frontiers in Microbiology

ORIGINAL RESEARCH
published: 08 February 2022
doi: 10.3389/fmicb.2021.813094

# T1SEstacker: A Tri-Layer Stacking Model Effectively Predicts Bacterial Type 1 Secreted Proteins Based on C-Terminal Non-repeats-in-Toxin-Motif Sequence Features

*Zewei Chen[1†], Ziyi Zhao[1†], Xinjie Hui[2†], Junya Zhang[1], Yixue Hu[1], Runhong Chen[1], Xuxia Cai[1], Yueming Hu[1] and Yejun Wang[1*]*

[1] Youth Innovation Team of Medical Bioinformatics, Shenzhen University Health Science Center, Shenzhen, China,
[2] Department of Respiratory Medicine, Xuanwu Hospital, Capital Medical University, Beijing, China

OPEN ACCESS

**Edited by:**
*Mohammad-Hossein Sarrafzadeh, University of Tehran, Iran*

**Reviewed by:**
*Timothy James Wells, The University of Queensland, Australia*
*Jack Christopher Leo, Nottingham Trent University, United Kingdom*

***Correspondence:**
*Yejun Wang*
*wangyj@szu.edu.cn*

*† These authors have contributed equally to this work*

**Specialty section:**
*This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology*

**Received:** *11 November 2021*
**Accepted:** *20 December 2021*
**Published:** *08 February 2022*

**Citation:**
*Chen Z, Zhao Z, Hui X, Zhang J, Hu Y, Chen R, Cai X, Hu Y and Wang Y (2022) T1SEstacker: A Tri-Layer Stacking Model Effectively Predicts Bacterial Type 1 Secreted Proteins Based on C-Terminal Non-repeats-in-Toxin-Motif Sequence Features. Front. Microbiol. 12:813094. doi: 10.3389/fmicb.2021.813094*

Type 1 secretion systems play important roles in pathogenicity of Gram-negative bacteria. However, the substrate secretion mechanism remains largely unknown. In this research, we observed the sequence features of repeats-in-toxin (RTX) proteins, a major class of type 1 secreted effectors (T1SEs). We found striking non-RTX-motif amino acid composition patterns at the C termini, most typically exemplified by the enriched "[FLI][VAI]" at the most C-terminal two positions. Machine-learning models, including deep-learning ones, were trained using these sequence-based non-RTX-motif features and further combined into a tri-layer stacking model, T1SEstacker, which predicted the RTX proteins accurately, with a fivefold cross-validated sensitivity of ∼0.89 at the specificity of ∼0.94. Besides substrates with RTX motifs, T1SEstacker can also well distinguish non-RTX-motif T1SEs, further suggesting their potential existence of common secretion signals. T1SEstacker was applied to predict T1SEs from the genomes of representative *Salmonella* strains, and we found that both the number and composition of T1SEs varied among strains. The number of T1SEs is estimated to reach 100 or more in each strain, much larger than what we expected. In summary, we made comprehensive sequence analysis on the type 1 secreted RTX proteins, identified common sequence-based features at the C termini, and developed a stacking model that can predict type 1 secreted proteins accurately.

**Keywords: T1SS, T1SE, RTX proteins, T1SEstacker, prediction, deep learning**

## INTRODUCTION

Type 1 secretion systems (T1SSs) are uniquely distributed in Gram-negative bacteria, which can secrete various substrate proteins through the two bacterial cell membranes by one step (classical) or two steps (non-classical) into extracellular milieu (Smith et al., 2018b; Spitz et al., 2019). A T1SS is composed by three elementary components—an ATP-binding cassette (ABC) transporter located

in the inner membrane, an outer membrane factor (OMF), and a membrane fusion protein (MFP) connecting the ABC transporter (Kanonenberg et al., 2018). A wide variety of proteins are secreted through this oligomeric secretion channel to play their biological roles. Due to the simple structure of the system, T1SSs have been widely applied in biomedical engineering applications (Schwarz et al., 2012; Ryu et al., 2015; Park et al., 2020).

The T1SS substrates, also called type 1 secreted effectors (T1SEs), have various biological functions, such as host invasion (virulence factors, e.g., HlyA) (Felmlee et al., 1985), enzymolysis (digestion enzymes, e.g., TliA and PrtA) (Son et al., 2012), nutrient acquisition (iron-scavenger proteins, e.g., HasA) (Kanonenberg et al., 2013), and biofilm formation (adhesins, e.g., LapA) (Guo et al., 2019). Since the first T1SS substrate, hemolysin A (HlyA), was discovered in 1979 and its nucleotide sequence was determined in 1985 (Noegel et al., 1979; Felmlee et al., 1985), the structural characteristics and function of T1SEs have been studied extensively. Typical T1SEs can be classified into three classes simply according to their T1SS ABC transporter types: C39-containing ABC transporters with hydrolase activity, C39-like domain (CLD)–containing ABC transporters without hydrolase activity, and a third type of ABC transporters without any additional N-terminal domain (Hui et al., 2021). Class 1 T1SEs, known as the smallest T1SS substrates, normally contain N-terminal leader peptides. The C termini of the leader peptides contain a canonical double glycine ("GG") motif, which can be recognized and cleaved by the C39 domains of corresponding ABC transporters before the mature proteins are secreted through T1SSs (Kanonenberg et al., 2013). Class 2 T1SEs have remarkable repeats-in-toxin (RTX) domains and are also known as RTX proteins. The glycine-rich nanopeptide repeats in RTX domains show a "GGxGxDxUx" consensus sequence motif where "x" is any amino acid and "U" represents a large or hydrophobic amino acid. Class 3 T1SEs may also contain RTX repeat sequences but not necessarily. The last two categories do not contain N-terminal leader peptides, but instead potentially have secretion signal sequences in the C termini. However, the C-terminal signal patterns and function mechanisms remain to be clarified (Kanonenberg et al., 2013). Recently, a group of non-classical T1SEs named RTX adhesins (class 4) have been reported, which are closely related to biofilm formation (Smith et al., 2018b). Different from class 1–3 T1SEs, the RTX adhesins are transported from cytoplasm to extracellular environment by a two-step secretion mechanism, which involves periplasmic intermediates. This subgroup of T1SS machinery is linked with a bacterial transglutaminase-like cysteine proteinase (BTLCP) (Smith et al., 2018b). The RTX adhesion proteins have dialanine BTLCP cleavage sites in the N-terminal retention module that can be recognized and cleaved by the machinery-coupled BTLCP in periplasm before the cross-outer-membrane transport (Boyd et al., 2014; Smith et al., 2018b). The currently known RTX adhesins also have RTX repeats and signal sequences in the C termini (Boyd et al., 2014; Smith et al., 2018b).

Both the function and sequences of T1SEs show large diversity, and until now only ∼100 T1SEs have been validated, which are homology-not-filtered, i.e., being redundant with high sequence homology, and therefore could represent fewer independent validated effectors[1]. Bioinformatic strategies have also been tried to predict novel T1SEs, but mainly focused on the RTX proteins with the consensus RTX motifs (Linhartova et al., 2010; Luo et al., 2015). For instance, Linhartova et al. (2010) combined pattern searching, Hidden Markov Model profiles, and the RPS-BLAST tool finding conversed domains to predict 1,024 candidate RTX proteins from 840 bacterial genomes, as comprised the most comprehensive list of RTX T1SE candidates. Luo et al. (2015) made the first attempt to develop a machine-learning model to predict RTX proteins. The random forest–based model learned amino acid sequence–derived features extracted from the full-length and C-terminal sequences of T1SE candidates predicted by Luo et al. (2015). Regretfully, neither a software tool nor a web server was provided for users to implement the method. Besides, both the homology-based and machine-learning methods completely focused on the RTX proteins and the conserved RTX motif was placed with a large weight. The methods are hardly generalized to find more novel T1SEs without RTX motif features.

By careful sequence pattern analysis, previously, we identified the position-specific amino acid composition (Aac), secondary structure element (Sse), and solvent accessibility (Acc) features of type 3 secreted effectors within their N termini and the various Aac, Sse, and Acc profiles of type 4 secreted effectors within their C termini (Wang et al., 2011, 2014). Given the evidence about the potential C-terminal secretion signals of T1SEs (Koronakis et al., 1989; Masure et al., 1990; Zhang et al., 1995; Delepelaire, 2004; Holland et al., 2005; Thomas et al., 2014), in this research, we comprehensively observed the amino acid sequence patterns, especially non-RTX-motif features within the C termini of RTX proteins, and also the Sse and Acc property. Furthermore, we developed machine-learning models to learn the newly observed sequence-derived features and predicted T1SEs with or without typical RTX motifs. Deep learning models and ensemblers have recently been widely used to predict bacterial secretion signals and achieved good performance (Wang et al., 2018, 2019; Almagro Armenteros et al., 2019; Xue et al., 2019; Hui et al., 2020). We also tested Deep Neural Network models and integrated them and others within a stacked model to improve the prediction performance.

# MATERIALS AND METHODS

## Datasets

Bacterial RTX proteins were collected from Linhartova et al. (2010). In total, there were 1,024 RTX proteins predicted from 840 bacterial genomes (Linhartova et al., 2010). CD-HIT was used to detect homology among the RTX proteins, while 30% was considered as the similarity cutoff and only one representative was retained if there were multiple proteins showing sequence similarity above the cutoff (Li and Godzik, 2006). Proteins were also sampled randomly from the whole proteomes derived from various bacterial genome sequences. The known T1SEs, RTX proteins, and their homologs with >30% blastp similarity were

[1]http://61.160.194.165/TxSEdb

removed, and a homology filtering strategy similar to that applied for RTX proteins were used to identify the non-redundant non-RTX proteins. In total, 512 non-redundant RTX proteins were retained, which were considered as the positive dataset (p). A total of 2,000 proteins were also randomly selected from the processed non-RTX proteins, and three groups, each with 512 proteins, were further picked out to match the number and general length distribution of the RTX proteins, forming the negative datasets ($n1 \sim n3$). The $p$ and $n1$ were used as the main observation datasets. A fivefold cross-validation strategy was used for training the machine-learning prediction models, for which both the positive and negative datasets were split into five subsets of equal size of protein sequences, with four of them being served as training datasets and the remaining one as testing datasets in each round of model analysis. Experimentally validated T1SEs were also annotated manually from literature. These proteins could be RTX or other type of proteins with experimental evidence to be transported through T1SSs. All the datasets were publically available together with the standalone T1SEstacker package (see Section "Software Availability"; see Text Footnote 1).

Once the datasets were collected and annotated, the sequence-based features were analyzed with in-house scripts. The secondary structure and solvent accessibility were predicted with SSpro/ACCpro5, with three elements encoded for secondary structure ("H" for helix, "E" for strand, and "C" for coil) and two elements for accessibility ("B" for buried and "E" for exposed) (Magnan and Baldi, 2014).

## Sequential and Position-Specific Amino Acid Composition Feature–Based Non-deep-Learning Models

The number and position distribution of RTX motifs featured as "GGxGxD" was observed within the RTX and non-RTX proteins. Sequential Aac, continuous and 1 or 2 amino acid interrupted bi-residue amino acid composition (bAac) features were extracted from the C-terminal 20- or 60-residue fragments of both the positive and negative datasets, respectively, observed, and compared. The features were used for training Random Forest (RF), Support Vector Machine (SVM), and Naive Bayesian (NB) models, with R packages of "randomForest," "e1071," and the "e1071" method "naiveBayes," respectively[2]. The neighbor-position Aac conditional constraint features in the C termini were learned in Markov models (Wang et al., 2013). Bi-profile Bayesian position-specific Aac features were extracted and trained with SVM models (Wang et al., 2011). For the SVM models, four kernels ("linear," "polynomial," "sigmoid," and "radial") were tested and the corresponding parameters, e.g., *gamma* and/or *cost*, were optimized using a 10-fold cross-validation grid search strategy within each training dataset. For the other models, the features were also extracted based on each training dataset. The details about the models and the optimized parameters refer to the website of T1SEstacker (see Section "Software Availability").

[2]https://www.r-project.org/

## Deep Learning Models

Deep learning models were trained with the Aac features of RTX proteins within the C-terminal 20 (C20) and 60 amino acid positions (C60). Each position was represented by a 20-element feature vector describing the composition of amino acids. An $m \times 20$ $L$ matrix was built to represent the original Aac features of training datasets, where $m$ is the number of training proteins and $L$ is 20 or 60 for C20 or C60 models, respectively. Fully connected Deep Neural Network (DNN), Self Attention (SelfAttention), and models with Long-Short Term Memory (LSTM) cells (RNN) were trained and tested with a fivefold cross-validation strategy. The details about the models and the optimized parameters refer to the website of T1SEstacker (see Section "Software Availability").

## A Stacked Model Featured by the Prediction Results of Individual Models

To achieve better prediction performance, we proposed a new stacking scheme to integrate prediction results of individual models (**Figure 1**). A primary stacked model was built for each original fivefold training dataset and its based individual models. For each original fivefold testing dataset, an embedded fivefold cross-validation was adopted to evaluate the performance of stacked models. The prediction result of each-fold best-trained model of individual algorithms on each protein of the corresponding testing dataset was based, and encoded as 1 (RTX) or 0 (non-RTX) according to the model-specific optimized cutoff score. Each protein within an embedded fivefold training dataset was represented as a feature vector of "0" and "1," and an $m' \times n$ matrix was generated for the whole training dataset, where $m'$ is the protein number of the embedded training dataset and $n$ is the number of individual machine-learning models. SVM models with "linear" kernels were trained and the parameters (costs) were optimized with a 10-fold cross-validation grid-searching strategy.

A voting strategy was used to integrate the five primary stacked models, with the same weight assigned for each model.

## Performance Evaluation of the Individual and Stacked Models

Sensitivity (Sn), specificity (Sp), accuracy (ACC), the area under the curve of receiver operating characteristic (rocAUC), and Matthews correlation coefficient (MCC) were defined and used as measures to assess the performance of models based on a fivefold cross-validation strategy.

$$Sn = TP/(TP + FN)$$

$$Sp = TN/(TN + FP)$$

$$ACC = (TP + TN)/(TP + FN + TN + FP)$$

$$MCC = [(TP \times TN) - (FN \times FP)]/\mathrm{sqrt}[(TP + FN) \times (TN + FP) \times (TP + FP) \times (FN + FN)].$$

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

## Statistics

Individual amino acids were counted within C-terminal 20, 60, or 110-aa fragments, and Mann–Whitney tests were performed to compare their distribution between RTX and non-RTX proteins,

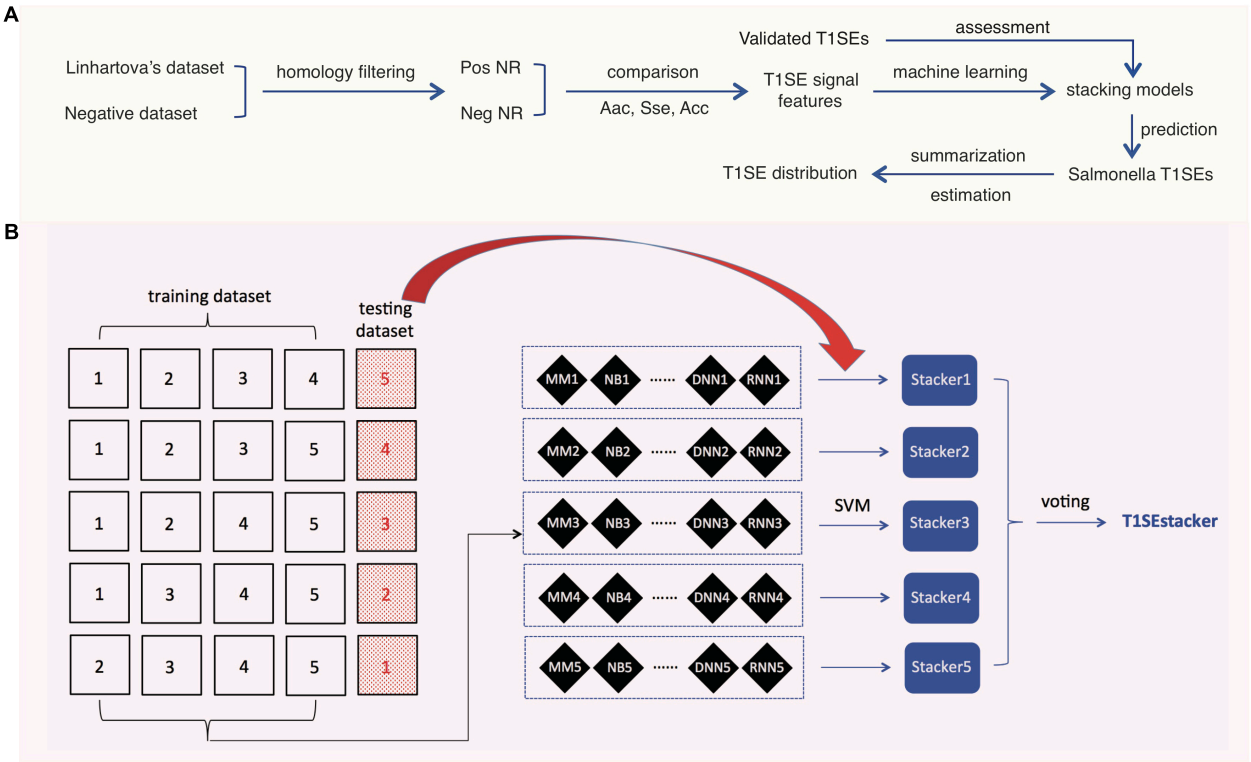


**FIGURE 1 |** Research design and scheme of the tri-layer staking model T1SEstacker. **(A)** Research design. Pos, positive; Neg, negative; NR, non-redundant. **(B)** Tri-layer stacking model. The original positive and negative datasets were fivefold divided and four of them were used as training datasets for training different machine-learning models, e.g., MM, NB, etc. The remaining sub-divided dataset was used as the corresponding testing dataset. For each round of cross-validation, the prediction results of the individual models for the testing dataset were used for training and testing the primary stacker models using SVM with an embedded fivefold cross-validation strategy. Prediction results of the primary stackers were further stacked in T1SEstacker with a voting strategy.

followed by Bonferroni corrections. For two continuous or non-continuous amino acids (bi-AAs), the composition was also compared between the C termini of RTX and non-RTX proteins using the same statistical methods. Another balanced rate comparison method, EBT, was also adopted to compare the C-terminal occurrence of bi-AAs between the two classes of proteins (Hui et al., 2017). The alpha levels for all tests were preset as 0.05.

## Software Availability

T1SEstackers and its modules were developed with Python, Perl, and R. The packages and user manual can be downloaded freely via the link, http://www.szu-bioinf.org/tools/T1SEstacker. A web server was also initiated to make internet-based prediction service: http://www.szu-bioinf.org/T1SEstacker.

## *Salmonella* Genomes

In total, 26 representative strains were included, which covered the known *Salmonella* phylogenetic groups. N268_08, NCTC12419, and RKS3044 belong to *Salmonella bongori*; RKS2983 and RSK2980 belong to *Salmonella enterica* subsp. *arizonae*; ATCC_BAA_1581 and RKS3027 belong to *Salmonella enterica* subsp. *houtenae*; 2439-64 and RKS3013 belong to *Salmonella enterica* subsp. *vii*; 11_01853, 11_01854, 11_01855, and RKS2978 belong to *S. enterica* subsp. *diarizonae*; RKS2986 and ST114 belong to *Salmonella enterica* subsp. *salamae*; 1121 and RKS3057 belong to *Salmonella enterica* subsp. *indica*; while P12519, 287/91, ATCC9150, SPB7, RKS4594, ATCC9120, CT18, 14028S, and LT2 represent various serovars of *Salmonella enterica* subsp. *enterica*. The genome and genome-encoding proteome were downloaded from NCBI genome database[3]. T1SEstacker was applied to predict the T1SE candidates with default settings.

# RESULTS

## Research Design

The major obstacles for training machine-learning models in prediction of bacterial T1SEs include (1) the limited number of experimentally validated positive proteins and (2) the large sequence diversity of T1SE groups. Comprehensive literature searching and manual annotation only curated 99 validated T1SEs, and only 49 were retained after a strict homology-filtering process, which were distributed in all the four major T1SE groups (see Text Footnote 1). To better analyze the likely novel sequential features that could facilitate understanding the mechanisms of

[3]https://www.ncbi.nlm.nih.gov/genome

type 1 secretion and prediction of new T1SEs, and as performed by others previously (Luo et al., 2015), we took the larger-scale RTX T1SE candidates identified by Linhartova et al. (2010) as training data for analysis of features other than RTX motifs and building models to predict novel T1SEs.

After removing the homologs, the remaining non-redundant T1SEs and paired non-T1SEs were compared for their sequential and position-specific Aac, Sse, and Acc features, especially non-RTX motif features (**Figure 1A**). With the sequence-based features, a stacking model was developed to predict T1SEs (**Figure 1B**). Representative strains of *Salmonella* phylogenetic branches were predicted with the newly developed model, and the possible number and distribution of candidate T1SEs were evaluated (**Figure 1A**).

## Distance Distribution of Repeats-in-Toxin Motifs to the C Termini in Repeats-in-Toxin Proteins

The 512 non-redundant RTX proteins show a length distribution from 70 to 36,805 amino acids, with a median of 1,112 residues and 7 super-long proteins with larger than 10,000 amino acids (**Figure 2A**). In addition, 494 from the 512 positive proteins could be found with at least one RTX motif within each protein sequence (**Figure 2B**). As a control, only 13 from the total 2,341 non-redundant negative proteins contained RTX motifs, which were filtered for further comparative or model-training analysis. The most C-terminal residue of each most C-terminal RTX motif shows a distance of 1–21,948 amino acids to the C terminus of the corresponding full-length protein, with a median of 110 amino acids (**Figure 2C**). Fewer than 9% of the C-terminal RTX motifs have a distance of smaller than 60 amino acids from the protein C termini, and only ∼5% are shorter than 20 amino acids (**Figure 2D**).

## Sequential Amino Acid Composition Features Buried in the C Termini of Repeats-in-Toxin Proteins

We compared the composition of individual amino acids (Aac) and two continuous or non-continuous amino acids (bAac) among the C termini of RTX proteins since there were possibly atypical secretion signals (Boyd et al., 2014; Smith et al., 2018a). To avoid the possible misinterpretation caused by RTX motifs, we mainly observed the Aac and bAac profiles within the C-terminal 20 (C20) and C-terminal 60 (C60) residues (**Supplementary Dataset 1**). Within C20, most individual amino acids show different compositions between the positive and negative proteins, with aspartic acid (D), leucine (L), threonine (T), valine (V), isoleucine (I), and phenylalanine (F) being most typically enriched and arginine (R), lysine (K), glutamic acid (E), and proline (P) being most strikingly depleted in RTX proteins (**Figure 3A**; Mann–Whitney *U*-tests with Bonferroni correction, $p < 0.001$). Glycine (G) was not different between the two types of proteins (**Figure 3A**; $p = 1$). When the observed length increases to C-terminal 60-aa, most of the featured residues identified from shorter fragments remain different between groups for the composition, whereas some others start to show difference or no difference, e.g., "G" being enriched in RTX proteins and "L," "V," and "I" becoming no difference (**Figure 3A**). The enrichment of "G" in RTX C60 fragments is not likely due to the increasing occurrence of RTX motifs, which is enriched with "G," since the RTX motifs are lowly represented and the RTX motif featured "GG" is either not strikingly higher in the C60 fragments of RTX proteins (**Figures 2D**, **3B**). For the C-terminal 110-aa fragments, the amino acid species with significantly different composition and the amplitude of difference further increase (**Figure 3A**). It cannot be excluded that the increased number of RTX motifs leads to the most striking composition amplitude change of "D" and "G," especially in C110, for which half of the sequences contained the RTX motifs. However, the "L" composition change is interesting, which shows higher composition in C20, no difference in C60, and lower composition in C110 of RTX proteins (**Figure 3A**).

The continuous and interrupted bAac profile also shows difference in C termini between RTX and non-RTX proteins. For example, "D[FL]," "TL/LT," "AxD," "Tx[LT]," "TxxD," and "Dxx[FI]" most frequently occur, whereas "R[RK]," "K[KA]," "AxR," "Rx[RL]," "AxxR," "Kxx[KE]," and "Rxx[QR]" are most strikingly depleted in the C-terminal 20-aa fragments of RTX proteins in contrast to non-RTX proteins (**Figures 3B–D**; Mann–Whitney *U*-tests with Bonferroni correction, $p < 0.001$; EBT_$p < 0.001$). As the observed C-terminal length increases (to 60 aa), the general bAac profile difference between RTX and non-RTX proteins remains or becomes more typical, with only a few changes. The main changes involve the reduced "L" and increased "G" combinations in the RTX C60 enriched list (**Figures 3B–D**). It is noted that either "GG" or "GxG," which is supposed to be highly represented by RTX motifs, does not show the most significant different composition or occurrence in C60 between RTX and non-RTX proteins, suggesting that the observed different "G"-combination compositions are not due to the increased percent of RTX motifs in C60 of RTX proteins. In C110, however, the composition shows striking difference for both "GG" and "GxG" between RTX and non-RTX proteins (**Supplementary Dataset 1**).

Other independent non-RTX proteins datasets are also paired and the profile difference for Aac and bAac in C termini between RTX and non-RTX proteins shows large consistence.

## Position-Specific Amino Acid Composition Features Buried in the C Termini of Repeats-in-Toxin Proteins

The C-terminal position-specific amino acid composition (psAac) profiles were also compared between RTX and non-RTX proteins. Generally, RTX proteins show much larger amino acid composition preference (**Figure 4A**). C20 and C21-60 in RTX proteins also show different preference profiles. C20 shows apparent preference for non-polar "L" and "A" while C21-60 more prefers polar "G" (**Figure 4A**). "D," "S," and "T" are preferred in both C20 and C21-60 of RTX proteins. The results are consistent with and explain the observations on sequential Aac and bAac in C termini of RTX and non-RTX proteins. Remarkably, the C-terminal endmost two positions in RTX proteins show the

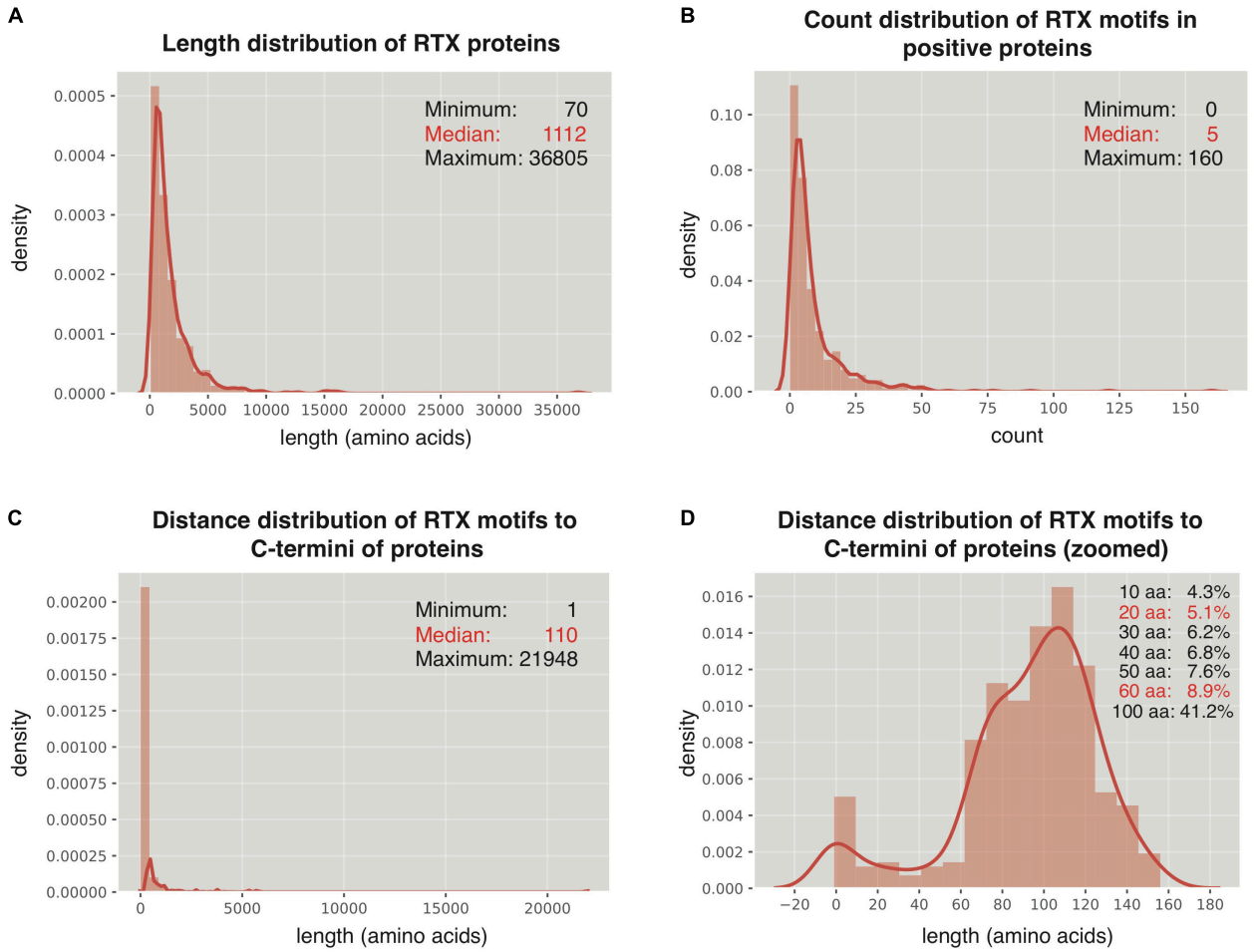


**FIGURE 2 |** Distribution of RTX motifs in RTX proteins. **(A)** Length distribution of the RTX proteins. **(B)** Count distribution of RTX motifs in the RTX proteins. **(C,D)** Distance distribution of RTX motifs to the C-terminal ends of the RTX proteins. The accumulated probabilities of the proteins with the RTX motif distance of ≤10 ~ 60-aa and 100-aa from the C termini are shown in **(D)**.

most typical psAac bias, with a pattern of non-polar hydrophobic "[FLI][VAI]" motif (**Figure 4A**).

The psAac profile of C termini of RTX proteins and the difference between them and non-RTX proteins were confirmed with other, paired, independent negative datasets (**Supplementary Figure 1**). We also compared the psAac profile of N termini of RTX and non-RTX proteins (**Supplementary Figure 2**). There was a difference, but not as typical as that observed within the C termini. Moreover, until now there is no evidence suggesting the existence of type 1 secretion signals within N termini of the substrate proteins. Therefore, the N termini were not further studied in this study.

## Enrichment of β-Strands and Depletion of α-Helices Within the C Termini of Repeats-in-Toxin Proteins

An apparent difference between the C termini of RTX T1SEs and non-T1SEs was the depletion of α-helices or enrichment of β-strands and coiled coils, no matter in C20 or C60 (**Figure 4B**). The solvent accessibility was not different between the RTX and non-RTX proteins within the C termini (data not shown). The different forms of secondary structure are likely related with the composition preference of residues. For instance, both polar "G" and non-polar "A" are enriched in β-strands, while "F" and "I" are not for beneficial for maintenance of the stability of α-helices (**Figure 4A**). It remains to be clarified whether the residue composition and structure features are associated with specific recognition of the proteins for specific type 1 secretion.

## C-Terminal Non-repeats-in-Toxin Motif Features Accurately Classify Repeats-in-Toxin From Non-repeats-in-Toxin Proteins

A list of machine-learning models were trained to learn the sequence-based non-RTX motif features buried within the C termini of RTX proteins, including NB, RF, and SVM models learning sequential Aac and bAac features, MM models using adjacent amino acid dependent Aac features, and SVM models

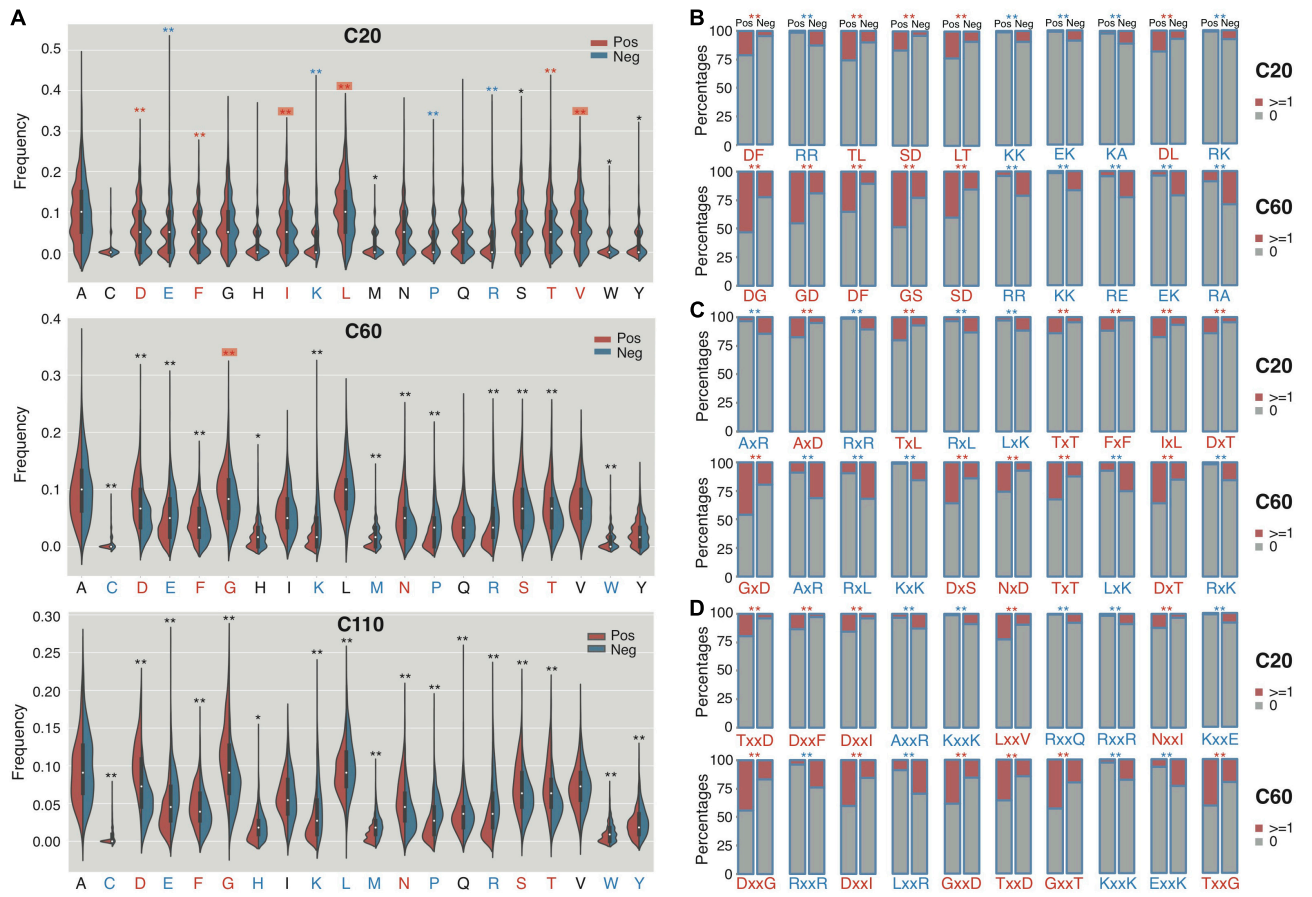


**FIGURE 3 |** The composition profile difference between the C termini of RTX and non-RTX proteins for single AAs, continuous and interrupted bi-AAs. **(A)** Single AAs. Bonferroni-corrected Mann–Whitney tests were performed. **(B)** Continuous bi-AAs. **(C)** One amino acid interrupted bi-AAs. **(D)** Two amino acids interrupted bi-AAs. EBT tests were performed for the bi-AAs. Most significantly enriched or depleted single AAs or bi-AAs in RTX C termini are shown in red and blue, respectively. The single AAs with biased composition changing among the C termini of different lengths are shown with red background. $^{**}p < 0.0001$, $^{*}p < 0.05$.

analyzing position-specific Aac features (**Table 1**). Moreover, five types of DL models were trained, with three among them of best performance retained (DNN, Attention, and RNN), which also learned the C-terminal Aac features of RTX proteins (**Table 1**). Secondary structure features were not learned in the models since they are not stable, which were predicted with varied accuracy using different software tools.

All the models showed certain ability to classify RTX proteins from the non-RTX ones correctly only based on the Aac features within C-terminal 20-aa peptide fragments of known RTX proteins (**Table 2** and **Figure 5A**). RNN, MM, RF, and seqSVM showed best prediction performance with the same average rocAUC of 0.88, while BPBAac and DNN appeared poorest with a rocAUC of 0.85 (**Table 2** and **Figure 4A**). C60 models outperformed C20 ones obviously, and MM, RF, and seqSVM remained the best-performed models, reaching a rocAUC of 0.94 (**Table 2** and **Figure 5B**).

Taken together, the results demonstrate that the C termini of RTX proteins contain non-RTX Aac signals, which can be used to recognize RTX proteins accurately. The signals are likely distributed along the C-terminal 60-aa positions.

## A Stacked Model Shows Striking Performance Improvement in Prediction of Repeats-in-Toxin Types of Type 1 Secreted Effectors

To achieve better performance, we designed a tri-layer stacking model, which integrates the prediction results of individual models learning sequence-based features, to classify RTX and non-RTX proteins (**Figure 1**). The primary SVM-based stacked models (pT1SEstacker) trained with the prediction results of original fivefold cross-validated testing datasets showed better performance than individual models for both C20 and especially C60, with average rocAUC of 0.85 and 0.95, respectively (**Table 2** and **Figure 5C**). The prediction results of the primary stacked models based on cross-validated testing datasets were assembled in the final model (T1SEstacker) with a voting strategy. It is noted that, with an independent dataset, which will be explained in the next section, the voting-based tri-layer stacker T1SEstacker generally balanced the effect of individual pT1SEstacker models and always achieved slightly better performance when voting cutoff was set as 0.6 (**Figure 6A** and **Supplementary Figure 3**).

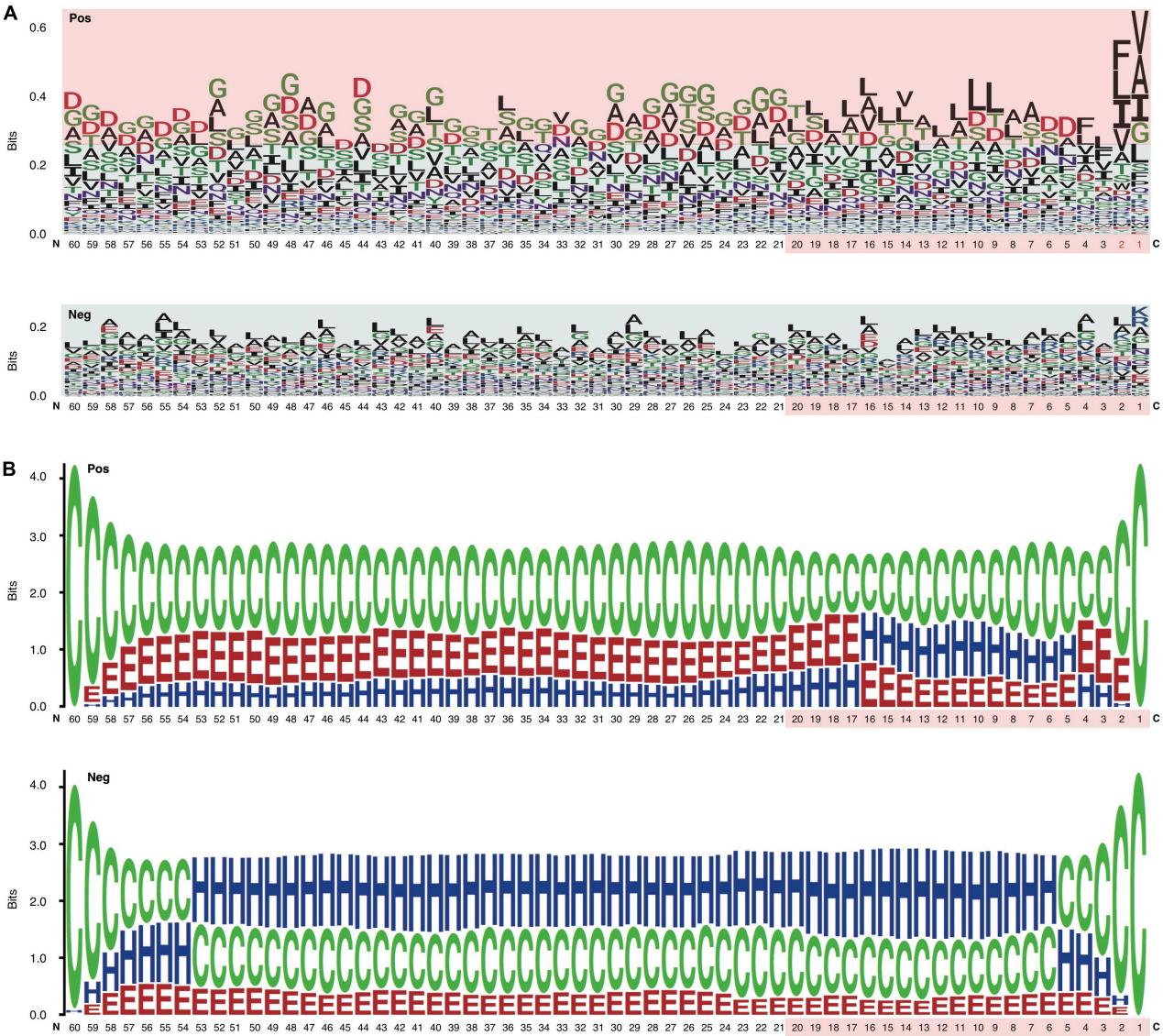


**FIGURE 4 |** Position-specific Aac and Sse profile difference between the C termini of RTX and non-RTX proteins. **(A)** Position-specific Aac profile of RTX and non-RTX proteins at C-terminal 60 positions. The strikingly specific bias Aac preference profile of RTX proteins with C termini and the C-terminal 20 positions of both RTX and non-RTX proteins are shown with pink background. The endmost C-terminal two positions with most typical sequence patterns are shown in red. **(B)** Position-specific Sse profile of RTX and non-RTX proteins at C-terminal 60 positions.

## T1SEstacker Can Recognize the Common Secretion Signals Among Different Types of Type 1 Secreted Effectors

We curated experimentally validated T1SEs and applied the RTX protein prediction models to identify them. It should be noted that none of the C60 or C20 of the verified T1SEs contained any RTX motif. Both T1SEstacker_C20 and T1SEstacker_C60 could well predict the T1SEs (**Figure 6A** and **Supplementary Figure 3**). The recalling rate of T1SEstacker_C20 and T1SEstacker_C60 reached 77 and 81%, respectively (**Figure 6B**). As a control, we used an independent negative dataset, and the specificity of T1SEstacker_C20 and T1SEstacker_C60 was 89 and 96%, respectively (**Figure 6B**).

Among the validated T1SEs, 25% (25/99) do not contain any putative RTX motif along the full-length protein sequences (**Figure 6C** and **Supplementary Dataset 2**). Interestingly, T1SEstacker_C60 correctly recalled 52% (13/25) of the non-RTX-motif T1SEs (**Figure 6C**). Another non-RTX-motif T1SE was predicted to be non-effector by the final T1SEstacker_C60 model, yet it was correctly recalled by two primary models. The recalling rates of non-RTX-motif T1SEs are much higher than the false-positive rates of the negative dataset for both C60 and C20 models (**Figures 6B,C**). Therefore, the results further suggested that C termini of T1SEs, with-RTX-motif or non-RTX-motif

**TABLE 1 |** Models and the optimized parameters.

| Model | Algorithm\| Features |
|---|---|
| MM | Markov model\| Aac conditional on that of the preceding position. |
| RF | Random forest\| AAs, continuous and interrupted bi-AAs with striking sequential composition difference between positive and negative sequences. |
| NB | Naïve Bayes\| same with RF. |
| seqSVM | Support vector machine\| same with RF. |
| BPBAac | Support vector machine\| bi-profile position-specific Aac profiles. |
| DNN | Simple full-connected deep neural network\| Aac profiles. |
| SelfAttention | Softmax deep neural network\| Aac profiles. |
| RNN | Deep neural network with LSTM cells\| Aac profiles. |

type, potentially contained common signals, which can guide the accurate prediction of these proteins.

Most of the validated T1SEs were not well classified into one of the four T1SE classes, except for seven being clear class 4 effectors, including enterotoxigenic *Escherichia coli* CexE (accession: ABM92275.1), *Gallibacterium anatis* GtxA (OBW99045.1), *Pseudomonas fluorescens* LapA (ABA71877.1), *Legionella pneumophila* RtxA (CAH11847.1), *Bordetella bronchiseptica* BrtA (CAE31684.1), *Shewanella oneidensis* BpfA (Q8EIX3.1), and *Vibrio cholera* FrhA (AWB74152.1). Five could be predicted by T1SEstacker_C60 correctly and only two (BpfA and CexE) were not recalled (**Supplementary Dataset 2**). The well-known class 2 effector, *E. coli* HlyA (P08715.1), other two class 2 effectors, *Aggregatibacter actinomycetemcomitans* LtxA (WP_148335754.1) and *Neisseria meningitides* FrpC (AAA99902.1), and one typical class 3 effector, *Serratia marcescens* LipA (Q59933), were all correctly predicted (**Supplementary Dataset 2**). Because the other effectors were not well classified, we did not further compare the prediction performance of T1SEstacker on different T1SE classes. Interestingly, five validated T1SEs were annotated to be bacteriocins, including *Rhizobium leguminosarum* RzcA (AAF36415.1), *Bradyrhizobium elkanii* BAB55900.1, *Xylella fastidiosa* XF2407 (AAF85206.1) and XF2759 (AAF85544.1), *Xanthomonas oryzae* AAW74644.1, and *Agrobacterium tumefaciens* RzcA (AAK89027.2). Four of the bacteriocins were correctly predicted, except for RzcA (**Supplementary Dataset 2**).

## Large Variation of Type 1 Secreted Effectors Composition in *Salmonella* Strains

The chromosomes of 26 representative strains from all *Salmonella* major phylogenetic branches were scanned with T1SEstacker C60 model (**Supplementary Dataset 3**). In each strain, 269 ± 22 T1SE candidates were predicted (**Figure 7A**). With the recalling rate of 0.81 and false-positive rate of 0.04 evaluated previously on the validated T1SE dataset, the real number of T1SEs was estimated to reach 88 to 154, with an average of 123, in *Salmonella* strains (**Figure 7A**). The precision of predicted T1SE candidates was only ∼0.37 (123 × 0.81/269). However, it is difficult to improve the precision by shifting the decision cutoff values or to distinguish the true positives from the false ones. Moreover, most of the real T1SEs were included in the predictions. Therefore, we used the original T1SEstacker predictions to analyze the distribution of T1SE candidates among the *Salmonella* strains.

Despite a relatively stable number of T1SE candidates in different strains, the protein composition varied a lot. The candidates were clustered into 1,004 orthologous families, among which 240 (24%) were strain-specific proteins, 670 (67%) were

**TABLE 2 |** Performance of models.

| Model | SN | SP | ACC | rocAUC | MCC |
|---|---|---|---|---|---|
| MM_C20 | 0.81 ± 0.06 | 0.81 ± 0.05 | 0.81 ± 0.02 | **0.88 ± 0.02** | 0.62 ± 0.04 |
| RF_C20 | 0.79 ± 0.06 | **0.82 ± 0.09** | 0.80 ± 0.06 | **0.88 ± 0.04** | 0.61 ± 0.12 |
| NB_C20 | **0.89 ± 0.05** | 0.69 ± 0.06 | 0.79 ± 0.04 | 0.87 ± 0.03 | 0.59 ± 0.09 |
| seqSVM_C20 | 0.79 ± 0.05 | 0.82 ± 0.06 | **0.81 ± 0.04** | **0.88 ± 0.04** | 0.61 ± 0.09 |
| BPBAac_C20 | 0.72 ± 0.06 | **0.82 ± 0.02** | 0.77 ± 0.03 | 0.85 ± 0.03 | 0.55 ± 0.06 |
| DNN_C20 | 0.77 ± 0.05 | 0.75 ± 0.05 | 0.76 ± 0.04 | 0.85 ± 0.03 | 0.53 ± 0.07 |
| SelfAttention_C20 | 0.80 ± 0.03 | 0.80 ± 0.05 | 0.80 ± 0.04 | 0.87 ± 0.02 | 0.60 ± 0.07 |
| RNN_C20 | 0.82 ± 0.05 | 0.80 ± 0.05 | **0.81 ± 0.04** | **0.88 ± 0.04** | **0.63 ± 0.07** |
| pT1SEstacker_C20 | 0.83 ± 0.06 | **0.85 ± 0.04** | **0.84 ± 0.04** | **0.88 ± 0.06** | **0.69 ± 0.09** |
| MM_C60 | 0.86 ± 0.06 | **0.93 ± 0.04** | **0.89 ± 0.02** | **0.94 ± 0.02** | **0.79 ± 0.02** |
| RF_C60 | 0.85 ± 0.06 | 0.90 ± 0.02 | 0.88 ± 0.03 | **0.94 ± 0.03** | 0.76 ± 0.05 |
| NB_C60 | 0.86 ± 0.03 | 0.83 ± 0.05 | 0.84 ± 0.04 | 0.92 ± 0.02 | 0.69 ± 0.07 |
| seqSVM_C60 | 0.84 ± 0.08 | 0.92 ± 0.02 | 0.88 ± 0.04 | **0.94 ± 0.02** | 0.77 ± 0.07 |
| BPBAac_C60 | 0.84 ± 0.04 | 0.89 ± 0.02 | 0.87 ± 0.02 | 0.93 ± 0.01 | 0.73 ± 0.03 |
| DNN_C60 | 0.87 ± 0.06 | 0.85 ± 0.04 | 0.86 ± 0.02 | 0.92 ± 0.02 | 0.72 ± 0.04 |
| SelfAttention_C60 | **0.89 ± 0.02** | 0.89 ± 0.03 | **0.89 ± 0.02** | 0.93 ± 0.01 | 0.78 ± 0.03 |
| RNN_C60 | 0.85 ± 0.07 | 0.90 ± 0.07 | 0.87 ± 0.03 | 0.93 ± 0.03 | 0.75 ± 0.06 |
| pT1SEstacker_C60 | **0.89 ± 0.04** | **0.94 ± 0.02** | **0.91 ± 0.02** | **0.95 ± 0.02** | **0.83 ± 0.03** |

*Sn, Sensitivity; Sp, specificity; ACC, accuracy; rocAUC, the area under the curve of receiver operating characteristic; MCC, Matthews correlation coefficient. The best performance was highlighted in bold font.*

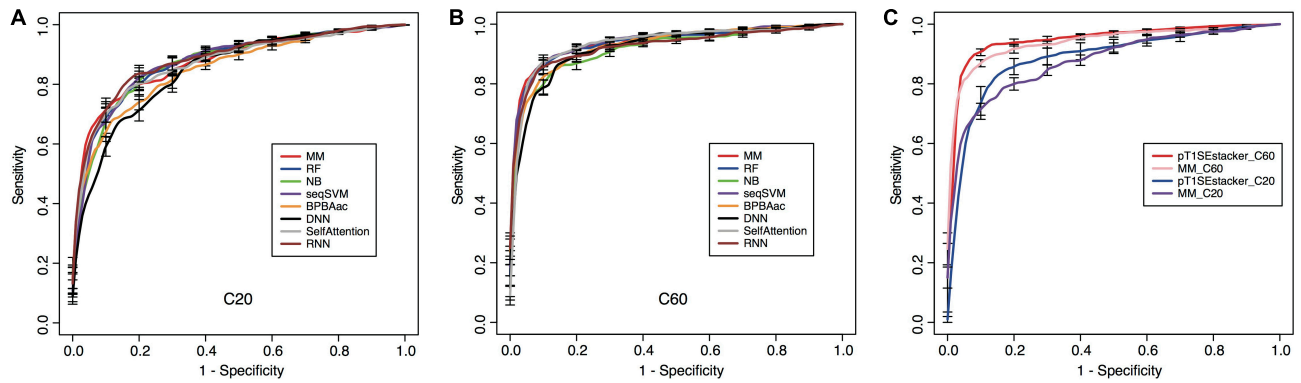


**FIGURE 5 |** Performance of individual and stacking models on prediction of RTX proteins. **(A)** Fivefold cross-validation ROC curves of individual machine-learning models predicting RTX and non-RTX proteins based on the C-terminal 20-aa features. **(B)** Fivefold cross-validation ROC curves of individual machine-learning models predicting RTX and non-RTX proteins based on the C-terminal 60-aa features. **(C)** Performance comparison of the primary stacking models (pT1SEstacker_C20 and pT1SEstacker_C60) and the representative individual machine-learning models (MM_C20 and MM_C60), based on the average fivefold cross-validation results.

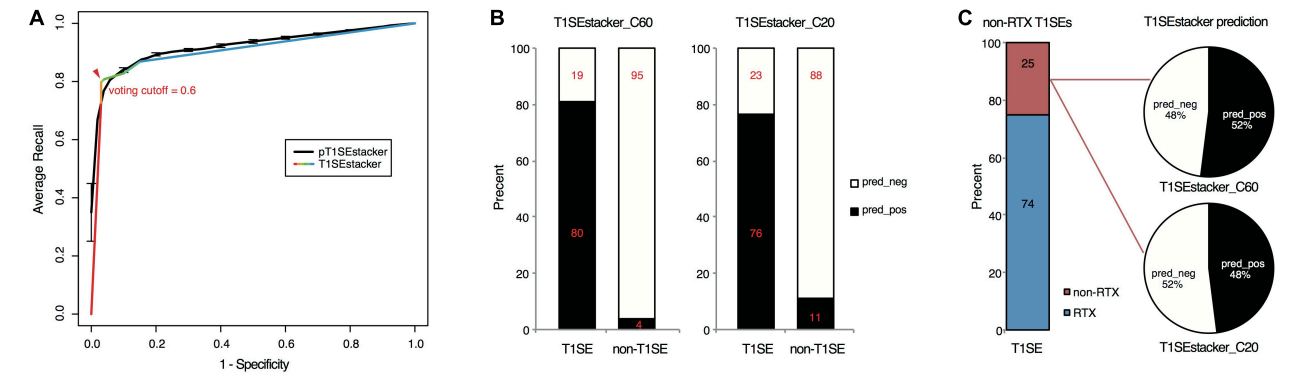


**FIGURE 6 |** Prediction performance of T1SEstacker on experimentally verified T1SEs and the paired non-T1SEs. **(A)** ROC curves of the final T1SEstacker_C60 and the primary pT1SEstacker_C60 models on prediction of experimentally verified T1SEs and the paired non-T1SEs. The number for both T1SEs and non-T1SEs was 99. The best-optimized cutoff for the decision of T1SEstacker is indicated with a red arrow. **(B)** The recalling T1SEs and false-positive T1SE predictions of T1SEstacker_C60 and T1SEstacker_C20. **(C)** The T1SE proteins without putative RTX motifs (non-RTX T1SEs) and the correctly predicted percent of T1SEstacker on them.

present in fewer than half of the strains, and only 179 (18%) were distributed in the core genome of the *Salmonella* strains (**Figure 7B** and **Supplementary Dataset 3**). For the core-genome hits, only 49% (88/179) were recognized as T1SEs in all the strains, and 31% (55/179) of the families were predicted as T1SEs only in fewer than half of the strains (**Figure 7C** and **Supplementary Dataset 3**). The results suggested that there is a large variety for the composition of T1SEs in different bacterial strains, and that a T1SE homolog does not necessarily remain a T1SE since mutations in the C terminus could frequently avoid the recognition of T1SS.

## DISCUSSION

Like other secreted proteins, bacterial type 1 secreted proteins (T1SEs) also play important roles in various infection diseases. Some T1SEs, e.g., bacteriocins, show non-self bacteria-killing activities and therefore have been used for anti-bacteria drug or probiotic development. How many T1SEs are there in each bacterial strain? How diverse is their function? The questions remain unanswered since we are still at the very beginning on understanding the mechanisms of type 1 secretion. Only around 100 T1SEs have been verified by experiments, and many of them contain RTX motifs nearby the C termini of protein sequences. However, not all T1SEs contain RTX motifs, while the proteins with RTX motifs, although more likely to be, are not necessarily T1SEs. Therefore, T1SEs could have other common targeted signals that mediate their specific type 1 secretion. More novel T1SEs could be identified based on these common signals.

Previous studies suggested possible signals within C termini of RTX and non-RTX T1SEs (Delepelaire, 2004; Huang et al., 2010; Wakeel et al., 2011). In this research, we focused on RTX T1SEs, observed the Aac features within their C termini comprehensively, and compared them with the C termini of non-RTX proteins or N termini of the RTX and non-RTX

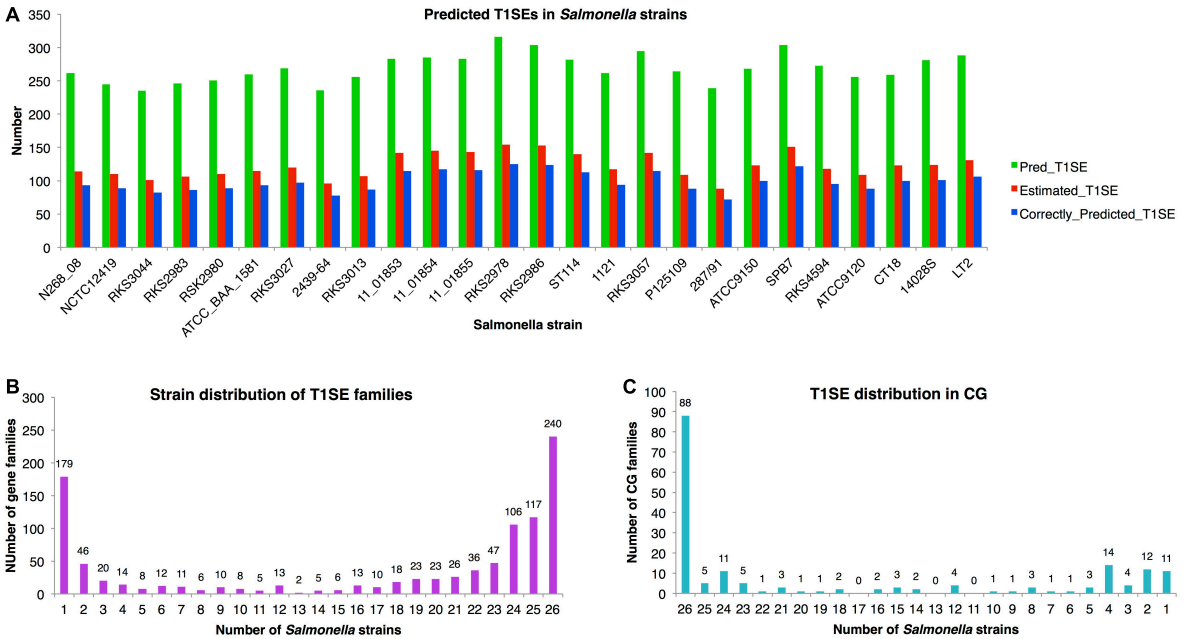


**FIGURE 7 |** Distribution of T1SE candidates predicted from *Salmonella* strains with T1SEstacker. **(A)** Number of predicted T1SEs, estimated real T1SEs, and correctly predicted T1SEs in *Salmonella* strains. **(B)** The orthologous family distribution of T1SEs in *Salmonella* strains. **(C)** Distribution of core-genome T1SE hits in *Salmonella* strains. CG, core genome.

proteins. It was interesting to identify specific Aac preference in C termini of RTX proteins (**Figure 4A**). As control, no apparent difference was found between the N termini of RTX and non-RTX proteins (**Supplementary Figures 1, 2**). The Aac preference profile was not biased by possibly included RTX motifs. On one hand, very few RTX motifs were retained in the observed length of C-terminal sequences (both C20 and C60) (**Figures 2C,D**). On the other hand, the motif-enriched bi-AAs were not as strikingly different as other bi-AAs (**Figure 2B**). Moreover, the real occurrence of some individual AAs or bi-AAs within C termini of RTX proteins, especially C60, e.g., "G" and "D," was much higher than the percentage of proteins with putative RTX motifs within the region. Therefore, such Aac preference could be independent of RTX motif. Alternatively, RTX motifs could also represent the preference, but a more specific and conserved pattern. Besides the enriched Aac, significantly depleted Aac should also be noted, e.g., "E," "K," "R," and "P." In the research, by observing the position-specific Aac profiles, we also identified a typical amino acid composition pattern at the C termini of RTX proteins, with a motif feature of "[FLI][VAI]." Previous studies on known T1SEs found the enrichment of "[LDAVTSIF]" residues in C-terminal signal regions (Delepelaire, 2004; Huang et al., 2010; Wakeel et al., 2011). The features were also evident in our C-terminal sequence–based or position-specific Aac analysis on the T1SEs. HlyA and its homologs in *E. coli*, *Proteus vulgaris*, and *Morganella morganii* were all shown with a preference of "[LS][AV]" at the C termini (Koronakis et al., 1989), consistent with our position-specific Aac observation. We also found that the C termini of RTX proteins preferred β-strands rather than α-helices as in non-RTX proteins (**Figure 4B**). It is intriguing to further investigate whether the unique amino acid composition and secondary structure contribute to the specificity of signal recognition of type 1 secretion.

Machine-learning models based on the C-terminal non-RTX-motif Aac features well predicted RTX proteins from non-RTX proteins (**Figure 5** and **Table 2**). The features within C20 showed certain power, while those buried in C60 showed better distinguishing capability (**Figure 5** and **Table 2**). The C60 models could also accurately recall verified T1SEs at high prediction specificity (larger than 95%) (**Figure 6B**). It should be pointed out again that none of the verified T1SEs contained any RTX motif within C20 or C60 regions. More interestingly, 25 of the verified T1SEs do not contain RTX motif throughout their full-length sequences, and yet 12 and 13 were still predicted by C20 and C60 models, respectively, as positive results (**Figure 6C**). Among the correctly predicted T1SEs, some are bacteriocins and others are not putative RTX proteins. Therefore, the features identified in this study can be used for development of general T1SE prediction models. In future studies and as more non-RTX T1SEs have been identified, the common features can be reanalyzed, with a more balanced training dataset of different types of T1SEs.

We developed a tri-layer stacking model, T1SEstacker, and showed that the stackers generally outperformed the individual machine-learning models (**Table 2** and **Figure 5C**). However, some individual models also showed good performance, e.g., MM, RNN, SelfAttention, and RF, but generally not as good or stable as the stackers, pT1SEstacker (**Table 2** and **Figure 5C**). We made a second round of stacking for the pT1SEstackers trained with sub-divided cross-validated datasets because for pT1SEstackers, we adopted a SVM model to

integrate the prediction results of individual machine-learning models (**Figure 1**). Similar with T1SEstacker that integrates pT1SEstacker results, other ensemblers often use voting strategy (Wang et al., 2019) or linearly weight each individual model (Hui et al., 2020). The parameters, i.e., linear weights for individual models and decision cutoffs for those models, were generally stable and not very sensitive to the sub-divided or full training datasets. However, for pT1SEstacker models, we trained the prediction results of individual models using SVM, and the parameters were pretty sensitive to the training datasets. Therefore, the five pT1SEstackers were each with different optimized parameters. To integrate their respective prediction results, another round of stacking had to be performed. The final model T1SEstacker appeared not apparently better than the pT1SEstacker models. However, once the optimized voting cutoff was selected ($\geq$0.6, 3/5, consensus prediction), the prediction of T1SEstacker always showed best performance, with a compromise of sensitivity and specificity (**Figure 6A** and **Supplementary Figure 3**).

The false-positive rate (FPR) of T1SEstacker_C60 was low and close to 0.04. It is important since many tools predicting bacterial secreted proteins showed a high FPR and the experimental research seldom benefited from the tool (Hui et al., 2020). As an example, we showed the influence of FPR on the final prediction performance, by prediction and estimation of T1SE candidates in *Salmonella* with T1SEstacker (**Figure 7A**). Despite the high specificity (0.96), among the predicted T1SE candidates, majority were false positives, and the precision was only $\sim$0.37 (**Figure 7A**). It is largely because for each genome, most genes are non-T1SEs, and even 1% FPR could generate 50–100 false-positive predictions, for which the number is close to that of true T1SEs. Therefore, it appears essential and urgent to further reduce FPR in predictor development, not merely for T1SE, but also for all types of secreted proteins.

Currently, there is still a lack of computational methods predicting T1SEs (Hui et al., 2021). Although Luo et al. (2015) developed a random forest predictor, the tool or codes were not publically available and therefore a direct comparison could not be performed. An important factor that impedes development of prediction tools for T1SEs is the very limited number of experimentally validated T1SE proteins. Linhartova et al. (2010) and Luo et al. (2015) we in this research used Linhartova's RTX proteins as the positive dataset. In fact, we also used the validated T1SEs to build a similar model, and the performance was only slightly inferior to T1SEstacker but the variance was much larger among the cross-validated replicates. Moreover, the T1SEstacker could accurately predict the novel ones in the validated effector dataset at a high specificity. Therefore, we presented the T1SEstacker based on Linhartova's RTX proteins finally. With T1SEstacker and *Salmonella* strains, we also made estimation on the distribution of T1SEs. Roughly, there could be $\sim$100 T1SEs in each bacterial strain. Therefore, the current T1SEs and function of T1SSs could be largely underestimated and underinvestigated. We also found that the T1SE composition varied a lot among different bacterial strains, suggesting they could exert specific function for better fitting and bacterial survival. Therefore, it is of great significance to identify and investigate the function of T1SEs for both microbiologists and computational biologists.

Very few T1SEs have been validated from *Salmonella* spp., and SiiE represents the most well-known one, a large non-fimbrial adhesin of 600 kDa consisting of 53 repeats of Ig domains, which is encoded in an T1SS operon within *Salmonella* Pathogenicity Island 4 (SPI-4) of *S. enterica* strains (Gerlach et al., 2007; Barlag and Hensel, 2015; Klingl et al., 2020). We found that it was conserved in 19 out of the total 26 *Salmonella* strains (ID: 19CG0093; **Supplementary Dataset 3**). Interestingly, the gene was also detected from *S. bongori* besides all the seven subspecies of *S. enterica*. However, for *S. bongori*, *S. enterica* subsp. *diarizonae*, *indica*, and *enterica*, there were always representative strains missing the gene (**Supplementary Dataset 3**). More efforts should be placed to check whether there is the gene but mis-annotated or the gene has been actually lost. If the gene is lost, it is also interesting to know how its function is complemented in the corresponding strains. In this research, we also provided a list of possible T1SE candidates and their distribution among *Salmonella* spp., which comprise a valuable resource for the research community to further investigate *Salmonella* T1SEs and their function in bacterial pathogenicity.

T1SEstacker is one of the earliest machine-learning models predicting T1SEs. The performance requires further assessment and improvement. In this study, only sequence-derived features of T1SEs were analyzed and learned. Integration of other features such as the genomic context, i.e., proximity of the candidate genes to those encoding secretion components (Glaser et al., 1988; Welch and Pellett, 1988; Welch, 1991), common motifs located in promoters for transcription co-regulation (Mukherjee et al., 2015), physiochemical properties of proteins (Welch et al., 1983), and so on, may be helpful in improving the prediction performance. In addition, T1SS type-specific or species-specific substrate feature analysis and model development could further improve the precision of prediction. Despite the functional relevance, what we have known on T1SSs and T1SEs remains much fewer than unknowns (Alav et al., 2021). It remains a big challenge for computational biologists to make thorough and systematic analysis of T1SE features and develop more effective prediction models.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

## AUTHOR CONTRIBUTIONS

YW conceived and designed the project. ZC, ZZ, and YW developed the models and evaluated the performance. XH developed the web server. JZ, YXH, RC, XC, and YMH annotated the data and performed the analysis. All authors participated in the manuscript preparation.

## FUNDING

This project was supported by the Funds for Medical Bioinformatics Youth Innovation Team of Shenzhen University (406/0000080805) and Natural Science Funds of Shenzhen (JCYJ201607115221141 and JCYJ20190808165205582). ZC and XH were supported by Special Funds for the Cultivation of Guangdong College Students' Scientific and Technological Innovation, Climbing Program (pdjha0427). ZC was supported by a National Undergraduate Training Program of China for Innovation and Entrepreneurship (no. 201910590003).

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.813094/full#supplementary-material

**Supplementary Figure 1 |** Position-specific Aac profile difference between the C termini of repeats-in-toxin (RTX) proteins and three independent groups of non-RTX proteins.

**Supplementary Figure 2 |** Position-specific Aac profile difference between the N termini of RTX proteins and three independent groups of non-RTX proteins.

**Supplementary Figure 3 |** ROC curves of T1SEstacker and pT1SEstacker models on the verified T1SEs and non-T1SEs. **(A)** Performance of T1SEstacker_C60 and pT1SEstacker_C60 on 99 verified T1SEs and 512 non-T1SEs. **(B)** Performance of T1SEstacker_C20 and pT1SEstacker_C20 on 99 verified T1SEs and paired 99 non-T1SEs. **(C)** Performance of T1SEstacker_C20 and pT1SEstacker_C20 on 99 verified T1SEs and 512 non-T1SEs. The best-optimized cutoff for the decision of T1SEstacker models are indicated with red arrows.

**Supplementary Dataset 1 |** Sequential Aac comparison between the C termini of repeats-in-toxin (RTX) and non-RTX proteins.

**Supplementary Dataset 2 |** Repeats-in-toxin motif distribution within the experimentally verified T1SEs and the prediction results of T1SEstacker_C60 and T1SEstacker_C20.

**Supplementary Dataset 3 |** *Salmonella* T1SEs predicted with T1SEstacker.

## REFERENCES

Alav, I., Kobylka, J., Kuth, M. S., Pos, K. M., Picard, M., Blair, J. M. A., et al. (2021). Structure, assembly, and function of tripartite efflux and type 1 secretion systems in gram-negative bacteria. *Chem. Rev.* 121, 5479–5596. doi: 10.1021/acs.chemrev.1c00055

Almagro Armenteros, J. J., Tsirigos, K. D., Sonderby, C. K., Petersen, T. N., Winther, O., Brunak, S., et al. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* 37, 420–423. doi: 10.1038/s41587-019-0036-z

Barlag, B., and Hensel, M. (2015). The giant adhesin SiiE of *Salmonella enterica*. *Molecules* 20, 1134–1150. doi: 10.3390/molecules20011134

Boyd, C. D., Smith, T. J., El-Kirat-Chatel, S., Newell, P. D., Dufrene, Y. F., and O'Toole, G. A. (2014). Structural features of the *Pseudomonas* fluorescens biofilm adhesin LapA required for LapG-dependent cleavage, biofilm formation, and cell surface localization. *J. Bacteriol.* 196, 2775–2788. doi: 10.1128/JB.01629-14

Delepelaire, P. (2004). Type I secretion in gram-negative bacteria. *Biochim. Biophys. Acta* 1694, 149–161.

Felmlee, T., Pellett, S., and Welch, R. A. (1985). Nucleotide sequence of an *Escherichia coli* chromosomal hemolysin. *J. Bacteriol.* 163, 94–105.

Gerlach, R. G., Jäckel, D., Stecher, B., Wagner, C., Lupas, A., Hardt, W. D., et al. (2007). *Salmonella* Pathogenicity Island 4 encodes a giant non-fimbrial adhesin and the cognate type 1 secretion system. *Cell Microbiol.* 9, 1834–1850. doi: 10.1111/j.1462-5822.2007.00919.x

Glaser, P., Sakamoto, H., Bellalou, J., Ullmann, A., and Danchin, A. (1988). Secretion of cyclolysin, the calmodulin-sensitive adenylate cyclase-haemolysin bifunctional protein of Bordetella pertussis. *EMBO J.* 7, 3997–4004.

Guo, S., Vance, T. D. R., Stevens, C. A., Voets, I., and Davies, P. L. (2019). RTX Adhesins are key bacterial surface megaproteins in the formation of biofilms. *Trends Microbiol.* 27, 453–467.

Holland, I. B., Schmitt, L., and Young, J. (2005). Type 1 protein secretion in bacteria, the ABC-transporter dependent pathway. *Mol. Membr. Biol.* 22, 29–39. doi: 10.1080/09687860500042013

Huang, B., Troese, M. J., Howe, D., Ye, S., Sims, J. T., Heinzen, R. A., et al. (2010). Anaplasma phagocy- tophilum APH_0032 is expressed latedur ing infection and localizes to the pathogen-occupied vacuolar membrane. *Microb. Pathog.* 49, 273–284. doi: 10.1016/j.micpath.2010.06.009

Hui, X., Chen, Z., Lin, M., Zhang, J., Hu, Y., Zeng, Y., et al. (2020). T3SEpp: an integrated prediction pipeline for bacterial type III secreted effectors. *mSystems* 5, e00288–20. doi: 10.1128/mSystems.00288-20

Hui, X., Chen, Z., Zhang, J., Lu, M., Cai, X., Deng, Y., et al. (2021). Computational prediction of secreted proteins in gram-negative bacteria. *Comput. Struct. Biotechnol. J.* 19, 1806–1828.

Hui, X., Hu, Y., Sun, M. A., Shu, X., Han, R., Ge, Q., et al. (2017). EBT: a statistic test identifying moderate size of significant features with balanced power and precision for genome-wide rate comparisons. *Bioinformatics* 33, 2631–2641. doi: 10.1093/bioinformatics/btx294

Kanonenberg, K., Schwarz, C. K., and Schmitt, L. (2013). Type I secretion systems - a story of appendices. *Res. Microbiol.* 164, 596–604. doi: 10.1016/j.resmic.2013.03.011

Kanonenberg, K., Spitz, O. I, Erenburg, N., Beer, T., and Schmitt, L. (2018). Type I secretion system-it takes three and a substrate. *FEMS Microbiol. Lett.* 365:fny094. doi: 10.1093/femsle/fny094

Klingl, S., Kordes, S., Schmid, B., Gerlach, R. G., Hensel, M., and Muller, Y. A. (2020). Recombinant protein production and purification of SiiD, SiiE and SiiF - Components of the SPI4-encoded type I secretion system from *Salmonella* Typhimurium. *Protein Expr. Purif.* 172:105632. doi: 10.1016/j.pep.2020.105632

Koronakis, V., Koronakis, E., and Hughes, C. (1989). Isolation and analysis of the C-terminal signal directing export of *Escherichia coli* hemolysin protein across both bacterial membranes. *EMBO J.* 8, 595–605.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Linhartova, I., Bumba, L., Masin, J., Basler, M., Osicka, R., Kamanova, J., et al. (2010). RTX proteins: a highly diverse family secreted by a common mechanism. *FEMS Microbiol. Rev.* 34, 1076–1112. doi: 10.1111/j.1574-6976.2010.00231.x

Luo, J., Li, W., Liu, Z., Guo, Y., Pu, X., and Li, M. (2015). A sequence-based two-level method for the prediction of type I secreted RTX proteins. *Analyst* 140, 3048–3056. doi: 10.1039/c5an00311c

Magnan, C. N., and Baldi, P. (2014). SSpro/ACCpro5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinformatics* 30, 2592–2597. doi: 10.1093/bioinformatics/btu352

Masure, H. R., Au, D. C., Gross, M. K., Donovan, M. G., and Storm, D. R. (1990). Secretion of the Bordetella pertussis ade-nylate cyclase from *Escherichia coli* containing the hemolysin operon. *Biochemistry* 29, 140–145. doi: 10.1021/bi00453a017

Mukherjee, D., Pal, A., Chakravarty, D., and Chakrabarti, P. (2015). Identification of the target DNA sequence and characterization of DNA binding features of HlyU, and suggestion of a redox switch for hlyA expression in the human pathogen *Vibrio cholerae* from in silico studies. *Nucleic Acids Res.* 43, 1407–1417. doi: 10.1093/nar/gku1319

Noegel, A., Rdest, U., Springer, W., and Goebel, W. (1979). Plasmid cistrons controlling synthesis and excretion of the exotoxin alpha-haemolysin of *Escherichia coli*. *Mol. Gen. Genet.* 175, 343–350. doi: 10.1007/BF00397234

Park, Y., Eom, G. T., Oh, J. Y., Park, J. H., Kim, S. C., Song, J. K., et al. (2020). High-level production of bacteriotoxic phospholipase A1 in bacterial host *Pseudomonas* fluorescens via ABC transporter-mediated secretion and inducible expression. *Microorganisms* 8:239. doi: 10.3390/microorganisms8020239

Ryu, J., Lee, U., Park, J., Yoo, D. H., and Ahn, J. H. (2015). A vector system for ABC transporter-mediated secretion and purification of recombinant proteins in *Pseudomonas* species. *Appl. Environ. Microbiol.* 81, 1744–1753. doi: 10.1128/AEM.03514-14

Schwarz, C. K. W., Landsberg, C. D., Lenders, M. H. H., Smits, S. H. J., and Schmitt, L. (2012). Using an E. coli Type 1 secretion system to secrete the mammalian, intracellular protein IFABP in its active form. *J. Biotechnol.* 159, 155–161. doi: 10.1016/j.jbiotec.2012.02.005

Smith, T. J., Sondermann, H., and O'Toole, G. A. (2018b). Type 1 does the two-step: type 1 secretion substrates with a functional periplasmic intermediate. *J. Bacteriol.* 200, e00168–18. doi: 10.1128/JB.00168-18

Smith, T. J., Font, M. E., Kelly, C. M., Sondermann, H., and O'Toole, G. A. (2018a). An N-Terminal retention module anchors the giant adhesin LapA of *Pseudomonas* fluorescens at the cell surface: a novel subfamily of type I secretion systems. *J. Bacteriol.* 200, e00734–17. doi: 10.1128/JB.00734-17

Son, M., Moon, Y., Oh, M. J., Han, S. B., Park, K. H., Kim, J. G., et al. (2012). Lipase and protease double-deletion mutant of *Pseudomonas* fluorescens suitable for extracellular protein production. *Appl. Environ. Microbiol.* 78, 8454–8462. doi: 10.1128/AEM.02476-12

Spitz, O., Erenburg, N. I., Beer, T., Kanonenberg, K. I, Holland, B., and Schmitt, L. (2019). Type I secretion systems-one mechanism for all? *Microbiol. Spectr.* 7:PSIB-0003-2018. doi: 10.1128/microbiolspec.PSIB-0003-2018

Thomas, S. I, Holland, B., and Schmitt, L. (2014). The Type 1 secretion pathway - the hemolysin system and beyond. *Biochim. Biophys. Acta* 1843, 1629–1641. doi: 10.1016/j.bbamcr.2013.09.017

Wakeel, A., den Dulk-Ras, A., Hooykaas, P. J. J., and McBride, J. W. (2011). Ehrlichia chaffeensis tandem repeat proteins and Ank200 are type 1 secretion system substrates related to the repeats-in-toxin exoprotein family. *Front. Cell. Infect. Microboiol.* 1:22. doi: 10.3389/fcimb.2011.00022

Wang, J., Yang, B., An, Y., Marquez-Lago, T., Leier, A., Wilksch, J., et al. (2019). Systematic analysis and prediction of type IV secreted effector proteins by machine learning approaches. *Brief Bioinform.* 20, 931–951. doi: 10.1093/bib/bbx164

Wang, J., Yang, B., Leier, A., Marquez-Lago, T. T., Hayashida, M., Rocker, A., et al. (2018). Bastion6: a bioinformatics approach for accurate prediction of type VI secreted effectors. *Bioinformatics* 34, 2546–2555. doi: 10.1093/bioinformatics/bty155

Wang, Y., Sun, M., Bao, H., and White, A. P. (2013). T3_MM: a Markov model effectively classifies bacterial type III secretion signals. *PLoS One* 8:e58173. doi: 10.1371/journal.pone.0058173

Wang, Y., Wei, X., Bao, H., and Liu, S. L. (2014). Prediction of bacterial type IV secreted effectors by C-terminal features. *BMC Genomics* 15:50. doi: 10.1186/1471-2164-15-50

Wang, Y., Zhang, Q., Sun, M. A., and Guo, D. (2011). High-accuracy prediction of bacterial type III secreted effectors based on position-specific amino acid composition profiles. *Bioinformatics* 27, 777–784. doi: 10.1093/bioinformatics/btr021

Welch, R. A. (1991). Pore-forming cytolysins of gram-negative bacteria. *Mol. Microbiol.* 5, 521–528. doi: 10.1111/j.1365-2958.1991.tb00723.x

Welch, R. A., Hull, R., and Falkow, S. (1983). Molecular cloning and physical characterization of a chromosomal hemolysin from *Escherichia coli*. *Infect. Immun.* 42, 178–186. doi: 10.1128/iai.42.1.178-186.1983

Welch, R. A., and Pellett, S. (1988). Transcriptionalorganization of the *Escherichia coli* hemolysin genes. *J. Bacteriol.* 170, 1622–1630.

Xue, L., Tang, B., Chen, W., and Luo, J. (2019). DeepT3: deep convolutional neural networks accurately identify Gram-negative bacterial type III secreted effectors using the N-terminal sequence. *Bioinformatics* 35, 2051–2057. doi: 10.1093/bioinformatics/bty931

Zhang, F., Yin, Y., Arrowsmith, C. H., and Ling, V. (1995). Secretion and circular dichroism analysis of the C-terminal signal peptides of HlyA and LktA. *Biochemistry* 34, 4193–4201. doi: 10.1021/bi00013a007

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.