22125 Algorithms in Bioinformatics

Method Evaluation Using Cross Validation

Authors

Eskild Fisker Angen (s184241)

Esben Vestergaard Øyan (s194564)

Chen Chen (s210168)

Yi Huang (s210304)

21 June, 2022

Contents

1 Abstract

Constructing accurate and robust prediction tools is an essential task in many scientific fields, with good prediction scores being treated as a prerequisite for publishing machine learning studies. Because of this, the aim has in some instances become to produce the model with the highest prediction score with insufficient regard for whether this score is actually correct. In this paper we tested the predictive abilities of an Artificial Neural Network and a Stabilization Matrix Method in determining binding of peptides with 9 amino acids long to HLA class I molecules. We tested the effect of extracting an evaluation data set pre-cross-validation, as a safeguard against over-fitting and the resultant blown-up Pearson Correlation Coefficient score of a model. We found that an evaluation set has a significant effect on the prediction score of a feed-forward neural network, indicating an increase in accuracy, and that of the two tested models, the neural network performed the best. During this analysis we also found the indication that it is not the size of the data sets, but rather the number of binders present in it, that is decisive for the prediction capabilities of the model.

2 Introduction

The vertebrae immune system is one of the most complex research fields in current-day bioinformatics. One of the layers of its complexity stems from the presence of the polyallelic Major Histocompatibility Complexes (MHC). There are in humans two classes of MHCs, MHC-I and MHC-II. MHC-I is expressed on the surface of all nucleated cells and informs the T-cells of possible infections inside the cells. MHC-II is expressed on all antigen-presenting cells (APCs) to elicit immune responses via helper cells in the immune system. The MHCs present peptides from protein degradation inside the cell, or from the extracellular milieu, which can alert immune cells, and due to the highly polymorphic nature of the responsible genes, the space of presentable peptides varies significantly between individuals.

In humans MHC's are called Human Leukocyte Antigens (HLA's) HLA-I's are classified further in groups. The three most common of these are HLA-A, HLA-B, and HLA-C, each of which most humans have two types. The wide range of possible ligands for each of them, necessitates the use of machine learning methods to predict binding, a feat that is useful concerning e.g. vaccines, and doable through the use of Artificial Neural Networks (ANNs). The use of ANNs, however, requires a method of comparison, to determine the method with the best prediction rates, and the largest utility in a given area. For this the value of Pearsons' correlation coefficient (PCC) is used. The PCC is used to describe the correlation between the predicted affinity of the peptides and their measured affinity, and is as such a measure of a models accuracy. The PCC ranges from +1 to -1, indicating the severity and nature of the relationship, with +1 indicating a perfect correlation and 0 indicating random correlation. For evaluation of model predictions, a PCC between 0 and 1 is expected.

When comparing different prediction methods, it is important that the reported performance is on a test set that was not included in the cross validation optimization. It is a common problem that occurs during

data partitioning for cross validation, that a partition of the data was not held out entirely, but has been included in the training data (Alvarez et al. 2019). Doing so leads to a prediction score that is larger than the actual score of the model. We will denote this practice as "conventional" cross validation, with "true" cross validation being the model tested on data different from that on which it was trained.

3 Materials and methods

3.1 Data sets

The data chosen for this project include 35 sets of HLA binding peptides with a length of 9, of which there are 19 HLA-A alleles and 16 HLA-B alleles derived from Peters et al. 2006. The number of binders is between 29 and 1181 and that of peptides is between 59 and 3089 in different data sets. The distribution of the number of sequences and binders in the data set is shown in Figure 1, which suggests a trend that larger data sets almost contain more number of binders. The statistically significant correlation between the binder and sequence numbers suggests a t-value of 15.375 and a p-value of 2 E-16 (See Appendix 6.1 for more information). Both binder and sequence numbers were utilized to investigate their influence on the evaluated performance of the two prediction methods.

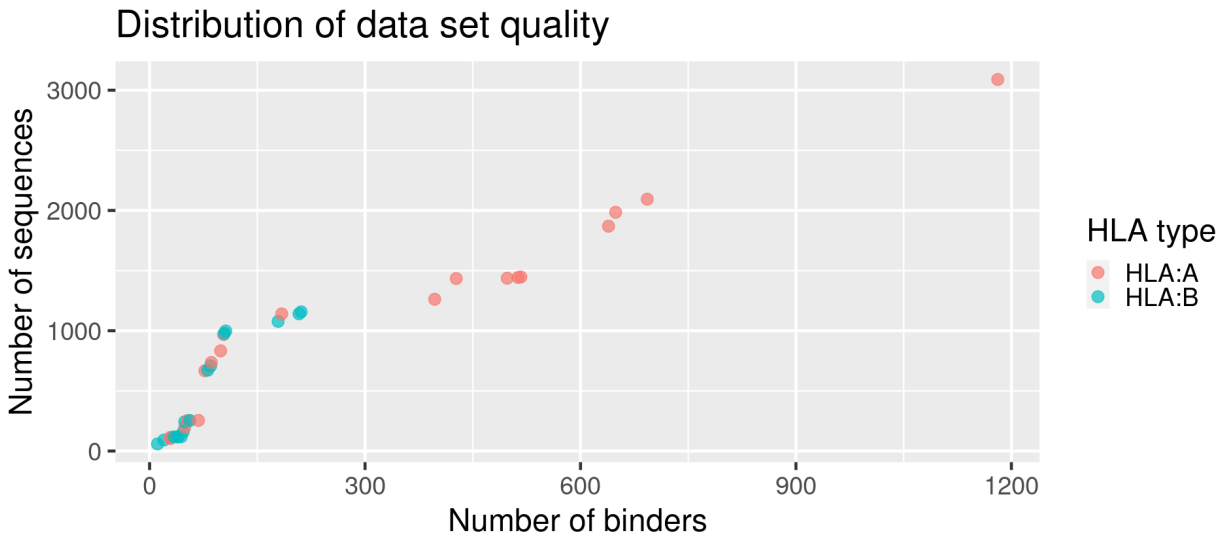

Figure 1: The distribution of the data points based on number of binders and sequences.

3.2 Stabilization Matrix Method

One method to predict binding is using the Stabilization Matrix Method (SMM), also known as ridge regression, which generates a positional weight matrix in the end for prediction. Compared with linear regression, a penalty term has been added to the cost function of the model to reduce the model complexity by shrinking the model parameters. In our case, peptides are encoded using the sparse scheme from a matrix of dimensions

9 by 20 to a vector of 1 by 180, and the model weights are a vector of 180 by 1. To optimize the model, gradient descent is used every time after the binding affinity of the input peptide has been predicted. For each target i, the error function is shown in Eq. 3.1, where O_i is the prediction, t_i is the measurement, λ is the penalty parameter, N is the number of the whole peptides in the data set, and w_l is the l^{th} model parameter. Gradient descent for each iteration is based on calculating the partial derivative of error E to the weight for each position.

$$E_i = \frac{1}{2} \cdot (O_i - t_i)^2 + \frac{\lambda}{N} \sum_l w_l^2 \tag{3.1}$$

3.3 Artificial neural networks

ANNs are a competitive machine learning model for classification and prediction. Due to its excellent properties of adaptivity, nonlinearity, self-learning, and advancement in input to output mapping, it is implemented to a large extent in universal function approximation. An ANN can function in a similar way to how the human brain performs information processing upon receiving a complex signal input. The main task of the brain or ANN is to design a unique model for processing the information. Evaluating ANN models is readily done with cross-validation, which is already covered and can be done correctly and incorrectly (Abiodun et al. 2018).

One type of ANN is the Feed Forward Neural Network (FFNN). It consists of an input layer, several hidden layers and an output layer where each layer contains certain units. In an FFNN, each unit in a layer is related to all other units in that layer, and the weight is usually unequal for the unit-unit connection in which each unit is from different layers. The weights are used to measure the potential knowledge of the network to allow information to be processed by flowing from the input layer to the output layer by passing through the hidden layers uniquely. To find the correct combination of the weights, the Backwards Propagation Neural Network is introduced to modify the weights by minimizing the cost function (Hecht-Nielsen 1992; Rojas 1996).

3.4 Five-fold cross validation

Cross-validation is a method to evaluate the learning algorithms by splitting the data into sets of equal sizes, typically segmented into sets used to train the model and others which are used to validate the model. Each data point is validated by comparing the training and validation sets iteratively.

The basic form of cross-validation is k-fold cross-validation, which means the data is partitioned into k equally or nearly equally sized sets with comparable data distributions (Refaeilzadeh et al. 2009). We have chosen five-fold cross-validation in our work.

3.4.1 Conventional cross validation

In conventional cross validation, the dataset ($\boldsymbol{D}$) is split into $\boldsymbol{K}$ partitions and $\boldsymbol{K}$ iterations of training and testing will be performed so each fold of the data acts as test set($\boldsymbol{R}$) for the specific model trained on the

remaining k-1 folds of data ($\boldsymbol{T}$) in successive rounds. The performance of the the learning algorithms can be presented by the accuracy, such as PCC (Refaeilzadeh et al. 2009).

This approach results in $\boldsymbol{K}$ different models, each of which has a different weighting. According to the "wisdom of crowds" sentiment, the $\boldsymbol{K}$ different models can then be applied to a data-set, and the biases of the $\boldsymbol{K}$ different models will be somewhat reduced by using the mean of the predictions actual score.

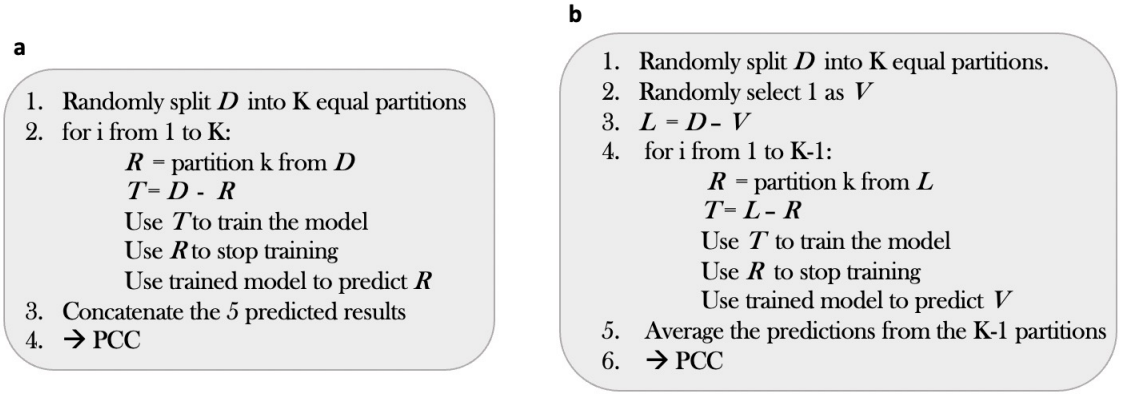

Figure 2: Cross validation.

(a) Conventional cross validation (b) Nested cross validation

3.4.2 Nested cross validation

Nested cross-validation (or true cross-validation) is a special case of k-fold cross validation made by splitting the data ($\boldsymbol{D}$) into 3 subsets: training, testing and evaluation sets. The fixed evaluation set ($\boldsymbol{V}$) is used after training each model to test prediction on not previously seen data.

For the $\boldsymbol{K}$-1 iterations, The training set ($\boldsymbol{T}$) is used to train the model with different parameters, and then the trained model with specific parameters will be evaluated on the test set ($\boldsymbol{R}$) which applies early stopping to avoid overfitting with an iterative method. After determining the optimal parameters, the model is finally tested on the fixed evaluation set, and the performance is the average for all the $\boldsymbol{K}$-1 models (Refaeilzadeh et al. 2009; Ghojogh and Crowley 2019).

3.5 Statistical test

The linear model was introduced for modeling the relationship between the performance of the prediction methods and either the number of binders or the number of sequences. In addition, The paired t-test and the Binomial test were both utilized in our project to evaluate whether there is a significant difference in the way of cross-validation for both ANN and SMM models. The typical steps for applying a t-test are proposing the null hypothesis, establishing the test level or the significance level by having the alpha level and standardizing the sample to get the T statistic and obtain the P-value. We set $p - value = 0.05$ to be the cut-off for all the tests we did in our analyses.

4 Results

4.1 Comparison between conventional and true cross-validated performance

The comparative performance of ANN and SMM on the data we selected is shown in the figures 3, 4, 5, and 6. We observe an overall increasing trend of PCC values regardless of model type and whether the models are ranked based on the number of binders or number of sequences in the data.

Figure 3 and 4 illustrate the performance of the ANNs, sorted either by the number of binders or the number of sequences in the data set. There is an observable trend that the PCC is larger in the case of conventional cross-validation compared to true cross-validation, in correspondence with what would be expected.

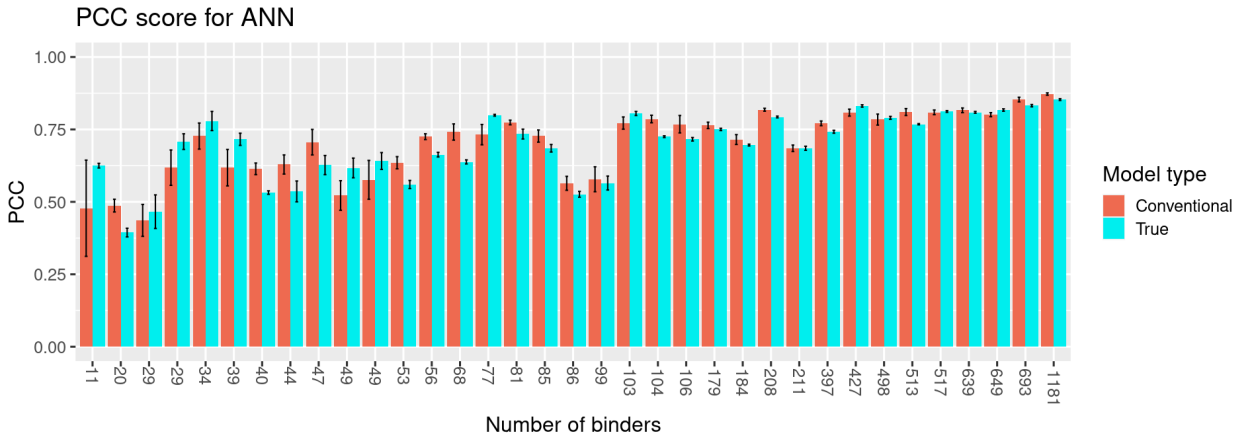

Figure 3: PCC for an ANN model trained with both conventional and true cross validation, ordered by number of binders. Error bars are calculated as mean error.

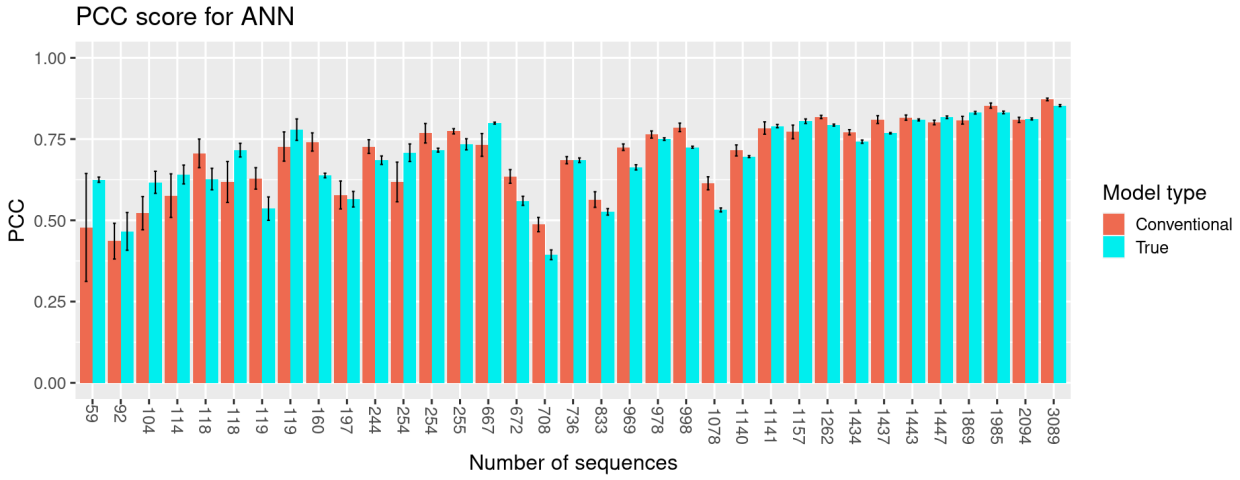

Figure 4: PCC for an ANN model trained with both conventional and true cross validation, ordered by number of sequences. Error bars are calculated as mean error.

Figure 5 and 6 show the calculated performance of the SMMs ordered by the number of binders and sequences, respectively. These figures also show the trend of increasing PCC with increasing sequences or binders.

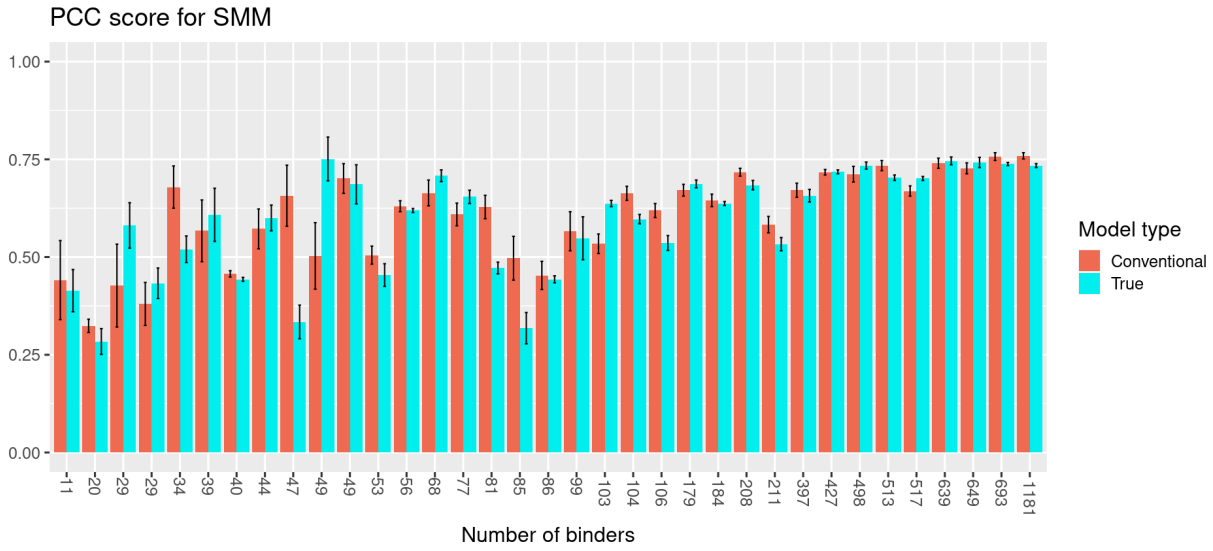

Figure 5: PCC for an SMM model trained with both conventional and true cross validation, ordered by number of binders. Error bars are calculated as mean error.

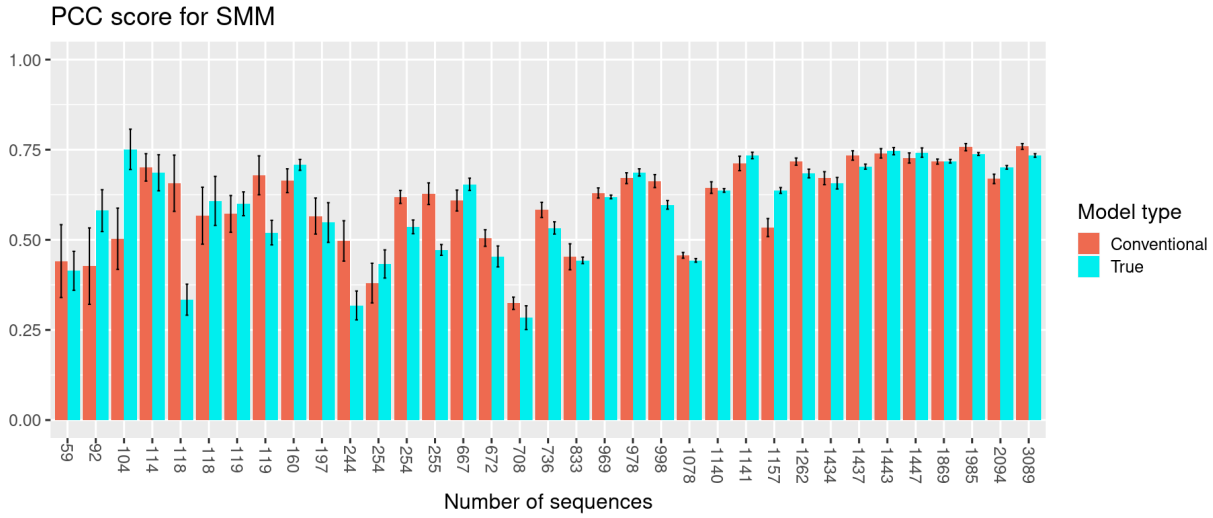

Figure 6: PCC for an SMM model trained with both conventional and true cross validation, ordered by number of sequences. Error bars are calculated as mean error.

	Number of binders		Number of sequences	
Model	t-value	p-value	t-value	p-value
ANN True	4,682	4,70 E-5	4,083	2,65 E-4
SMM True	4,133	2,30 E-4	3,428	1,65 E-3
ANN Conventional	4,986	1,92 E-5	5,136	1,23 E-5
SMM Conventional	4,824	3,10 E-5	3,571	1,12 E-3

Table 1: The result of a comparison of the models in figures 3, 4, 5, and 6 performed as a Students' t-test in R©. The null hypothesis was that there was no correlation between the PCC of the model in question and either the number of binder or sequences.

Table 1 is a summation of the significance of the models trained. It was found that the PCC values of all models had a correlation with both the number of binders and sequences in a statistically significant manner with p-values less than 0.05. The models trained with conventional cross-validation had a more significant correlation both when considering the number of binders or sequences as evaluation parameters for both ANN and SMM.

Considering only the models using true cross-validation, all were found to be statistically significant, but with stronger proof for the models when testing on the number of binders than the number of sequences. The correlation with the number of sequences might be derived from the trend observed in figure 1, where we observed that a higher number of sequences follow a higher number of binders.

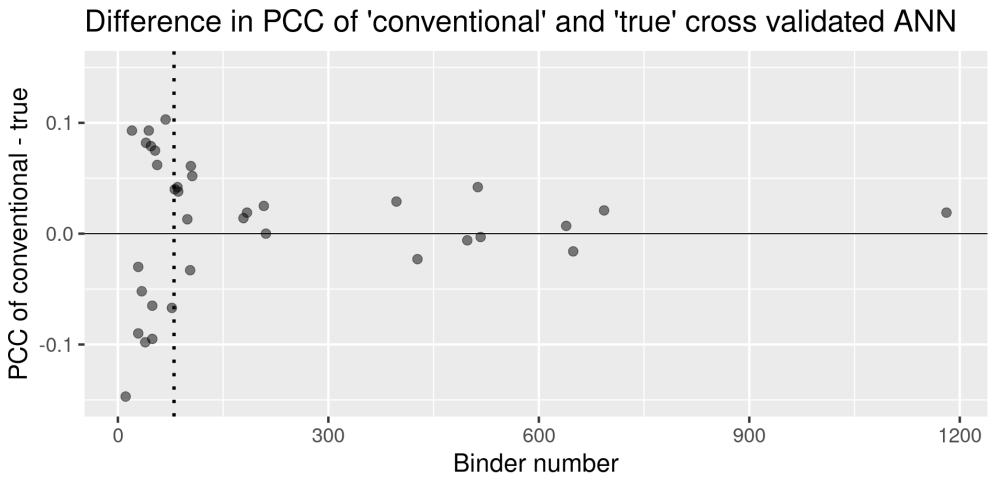

Figure 7: A comparison of the PCC values of the true and conventional ANN models illustrating a decrease in difference at higher number of binders. The difference is calculated as the PCC of the true cross validated model substracted from the PCC of the conventionally cross validated model.

Figure 7 shows the changing difference in the PCC of the true and conventionally cross-validated ANN model as a function of the number of binders. At low values of binders, the differences fluctuate to a high degree, which is why we have inserted the dotted line which is the cutoff applied in table 2.

A regression considering the top-20 points was made, resulting in a trendline with negative gradient of -2 E-5 and an R squared value of 0.073. Given the very small gradient and R value there cannot be drawn a conclusion on whether there is a decrease in difference between conventional and true with increasing binder numbers.

Paired t.test	Model 1	Model 2	t-value	p-value
Top 20 no. binders	ANN Conventional Binders	ANN True Binders	3,0159	0,007107
Top 20 no. binders	SMM Conventional Binders	SMM True Binders	1,7792	0,09121
Top 20 no. sequences	ANN Conventional Sequences	ANN True Sequences	3,1994	0,00472
Top 20 no. sequences	SMM Conventional Sequences	SMM True Sequences	1,0733	0,2966
Binomial	Model 1	Model 2		p-value
Top 20 no. binders	ANN Conventional Binders	ANN True Binders		0,1153
Top 20 no. binders	SMM Conventional Binders	SMM True Binders		0,2632
Top 20 no. sequences	ANN Conventional Sequences	ANN True Sequences		0,1153
Top 20 no. sequences	SMM Conventional Sequences	SMM True Sequences		0,2632

Table 2: Difference in prediction values of models as found with a paired t-test and a Binomial test. In either case the 15 least explained data points were omitted. The t-test derived p-values for both ANN comparisons indicate significant difference, while neither of the SMM comparisons do. The binomial tests indicate no significant difference.

In Table 2 the results of paired t-tests and binomial tests on the different versions of conventional and true models are explored. The results show the ANN-true-models to be significantly different in evaluated PCC when compared to the conventional method, but only when t-testing, and the same conclusion was not reached with a binomial test. Only the top 20 data points, i.e. the results of models on data sets with the 20 highest binder numbers or sequence space, are presented here, as the bottom 15 data points were deemed to add too much noise.

Due to these points, and the artificially high PCC values of the conventionally cross-validated models, we choose to continue working only with the models based on true cross validation.

4.2 Comparison between method performance using the true cross validation

In figure 8 we have compared the ANN and SMM models that were cross-validated using true cross-validation and based on the number of binders. The comparison shows that the ANN model has an overall higher performance, with a paired t-test derived p-value of 1.325 E-4, and that both models increase in performance with additional binders. At a low level of binders, both models have predictions that vary wildly and do not seem to be following a clear trend when considering data with fewer than 250 binders.

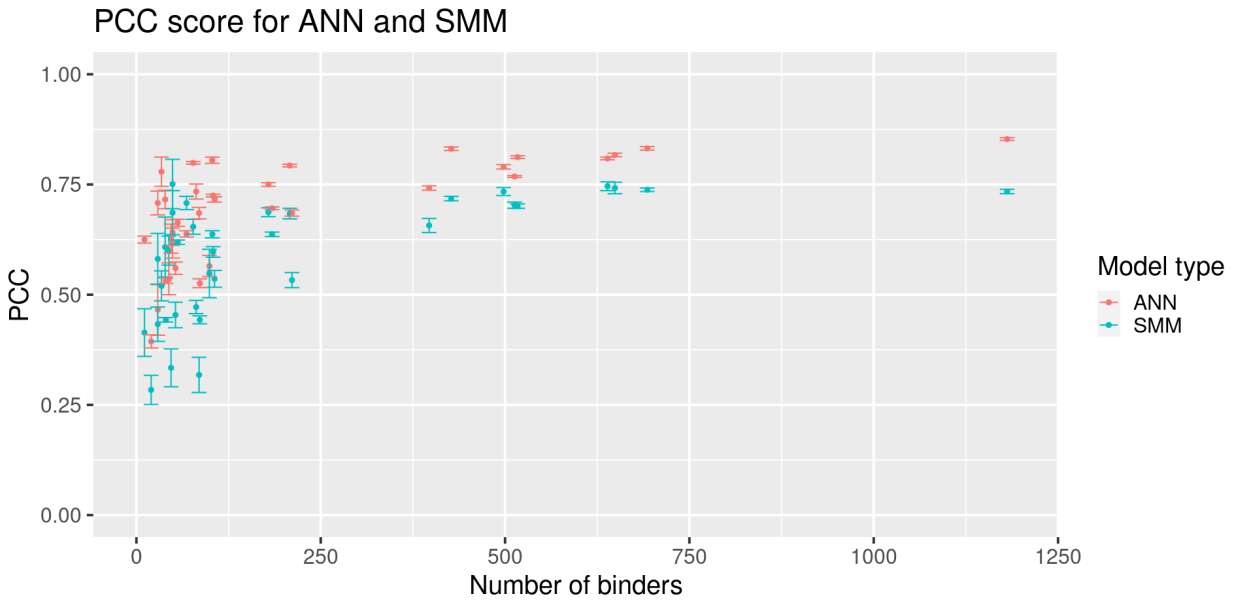

Figure 8: Pearson correlation coefficient (PCC) for an ANN and SMM model trained on the same data using an evaluation data set. The ANN model performs significantly better with a paired t-test derived p-value of 1.325e-4.

5 Discussion

We have found that there is a statistically significant relationship between the prediction scores of the models that we have tested and both the number of binders and sequences in the data they were trained on. This relationship is less pronounced when considering the number of sequences, and this correlation might be entirely the result of the correlation observed in Figure 1.

The different outcomes of the paired t-test and binomial test in Table 2, regarding the significance of the conventional ANN model vs the true ANN model, seem conflicting at first but might convey some deeper connection. While the binomial test only takes into account whether the conventional model scores higher or not, much like a coin-flip, the paired t-test also considers the numerical difference in PCC score. This indicates the importance, not of consistently higher scores, but the consistently well-defined difference in the scores. That the reliable difference between the two scores is only pronounced for the top half of the data set was not what we expected, but aligns with what we know. The volatility of the two models' PCC scores when trained on smaller data sets makes sense because the small number of binders would give biased training data. The resulting models would be very hit-or-miss, with unreliable PCC evaluations, and thus the random distribution is not surprising.

What is somewhat surprising is the results exemplified in Figure 7. The decrease in difference between conventional and true prediction scores with the increase of binder number is not as expressed as expected. While the regression line is negative, it is negligible, and thus the results on whether the difference decreases as a function of number of binders is inconclusive.

Using conventional cross-validation is bad scientific practice causing the performance of models trained this way will be artificially higher than the actual performance of the model. Understanding the models that one uses is essential to be able to apply them correctly and make valid conclusions. As a result of this, we consider the most efficient tool for predicting would be an ANN model trained with regards to the number of binders using true cross-validation.

While there is a significant difference between the ANN models when using the t-test, there is no significance regarding SMM. This could be a result of an ANN's inherent ability to confer higher-order correlations. This is supported further in Figure 8, where the significant difference between SMM and ANN performance when compared to each other. In other words, the use of true cross-validation might carry more impact on neural network models due to their more sophisticated structure.

6 Appendix

6.1 Data sets information

HLA allele	**No. peptides**	**No.binders**	**HLA allele**	**No. peptides**	**No.binders**
HLA-A0101	1157	103	HLA-A6901	833	86
HLA-A0201	3089	1181	HLA-B0702	1262	208
HLA-A0202	1447	649	HLA-B0801	708	20
HLA-A0203	1443	639	HLA-B1501	978	179
HLA-A0206	1437	513	HLA-B1801	118	47
HLA-A0301	2094	517	HLA-B2705	969	56
HLA-A1101	1985	693	HLA-B3501	736	211
HLA-A2301	104	49	HLA-B4001	1078	40
HLA-A2402	197	99	HLA-B4002	118	39
HLA-A2403	254	29	HLA-B4402	119	44
HLA-A2601	672	53	HLA-B4403	119	34
HLA-A2902	160	68	HLA-B4501	114	49
HLA-A3001	669	77	HLA-B5101	244	85
HLA-A3002	92	29	HLA-B5301	254	106
HLA-A3101	1869	427	HLA-B5401	255	81
HLA-A3301	1140	184	HLA-B5701	59	11
HLA-A6801	1141	498	HLA-B5801	988	104
HLA-A6802	1434	397			

Table 3: The selected datasets of the alleles with the total peptide numbers and the binder numbers.

References

Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A., and Arshad, H. (2018). "State-of-the-art in artificial neural network applications: A survey". In: *Heliyon* 4.11, e00938. ISSN: 2405-8440. DOI: https://doi.org/10.1016/j.heliyon.2018.e00938. URL: https://www.sciencedirect.com/science/article/pii/S2405844018332067.

Alvarez, B., Reynisson, B., Barra, C., Buus, S., Ternette, N., Connelley, T., Andreatta, M., and Nielsen, M. (2019). "NNAlign-MA; MHC peptidome deconvolution for accurate MHC binding motif characterization and improved t-cell epitope predictions". In: *Molecular and Cellular Proteomics* 18 (12), pp. 2459–2477. ISSN: 15359484. DOI: 10.1074/MCP.TIR119.001658.

Ghojogh, B. and Crowley, M. (2019). "The theory behind overfitting, cross validation, regularization, bagging, and boosting: tutorial". In: *arXiv preprint arXiv:1905.12787*.

Hecht-Nielsen, R. (1992). "Theory of the backpropagation neural network". In: *Neural networks for perception*. Elsevier, pp. 65–93.

Peters, B., Bui, H.-H., Frankild, S., Nielsen, M., Lundegaard, C., Kostem, E., Basch, D., Lamberth, K., Harndahl, M., Fleri, W., et al. (2006). "A community resource benchmarking predictions of peptide binding to MHC-I molecules". In: *PLoS computational biology* 2.6, e65.

Refaeilzadeh, P., Tang, L., and Liu, H. (2009). "Cross-validation." In: *Encyclopedia of database systems* 5, pp. 532–538.

Rojas, R. (1996). "The backpropagation algorithm". In: *Neural networks*. Springer, pp. 149–182.