

GEFÖRDERT VOM



Bundesministerium  
für Bildung  
und Forschung

## **Zuwendungsempfänger**

Technische University München

## **Thema der Förderung**

**CompLS - Runde 4 - Verbundprojekt: GENImmune - Generische Methoden zu  
Modellierung von Host-Pathogen Interaktionen angewandt zur effizienten  
Immunisierung gegen SARS-CoV-2 - Teilprojekt E**

## **Verantwortlicher**

Prof. Dr. Dmitrij Frishman

## **Förderkennzeichen**

031L0292E

# Schlussbericht zum Forschungsprojekt

## „Machine learning for epitope optimization“

Teilprojekt E	Machine learning for epitope optimization
Förderkennzeichen	031L0292E
Ausführende Stelle	Department of Bioinformatics School of Life Sciences Technische Universität München
Projektleiter	Prof. Dr. Dmitrij Frishman
Gesamtförderbetrag	€ 189.370,56
Laufzeit	01.01.2022-31.12.2023

## Vorhabendurchführung

### **Biological background, state of the art, and motivation for the project work**

B-cells provide long-term immunity against cancerous cells and pathogens/antigens and hence, are a vital component of the adaptive immune system. B-cells get activated when the B-cell receptors, a transmembrane protein present on the B-cell surface, interact with the B-cell epitopes, a specific portion of the antigen. B-cell epitopes can be classified into two categories – linear or continuous epitopes and conformational or discontinuous epitopes. Linear epitopes are sequential fragments of the antigen's protein sequence. Widely-used less stringent definitions of linear epitopes allow the existence of a small number of non-epitope residues within the continuous segment. Discontinuous epitopes are composed of several fragments distributed along the antigen's protein sequence which are brought together in spatial proximity due to protein folding. Although almost 90% of the B-cell epitopes are discontinuous, most of these consist of at least a few continuous segments.

Recent advances in the field of machine learning led to the development of models trained on huge datasets of protein sequences and structures that can be utilized for an accurate prediction of local and global structural features from the protein amino acid sequences only. Such models, termed as the protein Language Models (pLMs), generate numerical embeddings from the protein sequences. These numerical values or feature vectors can be exploited during training different machine learning algorithms for several prediction tasks related to biological research. Recently developed widely-used protein language models include ESM-2 and ProtTrans. The strong role played by pLM in B-cell epitope prediction has already been shown in a recent publication where the authors developed the highly performing Bepipred-3.0 using ESM-2 embeddings on the protein structure information. Antibody epitope prediction utilizing ProtTrans pLM has not been tested yet.

In this project we developed BLMPred for predicting linear B-cell epitopes by utilizing the embeddings generated by the ProtTrans protein language model. BLMPred has been trained only on peptide sequence data and protein structure information is not needed for using the model. BLMPred has been developed by training a Support Vector Machine on the ProtTrans generated numerical representations of the input dataset. BLMPred has an excellent predictive capability as evident from the high values of the performance metrics.

### **Dataset of linear B-cell epitopes**

We downloaded 208265 experimentally validated linear B-cell epitope sequences (positive samples) and 487127 non-B-cell epitope sequences (negative samples) from the Immune Epitope Database. We removed 1075 duplicate entries, 215 peptides containing non-standard amino acids symbols (Z, B, J, O, U, X), 512 peptides with a length less than the typical minimum length of 5 amino acids for linear B-cell epitopes, as well as 95971 peptides found both in the positive and the negative dataset. Furthermore, all peptides in the positive dataset longer than 60 amino acids were excluded from consideration since the negative dataset contained peptides with the lengths of up to 60 amino acids. This dataset was named BLMPred\_5to60.

Recent reports suggest that the length of B-cell epitopes varies between 5-8 and 25 amino acid residues. This length range is dictated by the structural requirements of epitope binding to the Complementarity Determining Regions (CDRs) of the B-cell receptors. According to the INDI database, CDR1, CDR2, and CDR3 vary in lengths between 4-17, 5-17, and 5-38 amino acids, respectively. We therefore created an alternative dataset, BLMPred\_8to25, consisting of peptides varying in length between 8 and 25. After data cleaning steps, BLMPred\_5to60 and BLMPred\_8to25 contained 111015 (390589) and 102023 (387155) positive (negative) samples, respectively.

### **Preparation of training and test datasets**

Since the BLMPred\_5to60 and BLMPred\_8to25 datasets were initially imbalanced, with the positive to negative sample ratio of 1:3, we made them balanced by drawing 111015 and 102023 negative

samples from the pool of 390589 and 387155 negative samples in these two datasets, respectively. Additionally, we ensured that the sequence length distribution of these negative samples closely matched that of the positive samples. Both datasets were split into training (90%) and test (10%) datasets while retaining similar length distributions. The training and test datasets constructed from BLMPred\_5to60 (BLMPred\_8to25) datasets are referred to as BLMPred\_5to60\_training (BLMPred\_8to25\_training) and BLMPred\_5to60\_test (BLMPred\_8to25\_test), respectively. The training datasets were used for cross-validation while the test datasets were solely used as independent datasets for assessing the performance of the final trained methods.

### Preparation of benchmarking dataset

We prepared a separate independent dataset (BLMPred\_benchmark) for comparing the performance of BLMPred with other existing tools. Since the training and test datasets described above are based on the IEDB release of March 2023, we downloaded sequences of 1105 linear B-cell epitopes and 822 non-B-cell epitopes deposited with IEDB between April, 2023 and August, 2023. In spite of the different submission dates, we nevertheless identified and removed 70 B-cell epitopes and 159 non-B-cell epitopes that were also found in BLMPred\_5to60 or BLMPred\_8to25, so that the final BLMPred\_benchmark dataset contained 1035 positive and 663 negative samples.

### Language Model (LM) embeddings

For each peptide in our dataset, we generated embeddings of length 1024 by utilizing the ProtT5-XL-U50 model of the ProtTrans Protein Language Model.

### Machine learning

Identification of linear B-cell epitopes was cast as a binary classification problem where an input peptide was either a B-cell epitope or not. A broad range of traditional machine learning models implemented in the Scikit-learn package was tested, including adaboost classifier, bagging classifier, extra trees classifier, Gaussian Naïve Bayes, histogram-based gradient boosting classifier, k-nearest neighbors, linear discriminant analysis, logistic regression, multi-layer perceptron, quadratic discriminant analysis, random forest, and support vector machine. Also, we tested the XGBoost classifier from the XGBoost package on the Python language platform. Furthermore, we trained an Explainable Boosting Machine (EBM) learning classifier provided by an open source Python package, InterpretML. We utilized RAPIDS, a data science framework capable of executing end-to-end pipelines completely in the GPU for an accelerated training of the machine learning models.

### Performance evaluation metrics

In order to assess the model performance on the test dataset, we calculated several performance metrics including accuracy (ACC), precision (P), sensitivity or recall (R), F1 score (F1), specificity (S), Matthew's Correlation Coefficient (MCC), area under the ROC curve (AUROC), and the area under the precision-recall curve (AUPRC) as follows:

$$\begin{aligned}
 ACC &= \frac{TP + TN}{TP + TN + FP + FN} \\
 P &= \frac{TP}{TP + FP} \\
 R &= \frac{TP}{TP + FN} \\
 F1 &= \frac{2 \times P \times R}{P + R} \\
 S &= \frac{TN}{TN + FP} \\
 MCC &= \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}
 \end{aligned}$$

where TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively. Among all the evaluation metrics, MCC has been reported to be more informative in evaluating binary classification problems. Hence, although we report our results based on the full set of metrics, we selected the trained model with the highest MCC measure as the optimal classifier for further processing.

### **Performance of models trained on the BLMPred\_5to60 dataset**

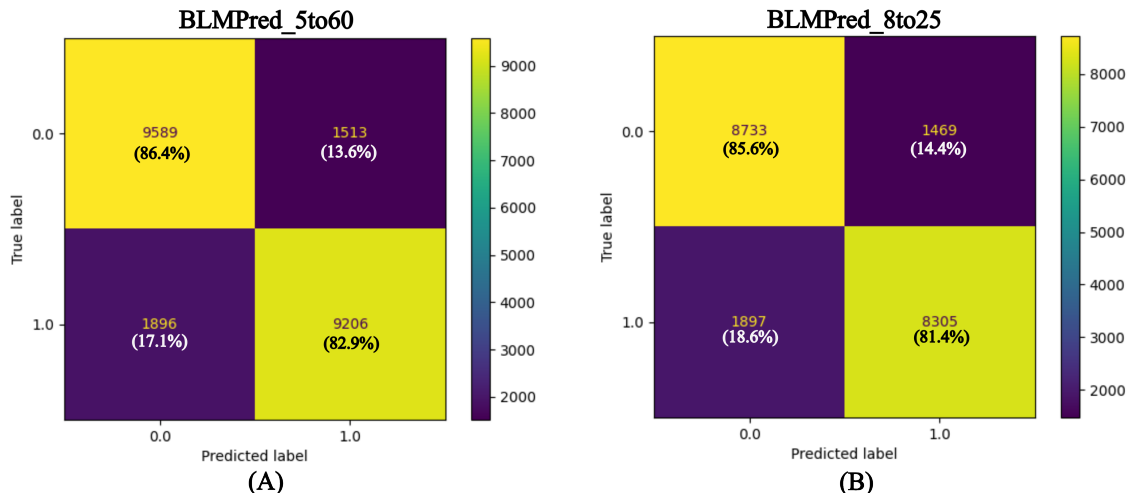
Machine learning models listed in Section 2.4 were trained on the embeddings derived from the BLMPred\_5to60 dataset and their performance was assessed by 10-fold cross validation. The best results were achieved with Support Vector Machine, with the mean±std values of accuracy, precision, recall, F1 score, specificity, MCC, AUROC and AUPRC of 0.839±0.0027, 0.849±0.0046, 0.824±0.0037, 0.837±0.0033, 0.855±0.0041, 0.679±0.0055, 0.839±0.0028, and 0.788±0.0047, respectively. For each classifier, the 10 trained models resulting from each fold of the 10-fold cross-validation were tested on the independent BLMPred\_5to60\_test dataset and SVM outperformed all other classifiers. Thus, when utilizing ProtTrans embeddings as input, SVM was clearly the best performing method on the BLMPred\_5to60 dataset. In general, we found that none of the models were overfitted and their results were quite robust, as evidenced by the low standard deviation values of the performance metrics.

### **Performance of models trained on the BLMPred\_8to25 dataset**

SVM was also the best performing method among the machine learning models trained on the embeddings derived from the BLMPred\_8to25 dataset. Its accuracy was around 5% higher than that of the next best performing model, k-nearest neighbors. SVMs trained on the BLMPred\_8to25 and BLMPred\_5to60 datasets achieved a similar performance. This is not surprising as the BLMPred\_8to25 and BLMPred\_5to60 strongly overlap: only approximately 9% of the BLMPred\_5to60 dataset is constituted by B-cell epitopes with lengths outside of the 8-25 range which have been eliminated to create the BLMPred\_8to25 dataset.

### **BLMPred models**

Based on the model assessment presented above, we selected SVM for further analyses. SVM trained on the entire BLMPred\_5to60\_training and BLMPred\_8to25\_training datasets will be referred to as BLMPred\_5to60 and BLMPred\_8to25 models, respectively. The BLMPred\_5to60 model, when tested on the independent BLMPred\_5to60\_test dataset, exhibited the accuracy, precision, recall, F1 score, specificity, MCC, AUROC and AUPRC of 0.846, 0.859, 0.829, 0.844, 0.864, 0.693, 0.846, and 0.798, respectively. Similarly, the values of accuracy, precision, recall, F1 score, specificity, MCC, AUROC and AUPRC achieved by the BLMPred\_8to25 model when tested on the independent BLMPred\_8to25\_test dataset were 0.835, 0.849, 0.814, 0.831, 0.856, 0.671, 0.835, and 0.785, respectively. The BLMPred\_5to60 and BLMPred\_8to25 models accurately predicted 82.9% (86.4%) and 81.4% (85.6%) of the B-cell epitopes (non-B-cell epitopes) present in the BLMPred\_5to60\_test and BLMPred\_8to25\_test datasets, respectively, with low Type I and Type II error levels (Figure 1).



**Figure 1.** Confusion matrices obtained (A) on the independent BLMPred\_5to60\_test dataset by the BLMPred\_5to60 model, and (B) on the BLMPred\_8to25\_test dataset by the BLMPred\_8to25 model.

**Performance of BLMPred compared with the reported performance of other linear B-cell epitope prediction tools**

For a detailed comparison of model performance, we selected 12 freely available linear B-cell epitope prediction tools that we were able to install and execute: Bepipred-3.0, Bepipred-2.0, Bepipred-1.0, Chou and Fasman beta turn prediction-based method, Emini surface accessibility scale-based method, Karplus and Schulz flexibility scale-based method, Kolaskar and Tongaonkar antigenicity scale-based method, Parker hydrophilicity prediction-based method, Support Vector Machine based on Tri-peptide similarity and Propensity scores (SVMTriP), Linear B-Cell Exact Epitope Predictor (LBEEP), epitope1D, and EpiDope. Among these tools, Chou and Fasman beta turn prediction, Emini surface accessibility scale, Karplus and Schulz flexibility scale, Kolaskar and Tongaonkar antigenicity scale, Parker hydrophilicity prediction, Bepipred-1.0, and Bepipred-2.0 have been implemented and hosted by the Immune Epitope Database Analysis Resource (IEDB-AR). A detailed summary of these tools is presented in Table 1, including the specific algorithms, datasets and features utilized as well as the performance metrics reported in the corresponding original publications. According to these data epitope1D reportedly outperforms other methods in terms of MCC and AUROC, while our method, BLMPred\_5to60, performs better than all existing tools in terms of accuracy, precision, recall, specificity, F1 score, and AUPRC. We attribute this high performance level of BLMPred\_5to60 to a large number of up-to-date experimentally verified linear B-cell epitope data collected from IEDB, extensive data filtering, and the utilization of ProtTrans embeddings.

**Table 1.** Detailed summary of the computational tools considered in our comparative study. Performance metrics of the tools as reported in their respective publications are presented, with the best values highlighted in bold.

Model	Minimum epitope length supported	Maximum epitope length supported	Algorithm	Dataset	Features	Reported performance							
						Accuracy	Precision	Recall	MCC	Specificity	F1 score	AUROC	AUPRC



epiDope	NA	NA	Deep neural networks	IEDB	ELMo Embeddings	NA	NA	NA	NA	NA	NA	0.625	NA	
epitope1D	6	NA	Random Forest and Explainable Boosting Machine	IEDB	Graph-based signature representation of protein sequences and organism ontology information	NA	NA	NA	<b>0.72</b>	NA	NA	NA	<b>0.93</b>	NA
LBEEP	6	15	Support Vector Machine & AdaBoost Random Forest	Linear BCE	Amino acid composition-based feature	0.73	NA	NA	NA	NA	NA	NA	NA	NA
SVMTriP	10	20	Support Vector Machine	Linear BCE from IEDB	Tri-peptide similarity and propensity scores	NA	0.552	0.801	NA	NA	0.693	0.702	NA	
Parker	7	NA	NA	Protein X-ray crystallographic data	Hydrophilic scale	NA	NA	NA	NA	NA	NA	NA	NA	NA
Kolaskar & Tongaonkar	7	NA	NA	NA	Physicochemical properties of amino acids and their frequencies of occurrence	0.75	NA	NA	NA	NA	NA	NA	NA	NA

### BLMPred compared with other existing tools on an independent dataset

We conducted a detailed comparative performance analysis of BLMPred and a large number of existing tools listed in Table 1. Some of these tools have restrictions on the minimum and/or maximum length of the B-cell epitopes. LBEEP, SVMTriP, BLMPred\_5to60, and BLMPred\_8to25 only accept as input peptides within the length ranges 6-15, 10-20, 5-60, and 8-25, respectively. Bepipred-2.0, Bepipred-1.0, Chou & Fasman, Karplus & Schulz, Kolaskar & Tongaonkar, and Parker

perform predictions on input peptides whose length is at least 7, while no limit on the maximum length is imposed. Likewise, Emini and epitope1D require input peptides of at least 6 amino acids in length. Bepipred-3.0 and epiDope do not have any length restrictions. For SVMTriP, six separate models were trained separately on epitopes of length 10, 12, 14, 16, 18, and 20 amino acids. To achieve the maximal performance for a peptide of a given length, the SVMTriP authors recommended using the model trained on epitopes of the same length. Accordingly, we selected the reportedly best-performing SVMTriP model trained on epitopes of 20 amino acids. Hence, we utilized different groups of length-based samples from the BLMPred\_benchmark dataset to fulfill the length-based restrictions of the selected tools for an unbiased detailed comparative analysis of their performances.

BLMPred\_5to60, BLMPred\_8to25, epitope1D, SVMTriP are binary classifiers that predict the input peptide as either a B-cell epitope or as a non-epitope. LBEPP classifies the entire input peptide as an antibody epitope if the corresponding score is greater than a specific threshold. All other tools considered in this comparative study predict the likelihood of individual residues to be part of an epitope. Bepipred-3.0, Bepipred-2.0, Bepipred-1.0, and epiDope assign residue positions as belonging to an epitope if their scores are over a fixed threshold. On the other hand, Chou & Fasman, Karplus & Schulz, Kolaskar & Tongaonkar, Emini, and Parker implemented by IEDB-AR rely on the average residue score for a particular peptide for threshold-based classification. Note that for residue-based methods, we get information on which peptide residues may be a part of an epitope but no information on whether the entire input peptide can be predicted as an epitope or not. For the ease of comparison, we devised an approach so that we are able to compare the per-peptide-based methods with the per-residue-based methods on a common platform. For the per-residue-based methods including Bepipred-3.0, Bepipred-2.0, Bepipred-1.0, epiDope, Chou & Fasman, Karplus & Schulz, Kolaskar & Tongaonkar, Emini, and Parker, we predict the input peptide as an epitope only if at least 50% of its residues are predicted to be part of an epitope by the corresponding methods and vice-versa.

We assessed the performance of models on appropriate length-based partitions of the BLMPred\_benchmark dataset for an unbiased comparison (Table 2). When tested on peptides of length 5-60 from the BLMPred\_benchmark dataset, the precision and specificity of BLMPred model was highest. Although Bepipred-3.0 also achieved comparable performance in this case, its false positive rate (type-I error) was very high, 97%. For other peptide length ranges tested, Bepipred-3.0 provides comparable performance like BLMPred model but its type-I error stays high. Except when tested on 20-length peptides, for all the other cases, BLMPred\_5to60 clearly outperforms other tools w.r.t. MCC, AUROC, and AUPRC.

**Table 2.** Comparison of BLMPred with the previously published methods. The best achieved values of for each performance metrics are highlighted in bold font.

Model		Sample selection from the BLMPred_benchmark dataset			Performance metrics											
Name	Length-based restriction (min-max)	Length range	#of epitopes	#of non-epitopes	Accuracy	Precision	Recall	F1 score	Specificity	MCC	AUROC	AUPRC	TP	FP	TN	FN
BLMPred_5to60	5-60	5-60	1033	663	0.593	<b>0.702</b>	0.576	0.633	<b>0.619</b>	<b>0.191</b>	<b>0.598</b>	<b>0.663</b>	595 (57.6%)	<b>252 (38%)</b>	<b>411 (62%)</b>	438 (42.4%)
Bepipred-3.0	NA				<b>0.608</b>	0.611	<b>0.979</b>	<b>0.753</b>	0.03	0.028	0.504	0.611	<b>1011 (97.9%)</b>	643 (97.0%)	20 (3.0%)	<b>22 (2.1%)</b>



BLMPred_8to25	BLMPred_5to6 0	epiDope	epitope1D	Parker	Kolaskar & Tongaonkar	Karplus & Schulz	Emini	Chou & Fasman	Bepipred-1.0	Bepipred-2.0	Bepipred-3.0	BLMPred_5to6 0
8-25	5-60	NA	6-	7-	7-	7-	6-	7-	7-	7-	NA	5-60
	20						7-60					
	74						1015					
	21						662					
0.611	0.621	0.549	0.417	0.597	0.506	0.565	0.566	0.494	0.538	0.505	0.604	0.589
0.785	0.806	0.594	<b>0.771</b>	0.605	0.592	0.618	0.601	0.585	0.658	0.625	0.607	0.697
0.689	0.676	0.809	0.053	0.962	0.587	0.736	0.847	0.561	0.496	0.455	<b>0.978</b>	0.568
0.734	0.735	0.685	0.099	0.743	0.589	0.672	0.703	0.573	0.565	0.527	<b>0.749</b>	0.626
0.333	0.429	0.153	<b>0.976</b>	0.038	0.38	0.304	0.136	0.391	0.604	0.582	0.03	0.621
0.02	0.091	-0.049	0.071	-0.002	-0.032	0.043	-0.023	-0.048	0.098	0.036	0.027	<b>0.185</b>
0.511	0.552	0.481	0.515	0.499	0.484	0.519	0.492	0.476	0.549	0.518	0.504	<b>0.595</b>
0.783	0.798	0.596	0.614	0.605	0.598	0.615	0.601	0.594	0.631	0.614	0.607	<b>0.657</b>
51 (68.9%)	50 (67.6%)	821 (80.9%)	54 (5.3%)	976 (96.2%)	596 (58.7%)	747 (73.6%)	860 (84.7%)	569 (56.0%)	503 (49.6%)	462 (45.5%)	<b>993 (97.8%)</b>	577 (56.8%)
14 (66.7%)	12 (57.1%)	561 (84.7%)	<b>16 (2.4%)</b>	637 (96.25)	410 (62.0%)	461 (69.6%)	572 (86.4%)	403 (60.9%)	262 (39.6%)	277 (41.8%)	642 (97.0%)	251 (37.9%)
7 (33.3%)	9 (42.9%)	101 (15.3%)	<b>646 (97.6%)</b>	25 (3.8%)	252 (38.0%)	201 (30.4%)	90 (13.6%)	259 (39.1%)	400 (60.4%)	385 (58.2%)	20 (3.0%)	411 (62.1%)
23 (31.1%)	24 (32.4%)	194 (19.1%)	961 (94.7%)	39 (3.8%)	419 (41.3%)	268 (26.4%)	155 (15.3%)	446 (44.0%)	512 (50.4%)	553 (54.5%)	<b>22 (2.2%)</b>	438 (43.2%)





programme and expenses. The only (rather typical) deviation is that we are somewhat late with submitting the papers. The main BLMPred paper is completely written and will be submitted shortly. The joint consortium paper is at a rather early stage and is being actively worked on.

### Während der Durchführung des Vorhabens bekannt gewordene Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen

Clifford et al. published a new version of BepiPred, which also relies on large language models:

Clifford JN, Høie MH, Deleuran S, Peters B, Nielsen M, Marcatili P. BepiPred-3.0: Improved B-cell epitope prediction using protein language models. *Protein Sci.* 2022:e4497.

Our approach is different from Clifford et al. as they are using the ESM-2 model while in our work ProfTrans was employed. A detailed comparison of BibiPred 3.0 and BLMPred is presented above.

### Wissenschaftliche und wirtschaftliche Anschlussfähigkeit / Verwertung der Ergebnisse nach Projektende

The B-cell epitopes predicted by BLMPred can be utilized in the fields of immunology and biotechnology for vaccine development and antibody engineering. As a future vision, we can anticipate that combining structure-based embeddings and sequence-based embeddings together may further improve the predictive potential of these computational tools. Also, it would be interesting to extend and test our approach for the computational prediction of T-cell epitopes.

### Die geplanten Veröffentlichungen des Ergebnisses

Das B. and Frishman D. (2024) BLMPred: predicting linear B-cell epitopes using pre-trained protein language models and machine learning, in preparation.

Geneimmune Consortium. (2024) Computational prediction of pan-coronavirus-vaccine candidates and their experimental validation for protection against SARS-CoV-2 variants, in preparation.

BLMPred model is available as a Github repository (<https://github.com/bdbarnalidas/BLMPred.git>) with thorough instructions for the users which can be easily cloned/downloaded and executed.