# Studying signal peptides with attention neural networks informs cleavage site predictions

**Patrick Bryant and Arne Elofsson**
Dep of Biochemistry and Biophysics and Science for Life Laboratory
Stockholm University, Solna, 171 21
patrick.bryant@scilifelab.se and arne@bioinfo.se

## Abstract

Signal peptides are essential for protein sorting and processing. Evaluating signal peptides experimentally is difficult and prone to errors, therefore the exact cleavage sites are often misannotated. Here, we describe a novel explainable method to identify signal peptides and predict the cleavage site, with a performance similar to state-of-the art methods. We treat each amino acid sequence as a sentence and its annotation as a translation problem. We utilise attention neural networks in a transformer model using a simple one-hot representation of each amino acid, without including any evolutionary information. By analysing the encoder-decoder attention of the trained network, we are able to explain what information in the peptide is used to annotate the cleavage site. We find the most common signal peptide motifs and characteristics and confirm that the most informative amino acid sites vary greatly between kingdoms and signal peptide types as previous studies have shown. Our findings open up the possibility to gain biological insight using transformer neural networks on small sets of labelled information.

## 1 Introduction

Signal peptides (SPs) are found in Archaea (1), Eukarya (2) and Bacteria (3) and are important for protein sorting and processing (4). Annotation of SPs is a two-fold problem, distinguishing between presence or absence of SPs and determining the SP cleavage site (CS) (5). These problems have proven to be difficult, in particular, finding the CS. The CSs arise from that signal peptidases (SPases) remove SPs after they have fulfilled their function of e.g. assuring proteins are translocated across membranes. The main pathway directing translocation is called the "general secretory pathway" (Sec). Although, another pathway, only present in Archaea, Chloroplasts, Mitochondria and Bacteria, called the twin-arginine translocation (Tat) pathway exists (6). The peptides in this pathway have been found to have two consecutive arginines near the N-terminal, hence the name.

The most common SPase is SPase I and exists in Archaea, Eukarya and Bacteria for cleaving SPs in the Sec pathway. In Bacteria, lipoproteins in the Sec pathway are cleaved by another SPase, SPase II, which recognizes a special cysteine motif immediately after the CS (7). The Tat SPs are only processed by either SPase I or II, although another SPase, SPase III, exists for cleavage of archaeal and bacterial pilins in the Sec pathway (8). Peptides in the Sec pathway can thus be cleaved by SPase I-III, while peptides in the Tat pathway can only be cleaved by SPase I or II.

Signal peptides have been found to be highly variable in sequence, although some motifs such as the mentioned cysteine and twin-arginine exist. However, a tripartite structure consisting of a (1) positively charged region near the N-terminal, (2) a span of 10-15 hydrophobic amino acids (AAs) in the SP center and (3) more polar AAs with decreasing size towards the C-terminal, is a classical characteristic (9; 10). Due to the presence of these characteristics, many methods have been developed to predict SPs and CSs, the most recent and successful using deep learning (5). The best current

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

method overall (although the performance is not the best for all SP/CS types) for predicting various types of SPs and their CSs is SignalP 5.0 (11). SignalP 5.0 distinguishes SPs in the Sec pathway cleaved by SPase I (Sec/SPI) or Spase II (Sec/SPII) and SPs in the Tat pathway cleaved by SPase I (Tat/SPI). Eukaryotic SPs can only be predicted for Sec/SPI, while archaeal and bacterial can be predicted for all types (Sec/SPI, Sec/SPII and Tat/SPI).

Here, we use a pure attention based language modelling approach to annotate SPs and CSs. We translate the meaning of each amino acid into six different annotation categories: Sec/SPI signal peptide, Tat/SPI signal peptide, Sec/SPII signal peptide, cytoplasm, transmembrane and extracellular. Compared to other language models pre-trained on all proteins, to be fine-tuned in a later stage for a certain task (12), our model is trained end-to-end with a more limited dataset. This enables the model to learn specific aspects of the SP "language" and what it means in regards to SP and CS annotation, similar to learning the semantics of a language, compared to just learning its grammar.

## 2 Methods

### 2.1 Data

To assess the performance of our network in comparison to the best available methods, we use the same dataset as in SignalP 5.0 (11) for training, testing and benchmarking. Briefly, this set consists of eukaryotic, archaeal and bacterial peptides from UniProtKB 2018_04 longer than 30 AAs (Table S1). Only the first 70 N-terminal AAs were used for all peptides. If the peptides were shorter than 70 AAs, they were padded with the character "X". For the signal peptides, the classes Sec/SPI, Tat/SPI and Sec/SPII exist. The Eukaryotic proteins all belong to the class Sec/SPI, while the other kingdoms have peptides in all classes. The negative set consists of globular and membrane proteins. Training, testing and benchmarking of the model was performed exactly like in SignalP 5.0, using a nested 5-fold cross-validation procedure. The data was homology partitioned on 20% sequence identity, which ensures low bias for the testing and benchmarking. Further, the nesting procedure ensures that the models predicting for each fold during testing and benchmarking have not seen any of the data.

### 2.2 Network architecture

The network is a pure attention based transformer model (13) (Figure1), where the input AA sequence of length 70 is transformed to an output annotation of length 70 consisting of six classes: Sec/SPI signal peptide, Tat/SPI signal peptide, Sec/SPII signal peptide, cytoplasm, transmembrane and extracellular. The input AA sequence is first embedded using token-position embeddings, which are added together, in the encoder block. These embeddings are learned during training and passed through attention layers consisting of multi-headed self-attention followed by skip-connection and normalization, feed-forward layer with ReLU activation and again skip-connection and normalization. The normalization transforms its input so a mean close to zero and standard deviation close to one are maintained. After each attention and feed-forward layer, dropout with rate 0.1 is applied throughout the network. N encoder blocks are applied before the information is passed to the decoder blocks.

To be able to assess the attention on the input sequence towards the output annotations, we pass a random annotation tensor as input to the decoder block. This tensor is embedded with a separate learned additive token-position embedding and passed through a multi-headed self-attention layer followed by skip-connection and normalization, feed-forward layer with ReLU activation and again skip-connection and normalization. The resulting tensors are joined with the encoded AA sequence in the decoder to utilize the concept of encoder-decoder attention. There, the key and value consists of the encoded AA sequence and the query of the multi-headed self-attention of the annotation tensor. The output from encoder-decoder attention is subject to skip-connection and normalization, feed-forward layer with ReLU activation and again skip-connection and normalization. N such decoder blocks are applied, where all but the first one receive their input from the previous decoder block and all receive keys and values from the N encoder blocks.

A softmax layer is applied on the decoder output, resulting in probabilities over the six different annotation classes. The network can, therefore, iteratively modify the annotations and obtain feedback on its decisions throughout the encoder-decoder attention. After iterating, a final decoder block, followed by softmax activation is applied resulting in the predicted annotation output. The loss is

calculated towards the true annotations across all six classes for all 70 positions in the annotation output (see section 6.2 for further details).

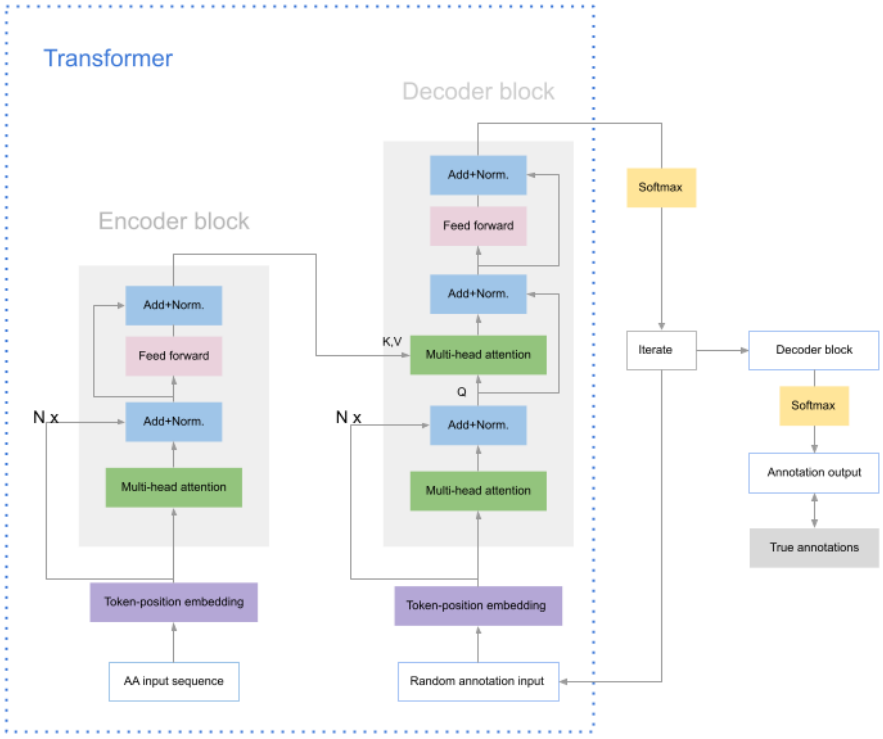


Figure 1: Transformer network architecture [13]. The AA input sequence is passed through the N Encoder blocks. The Random annotation input is passed through a multi-head attention layer in the decoder block and then joined with the output from the encoder blocks in a encoder-decoder multi-head attention layer. The encoder information acts as key (K) and value (V), while the multi-head attention of the annotation input as as query (Q). All N decoder blocks receive the same K and V from the encoder blocks, but the Q are different as they are decoded sequentially. After the N decoder blocks, a softmax layer is applied, allowing predictions of the 6 different annotation classes, which is then fed as input to the decoder again in an iterative fashion. Finally, a decoder block is applied, resulting in the annotation output. The loss is calculated across all six classes towards all 70 positions in the annotation output.

## 3 Results and Discussion

### 3.1 Attention focus

To analyse if the trained model has learned the common characteristics of SPs, we compare the bit-information from the attention on AAs used to annotate the SP with the occurrence of the AAs themselves, in the form of sequence logos. The peptides have been aligned at the CS, showing an area of the median SP length before the CS and three AAs after the CS for each respective kingdom and SP type (Figures 2 and S4, section 6.5). Only the TP CS predictions from the benchmark dataset are displayed, ranging from 50-67% for Sec/SPI (Table S4), 91-94% (Sec/SPII) and 46-56% (Tat/SPI), representing state-of-the-art performance (Figure S2). This means that the logos made, based on TP type predictions, will also contain information for informing the CSs.

The attention sequence logos represent the conservation and thus importance of specific amino acids at different positions relative to the cleavage site. However, the attention itself, regardless of amino acid, represents the most important positions overall. Previously, mutational studies have shown that residues -7 to -14 in prokaryotes and -6 to -13 in eukaryotes are the most important

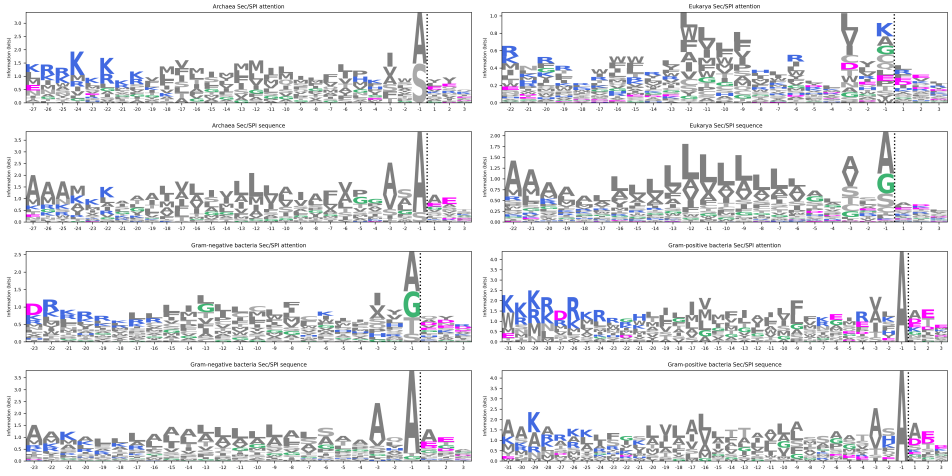


Figure 2: Attention and sequence logos for all Sec/SPI TP predictions in Archaea, Eukarya, Gram-negative and Gram-positive bacteria ordered in the N→C direction. The CS has been marked with a dashed black line. An area of the median SP length before the CS and three AAs after the CS is shown.

SP residues constituting a minimal hydrophobic region (14). These hydrophobic residues are the major determinants for signal recognition by the Signal recognition particle (SRP). However, the hydrophobic region is found not to be sufficient for this recognition alone, as positive charges act to fine-tune the SRP-SP affinity and targeting to the translocon. (15; 16)

As can be seen for Sec/SPI (Figure 2), the known tripartite structure has been learned. Towards the N-terminal, positive AAs obtain most attention, while the highly present alanines are disregarded. Towards the middle, the classic hydrophobic patch appears (5-16 AAs long) [24], both in sequence and attention logos. Right before the CS, motifs in positions -3 and -1 appear. The difference between attention and sequence and attention logos is stark here, displaying strong AxA motifs in the sequence and highly variable motifs in the attention logos. This suggests the network has learned the highly variable structure of signal peptides [14,17], allowing the possibility of a small or a polar or charged hydrophilic residue the same amount of attention. It is clear that the network has learned the importance of these positions due to the relative sum of the information there being consistently 3-6 times higher than the lowest. This is also true for the hydrophobic patch, known to be of various lengths and residues, as many different polar residues have high bit-information in the attention logos regardless of their statistical signal in the sequence logos.

The main attention is towards the hydrophobic region in Sec/SPI (Figure S4). The variability between kingdoms and range of attention focus reflects the known variability and length in the hydrophobic patch recognized by the SRP. Due to the continuity in the focus, it is hard to determine precise areas of focus for most kingdoms and types. For Sec/SPI, Archaea, the main attention focus ranges from -4 to -14, with the strongest focus on positions -7 and -9. In Eukarya, -6 to -17, with the strongest focus on position -7. In Gram-negative bacteria, the focus is wide, stretching from -2 to -17 with strongest focus on -2 and -7. In Gram-positive bacteria, the focus ranges from -7 to -16, with strong focus on positions -7,-8,-10 to -14.

## 4 Conclusions

The results show that analysing the encoder-decoder attention towards the CS provides explainable and meaningful biological insights. Calculating the bit-information in the attention matrix elucidates the tripartite structure and -3,-1 (Sec/SPI), cysteine (Sec/SPII) and Tat (Tat/SPI) motifs. Analysing the attention positional focus captures the most important SP residues constituting a minimal hydrophobic region for SRP interaction and the importance of the cysteine and Tat motifs. The high variability of SP sequences and lengths, especially of the important hydrophobic region, is also highlighted in the attention logos and attention positional focus. In addition, we train a transformer that is completely attention based, obtaining close to state-of-the-art performance.

## 5 Broader Impact

Recently, the use of language models for different protein related applications has increased substantially. It has been demonstrated that language models can be used in an unsupervised manner to learn protein features (12; 17) and for structure prediction (18) by pre-training on very large datasets. However, none of these studies have investigated the possibility to train on small targeted datasets in an end-to-end fashion. We show that it is possible to gain biological insights directly from small sets of labeled data by constructing a transformer neural network and analysing the encoder-decoder attention of this. It is possible that there exist unknown biological phenomena that can be learned in a similar fashion at a relatively small cost. What the trained transformer learns is highly dependent on the input data, why inherent biases in this will impact any conclusions drawn. It is therefore paramount that any conclusions are checked experimentally so that wrong conclusions are not propagated. We suggest this type of information extraction using neural networks should rather aim to ease the human workload in analysing biological data and not replace it.

## References


[1] V. Irihimovitch and J. Eichler, "Post-translational secretion of fusion proteins in the halophilic archaea haloferax volcanii," *J. Biol. Chem.*, vol. 278, Apr. 2003.

[2] G. Blobel and D. D. Sabatini, "Ribosome-Membrane interaction in eukaryotic cells," in *Biomembranes*, pp. 193–195, Springer, Boston, MA, 1971.

[3] H. Inouye and J. Beckwith, "Synthesis and processing of an escherichia coli alkaline phosphatase precursor in vitro," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 74, Apr. 1977.

[4] T. A. Rapoport, "Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes," *Nature*, vol. 450, pp. 663–669, Nov. 2007.

[5] H. Nielsen, K. D. Tsirigos, S. Brunak, and G. von Heijne, "A brief history of protein sorting prediction," *Protein J.*, vol. 38, pp. 200–216, May 2019.

[6] B. C. Berks, "The twin-arginine protein translocation pathway," *Annu. Rev. Biochem.*, vol. 84, 2015.

[7] G. von Heijne, "The structure of signal peptides from bacterial lipoproteins," *Protein Eng. Des. Sel.*, vol. 2, pp. 531–534, May 1989.

[8] Z. Szabó, A. O. Stahl, S.-V. Albers, J. C. Kissinger, A. J. M. Driessen, and M. Pohlschröder, "Identification of diverse archaeal proteins with class III signal peptides cleaved by distinct archaeal prepilin peptidases," *J. Bacteriol.*, vol. 189, pp. 772–778, Feb. 2007.

[9] G. von Heijne and C. Blomberg, "Trans-membrane translocation of proteins. the direct transfer model," *Eur. J. Biochem.*, vol. 97, pp. 175–181, June 1979.

[10] B. M. Austen, "Predicted secondary structures of amino-terminal extension sequences of secreted proteins," 1979.

[11] J. J. A. Armenteros, K. D. Tsirigos, C. K. Sønderby, T. N. Petersen, O. Winther, S. Brunak, G. von Heijne, and H. Nielsen, "SignalP 5.0 improves signal peptide predictions using deep neural networks," *Nat. Biotechnol.*, vol. 37, pp. 420–423, Feb. 2019.

[12] M. Heinzinger, A. Elnaggar, Y. Wang, C. Dallago, D. Nechaev, F. Matthes, and B. Rost, "Modeling aspects of the language of life through transfer-learning protein sequences," *BMC Bioinformatics*, vol. 20, pp. 1–17, Dec. 2019.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," June 2017.

[14] G. von Heijne, "Signal sequences: The limits of variation," *J. Mol. Biol.*, vol. 184, pp. 99–105, July 1985.

[15] N. Zheng and L. M. Gierasch, "Signal sequences: The same yet different," *Cell*, vol. 86, pp. 849–852, Sept. 1996.

[16] I. Nilsson, P. Lara, T. Hessa, A. E. Johnson, G. von Heijne, and A. L. Karamyshev, "The code for directing proteins for translocation across ER membrane: SRP cotranslationally recognizes specific features of a signal sequence," *J. Mol. Biol.*, vol. 427, p. 1191, Mar. 2015.

[17] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. Lawrence Zitnick, J. Ma, and R. Fergus, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 118, Apr. 2021.

[18] R. Chowdhury, N. Bouatta, S. Biswas, C. Rochereau, G. M. Church, P. K. Sorger, and M. AlQuraishi, "Single-sequence protein structure prediction using language models from deep learning." Aug. 2021.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," Dec. 2014.

[20] J. Duchi, "Adaptive subgradient methods for online learning and stochastic optimization," 2011.

[21] T.-Y. Lin, "Focal loss." `https://arxiv.org/pdf/1708.02002.pdf`, 2018. Accessed: 2021-3-22.

# 6 Supplementary material

## 6.1 Data

Table S1: Composition of the training set used for the 5-fold cross-validation procedure and the benchmark set in parentheses.

| Type | Archaea | Eukaryotes | Gram-negative bacteria | Gram-positive bacteria |
|---|---|---|---|---|
| Sec/SPI | 60 (50) | 2614 (210) | 509 (90) | 189 (25) |
| Sec/SPII | 28 (19) | N.A. | 1063 (442) | 449 (201) |
| Tat/SPI | 27 (22) | N.A. | 334 (98) | 95 (74) |
| Globular | 78 (63) | 13612 (6929) | 202 (103) | 140 (64) |
| Membrane | 44 (28) | 1044 (318) | 220 (50)6 | 50 (25) |
| Total | 237 (182) | 17270 (7457) | 2328 (783) | 923 (389) |

## 6.2 Optimization

The network was constructed using tensorflow version 2.5. The same nested cross-validation procedure as in SignalP(11) was used. Five homology partitions had been constructed on 20% sequence identity, using equal portions. Leaving each of these partitions out for testing, we trained four models on all 3+1 combinations of the remaining 4/5 of the data. For each validation run, we thus trained on 3/5 of the data and validated on 1/5. We did this in all combinations, in total 5 test partitions times 4 validation partitions, equalling 20 models in total. Using a grid-search approach, we optimised the following parameters in all combinations (in total 144 combinations):

```
embed dimensions = [16,32] (number of dimensions for embedding the amino acid and positions)

num heads = [1,2] (number of attention heads)

ff dim = [16,32] (number of dimensions in the feed-forward network)

num layers = [1,2,4] (number of transformer blocks)

batch sizes = [16,32] (batch size)

num iterations = [1,2,4] (number of iterations over the whole transformer)
```

All models were trained with a fixed learning rate of 0.001, using the Adam optimizer (19) with adagrad (20) and exponential decay (10000 decay steps, decay rate of 0.96, staircase=True) for 100 epochs. For the loss function we used focal loss (21) with $\gamma = 2$, a function where the misclassified examples are penalized, resulting in the model focusing on harder classes (see equation (i)).

$FL(pt) = -(1-p_t)^{\gamma} log(p_t)$ (i)

The best parameters on average over all four validation partitions, for each test partition, during the 100 epochs were chosen (Figure S1 and table S2). Using these parameters, we retrained the models for 50 epochs, as the models tended to overfit at this threshold, saving the ones with the best validation performance in these 50 epochs. The resulting 20 models were used for testing and benchmarking, using the average prediction of the four models for each test and benchmark partition. The signal peptide type was taken as the highest occurring signal annotation, if present. If no signal peptide annotation was present, the type was regarded as having no signal peptide. The cleavage site was evaluated only if a signal peptide type was predicted and in such cases taken as the last position of a signal annotation, starting from the N-terminal. A correct CS annotation was given in a window of ±3 AA around the annotated CS, just like in SignalP 5.0(11), as the experimental data has high variance.

## 6.3 Benchmark study

The transformer network does not outperform the best available methods (Figure S2). However, the results are comparable to the top methods across almost all comparisons, in most cases differing only to a small fraction. The highest performance is obtained for the Sec/SPII peptides and the lowest for Tat/SPI peptides. We note that SignalP, which dataset we have used in this study is outperformed in some regards (Gram-negative bacteria CS

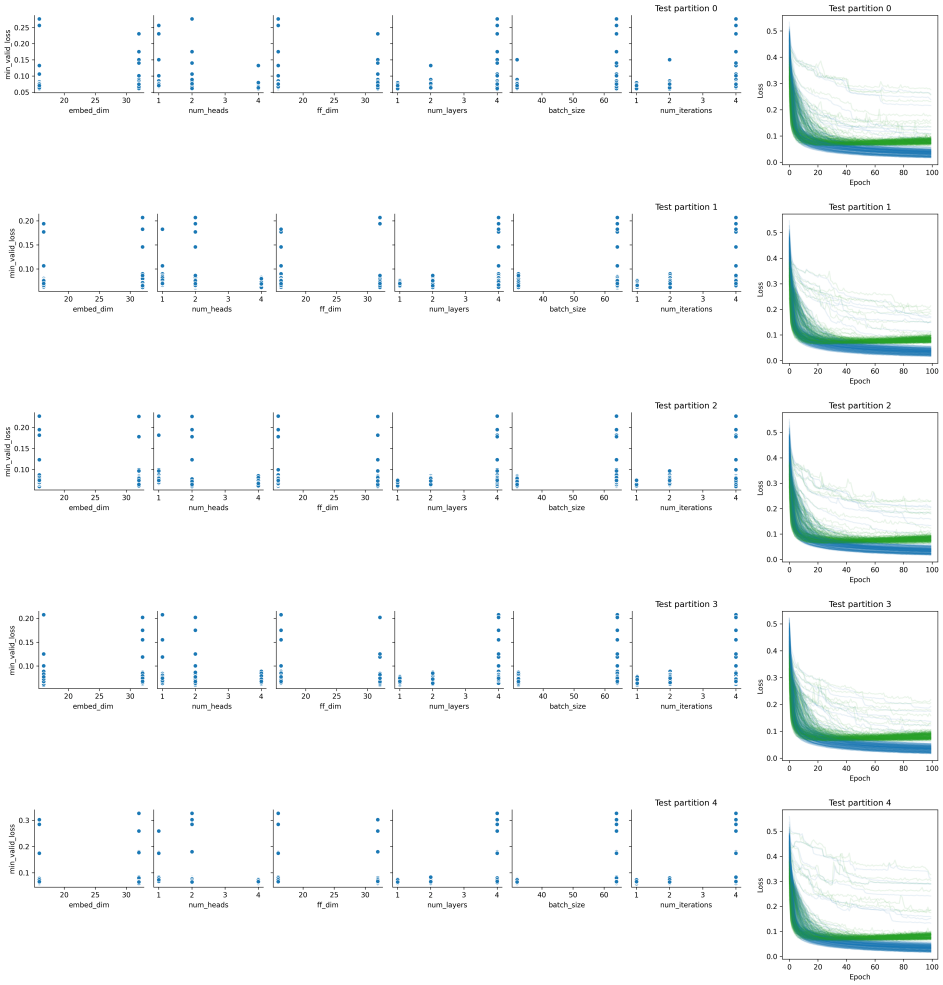


Figure S1: Optimisation results from the nested cross-validation for all five test partitions. The green curves represent the average validation performance and the blue the average train-performance for the four validation models for each parameter combination.

Table S2: The best parameter combinations over the validation for the five different test partitions.

| embed_dim | num_heads | ff_dim | num_layers | batch_size | num_iterations | test_partition | min_valid_loss |
|---|---|---|---|---|---|---|---|
| 32 | 2 | 32 | 4 | 64 | 1 | 0 | 59 |
| 32 | 2 | 16 | 1 | 64 | 1 | 1 | 61 |
| 16 | 2 | 32 | 1 | 32 | 1 | 2 | 60 |
| 16 | 2 | 32 | 1 | 32 | 1 | 3 | 60 |
| 16 | 2 | 32 | 1 | 32 | 1 | 4 | 59 |

precision and recall) by the TAT-type protein specific predictor PRED-TAT as well. The exact numbers in terms of cleavage site recall and precision and detection MCC can be found in tables S3-S9.

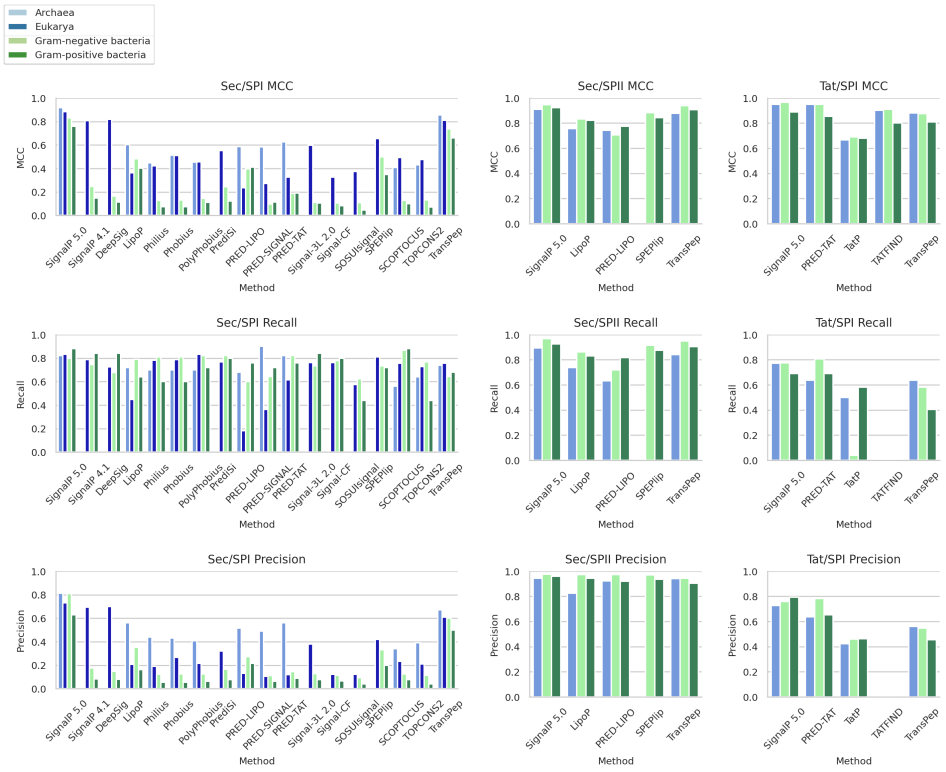


Figure S2: Benchmarking results for all signal peptides in terms of detection MCC, CS recall and CS precision. In the precision and recall, an error of ±3 positions have been allowed to handle potential errors in the experimental CS annotation. Our method, TransPep, does not outperform any of the best methods, although it is consistently among the top performing methods and equals these in many cases. For Sec/SPII the performance is almost identical with the best performance and for Tat/SPI the performance is worse in terms of CS recall and precision. For the exact numbers see tables S3-S9.

## 6.4 Encoder-Decoder attention

To see what information is used to make different decisions, we analysed the encoder-decoder attention of all correctly predicted peptides in the benchmark dataset. The maximum across all attention heads were taken for each position in the encoder-decoder attention matrix, resulting in a vector of 70x70 numbers for each peptide-annotation pair. To see what information the network uses to predict each type, we analysed the attention across the AA sequence towards an area of -25 to +3 amino acids around the true annotated cleavage sites and ±3 annotations, as well as the attention across the whole unaligned peptides. We build logos using the maximal amino acid column attention across all TP predictions with the python package Logomaker [23]. We added a pseudo probability of 0.0001 to the attention matrix to handle otherwise infinite negative logarithms. We also create logos from the raw AA sequences for comparison. All logo comparisons are made using bit-information (Shannon entropy).

## 6.5 Attention and sequence logos

For Eukarya, lysine (K) receives very high attention compared to its statistical signal in the -1 position in Sec/SPI. Only 2 out of 159 TP CS sequences in the benchmark dataset have a K in the -1 position. These are (the bold AAs represent the AAs used in the logos and the -1 K is underlined):

>A0A0F7YYX3|Conus victoriae (Queen Victoria cone)

MSS**PLRMDVTFLLAAIAVTWVCGL<u>K</u>IGF**PGFSTPPRSFIQHPKRTLCPEDCDIASPFKCEESPTCLRLFQ

>Q94K85|Arabidopsis thaliana (Mouse-ear cress)

MAVY**NTKLCLASVFLLLGLLLAFDLKGIE**AESLTKQKLDSKILQDEIVKKVNENPNAGWKAAINDRFSNA

Interestingly, these are both in plants and have SPs that are 3 (25 AAs) resp 4 residues (26 AAs) longer than the median Sec/SPI SP (22 AAs) in Eukarya. This is within one standard deviation (4.9 AAs), suggesting that it is the plant type itself and not the length of the SP that is important for the high attention. That these would both be in plants and the K not having a biological function is highly unlikely, as only 16/210 (7.6%, 171 Metazoa, 20 Fungi, 3 unclassified) of the Sec/SPI SPs are in plants.

Figure S3 displays the same comparison as for Sec/SPI in Figure 2 for Sec/SPII and Tat/SPI. Again, the tripartite structure and disregard for N-terminal alanines manifests itself. The Tat signals (R-R) are learned and given much attention for the Tat/SPI SPs. Both negatively and positively charged amino acids are given importance towards the N-terminal, suggesting the importance of the charge and hydrophilicity itself. The charged/hydrophilic N-terminal region is less present for Sec/SPII in Bacteria, where the +1 cysteine motif [7] is given almost all attention both in attention and sequence logos. The network has learned to provide more even attention to the -1 residues, showing the importance and variability of AAs in that position and disregard for statistical bias. Showing similar information for G,S,A and I in Gram-negative bacteria and towards G,S and A in Gram-positive enables the network to find the CS despite the variation. Just as for Sec/SPI, the network gives similar attention to a variation of residues for the -1,-3 motif in Tat/SPI, avoiding the strong statistical alanine bias. The -1,-3 motif seems much less important for Archaea though, especially the -1 signal is barely visible in both sequence and attention logos, where a strong hydrophobic patch obtains most information instead.

In the Tat/SPI SPs, the -1 arginine (R) and lysine (K) obtain much attention, compared to their statistical signals in the sequence logos. For R, his may be due to the importance of the R-R motif for recognition.

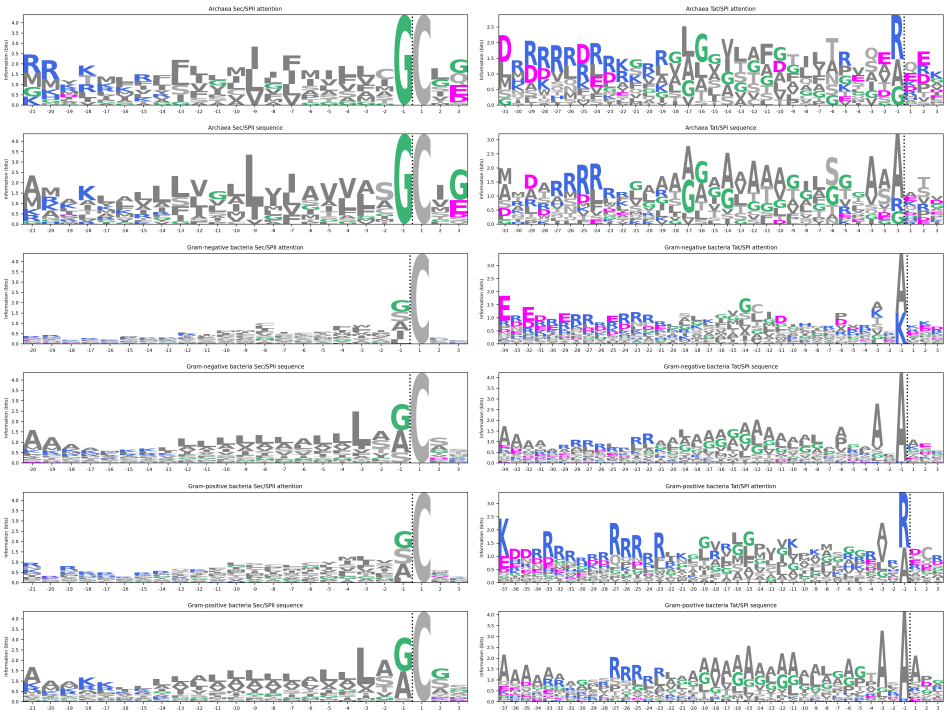


Figure S3: Attention and sequence logos for all Sec/SPII and Tat/SPI TP predictions in Archaea, Gram-negative and Gram-positive bacteria ordered in the N→C direction. The CS has been marked with a dashed black line. An area of the median SP length before the CS and three AAs after the CS is shown.

## 6.6 Attention focus

For Sec/SPII and Tat/SPI (Figure S5) as compared to Sec/SPI (Figure S4), the focus is entirely different as these types are more motif-driven with the highly conserved cysteine and R-R motifs. Due to almost all attention being towards the +1 cysteine motif in Sec/SPII, the attention towards the hydrophobic patch is harder to spot. In Eukarya, this focus is wide with the strongest signal towards -2 to -10. In both Gram-positive bacteria the focus is towards -3 to -15 and in Gram-negative bacteria -3 to -18. For Tat/SPI, the focus varies between the hydrophobic patch and the C-terminal where the R-R motif is to be found. In bacteria, there is also much focus on position -2, while there is no such focus in Archaea.

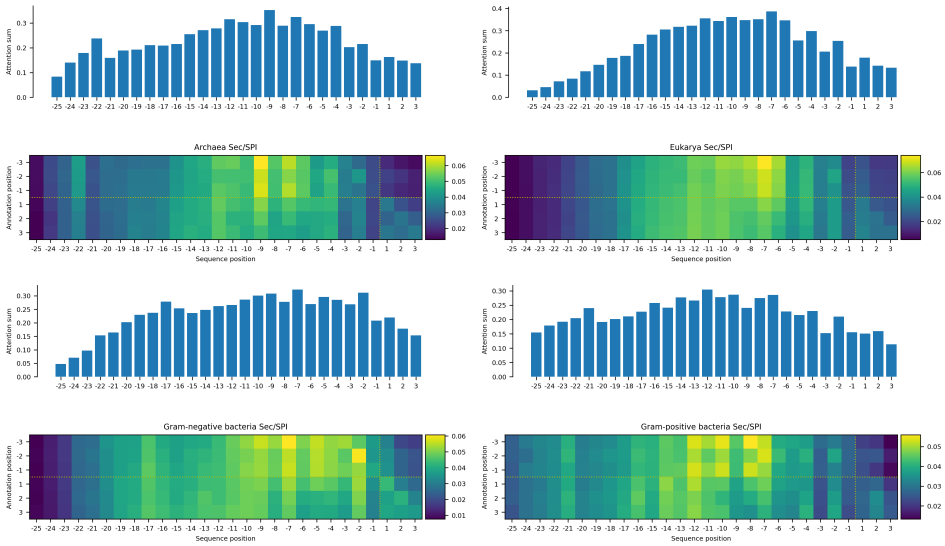


Figure S4: The average attention focus per position (matrix) and sum (bar chart) for all TP CS predictions in Sec/SPI.

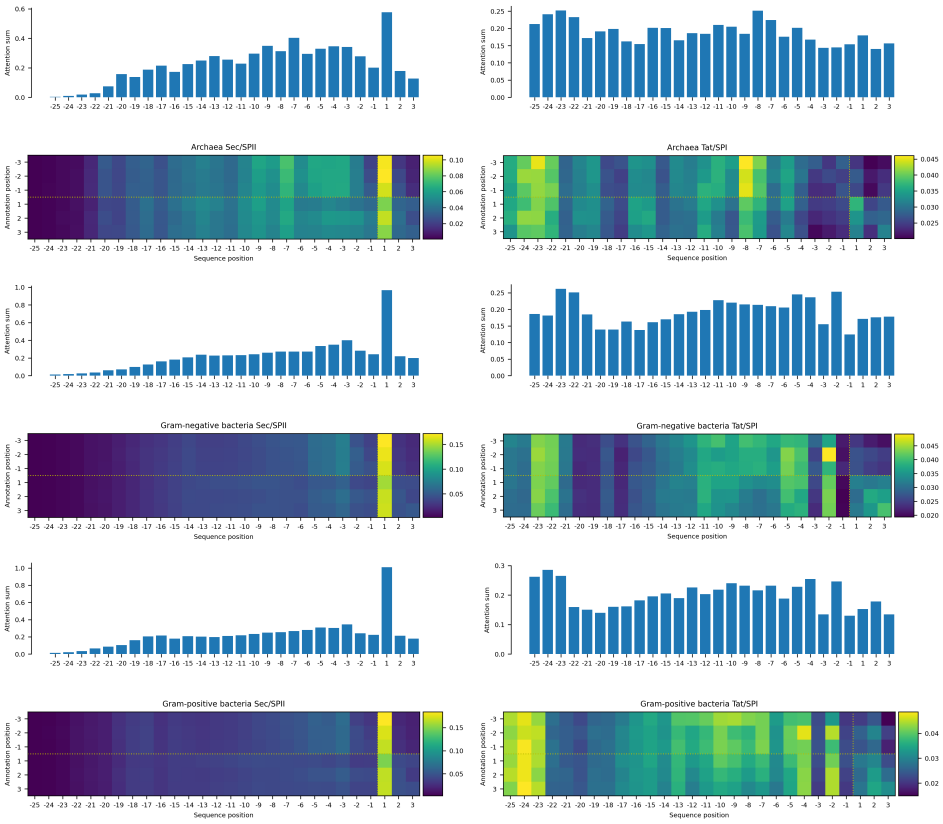


Figure S5: The average attention focus per position (matrix) and sum (bar chart) for all TP CS predictions in Sec/SPII and Tat/SPI.

### 6.7 Tables for the benchmark study

Table S3: Sec/SPI detection MCC

| Method | Archaea | Eukaryotes | Gram-negative bacteria | Gram-positive bacteria |
|---|---|---|---|---|
| SignalP 5.0 | 0.917 | 0.883 | 0.83 | 0.76 |
| SignalP 4.1 | n.d | 0.808 | 0.248 | 0.148 |
| DeepSig | n.d | 0.819 | 0.166 | 0.115 |
| LipoP | 0.604 | 0.363 | 0.483 | 0.403 |
| Philius | 0.447 | 0.421 | 0.127 | 0.075 |
| Phobius | 0.514 | 0.51 | 0.132 | 0.074 |
| PolyPhobius | 0.453 | 0.456 | 0.144 | 0.111 |
| PrediSi | n.d. | 0.553 | 0.244 | 0.121 |
| PRED-LIPO | 0.586 | 0.234 | 0.398 | 0.41 |
| PRED-SIGNAL | 0.584 | 0.272 | 0.098 | 0.114 |
| PRED-TAT | 0.626 | 0.326 | 0.187 | 0.189 |
| Signal-3L 2.0 | n.d. | 0.597 | 0.11 | 0.106 |
| Signal-CF | n.d. | 0.326 | 0.106 | 0.084 |
| SOSUIsignal | n.d. | 0.375 | 0.108 | 0.047 |
| SPEPlip | n.d. | 0.655 | 0.498 | 0.35 |
| SCOPTOCUS | 0.408 | 0.492 | 0.127 | 0.1 |
| TOPCONS2 | 0.432 | 0.477 | 0.131 | 0.071 |
| TransPep | 0.855 | 0.81 | 0.737 | 0.661 |

Table S4: Sec/SPI CS recall. The 0,±1,±2,±3 columns indicate the allowed error in CS annotation.

| Method | Archaea | | | | Eukaryotes | | | |
|---|---|---|---|---|---|---|---|---|
| | 0 | ±1 | ±2 | ±3 | 0 | ±1 | ±2 | ±3 |
| SignalP 5.0 | 0.66 | 0.74 | 0.78 | 0.82 | 0.729 | 0.762 | 0.795 | 0.833 |
| SignalP 4.1 | n.d. | n.d. | n.d. | n.d. | 0.695 | 0.729 | 0.762 | 0.786 |
| DeepSig | n.d. | n.d. | n.d. | n.d. | 0.624 | 0.652 | 0.69 | 0.724 |
| LipoP | 0.48 | 0.62 | 0.66 | 0.72 | 0.343 | 0.386 | 0.419 | 0.448 |
| Philius | 0.58 | 0.68 | 0.8 | 0.7 | 0.619 | 0.686 | 0.743 | 0.781 |
| Phobius | 0.54 | 0.64 | 0.66 | 0.7 | 0.667 | 0.7 | 0.738 | 0.786 |
| PolyPhobius | 0.56 | 0.68 | 0.68 | 0.7 | 0.681 | 0.733 | 0.776 | 0.833 |
| PrediSi | n.d. | n.d. | n.d. | n.d. | 0.652 | 0.695 | 0.719 | 0.767 |
| PRED-LIPO | 0.48 | 0.6 | 0.66 | 0.68 | 0.095 | 0.114 | 0.152 | 0.181 |
| PRED-SIGNAL | 0.8 | 0.9 | 0.9 | 0.9 | 0.224 | 0.29 | 0.329 | 0.362 |
| PRED-TAT | 0.58 | 0.72 | 0.0 | 0.82 | 0.41 | 0.51 | 0.571 | 0.614 |
| Signal-3L 2.0 | n.d. | n.d. | n.d. | n.d. | 0.648 | 0.686 | 0.733 | 0.762 |
| Signal-CF | n.d. | n.d. | n.d. | n.d. | 0.652 | 0.676 | 0.724 | 0.762 |
| SOSUIsignal | n.d. | n.d. | n.d. | n.d. | 0.176 | 0.329 | 0.467 | 0.576 |
| SPEPlip | n.d. | n.d. | n.d. | n.d. | 0.71 | 0.733 | 0.771 | 0.81 |
| SCOPTOCUS | 0.34 | 0.48 | 0.52 | 0.56 | 0.39 | 0.533 | 0.686 | 0.757 |
| TOPCONS2 | 0.48 | 0.6 | 0.62 | 0.64 | 0.371 | 0.505 | 0.638 | 0.729 |
| TransPep | 0.26 | 0.52 | 0.7 | 0.74 | 0.252 | 0.467 | 0.686 | 0.757 |
| Method | Gram-negative bacteria | | | | Gram-positive bacteria | | | |
| | 0 | ±1 | ±2 | ±3 | 0 | ±1 | ±2 | ±3 |
| SignalP 5.0 | 0.66 | 0.74 | 0.78 | 0.82 | 0.729 | 0.762 | 0.795 | 0.833 |
| SignalP 4.1 | n.d. | n.d. | n.d. | n.d. | 0.695 | 0.729 | 0.762 | 0.786 |
| DeepSig | n.d. | n.d. | n.d. | n.d. | 0.624 | 0.652 | 0.69 | 0.724 |
| LipoP | 0.48 | 0.62 | 0.66 | 0.72 | 0.343 | 0.386 | 0.419 | 0.448 |
| Philius | 0.58 | 0.68 | 0.8 | 0.7 | 0.619 | 0.686 | 0.743 | 0.781 |
| Phobius | 0.54 | 0.64 | 0.66 | 0.7 | 0.667 | 0.7 | 0.738 | 0.786 |
| PolyPhobius | 0.56 | 0.68 | 0.68 | 0.7 | 0.681 | 0.733 | 0.776 | 0.833 |
| PrediSi | n.d. | n.d. | n.d. | n.d. | 0.652 | 0.695 | 0.719 | 0.767 |
| PRED-LIPO | 0.48 | 0.6 | 0.66 | 0.68 | 0.095 | 0.114 | 0.152 | 0.181 |
| PRED-SIGNAL | 0.8 | 0.9 | 0.9 | 0.9 | 0.224 | 0.29 | 0.329 | 0.362 |
| PRED-TAT | 0.58 | 0.72 | 0.0 | 0.82 | 0.41 | 0.51 | 0.571 | 0.614 |
| Signal-3L 2.0 | n.d. | n.d. | n.d. | n.d. | 0.648 | 0.686 | 0.733 | 0.762 |
| Signal-CF | n.d. | n.d. | n.d. | n.d. | 0.652 | 0.676 | 0.724 | 0.762 |
| SOSUIsignal | n.d. | n.d. | n.d. | n.d. | 0.176 | 0.329 | 0.467 | 0.576 |
| SPEPlip | n.d. | n.d. | n.d. | n.d. | 0.71 | 0.733 | 0.771 | 0.81 |
| SCOPTOCUS | 0.34 | 0.48 | 0.52 | 0.56 | 0.39 | 0.533 | 0.686 | 0.757 |
| TOPCONS2 | 0.48 | 0.6 | 0.62 | 0.64 | 0.371 | 0.505 | 0.638 | 0.729 |
| TransPep | 0.26 | 0.52 | 0.7 | 0.74 | 0.252 | 0.467 | 0.686 | 0.757 |

Table S5: Sec/SPI CS precision. The 0,±1,±2,±3 columns indicate the allowed error in CS annotation.

| Method | Archaea | | | | Eukaryotes | | | |
|---|---|---|---|---|---|---|---|---|
| | 0 | ±1 | ±2 | ±3 | 0 | ±1 | ±2 | ±3 |
| SignalP 5.0 | 0.771 | 0.688 | 0.812 | 0.812 | 0.671 | 0.702 | 0.732 | 0.732 |
| SignalP 4.1 | n.d. | n.d. | n.d. | n.d. | 0.613 | 0.643 | 0.672 | 0.693 |
| DeepSig | n.d. | n.d. | n.d. | n.d. | 0.604 | 0.631 | 0.668 | 0.7 |
| LipoP | 0.484 | 0.375 | 0.516 | 0.562 | 0.159 | 0.178 | 0.194 | 0.207 |
| Philius | 0.425 | 0.362 | 0.438 | 0.438 | 0.151 | 0.168 | 0.182 | 0.191 |
| Phobius | 0.395 | 0.333 | 0.407 | 0.432 | 0.226 | 0.237 | 0.25 | 0.267 |
| PolyPhobius | 0.395 | 0.326 | 0.395 | 0.407 | 0.176 | 0.19 | 0.201 | 0.216 |
| PrediSi | n.d. | n.d. | n.d. | n.d. | 0.273 | 0.291 | 0.301 | 0.321 |
| PRED-LIPO | 0.455 | 0.364 | 0.5 | 0.515 | 0.069 | 0.083 | 0.11 | 0.131 |
| PRED-SIGNAL | 0.435 | 0.489 | 0.489 | 0.489 | 0.066 | 0.085 | 0.096 | 0.106 |
| PRED-TAT | 0.397 | 0.493 | 0.548 | 0.562 | 0.08 | 0.099 | 0.111 | 0.119 |
| Signal-3L 2.0 | n.d. | n.d. | n.d. | n.d. | 0.322 | 0.341 | 0.365 | 0.379 |
| Signal-CF | n.d. | n.d. | n.d. | n.d. | 0.105 | 0.109 | 0.117 | 0.123 |
| SOSUIsignal | n.d. | n.d. | n.d. | n.d. | 0.037 | 0.069 | 0.098 | 0.121 |
| SPEPlip | n.d. | n.d. | n.d. | n.d. | 0.366 | 0.378 | 0.398 | 0.418 |
| SCOPTOCUS | 0.207 | 0.293 | 0.317 | 0.341 | 0.12 | 0.164 | 0.211 | 0.233 |
| TOPCONS2 | 0.293 | 0.366 | 0.378 | 0.39 | 0.107 | 0.146 | 0.184 | 0.21 |
| TransPep | 0.236 | 0.472 | 0.636 | 0.672 | 0.203 | 0.375 | 0.551 | 0.609 |
| Method | Gram-negative bacteria | | | | Gram-positive bacteria | | | |
| | 0 | ±1 | ±2 | ±3 | 0 | ±1 | ±2 | ±3 |
| SignalP 5.0 | 0.742 | 0.775 | 0.809 | 0.809 | 0.6 | 0.6 | 0.629 | 0.629 |
| SignalP 4.1 | 0.151 | 0.167 | 0.172 | 0.175 | 0.083 | 0.083 | 0.083 | 0.083 |
| DeepSig | 0.131 | 0.144 | 0.146 | 0.148 | 0.073 | 0.073 | 0.08 | 0.08 |
| LipoP | 0.327 | 0.342 | 0.351 | 0.351 | 0.153 | 0.153 | 0.163 | 0.163 |
| Philius | 0.106 | 0.112 | 0.119 | 0.122 | 0.054 | 0.054 | 0.054 | 0.054 |
| Phobius | 0.098 | 0.11 | 0.12 | 0.124 | 0.054 | 0.054 | 0.054 | 0.054 |
| PolyPhobius | 0.097 | 0.11 | 0.122 | 0.124 | 0.06 | 0.06 | 0.063 | 0.063 |
| PrediSi | 0.144 | 0.157 | 0.162 | 0.164 | 0.062 | 0.062 | 74.0 | 0.078 |
| PRED-LIPO | 0.212 | 0.237 | 0.258 | 0.273 | 0.216 | 0.216 | 0.216 | 0.216 |
| PRED-SIGNAL | 0.076 | 0.089 | 0.106 | 0.11 | 0.06 | 0.06 | 0.064 | 0.064 |
| PRED-TAT | 0.125 | 0.135 | 0.141 | 0.145 | 0.082 | 0.082 | 0.087 | 0.087 |
| Signal-3L 2.0 | 0.113 | 0.123 | 0.127 | 0.129 | 0.074 | 0.074 | 0.08 | 0.078 |
| Signal-CF | 0.102 | 0.105 | 0.11 | 0.115 | 0.059 | 0.059 | 0.065 | 0.065 |
| SOSUIsignal | 0.04 | 0.055 | 0.086 | 0.094 | 0.018 | 0.021 | 0.025 | 0.039 |
| SPEPlip | 0.276 | 0.307 | 0.327 | 0.332 | 0.187 | 0.187 | 0.198 | 0.198 |
| SCOPTOCUS | 0.067 | 0.098 | 0.119 | 0.124 | 0.056 | 0.066 | 0.07 | 0.077 |
| TOPCONS2 | 0.081 | 0.093 | 0.11 | 0.115 | 0.022 | 0.029 | 0.036 | 0.039 |
| TransPep | 0.232 | 0.389 | 0.526 | 0.6 | 0.147 | 0.324 | 0.441 | 0.5 |

Table S6: Sec/SPII detection MCC.

| Method | Archaea | Gram-negative bacteria | Gram-positive bacteria |
|---|---|---|---|
| SignalP 5.0 | 0.91 | 0.946 | 0.923 |
| LipoP | 0.755 | 0.833 | 0.822 |
| PRED-LIPO | 0.743 | 0.707 | 0.775 |
| SPEPlip | n.d. | 0.884 | 0.843 |
| TransPep | 0.878 | 0.938 | 0.907 |

Table S7: Sec/SPII recall and precision. The 0,±1,±2,±3 columns indicate the allowed error in CS annotation. G- and G+ are shortenings for Gram-negative and Gram-positive bacteria respectively.

| Method | Archaea | | | | G- | | | | G+ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | ±1 | ±2 | ±3 | 0 | ±1 | ±2 | ±3 | 0 | ±1 | ±2 | ±3 |
| CS recall | | | | | | | | | | | | |
| SignalP | 0.895 | 0.895 | 0.895 | 0.895 | 0.964 | 0.964 | 0.964 | 0.968 | 0.925 | 0.925 | 0.925 | 0.925 |
| LipoP | 0.684 | 0,,684 | 0.737 | 0.737 | 0.86 | 0.86 | 0.86 | 0.862 | 0.831 | 0.831 | 0.831 | 0.831 |
| PRED-LIPO | 0.632 | 0.632 | 0.632 | 0.632 | 0.717 | 0.717 | 0.717 | 0.719 | 0.816 | 0.816 | 0.816 | 0.816 |
| SPEPlip | n.d. | n.d. | n.d. | n.d. | 0.912 | 0.912 | 0.914 | 0.914 | 0.876 | 0.876 | 0.876 | 0.876 |
| TransPep | 0.684 | 0.789 | 0.842 | 0.842 | 0.672 | 0.749 | 0.912 | 0.948 | 0.682 | 0.736 | 0.866 | 0.905 |
| CS precision | | | | | | | | | | | | |
| SignalP | 0.944 | 0.944 | 0.944 | 0.944 | 0.97 | 0.97 | 0.97 | 0.975 | 0.959 | 0.959 | 0.959 | 0.959 |
| LipoP | 0.765 | 0.765 | 0.824 | 0.824 | 0.969 | 0.969 | 0.969 | 0.972 | 0.944 | 0.944 | 0.944 | 0.944 |
| PRED-LIPO | 0.923 | 0.923 | 0.923 | 0.923 | 0.969 | 0.969 | 0.969 | 0.972 | 0.921 | 0.921 | 0.921 | 0.921 |
| SPEPlip | n.d. | n.d. | n.d. | n.d. | 0.969 | 0.969 | 0.971 | 0.971 | 0.936 | 0.936 | 0.936 | 0.936 |
| TransPep | 0.765 | 0.882 | 0.941 | 0.941 | 0.669 | 0.745 | 0.908 | 0.944 | 0.682 | 0.736 | 0.866 | 0.905 |

Table S8: Tat/SPI detection MCC.

| Method | Archaea | Gram-negative bacteria | Gram-positive bacteria |
|---|---|---|---|
| SignalP 5.0 | 0.948 | 0.965 | 0.889 |
| PRED-TAT | 0.948 | 0.948 | 0.853 |
| TatP | 0.667 | 0.689 | 0.68 |
| TATFIND | 0.902 | 0.91 | 0.8 |
| TransPep | 0.88 | 0.875 | 0.81 |

Table S9: Tat/SPI CS recall and CS precision. The 0,±1,±2,±3 columns indicate the allowed error in CS annotation. G- and G+ are shortenings for Gram-negative and Gram-positive bacteria respectively.

| Method | Archaea | | | | G- | | | | G+ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | ±1 | ±2 | ±3 | 0 | ±1 | ±2 | ±3 | 0 | ±1 | ±2 | ±3 |
| CS recall | | | | | | | | | | | | |
| SignalP 5.0 | 0.591 | 0.636 | 0.727 | 0.773 | 0.684 | 0.724 | 0.745 | 0.776 | 0.595 | 0.622 | 0.676 | 0.689 |
| PRED-TAT | 0.5 | 0.545 | 0.636 | 0.636 | 0.735 | 0.735 | 0.776 | 0.806 | 0.622 | 0.622 | 0.635 | 0.689 |
| TatP | 0.318 | 0.409 | 0.5 | 0.5 | 0.653 | 0.673 | 0.694 | 0.04 | 0.446 | 0.43 | 0.514 | 0.581 |
| TATFIND | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. |
| TransPep | 0.227 | 0.364 | 0.5 | 0.636 | 0.204 | 0.388 | 0.5 | 0.582 | 0.122 | 0.257 | 0.351 | 0.405 |
| CS precision | | | | | | | | | | | | |
| SignalP 5.0 | 0.591 | 0.636 | 0.727 | 0.727 | 0.698 | 0.74 | 0.76 | 0.76 | 0.698 | 0.73 | 0.794 | 0.794 |
| PRED-TAT | 0.5 | 0.545 | 0.636 | 0.636 | 0.713 | 0.733 | 0.752 | 0.782 | 0.59 | 0.59 | 0.603 | 0.654 |
| TatP | 0.269 | 0.346 | 0.423 | 0.423 | 0.427 | 0.44 | 0.453 | 0.46 | 0.355 | 0.376 | 0.409 | 0.462 |
| TATFIND | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. | n.d. |
| TransPep | 0.2 | 0.302 | 0.44 | 0.56 | 0.192 | 0.365 | 0.471 | 0.548 | 0.136 | 0.288 | 0.394 | 0.455 |

Table S10: Precision for each SP type in all four kingdoms. The precision represents the fraction of all positive types that are displayed in the attention and sequence logos from the benchmark dataset.

| Type | Archaea | Eukaryotes | Gram-negative bacteria | Gram-positive bacteria |
|---|---|---|---|---|
| Sec/SPI | 0.94 | 0.914 | 0.778 | 0.8 |
| Tat/SPI | 0.842 | N.A. | 0.918 | 0.797 |
| Sec/SPII | 0.955 | N.A. | 0.973 | 0.955 |