	0	0	...	0	0	0	0
0	0	0	0	0	0	0	0	...	0	0	0	0
0	0	0	0	1	0	0	0	...	0	0	0	0
0	0	0	0	0	0	0	0	...	0	0	0	0
0	0	0	0	0	0	0	0	...	0	0	0	1
0	0	0	0	0	0	0	0	...	0	1	0	0
0	0	0	0	0	0	0	0	...	0	0	0	1
0	0	0	0	0	0	0	0	...	0	0	1	0

Key ideas:

- each row is a 20-d vector with one-hot encoding represents presence of one AA given 20 natural AAs
- 15 rows presents AA positions in the whole sequence
- Reserve positional relationships and diversity representation of AAs

[1] Mason, D.M., Friedensohn, S., Weber, C.R. *et al.* Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nat Biomed Eng* **5**, 600–612 (2021).

Task 1: Predicting the binding status of novel Trastuzumab Variants

Classification

Classification with NLP-learned & one-hot encoded representations: Positional one-hot encoding (OHE) outperforms NLP-learned representations

Dataset:
- 3725 binding sequences
- 3620 non-binding sequences
  train:test = 80:20

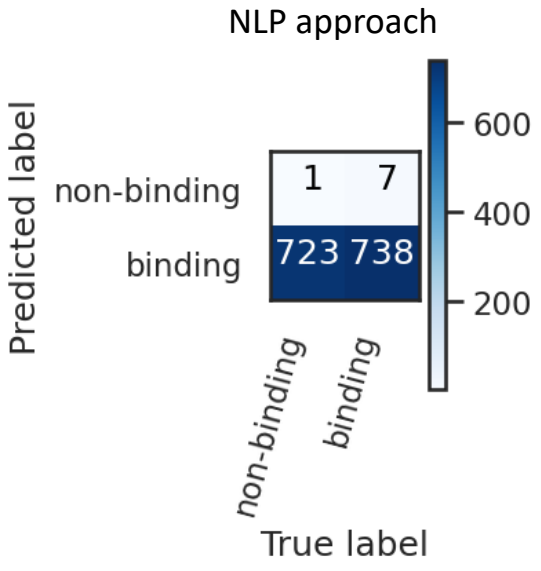
Input:
- 1024-dimensional feature vector
- 15 X 20 positional encoded matrix
  - Flattened to 300-d vector for algorithmic convenience

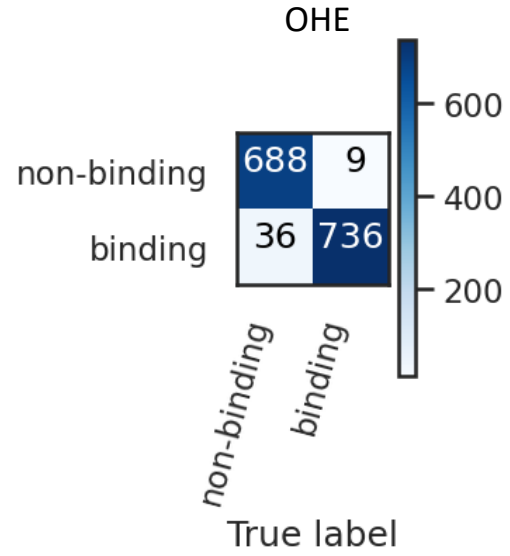
Model:
- Random forest classifier
  - Parameters tuned by grid-search

Performance evaluation
- 10-fold cross-validated
- f1, accuracy score

Test set result

	Test - F1 score	Test - accuracy
NLP	0.669	0.503
OHE	**0.970**	**0.968**

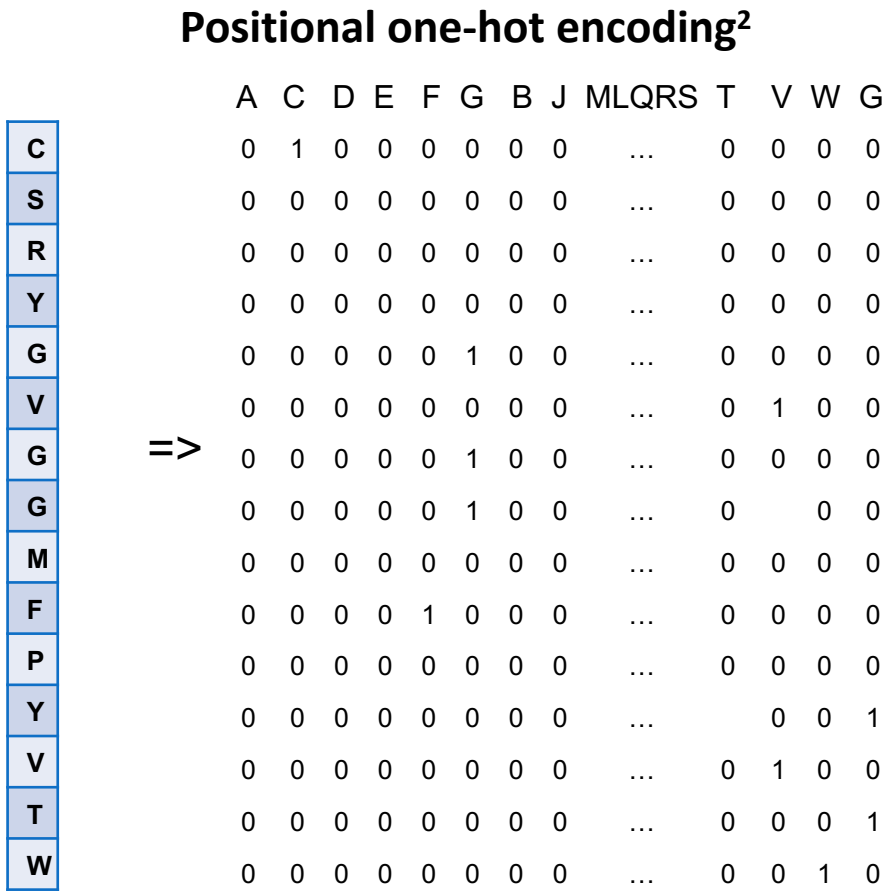




Therefore, positional one-hot encoder was seleceted for Further experiments

Learning with positional one-hot encoding

- Positional one-hot encoding deriving 2-dimensional 15 X 20 matrix to represent each sequence:

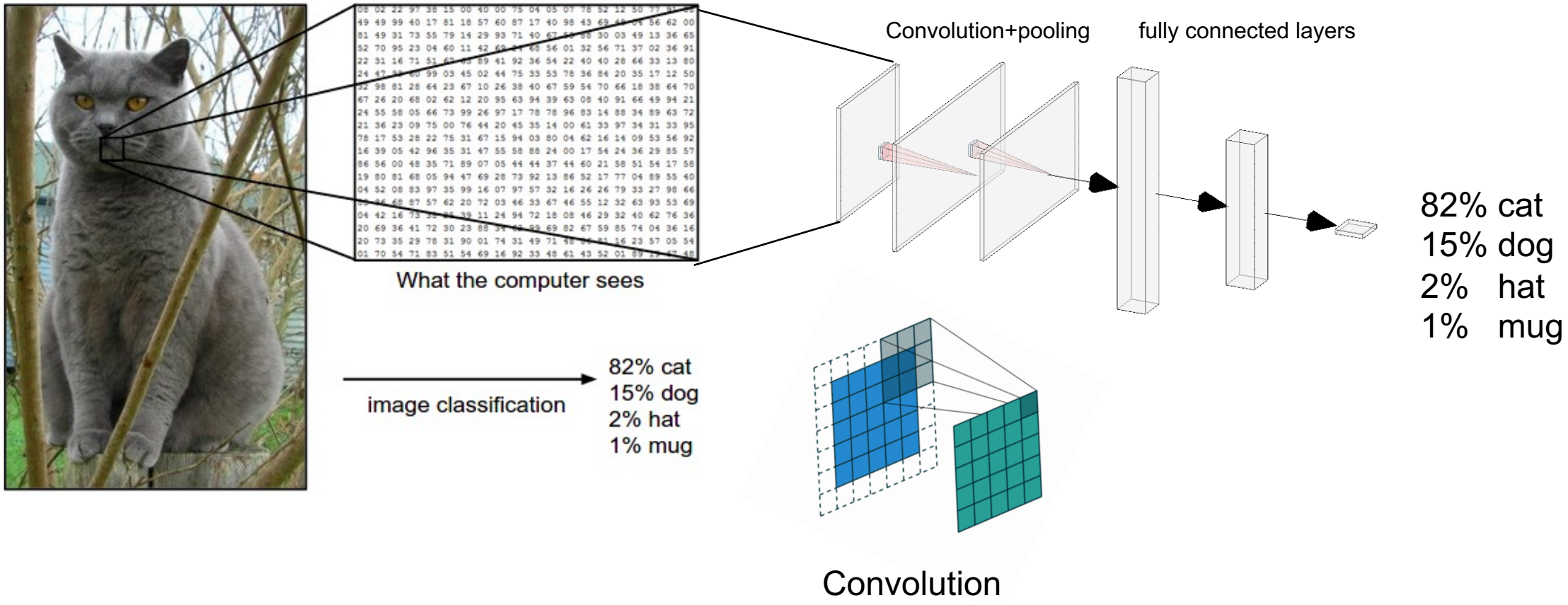


Best way to learn from spatial features?

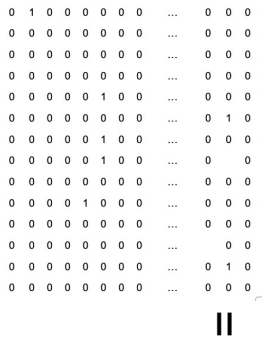
[1] A. Elnaggar *et al.*, "ProtTrans: Towards Cracking the Language of Lifes Code Through Self-Supervised Deep Learning and High Performance Computing," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, doi: 10.1109/TPAMI.2021.3095381.
[2] Mason, D.M., Friedensohn, S., Weber, C.R. *et al.* Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nat Biomed Eng* **5,** 600–612 (2021).

Classification with Convolutional Neural Networks (CNNs)

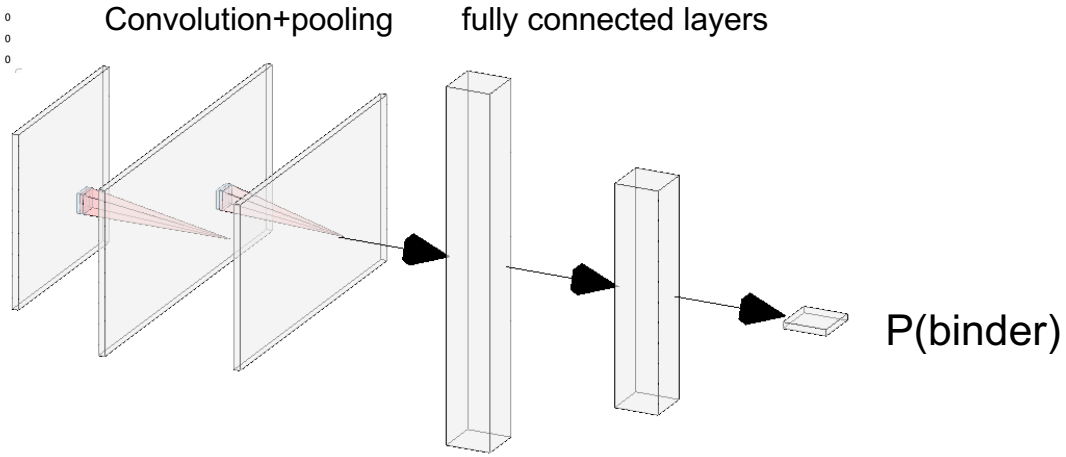
- CNNs is inspired from how brain perceives/recognize images



Classification with one-hot encoded representation and Convolutional Neural Networks (CNNs)



Vannilla CNN



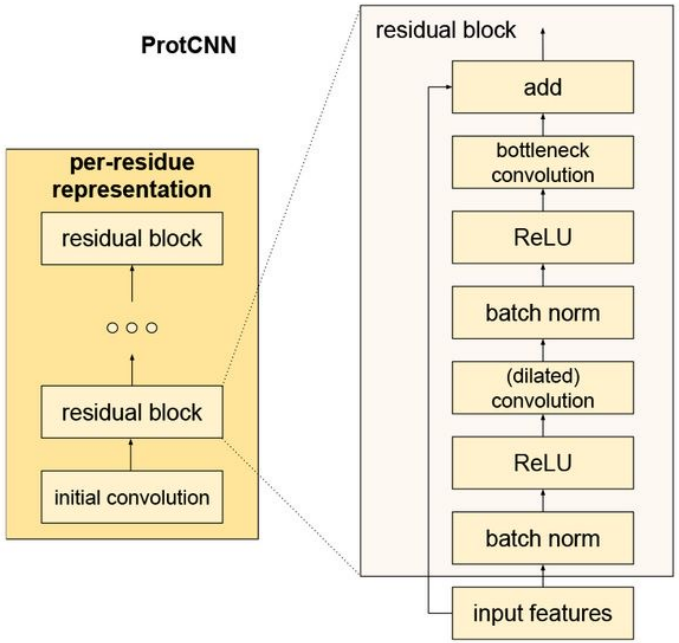
Simple architecture:

- Two convolution + max pooling layers
  - Extract spatial features and reserve spatial relationships (fits AA sequence structure)
- Two dense layers with batch normalization and dropouts to avoid overfitting

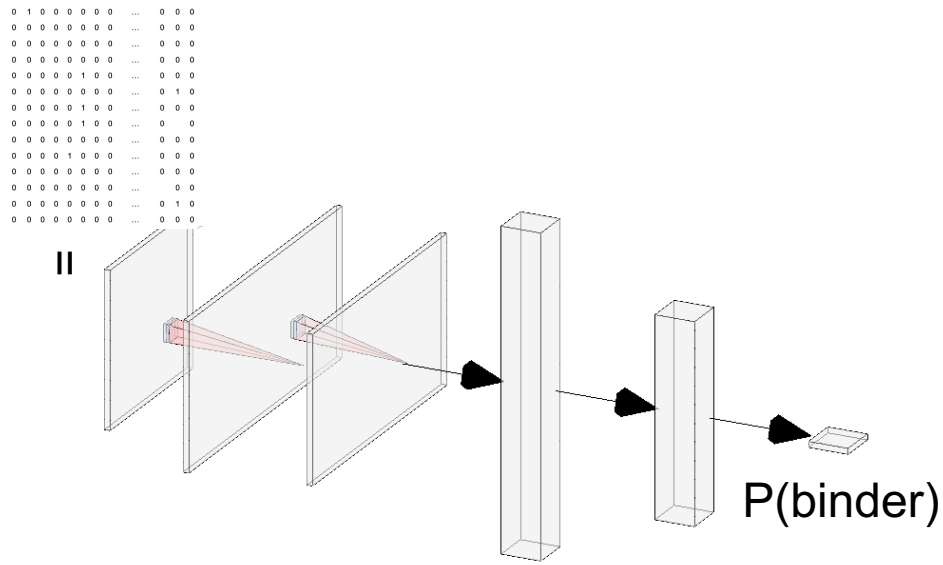
[1] Maxwell L. Bileschi, David Belanger, Drew Bryant, Theo Sanderson, Brandon Carter, D. Sculley, Mark A. DePristo, Lucy J. Colwell bioRxiv 626507; doi: https://doi.org/10.1101/626507
[2] arXiv:1512.03385

Classification with one-hot encoded representation and CNNs

Res-CNN[1]



Vannilla CNN



Key architecture: CNN + residual blocks + dilated convolution

- Residual blocks: avoid vanishing/exploding gradient problems[2]
- Dilated convolution : increase receptive field size, to include more spatial features with less computations

[1] Maxwell L. Bileschi, David Belanger, Drew Bryant, Theo Sanderson, Brandon Carter, D. Sculley, Mark A. DePristo, Lucy J. Colwell bioRxiv 626507; doi: https://doi.org/10.1101/626507
[2] arXiv:1512.03385

Classification with one-hot encoded representation:

Res-CNN slightly outperforms vanilla CNN architecture

Dataset:
- 3725 binding sequences
- 3620 non-binding sequences

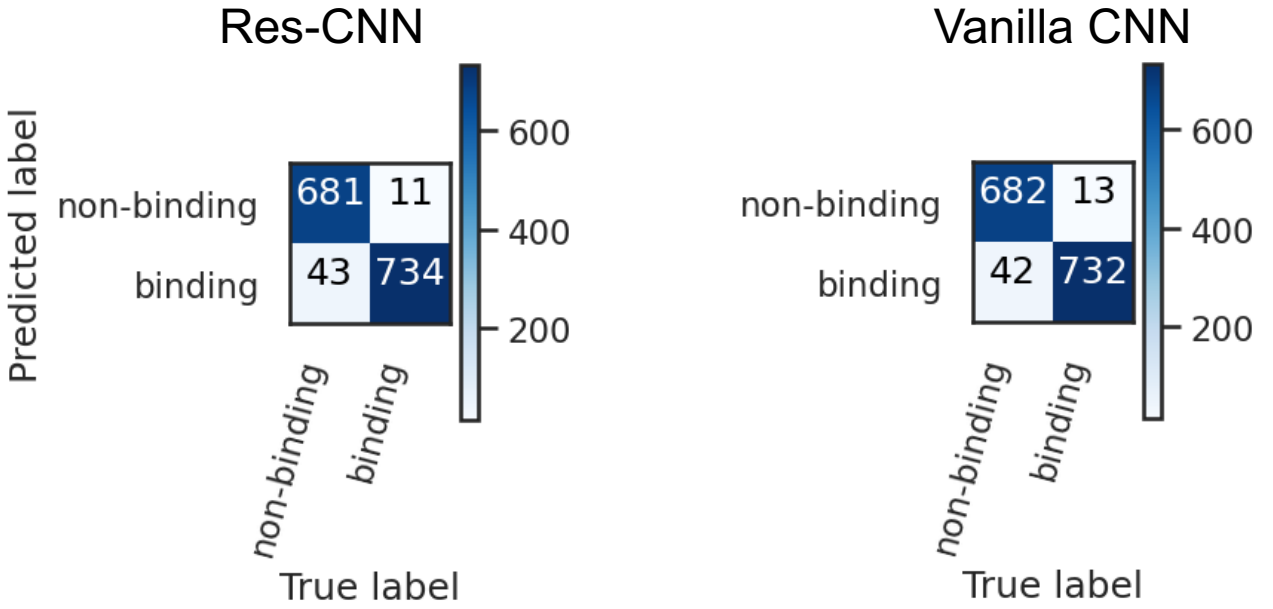
  Train : validation : test = 60:20:20

Input:
- 15 X 20 positional encoded matrix

Model:
- ResCNN & CNN
  - Trained for 200 epochs with early stopping and regularization

Performance evaluation
- Binary entropy loss
- F1, accuracy score

	Test - F1 score	Test - accuracy
Res-CNN	0.9634	0.9632
Vanilla CNN	0.9627	0.9626



Task 1: Predicting the binding status of novel Trastuzumab Variants

Conclusion:

Task 1 is a friendly start ☺, the key is to extract feature from the AA sequences

1. Highly accurate prediction of binding status of Trastuzumab can be achieved:
   - Highest: with random forest classifier (positional one-hot encoded) f1 score 0.97, accuracy 0.968
   - Res-CNN slightly outperforms vanilla CNN, slightly less performs the random forest classifier

2. CDR3 sequence embedding with positional one-hot encoding outperforms representations learned by pretraind BERT model

	Test - F1 score	Test - accuracy
Res-CNN with OHE	0.9634	0.9632
Vanilla CNN with OHE	0.9627	0.9626
Random forest with NLP encoding	0.669	0.503
Random forest with OHE	**0.970**	**0.968**

Task 1: Predicting the binding status of novel Trastuzumab Variants

Discussion & future perspective

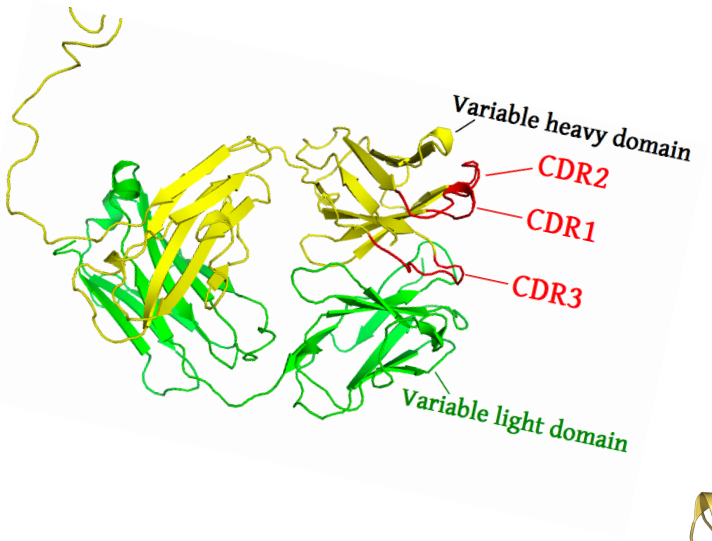
1. Positional one-hot encoding scheme retains positional relationships between AAs in a sequence
2. Learned sequence representations are 1024-dimensional numerical vector, might not reserve structural information
   ⇒ Averaging the AA representation might lead to loss of information
   ⇒ CDR3 Sequences are too short in comparison with protein that are used for pre-training
   ⇒ Needs for fine-tuning
   ⇒ Other representations can be experimented: such as protein descriptors
3. Skipping conventional tokenized & padding approach because:
   1. Sequence is too short (15 residues) to be padded
   2. Padding might lead to incomplete truncation on sequence, results in inaccurate feature extraction

*NLP-based approach was chosen out of personal interest/curiosity in experimenting new methods; positional OHE was chosen by its demonstrated good performance in literatures[1] and importance of AA and structural information to protein-antibody binding

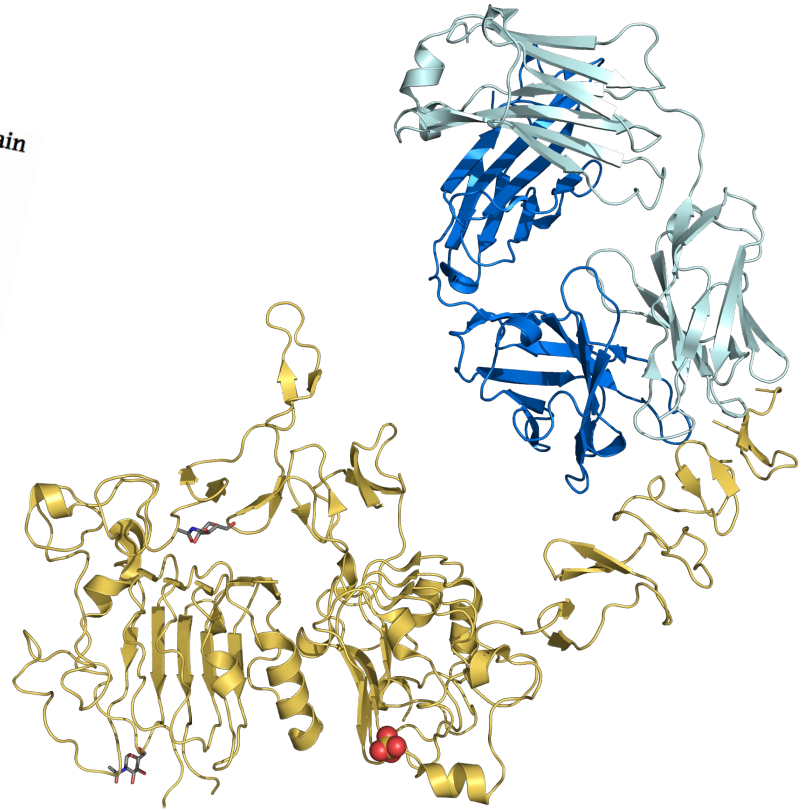
[1] Mason, D.M., Friedensohn, S., Weber, C.R. *et al.* Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nat Biomed Eng* **5,** 600–612 (2021).

Task 2: Predicting the Affinity of Novel Trastuzumab Variants Towards HER2

- Very challenging problem due to small dataset (90 entries)



CDR3 sequences
of Trastuzumab variants



HER2

Regression

Binding affinities:
KD (nm)

https://en.wikipedia.org/wiki/HER2/neu

Task 2: Predicting the Affinity of Novel Trastuzumab Variants Towards HER2

Key outline:

1. Data quality check
2. Log normalization of KD(nm)
3. Sequence diverisity analysis

1. Data augmentation with AA replacement

1. CNN
2. K-nearest neighbor

Final prediction see: https://github.com/PinaColadast/CDR3_prediction_case/blob/main/task2_predicted.csv
Implementation detals see https://github.com/PinaColadast/CDR3_prediction_case/blob/main/deepCDR_task2.ipynb

Task 2: Predicting the Affinity of Novel Trastuzumab Variants Towards HER2

Data Augmentation

Data augmentation scheme

- Key challenges: 90 entry points (a few thousands is ideal for pattern prediction task)
  - predict numerical values with low data regime
    - Challenging to recover true distribution of datasets

Basic question: Can we collect more data?

If Yes

- Go and collect more data

=> No HER2 antibody binding affinities data available

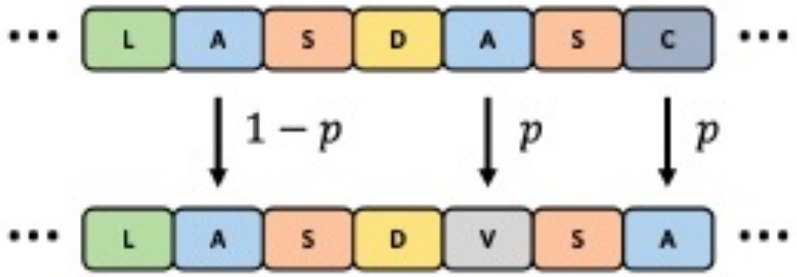
No

1. **Learning with original data**
2. **Learning with data augmentation**
3. Transfer learning with pretrained affinity model (no large-scale model founded)

Protein/AA sequence augmentation methods

Protein/AA sequence augmentation methods

- **Replacing AA residues (with probability *p*) with residues of most similar chemical properties[1,2]**
  - Retains AA properties that are crucial to bindng such as. Hydrogen donor/receptors



A:V, S:T, F:Y, K:R, C:M, D:E, N:Q, V:I, P:P, G:G, W:W, H:H, T:S, Y:F, M:C, E:D, Q:N, L:L, R:K, I:V

1. Pairing each naturally-occuring AA with a partner that belongs to the same class (aliphatic, hydroxyl, cyclic, aromatic, basic, or acidic)
2. Proline (only backbone cyclic), glycine (only an H side-chain), trytophan (indole side chain), histidine (basic side chain with no size or chemical reaction equivalent) remains same residue*

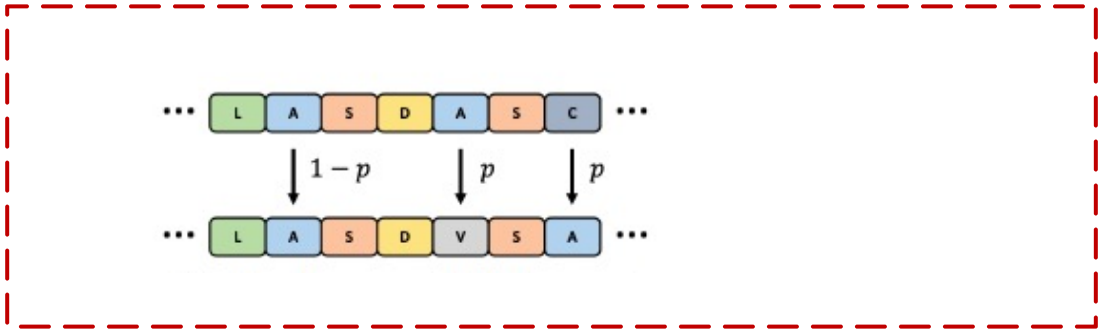
[1] Hongyu Shen, Layne C. Price, Taha Bahadori, Franziska Seeger **doi:** https://doi.org/10.1101/2021.02.18.431877
[2] French, S., Robson, B. What is a conservative substitution?. *J Mol Evol* **19,** 171–175 (1983). https://doi.org/10.1007/BF02300754

Protein/AA sequence augmentation methods with replacement with chemically similar residues

- The adaptation of replacement augmentation has following steps:
  1. Randomly select 1 position in each AA sequence (typically position 4-13 that are highly variable)
  2. With probability 0.1, the residue will be replaced, chance of 0.9 it won't be replaced
  3. Replacement follows pairwise rule:

```
A:V, S:T, F:Y, K:R, C:M, D:E, N:Q, V:I, P:P, G:G, W:W, H:H, T:S, Y:F, M:C, E:D, Q:N, L:L, R:K, I:V
```

  4. Repeat 1-3 steps 100 times for each sequence



Repeat 100 times for each sequence

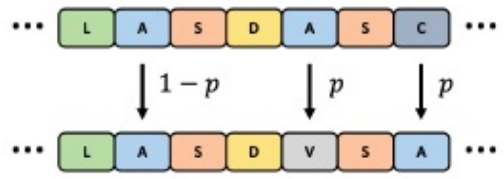
Protein/AA sequence augmentation methods with replacement with chemically similar residues

- The adaptation of replacement augmentation has following steps:
  1. Randomly select 1 position in each AA sequence (typically position 4-13 that are highly variable)
  2. With probability 0.1, the residue will be replaced, chance of 0.9 it won't be replaced
  3. Replacement follows pairwise rule:

```
A:V, S:T, F:Y, K:R, C:M, D:E, N:Q, V:I, P:P, G:G, W:W, H:H, T:S, Y:F, M:C, E:D, Q:N, L:L, R:K, I:V
```

  4. Repeat 1-3 steps 100 times for each sequence

- Corresponding affinity value is simulated from parent log KD value plus 1% Gaussian noise ~N(0,1)



Repeat 100 times for each sequence

Affinity' = 99% * Affinity + 1% * Y~ N(0,1)

=> To simulate experiments measure variance

=> 5700 CDR3 sequences in total for training

[1] Hongyu Shen, Layne C. Price, Taha Bahadori, Franziska Seeger **doi:** https://doi.org/10.1101/2021.02.18.431877

Learning with positional one-hot encoding of CDR3 sequences

- Positional one-hot encoding deriving 15 X 20 matrix to represent each sequence:
  - Selected by task 1 benchmark result

Positional one-hot encoding

		A	C	D	E	F	G	B	J	MLQRS	T	V	W	G
C		0	1	0	0	0	0	0	0	…	0	0	0	0
S		0	0	0	0	0	0	0	0	…	0	0	0	0
R		0	0	0	0	0	0	0	0	…	0	0	0	0
Y		0	0	0	0	0	0	0	0	…	0	0	0	0
G		0	0	0	0	0	1	0	0	…	0	0	0	0
V		0	0	0	0	0	0	0	0	…	0	1	0	0
G	=>	0	0	0	0	0	1	0	0	…	0	0	0	0
G		0	0	0	0	0	1	0	0	…	0		0	0
M		0	0	0	0	0	0	0	0	…	0	0	0	0
F		0	0	0	0	1	0	0	0	…	0	0	0	0
P		0	0	0	0	0	0	0	0	…	0	0	0	0
Y		0	0	0	0	0	0	0	0	…		0	0	1
V		0	0	0	0	0	0	0	0	…	0	1	0	0
T		0	0	0	0	0	0	0	0	…	0	0	0	1
W		0	0	0	0	0	0	0	0	…	0	0	1	0

[1] Mason, D.M., Friedensohn, S., Weber, C.R. *et al.* Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nat Biomed Eng* **5**, 600–612 (2021).

Task 2: Predicting the Affinity of Novel Trastuzumab Variants Towards HER2

Regression

Regression with original data and data augmentation:

With simplest K-nearest neighbor regressor (no restriction on data size)

Dataset:
- original data: 90
  - train:test = 60:20:20
- Augmented training set: 4000
  - Randomly selected from 5700 augmented training data

Input:
- 15 X 20 positional encoded matrix
  - Flattened to 300-d vector for algorithmic convenience

Model:
- KNN regressor: k=4
  - Hamming loss as distance metrics
    - Simiar to levenstein distance but normalized by length of vector
  - Parameters tuned

Performance evaluation
- Spearman rank correlation & mean squared error (MSE) (based on natural logarithm level)

Test set performance

	Rank correlation	MSE (log KD)
With augmentation	0.324**	1.200
Without augmentation	0.690**	0.760

KNN performs better without data augmentation
- **Rank correlation shows moderate to strong correlation**
- **KNN is not immune to noises in datasets**

**: statistically significant (p-value <0.01)

Regression with simple KNN exhibits comparable performance as Res-CNN

Dataset:
- Augmented training set: 4000
  - Randomly selected from 5700 augmented training data
  - Train: Val: Test = 60:20:20

Input:
- 15 X 20 positional encoded matrix

Model:
- Res-CNN
  - Similar model architecture to task1 adapted to regression

Performance evaluation
  - Spearman rank correlation & MSE (based on natural logarithm level)

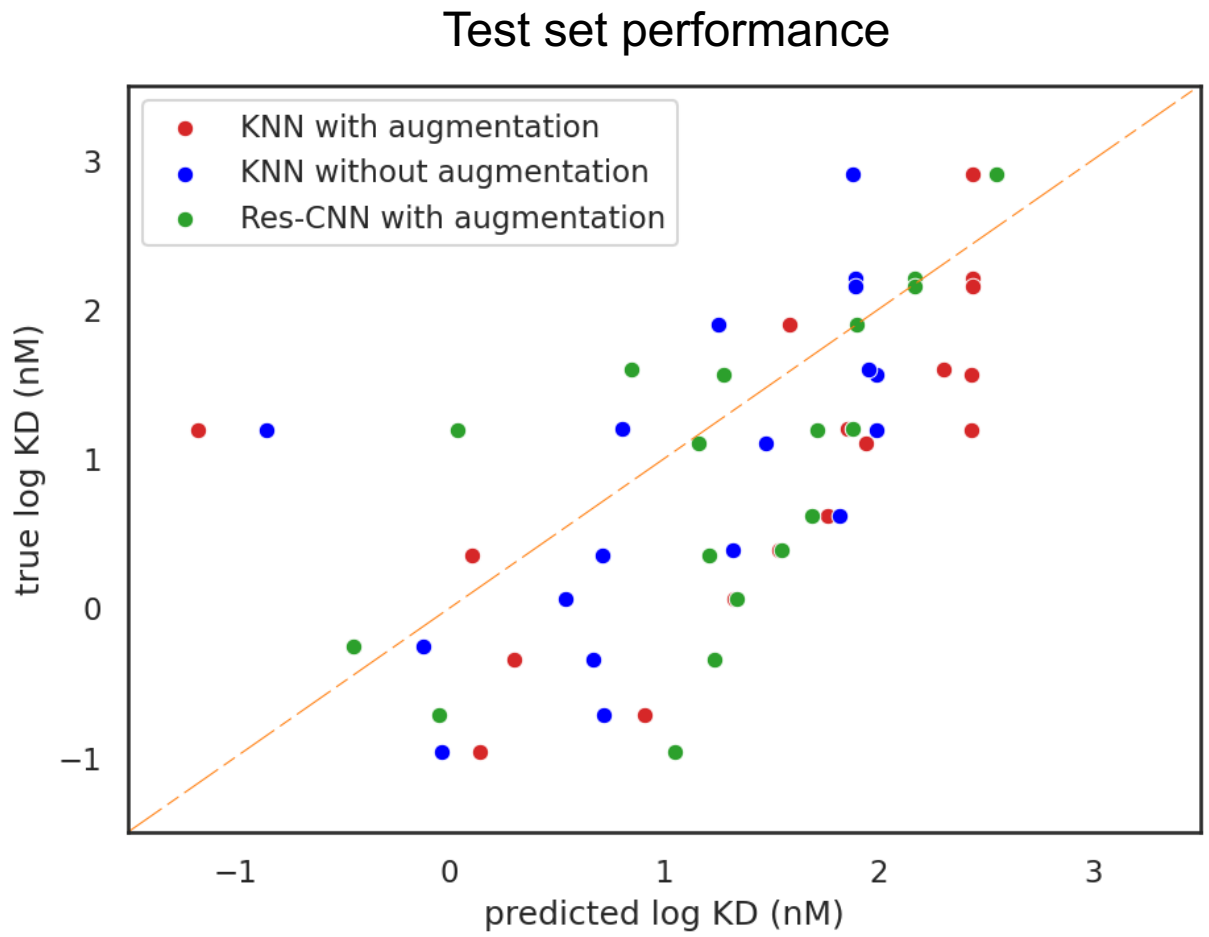
Test set performance comparison

	Rank correlation	MSE (log KD)
KNN With augmentation	0.324**	1.200
Res-CNN with augmentation	**0.702****	**0.82**
KNN without augmentation	**0.690****	**0.760**
Res-CNN Without augmentation	NA, model not converge due to too small size of data points	

❖ Res-CNN trained with data augmentation shows worse MSE but slightly better rank correlation

** statisitacally significant (p-value <0.01)

Regression with simple KNN slightly outperforms Res-CNN with data augmentation

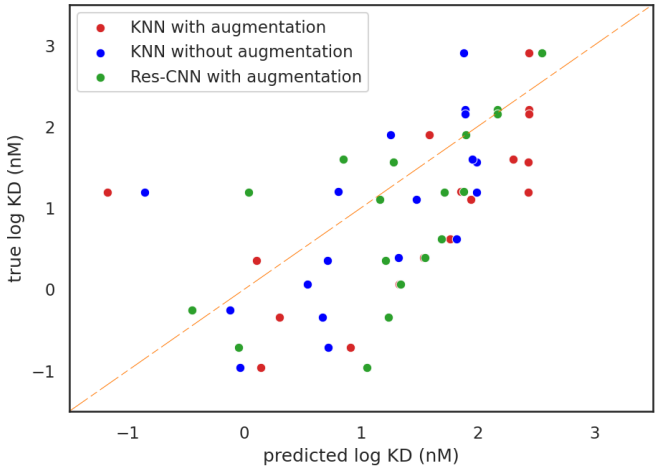


Task 2: Predicting the Affinity of Novel Trastuzumab Variants Towards HER2

Conclusion:

1. Challenging task to predict accurately with low data regime, lower accuracy than task 1
2. With data augmentation:
    i. KNN performs worse with data augmentation
        i. Pros: Might perform well under low data regime given selection of representative training data
        ii. Cons: KNN predicts by extropolating distance relationships between data points and performs weighted sum of neighbors, not robust against noise

    ii. Res-CNN performs better with augmentation
        i. Pros: Neural networks learns well with noisy and synthetic data
        ii. Cons: Deep learning requires larger dataset



Task 2: Predicting the Affinity of Novel Trastuzumab Variants Towards HER2

Future perspective:

1. More robust data augmentation techniques can be experimented
   - e.g. masking tokens in AA sequences
   - Literature[1] highlights other robust augmentation methods however without providing example code and algorithm, challenging to implement within limited timeframe

2. Transfer learning with pretrained protein properties model
   - Not so many large-scale affinity prediction data available
   - there might be other protein properties prediction dataset available, pretrained model can be used to initiate training => pretrained on task 1 dataset
3. more robust prediction algorithms (or postulate mathematical relationships between features and affinity)
4. Check prediction outliers to diagnose wrong prediction

5. **Best approach:**
   I. **Generate more data with experiment measurement ☺ (if possible)**
   II. Molecular simulation to mimic protein-antibody binding affnity as ground truth…(less recommended)

[1] Hongyu Shen, Layne C. Price, Taha Bahadori, Franziska Seeger **doi:** https://doi.org/10.1101/2021.02.18.431877

Take-away messages

- Data size/quality determines model performance
- Data-centric approach improves model performance more than model-centric approach
- **Protein/antibody engineering with machine learning is FUN ! ☺**

ETH zürich

Questions?

ETH zürich

Thank you!

For details of implementation (code) please check:
https://github.com/PinaColadast/CDR3_prediction_case

Back-up Slides

Machine/Deep learning optimizes antibody specificity by in-silico prediction

- Machine learning models trained from sequence and binding properties data can predict binding affinities
- Predicted performance serves as criterion to pre-select variants from a large candidates pool

=> Acclerate antiboody discovery and optimization

CDR3 sequences

CSRYGVGGMFPYVYW
CSRFGQAALYAYCYW

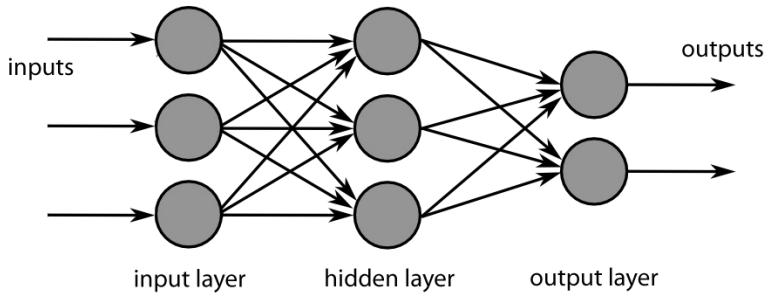
.....
CSRWVASGFFVYLYW
CSRYNNGGLYVYRYW

X = sequence, Y = affinity

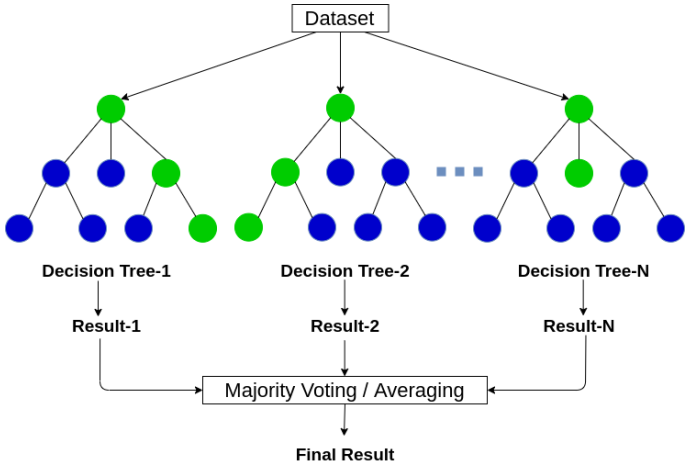
$$Y = f(X, \beta) + \varepsilon$$

Affinity

Strong binders?
Weak binders?
Non-binding?



Deep neural network



Random Forest

https://de.wikipedia.org/wiki/Deep_Learning

Feature engineering of CDR3 AA sequences representation

- Key challenges:
  1. To transform string-like AA sequences to machine readable numeric values

  ✓ Solutions:

  I. Natural language processing (NLP) approach to learn the data representation
  II. Binary encoding representing ab /presence of AAs
  III. Protein descriptors (protein fingerprints, electrostatic potential properties descriptors etc.)

  2. To reserve positional relationships of AA

  ✓ Solutions:

  1. Positional one-hot encoding of AAs
  2. Explicit protein descriptors

  3. To reserve functional properties of AA

  ✓ Solutions:

  I. Implicit categorical encoding of 20 natural AAs
  II. Explicit protein descriptors

❖ select AA sequences rather than nucleotide sequencs, as:

- AA impacts protein interaction directly
- More diverse categories (20) to encode

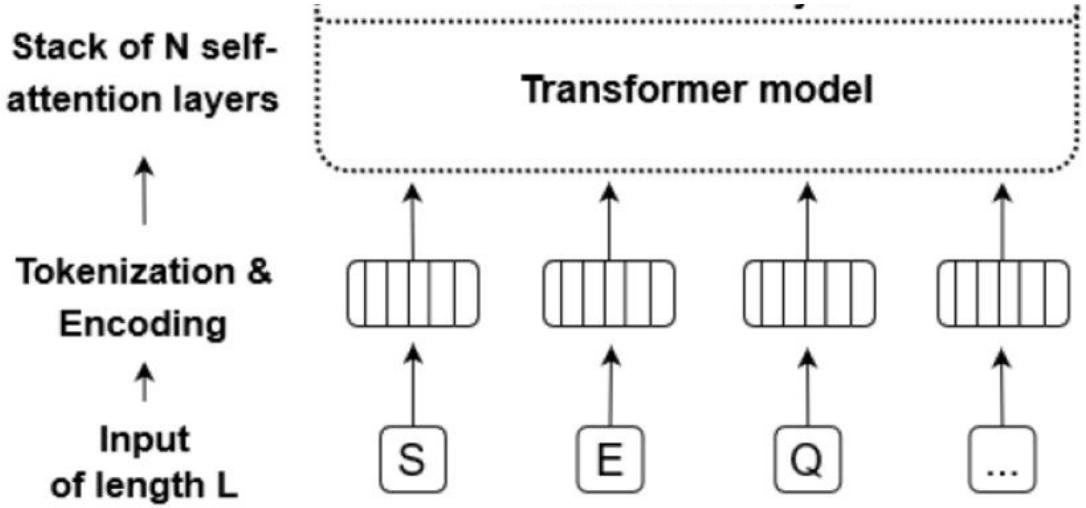
[1] Journal of Cheminformatics volume 5, Article number: 41 (2013)

Feature engineering of CDR3 AA sequences

I. Natural language processing (NLP) approach to learn 1024-vector numeric representation for each sequence

- AA sequence is a language with unique syntax and semantics

NLP-based representation learning[1]



Key ideas:
Leverage BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers) model pretrained with large-scale protein sequences to generate representation

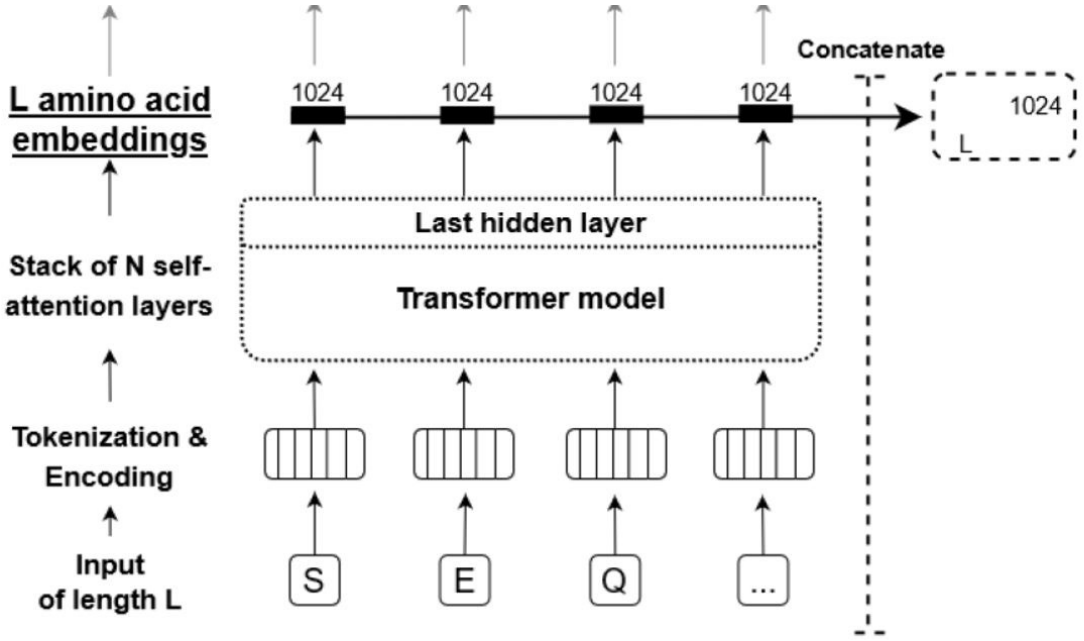
Key steps:

1. Tokenize protein sequence and create positional encoding, feed forward to BERT

[1] A. Elnaggar *et al.*, "ProtTrans: Towards Cracking the Language of Lifes Code Through Self-Supervised Deep Learning and High Performance Computing," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, doi: 10.1109/TPAMI.2021.3095381.

Feature engineering of CDR3 AA sequences

I. Natural language processing (NLP) approach to learn 1024-vector numeric representation for each sequence



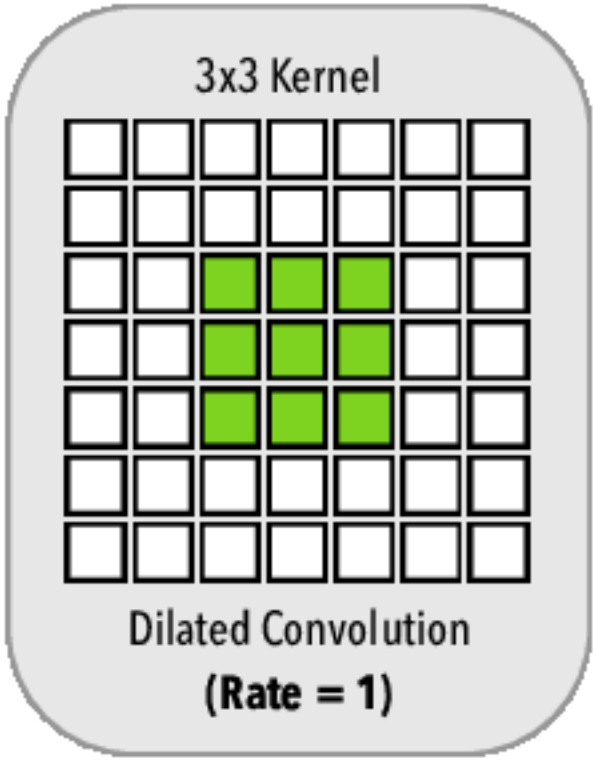
Key ideas:
Leverage BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers) model pretrained with large-scale protein sequences to generate representation

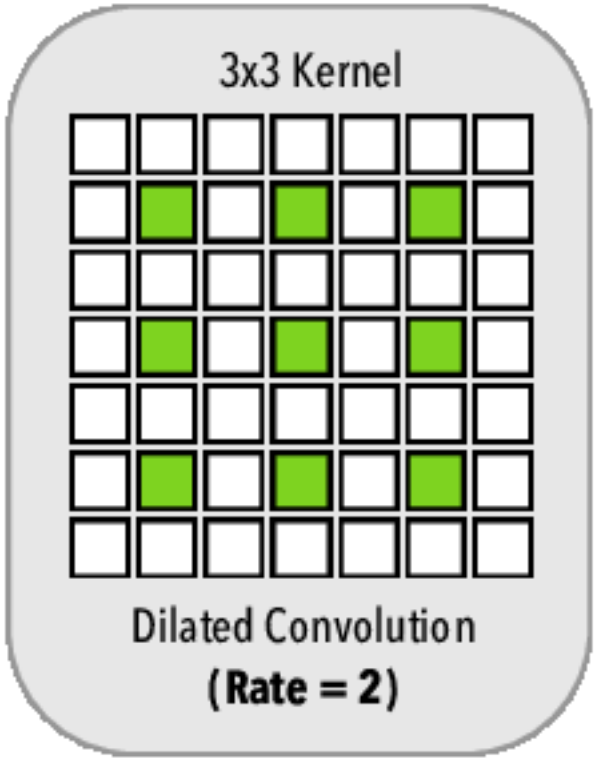
Key steps:

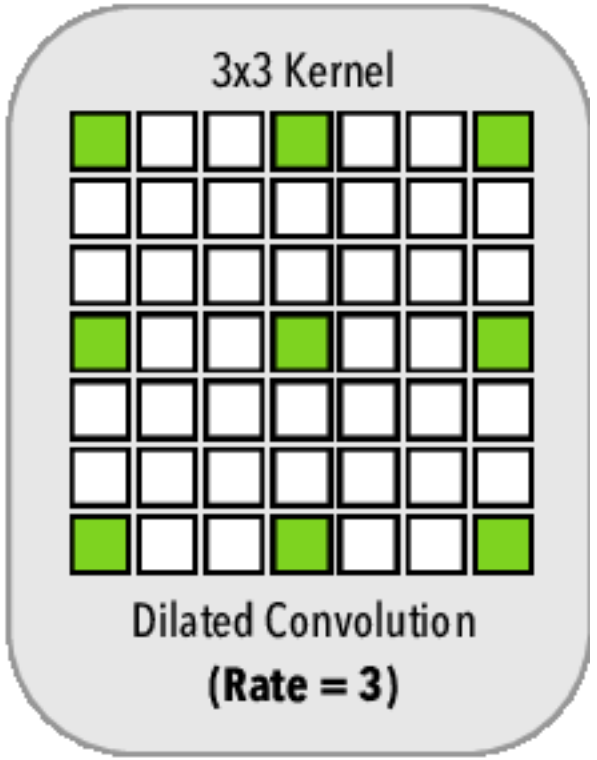
1. Tokenize protein sequence and create positional encoding, feed forward to BERT
2. Generate 1024- long vector embedding for each AA

[1] A. Elnaggar *et al.*, "ProtTrans: Towards Cracking the Language of Lifes Code Through Self-Supervised Deep Learning and High Performance Computing," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, doi: 10.1109/TPAMI.2021.3095381.

Dialated convolutions

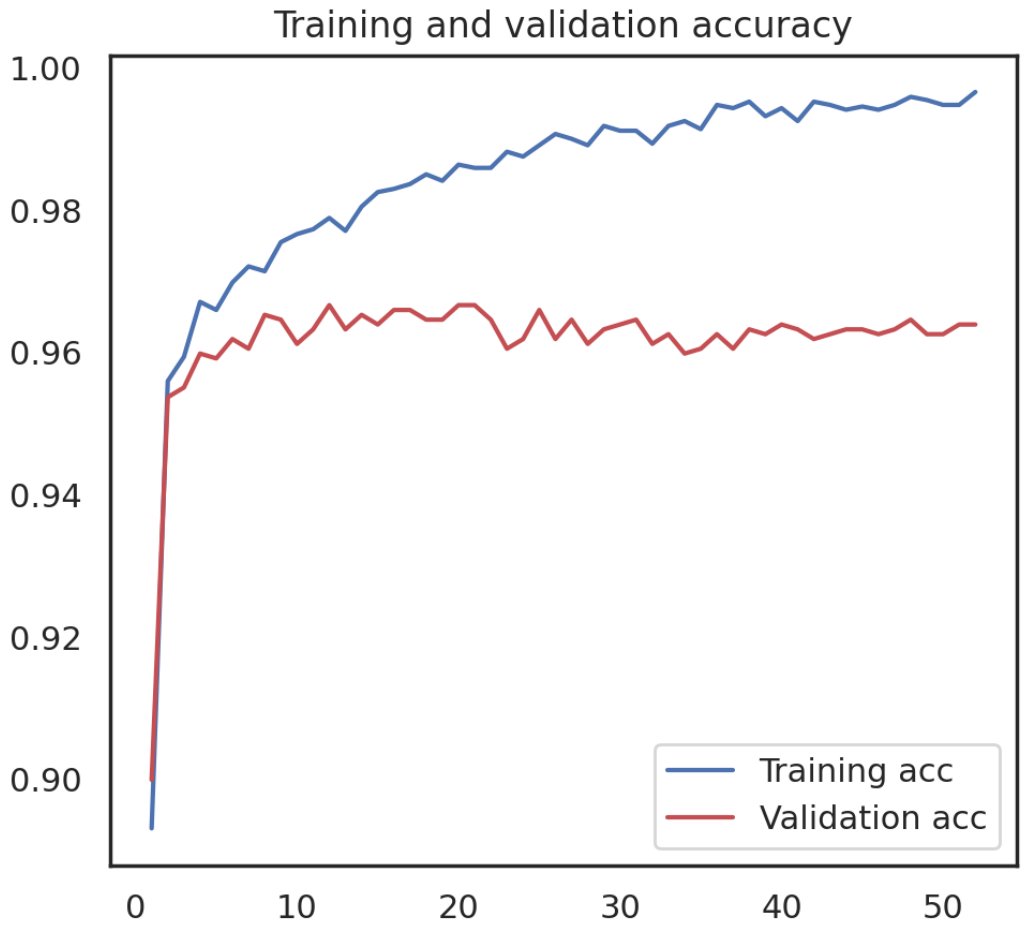


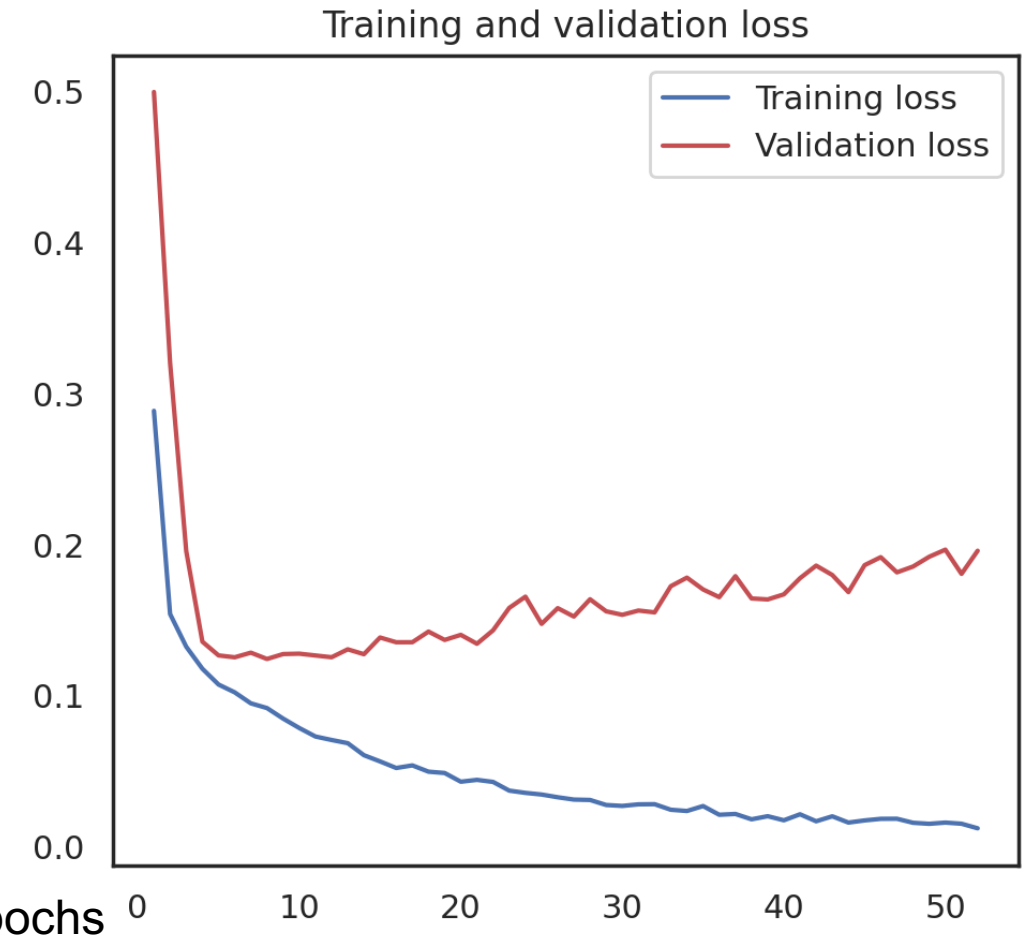




Classification with one-hot encoded representation:

Vanilla CNN architecture training history shows train loss approaches 0 with more epochs, validation loss oscillates after epoch 20

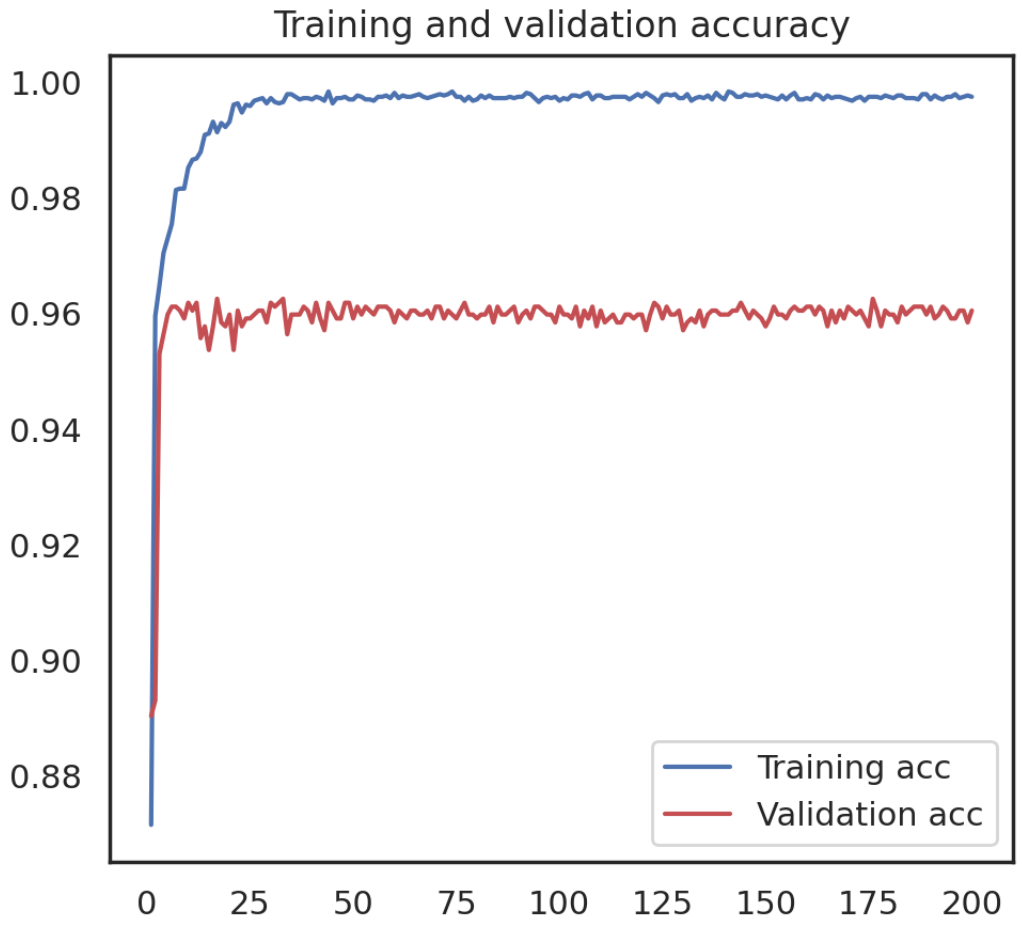


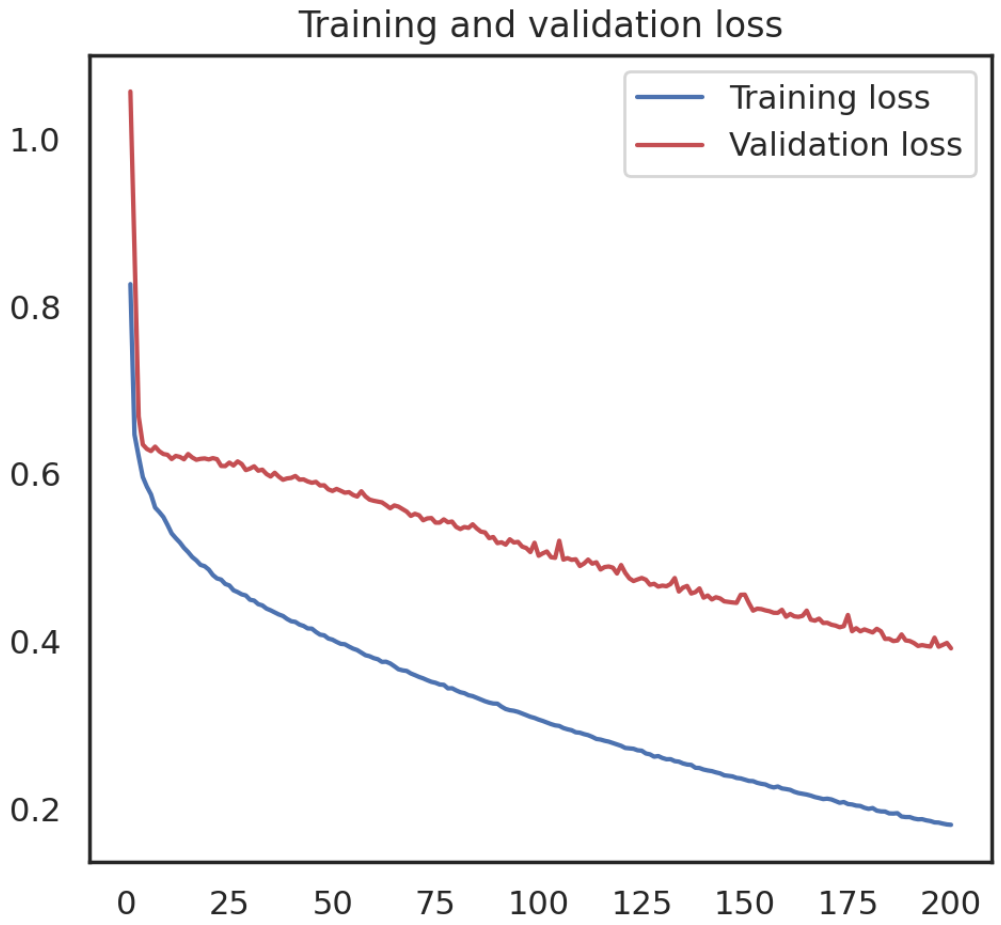


Epochs

Classification with one-hot encoded representation:

Res-CNN architecture training history shows train loss approaches 0 with more epochs, validation loss keep decreasing with accuracy reaching plateau after epoch 50





Epochs

Classification with NLP-learned & one-hot encoded representations: Positional one-hot encoding outperforms NLP-based representations

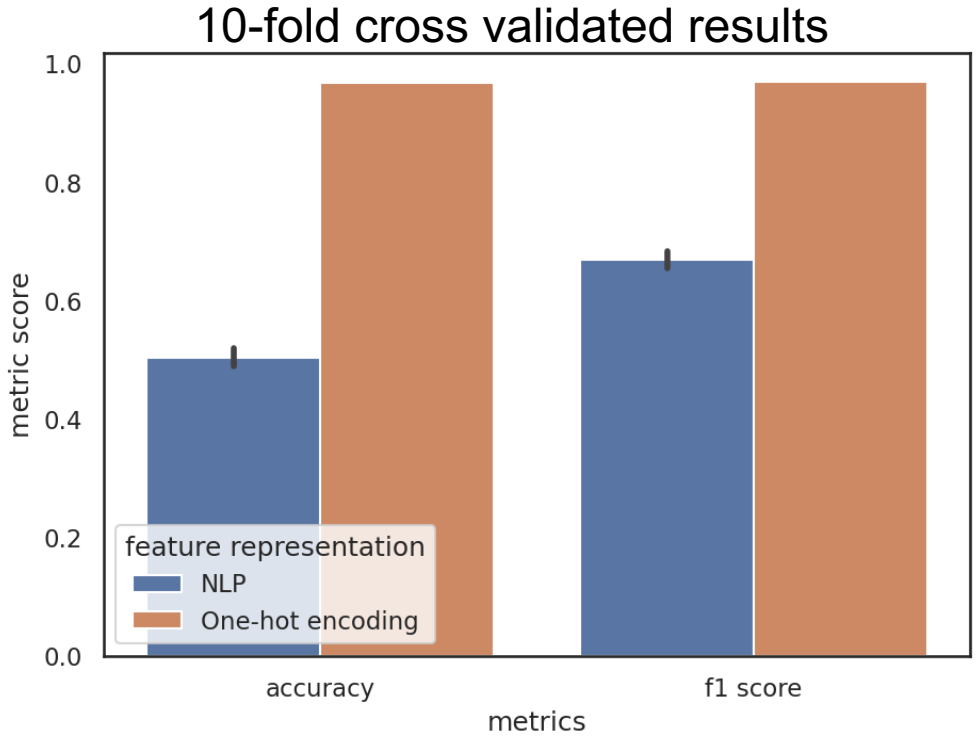
Dataset:
- 3725 binding sequences
- 3620 non-binding sequences

train:test = 80:20

Input:
- 1024-dimensional feature vector
- 15 X 20 positional encoded matrix
  - Flattened to 300-d vector for algorithmic convenience

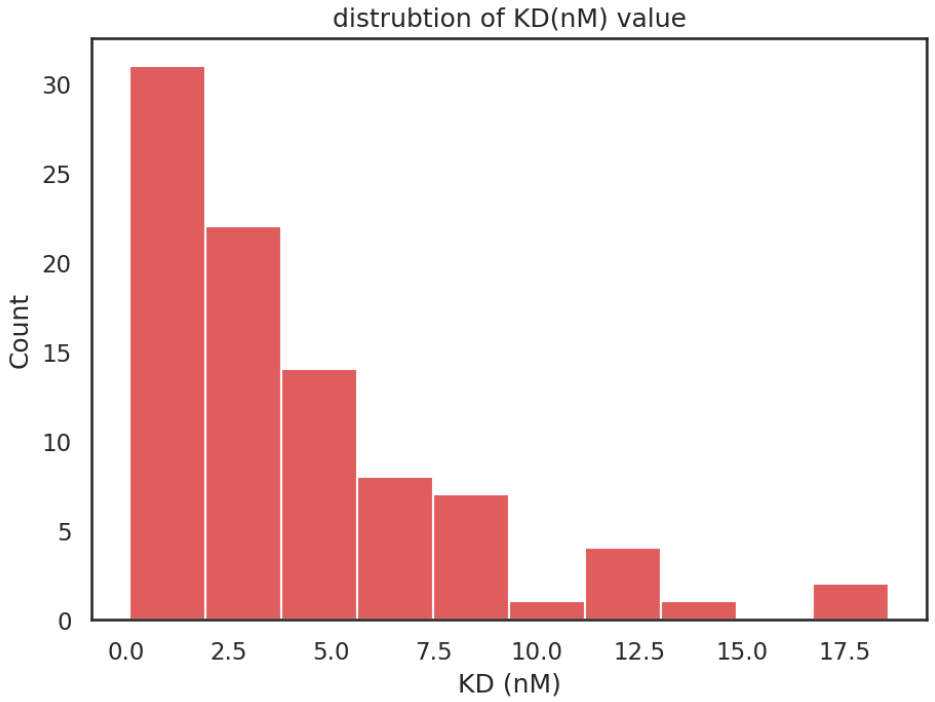
Model:
- random forest classifier
  - Parameters tuned by grid-search

Performance evaluation
- 10-fold cross-validated
- f1, accuracy score



Log(natural) normalization of KD values

- KD values exhibits right skewed distibution and with large fold gap between small and large values
- Log normalization of KD values gives less sample variance and smaller fold gap
  - Sequencing quality control was skipped due to absence of sequencing quality data (assuming high quality)
  - No repetitive data entries



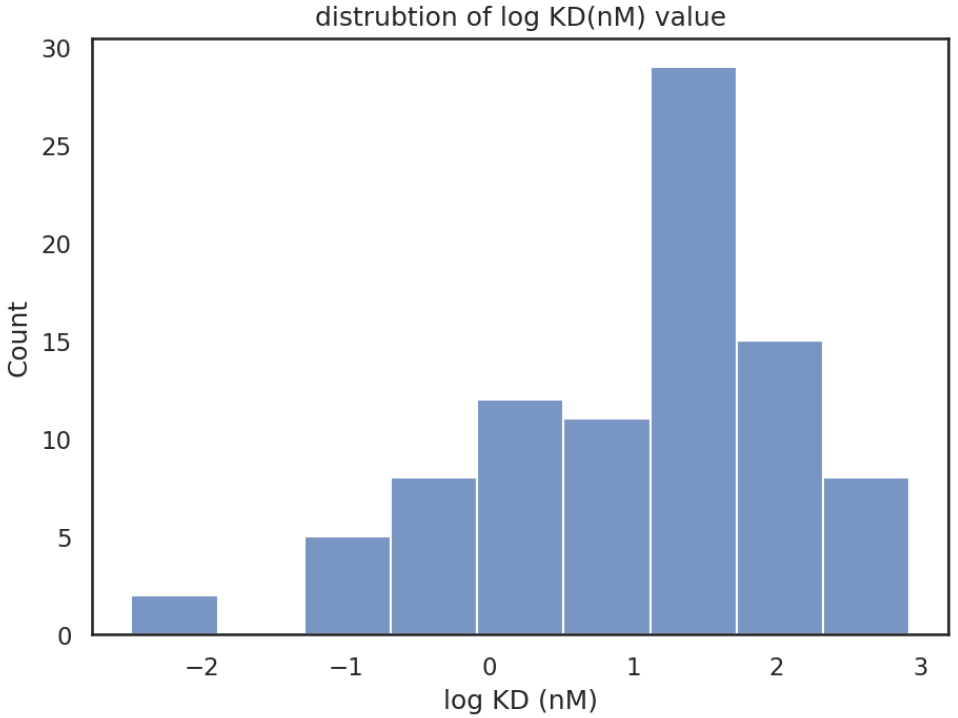
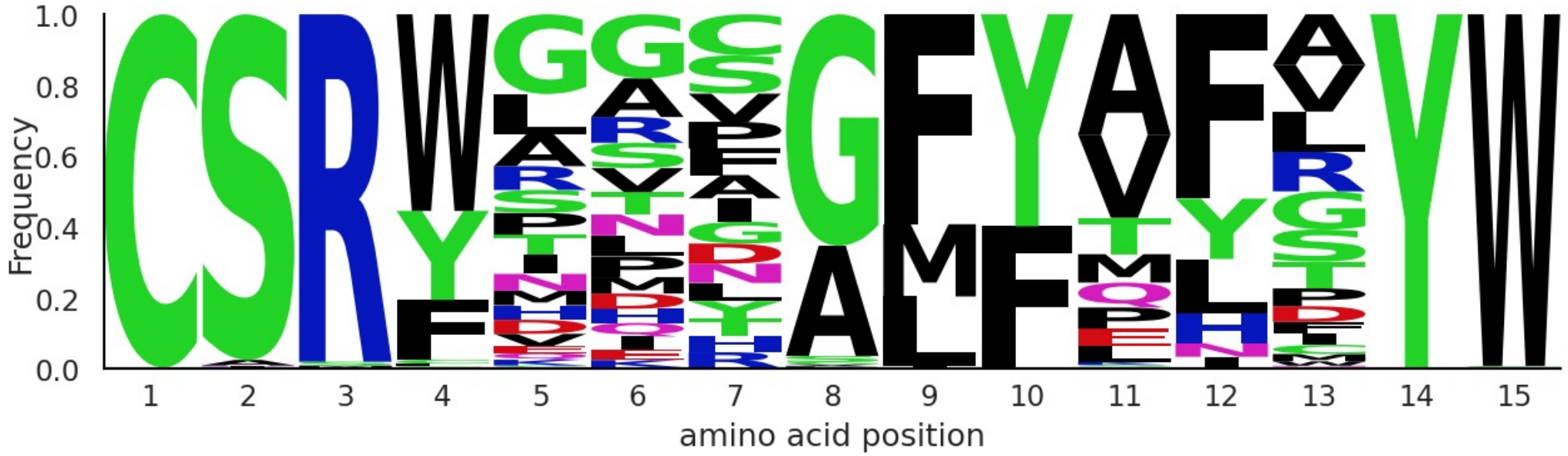


Figure: histgrams of binding affinity distributions

Sequence diversity analysis

- Length of AA sequences : 15 residues
- similar residue compositions and positions as data from task 1
  - Pos 1-3 : CSR mostly
  - Pos 14-15: YW



Protein/AA sequence augmentation methods

- Literature research has benchmarked AA sequence augmentation methods and highlighted its generailizibitliy
  - ❖ **Replacing residues (with probability p) with residues of most similar chemical properties[1,2]**
    - **Retains AA properties that are crucial to bindng such as. Hydrogen donor/receptors**

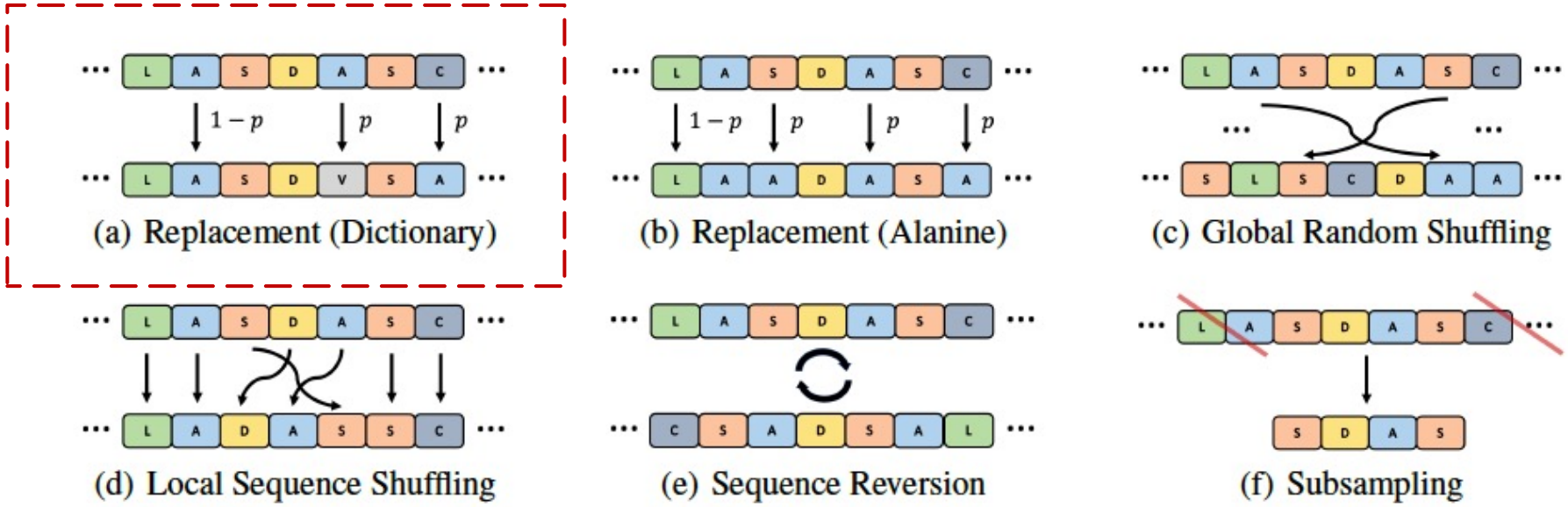


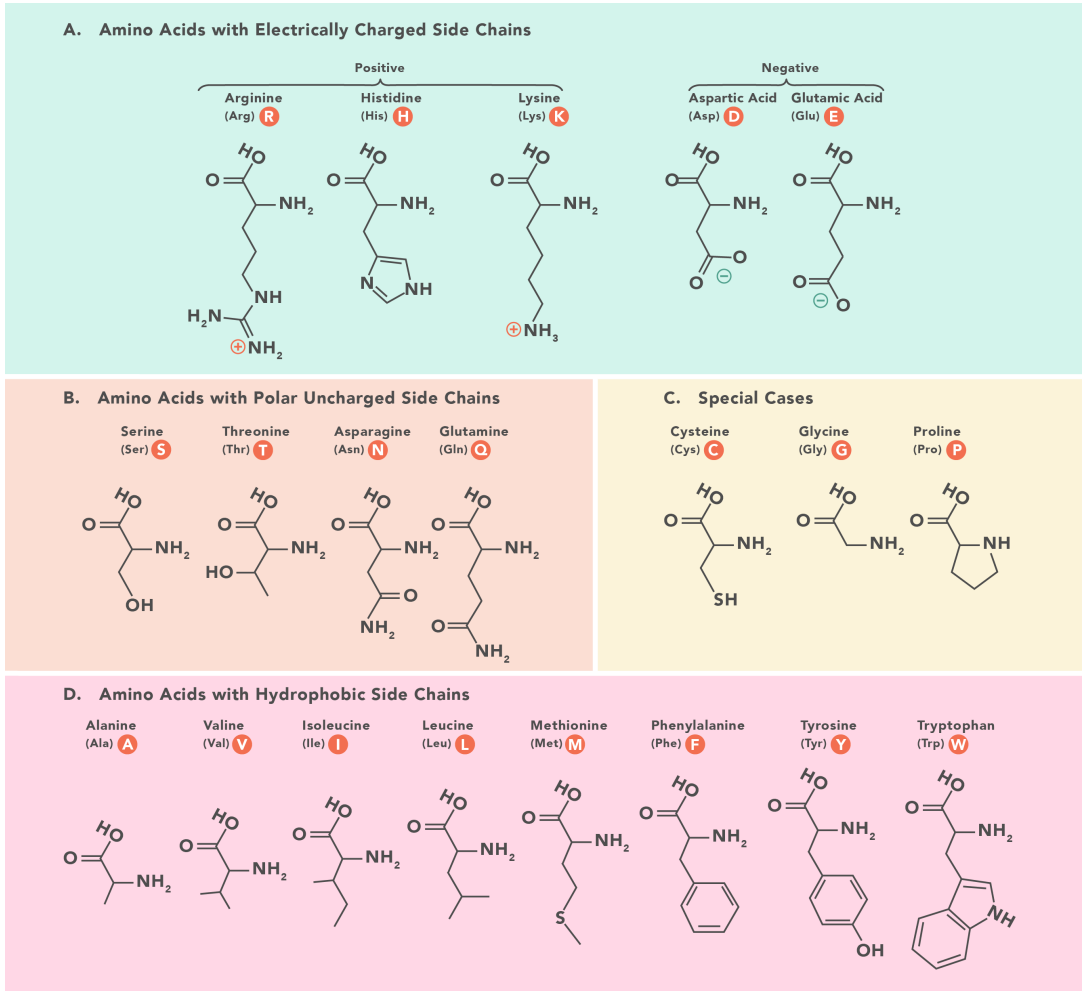
Figure 1: Diagram of data augmentations. We study randomly replacing residues (with probability p) with (a) a chemically-motivated dictionary replacement or (b) the single amino acid alanine. We also consider randomly shuffling either (c) the entire sequence or (d) a local region only. Finally, we look at (e) reversing the whole sequence and (f) subsampling to a subset of the original.

[1] Hongyu Shen, Layne C. Price, Taha Bahadori, Franziska Seeger **doi:** https://doi.org/10.1101/2021.02.18.431877
[2] French, S., Robson, B. What is a conservative substitution?. *J Mol Evol* **19,** 171–175 (1983). https://doi.org/10.1007/BF02300754

Protein/AA sequence augmentation methods

- **Replacing AA residues (with probability p) with residues of most similar chemical properties[1,2]**
  - Retains AA properties that are crucial to bindng such as. Hydrogen donor/receptors

A:V, S:T, F:Y, K:R, C:M, D:E, N:Q, V:I, P:P, G:G, W:W, H:H, T:S, Y:F, M:C, E:D, Q:N, L:L, R:K, I:V



[1] Hongyu Shen, Layne C. Price, Taha Bahadori, Franziska Seeger **doi:** https://doi.org/10.1101/2021.02.18.431877
[2] French, S., Robson, B. What is a conservative substitution?. *J Mol Evol* **19,** 171–175 (1983). https://doi.org/10.1007/BF02300754

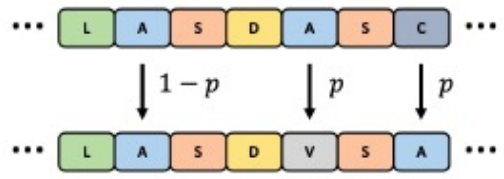
Protein/AA sequence augmentation methods with replacement with chemically similar residues

- The adaptation of replacement augmentation has following steps:
  1. Randomly select 1 position in each AA sequence (typically position 4-13, as 1-3, 14-15 are highly conserved)
  2. With probability 0.1, the residue will be replaced, chance of 0.9 it won't be replaced
  3. Replacement follows pairwise rule: *

     A:V, S:T, F:Y, K:R, C:M, D:E, N:Q, V:I, P:P, G:G, W:W, H:H, T:S, Y:F, M:C, E:D, Q:N, L:L, R:K, I:V

  4. Repeat 1-3 steps 100 times for each sequence

- Corresponding affinity value is simulated from parent log KD value plus 1% Gaussian noise ~N(0,1)



Repeat 100 times for each sequence

Affinity' = 99% * Affinity + 1% * N(0,1)

=> To simulate experiments measure variance

* note: proline (only backbone cyclic), glycine (only an H side-chain), tryptophan (indole side chain), or histidine (basic side chain with no size or chemical reaction equivalent) remains same residue

[1] Hongyu Shen, Layne C. Price, Taha Bahadori, Franziska Seeger **doi:** https://doi.org/10.1101/2021.02.18.431877

Regression with data augmentation and Res-CNN

Dataset:
- Augmented training set: 4000
  - Randomly selected from 5700 augmented traning data
  - Train: Test = 80:20

Input:
- 15 X 20 positional encoded matrix

Model:
- Res-CNN
  - Similar model architecture to task1 adapted to regression

Performance evaluation
  - Spearman rank correlation & mean squared error (MSE) (based on natural logarithm level)

Test set performance

	Rank correlation	MSE (log KD)
With augmentation	0.702	0.82
Without augmentation	NA, model not converge due to too small data points	

Regression with simple KNN exhibits comparable performance as Res-CNN

Test set performance comparison

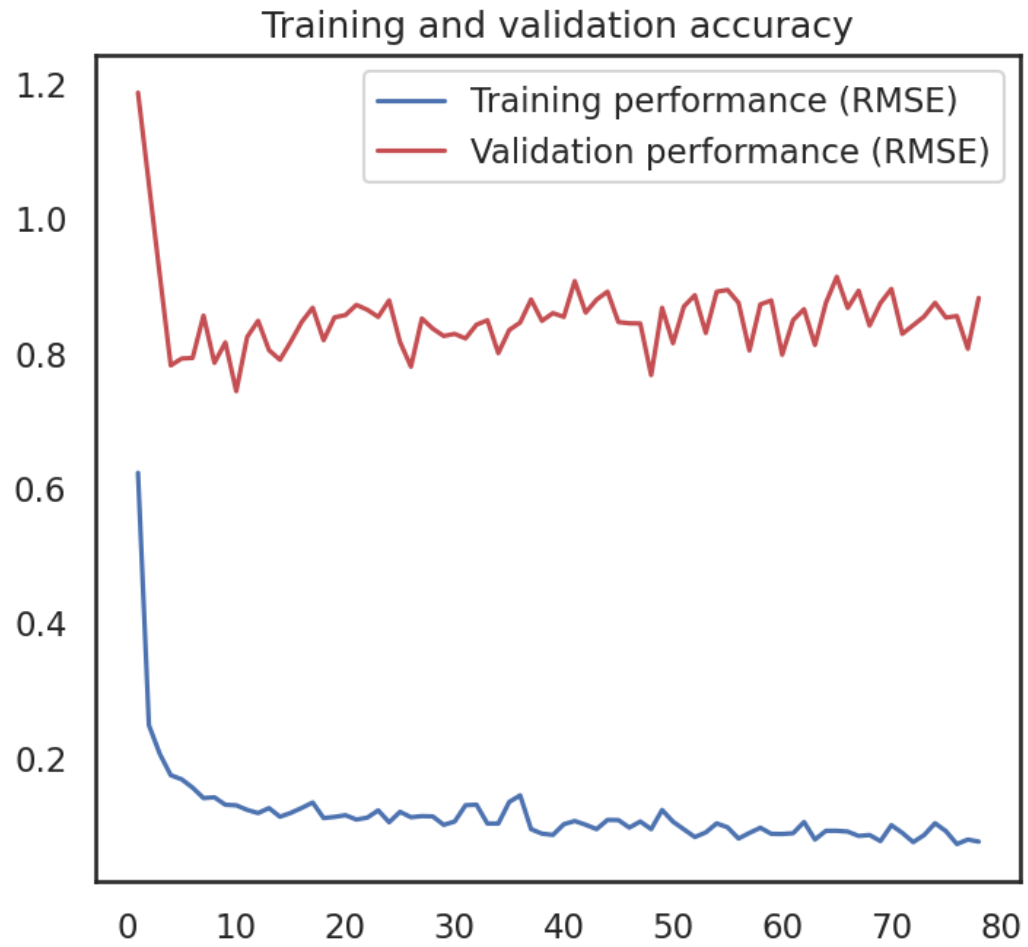
	Rank correlation	MSE (log KD)
KNN With augmentation	0.324**	1.200
Res-CNN with augmentation	**0.702****	**0.82**
KNN without augmentation	**0.690****	**0.760**
Res-CNN Without augmentation	NA, model not converge due to too small size of data points	

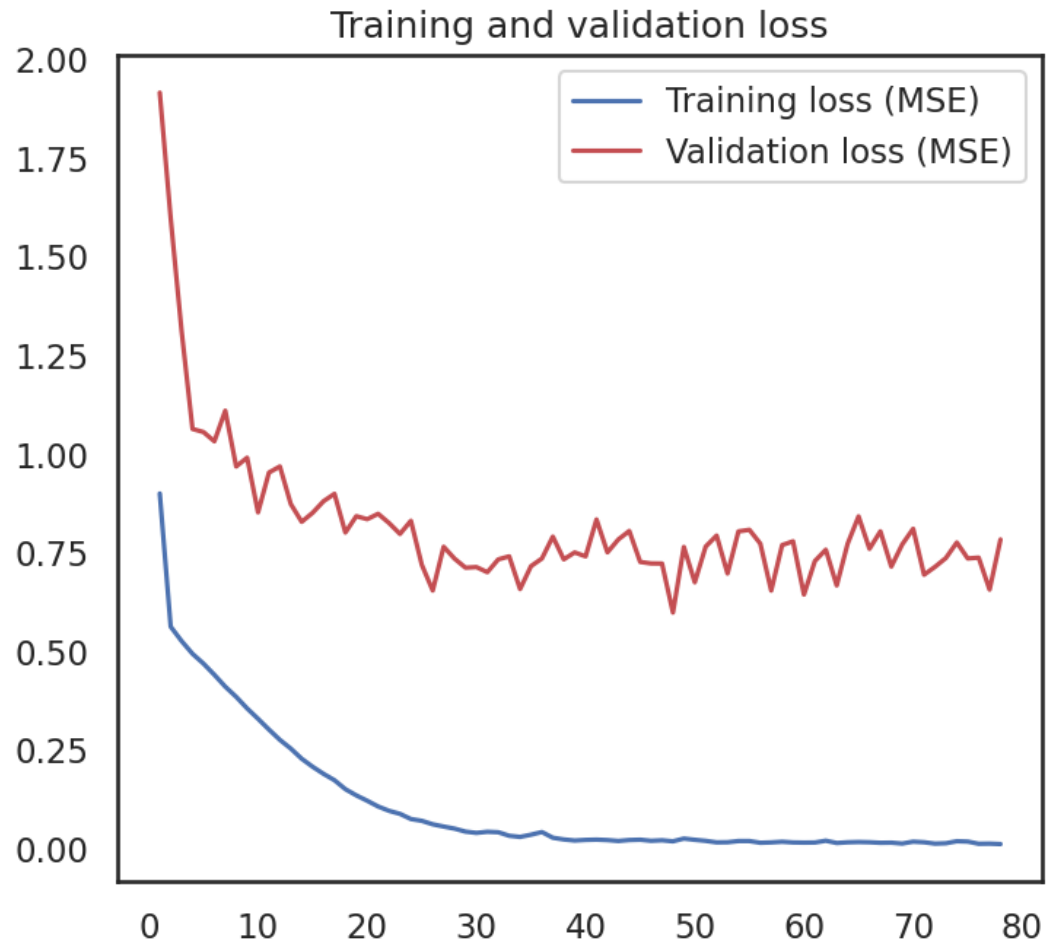
- Res-CNN trained with data augmentation shows worse MSE but slightly better rank correlation

** statisitacally significant (p-value <0.01)

Regression with Res-CNN

Res-CNN architecture training history shows training loss approaches 0 with more epochs, validation loss oscillates, perhaps due to stuck in local minimum or overfits the training set





Epochs

Regression with original data and data augmentation:

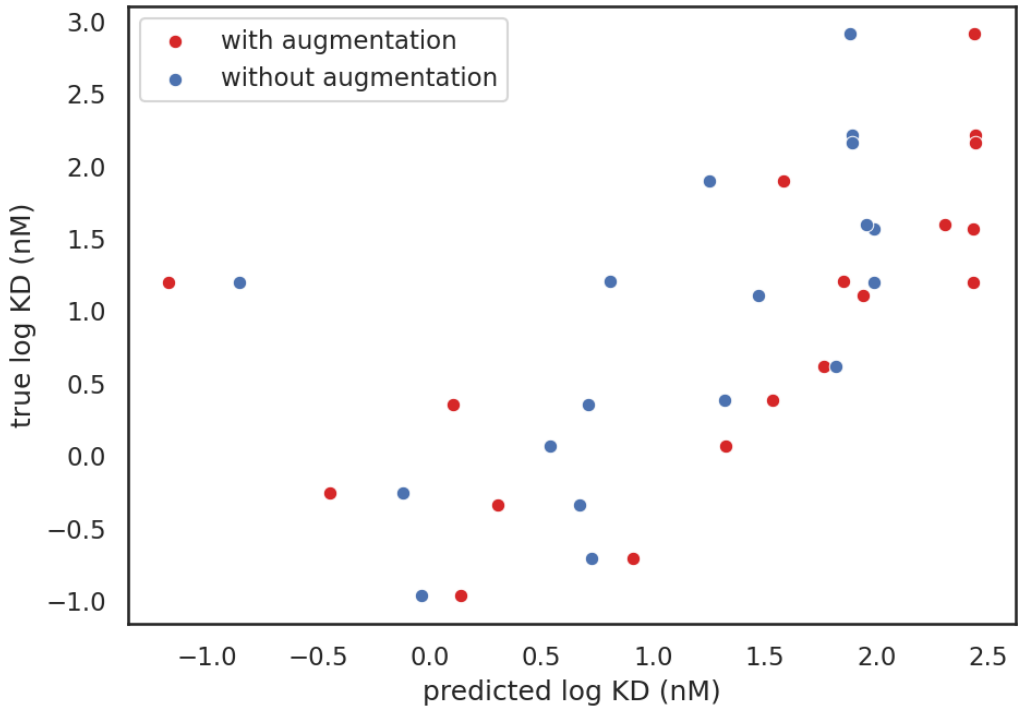
With simplest K nearest neighbor regressor (no restriction on data size)

Dataset:
- original data: 90
  - train:test = 80:20
- Augmented training set: 4000
  - Randomly selected from 5700 augmented traning data

Input:
- 15 X 20 positional encoded matrix
  - Flattened to 300-d vector for algorithmic convenience

Model:
- KNN regressor : k = 4
  - Hamming loss as distance metrics
    - Simiar to levenstein distance but normalized by length of vector
  - Parameters tuned

Performance evaluation
  - Spearman rank correlation & mean squared error (MSE) (based on natural logarithm level)



KNN performs better without data augmentation
- **predidcted value shifted by larger error with augmentation**
- **KNN can't handle noise very well, only extropolate values from distance relationships**