Proteasomal cleavage prediction: state-of-the-art and future directions

Ingo Ziegler[1] **Bolei Ma**[2] **Bernd Bischl**[2,3] **Emilio Dorigatti**[2,3,4,*,†] **Benjamin Schubert**[4,5,†]
[1] Center for Information and Language Processing, Ludwig-Maximilians-Universität München
[2] Department of Statistics, Ludwig-Maximilians-Universität München
[3] Munich Center for Machine Learning (MCML)
[4] Computational Health Center, Helmholtz Zentrum München – German Research Center for Environmental Health
[5] Department of Mathematics, Technical University of Munich
[*] Corresponding author: emilio.dorigatti@stat.uni-muenchen.de
[†] Share last authorship

Abstract

Epitope vaccines are a promising approach for precision treatment of pathogens, cancer, autoimmune diseases, and allergies. Effectively designing such vaccines requires accurate proteasomal cleavage prediction to ensure that the epitopes included in the vaccine trigger an immune response. The performance of proteasomal cleavage predictors has been steadily improving over the past decades owing to increasing data availability and methodological advances. In this review, we summarize the current proteasomal cleavage prediction landscape and, in light of recent progress in the field of deep learning, develop and compare a wide range of recent architectures and techniques, including long short-term memory (LSTM), transformers, and convolutional neural networks (CNN), as well as four different denoising techniques. All open-source cleavage predictors re-trained on our dataset performed within two AUC percentage points. Our comprehensive deep learning architecture benchmark improved performance by 1.7 AUC percentage points, while closed-source predictors performed considerably worse. We found that a wide range of architectures and training regimes all result in very similar performance, suggesting that the specific modeling approach employed has a limited impact on predictive performance compared to the specifics of the dataset employed. We speculate that the noise and implicit nature of data acquisition techniques used for training proteasomal cleavage prediction models and the complexity of biological processes of the antigen processing pathway are the major limiting factors. While biological complexity can be tackled by more data and, to a lesser extent, better models, noise and randomness inherently limit the maximum achievable predictive performance. All our datasets and experiments are available at https://github.com/ziegler-ingo/cleavage_benchmark.

1. Introduction

Epitope vaccines (EV) are a promising avenue for precision medical treatment of conditions such as autoimmune diseases and cancer, as they offer the possibility of including highly patient- and disease-specific antigens [Boniolo et al., 2021]. EVs are long polypeptides that contain the very specific parts of an antigen that are recognized by the immune system and trigger an immune response, the so-called epitopes. Proteasomal digestion plays a major role in EV efficacy [Cornet et al., 2006] (Fig. 1). By cleaving the EV incorrectly, therapeutic epitopes might get destroyed and will not be presented at the surface by the Major Histocompatibility Complex (MHC) to T cells.

Several strategies have been tested to deliver EVs in the most effective way [Sette and Fikes, 2003], and it has been experimentally shown that epitope ordering in an EV affects its efficacy by determining the specificity of the immune response to antigen stimulation after vaccination [Cornet et al., 2006]. Epitope ordering affects cleavage efficiency at their terminals and thus the probability that they are correctly separated from the EV construct. Modern EV design frameworks [Dorigatti and Schubert, 2020a,b], therefore, leverage proteasomal cleavage predictors to improve epitope recovery rates, thereby improving their effectiveness at lower dosages and thus reducing the cost to patients.

In this review, we give an overview of all known proteasomal cleavage site predictors and benchmark all currently

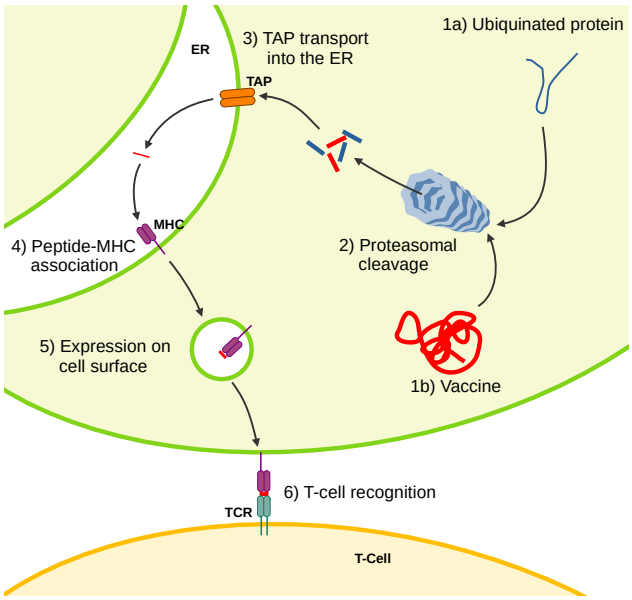

Figure 1. The events in the antigen processing pathway determine what is presented to T cells, thus accurate predictions are essential to enable the design of epitope vaccines. The original purpose of proteasomal cleavage (2) is to digest ubiquinated (old or misfolded) proteins into shorter peptides, but epitope vaccines (1b) undergo the same process. Some of these peptides enter the endoplasmic reticulum (ER) through the Transporter associated with Antigen Processing (TAP, 3), and a fraction of these bind to the Major Histocompatibility Complex (MHC, 4). The resulting complex is then presented on the cell surface (5) enabling immunosurveillance by T cells (6).

accessible ones. In light of the recent developments in the field of deep learning, we additionally benchmark a wide variety of neural architectures, denoising methods, amino acid embedding methods, and training regimes, that have never been applied to proteasomal cleavage prediction before and identify parameters and architecture choices that impact performance. Based on the benchmark results and ablation study, we finally discuss the future direction of the field and give recommendations for the development of novel proteasomal cleavage predictors.

2. Evolution of proteasomal cleavage predictors

Proteasomal cleavage prediction efforts (Table 1) were initiated by the Prediction Algorithm for Proteasomal Cleavages (PAProC), predicting proteasomal cleavage sites of human and yeast proteasomes [Kuttler et al., 2000], which later has been made publicly available as a web-server [Nussbaum et al., 2001]. The model encodes hand-picked features such as distance information from one cut to the next as well as probabilities about neighboring amino acids and the resulting likelihood leading to another cut. While PAProC was only trained on *in vitro* data, Keşmir et al. [2002] also used natural MHC-I ligands and extended their initial multi-layer-perceptron to an ensemble of neural networks to further improve performance [Nielsen et al., 2005a]. Cleavage prediction was integrated with the two other main processes of the antigen processing pathway, namely TAP transport and MHC binding, almost concurrently by [Tenzer et al., 2005b, ProteaSMM] and [Dönnes and Kohlbacher, 2005, PCM]. Both methods are based on position-specific scoring matrices (PSSM), where a separate coefficient is associated with each amino acid position combination and the final cleavage probability is derived linearly from such scores. Bhasin and Raghava [2005] proposed Pcleavage, a cleavage predictor based on Support Vector Machines [Cortes and Vapnik, 1995, SVM], which has been made available as a web-server. Along Pcleavage, the authors benchmarked alternative classifier systems based on nearest-neighbors, logistic regression, naive Bayes, and decision trees, but found the SVM predictor to perform best, with little performance difference between polynomial and radial basis function (RBF) kernels. Ginodi et al. [2008] approached cleavage prediction from a uniquely different perspective and tried to predict the probability that a given peptide was generated by proteasomal cleavage event, rather than the cleavage probability at each position of an antigen, with a similar approach to that of PCM.

While most previous and following cleavage predictors focused on the C-terminal, Schatz et al. [2008] used the PSSM technique introduced previously to study N-terminal cleavage patterns and used the insights derived from their models to propose a more comprehensive modeling approach of the antigen processing pathway comprising epitope trimming besides proteasomal cleavage, TAP transport, and MHC binding. With PCPS, Diez-Rivero et al. [2010] used a model based on n-gram frequencies due to their success in computational linguistics and showed that their model in conjunction with MHC-I binding prediction greatly increase the discovery rate of MHC-restricted CD8 T cell epitopes. Xie et al. [2013] departed from the one-hot encoding of amino acids used in the previous methods and instead used amino acid features summarizing their physio-chemical properties [Mei et al., 2005, VHSE] via principal-component analysis [Jolliffe, 2002, PCA]. By interpreting the importance of the features used by their proposed SVM-based predictor, the authors identified proteasome specificities in terms of hydrophobicity of the amino acids in the flanking regions. NetCleave [Amengual-Rigo and Guallar, 2021] extended cleavage prediction to MHC-II and H2 ligands and returned to neural network models, while also using VHSE amino acid descriptors. Pepsickle [Weeder et al., 2021] collected a larger dataset of *in vitro* cleavage sites and showed that using predicted C-terminal cleavage probabilities helped

Table 1. Chronological overview of existing proteasomal cleavage predictors, the modeling technique they adopted, which epitope terminals were modeled, the window size before (B) and after (A) the cleavage site, whether they used *in vitro* or mass-spectrometry (MS) training data, and the public availability of the predictor. Links to model, code and/or web server functionality, when available, are found in Appendix A.1.

Model	Modeling Technique	Epitope Terminals	Window size		Training data		Availability		
			B	A	*in vitro*	MS	Model	Code	Web
PAProC [Kuttler et al., 2000]	MLP	N/C	6	4	✓				
NetChop [Keşmir et al., 2002]	MLP	C	9	9	✓	✓			
ProteaSMM [Tenzer et al., 2005a]	Linear	C	6	4	✓		✓	✓	
PCM [Dönnes and Kohlbacher, 2005]	Linear	C	4	2	✓				
NetChop 3.1 [Nielsen et al., 2005a]	MLP Ens.	C	8	8	✓	✓	✓		✓
Pcleavage [Bhasin and Raghava, 2005]	SVM	C	9	9	✓	✓		✓	
[Schatz et al., 2008]	Linear	N/C	5	0		✓			
PCPS [Diez-Rivero et al., 2010]	N-grams	C	6	6		✓			✓
[Xie et al., 2013]	SVM	C	14	14	✓	✓			
NetCleave [Amengual-Rigo and Guallar, 2021]	MLP	C	4	3		✓	✓	✓	
Pepsickle [Weeder et al., 2021]	MLP	C	8	8	✓	✓	✓	✓	
PUUPL [Dorigatti et al., 2022]	MLP Ens.	N/C	6	4		✓	✓	✓	
This paper	BiLSTM	N/C	6	4		✓	✓	✓	

isolating immune-responsive epitopes, thus improving EV design. Finally, Dorigatti et al. [2022] proposed to treat decoy samples as unlabeled, rather than negatives, and used specific techniques that do not require labeled negatives to learn a classifier [Bekker and Davis, 2020].

2.1. Data types for training proteasomal cleavage prediction models

Proteasomal cleavage sites are usually inferred implicitly from MHC ligandomics data [Purcell et al., 2019] by matching eluted ligands to their progenitor protein to recover sequence information surrounding the terminals [Keşmir et al., 2002]. This procedure, however, does not give an indication of which amino acid sequences *cannot* result in a cleavage event, since missed cleavage sites are not observed in MHC ligands, and presents a biased picture of cleavage patterns as not all proteolytically generated peptides will be presented by the MHC. Alternatively, *in vitro* collected proteasomal digested proteins are used to train cleavage predictors. These data are more comprehensive and less biased, however, low-throughput and are harder to obtain compared to MHC ligandomics. Additionally, such data do not contain any information about alternative antigen-processing pathways that lead to peptide presentation and successive antigen-processing steps. This complicates the use of predictors trained on *in vitro* data alone in pipelines aimed at identifying immunogenic peptides as it creates the necessity of modeling successive pathway events separately, as done for example by NetCTLpan [Stranzl et al., 2010]. Therefore, most modern proteasomal cleavage predictors are trained solely on MHC ligandomics data or integrate both *in vitro* and *in vivo* data.

To bypass MHC liganomdics' inability to identify true non-cleaved sites, decoy negative samples are usually generated synthetically either by randomly shuffling the amino acids in a short window around the cleavage site or by considering artificial negative sites located around observed cleavage events [Calis et al., 2014] (Fig. 2). Even though such negative samples are not entirely reliable, the growing availability of this kind of data [Vita et al., 2018] spurred continuous development and improvement of proteasomal cleavage predictors [Keşmir et al., 2002, Kuttler et al., 2000, Dönnes and Kohlbacher, 2005, Nielsen et al., 2005b], which have been recently revised in light of new innovations in the deep learning field [Amengual-Rigo and Guallar, 2021, Dorigatti et al., 2022, Weeder et al., 2021].

3. Benchmarking of proteasomal cleavage prediction strategies

All existing methods differ by subtle details such as the size and composition of the training data, the dimension of the window surrounding the cleavage site, and the way amino acids are encoded and presented to the model. Therefore, we performed a comprehensive benchmark of publicly available approaches on comparable datasets to compare and identify beneficial modeling approaches. Additionally, we extended this benchmark to neural architectures and model-

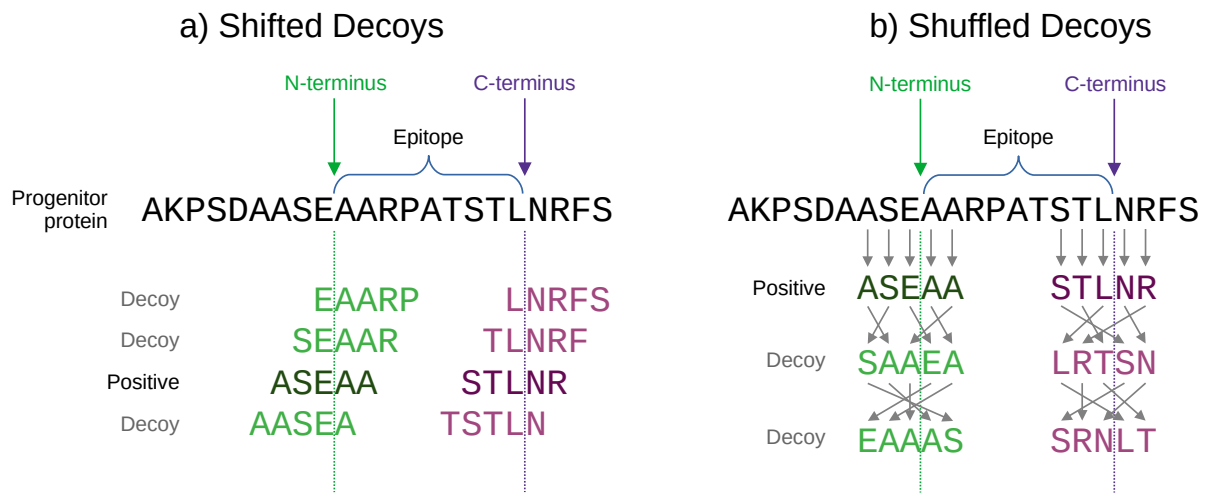

Figure 2. Two alternative ways of creating decoy negative samples from MHC ligandomics data. MHC-I epitopes are matched to their progenitor protein and short windows around its terminals, in the figure composed of three residues before and two after, used to construct decoys. Shifted decoys (a) are substrings of the original sequence starting earlier or after the observed cleavage site, whereas shuffled decoys (b) contain the same amino acids as the positive example but shuffled in random order. In this work, we consider shifted decoys as they are more realistic (see main text).

ing techniques that have not yet been used in the context of proteasomal cleavage site prediction. We considered three main axis of variation: the initial embedding of amino acids, the neural architecture of the predictor, and their training regime via noise handling and data augmentations.

3.1. Embedding

The concept of embeddings originally stems from natural language processing (NLP) and language models, but targeted adaptations for amino acid sequence data have already been presented [Ibtehaz and Kihara, 2021]. By associating numerical vectors to *tokens* such as characters or words, embeddings constitute the first encoding step of a model. The method employed to generate embeddings is crucial as it influences what intrinsic information a model can exploit for classification [Ibtehaz and Kihara, 2021]. We, thus, considered various embeddings in our analysis, while keeping the base architecture equal. Specifically, we analyzed the performance of a randomly initialized embedding layer that was optimized in conjunction with the whole model, as well as dedicated, pre-trained large protein language models such as Prot2Vec [Asgari and Mofrad, 2015]. Furthermore, we also employed sequence embeddings obtained by concatenating independently-trained forward and backward amino acid representations of each input [Heigold et al., 2016], as well as embeddings of the N most common variable-length substrings identified via byte-level byte pair encoding [Sennrich et al., 2016, BBPE] and WordPiece [Schuster and Nakajima, 2012, WP]. For BBPE, we try both $N = 1000$ and $N = 50\,000$, while for WP we use $N = 50\,000$.

3.2. Neural architectures

Different neural architectures embody different inductive biases that may or may not help a learning task. Most proteasomal cleavage methods made use of multi-layer perceptrons (Table 2), however, other types of architectures also provided high performance in sequence-based classification tasks. We, therefore, evaluated all modern types of neural architectures, from recurrent to transformer-based, convolutional neural networks, and perceptrons.

Recurrent: Bidirectional long short-term memory (BiLSTM) networks [Graves and Schmidhuber, 2005] are well suited for a wide range of text classification tasks, thus, as proteasomal cleavage prediction is performed on amino acid sequences in a sequence-to-one classification setting similar to text classification, we based nine of 12 model architectures around BiLSTMs. The fundamental structure for our BiLSTMs was built around the architecture proposed by [Ozols et al., 2021], in which multiple sequential BiLSTMs were followed by a hidden and an output layer. For eight of our nine BiLSTM-related experiments, we chose two sequential BiLSTMs in which sequence dimensionality was reduced by taking the maximum value of the depth-wise per-residue output of the last layer. For the hidden layer, we used the Gaussian Error Linear Units [Hendrycks and Gimpel, 2016, GELU] activation function. As the attention mechanism has also been applied to non-transformer architectures, we additionally employed a modified version of the BiLSTM with extended residual connections and an attention mechanism in order to improve its performance on protein sequence data [Liu and Gong, 2019].

Transformers: Besides RNNs, the attention mechanism introduced by [Vaswani et al., 2017] enabled a whole new architecture capable of processing sequences: the Transformer. These complex multi-million and up to multi-billion parameter models feature an encoder and decoder system and are frequently pre-trained on large datasets in a self-supervised manner, for example by predicting randomly masked-out tokens from an input sequence. The encoders and/or decoders of such pre-trained models are then used as the basis of new models that are fine-tuned for a variety of specific downstream tasks. We used ProtTrans' T5-XL encoder-only model [Elnaggar et al., 2022] featuring 1.2 billion parameters and pre-trained on the UniRef dataset, as well as the ESM2 Transformer [Lin et al., 2022] in its 150 million parameter version. Additionally, we included a fine-tuning performance of ESM2 by adding a linear layer projection from its vocabulary-sized per-residue RoBERTa Language Model Head [Liu et al., 2019, Rives et al., 2021] to our binary classification target.

Convolutional and Perceptron: While convolutional neural networks [LeCun et al., 1998, CNN] are predominantly used in computer vision problems, the convolution filters can also be seen as a sliding k-mer window over the input sequence [Cavnar and Trenkle, 2021]. While CNNs have already been used in conjunction with attention for protease-specific substrates and cleavage sites [Li et al., 2019], their use for proteasomal cleavage prediction remains unexplored (Table 2). Furthermore, as most of the proposed proteasomal cleavage predictors make use of multi-layer perceptrons (Table 2), we also included a single hidden layer perceptron [Rumelhart et al., 1986] with Rectified Linear Units [Agarap, 2018] as activation function into the analysis.

3.3. Training

The training dataset and procedure also impacts the final performance. As described in the introduction, decoy negative samples may contain some positives. To account for negative label uncertainty, we tested multiple procedures that account for this fact. Data augmentation also proved successful to reduce overfitting and improve generalization.

Dataset: We used the dataset introduced in Dorigatti et al. [2022], which contains 229 163 and 222 181 N- and C-terminals cleavage sites, respectively. Each cleavage site is captured in a window comprising six amino acids to its left and four to its right and is associated with six decoy negative samples obtained by considering the three residues preceding and following it (Fig. 2), resulting in a total of 1 434 989 and 1 419 501 samples after deduplication for N- and C-terminals. As the decoy negatives are situated in close proximity to real cleavage sites and due to the probabilistic nature of proteasomal cleavage, some of the negative samples are likely to be actual, unmeasured cleavage sites, and may influence the performance of predictors trained using such data. However, we prefer this method of generating decoys compared to the popular alternative of randomly shuffling amino acids in the window as the latter results in sequences that are potentially not observed in nature, thus making the prediction problem artificially easier and overestimating the models' performance on natural sequences. For the ablation analyses, we used the above-described dataset to investigate the impact of varying the window size around the cleavage site as well as the number of training samples used to train the predictors.

Noisy labels: To reduce the impact of asymmetric label noise on the performance of our classifiers, we took five recent deep learning-specific denoising approaches into consideration: a noise adaptation layer, which attempts to learn the noise distribution in the data [Goldberger and Ben-Reuven, 2017], co-teaching, where two models are trained simultaneously by deciding for the respective other network which samples from a mini-batch to use for training [Han et al., 2018], and co-teaching-plus [Yu et al., 2019], which extends co-teaching with the disagreement learning approach of decoupling [Malach and Shalev-Shwartz, 2017]. We additionally considered a joint training method with co-regularization [Wei et al., 2020, JoCoR] that, unlike co-teaching and its variations, aims at reducing the disagreement of two networks by calculating a joint loss under co-regularization and choosing small-loss samples to update the parameters of both networks simultaneously. We also tried DivideMix [Li et al., 2020a], a holistic approach originally developed for computer vision, which integrates multiple frameworks such as co-teaching and MixMatch [Berthelot et al., 2019] into one. Following the co-teaching approach, each network feeds the other network parts of the training dataset. In this case, however, the division of samples happens at the beginning of an epoch by fitting a Gaussian Mixture Model [Reynolds, 2009, GMM] and splitting the dataset at a chosen threshold. High-loss samples, as classified by the GMM, are treated as unlabeled and semi-supervised learning techniques based on MixMatch are followed. MixMatch unifies consistency regularization, where different augmented views of the same example are constrained to lead to the same prediction, with entropy regularization, which forces the classifier's predictions to have low entropy. As MixMatch was developed for image data, we adjust it for sequence data by mixing up the embedded sequence representation [Guo et al., 2019] instead of the pixel input in the data loading process.

Data augmentation: For all models except external baselines as well as models using BytePair or WordPiece tokenization, we applied data augmentation directly on the input sequences to combat overfitting and improve generalization by randomly masking $\max(1, \lfloor \ell/10 \rfloor)$ amino acids

for a sequence of length ℓ as unknown [Shen et al., 2021]. The models featuring BytePair or WordPiece tokenization do not benefit from additional data augmentation because the trainable tokenization process with large vocabulary sizes might assign a single token to a whole input sequence. Consequently, the augmentation process would mask the sole token, which represents the whole sequence, as unknown.

All predictors except ESM2 fine-tuning used adaptive momentum [Kingma and Ba, 2015] as their optimization technique, whereas ESM2 fine-tuning used adaptive momentum with decoupled weight decay [Loshchilov and Hutter, 2017]. All models without denoising techniques used binary cross-entropy loss [Cox, 1958], while all denoising models calculated dedicated losses.

3.4. Baselines

As baseline models, we considered a logistic regression classifier with L2 regularization, which is essentially equivalent to ProteaSMM, tuned via 10-folds cross-validation, and a SVM model fitted using 15 000 samples to computational limitations, optimizing its hyperparameters via three-folds cross-validation. As for published predictors, we benchmarked against NetCleave [Amengual-Rigo and Guallar, 2021], Pepsickle [Weeder et al., 2021], and PUUPL [Dorigatti et al., 2022], retraining them on our dataset of both N- and C-terminals, as well as comparing a retrained version of PCM [Dönnes and Kohlbacher, 2005] with the published pretrained model. We also considered PCPS (Model 1) [Diez-Rivero et al., 2010], NetChop 20S and NetChop C-term [Nielsen et al., 2005a], however, we could not re-train these due to the unavailability of the source code. We could not use PAProC as the web-server [Nussbaum et al., 2001] is currently not functional, while [Xie et al., 2013] never published code nor pretrained models. Both authors did not respond to our inquiries.

3.5. Experimental protocol

The experimental protocol determines how we optimized and evaluated the models.

Hyperparameter optimization: To gather initial broad ranges of well-performing hyperparameters, we performed manual try-out runs in a grid search approach. Due to computational limitations, we divided the hyperparameter search into two priority groups: group one used Ray Tune's [Moritz et al., 2018] implementation of the asynchronous Hyperband algorithm [Li et al., 2020b] and evaluated each configuration in a ten-folds cross-validation (CV), while for group two we chose hyperparameters according to the initial grid search results and evaluated each configuration with five-folds CV(as shown in Table 16 of Appendix A.3). We then used the best hyperparameter combination to train each architecture with all denoising methods, except for DivideMix where we only trained the overall best-performing architecture due to computational limitations. Information on the different architectures and the exact Hyperband ranges and chosen hyperparameters for all models are documented in Appendix A.2.

Evaluation: As previously mentioned, some negative samples may actually result in a proteasomal cleavage event *in vivo* due to the way these negative samples are generated. For this reason, traditional binary classification metrics such as accuracy, precision, recall, etc., are misleading and model evaluation should instead be based on the AUC [Menon et al., 2015]. We reserved a random 10% of each terminal dataset as test dataset used for the final evaluation of the best hyperparameters.

4. Results

General results: The overall best-performing model architecture for both C- and N-terminus as measured by AUC was the BiLSTM at 89.85% (C-terminus) and 80.62% (N-terminus) without any denoising methods (Fig. 3 and Appendix A.3 and A.4). If denoising techniques were applied, the noise adaptation layer consistently performed best for both the C- and N-terminal. For the C-terminus, the noise adaptation layer showed improved performance in 5 out of 12 models. However, within these 5 models, improvements were minimal, with 4 of 5 models exhibiting less than 0.1 percentage points (pp.) enhancement. This level of improvement is negligible, particularly given that the standard deviation of the models was higher in 2 of the 4 models. In the case of the 1.2 billion BiLSTM+T5 model, the noise adaptation layer led to a more substantial improvement of 0.51 pp, resulting in an AUC of 87.21%, which is still 2.64 pp. below the 4.66 million parameter BiLSTM with 89.85 % AUC.

On the other hand, for models in the N-terminus category also utilizing the noise adaptation layer, improvements were similarly marginal. The BiLSTM+Prot2Vec architecture experienced a modest increase of 0.15 pp., narrowly above its standard deviation of 0.12 pp. The BiLSTM+WP50 model exhibited an improvement of 0.07 pp., which falls below its standard deviation of 0.11 pp. In all other experiments, none of the denoising methods yielded superior performance.

Denoising methods: In the context of experiments with other denoising methods, it is noteworthy that co-teaching-plus yielded identical outcomes in 11 out of 12 models for both the C- and N-terminals. This indicates that the decoupling-based disagreement learning approach within co-teaching-plus was not triggered except once, resulting in co-teaching-plus being equal to co-teaching. In the one remaining scenario of the C-terminus, co-teaching-plus

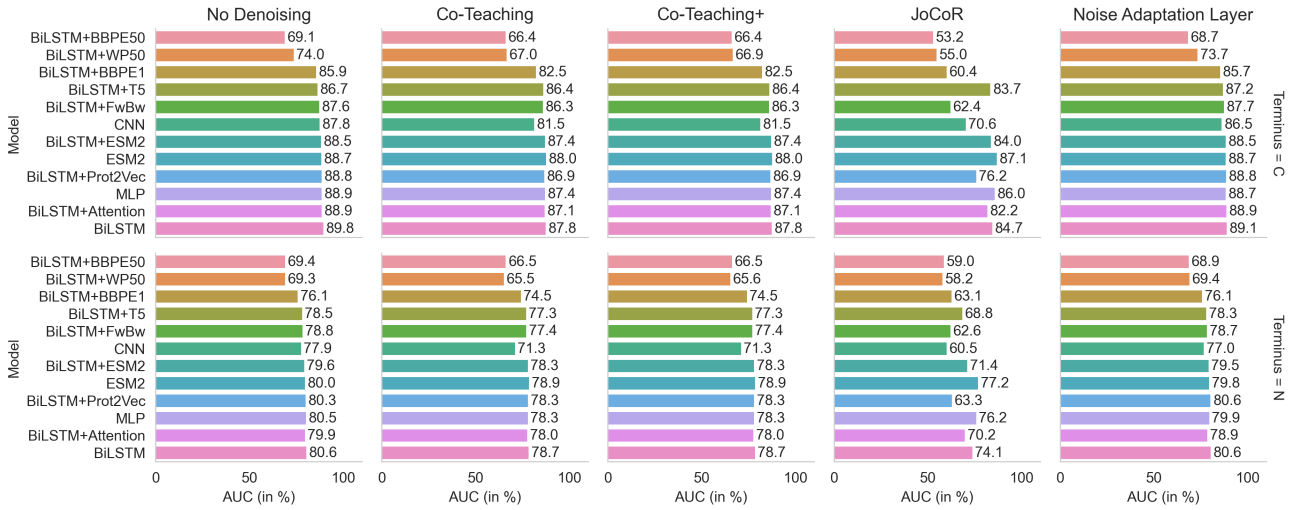

Figure 3. Model performances on C- and N-terminal (rows) for different denoising methods (columns).

showed slightly inferior results, trailing co-teaching by 0.14 pp. Nevertheless, this difference was still within the narrow standard deviation range of 0.15 pp. For the N-terminus, AUC marginally increased by 0.1 pp. for the case where co-teaching-plus took effect. This time, the performance gain of 0.1 pp. exceeded the model's standard deviation range of $\pm$ 0.08 pp. For both the C- and N-terminal, the BiLSTM+WP50 model activated co-teaching-plus's disagreement learning approach.

JoCoR appeared to significantly hinder model performance in all architectures and across both terminals. For the C-terminus, the losses range from -1.62 pp. for the ESM2 model to around -25 pp. for both the BiLSTM+FwBw and BiLSTM+BBPE1 models. The N-terminus saw similar but less pronounced performance deficiencies between -2.79 pp. on the low end for the ESM2 model and up to -20.58 pp. for the BiLSTM+WP50 model. DivideMix, as the most encompassing denoising method, also reduced the BiLSTM AUC score by around 3.6 pp. in both terminals.

Performance plateaus: While the best-performing BiLSTM consisted of 4.6 million parameters, the MLP with 30 529 parameters only lacked 0.99 pp. AUC behind. Most surprisingly, the MLP outperformed 9 out of the remaining 11 architectural configurations, achieving an AUC of 88.86%. This implies that a small, from-scratch-trained model with 30 529 parameters surpassed all the pre-trained Prot2Vec embedding models, the 148 million pre-trained and fine-tuned ESM2 transformer, the 152 million BiLSTM+ESM2 combination, and the 1.2 billion BiLSTM+T5 model, as well as all other BiLSTM and tokenizer combinations. Only the BiLSTM and BiLSTM+Attention models exhibited superior performance, suggesting that the general pre-trained Prot2Vec embeddings may have hindered the performance of the third BiLSTM model, the BiLSTM+Prot2Vec, in this specific task. Although the test set performance of the MLP, indeed, ranked 3rd, it is noteworthy that the performance differences to the non-trainable-tokenizer models ranged only from 0.06 pp. to 2.95 pp. This highlights more of a general performance plateau rather than the dominance of a single architecture. Models trained on the N-terminal followed a similar pattern.

Trainable tokenizers: For the C-terminal, models including trainable tokenizers dropped to their worst-performing state compared to their fixed-vocabulary counterpart, especially when increasing the number of amino acid substring combinations that had to be learned. Whereas the BiLSTM with BBPE vocabulary size 1000 dropped 3.9 p.p. to 85.94% AUC, the same model architecture with 50 000 learned substring combinations was only able to reach 69.14% AUC. A similar but less severe pattern could be observed with WordPiece encodings, where the version with vocabulary of size 50 000 reached 73.95% AUC. If these models additionally featured denoising methods, the performance loss intensified up to a level of almost random-guessing with performances losses in the ranges of up to 37.65 pp. (53.20% AUC for 50 000 BBPE and JoCoR, 55.01% AUC for WP and JoCoR).

On the other hand, the performance loss from to trainable tokenizers for N-terminal cleavage sites was less severe. Although the direct comparison between the BiLSTM and BiLSTM+BBPE1 for C- and C-terminal was larger with -4.5 pp. instead of -3.9 pp., the maximum performance losses are limited at 11.2 pp. for both BiLSTM+BBPE50 and BiLSTM+WP50. When trainable tokenizers are paired with JoCoR, the performance also drops to their lowest level. However, the largest performance loss is capped at 22.45 pp.

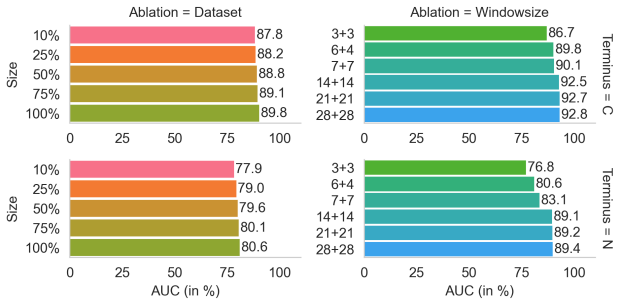

Figure 4. BiLSTM performance with smaller training datasets (left) and different window sizes (right) on C- and N-terminals (top and bottom rows).

for BiLSTM+WP50 compared with the BiLSTM+BBPE50 at 37.65 pp. of the C-terminal. Replacing the embedding layer with a forward-backward representation does not benefit the classification task. Compared with the BiLSTM, not only does performance drop around 2.3 pp., the runtime also increases by a factor of six due to the necessity of increased sequential modeling of each forward-backward token representation.

4.1. Ablation results

To study the influence of the dataset on the predictive performance, we trained the best-performing model identified in the comprehensive benchmark, the BiLSTM, with fewer and fewer training examples as well as with different window sizes around cleavage sites. Due to the small variation in performance discussed earlier, it is likely that repeating these experiments with different methods would result in qualitatively identical conclusions.

Training set size: The size of the training dataset had a negligible impact on performance, with a decrease of only 1.08 and 0.98 p.p. in AUC with half as many training samples for C- and N-terminals respectively (Fig. 4 left, and Table 29 in Appendix A.5). The performance decrease when training with only 10% of the full dataset (about 22 000 cleavage sites and 140 000 decoys) was 2.7 p.p. for N-terminals and 2.06 for C-terminals. We monitored training and validation losses throughout to ensure that no overfitting occurred with the smaller training sets.

Window size: As for window size, performance seemed to plateau when 14 residues were used at each side of the cleavage site, with an improvement of only around 0.3 p.p. when doubling the window size (Fig. 4 right, and Table 30 in Appendix A.5). Starting at a half-window size of 21, the BiLSTM models failed to learn anything, achieving no more than 54% training AUC. Introducing batch normalization layers [Ioffe and Szegedy, 2015] resolved the issue, and in this way the overall best AUC achieved with a window of 56 residues was 92.81% for C-terminals and 89.41% for N-terminals. Increasing the window size from 6 to 56 improved predictive performance by 12.61 pp. for N-terminals and only by 6.14 pp. for C-terminals.

Table 2. AUC (in %) of several cleavage prediction methods on our dataset. We present our BiLSTM results with a window size (in parentheses) comparable to the other methods as well as with the largest window size we could use for training. † SVM, logistic regression, NetCleave, Pepsickle, and PUUPL were re-trained from scratch on our dataset, while we used the published NetChop pre-trained models and the PCPS web-server functionality to obtain their performance on our test set. We present the performance of the original PCM as well as a version re-trained on our dataset.

Method	Dataset	
	C-terminal	N-terminal
PCPS	51.3	50.0
PCM	64.5	52.4
NetChop 3.1 (20S)	66.1	52.7
NetChop 3.1 (C-term)	81.5	51.0
SVM†	84.8	73.2
PCM†	85.3	75.5
Logistic Regression†	86.2	76.2
NetCleave†	86.9	76.4
PUUPL†	87.2	78.0
Pepsickle†	88.1	78.9
BiLSTM (6+4)	**89.8**	**80.6**
BiLSTM (28+28)	**92.8**	**89.4**

Model size: To ensure that the performance numbers just mentioned were not limited by the capacity of the model, we tried to increase the number of units and layers by 50% and 100%, however, both models started to overfit before reaching the performance level of our base BiLSTM model. As we already applied 50% dropout [Srivastava et al., 2014], we progressively increased weight decay [Krogh and Hertz, 1991] and repeated training. In no scenario did any combination of weight decay and increased model size achieve the performance of our base BiLSTM model. For higher levels of weight decay, e.g., 1×10^{-2}, training was severely limited and learning stopped at 87% AUC. Lower levels of weight decay, e.g., 1×10^{-4} were more competitive but still could not match the base BiLSTM performance, even in a setting with extended training epochs. The learning and generalization curves for the 50% size increase model were mirrored by those with 100% increased size. Exact results for each weight decay setting can be found in Table 31 in Appendix A.5.

4.2. Baseline results

Models re-trained on our dataset (PCM, Logistic regression, SVM, NetCleave, PUUPL, and Pepsickle) outperformed available pre-trained models (both versions of NetChop, PCPS and PCM). In all cases, our BiLSTM models outperform all to-date published methods and thus constitute a new state-of-the-art (Table 2). Interestingly, both PCM and the logistic regression classifier performed better than NetChop but worse than NetCleave. Both NetChop and NetCleave are based on MLPs, but while NetCleave was re-trained, NetChop was not, suggesting that retraining NetChop on our dataset would obtain higher performance than shown in the table. Importantly, such simple linear models achieved only three or four pp. fewer than the considerably more complicated BiLSTM. The fact that both linear models performed very competitively compared to the other deep architectures suggests that improvements in the training data were the major driver of the improved performance observed over the years. The SVM model performed slightly worse than logistic regression even though the hyperparameter optimization preferred a RBF kernel over the linear and polynomial kernels. The number of training samples was the bottleneck, and performance could be increased at great computational cost.

5. Conclusion

We benchmarked various deep learning architectures, including MLPs, CNNs, LSTMs, and Transformers, for the task of proteasomal cleavage prediction, and showed that several embedding techniques in combination with model architectures of vastly different scale and complexity all reach similar performance levels around 89.8% AUC for C-terminals and 80.6% AUC for N-terminals, while a simple linear model only achieves three or four p.p. fewer compared to such complicated models. Denoising techniques as well as trainable tokenizers appeared to offer limited to no, or even negative benefits, possibly because the MHC ligandomics dataset already covers most of the cleavage sites that can result in an MHC ligand. By increasing the size of the window around the cleavage site, we reached up to 92.8% and 89.4% AUC for C- and N-terminals, with evident plateaus in terms of training dataset and window size. Performance on the N-terminals dataset was more sensitive both to window size and the number of training samples, confirming the higher difficulty of predicting this process compared to C-terminus cleavage. When maximum predictive performance is desired, we suggest using our best-performing BiLSTM model with the largest possible window size that is feasible for the application at hand, while single-layer or multi-layer perceptrons can be used when increased efficiency is desired with only a slight penalty in predictive performance. The logistic regression classifier also presents a viable option when efficiency is the most important concern as its predictive performance is very competitive with all other deep neural network models.

Such saturated results suggest that different modeling choices of architectures, embeddings, or training regimes are unlikely to yield significantly better predictive performance in the future, although marginal gains could still be possible with even larger datasets. Based on this insight we speculate that further efforts to improve proteasomal cleavage prediction could focus on a more comprehensive modeling of the antigen pathway. For example, it has been shown that cleavage prediction benefits from additionally predicting the MHC binding affinity of the resulting peptides [Keşmir et al., 2002, Diez-Rivero et al., 2010]. Another possibility is that biological processes such as proteasomal cleavage are simply too noisy and random to allow more accurate predictions, in which case we may already be close to the boundary of what is possible to achieve *in silico*.

Competing interests

No competing interest is declared.

Author contributions statement

I.Z. and B.M. conducted the experiments, analyzed the results, and wrote the manuscript. B.B. advised on the experimental setup and assisted in the interpretation of the results. E.D. conceived the experiments, prepared the dataset, advised on the experimental setup, assisted in the interpretation of the results, and wrote the manuscript. B.S. advised on the experimental setup, assisted in the interpretation of the results, and wrote the manuscript. All authors reviewed and approved the manuscript.

Acknowledgments

E. D. was supported by the Helmholtz Association under the joint research school “Munich School for Data Science - MUDS” (Award Number HIDSS-0006). B. B. was supported by the Bavarian Ministry of Economic Affairs, Regional Development and Energy through the Center for Analytics – Data – Applications (ADA-Center) within the framework of BAYERN DIGITAL II (20-3410-2-9-8). B. B. was supported by the German Federal Ministry of Education and Research (BMBF) and the Munich Center for Machine Learning (MCML).

References

Fabio Boniolo, Emilio Dorigatti, Alexander J. Ohnmacht, Dieter Saur, Benjamin Schubert, and Michael P. Menden. Artificial intelligence in early drug discovery enabling precision medicine. *Expert Opinion on Drug Discovery*, 16(9):991–1007, June 2021. doi: 10.1080/17460441.2021.1918096. URL https://doi.org/10.1080/17460441.2021.1918096.

Sébastien Cornet, Isabelle Miconnet, Jeanne Menez, François Lemonnier, and Kostas Kosmatopoulos. Optimal organization of a polypeptide-based candidate cancer vaccine composed of cryptic tumor peptides with enhanced immunogenicity. *Vaccine*, 24(12):2102–2109, 2006.

Alessandro Sette and John Fikes. Epitope-based vaccines: an update on epitope identification, vaccine design and delivery. *Current Opinion in Immunology*, 15(4):461–470, aug 2003. doi: 10.1016/s0952-7915(03)00083-9. URL https://doi.org/10.1016%2Fs0952-7915%2803%2900083-9.

Emilio Dorigatti and Benjamin Schubert. Graph-theoretical formulation of the generalized epitope-based vaccine design problem. *PLOS Computational Biology*, 16(10):e1008237, October 2020a. doi: 10.1371/journal.pcbi.1008237.

Emilio Dorigatti and Benjamin Schubert. Joint epitope selection and spacer design for string-of-beads vaccines. *Bioinformatics*, 36(Supplement_2):i643–i650, December 2020b. doi: 10.1093/bioinformatics/btaa790.

Christina Kuttler, Alexander K Nussbaum, Tobias P Dick, Hans-Georg Rammensee, Hansjörg Schild, and Karl-Peter Hadeler. An algorithm for the prediction of proteasomal cleavages. *Journal of Molecular Biology*, 298(3):417–429, May 2000. doi: 10.1006/jmbi.2000.3683.

Can Keşmir, Alexander K. Nussbaum, Hansjörg Schild, Vincent Detours, and Søren Brunak. Prediction of proteasome cleavage motifs by neural networks. *Protein Engineering, Design and Selection*, 15(4):287–296, April 2002. doi: 10.1093/protein/15.4.287.

S. Tenzer, B. Peters, S. Bulik, O. Schoor, C. Lemmel, M. M. Schatz, P.-M. Kloetzel, H.-G. Rammensee, H. Schild, and H.-G. Holzh"utter. Modeling the MHC class I pathway by combining predictions of proteasomal cleavage,TAP transport and MHC class I binding. *CMLS Cellular and Molecular Life Sciences*, 62(9):1025–1037, May 2005a. ISSN 1420-682X, 1420-9071. doi: 10.1007/s00018-005-4528-2.

Pierre Dönnes and Oliver Kohlbacher. Integrated modeling of the major events in the MHC class I antigen processing pathway. *Protein Science*, 14(8):2132–2140, August 2005. ISSN 09618368, 1469896X. doi: 10.1110/ps.051352405.

Morten Nielsen, Claus Lundgegaard, Ole Lund, and Can Keşmir. The role of the proteasome in generating cytotoxic t-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. *Immunogenetics*, 57:33–41, 1-2 2005a. doi: https://doi.org/10.1007/s00251-005-0781-7.

M. Bhasin and G. P. S. Raghava. Pcleavage: an SVM based method for prediction of constitutive proteasome and immunoproteasome cleavage sites in antigenic sequences. *Nucleic Acids Research*, 33(Web Server):W202–W207, jul 2005. doi: 10.1093/nar/gki587. URL https://doi.org/10.1093%2Fnar%2Fgki587.

Mark M. Schatz, Björn Peters, Nadja Akkad, Nina Ullrich, Alejandra Nacarino Martinez, Oliver Carroll, Sascha Bulik, Hans-Georg Rammensee, Peter van Endert, Hermann-Georg Holzhütter, Stefan Tenzer, and Hansjörg Schild. Characterizing the n-terminal processing motif of MHC class i ligands. *The Journal of Immunology*, 180(5):3210–3217, mar 2008. doi: 10.4049/jimmunol.180.5.3210.

Carmen M Diez-Rivero, Esther M Lafuente, and Pedro A Reche. Computational analysis and modeling of cleavage by the immunoproteasome and the constitutive proteasome. *BMC Bioinformatics*, 11(1), sep 2010. doi: 10.1186/1471-2105-11-479. URL https://doi.org/10.1186%2F1471-2105-11-479.

Jiangan Xie, Zhiling Xu, Shangbo Zhou, Xianchao Pan, Shaoxi Cai, Li Yang, and Hu Mei. The vhse-based prediction of proteasomal cleavage sites. *PLOS ONE*, 8(9):1–8, 09 2013. doi: 10.1371/journal.pone.0074506. URL https://doi.org/10.1371/journal.pone.0074506.

Pep Amengual-Rigo and Victor Guallar. NetCleave: an open-source algorithm for predicting c-terminal antigen processing for MHC-i and MHC-II. *Scientific Reports*, 11(1), June 2021. doi: 10.1038/s41598-021-92632-y.

Benjamin R Weeder, Mary A Wood, Ellysia Li, Abhinav Nellore, and Reid F Thompson. pepsickle rapidly and accurately predicts proteasomal cleavage sites for improved neoantigen identification. *Bioinformatics*, 37(21):3723–3733, September 2021. doi: 10.1093/bioinformatics/btab628. URL https://doi.org/10.1093/bioinformatics/btab628.

Emilio Dorigatti, Bernd Bischl, and Benjamin Schubert. Improved proteasomal cleavage prediction with positive-unlabeled learning. *arXiv preprint arXiv:2209.07527*, 2022.

A. Nussbaum, C. Kuttler, KP. Hadeler, HG. Rammensee, and H. Schild. Paproc: a prediction algorithm for proteasomal cleavages available on the www. *Immunogenetics*, 53:87–94, 2001. doi: https://doi.org/10.1007/s002510100300.

S. Tenzer, B. Peters, S. Bulik, O. Schoor, C. Lemmel, M. M. Schatz, P.-M. Kloetzel, H.-G. Rammensee, H. Schild, and H.-G. Holzh"utter. Modeling the MHC class i pathway by combining predictions of proteasomal cleavage,TAP transport and MHC class i binding. *CMLS, Cell. Mol. Life Sci. Cellular and Molecular Life Sciences*, 62(9):1025–1037, may 2005b. doi: 10.1007/s00018-005-4528-2.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Ido Ginodi, Tal Vider-Shalit, Lea Tsaban, and Yoram Louzoun. Precise score for the prediction of peptides cleaved by the proteasome. *Bioinformatics*, 24(4):477–483, jan 2008. doi: 10.1093/bioinformatics/btm616.

Hu Mei, Zhihua Liao, Yuan Zhou, and Shengshi Z. Li. A new set of amino acid descriptors and its application in peptide qsars. *Biopolymers*, 80 6:775–86, 2005.

Ian T. Jolliffe. Principal component analysis. In *International Encyclopedia of Statistical Science*, 2002.

Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: A survey. *Machine Learning*, 109(4):719–760, 2020.

Anthony W. Purcell, Sri H. Ramarathinam, and Nicola Ternette. Mass spectrometry–based identification of MHC-bound peptides for immunopeptidomics. *Nat Protoc*, 14(6):1687–1707, may 2019. doi: 10.1038/s41596-019-0133-y.

Thomas Stranzl, Mette Voldby Larsen, Claus Lundegaard, and Morten Nielsen. NetCTLpan: pan-specific MHC class i pathway epitope predictions. *Immunogenetics*, 62(6):357–368, apr 2010. doi: 10.1007/s00251-010-0441-4.

Jorg J. A. Calis, Peter Reinink, Christin Keller, Peter M. Kloetzel, and Can Keşmir. Role of peptide processing predictions in t cell epitope identification: contribution of different prediction programs. *Immunogenetics*, 67(2):85–93, December 2014. doi: 10.1007/s00251-014-0815-0. URL https://doi.org/10.1007/s00251-014-0815-0.

Randi Vita, Swapnil Mahajan, James A Overton, Sandeep Kumar Dhanda, Sheridan Martini, Jason R Cantrell, Daniel K Wheeler, Alessandro Sette, and Bjoern Peters. The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Research*, 47 (D1):D339–D343, 10 2018. ISSN 0305-1048. doi: 10.1093/nar/gky1006.

Pierre Dönnes and Oliver Kohlbacher. Integrated modeling of the major events in the MHC class i antigen processing pathway. *Protein Science*, 14(8):2132–2140, August 2005. doi: 10.1110/ps.051352405.

Morten Nielsen, Claus Lundegaard, Ole Lund, and Can Keşmir. The role of the proteasome in generating cytotoxic t-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. *Immunogenetics*, 57(1-2):33–41, mar 2005b. doi: 10.1007/s00251-005-0781-7.

Nabil Ibtehaz and Daisuke Kihara. Application of sequence embedding in protein sequence-based predictions. *arXiv preprint arXiv:2110.07609*, 2021.

Ehsaneddin Asgari and Mohammad RK Mofrad. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PloS one*, 10(11):e0141287, 2015.

Georg Heigold, Guenter Neumann, and Josef van Genabith. Neural morphological tagging from characters for morphologically rich languages. *CoRR*, abs/1606.06640, 2016. URL http://arxiv.org/abs/1606.06640.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL http://aclweb.org/anthology/P16-1162.

Mike Schuster and Kaisuke Nakajima. Japanese and korean voice search. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5149–5152. IEEE, 2012.

Alex Graves and J"urgen Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6):602–610, July 2005. ISSN 08936080. doi: 10.1016/j.neunet.2005.06.042. URL https://linkinghub.elsevier.com/retrieve/pii/S0893608005001206.

Matiss Ozols, Alexander Eckersley, Christopher I. Platt, Callum Stewart-McGuinness, Sarah A. Hibbert, Jerico Revote, Fuyi Li, Christopher E. M. Griffiths, Rachel E. B. Watson, Jiangning Song, Mike Bell, and Michael J. Sherratt. Predicting proteolysis in complex proteomes using deep learning. *International Journal of Molecular Sciences*, 22(6), 2021. ISSN 1422-0067. doi: 10.3390/ijms22063071. URL https://www.mdpi.com/1422-0067/22/6/3071.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Jiale Liu and Xinqi Gong. Attention mechanism enhanced lstm with residual architecture and its application for protein-protein interaction residue pairs prediction. *BMC bioinformatics*, 20(1): 1–11, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *arXiv:1706.03762 [cs]*, December 2017. URL http://arxiv.org/abs/1706.03762. arXiv: 1706.03762.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 7112–7127, 2022. doi: 10.1109/TPAMI.2021.3095381.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, and Alexander Rives. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022. doi: 10.1101/2022.07.20.500902. URL https://www.biorxiv.org/content/early/2022/07/21/2022.07.20.500902.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*, July 2019. URL http://arxiv.org/abs/1907.11692. arXiv: 1907.11692.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021. doi: 10.1073/pnas.2016239118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2016239118.

Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-Based Learning Applied to Document Recognition. *Proc. of the IEEE*, pages 1–46, 1998.

William B Cavnar and John M Trenkle. N-Gram-Based Text Categorization. *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval*, May 2021. URL https://www.researchgate.net/publication/2375544_N-Gram-Based_Text_Categorization.

Fuyi Li, Jinxiang Chen, André Leier, Tatiana Marquez-Lago, Quanzhong Liu, Yanze Wang, Jerico Revote, A Ian Smith, Tatsuya Akutsu, Geoffrey I Webb, Lukasz Kurgan, and Jiangning Song. DeepCleave: a deep learning predictor for caspase and matrix metalloprotease substrates and cleavage sites. *Bioinformatics*, 36(4):1057–1065, 09 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz721. URL https://doi.org/10.1093/bioinformatics/btz721.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

Abien Fred Agarap. Deep learning using rectified linear units (relu). *CoRR*, abs/1803.08375, 2018. URL http://arxiv.org/abs/1803.08375.

Jacob Goldberger and Ehud Ben-Reuven. Training deep neural-networks using a noise adaptation layer. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=H12GRgcxg.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018.

Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *International Conference on Machine Learning*, pages 7164–7173. PMLR, 2019.

Eran Malach and Shai Shalev-Shwartz. Decoupling" when to update" from" how to update". *Advances in neural information processing systems*, 30, 2017.

Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13726–13735, 2020.

Junnan Li, Richard Socher, and Steven C.H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2020a. URL https://openreview.net/forum?id=HJgExaVtwr.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

Douglas A. Reynolds. Gaussian mixture models. In *Encyclopedia of Biometrics*, 2009.

Hongyu Guo, Yongyi Mao, and Richong Zhang. Augmenting data with mixup for sentence classification: An empirical study. *arXiv preprint arXiv:1905.08941*, 2019.

Hongyu Shen, Layne C. Price, Taha Bahadori, and Franziska Seeger. Improving generalizability of protein sequence models with data augmentations. *bioRxiv*, 2021. doi: 10.1101/2021.02.18.431877. URL https://www.biorxiv.org/content/early/2021/02/18/2021.02.18.431877.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

D.R. Cox. The Regression Analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 2 (2), 1958. ISSN 00359246.

Philipp Moritz, Robert Nishihara, Stephanie Wang, Alexey Tumanov, Richard Liaw, Eric Liang, Melih Elibol, Zongheng Yang, William Paul, Michael I Jordan, et al. Ray: A distributed framework for emerging {AI} applications. In *13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18)*, pages 561–577, 2018.

Liam Li, Kevin Jamieson, Afshin Rostamizadeh, Ekaterina Gonina, Jonathan Ben-Tzur, Moritz Hardt, Benjamin Recht, and Ameet Talwalkar. A system for massively parallel hyperparameter tuning. *Proceedings of Machine Learning and Systems*, 2:230–246, 2020b.

Aditya Menon, Brendan Van Rooyen, Cheng Soon Ong, and Bob Williamson. Learning from corrupted binary labels via class-probability estimation. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 125–134, Lille, France, 07–09 Jul 2015. PMLR.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/ioffe15.html.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

Anders Krogh and John Hertz. A simple weight decay can improve generalization. In J. Moody, S. Hanson, and R.P. Lippmann, editors, *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann, 1991. URL https://proceedings.neurips.cc/paper/1991/file/8eefcfdf5990e441f0fb6f3fad709e21-Paper.pdf.

A. Appendix

A.1. Availability

As of October 2023, the publicly available proteasomal cleavage predictors listed in Table 2 are found at:

- ProteaSMM: The official website http://www.mhc-pathway.net/ does not reply. The model is re-implemented in the epytope framework: https://github.com/KohlbacherLab/epytope
- PCM is implemented in the epytope framework: https://github.com/KohlbacherLab/epytope
- NetChop 3.1: https://services.healthtech.dtu.dk/service.php?NetChop-3.1
- Pcleavage: http://crdd.osdd.net/raghava/pcleavage/
- PCPS: http://imed.med.ucm.es/Tools/pcps/
- NetCleave: https://github.com/pepamengual/NetCleave
- Pepsickle: https://github.com/pdxgx/pepsickle
- PUUPL: https://github.com/SchubertLab/proteasomal-cleavage-puupl
- This paper: https://github.com/ziegler-ingo/cleavage_benchmark

A.2. Hyperparameters

Table 3. BiLSTM

Hyperparameter	**Range** (Uniformly random choice)	**Final value**	
		C-terminal	**N-terminal**
Epochs trained	≤ 25	15	15
Learning rate	$\{5 \times 10^{-5}, 10^{-4}, 3 \times 10^{-4}\}$	3×10^{-4}	3×10^{-4}
Dropout rate	$\{0.45, 0.46, \ldots, 0.51, 0.52\}$	0.5	0.5
Linear layer size	$[120, 181)$	164	179
Embedding dimension	$[50, 201)$	91	76
LSTM size 1	$[220, 281)$	228	252
LSTM size 2	$[450, 520)$	506	518

Table 4. BiLSTM+Attention

Hyperparameter	**Range** (Uniformly random choice)	**Final value**	
		C-terminal	**N-terminal**
Epochs trained	≤ 25	20	20
Learning rate	$\{3 \times 10^{-5}, 5 \times 10^{-5}, 8 \times 10^{-5}, 10^{-4}\}$	10^{-4}	10^{-4}
Dropout rate	$\{0.45, 0.46, \ldots, 0.51, 0.52\}$	0.5	0.5
Linear layer size	$[100, 181)$	147	150
Embedding dimension	$\{120, 124, \ldots, 216, 220\}$	216	216
LSTM size	$[64, 131)$	108	111
Attention heads	$\{1, 2, 4\}$	4	1

Table 5. BiLSTM+Prot2Vec

Hyperparameter	**Range** (Uniformly random choice)	**Final value**	
		C-terminal	**N-terminal**
Epochs trained	≤ 60	60	60
Learning rate	$\{8 \times 10^{-5}, , 10^{-4}, 3 \times 10^{-4}, 5 \times 10^{-4}\}$	3×10^{-4}	3×10^{-4}
Dropout rate	$\{0.45, 0.46, \ldots, 0.51, 0.52\}$	0.5	0.5
Linear layer size	$[120, 180)$	145	139
LSTM size	$[480, 531)$	480	531

Table 6. CNN

Hyperparameter	Range (Uniformly random choice)	Final value C-terminal	Final value N-terminal
Epochs trained	≤ 60	60	60
Learning Rate	$\{8 \times 10^{-5}, 10^{-4}, 3 \times 10^{-4}, 5 \times 10^{-4}\}$	3×10^{-4}	3×10^{-4}
Dropout Rate	$\{0, 0.02, \ldots, 0.08, 0.1\}$	0.04	0.08
Linear layer size 1	$[64, 101)$	79	89
Linear layer size 2	$[15, 33)$	24	15
Attention heads 1	$\{1, 2, 3, 4, 5, 6\}$	3	4
Attention heads 2	$\{1, 2, 3, 4, 5, 6\}$	2	3
Filter size 1	-	1	1
Number filters 1	$[220, 301)$	220	249
Number filters 2	$[220, 301)$	262	229
Filter size 2a	$\{1, 3, 5, 7\}$	3	1
Filter size 2b	$\{15, 17, 19, 21, 23, 25\}$	17	15
Filter size 3c	$\{13, 15, 17, 19, 21, 23\}$	13	13
Number filters 3	$[350, 431)$	398	400
Filter size 3a	$\{11, 13, 15, 17, 19\}$	11	13
Filter size 3b	$\{13, 15, 17, 19, 21, 23\}$	15	21
Filter size 3c	$\{11, 13, 15, 17, 19\}$	19	15

Table 7. MLP

Hyperparameter	Range (Uniformly random choice)	Final value C-terminal	Final value N-terminal
Epochs trained	≤ 60	30	30
Learning rate	$\{10^{-4}, 5 \times 10^{-4}, 8 \times 10^{-4}, 10^{-3}\}$	10^{-3}	10^{-3}
Dropout rate	$\{0.1, 0.12, \ldots, 0.24, 0.26\}$	0.24	0.24
Linear layer size	$[120, 201)$	144	167

Table 8. Pepsickle

	Hyperparameter	Range (Uniformly random choice)	Final value C-terminal	Final value N-terminal
Sequence-Net	Epochs trained	≤ 100	20	10
	Learning rate	10^{-3}	10^{-3}	10^{-3}
	Dropout rate	$\{0.1, 0.12, \ldots, 0.34, 0.36\}$	0.34	0.32
	Linear layer size 1	$[120, 201)$	187	175
	Linear layer size 2	$[60, 141)$	71	96
	Linear layer size 3	$[20, 81)$	20	44
Motif-Net	Epochs trained	≤ 100	20	10
	Learning rate	$\{10^{-3}$	10^{-3}	10^{-3}
	Dropout rate	$\{0.1, 0.12, \ldots, 0.24, 0.26\}$	0.12	0.12
	Linear layer size 1	$[120, 201)$	200	200
	Linear layer size 2	$[20, 101)$	23	23

Table 9. BiLSTM+ESM2

Hyperparameter	C-terminal	N-terminal
Epochs trained	10	10
Learning rate	3×10^{-4}	3×10^{-4}
Dropout rate	0.5	0.5
Linear layer size	128	128
LSTM size	512	512

Table 10. ESM2

Hyperparameter	C-terminal	N-terminal
Epochs trained	3	3
Learning rate	2×10^{-5}	2×10^{-5}
Dropout rate	0.5	0.5

Table 11. BiLSTM+T5

Hyperparameter	C-terminal	N-terminal
Epochs trained	10	10
Learning rate	3×10^{-4}	3×10^{-4}
Dropout rate	0.5	0.5
Linear layer size	128	128
LSTM size	512	512

Table 12. BiLSTM+BBPE1, BiLSTM+BBPE50, BiLSTM+WP50

Hyperparameter	C-terminal	N-terminal
Epochs trained	15	15
Learning rate	10^{-4}	10^{-4}
Dropout rate	0.5	0.5
Embedding dimension	150	150
Linear layer size	128	128
LSTM size	512	512

Table 13. BiLSTM+FwBw

Hyperparameter	C-terminal	N-terminal
Epochs trained	15	15
Learning rate	10^{-4}	10^{-4}
Dropout rate	0.5	0.5
Linear layer size 1	128	128
LSTM size 1	128	128
LSTM size 2	512	512
Sequence encoding embedding dimension	100	100
Sequence encoding BiLSTM size	200	200

Table 14. Logistic Regression

Hyperparameter	C-terminal	N-terminal
Maximum Iteration	100	100
Cs	10	10
Penalty	L2	L2
Fit Intercept	Yes	Yes
Intercept Scaling	1	1
Solver	LBFGS	LBFGS
Tolerance	10^{-4}	10^{-4}

PUUPL: See configuration files at https://github.com/SchubertLab/proteasomal-cleavage-puupl/tree/main/configs

Table 15. Denoising methods (Co-Teaching, Co-Teaching+, JoCoR, DivideMix, Noise adaptation layer)

Hyperparameter	Co-teaching	Co-teaching+	JoCoR	DivideMix	Noise adaptation layer
Number scale-up epochs	10	10	10	5	-
Noise rate	0.2	0.2	0.2	-	-
Forget rate	0.2	0.2	0.1	-	-
Exponent	1	1	1	-	-
Number warm-up epochs	-	-	-	1	1
α	-	-	-	0.5	-
λ_u	-	-	-	0	-
Probability threshold	-	-	-	0.5	-
Temperature	-	-	-	0.5	-
β	-	-	-	-	0.8

A.3. Results without denoising methods

Table 16. Model performances without denoising

Terminal	Prio.	Models	AUC	ACC	s/epoch	Epochs	Parameters	Trainable
C-terminal	1	BiLSTM	89.85 ± 0.11	87.66 ± 0.07	25	15	4 655 984	4 655 984
		BiLSTM+Attention	88.90 ± 0.11	87.09 ± 0.09	20	20	1 632 391	1 632 391
		BiLSTM+Prot2Vec	88.80 ± 0.04	87.26 ± 0.06	20	15	16 009 371	5 830 371
		CNN	87.77 ± 0.14	86.25 ± 0.13	45	60	16 084 198	16 084 198
		MLP	88.86 ± 0.11	87.22 ± 0.08	4	30	30 529	30 529
	2	BiLSTM+ESM2	88.53 ± 0.07	87.01 ± 0.03	330	10	152 998 267	4 858 113
		ESM2	88.74 ± 0.13	86.37 ± 1.46	900	3	148 140 188	148 140 188
		BiLSTM+T5	86.70 ± 0.05	85.91 ± 0.14	780	10	1 214 572 801	6 430 977
		BiLSTM+BBPE1	85.91 ± 0.11	85.83 ± 0.10	13	15	5 319 409	5 319 409
		BiLSTM+BBPE50	69.14 ± 0.17	82.41 ± 0.11	12	15	12 669 409	12 669 409
		BiLSTM+WP50	73.95 ± 0.24	82.99 ± 0.08	16	15	12 669 409	12 669 409
		BiLSTM+FwBw	87.57 ± 0.07	86.75 ± 0.09	120	15	4 315 369	4 315 369
N-terminal	1	BiLSTM	80.62 ± 0.11	83.82 ± 0.12	25	15	5 096 135	5 096 135
		BiLSTM+Attention	79.92 ± 0.10	83.63 ± 0.13	20	20	1 718 233	1 718 233
		BiLSTM+Prot2Vec	80.31 ± 0.09	83.81 ± 0.07	20	15	16 772 049	6 593 049
		CNN	77.91 ± 1.17	82.95 ± 0.26	45	60	15 237 057	15 237 057
		MLP	80.46 ± 0.10	83.76 ± 0.13	4	30	35 405	35 405
	2	BiLSTM+ESM2	79.59 ± 0.12	83.55 ± 0.07	360	10	152 998 267	4 858 113
		ESM2	79.96 ± 0.14	82.52 ± 0.57	900	3	148 140 188	148 140 188
		BiLSTM+T5	78.48 ± 0.15	83.06 ± 0.18	780	10	1 214 572 801	6 430 977
		BiLSTM+BBPE1	76.12 ± 0.06	82.61 ± 0.07	13	15	5 319 409	5 319 409
		BiLSTM+BBPE50	69.42 ± 0.20	82.15 ± 0.11	12	15	12 669 409	12 669 409
		BiLSTM+WP50	69.33 ± 0.10	81.91 ± 0.14	15	15	12 669 409	12 669 409
		BiLSTM+FwBw	78.75 ± 0.13	82.23 ± 0.07	120	15	4 315 369	4 315 369

A.4. Results with denoising methods

Table 17. BiLSTM with denoising

Terminal	Denoising methods	AUC	ACC	s/epoch	Epochs	Parameters	Trainable
C-terminal	Co-Teaching	87.75 ± 0.09	86.80 ± 0.06	41	15	4 655 984	4 655 984
	Co-Teaching+	87.75 ± 0.09	86.80 ± 0.06	41	15	4 655 984	4 655 984
	JoCoR	84.70 ± 0.22	85.36 ± 0.06	41	15	4 655 984	4 655 984
	DivideMix	86.25	84.02	210	15	4 656 149	4 656 149
	Noise Adaptation Layer	89.11 ± 0.06	87.36 ± 0.04	22	15	4 656 149	4 656 149
N-terminal	Co-Teaching	78.74 ± 0.11	83.41 ± 0.11	44	15	5 096 135	5 096 135
	Co-Teaching+	78.74 ± 0.11	83.41 ± 0.11	44	15	5 096 135	5 096 135
	JoCoR	74.09 ± 0.25	82.83 ± 0.14	44	15	5 096 135	5 096 135
	DivideMix	77.08	81.52	210	15	5 096 315	5 096 315
	Noise Adaptation Layer	70.60 ± 0.10	83.81 ± 0.07	23	15	5 096 315	5 096 315

Table 18. BiLSTM+Attention with denoising

Terminal	Denoising methods	AUC	ACC	s/epoch	Epochs	Parameters	Trainable
C-terminal	Co-Teaching	87.10 ± 0.05	86.42 ± 0.10	40	20	1 632 391	1 632 391
	Co-Teaching+	87.10 ± 0.05	86.42 ± 0.10	40	20	1 632 391	1 632 391
	JoCoR	82.17 ± 0.29	85.46 ± 0.06	38	20	1 632 391	1 632 391
	Noise Adaptation Layer	88.93 ± 0.06	87.24 ± 0.08	18	20	1 632 539	1 632 539
N-terminal	Co-Teaching	77.96 ± 0.11	83.19 ± 0.07	43	20	1 718 233	1 718 233
	Co-Teaching+	77.96 ± 0.11	83.19 ± 0.07	42	20	1 718 233	1 718 233
	JoCoR	70.16 ± 0.70	82.12 ± 0.14	41	20	1 718 233	1 718 233
	Noise Adaptation Layer	78.89 ± 0.11	83.59 ± 0.06	22	20	1 718 384	1 718 384

Table 19. BiLSTM+Prot2Vec with denoising

Terminal	Denoising methods	AUC	ACC	s/epoch	Epochs	Parameters	Trainable
C-terminal	Co-Teaching	86.94 ± 0.12	87.30 ± 0.07	32	15	16 009 371	5 830 371
	Co-Teaching+	86.94 ± 0.12	87.30 ± 0.07	31	15	16 009 371	5 830 371
	JoCoR	76.18 ± 0.16	82.31 ± 0.12	32	15	16 009 371	5 830 371
	Noise Adaptation Layer	88.81 ± 0.08	87.21 ± 0.06	15	15	16 009 517	5 830 517
N-terminal	Co-Teaching	78.32 ± 0.12	83.56 ± 0.07	40	15	16 772 049	6 593 049
	Co-Teaching+	78.32 ± 0.12	83.56 ± 0.07	40	15	16 772 049	6 593 049
	JoCoR	63.29 ± 1.08	81.83 ± 0.12	40	15	16 772 049	6 593 049
	Noise Adaptation Layer	80.56 ± 0.15	83.71 ± 0.08	19	15	16 772 189	6 593 189

Table 20. CNN with denoising

Terminal	Denoising methods	AUC	ACC	s/epoch	Epochs	Parameters	Trainable
C-terminal	Co-Teaching	81.54 ± 0.54	85.51 ± 0.14	102	60	16 084 198	16 084 198
	Co-Teaching+	81.54 ± 0.54	85.51 ± 0.14	102	60	16 084 198	16 084 198
	JoCoR	70.58 ± 0.21	82.31 ± 0.07	102	60	16 084 198	16 084 198
	Noise Adaptation Layer	86.50 ± 5.75	85.55 ± 0.98	49	60	16 084 223	16 084 223
N-terminal	Co-Teaching	71.33 ± 0.80	82.61 ± 0.08	102	60	15 237 057	15 237 057
	Co-Teaching+	71.33 ± 0.80	82.61 ± 0.09	102	60	15 237 057	15 237 057
	JoCoR	60.48 ± 0.59	81.83 ± 0.12	102	60	15 237 057	15 237 057
	Noise Adaptation Layer	76.99 ± 2.52	82.91 ± 0.26	47	60	15 237 073	15 237 073

Table 21. MLP with denoising

Terminal	Denoising methods	AUC	ACC	s/epoch	Epochs	Parameters	Trainable
C-terminal	Co-Teaching	87.37 ± 0.10	86.72 ± 0.06	6	30	30 529	30 529
	Co-Teaching+	87.37 ± 0.10	86.72 ± 0.06	6	30	30 529	30 529
	JoCoR	85.99 ± 0.09	85.63 ± 0.04	5	30	30 529	30 529
	Noise Adaptation Layer	88.70 ± 0.09	87.19 ± 0.07	3	30	30 674	30 674
N-terminal	Co-Teaching	78.30 ± 0.06	83.38 ± 0.10	6	30	35 405	35 405
	Co-Teaching+	78.30 ± 0.06	83.38 ± 0.10	5	30	35 405	35 405
	JoCoR	76.21 ± 0.05	81.97 ± 0.14	5	30	35 405	35 405
	Noise Adaptation Layer	79.88 ± 0.10	83.65 ± 0.10	4	30	35 573	35 573

Table 22. BiLSTM+ESM2 with denoising

Terminal	Denoising methods	AUC	ACC	s/epoch	Epochs	Parameters	Trainable
C-terminal	Co-Teaching	87.37 ± 0.02	86.65 ± 0.05	660	10	152 998 267	4 858 113
	Co-Teaching+	87.37 ± 0.02	86.65 ± 0.05	660	10	152 998 267	4 858 113
	JoCoR	84.00 ± 0.11	84.82 ± 0.08	660	10	152 998 267	4 858 113
	Noise Adaptation Layer	88.55 ± 0.08	87.00 ± 0.06	330	10	152 998 396	4 858 242
N-terminal	Co-Teaching	78.30 ± 0.08	83.42 ± 0.08	660	10	152 998 267	4 858 113
	Co-Teaching+	78.30 ± 0.08	83.42 ± 0.08	660	10	152 998 267	4 858 113
	JoCoR	71.41 ± 0.36	81.86 ± 0.13	660	10	152 998 267	4 858 113
	Noise Adaptation Layer	79.51 ± 0.07	83.50 ± 0.11	360	10	152 998 396	4 858 242

Table 23. ESM2 with denoising

Terminal	Denoising methods	AUC	ACC	s/epoch	Epochs	Parameters	Trainable
C-terminal	Co-Teaching	87.99 ± 0.18	82.55 ± 0.83	2160	3	148 140 188	148 140 188
	Co-Teaching+	87.99 ± 0.18	85.55 ± 0.83	2100	3	148 140 188	148 140 188
	JoCoR	87.12 ± 0.22	86.29 ± 0.84	2100	3	148 140 188	148 140 188
	Noise Adaptation Layer	88.68 ± 1.90	77.71 ± 11.67	960	3	148 140 222	148 140 222
N-terminal	Co-Teaching	78.90 ± 1.03	81.83 ± 2.59	2220	3	148 140 188	148 140 188
	Co-Teaching+	78.90 ± 1.03	81.83 ± 2.59	2220	3	148 140 188	148 140 188
	JoCoR	77.17 ± 0.84	82.59 ± 0.28	2160	3	148 140 188	148 140 188
	Noise Adaptation Layer	79.82 ± 0.48	70.27 ± 14.85	930	3	148 140 222	148 140 222

Table 24. BiLSTM+T5 with denoising

Terminal	Denoising methods	AUC	ACC	s/epoch	Epochs	Parameters	Trainable
C-terminal	Co-Teaching	86.41 ± 0.20	86.31 ± 0.17	1500	10	1 214 572 801	6 430 977
	Co-Teaching+	86.41 ± 0.20	86.31 ± 0.17	1500	10	1 214 572 801	6 430 977
	JoCoR	83.72 ± 0.15	84.01 ± 0.06	1500	10	1 214 572 801	6 430 977
	Noise Adaptation Layer	87.21 ± 0.17	86.12 ± 0.30	720	10	1 214 572 930	6 431 106
N-terminal	Co-Teaching	77.28 ± 0.11	82.84 ± 0.10	1560	10	1 214 572 801	6 430 977
	Co-Teaching+	77.28 ± 0.11	85.91 ± 0.14	1560	10	1 214 572 801	6 430 977
	JoCoR	68.79 ± 0.17	81.83 ± 0.12	1500	10	1 214 572 801	6 430 977
	Noise Adaptation Layer	78.30 ± 0.05	82.82 ± 0.18	780	10	1 214 572 930	6 431 106

Table 25. BiLSTM+BBPE1 with denoising

Terminal	Denoising methods	AUC	ACC	s/epoch	Epochs	Parameters	Trainable
C-terminal	Co-Teaching	82.48 ± 0.12	85.42 ± 0.06	33	15	5 319 409	5 319 409
	Co-Teaching+	82.48 ± 0.12	85.42 ± 0.06	34	15	5 319 409	5 319 409
	JoCoR	60.39 ± 0.37	82.20 ± 0.07	33	15	5 319 409	5 319 409
	Noise Adaptation Layer	85.70 ± 0.21	85.72 ± 0.20	17	15	5 319 538	5 319 538
N-terminal	Co-Teaching	74.50 ± 0.27	82.59 ± 0.09	34	15	5 319 409	5 319 409
	Co-Teaching+	74.50 ± 0.27	82.59 ± 0.09	34	15	5 319 409	5 319 409
	JoCoR	63.11 ± 0.36	81.88 ± 0.12	33	15	5 319 409	5 319 409
	Noise Adaptation Layer	76.08 ± 0.14	82.53 ± 0.09	17	15	5 319 538	5 319 538

Table 26. BiLSTM+BBPE50 with denoising

Terminal	Denoising methods	AUC	ACC	s/epoch	Epochs	Parameters	Trainable
C-terminal	Co-Teaching	66.44 ± 0.18	82.28 ± 0.08	30	15	12 669 409	12 669 409
	Co-Teaching+	66.44 ± 0.18	82.28 ± 0.08	30	15	12 669 409	12 669 409
	JoCoR	53.20 ± 0.22	82.20 ± 0.07	30	15	12 669 409	12 669 409
	Noise Adaptation Layer	68.67 ± 0.13	82.38 ± 0.13	16	15	12 669 538	12 669 538
N-terminal	Co-Teaching	66.49 ± 0.23	82.05 ± 0.13	31	15	12 669 409	12 669 409
	Co-Teaching+	66.49 ± 0.23	82.05 ± 0.13	32	15	12 669 409	12 669 409
	JoCoR	58.99 ± 0.81	81.88 ± 0.12	31	15	12 669 409	12 669 409
	Noise Adaptation Layer	68.88 ± 0.37	82.14 ± 0.14	15	15	12 669 538	12 669 538

Table 27. BiLSTM+WP50 with denoising

Terminal	Denoising methods	AUC	ACC	s/epoch	Epochs	Parameters	Trainable
C-terminal	Co-Teaching	67.02 ± 0.15	82.25 ± 0.02	40	15	12 669 409	12 669 409
	Co-Teaching+	66.88 ± 0.23	82.78 ± 0.02	40	15	12 669 409	12 669 409
	JoCoR	55.01 ± 0.57	82.20 ± 0.07	39	15	12 669 409	12 669 409
	Noise Adaptation Layer	73.67 ± 0.55	82.79 ± 0.08	20	15	12 669 538	12 669 538
N-terminal	Co-Teaching	65.53 ± 0.10	81.99 ± 0.15	37	15	12 669 409	12 669 409
	Co-Teaching+	65.63 ± 0.12	81.84 ± 0.15	37	15	12 669 409	12 669 409
	JoCoR	58.17 ± 0.32	81.88 ± 0.12	36	15	12 669 409	12 669 409
	Noise Adaptation Layer	69.40 ± 0.16	81.88 ± 0.14	20	15	12 669 538	12 669 538

Table 28. BiLSTM+FwBw with denoising

Terminal	Denoising methods	AUC	ACC	s/epoch	Epochs	Parameters	Trainable
C-terminal	Co-Teaching	86.29 ± 0.13	85.81 ± 0.08	270	15	4 315 369	4 315 369
	Co-Teaching+	86.29 ± 0.13	85.81 ± 0.08	270	15	4 315 369	4 315 369
	JoCoR	62.41 ± 0.64	82.31 ± 0.07	270	15	4 315 369	4 315 369
	Noise Adaptation Layer	87.66 ± 0.07	86.45 ± 0.08	120	15	4 315 498	4 315 498
N-terminal	Co-Teaching	77.38 ± 0.10	82.89 ± 0.09	270	15	4 315 369	4 315 369
	Co-Teaching+	77.38 ± 0.10	82.89 ± 0.09	270	15	4 315 369	4 315 369
	JoCoR	62.56 ± 0.46	81.83 ± 0.12	270	15	4 315 369	4 315 369
	Noise Adaptation Layer	78.68 ± 0.12	83.19 ± 0.11	120	15	4 315 498	4 315 498

A.5. Ablation results

Table 29. Dataset size ablation of BiLSTM

Terminal	Size	AUC	ACC	s/epoch	Epochs	Parameters	Trainable
C-terminal	10%	87.79 ± 0.33	86.27 ± 0.29	2	15	4 655 984	4 655 984
	25%	88.16 ± 0.20	86.53 ± 0.16	6	15	4 655 984	4 655 984
	50%	88.77 ± 0.22	86.94 ± 0.12	12	15	4 655 984	4 655 984
	75%	89.12 ± 0.16	87.14 ± 0.11	18	15	4 655 984	4 655 984
	100%	89.85 ± 0.11	87.66 ± 0.07	25	15	4 655 984	4 655 984
N-terminal	10%	77.91 ± 0.65	82.99 ± 0.31	2	15	5 096 135	5 096 135
	25%	78.97 ± 0.29	83.12 ± 0.19	6	15	5 096 135	5 096 135
	50%	79.64 ± 0.17	83.21 ± 0.09	12	15	5 096 135	5 096 135
	75%	80.11 ± 0.14	83.36 ± 0.11	18	15	5 096 135	5 096 135
	100%	80.62 ± 0.11	83.82 ± 0.12	25	15	5 096 135	5 096 135

Table 30. Window size ablation of BiLSTM

Terminal	Halfwindow Size	AUC	ACC	s/epoch	Epochs	Parameters	Trainable
C-terminal	3	86.67 ± 0.12	85.62 ± 0.10	17	15	4 655 984	4 655 984
	7	90.07 ± 0.12	87.94 ± 0.11	30	15	4 655 984	4 655 984
	14	92.50 ± 0.09	89.71 ± 0.07	57	15	4 655 984	4 655 984
	21	92.74 ± 0.11	89.90 ± 0.09	84	15	4 656 648	4 656 648
	28	92.81 ± 0.17	90.01 ± 0.16	109	15	4 656 648	4 656 648
N-terminal	3	76.80 ± 0.24	82.76 ± 0.05	21	15	5 096 135	5 096 135
	7	83.12 ± 0.18	84.82 ± 0.09	34	15	5 096 135	5 096 135
	14	89.13 ± 0.11	88.17 ± 0.13	64	15	5 096 135	5 096 135
	21	89.24 ± 0.22	88.53 ± 0.11	90	15	5 096 829	5 096 829
	28	89.41 ± 0.11	88.78 ± 0.14	115	15	5 096 829	5 096 829

Table 31. Size variations with weight decay for BiLSTM on C-terminal with a halfwindow size of 28

Modal size	Weight decay	Max AUC	Epoch	s/epoch	Epochs trained	Parameters	Trainable
50%	0	91.89	5	210	15	10 464 925	10 464 925
	1×10^{-2}	87.25	15	210	15	10 464 925	10 464 925
	1×10^{-3}	89.55	15	210	15	10 464 925	10 464 925
	1×10^{-4}	91.68	19	210	20	10 464 925	10 464 925
	1×10^{-5}	91.85	6	210	15	10 464 925	10 464 925
100%	0	91.84	4	330	15	10 464 925	10 464 925
	1×10^{-2}	87.31	15	330	15	10 464 925	10 464 925
	1×10^{-3}	89.72	15	330	15	10 464 925	10 464 925
	1×10^{-4}	91.70	18	330	20	10 464 925	10 464 925
	1×10^{-5}	91.77	5	330	15	10 464 925	10 464 925