T4SEpp: a pipeline integrated with protein language models effectively predicting bacterial type IV secreted effectors

Yueming Hu[1,#], Yejun Wang[2,3,#], Xiaotian Hu[1], Haoyu Chao[1], Sida Li[1], Qinyang Ni[1], Yanyan Zhu[1], Yixue Hu[2], Ziyi Zhao[2], Ming Chen[1,*].

[1] Department of Bioinformatics, College of Life Sciences, Zhejiang University, Hangzhou, China.

[2] Youth Innovation Team of Medical Bioinformatics, Shenzhen University Medical School, Shenzhen, China.

[3] Department of Cell Biology and Genetics, College of Basic Medicine, Shenzhen University Medical School, Shenzhen, China.

[#] Y.H. and Y.W. contributed equally to this study.

[*]Correspondence: mchen@zju.edu.cn (M.C.); Tel./Fax: +86-(0)571-88206612 (M.C.)

Abstract

Many pathogenic bacteria use type IV secretion systems(T4SSs) to deliver effectors (T4SEs) into the cytoplasm of eukaryotic cells, causeing diseases. The identification of effectors is a crucial step in understanding the mechanisms of bacterial pathogenicity, but this remains a major challenge. In this study, we used the full-length embedding features generated by six pre-trained protein language models to train classifiers predicting T4SEs, and compared their performance. An integrated model T4SEpp was assembled by a module searching full-length, signal sequence and effector domain homologs of known T4SEs, a machine learning module based on the hand-crafted features extracted from the signal sequences, and the third module containing three best-performing protein language pre-trained models. T4SEpp outperformed the other state-of-the-art (SOTA) software tools, achieving ~0.95 sensitivity at a high specificity of ~0.99, based on the assessment of an independent testing dataset. Additionally, we performed a comprehensive search among 8,761 bacterial species, leading to the discovery of 227 species belonging to 3 phyla and 117 genera that possess T4SSs. Furthermore, leveraging the power of T4SEpp, we successfully identified a grand total of 12,622 plausible T4SEs. Overall, T4SEpp provides a better solution to assist in the identification of bacterial T4SEs, and facilitates studies of bacterial pathogenicity. T4SEpp is freely accessible at https://bis.zju.edu.cn/T4SEpp.

Key words: T4SEpp; Type IV Secreted Effector; Deep Learning; Protein Language Model; Prediction

Introduction

Gram-negative bacteria employ more than one dozen of secretion systems to transport proteins out of the cell envelope[1, 2]. Among them, the type IV secretion system (T4SS) is a complex molecular machine spanning both the inner and outer membranes, and translocate substrate proteins into eukaryotic host cells in only one step[3-9]. Protein-translocating T4SSs can be divided into two major families according to the composition of component elements: type IVA, exemplified by the *A. tumfaciens* VirB/VirD4 T4SS and *H. pylori* Cag T4SS, and type IVB exemplified by *Legionella* Dot/Icm T4SS[9]. Substrate proteins translocated by T4SSs, also called effectors, play important roles in bacterial infections and pathogenicity[1, 10, 11].

Effectors of T4SSs (T4SEs) are transported directly or as complexes with DNA in many pathogenic bacteria, such as *Helicobacter pylori*, *Legionella pneumophila*, *Bordetella pertussis*, *Coxiella*, *Brucella*, and *Bartonella*[12-17]. T4SS-mediated entry of effector proteins into recipient cells is contact-dependent[18]. Once they enter the eukaryotic host cytoplasm, they disrupt signal transduction and cause various host diseases. Identifying these effectors is crucial for understanding the mechanisms of infection and pathogenicity caused by these bacteria. However, because the composition and sequences vary significantly, it is challenging to identify new T4SEs experimentally. Although many T4SEs have been identified and characterized in a few model organisms[19-22], the exact mechanism remains unclear.

Since 2009 when the first machine-learning algorithm was introduced, tens of computational models have been developed to predict T4SEs[2, 23]. Early algorithms were mainly species-specific, such as those predicting T4SEs in *Legionella pneumophila*[23]. In another study, Wang *et al.* developed an SVM-based model, T4SEpre, which exhibited good overall and cross-species performance[24]. However, T4SEpre only considers the features buried in the C-terminal 100 amino acids[24]. More studies, especially ensemble models recently developed with multi-aspect features, learn features from full-length proteins to improve performance[25, 26]. Deep learning algorithms have also been applied in for the prediction of T4SEs. For example, CNN-T4SE integrated three convolutional neural network (CNN) models to learn the features of amino acid composition, solvent accessibility, and secondary structure of the full-length T4SEs[27]. T4SEfinder is a multi-layer perception (MLP) model that learns the features generated by a pre-trained BERT model[28], which can predict T4SEs accurately[29]. Notably, BERT is a natural language processing (NLP) model that is appealing in biology and other fields[30-35]. NLP models have been successfully applied to the prediction of protein subcellular localization[31, 32], secondary structure[32, 33, 35], and others[34]. Besides T4SEfinder, the NLP-based pre-trained transformers have also been used for the prediction of bacterial type III secreted effectors and Sec/Tat substrates, both achieving superior performance[36, 37].

Although machine learning strategies have achieved some success in the identification of T4SEs[2, 23, 24], the high false-positive rate has been a big challenge. To reduce the false-positive rate in predicting type III effectors, Hui et al. proposed a strategy to combine machine learning models with homology searching, and integrate multiple modules considering the multi-aspect biological features of the effector genes[38]. To improve model performance, other models have also considered the multiple features and a combination of homology-based strategies in the prediction of type III effectors[39-41]. For T4SE prediction, homology searching was also been applied independently. For example, S4TE integrates 13 sequence homology-based features, including homology to known effectors, homology to eukaryotic domains, presence of subcellular localization signals, and secretion signals, and develops a scoring scheme to predict T4SEs mainly from α- and γ-proteobacteria[42]. Despite the high precision, the sensitivity could be influenced by the large diversity of T4SE composition and sequences. Therefore, it could be a better solution to take the advantages of both machine learning approaches, especially ensemblers, and homology-based methods, designing an integrated T4SE prediction pipeline that combines various models and comprehensively considers various characteristics of effector sequences.

In this study, we proposed a hybrid strategy for predicting T4SEs. First, a homology searching strategy scanned both the global homology of full-length proteins and the local homology of domains to known effectors. Additionally, we retrained a machine learning module T4SEpre[24] with updated T4SE data and hand-crafted amino acid composition features in the C-termini. Furthermore, a group of transfer learning models was developed based on the features generated by various pretrained transformers. For the transfer learning models, we utilized the deep context protein language models ESM-1b, ProtBert, ProtT5-XL, and ProtAlbert to represent protein sequence features[32, 33]. These features can characterize the intrinsic but unclear properties of protein sequences and the interactions between positions. Based on these feature representations, application models were developed to classify T4SEs using a deep neural network architecture with an attention mechanism. Finally, we integrated the homology-based modules, machine learning models based on traditional handcrafted features, and transfer learning models with transformer-generated features into a pipeline, namely T4SEpp, which assembles the individual modules in a linear function to generate a prediction score reflecting the likelihood of a protein to be a T4SE. A web application for T4SEpp is also available via the link: https://bis.zju.edu.cn/T4SEpp.

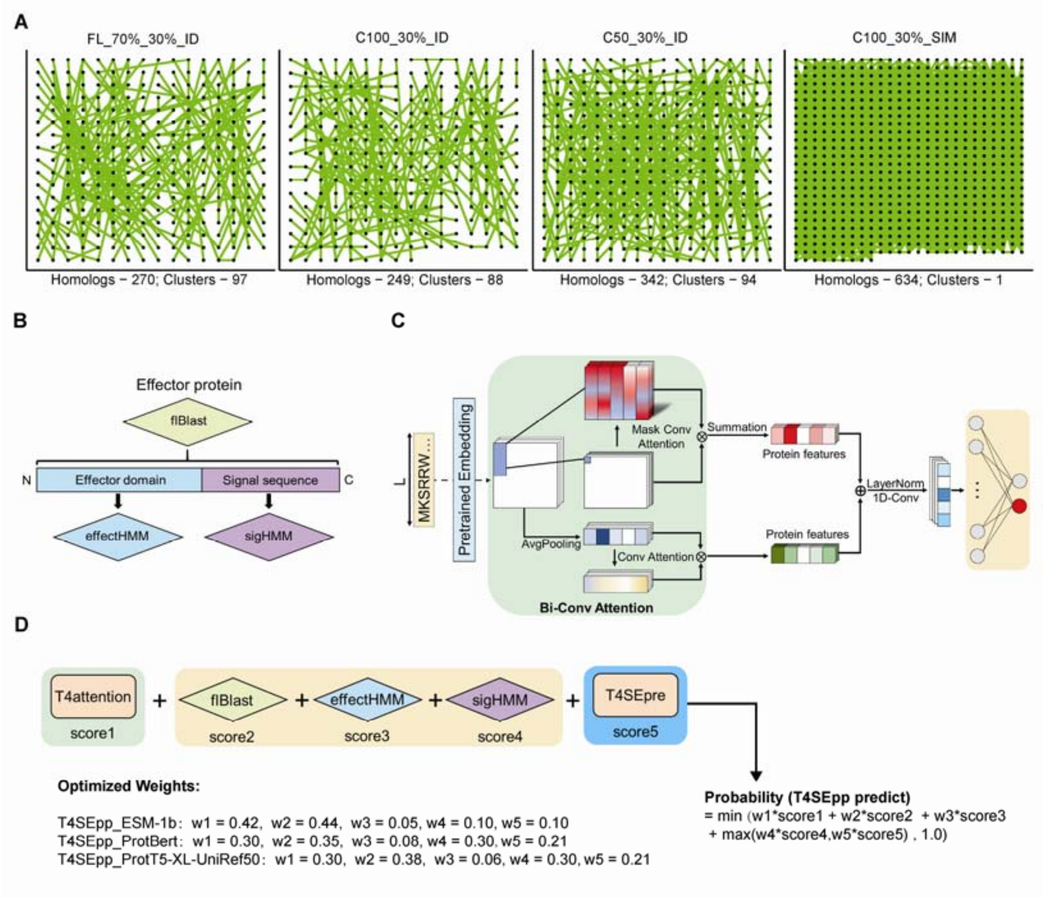
Results

Sequence homology among verified effectors and the integrated prediction framework

Experimentally verified effectors were collected from literature and databases, and 653 proteins were obtained after removing redundant sequences, representing the latest and most comprehensive list of experimentally verified T4SEs[26, 43] (see Materials and Methods). Pairwise sequence alignments of full-length (FL) effector proteins or their C-terminal peptides of 100 or 50 amino acids (C100 or C50, respectively) were performed. For the FL proteins, 481 non-homologous clusters were identified after homology filtering for the proteins with > 30% identity and > 70% length coverage of the pair of proteins (FL_70%_30%_ID) (Figure 1A). However, for the C100 sequences, 249 were homologous to others with an identity of > 30%, and 473 non-redundant clusters were retained from these sequences after homology filtering (C100_30%_ID) (Figure 1A). The reduction in the number of clusters indicated that the C-terminal 100 amino acids showed more homology than the full-length effector proteins, but there were no significant differences between them (473/654 vs. 481/654, EBT *P*= 0.614). The C50 sequences further reflected the typical C-terminal homology between effectors. A total of 342 peptides were found to have homology with the others, while 401 clusters remained for these peptides after homology filtering (C50_30%_ID, 401/654 vs. 481/654, EBT *P*=3.17e-03) (Figure 1A). Rigorous homology filtering is a prerequisite for the application of machine learning to sequence analysis and effector identification. Sequence homology is often measured using similarity (SIM) rather than identity, with a cut-off of ≤ 30% for proteins. Therefore, we also employed a loose measure of homology, defined as >30% similarity, to examine sequence similarity between validated effectors. Surprisingly, the homology network involved all the 634 C100 peptides (C100_30%_SIM) (Figure 1A). The results demonstrated that the validated T4SEs showed unexpectedly significant homology, especially for the C-terminus.

Taking full advantage of the fragmental similarity between T4SEs, combined with machine learning techniques, a comprehensive prediction pipeline (T4SEpp) was designed (Figure

1B and C). Several homology searching modules have been developed to detect full-length (flBlast), effector domain (effectHMM) and C-terminal signal region (sigHMM) homologs of known T4SEs. A previous machine learning model, T4SEpre, which predicts T4SEs based on the C-terminal hand-crafted features and fine-tuned based on an updated dataset [24]. Using the generative features from pre-trained transformers, we also developed a deep learning module, T4attention, incorporated with the Bi-Conv attention mechanism. Figure 1D showes the framework of T4SEpp, taking the prediction scores of the homology search module (flBlast, effectHMM, and sigHMM), T4SEpre, and T4attention into a linear model to generate the final score, which reflects the likelihood of an input protein to be an effector.



Optimized Weights:

T4SEpp_ESM-1b: w1 = 0.42, w2 = 0.44, w3 = 0.05, w4 = 0.10, w5 = 0.10
T4SEpp_ProtBert: w1 = 0.30, w2 = 0.35, w3 = 0.08, w4 = 0.30, w5 = 0.21
T4SEpp_ProtT5-XL-UniRef50: w1 = 0.30, w2 = 0.38, w3 = 0.06, w4 = 0.30, w5 = 0.21

Probability (T4SEpp predict)
= min (w1*score1 + w2*score2 + w3*score3
+ max(w4*score4,w5*score5) , 1.0)

Figure 1. Sequence homology among T4S effectors and an integrated prediction framework. (A) Sequence homology network of T4SE. The nodes represented effectors with homology with at least one other effector. The pairs with homology (identified by the criteria defined at the top) were connected by green lines. The cluster and homology represented the number of T4SE multi-member clusters and homologous proteins. (B) Homology-based modules developed for T4SEpp, based on the full-length effector proteins (flBlast) or signal sequence (sigHMM), and effector (effectHMM) domains. (C)

T4attention, a deep learning model framework based on Bi-Conv attention. (D) Flowchart of the T4SEpp prediction program. The weighted sum of the prediction scores from each individual module is incorporated into the probability that a protein is a T4SE.

T4SE families of signal sequences and functional domains

According to the homology of the C50 peptides, the effectors could be clustered into 405 signal sequence families, including 94 multi-member and 311 singlet families (Supplementary Table S3). After the signal sequences (C50) were removed, 640 effectors with a length of ≥ 30 amino acids remained, of which 270 were classified into 106 multi-member families and 370 represented singlet families (Supplementary Table S4). The sequences within each multi-component family showed striking similarity, and multiple positions appeared conserved, as shown for one example, sigFAM50 (Fig. 2A). The amino acid composition (AAC) showed apparent preference in multiple positions, e.g., leucine in positions 9, 24, and 37, serine in position 18, 30, and 64, and asparagine in position 11, 26, and 48, of sigFAM50 (Fig. 2A). Effectors of the same signal sequence family may belong to different effector functional domain families and *vice versa*. For example, six cytotoxin-associated gene A (CagA) effectors and two *Legionella* proteins contained the signal sequences of the same family (sigFAM50, Figure 2B; Supplementary Table S3), but they also fell into three different effector functional domain families (effectFAM73 for all the CagAs, and effectFAM19 and effectFAM57 for the other two proteins; Figure 2B; Supplementary Table S4). This could be related to frequent domain reshuffling events that have been reported in *Legionella*[44].

Furthermore, we searched for homologs of known T4SEs from the representative bacterial genomes downloaded from UniProt (8761 genomes; Supplementary Table S5). In total, 258 protein-translocating T4SSs were detected from 227 bacterial strains distributed in their phyla (*Proteobacteria*, *Fusobacteria* and *Nitrospirae*), six classes (*Alphaproteobacteria*, *Betaproteobacteria*, *Epsilonproteobacteria*, *Gammaproteobacteria* *Fusobacteriia*, and *Nitrospira*), 117 genera and 227 species (Figure 2C, Supplementary Table S6). In these strains with T4SSs, 10,130 proteins were detected with full-length or local homology to the known T4SEs using the individual homology searching modules, and 1,020 were identified by all the three modules (Figure 2D, Supplementary Table S7).

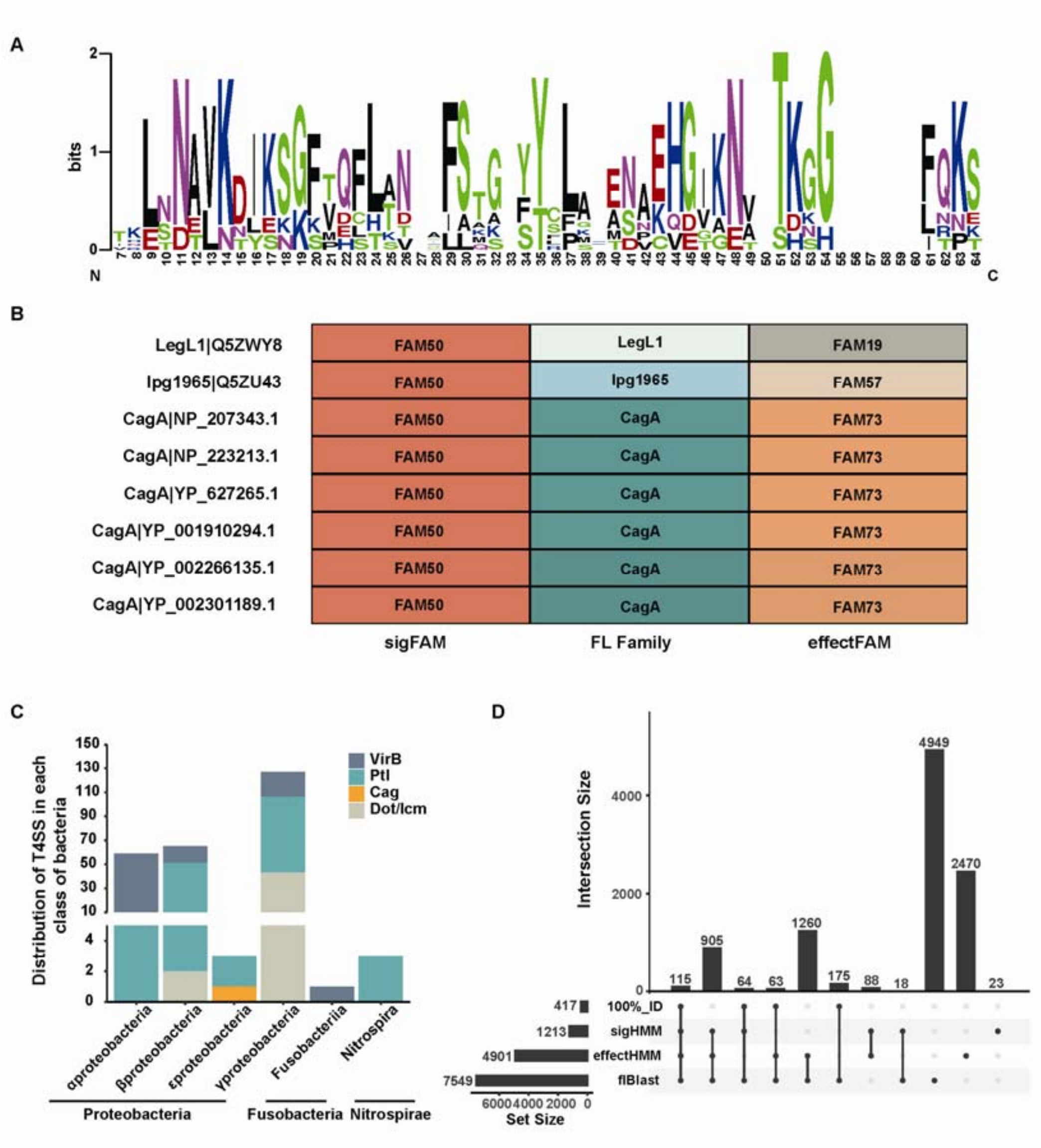


Figure 2. Search for T4SS and effectors in the UniProt reference proteome based on sequence homology. (A) Multiple-sequence alignment (MSA) of a homologous cluster (i.e., sigFAM50) of T4SE signal sequences. Then, utilize the sequence logo of position-specific Amino Acid Compositions (AAC) corresponding to the alignment. The height of the amino acid in each position indicated the AAC preference. (C) Using the core protein components of T4SS to construct a Hidden Markov Model (HMM) to predict the distribution of T4SS in the UniProt reference proteome. (D) Three homologous modules (sigHMM, effectHMM and flBlast) were used to predict the potential T4SE in the UniProt reference proteome containing T4SS, respectively. Where 100%_ID represents a known verified T4SE.

Prediction of T4SEs with pre-trained transformer-based models

Recently, protein language models have been successfully applied for structural prediction and sequence classification. In this research, we used six pre-trained models, ESM-1b, ProtAlbert, ProtBert-BFD, ProtBert-UniRef100, ProtT5-XL-BFD, and

ProtT5-XL-UniRef50, to generate features; based on this, we developed deep learning models (T4attention) based on Bi-Conv attention respectively to classify T4SEs and non-T4SEs. The T4attention models based on different sequence embedding features were compared for performance based on a five-fold cross-validation strategy (Table 1). Generally, T4attention_ESM-1b performed the best, followed by T4attention_ProtT5-XL-UniRef50, and T4attention_ProtAlbert showed the poorest performance, according to the Matthew's correlation coefficient (MCC) and F1-score (Table 1). T4attention_ESM-1b not only reached the highest MCC and F1-score (0.861 and 0.819, respectively), but required the lowest computational resources (Supplementary Figure S3). It was also noted that, for the same protein language model architecture, ProtBert or ProtT5-XL, for example, the generation of features from models pre-trained from various volumes of protein database required similar computational resource, but the smaller database-based pre-trained models always generated features for subsequent T4attention models with better performance (MCC of T4attention_ProtBert vs. T4attention_ProtBert-BFD, 0.814 vs. 0.797; T4attetion_ProtT5-XL-UniRef50 vs. ProtT5-XL-BFD, 0.818 vs. 0.800) (Table 1, Supplementary Figure S3). The redundancy of protein sequences in the BFD dataset might lead to biases in model training, and further compromise the performance of models addressing downstream tasks.

We also evaluated the performance and generalization abilities of these models on an independent testing dataset. T4attention_ProtBert showed the overall the best performance, for which the MCC, F1-score, and accuracy reached 0.917, 0.927, and 0.987, respectively (Table 2). T4attention_ESM-1b was unexpected and showed poor performance (Table 2). Consistent with the cross-validation results, the ProtBert and ProtT5-XL models, based on the features generated by transformers pre-trained from a smaller database (UniRef100/UniRef50), showed better performance (Table 2, Supplementary Figure S4).

Considering the performance of models based on both cross-validation results and the independent testing dataset, as well as the requirement of computational resources, we integrated three models, T4attention_ESM-1b, T4attention_ProtBert, and

T4attention_ProtT5-XL-UniRef50, into the pipeline to predict T4SEs.

An integrated pipeline predicting T4SEs with largely improved performance

In addition to the models based on the features generated by the transformer, we tested traditional machine learning models based on hand-crafted features. To this end, we fine-tuned two models of T4SEpre models (T4SEpre_psAac and T4SEpre_bpbAac) to learn the amino acid composition features in the C-termini of T4SEs[24]. Both models showed a certain performance in the prediction of T4SEs according to the cross-validation results or the independent testing dataset, although they were not comparable to the T4attention models (Tables 1 and 2).

To further improve the accuracy and reduce the false positive rate for T4SE prediction, we assembled a unified pipeline, T4SEpp, integrating the homology searching modules, machine learning models based on hand-crafted features and models based on transformer-generated features (Figure 1). The integrated pipeline showed strikingly better performance than the individual models, with MCC values of 0.930, 0.911 and 0.924 for T4SEpp_ESM-1b, T4SEpp_ProtBert, and T4SEpp_ProtT5-XL-UniRef50 based on the cross-validation evaluation and 0.883, 0.943, and 0.942 for the testing dataset, respectively (Tables 1 and 2).

T4SEpp was also compared to other state-of-the-art(SOTA) T4SE prediction models, such as Bastion4[26], CNNT4SE[27] and T4SEfinder[29]. Among these other models, Bastion4 showed the best performance, which was close to that of the T4attention models but was far inferior to the integrated T4SEpp (Table 2).

Genome-wide screening of T4SEs in *Helicobacter pylori* and other bacteria

H. pylori is a gram-negative, spiral-shaped bacterium that colonizes the stomach in approximately half of the world's population[45]. Although most individuals do not experience any adverse health outcomes attributable to *H. pylori*, the presence of these bacteria in the stomach increases the risk of developing gastric diseases[46-50]. *H. pylori* infection is also the strongest known risk factor for gastric cancer, the third leading cause of cancer-related death worldwide[51]. T4SS plays an important role in *H. pylori*[47-50]. However, to date, only one T4SE, CagA, has been identified for the T4SS in *H. pylori*[52]. Here, we applied T4SEpp to screen T4SE candidates from the proteins derived from the genome of *H. pylori* 26695, a model *H. pylori* strain (NCBI accession number: NC_000915.1). The three T4SEpp integrated models, T4SEpp_ESM-1b, T4SEpp_ProtBert, and T4SEpp_ProtT5-XL-UniRef50, predicted 55, 22, and 38 T4SE candidates, respectively, and 13 were shared by the prediction results of all the three models (Figure 3A-B; Supplementary Tables S6, S8). The 13 potential effector genes were scattered throughout the genome (Figure 3B). Notably, *HP_RS02695,* which encodes the only known effector CagA, was among the 13 candidates (Figure 3B).

Gene co-expression was analyzed for the 13 T4SE candidate genes in *H. pylori 26695* using an RNA-seq dataset sampled from the strain collected under 12 different conditions[53]. Except for *HP_RS06290*, *HP_RS03730*, *HP_RS04865*, and *HP_RS06295*, the remaining eight genes showed a strong expression correlation with *cagA* expression (Figure 3C). The genes co-expressed with *cagA* also showed a significant correlation with the expression of the core component genes of the Cag T4SS (Figure 3C). Furthermore, we annotated 12 human proteins that showed experimentally verified interactions with CagA by literature search, including ASPP2, c-Abl, c-Met, Crk, E-cadherin, GSK-3, PAR1, PRK2, SHP-1, SHP-2, TAK1, and ZO-1[54-65]. The interaction network between the 13 potential *H. pylori* 26695 T4SEs and the 12 human proteins was inferred (Figure 3D). Ten of the candidate T4SEs showed potential interaction with at least one of the human proteins (Figure 3D). Similar to CagA, HP_RS02225, HP_RS06295 and HP_RS03730 showed interacted with all the 12 human proteins (Figure 3D). Taken together, the proteins predicted by T4SEpp could potentially represented new T4SEs, or may be closely related to the pathogenicity of *H. pylori 26695*.

We also used T4SEpp to screen the T4SE candidates from the genomes of 227 bacterial strains bearing T4SSs. T4SEpp_ESM-1b, T4SEpp_ProtBert, and T4SEpp_ProtT5-XL-UniRef50 detected 16,972, 20,441 and 17,197 T4SE candidates respectively, with 12,622 common candidates co-predicted by all the three T4SEpp models (Supplementary Table S9, Supplementary Figure S5).

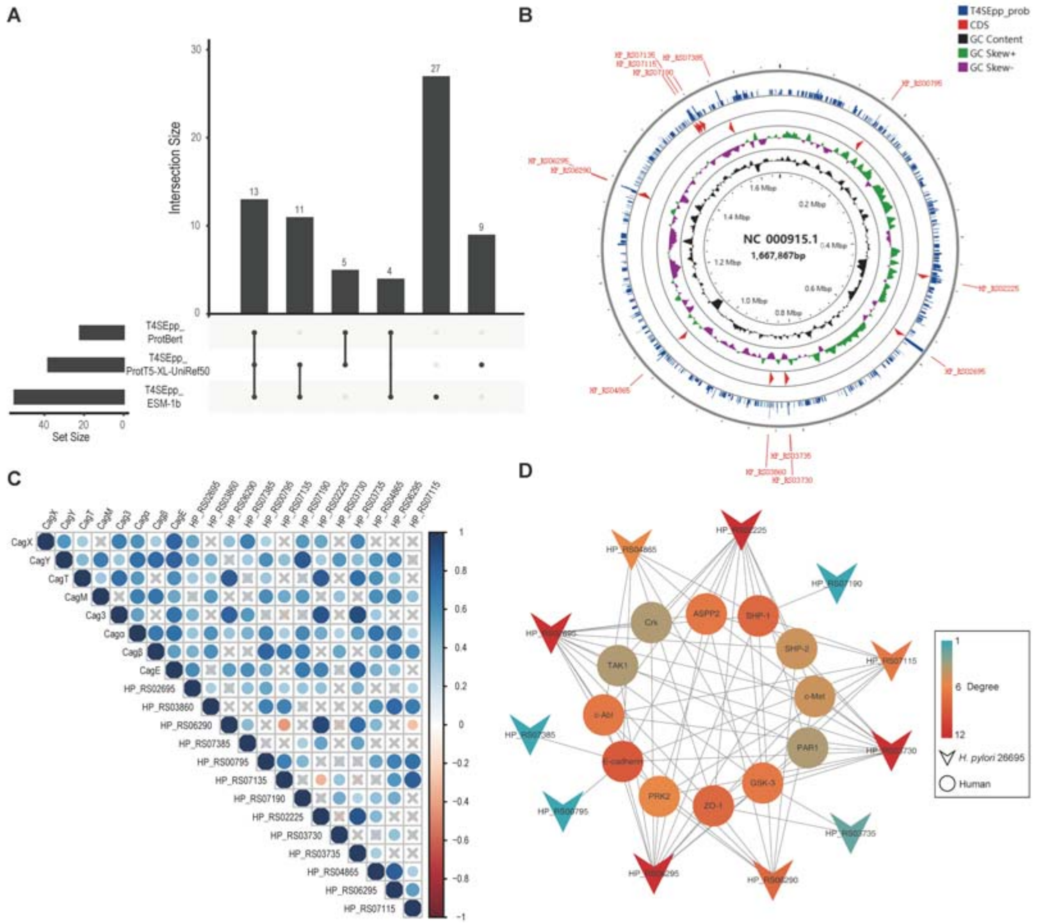


Figure 3. Whole-proteome detection for T4SEs in pathogenic bacteria (*H. pylori* 26695). (A) Prediction of potential T4SEs in the *H. pylori 26695* proteome using three T4SEpp models. (B) Use the circos diagram to show the distribution of potential T4SEs predicted by the three T4SEpp models on the *H. pylori 26695* chromosome (NC_000915.1), where T4SEpp_prob represents the mean value of the prediction results of the three T4SEpp models, and the outer circle of the circos diagram represents the three T4SEpp model predictions were all positive. (C) Under 12 different expression conditions of *H. pylori 26695*, the expression correlation of Cag T4SS core components with 12 potential T4SEs and CagA (HP_RS02695) predicted by three T4SEpp models were positive. (D) Prediction of potential interactions between 12 potential T4SEs in H. pylori 26695 and 12 human proteins using DeepHPI. These 12 human proteins are known to interact with CagA(HP_RS02695).

Web server and implementation of T4SEpp

To facilitate the implementation of T4SEpp, we developed a user-friendly web application (https://bis.zju.edu.cn/T4SEpp). The three T4SEpp integrated models, T4SEpp_ESM-1b, T4SEpp_ProtBert, and T4SEpp_ProtT5-XL-UniRef50 can be chosen and implemented by users. Both the overall prediction results and the results of the individual modules are displayed in table format, which can be downloaded and filtered easily.

Discussion

T4SS plays a crucial role in bacterial pathogenicity by secreting effectors into host cells. *L. pneumophila* can translocated more than 300 known effectors into human cells via the Dot/Icm T4SS system, causing legionellosis[66, 67]. In *H. pylori*, CagA is the only known T4SE that can hijack multiple signaling pathways in gastric epithelial cells, leading to gastritis, gastric ulcer and even gastric cancer[68, 69]. Identifying the full repertoire of T4SEs in a pathogen is important to understand its pathogenic mechanisms. Computational methods can assist with the effective identification of new effectors[70]. However, the currently available T4SE prediction tools still show high false positive rates[2]. To address this issue, we developed a unified T4SE prediction pipeline, T4SEpp, which includes homologous search modules, traditional machine learning modules and natural language processing-based modules. T4SEpp outperformed other SOTA methods for predicting T4SEs, with improved sensitivity and specificity. Furthermore, we initiated a web server that can conveniently implement the T4SEpp pipeline, providing the prediction results for each module.

Although the component modules of T4SEpp can be used for T4SE prediction, they often show higher false positive rates when used alone. This could be related to the low power of the individual dimensions of the features. Specifically, T4SE signal sequences were considered to contain important common features guiding T4SE secretion and translocation, which were used for effective T4SE prediction using tools such as T4SEpre[24]. However, the computational models based only on the signal sequences showed performance inferior to other models based on multiple-aspect features extracted from full-length proteins[26]. In this study, we discovered high sequence similarity in the C-terminal signal region among the proteins, without apparent homology to full-length effectors. Such undetected homology could have introduced bias and led to overfitting of various established machine learning algorithms and the discrepancy between the reported and actual accuracy of these methods. However, the C-terminal homology could also suggest the independent evolution of the signal sequences, and it could potentially be applied to facilitate the identification of new effectors[42].

In this study, three types of modules were integrated to predict T4SEs. Homology searching-based modules provide more accurate results, but they also show a lower capacity to detect new effectors with or without remote homology. The re-trained T4SEpre modules focused on the important features of the C-terminal signal sequences of T4SEs. T4attention learns from the full-length effector proteins the features generated by protein language models (pLMs) pre-trained with large-scale protein databases. These pLM-based models can learn new, previous unknown features that may involve position-position interactions, and have demonstrated outstanding performance in the prediction of proteins with various biological functions, such as subcellular localization and secondary structure. We used multiple pLMs to build transfer learning models, most of which exhibited excellent performance in T4SE prediction. Interestingly, we noticed that the pre-trained pLMs based on the larger datasets did not generate better prediction performance. pLMs pre-trained on smaller datasets are more efficient. Therefore, the transfer models were trained with the pLMs based on smaller non-redundant protein datasets. T4SEpp, which integrated all three types of modules, significantly outperformed both individual modules and other similar applications.

Using T4SEpp, we analyzed the potential new T4SEs in both *H. pylori* and other strains bearing T4SS. We identified 12 new T4SEs in *H. pylori*. We also identified 12,205 new T4SEs and 417 known T4SEs from 227 strains bearing a T4SS. The results suggested that there are many new effectors yet to be clarified.

Despite the significant performance improvement of T4SEpp, there remains a need to further improve the prediction of T4SEs. Other features that have been known to contribute to the recognition of T4SEs, such as the GC content of genomic loci, phylogenetic profiles, consensus regulatory motifs in promoters, physicochemical properties, secondary structures, homology to eukaryotic domains, and organelle-targeting signals, have not been integrated into the current version of the model[70]. Novel features that could be further integrated to improve the model performance remain to be disclosed. The different types (IVA and IVB) of effectors, chaperone-dependent or chaperone-independent effectors, or species-specific effectors can also be modeled and predicted separately to make more accurate prediction[70].

Materials and methods

Datasets

The 390 T4SEs used by Bastion4 as the positive training dataset[26] and 540 T4SEs annotated in SecReT4 v2.0[43] were collected and merged, and in total we got 653 non-identical, validated T4SEs. CD-HIT[71] was used to filter homology-redundant proteins with sequence identity ≥ 60%, generating 518 non-redundant T4SEs, which were used as the positive training dataset(Supplementary Figure S1A). For the negative training dataset, we collected 1112 and 1548 non-T4SE protein sequences from Bastion4[26] and PredT4SE-stack[72], respectively. The same procedure was used to eliminate the sequence redundancy among the non-T4SEs and between the non-T4SEs and T4SEs in the positive training dataset, generating 1590 non-redundant non-T4SEs (Supplementary Figure S1A). An independent validation dataset was also prepared, for which the T4SEs were collected from the testing dataset of Bastion4 (30) and others (74) annotated from literature published recently (Supplementary Table S1), and the 150 testing non-T4SEs of Bastion4 were also used as negative ones. CD-HIT was used to filter the redundant proteins with ≥60% sequence identity to the training proteins and among proteins in the validation dataset, resulting in 20 non-redundant T4SEs and 150 non-T4SEs (Supplementary Figure S1B).

Genome-wide screening of protein-translocation T4SSs

The conserved core component proteins were collected from four representative protein-translocation T4SSs, including the *Agrobacterium tumefaciens* VirB/VirD4 T4SS (inner membrane complex proteins VirB3, VirB6, VirB8, VirB10 and VirD4, and outer membrane complex proteins VirB7, VirB9 and VirB10)[16], the *Bordetella pertussis* Ptl T4SS (inner membrane complex proteins PtlB, PtlE and PtlH, and outer membrane complex proteins PtlF and PtlG)[73], the *Helicobacter pylori* Cag T4SS (inner membrane complex proteins Cagα, Cagβ and CagE, and outer membrane complex proteins CagX, CagY, CagT, CagM and Cag3)[18], the *Legionella pneumophila* Dot/Icm T4SS (inner membrane complex proteins IcmB, IcmG and DotB, and outer membrane complex proteins DotC, DotD, DotG and IcmK)[16]. Hidden Markov Model (HMM) profiles were built using HMMER 3.1 for the T4SS component protein families[74]. Protein sequences derived from the 8761 reference bacterial genomes curated in UniProt were scanned with HMMER and the HMM profiles to determine the distribution of homologs of T4SS core component proteins (Supplementary Table S5).

Homology networks of the T4SE peptide sequences

The sequences of 653 non-identical verified T4SE proteins were used to construct the homology networks. JAligner implemented the Smith-Waterman algorithm to determine the similarity between any pair of full-length effectors or peptide fragments of designated length (http://jaligner.sourceforge.net/). The identity and similarity percentages between any pair of sequences were used as measures to determine the homology level[38].

Homology-based T4SE detection modules

Diamond blastp was used to determine the homology and cluster the full-length effector proteins[75] and to screen new full-length homologs (flBlast). Two proteins showing ≥30% similarity for ≥70% of the full length of either protein were considered to be full-length homologs[38, 76]. The C-terminal 50-aa signal sequences of the verified effectors were clustered according to homology networks with 30% identity for 70% length aligned by JAligner. HMM profiles were built for each signal sequence family, and a sigHMM module was developed to screen for proteins with C-terminal sequences homologous to the profiles of known T4SE signal sequence families. The homology cutoff for HMM searching was optimized for each family, ensuring that all or most of the known effectors recalled and maintained a higher specificity. For effectHMM, we removed the C-terminal 50-aa signal from each known effector sequence, and the remaining peptide fragment with >30-aa length was used for domain clustering. Pairwise alignment was repeatedly performed with BLAST between the domain sequences, and the cutoff for homology was optimized based on the average coverage of the aligned length multiplied by the identity, that is, ≥10[38]. The HMM profiles were built for the effector domain families, and effectHMM was developed using a similar procedure as sigHMM to screen the proteins with homologous T4SE effector-domains. We used EBT to compare general homology between proteins[38, 77].

Fine-tune T4SEpre models with updated datasets

Fine-tune T4SEpre models (T4SEpre_psAac and T4SEpre_bpbAac) using the new training datasets of T4SEs and non-T4SEs. The original T4SEpre procedure was followed for feature representation, parameter optimization and model training[24]. Briefly, sequential amino acid, bi-residue and motif composition features and position-specific amino acid composition profile for the positive training dataset were represented for each C-terminal 100-aa sequence for the psAac model. For the bpbAac model, position-specific amino acid composition profiles of both the positive and the negative training datasets (Bi-Profile Bayesian features) were represented for each C-terminial 100-aa sequence. Support vector machine (SVM) models were trained for feature matrices. The kernel functions, that is, linear, polynomial, sigmoid, and radial base function (RBF), and corresponding parameters (cost and gamma) were optimized using a 5-fold cross-validation grid search strategy. The sklearn v1.0.1 was used for implementing SVM model training and kernel/parameter optimization.

The deep learning architecture of T4attention based on pre-trained protein language models

Input embeddings. Frozen embeddings were extracted directly from protein language models (pLMs) without fine-tuning the training data. Four different basic LMs were used in this study, and six different pLMs were pre-trained with different datasets. The basid LMs include, (i) “ESM-1b”[33], which is a Transformer model, (ii) "ProtBert" [32], which is a BERT-based encoder model[30]**,** generating two pLMs pre-trained on BFD[78] and UniRef100[79] data, respectively, (iii) ProtT5-XL[32], which is an encoder model based on T5[80], generating two pLMs pre-trained on BFD and UniRef50, respectively, and (iv) ProtAlbert[32], which is an encoder model based on Albert[81] and pre-trained only with UniRef100.

Optimization strategy. We use a BERT-like optimizer AdamW and a Cosine Warm-up strategy[30] to optimize the loss of the learning model. The initial learning rate is set to 0.0001, the batch size is set to 18, and the warm-up steps were set to 10. An early stopping strategy was applied to monitor the validation ACC with 30 epochs to prevent overfitting. To address the challenges of imbalanced positive and negative samples and the difficulty of training individual samples in deep learning model training, we adopted the Focal Loss method to mitigate the issue of gradient descent difficulty[82]. Focal Loss increases the hyperparameter γ (default γ=2) based on the weighted cross-entropy loss, which controls the shape of the curve.

$$\boldsymbol{FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t)}$$

$\boldsymbol{\alpha_t}$: Weight of the sample t,

$\boldsymbol{p_t}$: Binary cross entropy loss.

T4attention model. The input to T4attention (Figure 1C, Supplement Figure S2) is a protein embedding $\boldsymbol{E_0} \in \mathbb{R}^{\boldsymbol{n \times d_0}}$**,** where **n** is the sequence length and **d₀** is the size of the embedding (depending on the feature extraction model)**.** T4attention is a model based on Bi-Conv attention. In the protein embedding direction, average pooling is performed directly, and the input is transformed by two separate 1D convolutions, where the 1D convolution serves as the attention coefficient **e** and value **v** for computing the embedding dimension, $\boldsymbol{e, v} \in \mathbb{R}^{\boldsymbol{d_1}}$**.** Thus, we obtained the feature representation of the embedding dimension $\boldsymbol{x} = \boldsymbol{softmax(e)} \times \boldsymbol{v}$**.** In the direction of the protein sequence, we randomly intercept the length of **m** in the length direction of the protein-embedding sequence such that the protein embedding becomes $\boldsymbol{E_1} \in \mathbb{R}^{\boldsymbol{m \times d_0}}$**.** Similar to the convolutional attention calculation in the protein embedding direction, the attention coefficient **e'** and value **v'** are obtained, $\boldsymbol{e', v'} \in \mathbb{R}^{\boldsymbol{m \times d_1}}$**.** The difference is that the direction of the convolution is in the direction of the sequence length, so that we can obtain the feature representation of the protein sequence direction and converge according to the sequence length direction by $\boldsymbol{x'} = \sum_{\boldsymbol{i}}^{\boldsymbol{m}} \boldsymbol{softmax(e')} \times \boldsymbol{v'}$**.** The convolution attention results of the embedding direction and the protein sequence direction are merged and passed through the LayerNorm and the residual one-dimensional convolution, and the class probabilities are obtained through the two-layer multi-layer perceptron (MLP), $\boldsymbol{p}(\mathbf{c}|\boldsymbol{x}) = \boldsymbol{softmax}(\boldsymbol{MLP}(\boldsymbol{Conv}(\boldsymbol{x}+\boldsymbol{x}') + (\boldsymbol{x}+\boldsymbol{x}')))$, where **c** indicates the category of the output (i.e., T4SE or nonT4SE).

T4attention was developed using PyTorch v1.10.1. The models were trained and evaluated with 24-GB of memory and an NVIDIA GeForce RTX 3090 GPU for acceleration.

Integrated T4SE prediction model

T4SEpp is a linear model that integrates multiple prediction modules developed or re-trained in this study, including homology-searching modules for full-length or fragmented effector proteins, traditional machine-learning modules with hand-crafted features, and the attention-based transfer learning modules using the features generated by pre-trained protein language models. For any prediction module, the factor was set to 1.0 if there was a positive prediction result, and 0 otherwise. Weight **x** was assigned empirically to each module, where $\mathbf{x} \in (0,0.50)$. The maximum T4SEpp predicted value was set as 1.0. We trained the model using a grid search with 5-fold cross-validation to determine the optimal combination of weights. The early stopping strategy was similar to that used for T4attention. The final optimal parameters were shown in Figure 1D.

Assessment of model performance

Measures including accuracy (ACC), sensitivity (SN), specificity (SP), precision (PR), F1-score, Matthew's correlation coefficient (MCC), the area under the receiver operating characteristic curve (rocAUC), and the precision recall rate curve (AUPRC) were calculated to evaluate and compare the performance of models predicting T4SEs. Some of these measures are defined as follows:

$$\boldsymbol{ACC} = \frac{\boldsymbol{TP}+\boldsymbol{TN}}{\boldsymbol{TP}+\boldsymbol{TN}+\boldsymbol{FP}+\boldsymbol{FN}}$$

$$\boldsymbol{SN} = \frac{\boldsymbol{TP}}{\boldsymbol{TP}+\boldsymbol{FN}}$$

$$\boldsymbol{SP} = \frac{\boldsymbol{TN}}{\boldsymbol{TN}+\boldsymbol{FP}}$$

$$PR = \frac{TP}{TP + FP}$$

$$F1 - score = \frac{2 \times TP}{2 \times TP + FP + FN}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FN) \times (TN + FP)}}$$

where TP, TN, FP, and FN represent the number of true positives, true negatives, false positives, and false negatives, respectively.

RNA-seq analysis

RNA-seq datasets of *H. pylori* 26695 under different conditions were downloaded from the NCBI GEO DataSets database with accessions GSE165055 and GSE165056[53]. After removing the adapters and low-quality sequences with Trimmomatic v0.39[83], the cleaned reads were mapped to the *H. pylori* 26695 reference genome (NC_000915.1) using READemption (Version 2.0.0)[84]. The annotated genes were then quantified and analyzed. Protein-Protein Interaction (PPI) Networks were built and visualized using the Cytoscope v3.9.1[85].

Availability

The online version of the T4SEpp is freely accessible at https://bis.zju.edu.cn/T4SEpp. The standalone version of the T4SEpp model and the individual modules were are also deposited at https://github.com/yuemhu/T4SEpp. RNA-seq data are publicly available in the NCBI GEO DataSets database with accession numbers GSE165055 and GSE165056.

Funding

This work was supported by the National Key Research and Development Program of China (2016YFA0501704, 2018YFC0310602), the National Natural Sciences Foundation of China (31771477, 32070677), the Science and Technology Innovation Leading Scientist (2022R52035), and the 151 talent project of Zhejiang Province (first level), Collaborative Innovation Center for Modern Crop Production co-sponsored by province and ministry, and the Natural Science Fund of Shenzhen (JCYJ20190808165205582).

Authors' Contribution

MC conceived and supervised the project. YH, MC, and YW coordinated the project. YH, YZ, YH, and ZZ dataset collection. YH provided codes, models and software tools. YH, XH, and HC developed the website. YH and YW performed model comparison and RNA-seq data analyses. YH, XH, HC, SL, QN, YW, and MC wrote the first draft of this manuscript. YH, YW, and MC revised the manuscript accordingly.

Conflict of Interest: none declared.

ReferencesUncategorized References

1. Costa TR, Felisberto-Rodrigues C, Meir A, Prevost MS, Redzej A, Trokter M, et al. Secretion systems in Gram-negative bacteria: structural and mechanistic insights. Nat Rev Microbiol. 2015;13(6):343-59. doi: 10.1038/nrmicro3456. PMID: 25978706.

2. Hui X, Chen Z, Zhang J, Lu M, Cai X, Deng Y, et al. Computational prediction of secreted proteins in gram-negative bacteria. Comput Struct Biotechnol J. 2021;19:1806-28. doi: 10.1016/j.csbj.2021.03.019. PMID: 33897982.

3. Grohmann E, Christie PJ, Waksman G, Backert S. Type IV secretion in Gram-negative and Gram-positive bacteria. Mol Microbiol. 2018;107(4):455-71. doi: 10.1111/mmi.13896. PMID: 29235173.

4. Galan JE, Waksman G. Protein-Injection Machines in Bacteria. Cell. 2018;172(6):1306-18. doi: 10.1016/j.cell.2018.01.034. PMID: 29522749.

5. Waksman G. From conjugation to T4S systems in Gram-negative bacteria: a mechanistic biology perspective. EMBO Rep. 2019;20(2). doi: 10.15252/embr.201847012. PMID: 30602585.

6. Li YG, Hu B, Christie PJ. Biological and Structural Diversity of Type IV Secretion Systems. Microbiol Spectr. 2019;7(2). doi: 10.1128/microbiolspec.PSIB-0012-2018. PMID: 30953428.

7. Gonzalez-Rivera C, Bhatty M, Christie PJ. Mechanism and Function of Type IV Secretion During Infection of the Human Host. Microbiol Spectr. 2016;4(3). doi: 10.1128/microbiolspec.VMBF-0024-2015. PMID: 27337453.

8. Christie PJ. The Mosaic Type IV Secretion Systems. EcoSal Plus. 2016;7(1). doi: 10.1128/ecosalplus.ESP-0020-2015. PMID: 27735785.

9. Christie PJ, Gomez Valero L, Buchrieser C. Biological Diversity and Evolution of Type IV Secretion Systems. Curr Top Microbiol Immunol. 2017;413:1-30. doi: 10.1007/978-3-319-75241-9_1. PMID: 29536353.

10. Chandran Darbari V, Waksman G. Structural Biology of Bacterial Type IV Secretion Systems. Annu Rev Biochem. 2015;84:603-29. doi: 10.1146/annurev-biochem-062911-102821. PMID: 26034891.

11. Sheedlo MJ, Ohi MD, Lacy DB, Cover TL. Molecular architecture of bacterial type IV secretion systems. PLoS Pathog. 2022;18(8):e1010720. doi: 10.1371/journal.ppat.1010720. PMID: 35951533.

12. Ansari S, Yamaoka Y. Helicobacter pylori Virulence Factor Cytotoxin-Associated Gene A (CagA)-Mediated Gastric Pathogenicity. Int J Mol Sci. 2020;21(19). doi: 10.3390/ijms21197430. PMID: 33050101.

13. Hubber A, Roy CR. Modulation of host cell function by Legionella pneumophila type IV effectors. Annu Rev Cell Dev Biol. 2010;26:261-83. doi: 10.1146/annurev-cellbio-100109-104034. PMID: 20929312.

14. Wozniak RA, Waldor MK. Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. Nat Rev Microbiol. 2010;8(8):552-63. doi: 10.1038/nrmicro2382. PMID: 20601965.

15. Wallden K, Rivera-Calzada A, Waksman G. Type IV secretion systems: versatility and diversity in function. Cell Microbiol. 2010;12(9):1203-12. doi: 10.1111/j.1462-5822.2010.01499.x. PMID: 20642798.

16. Costa TRD, Harb L, Khara P, Zeng L, Hu B, Christie PJ. Type IV secretion systems: Advances in structure, function, and activation. Mol Microbiol. 2021;115(3):436-52. doi: 10.1111/mmi.14670. PMID: 33326642.

17. Burns DL. Type IV transporters of pathogenic bacteria. Curr Opin Microbiol. 2003;6(1):29-34. doi: 10.1016/s1369-5274(02)00006-1. PMID: 12615216.

18. Cover TL, Lacy DB, Ohi MD. The Helicobacter pylori Cag Type IV Secretion System. Trends Microbiol. 2020;28(8):682-95. doi: 10.1016/j.tim.2020.02.004. PMID: 32451226.

19. Ward DV, Zambryski PC. The six functions of Agrobacterium VirE2. Proc Natl Acad Sci U S A. 2001;98(2):385-6. doi: 10.1073/pnas.98.2.385. PMID: 11209039.

20. Schrammeijer B, den Dulk-Ras A, Vergunst AC, Jurado Jacome E, Hooykaas PJ. Analysis of Vir protein translocation from Agrobacterium tumefaciens using Saccharomyces cerevisiae as a model: evidence for transport of a novel effector protein VirE3. Nucleic Acids Res. 2003;31(3):860-8. doi: 10.1093/nar/gkg179. PMID: 12560481.

21. Hofreuter D, Odenbreit S, Puls J, Schwan D, Haas R. Genetic competence in Helicobacter pylori: mechanisms and biological implications. Res Microbiol. 2000;151(6):487-91. doi: 10.1016/s0923-2508(00)00164-9. PMID: 10961464.

22. Lee YW, Wang J, Newton HJ, Lithgow T. Mapping bacterial effector arsenals: in vivo and in silico approaches to defining the protein features dictating effector secretion by bacteria. Curr Opin Microbiol. 2020;57:13-21. doi: 10.1016/j.mib.2020.04.002. PMID: 32505919.

23. Burstein D, Zusman T, Degtyar E, Viner R, Segal G, Pupko T. Genome-scale identification of Legionella pneumophila effectors using a machine learning approach. PLoS Pathog. 2009;5(7):e1000508. doi: 10.1371/journal.ppat.1000508. PMID: 19593377.

24. Wang Y, Wei X, Bao H, Liu SL. Prediction of bacterial type IV secreted effectors by C-terminal features. BMC Genomics. 2014;15:50. doi: 10.1186/1471-2164-15-50. PMID: 24447430.

25. Zou L, Nan C, Hu F. Accurate prediction of bacterial type IV secreted effectors using amino acid composition and PSSM profiles. Bioinformatics. 2013;29(24):3135-42. doi: 10.1093/bioinformatics/btt554. PMID: 24064423.

26. Wang J, Yang B, An Y, Marquez-Lago T, Leier A, Wilksch J, et al. Systematic analysis and prediction of type IV secreted effector proteins by machine learning approaches. Brief Bioinform. 2019;20(3):931-51. doi: 10.1093/bib/bbx164. PMID: 29186295.

27. Hong J, Luo Y, Mou M, Fu J, Zhang Y, Xue W, et al. Convolutional neural network-based annotation of bacterial type IV secretion system effectors with enhanced accuracy and reduced false discovery. Brief Bioinform. 2020;21(5):1825-36. doi: 10.1093/bib/bbz120. PMID: 31860715.

28. Rao R, Bhattacharya N, Thomas N, Duan Y, Chen X, Canny J, et al. Evaluating Protein Transfer Learning with TAPE. Adv Neural Inf Process Syst. 2019;32:9689-701. PMID: 33390682.

29. Zhang Y, Zhang Y, Xiong Y, Wang H, Deng Z, Song J, et al. T4SEfinder: a bioinformatics tool for genome-scale prediction of bacterial type IV secreted effectors using pre-trained protein language model. Brief Bioinform. 2022;23(1). doi: 10.1093/bib/bbab420. PMID: 34657153.

30. Devlin J, Chang M-W, Lee K, Toutanova K, editors. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding2019 June; Minneapolis, Minnesota: Association for Computational Linguistics.

31. Stärk H, Dallago C, Heinzinger M, Rost B. Light attention predicts protein location from the language of life. Bioinform Adv. 2021;1(1):vbab035. doi: 10.1093/bioadv/vbab035. PMID: 36700108.

32. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. IEEE Trans Pattern Anal Mach Intell. 2022;44(10):7112-27. doi: 10.1109/TPAMI.2021.3095381. PMID: 34232869.

33. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci U S A. 2021;118(15). doi: 10.1073/pnas.2016239118. PMID: 33876751.

34. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234-40. doi: 10.1093/bioinformatics/btz682. PMID: 31501885.

35. Heinzinger M, Elnaggar A, Wang Y, Dallago C, Nechaev D, Matthes F, et al. Modeling aspects of the language of life through transfer-learning protein sequences. BMC Bioinformatics. 2019;20(1):723. doi: 10.1186/s12859-019-3220-8. PMID: 31847804.

36. Wagner N, Alburquerque M, Ecker N, Dotan E, Zerah B, Pena MM, et al. Natural language processing approach to model the secretion signal of type III effectors. Front Plant Sci. 2022;13:1024405. doi: 10.3389/fpls.2022.1024405. PMID: 36388586.

37. Teufel F, Almagro Armenteros JJ, Johansen AR, Gislason MH, Pihl SI, Tsirigos KD, et al. SignalP 6.0 predicts all five types of signal peptides using protein language models. Nat Biotechnol. 2022;40(7):1023-5. doi: 10.1038/s41587-021-01156-3. PMID: 34980915.

38. Hui X, Chen Z, Lin M, Zhang J, Hu Y, Zeng Y, et al. T3SEpp: an Integrated Prediction Pipeline for Bacterial Type III Secreted Effectors. mSystems. 2020;5(4). doi: 10.1128/mSystems.00288-20. PMID: 32753503.

39. Dong X, Lu X, Zhang Z. BEAN 2.0: an integrated web resource for the identification and functional analysis of type III secreted effectors. Database (Oxford). 2015;2015:bav064. doi: 10.1093/database/bav064. PMID: 26120140.

40. Goldberg T, Rost B, Bromberg Y. Computational prediction shines light on type III secretion origins. Sci Rep. 2016;6:34516. doi: 10.1038/srep34516. PMID: 27713481.

41. Wagner N, Avram O, Gold-Binshtok D, Zerah B, Teper D, Pupko T. Effectidor: an automated machine-learning-based web server for the prediction of type-III secretion system effectors. Bioinformatics. 2022;38(8):2341-3. doi: 10.1093/bioinformatics/btac087. PMID: 35157036.

42. Meyer DF, Noroy C, Moumene A, Raffaele S, Albina E, Vachiery N. Searching algorithm for type IV secretion system effectors 1.0: a tool for predicting type IV effectors and exploring their genomic context. Nucleic Acids Res. 2013;41(20):9218-29. doi: 10.1093/nar/gkt718. PMID: 23945940.

43. Bi D, Liu L, Tai C, Deng Z, Rajakumar K, Ou HY. SecReT4: a web-based bacterial type IV secretion system resource. Nucleic Acids Res. 2013;41(Database issue):D660-5. doi: 10.1093/nar/gks1248. PMID: 23193298.

44. Burstein D, Amaro F, Zusman T, Lifshitz Z, Cohen O, Gilbert JA, et al. Genomic analysis of 38 Legionella species identifies large and diverse effector repertoires. Nat Genet. 2016;48(2):167-75. doi: 10.1038/ng.3481. PMID: 26752266.

45. Hooi JKY, Lai WY, Ng WK, Suen MMY, Underwood FE, Tanyingoh D, et al. Global Prevalence of Helicobacter pylori Infection: Systematic Review and Meta-Analysis. Gastroenterology. 2017;153(2):420-9. doi: 10.1053/j.gastro.2017.04.022. PMID: 28456631.

46. Cover TL, Blaser MJ. Helicobacter pylori in health and disease. Gastroenterology. 2009;136(6):1863-73. doi: 10.1053/j.gastro.2009.01.073. PMID: 19457415.

47. Blaser MJ, Perez-Perez GI, Kleanthous H, Cover TL, Peek RM, Chyou PH, et al. Infection with Helicobacter pylori strains possessing cagA is associated with an increased risk of developing adenocarcinoma of the stomach. Cancer Res. 1995;55(10):2111-5. PMID: 7743510.

48. Figueiredo C, Machado JC, Pharoah P, Seruca R, Sousa S, Carvalho R, et al. Helicobacter pylori and interleukin 1 genotyping: an opportunity to identify high-risk individuals for gastric carcinoma. J Natl Cancer Inst. 2002;94(22):1680-7. doi: 10.1093/jnci/94.22.1680. PMID: 12441323.

49. Plummer M, van Doorn LJ, Franceschi S, Kleter B, Canzian F, Vivas J, et al. Helicobacter pylori

cytotoxin-associated genotype and gastric precancerous lesions. J Natl Cancer Inst. 2007;99(17):1328-34. doi: 10.1093/jnci/djm120. PMID: 17728213.

50. Cover TL. Helicobacter pylori Diversity and Gastric Cancer Risk. mBio. 2016;7(1):e01869-15. doi: 10.1128/mBio.01869-15. PMID: 26814181.

51. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394-424. doi: 10.3322/caac.21492. PMID: 30207593.

52. Knorr J, Ricci V, Hatakeyama M, Backert S. Classification of Helicobacter pylori Virulence Factors: Is CagA a Toxin or Not? Trends Microbiol. 2019;27(9):731-8. doi: 10.1016/j.tim.2019.04.010. PMID: 31130493.

53. Loh JT, Shum MV, Jossart SDR, Campbell AM, Sawhney N, McDonald WH, et al. Delineation of the pH-Responsive Regulon Controlled by the Helicobacter pylori ArsRS Two-Component System. Infect Immun. 2021;89(4). doi: 10.1128/IAI.00597-20. PMID: 33526561.

54. Nesic D, Buti L, Lu X, Stebbins CE. Structure of the Helicobacter pylori CagA oncoprotein bound to the human tumor suppressor ASPP2. Proc Natl Acad Sci U S A. 2014;111(4):1562-7. doi: 10.1073/pnas.1320631111. PMID: 24474782.

55. Poppe M, Feller SM, Romer G, Wessler S. Phosphorylation of Helicobacter pylori CagA by c-Abl leads to cell motility. Oncogene. 2007;26(24):3462-72. doi: 10.1038/sj.onc.1210139. PMID: 17160020.

56. Churin Y, Al-Ghoul L, Kepp O, Meyer TF, Birchmeier W, Naumann M. Helicobacter pylori CagA protein targets the c-Met receptor and enhances the motogenic response. J Cell Biol. 2003;161(2):249-55. doi: 10.1083/jcb.200208039. PMID: 12719469.

57. Suzuki M, Mimuro H, Suzuki T, Park M, Yamamoto T, Sasakawa C. Interaction of CagA with Crk plays an important role in Helicobacter pylori-induced loss of gastric epithelial cell adhesion. J Exp Med. 2005;202(9):1235-47. doi: 10.1084/jem.20051027. PMID: 16275761.

58. Murata-Kamiya N, Kurashima Y, Teishikata Y, Yamahashi Y, Saito Y, Higashi H, et al. Helicobacter pylori CagA interacts with E-cadherin and deregulates the beta-catenin signal that promotes intestinal transdifferentiation in gastric epithelial cells. Oncogene. 2007;26(32):4617-26. doi: 10.1038/sj.onc.1210251. PMID: 17237808.

59. Lee DG, Kim HS, Lee YS, Kim S, Cha SY, Ota I, et al. Helicobacter pylori CagA promotes Snail-mediated epithelial-mesenchymal transition by reducing GSK-3 activity. Nat Commun. 2014;5:4423. doi: 10.1038/ncomms5423. PMID: 25055241.

60. Saadat I, Higashi H, Obuse C, Umeda M, Murata-Kamiya N, Saito Y, et al. Helicobacter pylori CagA targets PAR1/MARK kinase to disrupt epithelial cell polarity. Nature. 2007;447(7142):330-3. doi: 10.1038/nature05765. PMID: 17507984.

61. Mishra JP, Cohen D, Zamperone A, Nesic D, Muesch A, Stein M. CagA of Helicobacter pylori interacts with and inhibits the serine-threonine kinase PRK2. Cell Microbiol. 2015;17(11):1670-82. doi: 10.1111/cmi.12464. PMID: 26041307.

62. Saju P, Murata-Kamiya N, Hayashi T, Senda Y, Nagase L, Noda S, et al. Host SHP1 phosphatase antagonizes Helicobacter pylori CagA and can be downregulated by Epstein-Barr virus. Nat Microbiol. 2016;1:16026. doi: 10.1038/nmicrobiol.2016.26. PMID: 27572445.

63. Higashi H, Tsutsumi R, Muto S, Sugiyama T, Azuma T, Asaka M, et al. SHP-2 tyrosine phosphatase as an intracellular target of Helicobacter pylori CagA protein. Science. 2002;295(5555):683-6. doi: 10.1126/science.1067147. PMID: 11743164.

64. Lamb A, Yang XD, Tsang YH, Li JD, Higashi H, Hatakeyama M, et al. Helicobacter pylori CagA

activates NF-kappaB by targeting TAK1 for TRAF6-mediated Lys 63 ubiquitination. EMBO Rep. 2009;10(11):1242-9. doi: 10.1038/embor.2009.210. PMID: 19820695.

65. Amieva MR, Vogelmann R, Covacci A, Tompkins LS, Nelson WJ, Falkow S. Disruption of the epithelial apical-junctional complex by Helicobacter pylori CagA. Science. 2003;300(5624):1430-4. doi: 10.1126/science.1081919. PMID: 12775840.

66. Goncalves IG, Simoes LC, Simoes M. Legionella pneumophila. Trends Microbiol. 2021;29(9):860-1. doi: 10.1016/j.tim.2021.04.005. PMID: 33994277.

67. Mondino S, Schmidt S, Rolando M, Escoll P, Gomez-Valero L, Buchrieser C. Legionnaires' Disease: State of the Art Knowledge of Pathogenesis Mechanisms of Legionella. Annu Rev Pathol. 2020;15:439-66. doi: 10.1146/annurev-pathmechdis-012419-032742. PMID: 31657966.

68. Hatakeyama M. Oncogenic mechanisms of the Helicobacter pylori CagA protein. Nat Rev Cancer. 2004;4(9):688-94. doi: 10.1038/nrc1433. PMID: 15343275.

69. Hatakeyama M. Helicobacter pylori CagA and gastric cancer: a paradigm for hit-and-run carcinogenesis. Cell Host Microbe. 2014;15(3):306-16. doi: 10.1016/j.chom.2014.02.008. PMID: 24629337.

70. Zhao Z, Hu Y, Hu Y, White AP, Wang Y. Features and algorithms: facilitating investigation of secreted effectors in Gram-negative bacteria. Trends Microbiol. 2023. doi: 10.1016/j.tim.2023.05.011. PMID: 37349207.

71. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22(13):1658-9. doi: 10.1093/bioinformatics/btl158. PMID: 16731699.

72. Xiong Y, Wang Q, Yang J, Zhu X, Wei DQ. PredT4SE-Stack: Prediction of Bacterial Type IV Secreted Effectors From Protein Sequences Using a Stacked Ensemble Method. Front Microbiol. 2018;9:2571. doi: 10.3389/fmicb.2018.02571. PMID: 30416498.

73. O'Callaghan D, Cazevieille C, Allardet-Servent A, Boschiroli ML, Bourg G, Foulongne V, et al. A homologue of the Agrobacterium tumefaciens VirB and Bordetella pertussis Ptl type IV secretion systems is essential for intracellular survival of Brucella suis. Mol Microbiol. 1999;33(6):1210-20. doi: 10.1046/j.1365-2958.1999.01569.x. PMID: 10510235.

74. Eddy SR. Profile hidden Markov models. Bioinformatics. 1998;14(9):755-63. doi: 10.1093/bioinformatics/14.9.755. PMID: 9918945.

75. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12(1):59-60. doi: 10.1038/nmeth.3176. PMID: 25402007.

76. Hu Y, Huang H, Cheng X, Shu X, White AP, Stavrinides J, et al. A global survey of bacterial type III secretion systems and their effectors. Environ Microbiol. 2017;19(10):3879-95. doi: 10.1111/1462-2920.13755. PMID: 28401683.

77. Hui X, Hu Y, Sun MA, Shu X, Han R, Ge Q, et al. EBT: a statistic test identifying moderate size of significant features with balanced power and precision for genome-wide rate comparisons. Bioinformatics. 2017;33(17):2631-41. doi: 10.1093/bioinformatics/btx294. PMID: 28472273.

78. Steinegger M, Soding J. Clustering huge protein sequence sets in linear time. Nat Commun. 2018;9(1):2542. doi: 10.1038/s41467-018-04964-5. PMID: 29959318.

79. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, UniProt C. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics. 2015;31(6):926-32. doi: 10.1093/bioinformatics/btu739. PMID: 25398609.

80. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the Limits of Transfer

Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research. 2020;21:1-67. doi: 10.48550/arXiv.1910.10683.

81. Lan Z, Chen M, Goodman. S, Gimpel. K, Sharma. P, Soricut. R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. International Conference on Learning Representations; 20 Dec2019.

82. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal Loss for Dense Object Detection. IEEE Trans Pattern Anal Mach Intell. 2020;42(2):318-27. doi: 10.1109/TPAMI.2018.2858826. PMID: 30040631.

83. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20. doi: 10.1093/bioinformatics/btu170. PMID: 24695404.

84. Forstner KU, Vogel J, Sharma CM. READemption-a tool for the computational analysis of deep-sequencing-based transcriptome data. Bioinformatics. 2014;30(23):3421-3. doi: 10.1093/bioinformatics/btu533. PMID: 25123900.

85. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498-504. doi: 10.1101/gr.1239303. PMID: 14597658.

Tables

Table 1. Performance comparison of the models in T4SEpp on 5-fold cross-validation dataset.

Method	ACC	SN	SP	PR	F1	MCC	rocAUC	AUPRC
T4attention_ESM-1b	**0.934±0.010**	0.844±0.017	**0.963±0.008**	**0.881±0.026**	**0.861±0.021**	**0.819±0.028**	0.950±0.008	**0.897±0.026**
T4attention_ProtBert	0.931±0.013	**0.859±0.030**	0.954±0.013	0.861±0.036	0.859±0.025	0.814±0.033	**0.954±0.010**	**0.897±0.028**
T4attention_ProtBert-BFD	0.924±0.007	0.846±0.015	0.950±0.009	0.848±0.023	0.847±0.012	0.797±0.017	0.939±0.006	0.848±0.047
T4attention_ProtT5-XL-UniRef50	0.933±0.015	0.844±0.021	0.962±0.016	**0.881±0.044**	**0.861±0.028**	0.818±0.038	0.949±0.007	0.895±0.030
T4attention_ProtT5-XL-BFD	0.925±0.021	0.847±0.017	0.950±0.025	0.851±0.065	0.849±0.037	0.800±0.051	0.949±0.011	0.887±0.032
T4attention_ProtAlbert	0.921±0.014	0.851±0.009	0.944±0.015	0.834±0.037	0.842±0.024	0.790±0.033	0.940±0.015	0.860±0.036
T4SEpre_psAac[a]	0.841±0.014	0.825±0.030	0.858±0.049	0.856±0.040	0.839±0.012	0.686±0.030	0.917±0.016	0.884±0.015
T4SEpre_bpbAac[a]	0.856±0.032	0.817±0.059	0.894±0.038	0.887±0.037	0.849±0.036	0.716±0.061	0.918±0.018	0.898±0.023
T4SEpp_ESM-1b	0.974±0.004	0.919±0.009	0.993±0.005	0.976±0.015	0.946±0.008	0.930±0.011	0.995±0.004	0.949±0.069
T4SEpp_ProtBert	0.967±0.006	0.909±0.005	0.986±0.007	0.956±0.022	0.932±0.011	0.911±0.016	0.994±0.003	0.964±0.038
T4SEpp_ProtT5-XL-UniRef50	0.972±0.006	0.917±0.009	0.990±0.006	0.968±0.019	0.942±0.012	0.924±0.015	0.994±0.003	0.957±0.049

ACC, Accuracy; SN, sensitivity; SP, specificity; PR, precision; F1, F1-score; MCC, Matthews correlation coefficient; rocAUC, area under the receiver operating characteristic curve; AUPRC, precision recall rate curve; [a], fine-tune the model.

Table 2. Performance comparison of the models in T4SEpp and other tools on the independent dataset.

Method	ACC	SN	SP	PR	F1	MCC	rocAUC	AUPRC
T4attention_ESM-1b	0.935	0.850	0.947	0.680	0.756	0.743	0.956	0.850
T4attention_ProtBert	0.982	**0.950**	0.987	0.905	0.927	0.917	**0.989**	0.936
T4attention_ProtBert-BFD	0.959	**0.950**	0.960	0.760	0.844	0.828	0.973	0.936
T4attention_ProtT5-XLUniRef50	0.959	0.900	0.967	0.783	0.837	0.816	0.973	0.880
T4attention_ProtT5-XL-BFD	0.929	**0.950**	0.927	0.633	0.760	0.741	0.973	0.930
T4attention_ProtAlbert	0.953	0.900	0.960	0.750	0.818	0.796	0.959	0.891
T4SEpp_ESM-1b	0.976	0.850	0.993	0.944	0.894	0.883	0.922	0.868

T4SEpp_ProtBert	**0.988**	**0.950**	0.993	0.950	**0.950**	**0.943**	0.974	**0.946**
T4SEpp_ProtT5-XL-UniRef50	**0.988**	0.900	**1.000**	**1.000**	0.947	0.942	0.948	0.901
T4SEfinder-TAPEBert_MLP	0.958	0.850	0.973	0.810	0.829	0.806	0.959	0.805
T4SEfinder-hybridbilstm	0.941	0.800	0.960	0.727	0.762	0.730	0.945	0.852
T4SEfinder-pssm_cnn	0.906	0.800	0.920	0.571	0.667	0.625	0.923	0.759
Bastion4	0.965	0.900	0.973	0.818	0.857	0.838	0.907	0.706
CNNT4SE	0.953	0.700	0.987	0.875	0.778	0.758	0.943	0.860
T4SEpre_psAac[a]	0.888	0.700	0.913	0.519	0.596	0.541	0.921	0.740
T4SEpre_bpbAac[a]	0.829	0.700	0.847	0.378	0.491	0.427	0.895	0.730

ACC, Accuracy; SN, sensitivity; SP, specificity; PR, precision; F1, F1-score; MCC, Matthews correlation coefficient; rocAUC, area under the receiver operating characteristic curve; AUPRC, precision recall rate curve; [a], fine-tune the model.

Supplementary data

Supplementary Figure S1. The workflow to construct the training(A) or independent testing(B) dataset in this study.

Supplementary Figure S2. Two modules used by the T4attention model.

Supplementary Figure S3. The relationship between the feature extraction time of 6 different protein natural language models and the prediction performance of T4attention model F1-score (A) and MCC (B) in the 5-fold cross-validation dataset.

Supplementary Figure S4. The relationship between T4attention model prediction performance F1-score (A) and MCC (B) in the independent test set and the overall time-consuming use of 6 different protein natural language models to extract features and their T4attention model predictions.

Supplementary Figure S5. Three T4SEpp model were used to predict the potential T4SE in the UniProt reference proteome containing T4SS, respectively. Where 100%_ID represents a known verified T4SE.

Supplementary Table S1. The 74 T4SEs independently collected from the literature.

Supplementary Table S2. Hyperparameters used in deep learning models of T4attention.

Supplementary Table S3. Homologous Clusters of T4S Effector Signal Sequences.

Supplementary Table S4. The distribution of effector domain families.

Supplementary Table S5. Distribution of the Uniprot Bacteria Reference Proteomes (Download date October 19, 2022).

Supplementary Table S6. Distribution of T4SS in the UniPort bacterial reference proteome.

Supplementary Table S7. Homology prediction results of T4SE in strains containing T4SS in the Uniport Bacteria Reference Proteomes.

Supplementary Table S8. Distribution of potential T4SEs in the *H. pylori_26695* (NC_000915.1).

Supplementary Table S9. Distribution of potential T4SEs in the Uniport Bacteria Reference Proteomes.