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 to deliver effectors (T4SEs) into the cytoplasm of eukaryotic cells and causes diseases. Identification of effectors is a crucial step in understanding the mechanisms of bacterial pathogenicity, but it remains a big 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 pipeline 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 containg three best-performed protein language pre-trained models. T4SEpp outperforms the other state-of-the-art software tools, achieving ~0.95 sensitivity at a high specificity of ~0.99 based on the assessment of an independent testing dataset. T4SEpp predicted 13 potential T4SEs, including the *H. pylori* cytotoxin-associated gene A (CagA). Among these, 10 T4SEs have the potential to interact with at least one human protein. This suggests that these potential T4SEs may be associated with the pathogenicity of *H. pylori*. Overall, T4SEpp provides a better solution to assist identification of bacterial T4SEs, and facilitates studies on 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(Costa, et al., 2015; Hui, et al., 2021). Among them, 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 by only one step(Christie, 2016; Christie, et al., 2017; Galan and Waksman, 2018; Gonzalez-Rivera, et al., 2016; Grohmann, et al., 2018; Li, et al., 2019; Waksman, 2019).

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(Christie, et al., 2017). The substrate proteins translocated by T4SSs, also called effectors, play important roles in bacterial infection and pathogenicity(Chandran Darbari and Waksman, 2015; Costa, et al., 2015; Sheedlo, et al., 2022).

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*(Ansari and Yamaoka, 2020; Burns, 2003; Costa, et al., 2021; Hubber and Roy, 2010; Wallden, et al., 2010; Wozniak and Waldor, 2010). T4SS-mediated entry of effector proteins into recipient cells is contact-dependent(Cover, et al., 2020). Once entering eukaryotic host cytoplasm, they can disrupt the signal transduction and cause various host diseases. To identify the effectors is crucial for understanding the mechanisms of infection and pathogenicity caused by these bacteria. However, since the composition and sequences vary a lot, it has been a big challenge to experimentally identify new T4SEs. Although many T4SEs have been identified and characterized in a few model organisms(Hofreuter, et al., 2000; Lee, et al., 2020; Schrammeijer, et al., 2003; Ward and Zambryski, 2001), there remain much more unclarified yet.

Since 2009 when the first machine learning algorithm was introduced, tens of computational models have been developed to predict T4SEs(Burstein, et al., 2009; Hui, et al., 2021). Early algorithms were mainly species-specific, such as those predicting T4SEs in *Legionella pneumophila*(Burstein, et al., 2009). In another study, Wang *et al.* developed an SVM-based model T4SEpre, which exhibited good overall and cross-species performance(Wang, et al., 2014). However, T4SEpre only considered the features buried in the C-terminal 100 amino acids(Wang, et al., 2014). More studies, especially the ensemble models recently developed with multi-aspect features, learn the features from full-length proteins to improve performance(Wang, et al., 2019; Zou, et al.,

2013). Deep learning algorithms have also been applied in 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(Hong, et al., 2020). T4SEfinder is a multi-layer perception (MLP) model learning the features generated by a pre-trained BERT model(Rao, et al., 2019), which could predict T4SEs accurately(Zhang, et al., 2022). Notably, BERT is one of the natural language processing (NLP) models, which are appealing in biology and other fields(Devlin, et al., 2019; Elnaggar, et al., 2022; Heinzinger, et al., 2019; Lee, et al., 2020; Rives, et al., 2021; Stärk, et al., 2021). NLP models have been successfully applied in prediction of protein subcellular localization(Elnaggar, et al., 2022; Stärk, et al., 2021), secondary structure(Elnaggar, et al., 2022; Heinzinger, et al., 2019; Rives, et al., 2021), and others(Lee, et al., 2020). Besides T4SEfinder, the NLP-based pre-trained transformers have also been used for prediction of bacterial type III secreted effectors and Sec/Tat substrates, both achieving superior performance(Teufel, et al., 2022; Wagner, et al., 2022).

Although machine learning strategies achieved some success in identification of T4SEs(Burstein, et al., 2009; Hui, et al., 2021; Wang, et al., 2014), the high false positive rate has been a big challenge. To reduce 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(Hui, et al., 2020). To improve model performance, other models also considered the multiple features and combination of homology-based strategies in prediction of type III effectors(Dong, et al., 2015; Goldberg, et al., 2016; Wagner, et al., 2022). For T4SE prediction, homology searching has 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, secretion signals, etc., and develops a scoring scheme to predict T4SEs mainly from α- and γ-proteobacteria(Meyer, et al., 2013). 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 the ensemblers, and the 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 to predict T4SEs. First, a homology searching strategy scaned both global homology of full-length proteins and local homology of domains to known effectors. Additionally, we retrained a machine learning module T4SEpre(Wang, et al., 2014) with updated T4SE data and hand-crafted amino acid composition features in C-termini. Furthermore, a group of transfer learning models were developed based on the features generated by various pre-trained transformers. For the transfer learning models, we utilized deep context protein language models ESM-1b, ProtBert, ProtT5-XL and ProtAlbert to represent protein sequence features(Elnaggar, et al., 2022; Rives, et al., 2021). These features can well characterize the intrinsic but unclarified properties of protein sequences and the interactions between positions. Based on these feature representations, subsequent 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 hand-crafted features and transfer learning models with transformer-generated features, into a pipeline, namely T4SEpp, which assembled the individual modules in a linear function to generate the 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(Bi, et al., 2013; Wang, et al.,

129 2019) (see Materials and Methods). Pairwise sequence alignments of full-length (FL)
130 effector proteins or their C-terminal peptides of 100 or 50 amino acids (C100 or C50) were
131 performed. For the FL proteins, 481 non-homologous clusters were identified after
132 homology filtering for the proteins with > 30% identity for > 70% length coverage of the
133 pair of proteins (FL_70%_30%_ID) (Figure 1A). However, for C100 sequences, 249 were
134 detected of homology to others with identity of larger than 30%, and 473 non-redundant
135 clusters were retained from these sequences after homology filtering (C100_30%_ID)
136 (Figure 1A). The reduction in the number of clusters indicated that the C-terminal 100
137 amino acids showed more homology than the full-length effector proteins, but there are no
138 significant differences between them (473/654 vs 481/654, EBT *P*= 0.614). The C50
139 sequences further reflected the more typical C-terminal homology between effectors. A
140 total of 342 peptides were found to have homology with others, while 401 clusters
141 remained for these peptides after homology filtering (C50_30%_ID, 401/654 vs. 481/654,
142 EBT *P*=3.17e-03) (Figure 1A). Rigorous homology filtering is a prerequisite for the
143 application of machine learning in sequence analysis and effector identification. Sequence
144 homology is often measured using similarity (SIM) rather than identity, with a cutoff ≤ 30%
145 for proteins. Therefore, we also employed a loose measure of homology to examine
146 sequence similarity between validated effectors, defined as >30% similarity. Surprisingly,
147 the homology network involved all the 634 C100 peptides (C100_30%_SIM) (Figure 1A).
148 The results demonstrated that the validated T4SEs showed unexpectedly significant
149 homology, especially for the C-termini.
150 Taking full advantage of the fragmental similarity between T4SEs, combined with machine
151 learning techniques, a comprehensive prediction pipeline (T4SEpp) was designed (Figure
152 1B and C). Several homology searching modules were developed to detect full-length
153 (flBlast), effector domain (effectHMM) and C-terminal signal region (sigHMM) homologs of
154 known T4SEs. A previous machine learning model T4SEpre, which predicts T4SEs based
155 on the C-terminal hand-crafted features, were re-trained with updated datasets(Wang, et
156 al., 2014). With the generative features from pre-trained transformers, we also developed
157 a deep learning module T4attention incorporated with the Bi-Conv attention mechanism.
158 Figure 1D showed the framework of T4SEpp, taking prediction scores of 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.

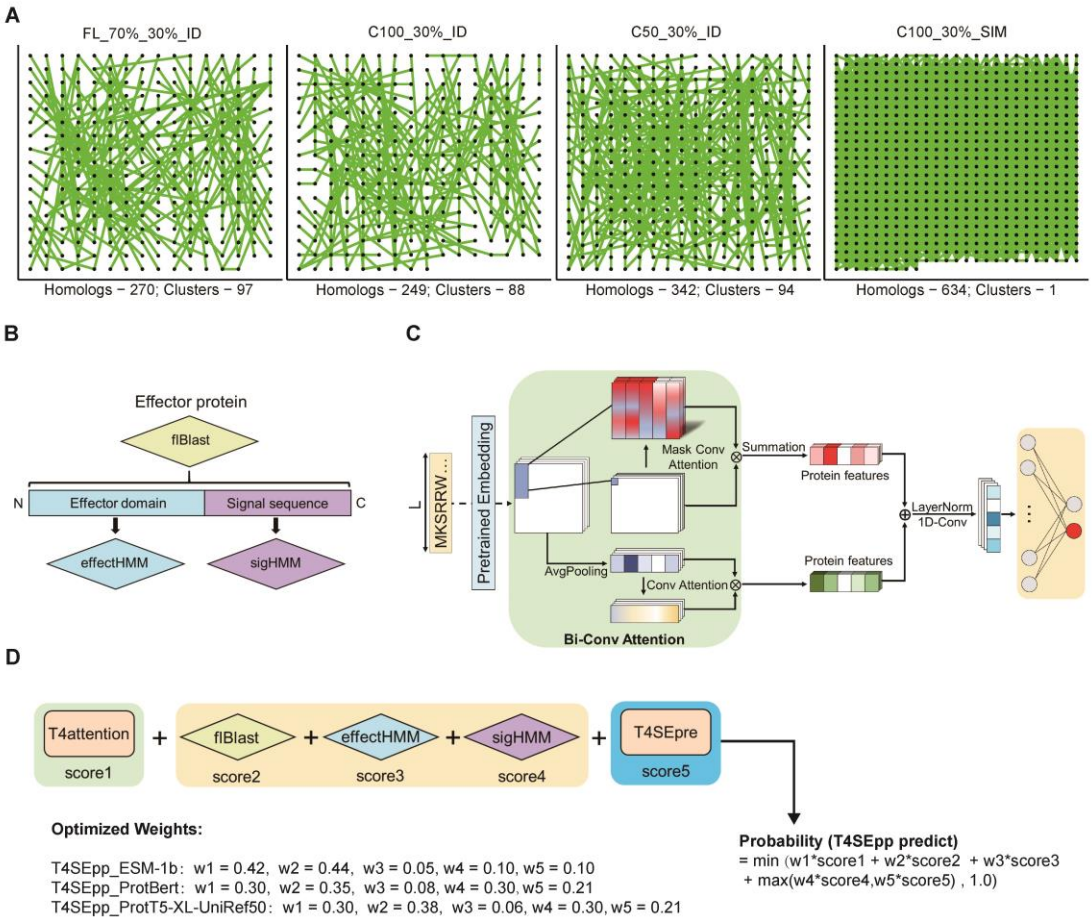


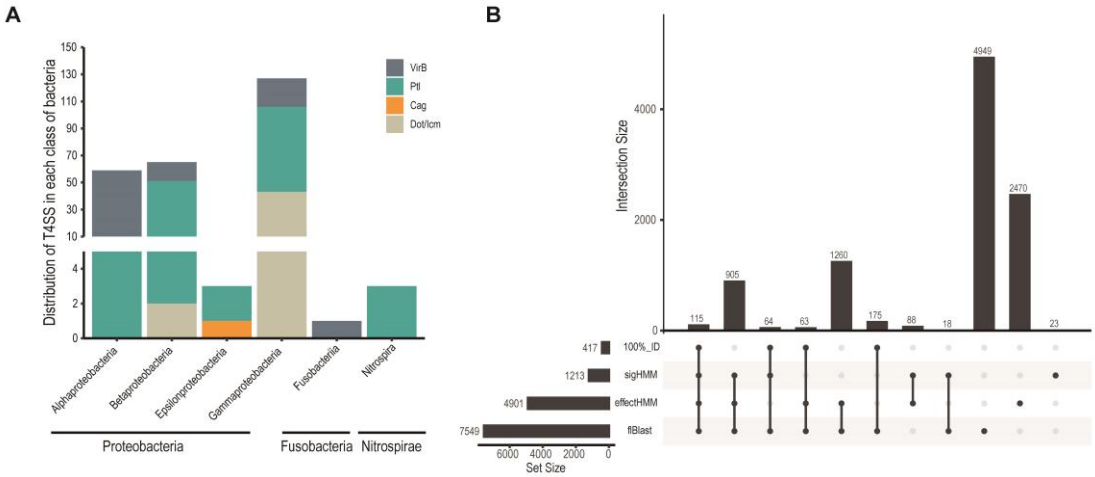
Figure 1. Sequence homology among T4S effectors and an integrated prediction framework. (A) Sequence homology network of T4SE. The nodes represented effectors with homology with at least one other effector. The pairs with homology (identified by the criteria defined at the top) were connected by green lines. (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 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, there remained 640 effectors with a length of >= 30 amino acids, among which 270 were classified into 106 multi-member families and 370 represented singlet families

176 (Supplementary Table S4). Effectors of the same signal sequence family could belong to
177 different effector functional domain families and *vice versa*. For example, six CagA
178 effectors and two *Legionella* proteins contained the signal sequences of the same family
179 (sigFAM50, Supplementary Table S3), but they also fell into three different effector
180 functional domain families (effectFAM73 for all the CagAs, and effectFAM19 and
181 effectFAM57 for the other two proteins; Supplementary Table S4). It could be related with
182 frequent domain reshuffling events that have been reported in *Legionella*(Burstein, et al.,
183 2016).

184 Furthermore, we searched the homologs of known T4SEs from the representative
185 bacterial genomes downloaded from UniProt (8761 genomes; Supplementary Table S5).
186 In total, 258 protein-translocating T4SSs were detected from 227 bacterial strains
187 distributed in 3 phyla (*Proteobacteria*, *Fusobacteria* and *Nitrospirae*), 6 classes
188 (*Alphaproteobacteria*, *Betaproteobacteria*, *Epsilonproteobacteria*, *Gammaproteobacteria*
189 *Fusobacteriia*, and *Nitrospira*), 117 genera and 227 species (Figure 2A, Supplementary
190 Table S6). In these strains with T4SSs, 10,130 proteins were detected with full-length or
191 local homology to the known T4SEs using the individual homology searching modules,
192 and 1,020 were identified by all the three modules (Figure 2B, Supplementary Table S7).



193
194 **Figure 2.** Search for T4SS and effectors in the UniProt reference proteome based on sequence
195 homology. (A) Using the core protein components of T4SS to construct a Hidden Markov Model (HMM) to
196 predict the distribution of T4SS in the UniProt reference proteome. (B) Three homologous modules
197 (sigHMM, effectHMM and flBlast) were used to predict the potential T4SE in the UniProt reference
198 proteome containing T4SS, respectively. Where 100%_ID represents a known verified T4SE.

199 **Prediction of T4SEs with pre-trained transformer-based models**

200 Recently, protein language models have been successfully applied in structure prediction
201 and sequence classification. In this research, we used six pre-trained models, ESM-1b,
202 ProtAlbert, ProtBert-BFD, ProtBert-UniRef100, ProtT5-XL-BFD and ProtT5-XL-UniRef50
203 to generate features, based on which we developed deep learning models (T4attention)
204 based on Bi-Conv attention respectively to classify T4SEs and non-T4SEs. The
205 T4attention models based on different sequence embedding features were compared for
206 the performance based on a five-fold cross-validation strategy (Table 1). Generally,
207 T4attention_ESM-1b performed best, followed by T4attention_ProtT5-XL-UniRef50, and
208 T4attention_ProtAlbert showed the poorest performance, according to MCC and F1-score
209 (Table 1). T4attention_ESM-1b did not only reach the highest MCC and F1-score (0.861
210 and 0.819 respectively), but required lowest computational resource (Supplementary
211 Figure S3). It was also noted that, for the same protein language model architecture,
212 ProtBert or ProtT5-XL for example, the generation of features from models pre-trained
213 from varied volume of protein database required similar computational resource but the
214 smaller database based pre-trained models always generated features for subsequent
215 T4attention models with better performance (MCC of T4attention_ProtBert vs
216 T4attention_ProtBert-BFD, 0.814 vs. 0.797; T4attetion_ProtT5-XL-UniRef50 vs.
217 ProtT5-XL-BFD, 0.818 vs. 0.800) (Table 1, Supplementary Figure S3). Redundancy of
218 protein sequences in the BFD dataset might lead to biases in model training, and further
219 caused the performance compromise of models addressing downstream tasks.

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

227 Considering the performance of models based on both cross-validation results and the
228 independent testing dataset, as well as the requirement of computational resources, we
229 integrated three models, including T4attention_ESM-1b, T4attention_ProtBert and
230 T4attention_ProtT5-XL-UniRef50, into the pipeline to predict T4SEs.

231 **An integrated pipeline predicting T4SEs with largely improved performance**

232 Besides the models based on the features generated with transformer, we also tested
233 traditional machine learning models based on hand-crafted features. To this end, we
234 retrained two models of T4SEpre (T4SEpre_psAac and T4SEpre_bpbAac) learning the
235 amino acid composition features in the C-termini of T4SEs(Wang, et al., 2014). Both
236 models showed certain performance in prediction of T4SEs according to the
237 cross-validation results or the independent testing dataset, although they appeared not
238 comparable to the T4attention models (Table 1; Table 2).

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

247 T4SEpp was also compared to other state-of-the-art(SOTA) T4SE prediction models,
248 such as Bastion4(Wang, et al., 2019), CNNT4SE(Hong, et al., 2020) and
249 T4SEfinder(Zhang, et al., 2022). Among these other models, Bastion4 showed the best
250 performance, which was close to that of the T4attention models but was far inferior to the
251 integrated T4SEpp (Table 2).

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

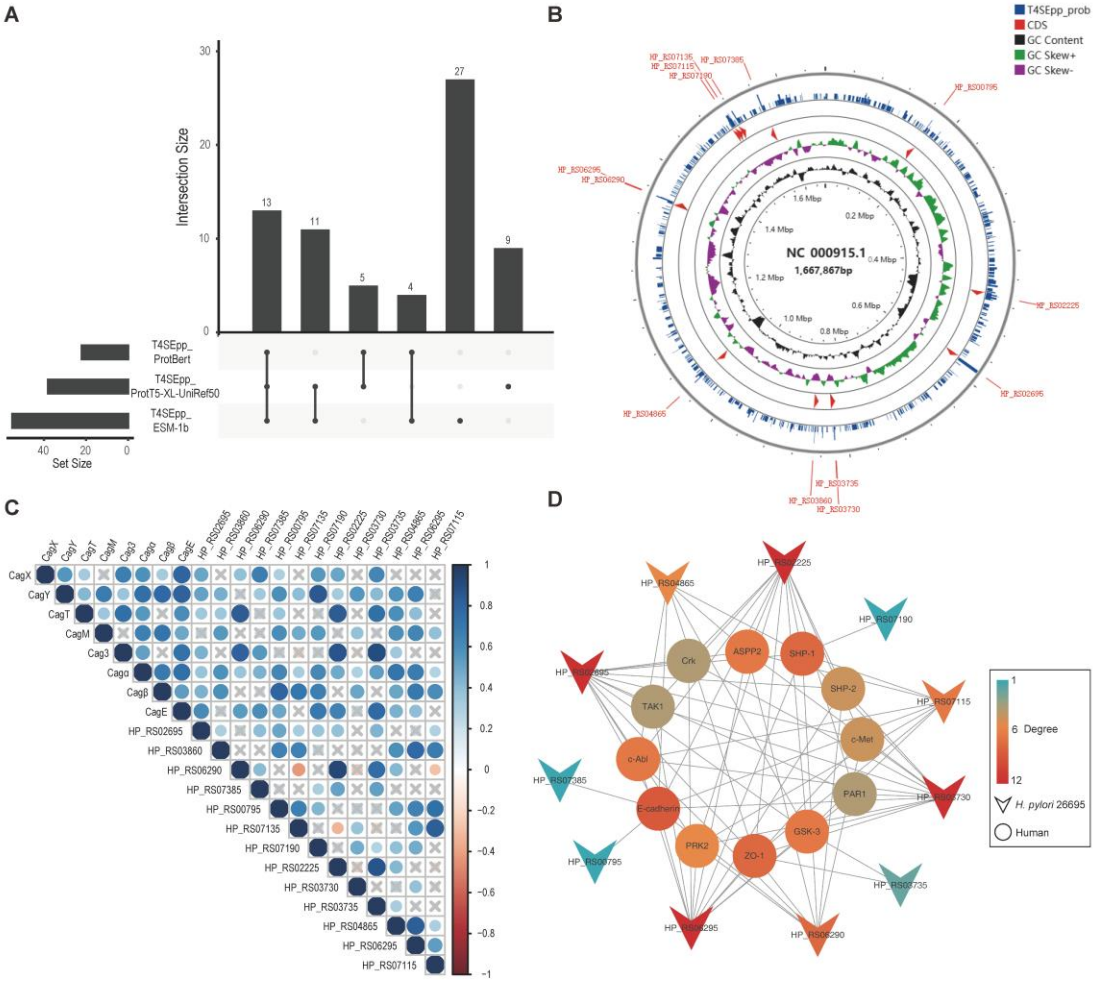
253 **bacteria**

254 *H. pylori* is a Gram-negative, spiral-shaped bacterium that colonizes the stomach of
255 approximately half of the world's population(Hooi, et al., 2017). Although most individuals
256 never experience any adverse health outcomes attributable to *H. pylori*, the presence of
257 these bacteria in the stomach increases the risk of developing gastric diseases(Blaser, et
258 al., 1995; Cover, 2016; Cover and Blaser, 2009; Figueiredo, et al., 2002; Plummer, et al.,
259 2007). *H. pylori* infection is also the strongest known risk factor of gastric cancer, the third
260 leading cause of cancer-related death worldwide(Bray, et al., 2018). The T4SS play an
261 important role in the pathogenicity of *H. pylori*(Blaser, et al., 1995; Cover, 2016; Figueiredo,
262 et al., 2002; Plummer, et al., 2007). However, to date, only one T4SE, CagA, has been
263 identified for the T4SS in *H. pylori*(Knorr, et al., 2019). Here, we applied T4SEpp to screen
264 T4SE candidates from the proteins derived from the genome of *H. pylori* 26695, a model
265 *H. pylori* strain (NCBI accession number: NC_000915.1). The three T4SEpp integrated
266 models, i.e., T4SEpp_ESM-1b, T4SEpp_ProtBert, and T4SEpp_ProtT5-XL-UniRef50,
267 predicted 55, 22 and 38 T4SE candidates respectively, and 13 were shared by the
268 prediction results of all the three models (Figure 3A-B; Supplementary Tables S6, S8).
269 The 13 potential effector genes were scattered in the genome (Figure 3B). Notably,
270 *HP_RS02695* encoding the only known effector CagA, was among the 13 candidates
271 (Figure 3B).

272 Gene co-expression was analyzed for the 13 T4SE candidate genes in *H. pylori 26695*
273 with an RNA-seq dataset sampled from the strain collected at 12 different conditions(Loh,
274 et al., 2021). Except for *HP_RS06290*, *HP_RS03730*, *HP_RS04865*, and *HP_RS06295*,
275 the rest eight genes showed strong expression correlation with *cagA* (Figure 3C). The
276 genes co-expressed with *cagA* also showed significant correlation with the expression of
277 the core component genes of the Cag T4SS (Figure 3C). Furthermore, we annotated 12
278 human proteins that showed experimentally verified interaction with CagA by literature
279 searching, including ASPP2, c-Abl, c-Met, Crk, E-cadherin, GSK-3, PAR1, PRK2, SHP-1,
280 SHP-2, TAK1, and ZO-1(Amieva, et al., 2003; Churin, et al., 2003; Higashi, et al., 2002;

Lamb, et al., 2009; Lee, et al., 2014; Mishra, et al., 2015; Murata-Kamiya, et al., 2007; Nesic, et al., 2014; Poppe, et al., 2007; Saadat, et al., 2007; Saju, et al., 2016; Suzuki, et al., 2005). 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 interaction with all the 12 human proteins (Figure 3D). Taken together, the proteins predicted by T4SEpp could potentially represented new T4SEs, or closely related with the pathogenicity of *H. pylori 26695*.

We also utilized 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 ones co-predicted by all the three T4SEpp models (Supplementary Table S9, Supplementary Figure S5).



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

305 **Web server and implementation of T4SEpp**

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

312 **Discussion**

313 T4SS plays a crucial role in bacterial pathogenicity by secreting effectors into host cells. *L.*
314 *pneumophila* can translocated more than 300 known effectors into human cells by the
315 Dot/Icm T4SS system, causing legionellosis(Goncalves, et al., 2021; Mondino, et al.,
316 2020). In *H. pylori*, CagA is the only known T4SE, which can hijack multiple signaling
317 pathways in gastric epithelial cells, leading to gastritis, gastric ulcer and even gastric
318 cancer(Hatakeyama, 2004; Hatakeyama, 2014). Identifying the full repertoire of T4SEs in
319 a pathogen is important to understand its pathogenic mechanisms. Computational
320 methods can assist identification of new effectors effectively(Zhao, et al., 2023). However,
321 the currently avaialbe T4SE prediction tools still show high false positive rates(Hui, et al.,
322 2021). To address this issue, we developed a unified T4SE prediction pipeline, T4SEpp,
323 which includes homologous search modules, traditional machine learning modules and
324 natural language processing-based modules. T4SEpp outperforms other SOTA methods

325 predicting T4SEs, with apparently improved sensitivity and specificity. Furthermore, we
326 initiated a web server, which can implement the T4SEpp pipeline conveniently, providing
327 the prediction results for each module meanwhile.

328 Although the component modules of T4SEpp can be used for T4SE prediction, they often
329 show higher false positive rates when used alone. It could be related with the low power of
330 individual dimensions of features. Specifically, T4SE signal sequences were considered to
331 contain important common features guiding T4SE secretion and translocation, which were
332 used for effective T4SE prediction by some tools such as T4SEpre(Wang, et al., 2014).
333 However, the computational models only based on the signal sequences showed
334 performance inferior to other models based on the multiple-aspect features extracted from
335 full-length proteins(Wang, et al., 2019). In this study, we discovered high sequence
336 similarity in the C-terminal signal region, among the proteins without apparent homology
337 for full-length effectors. Such undetected homology could have introduced bias and led to
338 overfitting of various established machine learning algorithms and the discrepancy
339 between the reported and actual accuracy of these methods. However, the C-terminal
340 homology could also suggest the independent evolution of the signal sequences, and it
341 could potentially be applied to facilitate identification of new effectors(Meyer, et al., 2013).

342 In this study, we integrated three types of modules to predict T4SEs. The homology
343 searching based modules provide more accurate results, but they also show lower
344 capacity to detect new effectors without or with remote homology. The re-trained T4SEpre
345 modules focus on the important features in C-terminal signal sequences of T4SEs.
346 T4attention learns from the full-length effector proteins the features generated by pLMs
347 pre-trained with large-scale protein databases. These pLM-based models can learn new,
348 previous unknown features that may involve position-position interactions, and have
349 demonstrated outstanding performance in prediction of proteins with various biological
350 function, such as subcellular localization, secondary structure, etc. We used multiple
351 pLMs to build transfer learning models, most of which showed excellent performance in
352 T4SE prediction. Interestingly, we noticed that the pre-trained pLMs based on the larger

datasets did not generate better prediction performance. The pLMs pre-trained on smaller datasets were more efficient. Therefore, the transferring models were trained with the pLMs based on smaller non-redudant protein datasets. T4SEpp that integrated all the three types of modules outperformed both the individual modules and other similar applications significantly.

With T4SEpp, we analyzed the potential new T4SEs in both *H. pylori* and other strains bearing a T4SS. In *H. pylori*, we disclosed 12 new T4SEs. We also disclosed 12,205 new T4SEs and 417 known T4SEs from all the 227 strains bearing a T4SS. The results suggested that there much more new effectors yet to be clarified.

Despite the large performance improvement of T4SEpp, there remains space 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 or organelle-targeting signals, etc, have not been integrated in the current version of the model(Zhao, et al., 2023). Novel features remain to be disclosed, which could further be integrated to improve the model performance. The different types (IVA and IVB) of effectors, chaperone-dependent or chaperone-independent effectors, or species-specific effectors could also be modeled and predicted separately, to make more accurate prediction(Zhao, et al., 2023).

373 **Materials and methods**

374 **Datasets**

375 The 390 T4SEs used by Bastion4 as the positive training dataset(Wang, et al., 2019) and
376 540 T4SEs annotated in SecReT4 v2.0(Bi, et al., 2013) were collected and merged, and
377 in total we got 653 non-identical, validated T4SEs. CD-HIT(Li and Godzik, 2006) was used
378 to filter the homology-redundant proteins with sequence identity ≥ 60%, generating 518
379 non-redundant T4SEs, which were used as the positive training dataset(Supplementary
380 Figure S1A). For the negative training dataset, we collected the 1112 and 1548 non-T4SE
381 protein sequences from Bastion4(Wang, et al., 2019) and PredT4SE-stack(Xiong, et al.,
382 2018), respectively. The same procedure was used to eliminate the sequence redundancy
383 among the non-T4SEs and between the non-T4SEs and the T4SEs in the positive training
384 dataset, generating 1590 non-redundant non-T4SEs (Supplementary Figure S1A). An
385 independent validation dataset was also prepared, for which the T4SEs were collected
386 from the testing dataset of Bastion4 (30) and others (74) annotated from literatures
387 published recently (Supplementary Table S1), and the 150 testing non-T4SEs of Bastion4
388 were also used as the negative ones. CD-HIT was used to filter the redundant proteins
389 with ≥60% sequence identity to the training proteins and among proteins of the validation
390 dataset, resulting in 20 non-redundant T4SEs and 150 non-T4SEs (Supplementary Figure
391 S1B).

392 **Homology networks of the T4SE peptide sequences**

393 The sequences of 653 non-identical verified T4SE proteins were used to construct the
394 homology networks. JAligner implementing the Smith-Waterman algorithm was used to
395 find the similarity between any pair of full-length effectors or peptide fragments with
396 designated length (http://jaligner.sourceforge.net/). The identity and similarity percentages
397 between any pair of sequences were used as measures to determine the homology
398 level(Hui, et al., 2020). Networks were built and visualized with Cytoscape
399 v3.9.1(Shannon, et al., 2003).

400 **Genome-wide screening of protein-translocation T4SSs**

401 The conserved core component proteins were collected from four representative
402 protein-translocation T4SSs, including the *Agrobacterium tumefaciens* VirB/VirD4 T4SS

403 (inner membrane complex proteins VirB3, VirB6, VirB8, VirB10 and VirD4, and outer
404 membrane complex proteins VirB7, VirB9 and VirB10)(Costa, et al., 2021), the *Bordetella*
405 *pertussis* Ptl T4SS (inner membrane complex proteins PtlB, PtlE and PtlH, and outer
406 membrane complex proteins PtlF and PtlG)(O'Callaghan, et al., 1999), the *Helicobacter*
407 *pylori* Cag T4SS (inner membrane complex proteins Cagα, Cagβ and CagE, and outer
408 membrane complex proteins CagX, CagY, CagT, CagM and Cag3)(Cover, et al., 2020),
409 the *Legionella pneumophila* Dot/Icm T4SS (inner membrane complex proteins IcmB,
410 IcmG and DotB, and outer membrane complex proteins DotC, DotD, DotG and
411 IcmK)(Costa, et al., 2021). Hidden Markov Model (HMM) profiles were built with HMMER
412 3.1 for the T4SS component protein families(Eddy, 1998). Protein sequences derived from
413 the 8761 reference bacterial genomes curated in UniProt were scanned with HMMER and
414 the HMM profiles for the distribution of homologs of T4SS core component proteins
415 (Supplementary Table S5).

416 **Homology-based T4SE detection modules**

417 The diamond blastp was used to determine the homology and cluster the full-length
418 effector proteins(Buchfink, et al., 2015), and to screen new full-length homologs (flBlast).
419 Two proteins showing ≥30% similarity for ≥70% of the full length of either protein were
420 considered to be full-length homologs(Hu, et al., 2017; Hui, et al., 2020). The C-terminal
421 50-aa signal sequences of verified effectors were clustered according to the homology
422 networks with 30% identity for 70% length aligned by JAligner. HMM profiles were built for
423 each signal sequence family and a module sigHMM was developed to screen proteins
424 with C-terminal sequences homologous to the profiles of known T4SE signal sequence
425 families. The homology cutoff for HMM searching was optimized to each family, ensuring
426 all or most of the known effectors recalled and maintaining a higher specificity. For
427 effectHMM, we removed the C-terminal 50-aa signal from each known effector sequence,
428 and the remaining peptide fragment with >30-aa length was used for domain clustering.
429 Pairwise alignment was repeatedly performed with BLAST between the domain
430 sequences, and the cutoff for homology was optimized based on the average coverage of
431 aligned length multiplying the identity, i.e., ≥10(Hui, et al., 2020). The HMM profiles were
432 built for the effector domain families and effectHMM was developed with a similar

433 procedure for sigHMM, to screen the proteins with homologous T4SE effector-domains.
434 We use a statistical method, EBT, to compare general homology between proteins(Hui, et
435 al., 2020; Hui, et al., 2017).
436
437 **Re-training of T4SEpre models with updated datasets**
438 T4SEpre models (T4SEpre_psAac and T4SEpre_bpbAac) were retrained with the new
439 training datasets of T4SEs and non-T4SEs. The procedure of original T4SEpre was
440 followed for feature representation, parameter optimization and model training(Wang, et
441 al., 2014). Briefly, sequential amino acid, bi-residue and motif composition features and
442 position-specific amino acid composition profile for the positive training dataset were
443 represented for each C-terminial 100-aa sequence for the psAac model. For the bpbAac
444 model, position-specific amino acid composition profiles of both the positive and the
445 negative training datasets (Bi-Profile Bayesian features) were represented for each
446 C-terminial 100-aa sequence. Support vector machine (SVM) models were trained for the
447 feature matrices. The kernel functions, i.e., linear, polynomial, sigmoid, and radial base
448 function (RBF), and corresponding parameters (cost and gamma) were optimized using a
449 5-fold cross-validation grid search strategy. The sklearn v1.0.1 was used for implementing
450 SVM model training and kernel/parameter optimization.
451 **The deep learning architecture of T4attention based on pre-trained protein**
452 **language models**
453 Input embeddings. We extracted frozen embeddings directly from protein language
454 models (pLMs) without fine-tuning on training data. Four different basic LMs were used in
455 this study, and in total six different pLMs were pre-trained with different datasets. The
456 basid LMs include, (i) “ESM-1b”(Rives, et al., 2021), which is a Transformer model, (ii)
457 "ProtBert" (Elnaggar, et al., 2022), which is a BERT-based encoder model(Devlin, et al.,
458 2019), generating two pLMs pre-trained on BFD(Steinegger and Soding, 2018) and
459 UniRef100(Suzek, et al., 2015) data, respectively, (iii) ProtT5-XL(Elnaggar, et al., 2022),
460 which is an encoder model based on T5(Raffel, et al., 2020), generating two pLMs
461 pre-trained on BFD and UniRef50, respectively, and (iv) ProtAlbert(Elnaggar, et al., 2022),

which is an encoder model based on Albert(Lan, et al., 2019) and pre-trained only with UniRef100.

Optimization strategy. We use a BERT-like optimizer AdamW and a Cosine Warm-up strategy(Devlin, et al., 2019) to optimize the loss of the learning model, the initial learning rate is set to 0.0001, batch size is set to 18 and the warm-up steps are set to 10. Early stopping strategy was applied to monitor the validation ACC, with a patience of 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(Lin, et al., 2020). Focal Loss is to increase the hyperparameter γ (default γ=2) based on 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 sample t,

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

T4attention model. The input to T4attention (Figure 1C, Supplement Figure S2) is a protein embedding $\boldsymbol{E_0} \in \mathbb{R}^{n\times d_0}$, where **n** is the sequence length and $\mathbf{d_0}$ is the size of the embedding (depending on the feature extraction model). T4attention is a model based on the 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, $\boldsymbol{e, v} \in \mathbb{R}^{d_1}$. Thus, we get the feature representation of the embedding dimension $\boldsymbol{x = softmax(e) \times v}$. In the direction of the protein sequence, we randomly intercept the length of **m** in the length direction of the protein embedding sequence, so that the protein embedding becomes $\boldsymbol{E_1} \in \mathbb{R}^{m\times d_0}$. Similar to the convolutional attention calculation in the protein embedding direction, the attention coefficient **e'** and value **v'** are obtained, $\boldsymbol{e', v'} \in \mathbb{R}^{m\times d_1}$, and the difference is that the direction of the convolution is in the direction of the sequence length, so that we can get the feature representation of the protein sequence direction, and converge according to the sequence length direction by $\boldsymbol{x' = \sum_i^m softmax(e') \times v'}$. The convolution attention results of the embedding direction

and the protein sequence direction are merged, passed through the LayerNorm and the residual one-dimensional convolution, and finally the class probabilities are obtained through the two-layer multi-layer perceptron (MLP), $\boldsymbol{p}(\mathbf{c}|\boldsymbol{x}) = \boldsymbol{softmax}(\boldsymbol{MLP}\big(\boldsymbol{Conv}(\boldsymbol{x} + \boldsymbol{x}') + (\boldsymbol{x} + \boldsymbol{x}')\big))$, where **c** indicates the category of the output (i.e., T4SE or nonT4SE). T4attention was developed with PyTorch v1.10.1. The models were trained and evaluated with 24-GB memory and 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 the homology-searching modules for full-length or fragmental effector proteins, the 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. A weight $\mathbf{x}$ was assigned empirically to each module, where $\mathbf{x} \in (\mathbf{0}, \mathbf{0.50})$. The maximum T4SEpp predicted value was set to 1.0. We trained the model using grid search with 5-fold cross-validation to find 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

The 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 the measures were defined as:

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

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

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

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

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

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

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 data sets of *H. pylori* 26695 under different conditions were downloaded from NCBI GEO DataSets database with accessions GSE165055 and GSE165056(Loh, et al., 2021). After removing the adapters and low-quality sequences with Trimmomatic v0.39(Bolger, et al., 2014), the cleaned reads were mapped to the *H. pylori* 26695 reference genome (NC_000915.1) with READemption (Version 2.0.0)(Forstner, et al., 2014). The annotated genes were further quantified and analyzed.

Availability

The online version of T4SEpp is freely accessible at https://bis.zju.edu.cn/T4SEpp. The standalone version of T4SEpp model and the individual modules are are also deposited at https://github.com/yuemhu/T4SEpp.

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; 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 made the model comparison and RNA-seq data analysis. YH, XH, HC, SL, QN, YW and MC wrote the first draft. YH, YW and MC revised the manuscript.

Conflict of Interest: none declared.

Key Points：

1. This study employed full-length embedding features from six pre-trained protein language models to train classifiers for T4SE prediction.
2. A pipeline for was developed for highly accurate T4SE prediction, which integrates modules for homology searching, machine learning modules based on hand-crafted features, and protein language pre-trained models.
3. T4SEpp exhibited superior performance compared to state-of-the-art software tools, demonstrating high sensitivity and specificity on an independent test dataset.
4. T4SEpp successfully predicted T4SEs in 220+ bacterial genomes from 3 phyla and 117 genera, identifying a total of 12,622 plausible T4SEs.
5. A publicly available standalone version and web server are provided for genome-scale T4SE identification.

561 **References**

562 Amieva, M.R., Vogelmann, R., Covacci, A.*, et al.* Disruption of the epithelial apical-junctional
563 complex by Helicobacter pylori CagA. *Science* 2003;300(5624):1430-1434.
564 Ansari, S. and Yamaoka, Y. Helicobacter pylori Virulence Factor Cytotoxin-Associated Gene A
565 (CagA)-Mediated Gastric Pathogenicity. *Int J Mol Sci* 2020;21(19).
566 Bi, D., Liu, L., Tai, C.*, et al.* SecReT4: a web-based bacterial type IV secretion system resource.
567 *Nucleic Acids Res* 2013;41(Database issue):D660-665.
568 Blaser, M.J., Perez-Perez, G.I., Kleanthous, H.*, et al.* Infection with Helicobacter pylori strains
569 possessing cagA is associated with an increased risk of developing adenocarcinoma of the
570 stomach. *Cancer Res* 1995;55(10):2111-2115.
571 Bolger, A.M., Lohse, M. and Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence
572 data. *Bioinformatics* 2014;30(15):2114-2120.
573 Bray, F., Ferlay, J., Soerjomataram, I.*, et al.* Global cancer statistics 2018: GLOBOCAN
574 estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J*
575 *Clin* 2018;68(6):394-424.
576 Buchfink, B., Xie, C. and Huson, D.H. Fast and sensitive protein alignment using DIAMOND.
577 *Nat Methods* 2015;12(1):59-60.
578 Burns, D.L. Type IV transporters of pathogenic bacteria. *Curr Opin Microbiol* 2003;6(1):29-34.
579 Burstein, D., Amaro, F., Zusman, T.*, et al.* Genomic analysis of 38 Legionella species identifies
580 large and diverse effector repertoires. *Nat Genet* 2016;48(2):167-175.
581 Burstein, D., Zusman, T., Degtyar, E.*, et al.* Genome-scale identification of Legionella
582 pneumophila effectors using a machine learning approach. *PLoS Pathog* 2009;5(7):e1000508.
583 Chandran Darbari, V. and Waksman, G. Structural Biology of Bacterial Type IV Secretion
584 Systems. *Annu Rev Biochem* 2015;84:603-629.
585 Christie, P.J. The Mosaic Type IV Secretion Systems. *EcoSal Plus* 2016;7(1).
586 Christie, P.J., Gomez Valero, L. and Buchrieser, C. Biological Diversity and Evolution of Type
587 IV Secretion Systems. *Curr Top Microbiol Immunol* 2017;413:1-30.
588 Churin, Y., Al-Ghoul, L., Kepp, O.*, et al.* Helicobacter pylori CagA protein targets the c-Met
589 receptor and enhances the motogenic response. *J Cell Biol* 2003;161(2):249-255.
590 Costa, T.R., Felisberto-Rodrigues, C., Meir, A.*, et al.* Secretion systems in Gram-negative
591 bacteria: structural and mechanistic insights. *Nat Rev Microbiol* 2015;13(6):343-359.
592 Costa, T.R.D., Harb, L., Khara, P.*, et al.* Type IV secretion systems: Advances in structure,
593 function, and activation. *Mol Microbiol* 2021;115(3):436-452.
594 Cover, T.L. Helicobacter pylori Diversity and Gastric Cancer Risk. *mBio*
595 2016;7(1):e01869-01815.
596 Cover, T.L. and Blaser, M.J. Helicobacter pylori in health and disease. *Gastroenterology*
597 2009;136(6):1863-1873.
598 Cover, T.L., Lacy, D.B. and Ohi, M.D. The Helicobacter pylori Cag Type IV Secretion System.
599 *Trends Microbiol* 2020;28(8):682-695.
600 Devlin, J., Chang, M.-W., Lee, K.*, et al.* BERT: Pre-training of Deep Bidirectional Transformers
601 for Language Understanding. In. Minneapolis, Minnesota: Association for Computational
602 Linguistics; 2019. p. 4171-4186.
603 Dong, X., Lu, X. and Zhang, Z. BEAN 2.0: an integrated web resource for the identification and
604 functional analysis of type III secreted effectors. *Database (Oxford)* 2015;2015:bav064.

605 Eddy, S.R. Profile hidden Markov models. *Bioinformatics* 1998;14(9):755-763.
606 Elnaggar, A., Heinzinger, M., Dallago, C.*, et al.* ProtTrans: Toward Understanding the
607 Language of Life Through Self-Supervised Learning. *IEEE Trans Pattern Anal Mach Intell*
608 2022;44(10):7112-7127.
609 Figueiredo, C., Machado, J.C., Pharoah, P.*, et al.* Helicobacter pylori and interleukin 1
610 genotyping: an opportunity to identify high-risk individuals for gastric carcinoma. *J Natl Cancer*
611 *Inst* 2002;94(22):1680-1687.
612 Forstner, K.U., Vogel, J. and Sharma, C.M. READemption-a tool for the computational analysis
613 of deep-sequencing-based transcriptome data. *Bioinformatics* 2014;30(23):3421-3423.
614 Galan, J.E. and Waksman, G. Protein-Injection Machines in Bacteria. *Cell*
615 2018;172(6):1306-1318.
616 Goldberg, T., Rost, B. and Bromberg, Y. Computational prediction shines light on type III
617 secretion origins. *Sci Rep* 2016;6:34516.
618 Goncalves, I.G., Simoes, L.C. and Simoes, M. Legionella pneumophila. *Trends Microbiol*
619 2021;29(9):860-861.
620 Gonzalez-Rivera, C., Bhatty, M. and Christie, P.J. Mechanism and Function of Type IV
621 Secretion During Infection of the Human Host. *Microbiol Spectr* 2016;4(3).
622 Grohmann, E., Christie, P.J., Waksman, G.*, et al.* Type IV secretion in Gram-negative and
623 Gram-positive bacteria. *Mol Microbiol* 2018;107(4):455-471.
624 Hatakeyama, M. Oncogenic mechanisms of the Helicobacter pylori CagA protein. *Nat Rev*
625 *Cancer* 2004;4(9):688-694.
626 Hatakeyama, M. Helicobacter pylori CagA and gastric cancer: a paradigm for hit-and-run
627 carcinogenesis. *Cell Host Microbe* 2014;15(3):306-316.
628 Heinzinger, M., Elnaggar, A., Wang, Y.*, et al.* Modeling aspects of the language of life through
629 transfer-learning protein sequences. *BMC Bioinformatics* 2019;20(1):723.
630 Higashi, H., Tsutsumi, R., Muto, S.*, et al.* SHP-2 tyrosine phosphatase as an intracellular target
631 of Helicobacter pylori CagA protein. *Science* 2002;295(5555):683-686.
632 Hofreuter, D., Odenbreit, S., Puls, J.*, et al.* Genetic competence in Helicobacter pylori:
633 mechanisms and biological implications. *Res Microbiol* 2000;151(6):487-491.
634 Hong, J., Luo, Y., Mou, M.*, et al.* Convolutional neural network-based annotation of bacterial
635 type IV secretion system effectors with enhanced accuracy and reduced false discovery. *Brief*
636 *Bioinform* 2020;21(5):1825-1836.
637 Hooi, J.K.Y., Lai, W.Y., Ng, W.K.*, et al.* Global Prevalence of Helicobacter pylori Infection:
638 Systematic Review and Meta-Analysis. *Gastroenterology* 2017;153(2):420-429.
639 Hu, Y., Huang, H., Cheng, X.*, et al.* A global survey of bacterial type III secretion systems and
640 their effectors. *Environ Microbiol* 2017;19(10):3879-3895.
641 Hubber, A. and Roy, C.R. Modulation of host cell function by Legionella pneumophila type IV
642 effectors. *Annu Rev Cell Dev Biol* 2010;26:261-283.
643 Hui, X., Chen, Z., Lin, M.*, et al.* T3SEpp: an Integrated Prediction Pipeline for Bacterial Type III
644 Secreted Effectors. *mSystems* 2020;5(4).
645 Hui, X., Chen, Z., Zhang, J.*, et al.* Computational prediction of secreted proteins in
646 gram-negative bacteria. *Comput Struct Biotechnol J* 2021;19:1806-1828.
647 Hui, X., Hu, Y., Sun, M.A.*, et al.* EBT: a statistic test identifying moderate size of significant
648 features with balanced power and precision for genome-wide rate comparisons. *Bioinformatics*

649 2017;33(17):2631-2641.
650 Knorr, J., Ricci, V., Hatakeyama, M., *et al.* Classification of Helicobacter pylori Virulence
651 Factors: Is CagA a Toxin or Not? *Trends Microbiol* 2019;27(9):731-738.
652 Lamb, A., Yang, X.D., Tsang, Y.H., *et al.* Helicobacter pylori CagA activates NF-kappaB by
653 targeting TAK1 for TRAF6-mediated Lys 63 ubiquitination. *EMBO Rep*
654 2009;10(11):1242-1249.
655 Lan, Z., Chen, M., Goodman. S, *et al.* ALBERT: A Lite BERT for Self-supervised Learning of
656 Language Representations. In, *International Conference on Learning Representations*. 2019.
657 Lee, D.G., Kim, H.S., Lee, Y.S., *et al.* Helicobacter pylori CagA promotes Snail-mediated
658 epithelial-mesenchymal transition by reducing GSK-3 activity. *Nat Commun* 2014;5:4423.
659 Lee, J., Yoon, W., Kim, S., *et al.* BioBERT: a pre-trained biomedical language representation
660 model for biomedical text mining. *Bioinformatics* 2020;36(4):1234-1240.
661 Lee, Y.W., Wang, J., Newton, H.J., *et al.* Mapping bacterial effector arsenals: in vivo and in
662 silico approaches to defining the protein features dictating effector secretion by bacteria. *Curr*
663 *Opin Microbiol* 2020;57:13-21.
664 Li, W. and Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein
665 or nucleotide sequences. *Bioinformatics* 2006;22(13):1658-1659.
666 Li, Y.G., Hu, B. and Christie, P.J. Biological and Structural Diversity of Type IV Secretion
667 Systems. *Microbiol Spectr* 2019;7(2).
668 Lin, T.Y., Goyal, P., Girshick, R., *et al.* Focal Loss for Dense Object Detection. *IEEE Trans*
669 *Pattern Anal Mach Intell* 2020;42(2):318-327.
670 Loh, J.T., Shum, M.V., Jossart, S.D.R., *et al.* Delineation of the pH-Responsive Regulon
671 Controlled by the Helicobacter pylori ArsRS Two-Component System. *Infect Immun*
672 2021;89(4).
673 Meyer, D.F., Noroy, C., Moumene, A., *et al.* Searching algorithm for type IV secretion system
674 effectors 1.0: a tool for predicting type IV effectors and exploring their genomic context.
675 *Nucleic Acids Res* 2013;41(20):9218-9229.
676 Mishra, J.P., Cohen, D., Zamperone, A., *et al.* CagA of Helicobacter pylori interacts with and
677 inhibits the serine-threonine kinase PRK2. *Cell Microbiol* 2015;17(11):1670-1682.
678 Mondino, S., Schmidt, S., Rolando, M., *et al.* Legionnaires' Disease: State of the Art
679 Knowledge of Pathogenesis Mechanisms of Legionella. *Annu Rev Pathol* 2020;15:439-466.
680 Murata-Kamiya, N., Kurashima, Y., Teishikata, Y., *et al.* Helicobacter pylori CagA interacts with
681 E-cadherin and deregulates the beta-catenin signal that promotes intestinal transdifferentiation
682 in gastric epithelial cells. *Oncogene* 2007;26(32):4617-4626.
683 Nesic, D., Buti, L., Lu, X., *et al.* Structure of the Helicobacter pylori CagA oncoprotein bound to
684 the human tumor suppressor ASPP2. *Proc Natl Acad Sci U S A* 2014;111(4):1562-1567.
685 O'Callaghan, D., Cazevieille, C., Allardet-Servent, A., *et al.* A homologue of the Agrobacterium
686 tumefaciens VirB and Bordetella pertussis Ptl type IV secretion systems is essential for
687 intracellular survival of Brucella suis. *Mol Microbiol* 1999;33(6):1210-1220.
688 Plummer, M., van Doorn, L.J., Franceschi, S., *et al.* Helicobacter pylori cytotoxin-associated
689 genotype and gastric precancerous lesions. *J Natl Cancer Inst* 2007;99(17):1328-1334.
690 Poppe, M., Feller, S.M., Romer, G., *et al.* Phosphorylation of Helicobacter pylori CagA by c-Abl
691 leads to cell motility. *Oncogene* 2007;26(24):3462-3472.
692 Raffel, C., Shazeer, N., Roberts, A., *et al.* Exploring the Limits of Transfer Learning with a

Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 2020;21:1-67.
Rao, R., Bhattacharya, N., Thomas, N.*, et al.* Evaluating Protein Transfer Learning with TAPE. *Adv Neural Inf Process Syst* 2019;32:9689-9701.
Rives, A., Meier, J., Sercu, T.*, 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).
Saadat, I., Higashi, H., Obuse, C.*, et al.* Helicobacter pylori CagA targets PAR1/MARK kinase to disrupt epithelial cell polarity. *Nature* 2007;447(7142):330-333.
Saju, P., Murata-Kamiya, N., Hayashi, T.*, et al.* Host SHP1 phosphatase antagonizes Helicobacter pylori CagA and can be downregulated by Epstein-Barr virus. *Nat Microbiol* 2016;1:16026.
Schrammeijer, B., den Dulk-Ras, A., Vergunst, A.C.*, et al.* 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):860-868.
Shannon, P., Markiel, A., Ozier, O.*, et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;13(11):2498-2504.
Sheedlo, M.J., Ohi, M.D., Lacy, D.B.*, et al.* Molecular architecture of bacterial type IV secretion systems. *PLoS Pathog* 2022;18(8):e1010720.
Stärk, H., Dallago, C., Heinzinger, M.*, et al.* Light attention predicts protein location from the language of life. *Bioinform Adv* 2021;1(1):vbab035.
Steinegger, M. and Soding, J. Clustering huge protein sequence sets in linear time. *Nat Commun* 2018;9(1):2542.
Suzek, B.E., Wang, Y., Huang, H.*, et al.* UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 2015;31(6):926-932.
Suzuki, M., Mimuro, H., Suzuki, T.*, et al.* 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):1235-1247.
Teufel, F., Almagro Armenteros, J.J., Johansen, A.R.*, et al.* SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat Biotechnol* 2022;40(7):1023-1025.
Wagner, N., Alburquerque, M., Ecker, N.*, et al.* Natural language processing approach to model the secretion signal of type III effectors. *Front Plant Sci* 2022;13:1024405.
Wagner, N., Avram, O., Gold-Binshtok, D.*, et al.* Effectidor: an automated machine-learning-based web server for the prediction of type-III secretion system effectors. *Bioinformatics* 2022;38(8):2341-2343.
Waksman, G. From conjugation to T4S systems in Gram-negative bacteria: a mechanistic biology perspective. *EMBO Rep* 2019;20(2).
Wallden, K., Rivera-Calzada, A. and Waksman, G. Type IV secretion systems: versatility and diversity in function. *Cell Microbiol* 2010;12(9):1203-1212.
Wang, J., Yang, B., An, Y.*, et al.* Systematic analysis and prediction of type IV secreted effector proteins by machine learning approaches. *Brief Bioinform* 2019;20(3):931-951.
Wang, Y., Wei, X., Bao, H.*, et al.* Prediction of bacterial type IV secreted effectors by C-terminal features. *BMC Genomics* 2014;15:50.
Ward, D.V. and Zambryski, P.C. The six functions of Agrobacterium VirE2. *Proc Natl Acad Sci U S A* 2001;98(2):385-386.

Wozniak, R.A. and Waldor, M.K. Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. *Nat Rev Microbiol* 2010;8(8):552-563.

Xiong, Y., Wang, Q., Yang, J.*, et al.* PredT4SE-Stack: Prediction of Bacterial Type IV Secreted Effectors From Protein Sequences Using a Stacked Ensemble Method. *Front Microbiol* 2018;9:2571.

Zhang, Y., Zhang, Y., Xiong, Y.*, 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).

Zhao, Z., Hu, Y., Hu, Y.*, et al.* Features and algorithms: facilitating investigation of secreted effectors in Gram-negative bacteria. *Trends Microbiol* 2023.

Zou, L., Nan, C. and Hu, F. Accurate prediction of bacterial type IV secreted effectors using amino acid composition and PSSM profiles. *Bioinformatics* 2013;29(24):3135-3142.

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: the area under the receiver operating characteristic curve, AUPRC: the precision recall rate curve, [a]: re-train 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-XL-UniRef50	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: the area under the receiver operating characteristic curve, AUPRC: the precision recall rate curve, [a]: re-train the model

761 **Supplementary data**

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

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

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

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

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

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

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

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

779 **Supplementary Table S4.** The distribution of effector domain families.

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

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

783 proteome.

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

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

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