frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 31 October 2022
DOI 10.3389/fpls.2022.1024405

OPEN ACCESS

EDITED BY
Dirk Walther,
Max Planck Institute of Molecular Plant Physiology, Germany

REVIEWED BY
Jiesi Luo,
Southwest Medical University, China
Lingyun Zou,
Chongqing University Central Hospital, China

*CORRESPONDENCE
Tal Pupko
talp@tauex.tau.ac.il

†These authors have contributed equally to this work

SPECIALTY SECTION
This article was submitted to Plant Systems and Synthetic Biology, a section of the journal Frontiers in Plant Science

RECEIVED 21 August 2022
ACCEPTED 11 October 2022
PUBLISHED 31 October 2022

CITATION
Wagner N, Alburquerque M, Ecker N, Dotan E, Zerah B, Pena MM, Potnis N and Pupko T (2022) Natural language processing approach to model the secretion signal of type III effectors. *Front. Plant Sci.* 13:1024405. doi: 10.3389/fpls.2022.1024405

Natural language processing approach to model the secretion signal of type III effectors

Naama Wagner[1†], Michael Alburquerque[1†], Noa Ecker[1], Edo Dotan[1], Ben Zerah[1], Michelle Mendonca Pena[2], Neha Potnis[2] and Tal Pupko[1*]

[1]The Shmunis School of Biomedicine and Cancer Research, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel, [2]Department of Entomology and Plant Pathology, Auburn University, Auburn, AL, United States

Type III effectors are proteins injected by Gram-negative bacteria into eukaryotic hosts. In many plant and animal pathogens, these effectors manipulate host cellular processes to the benefit of the bacteria. Type III effectors are secreted by a type III secretion system that must "classify" each bacterial protein into one of two categories, either the protein should be translocated or not. It was previously shown that type III effectors have a secretion signal within their N-terminus, however, despite numerous efforts, the exact biochemical identity of this secretion signal is generally unknown. Computational characterization of the secretion signal is important for the identification of novel effectors and for better understanding the molecular translocation mechanism. In this work we developed novel machine-learning algorithms for characterizing the secretion signal in both plant and animal pathogens. Specifically, we represented each protein as a vector in high-dimensional space using Facebook's protein language model. Classification algorithms were next used to separate effectors from non-effector proteins. We subsequently curated a benchmark dataset of hundreds of effectors and thousands of non-effector proteins. We showed that on this curated dataset, our novel approach yielded substantially better classification accuracy compared to previously developed methodologies. We have also tested the hypothesis that plant and animal pathogen effectors are characterized by different secretion signals. Finally, we integrated the novel approach in Effectidor, a web-server for predicting type III effector proteins, leading to a more accurate classification of effectors from non-effectors.

KEYWORDS

type III secretion system, secretion signal, machine learning, natural language processing, effectors, pathogenomics

1 Introduction

A large number of plant and animal bacterial pathogens use Type III Secretion Systems (T3SS), Type IV Secretion Systems (T4SS), and Type VI Secretion Systems (T6SS), to translocate bacterial proteins called effectors into host cells, thus promoting their pathogenicity (Green and Mecsas, 2016). While proteins encoding the secretion apparatus are relatively conserved among different bacterial clades, the effector repertoire is highly variable, even among closely related strains of the same species (Groisman and Ochman, 1996). Furthermore, experimental validation of putative effectors is both labor intensive and costly, motivating the development of bioinformatic algorithms for effector prediction. Detecting the complete repertoire of effectors encoded in a given pathogenic bacterial genome is a critical first step to elucidate the molecular mechanisms involved in the host-pathogen interactions.

Over a decade ago, others and we were the first to formulate the problem of effector identification as a machine-learning (ML) classification problem (Arnold et al., 2009; Burstein et al., 2009; Löwer and Schneider, 2009; Samudrala et al., 2009). Since then, numerous ML approaches have been applied to predict effectors for T3SS, T4SS and T6SS, for various animal and plant pathogenic bacteria (Yang et al., 2010; Niemann et al., 2011; Sato et al., 2011; Yang, 2012; Dong et al., 2013; Zou et al., 2013; Lifshitz et al., 2014; Burstein et al., 2015; Burstein et al., 2016; Hobbs et al., 2016; Teper et al., 2016; Ashari et al., 2018; Jiaweiwang et al., 2018; Nissan et al., 2018; Jiménez-Guerrero et al., 2020; Ruano-Gallego et al., 2021; Wagner et al., 2022).

Successful ML-based prediction relies on curated data to be used for training and validation, i.e., a set of known effectors as well as a set of non-effectors. Previous published work differed in the set of sequences used for training and testing, hampering fair comparisons among different ML classification tools. Moreover, numerous algorithms for ML-based classification were evaluated, including Naïve Bayes (Tay et al., 2010), Support Vector Machine (SVM) (Yang et al., 2010; Wang et al., 2013a; Goldberg et al., 2016), Random Forest (Yang X. et al., 2013), LightGBM (Wang et al., 2019), and recently deep learning approaches (Hobbs et al., 2016; Fu and Yang, 2019; Xue et al., 2019; Jing et al., 2021). Finally, some reports considered identifying effectors that share high sequence similarity with previously identified effectors as a success. In contrast, others specifically removed highly similar sequences from both train and test data, to emphasize the ability of their algorithm to detect novel effectors, i.e., effectors that share no sequence similarity with previously identified effectors.

Classic ML-tools extract different sets of features from each sequence. Common features used by many effector prediction algorithms include: (1) GC content; (2) Amino-acid composition of the N-terminus (see below); (3) Presence of sequence homology to other effectors, or to non-effectors. In essence, the ML classification task is to provide a function from each possible vector of features to a score, which reflects the likelihood that this sequence encodes an effector. Finding this function is computed based on training examples. These training examples are labeled, i.e., we know whether each sequence in this set encodes an effector or not. The function is then tested on labeled data that were not used for training, to evaluate the performance. Trained models can also be applied to unlabeled data for the task of discovering putative novel effectors, which are subsequently verified experimentally (Burstein et al., 2009; Burstein et al., 2015; Teper et al., 2016; Jiménez-Guerrero et al., 2020; Ruano-Gallego et al., 2021). However, for the purpose of comparing the accuracy of different classification algorithms, here we will only deal with labeled data.

Clearly, many features are highly informative for the task of differentiating type III effectors (T3Es) from non-effectors. For example, a common feature in ML-based T3E identification tools is sequence similarity to eukaryotic domains. As T3Es often interact directly with host proteins, they frequently have domains that resemble their host proteins, both in sequence and in structure. These domains are almost always found in eukaryotes only (Stebbins and Galaán, 2001; Desveaux et al., 2006; Jelenska et al., 2007; McCann and Guttman, 2008). While such features are highly informative for the identification of novel T3Es, it is clear that the interaction between the secretion apparatus and the T3E is not based on the presence of an eukaryotic domain. Extensive previous work has localized the secretion signal of T3Es to their N-terminal region. It is of high interest to elucidate the characteristics of this secretion signal, to better understand the biochemical mechanism by which the bacterial cell sorts its proteins to secreted versus non-secreted. A better understanding of the secretion signal will also improve ML-based methods that utilize N-terminus features as part of their prediction.

The secretion signal of T3Es was first shown to reside in their N-terminus by analyzing pathogenic *Yersinia* effectors. The N-terminus was shown to be both essential and sufficient for secretion. It was also shown that no clear sequence similarity exists between the N-termini of validated effectors, suggesting that the secretion signal is not a simple sequential motif (Michiels and Cornelis, 1991; Sory and Cornelis, 1994). Yet, the importance of specific amino acids for secretion was demonstrated both computationally and experimentally. For example, extensive mutagenesis of positions 2-9 of the *Yersinia* effector YopE has clearly shown the importance of serine residues within an amphipathic region for secretion (Lloyd et al., 2002). As stated above, efforts to characterize the secretion signal were often part of a more general task of developing ML algorithms for predicting novel effector proteins. Thus, for example, both the overall amino-acid composition within the N-terminus and the occurrence of specific residues at specific sites were considered as features in SIEVE, one of the first ML tools for predicting T3Es (Samudrala et al., 2009). Of note, in that work, different lengths

of the N-terminal regions were considered, and it was concluded that accounting for more than 29-31 residues from the N-terminus does not contribute to the ability to correctly classify effectors from non-effectors. An optimum length of 30 residues was also concluded by Löwer and Schneider (2009). They implemented one-hot encoding and a sliding window approach to capture the amino-acid composition of the effectors N-termini. Another early ML work used a reduced alphabet, and suggested that positions 1-30 from the N-terminus were important for plant T3Es and 1-50 for animal T3Es (Arnold et al., 2009). In that work it was also suggested that secondary structure features had no significant contribution to prediction. A sliding window approach for characterizing the secretion signal was also suggested, accounting for such factors as hydrophobicity, polarity, and the occurrence of beta turns (Tay et al., 2010). Such an approach can potentially capture spatial variation of the signal along the sequence, yet it allows some flexibility with regard to the location. In contrast to Arnold et al. (2009); Yang et al. (2010) reported the benefit of including predicted secondary-structure information and solvent accessibility in addition to amino-acid composition for accurate prediction of T3Es. In that work it was further claimed that including k-mer based features did not contribute to prediction accuracy.

Following these initial efforts, additional representation of the amino-acid composition, combined with various ML algorithms and train and test data were developed (Sato et al., 2011; Wang et al., 2011; Wang et al., 2013a; Wang et al., 2013b; Dong et al., 2015; Goldberg et al., 2016; Hobbs et al., 2016; Cheng et al., 2018; Fu and Yang, 2019; Wang et al., 2019; Hui et al., 2020; Ding et al., 2021; Li J. et al., 2021; Yu et al., 2021). These studies differed in: (1) the way the sequences of effectors and non-effectors were encoded; (2) the selection of training and test data; (3) the algorithms used for classification; (4) the hyper-parameters tuning these classifiers. In addition, many of these works included features that are not related to the secretion signal, e.g., sequence similarity to host proteins and for plant pathogens and the existence of regulatory elements such as the PIP-box (Fenselau and Bonas, 1995).

In recent years, large scale pre-trained models such as BERT (Devlin et al., 2019) and GPT (Brown et al., 2020) have allowed for great progress in the field of NLP. Their success has led to the adoption of large scale pre-trained models in other domains, e.g., in the field of computer vision (Simonyan and Zisserman, 2014). Such pre-trained models enable the encoding of large amounts of domain knowledge into millions of learned parameters (Han et al., 2021). The extensive data analyzed during training should allow capturing many nuances of the domain, which otherwise would take experts years to discover. This approach of using automatic features to capture a problem space, greatly differs from classical ML approaches, which require careful extraction of meaningful features to adequately represent the data. Moreover, these hand-crafted features are often very task specific, making them useless in tasks that operate on a different problem space. In the context of our study, the pre-trained models were not developed in the context of T3Es. Yet, as we demonstrate, they are useful outside the immediate context for which they were developed. Finally, we note that the extensive model training in such cases is done only once.

Recently Facebook's research team created such a large-scale pre-trained model on multiple sequence alignments (MSAs) (Rao et al., 2021). This "MSA-transformer" was pre-trained on a large database, 3.8TB in size, representing 28 million MSAs. Specifically, their data contained all the protein MSAs available in RefSeq (Li W. et al., 2021). The trained "MSA-transformer" can transform a user input MSA (or sequence) into a high-dimension vector, which ideally should capture the information in this MSA. In the original paper the utility of the transformer was demonstrated on two downstream tasks; protein contact prediction and secondary structure prediction. On both tasks the model accuracy was equivalent and sometimes even higher than state-of-the-art computational tools, showing the utility of this approach in biological domains as well.

In this work, we focused on ML-based analysis of the secretion signal of T3Es. We aimed to compare different methodologies to encode the secretion signal, on the same train and test data, using the same classifiers and hyperparameter tuning. As we aim to study the secretion signal harnessed within the N-terminal region of T3Es, we compared only methods that use the amino-acid sequence as input, i.e., we did not include features such as regulation and the presence of specific eukaryotic domains. In addition to previously-described features, we also aimed to test the utility of encoding the secretion signal using Natural Language Processing (NLP) approaches, specifically using large pre-trained models. Finally, we incorporated the optimal characterization of the secretion signal into Effectidor, our previously developed ML algorithm for predicting T3Es, which uses a host of features including regulation, spatial distribution of effectors within the genome, homology to known effectors and to non-effectors, to name but a few (Wagner et al., 2022).

2 Materials and methods

2.1 Data

Positive and negative datasets. For preparing the positive data, a total of 1,857 known effectors from plant and animal pathogens were first retrieved from the Effectidor web server (Wagner et al., 2022) at https://effectidor.tau.ac.il/T3Es_data/T3Es.faa. These data were filtered to remove closely related homologs by conducting a blastP search (all-against-all with an E value cutoff of 10^{-4}) and randomly selecting a single representative from each connected component and by only considering effectors of length equal or higher than 100 amino

acids. This resulted in a total of 641 positive effectors. From this dataset, we removed 84 *Xanthomonas* effectors to be served as the positive data for the "*Xanthomonas* dataset". From the remaining effectors we randomly removed 60 effectors to serve as the positive set for the "test dataset". The remaining 497 effectors serve as the positive "training data". Thus, the entire data are comprised of three independent datasets: train data, test data, and *Xanthomonas*.

To prepare the negative data we first extracted all open reading frames (ORFs) of *Escherichia coli* K12 substr. MG1655. This strain does not encode for T3SS, and hence it is assumed that all the proteins encoded in this genome are not T3Es. To define a negative set for the "*Xanthomonas* dataset", we searched for entries in *Xanthomonas campestris* strain 8004, using blastP with an E-value cutoff of 10^{-4}, which show significant similarity to ORFs encoded in the *E. coli* strain. As stated above, as these proteins have a homolog in an *E. coli* strain that does not encode a T3SS, they are also assumed to be true negatives. The number of negative entries in the *Xanthomonas* dataset was 2,300. Note, the positive and the negative datasets are disjoint. To construct the negative data for the training and test datasets, we initially considered all *E. coli* proteins. To avoid any possible overlap between the negatives of the train and test data and the negative samples of the *Xanthomonas* data, we excluded from these data all queries that showed significant homology with any *Xanthomonas* ORF (using blastP with an E-value cutoff of 10^{-4}). The remaining 2,490 entries were divided to 2,244 and 246 ORFs used as negatives for the training and test datasets, respectively. We note that the number of negative examples is an order of magnitude higher than the positive dataset. This situation better reflects the situation in empirical genomes, in which the number of effectors is a small fraction of the total number of proteins encoded within the genome.

All these datasets are available at: https://github.com/naamawagner/T3ES_secretion_signal_analysis/tree/main/data.

Positive datasets derived from plant and animal pathogens. The positive training and test sets were further divided to samples derived from plant pathogens and to samples derived from animal pathogens. This resulted in 268 and 40 samples derived from plant pathogens in the training set and test set, respectively, and 229 and 20 samples derived from animal pathogens in the training set and test set, respectively.

Dataset used for Effectidor runs. To evaluate the performance of Effectidor with and without the inclusion of the secretion signal score, we established data that included 80 *Ralstonia solanacearum* GMI1000 effectors (Peeters et al., 2013), and 2,661 non-effectors found by Effectidor based on similarity to *E. coli* K12 proteins.

2.2 Features considered

The following features were considered in this work. Of note, all these features were computed from the N-terminus of effectors and non-effectors. The length of the N-terminus considered is denoted as m below. The value of m was optimized as part of the cross-validation procedure on the training data (10 values of m were considered: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100).

2.2.1 Amino acid composition

Nine sets of features reflecting the amino acid composition of the peptide were considered:

2.2.1.1 Euclidean distance to known effectors versus known non-effectors based on amino-acid composition (1 feature)

As a preprocessing step, the frequency of each of the 20 amino acids in the entire set of known effectors was computed, and similarly for the entire set of known non-effectors, in the training data. Let aa_e^i and Let aa_{ne}^i be the frequency of amino acid i in effectors and non-effectors, respectively ($\sum_{i=1}^{20} aa_e^i = \sum_{i=1}^{20} aa_{ne}^i = 1$). Similarly, let $aa_{peptide}^i$ be the frequency of amino acid i in a given peptide that we aim to classify ($\sum_{i=1}^{20} aa_{peptide}^i = 1$). We next computed the Euclidean distance between the amino acid composition vector of the peptide and the amino acid composition vector of the known effectors:

$$d_e(peptide) = \sqrt{\sum_{i=1}^{20}(aa_{peptide}^i - aa_e^i)^2}$$

Similarly, we computed the distance between the amino acid composition vector of the peptide and the amino acid composition vector of known non-effectors:

$$d_{ne}(peptide) = \sqrt{\sum_{i=1}^{20}(aa_{ne}^i - aa_{peptide}^i)^2}$$

The feature we considered is the difference between these two distances:

$$Feature_1 = d_e(peptide) - d_{ne}(peptide)$$

2.2.1.2 Euclidean distance to known effectors and Euclidean distance to known non-effectors (2 features)

In the description above, the considered feature is the difference between two distances. Instead of forcing the minus operation between these two distances, we can let the ML classifier decide how to optimally weight the two distances. Thus, in this approach, we consider two different features: $d_e(peptide)$ and $d_{ne}(peptide)$.

2.2.1.3 Amino acid frequencies in peptide (20 features)

In the above descriptions, the individual amino-acid frequencies were forced into a distance formula. Here, we considered the 20 values of $aa_{peptide}^i$ as separate features, allowing the classifier to optimally weight them.

2.2.1.4 Position-specific score matrix (1 feature). Hyperparameter: *C*

In the above descriptions, the locations of the amino acids along the peptide were ignored, i.e., shuffling the peptide sequence would not change the value of any of the above features. However, it is reasonable to expect that the location within the peptide is important. Let $n_{aa_j^i}(e)$ and $n_{aa_j^i}(ne)$ be the counts of amino acid i at position j in the set of known effectors and the set of known non-effectors, respectively. The probability of each amino acid at each position can be estimated by:

$$f_{aa_j^i(e)} = \frac{n_{aa_j^i(e)}}{\sum_{k=1}^{20} n_{aa_j^k(e)}}$$

And similarly, for $f_{aa_j^i(ne)}$. This computation provides a probability for each amino acid in each position, a data structure called position-specific score matrix (PSSM) or position weight matrix (PWM). It is common to add pseudo-counts to avoid zero probabilities (Durbin et al., 1998). We thus modify the computation of the PSSM to include a pseudo-count.

$$f_{aa_j^i(e)} = \frac{n_{aa_j^i(e)} + C}{\sum_{K=1}^{20} (n_{aa_j^K(e)} + C)}$$

In our implementation, we used $C = 1$.

Next, to score a peptide, let p_1, p_2…, p_m be the amino acids in the peptide in each of its m positions. The contribution of p_1 to the score is its probability to appear in the first position, based on the PSSM. Formally, the entire score of the peptide when compared to the PSSM of effectors is:

$$PSSM_{Score}(e) = \sum_{k=1}^{m} \log\left(f_{aa_k^{p_k}(e)}\right)$$

And similarity, for non-effectors:

$$PSSM_{Score}(ne) = \sum_{k=1}^{m} \log\left(f_{aa_k^{p_k}(ne)}\right)$$

The feature we consider is the difference between the two PSSM scores:

$$PSSM_{Score} = PSSM_{Score}(e) - PSSM_{Score}(ne)$$

2.2.1.5 Position-specific score matrices (2 feature). Hyperparameter: *C*

As in the above, it may be more informative to consider the two PSSM scores, $PSSM_{Score}(e)$ and $PSSM_{Score}(ne)$ as two separate features.

2.2.1.6 Position-specific score matrices per position (*m* features). Hyperparameter: *C*

It may be more informative to consider the score of each position as a separate feature. Thus, in this representation, we consider the following m features: $\log\left(f_{aa_j^{p_j}(e)}\right) - \log\left(f_{aa_j^{p_j}(ne)}\right)$, where $j = 1,\ldots,m$.

2.2.1.7 Position-specific score matrices per position (2*m* features). Hyperparameter: *C*

In the above descriptions, the individual PSSM scores per each position were combined into a single feature by a subtraction operation. Here, we consider each score as an individual feature, thus allowing the classifier to optimally weight each of them. The $2m$ features in this case are the m values of $\log\left(f_{aa_j^{p_j}(e)}\right)$ and the m values of $\log\left(f_{aa_j^{p_j}(ne)}\right)$, where $j = 1,\ldots,m$.

2.2.1.8 One-hot encoding per position (20*m* features)

In One-hot encoding each amino acid in each position is represented as a binary vector of size 20, where each entry is 0 if the amino acid is absent and 1 if present. These vectors are than concatenated to create a $20m$ representation of the entire sequence, where m is the length of the peptide. Each coordinate of this vector is considered as a separate feature.

2.2.1.9 One-hot encoding with a sliding window (20w features); Hyperparameter: *w, l*

The above algorithm can be trivially extended to overlapping windows of size w. When using such a sliding window approach, the entire peptide sequences is divided to overlapping windows. The degree of overlapping is defined by the offset parameter l, the number of characters in the left window that are not included in the right window. In this work, we used $l = 1$. In such an approach each peptide contributes several windows to the learning (and for the testing) and each such a window is encoded by a vector of size 20w features. In other words, the trained classifier predicts for each sequence of length w whether or not it is a T3E. Once a new sequence is provided, the trained classifier predicts for each of its m-w+1 windows whether or not it is part of a T3E. If the majority of windows are predicted to be T3E, then the entire sequence is predicated to be T3E (Löwer and Schneider, 2009).

2.2.2 Hidden Markov Model (1 feature)

Hidden Markov models (HMMs) are probabilistic models. Each hidden state generates columns based on probabilities similar to a PSSM matrix, i.e., it emits characters (in our case, amino acids) based on a specific frequency distribution. The entire sequence is modeled by a Markov process over the hidden states, i.e., the sequence is represented by an ensemble of hidden states. Here, we trained the HMM model using the Baum-Welch algorithm (Rabiner, 1989), which is an expectation-maximization (EM) algorithm that iteratively improves the data likelihood function until convergence. The number of hidden states was determined using a 3-fold cross-validation.

To this end, the entire set of training data was randomly split to three folds. The HMM model was trained on the positive sequences of two of the folds with each possible value of number of hidden states between 1 and 20. The obtained HMM model for each number of hidden states was then used to evaluate the log-likelihood of each sequence in the third fold (both positive and negative). The performance of a classifier with a single feature (the HMM's log-likelihood score) was used to evaluate the performance (using the MCC value) of each possible number of hidden states. This was repeated three times, each time a different fold was used for evaluation. The number of hidden states that was selected was the one that yielded the highest average MCC. After the number of hidden states was determined (the optimum was 10), the HMM was trained on the entire set of positive training data. The score of this HMM for each sequence in the test data was treated as a feature for classification.

2.2.3 LSTM model

Long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) networks are a form of recurrent neural networks designed to process sequential data. The data must be segmented into tokens, and in the case of protein sequences usually each amino acid is referred to as a token. These networks operate on raw sequence data and learn how to represent these data as part of their training process. LSTM networks are often used for translation between two languages. To adapt this translation task for classifying proteins to either T3E or non-effector, the model was trained to translate the protein sequence into a language that has only two words, yes and no, corresponding to whether the sequence is a T3E or not, respectively. Here we used an LSTM with an encoder-decoder architecture, each having a single RNN layer with a hidden state of size 512 from the "fairseq" Python package (Ott et al., 2019).

2.2.4 Using Facebook's "MSA transformer" for classification – LME (1,280 features)

Facebook's "MSA transformer" allows to leverage a very extensive training process to extract meaningful features from protein sequence data (Rao et al., 2021). Facebook's "MSA transformer" receives as input an alignment in FASTA format and here we used as input an "alignment" of a single sequence. The output is a set of 1,280 weights extracted from the last (34^{th}) layer of the encoder. These weights are used by the neural network to encode the sequence, and can be used as a 1,280-dimensional feature vector.

2.2.5 Features from Hobbs et al. (17 features)

Hobbs et al. (2016) suggested a set of features that can potentially classify effectors from non-effectors. We tested whether the inclusion of these features improves classification accuracy. Specifically, the following sets of amino acids are defined: (1) Tiny: A, C, G, S, T; (2) Small: A, C, D, G, N, P, S, T, V, B; (3) Aliphatic: A, I, L, V; (4) Aromatic: F, H, W, Y; (5) Polar: D, E, H, K, N, Q, R, Z; (6) Non-polar: A, C, F, G, I, L, M, P, V, W, Y; (7) Charged: D, E, H, K, R, B, Z; (8) Basic: H, K, R; (9) Acidic: D, E, B, Z. Each such a set defines a single feature, which is the fraction of positions in which one of the amino acids in the set is present. These features were implemented in Python.

The following features were implemented using the Biopython package "ProteinAnalysis": (10) "Charge", which measured the total charge of the peptide (in pH = 7); (11) A_{280} molar extinction coefficient, which predicts the light absorbance of the protein in 280 nm; (12) Isoelectric point; (13) Instability index; (14) Aliphatic index; (15) GRAVY Score; (16) Molecular weight. (17) "Probability of expression in inclusion bodies (PEPIB)", which was implemented in Python based on Ahuja et al. (2006).

2.3 ML model

Protein sequences were classified to either effectors or non-effectors using LightGBM, a decision-tree classifier with gradient boosting (Ke et al., 2017), as implemented in the Python package lightgbm. The following LightGBM hyperparameters were optimized using ten-fold cross validation on the train data: type of boosting algorithm ('gbdt', 'dart', 'rf', 'goss'), number of leaves in each tree (10, 30, 50, 100, 200), tree depth (10, 100, 1,000, infinite), learning rate (0.1, 0.05, 0.005), number of tree estimators (10, 50, 100, 200, 1,000), alpha (0, 0.5, 1, 3, 10, 100), lambda (0, 0.5, 1.5, 3, 100, 500, 1,000, 1,200), and is-balanced (True/False). The is-balanced hyperparameter controls weights assigned to each class (in our case, effectors versus non-effectors), and may be highly important for unbalanced datasets.

2.4 Performance evaluation methods

Several scoring methods were used to evaluate performance. As the data are unbalanced, i.e., the number of negative samples is an order of magnitude higher than the number of positive samples, traditional scoring methods such as accuracy or the Area Under the Curve (AUC) are not well suited. Instead, the Area Under the Precision-Recall Curve (AUPRC) and Matthew's Correlation Coefficient (MCC) are more suitable for unbalanced data. While the AUPRC score is used for probabilistic predictions, the MCC score is used for binary predictions, and is calculated with the following formula:

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

Where TN (True Negative) is the number of non-effectors correctly classified as non-effectors, TP (True Positive) is the number of effectors correctly classified as effectors, FN (False Negative) is the number of effectors misclassified as non-

effectors, and FP (False Positive) is the number of non-effectors misclassified as effectors. In cases where the AUPRC was impossible to compute, i.e., in deterministic prediction rather than a probabilistic one, the F1 score was used. The F1 score is calculated using the following formula:

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

While precision and recall are given in the following formulas:

$$precision = \frac{TP}{TP + FP}, \quad recall = \frac{TP}{TP + FN}$$

2.5 Integrating the signal score into the Effectidor web server

Our results below show that the LME model (1,280 features) together with the 17 features from Hobbs et al. (2016), provides the best classifier based on the secretion signal. We term the lightGBM classifier, which was trained on the entire data based on these features, "trained LME". The hyperparameters of this model were the same as those found to yield the best performance using cross validation on the training data. This trained LME model provides for each possible protein a score that reflects its propensity to harbor a type III secretion signal within its 100 N-terminal positions. This score was added as an optional feature within the Effectidor web server (Wagner et al., 2022).

An overview of the entire pipeline implemented in this work is available in Figure 1.

3 Results

3.1 Performance of all algorithms

We have implemented three new ML-based approaches to model the secretion signal of T3Es: LME, LSTM, and HMM (see Methods). Our results clearly show that on both testing data, the accuracy is highest for LME, and lowest for HMM (Table 1). The LME accuracy was highest when all 100 amino acid residues of the N-terminus were considered (Figure 2). In all three methods, the accuracy was higher for the test dataset compared to the *Xanthomonas* dataset (for the LME methodology, the MCC and AUPRC were 0.81 and 0.88 for the test dataset, respectively, and 0.71 and 0.77 for the *Xanthomonas* datasets, respectively).

We then compared these methods to nine alternative methods to model the secretion signal (Table 2). The best of these methods performed substantially poorer compared to the LME method (Figure 2). For the test dataset, the highest accuracy among these nine methods was obtained using the 20 amino acid frequencies as features, yielding an MCC and AUPRC values of 0.65 and 0.71, respectively. For the *Xanthomonas* dataset, the best performing method was position-specific score matrices per position (*m* features), with an associated MCC and AUPRC values of 0.45 and 0.5, respectively. These results clearly show that our proposed novel LME methodology is well suited for modeling the secretion signals of T3Es.

3.2 Testing feature combinations

We next tested the hypothesis that the LME method can be further improved by integrating several features. As the various models for amino-acid composition (see section 2.2.1) are very similar, for the combination analysis we selected the best performing method among them, i.e., amino-acid frequencies. We tested seven combinations of feature groups. Results on the test dataset as well as the *Xanthomonas* dataset show that combining different groups of features had a marginal impact on accuracy: on the test dataset, the best performing combination increased the MCC and AUPRC scores from 0.81 and 0.88 to 0.83 and 0.91, respectively. On the *Xanthomonas* dataset, the MCC and AUPRC scores were improved from 0.71 and 0.77 to 0.72 and 0.87, respectively (Table 3). We conclude that the major improvement in modeling the secretion signal stems from the proposed LME method.

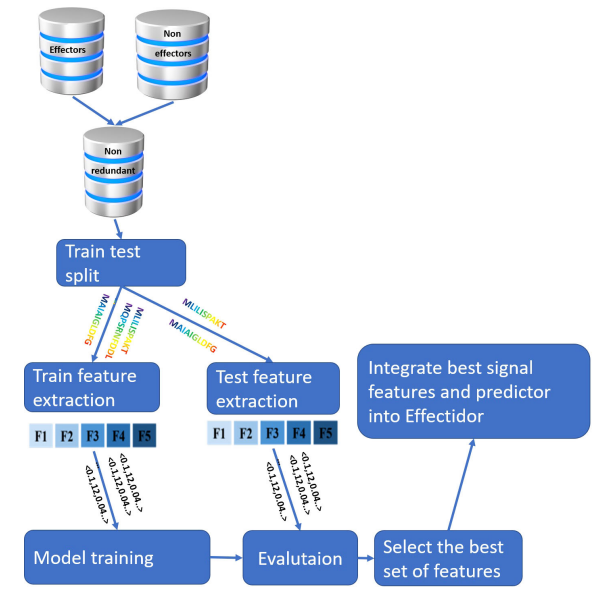

FIGURE 1
Overview of the pipeline implemented in this work. The input for the feature extraction are the N-terminal protein sequences, {F1,..., F5} represent different features, and the input to the model training and evaluation are the vectorial representations of the protein samples, as represented by the features.

TABLE 1 The performance of the three methods proposed in this work. MCC and AUPRC (in parenthesis) on the test datasets ("test dataset" and "*Xanthomonas* dataset").

	Hidden Markov Model	LSTM	LME
Test dataset	0.43 (0.4)	0.63 (0.7)	**0.81 (0.88)**
***Xanthomonas* dataset**	0.12 (0.05)	0.37 (0.44)	**0.71 (0.77)**

On the LSTM methods, instead of AUPRC, the F1 score is given, as it was a deterministic prediction. These results were obtained by analyzing the 100 N-terminal amino acids. In bold are the best scores for each of the test datasets.

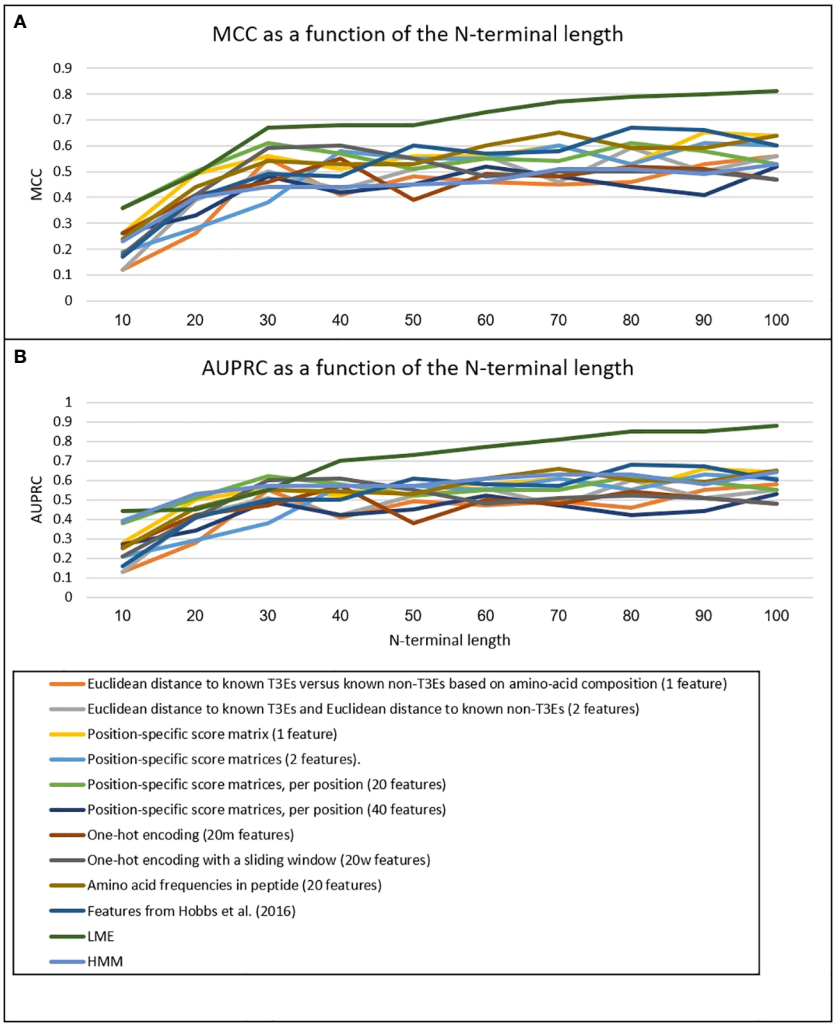

FIGURE 2
MCC **(A)** and AUPRC **(B)** of the different methods as a function of the N-terminal length. Scores are the mean scores of 10-fold Cross-Validation on the training data.

3.3 Performance when applied separately to plant/animal associated bacterial effectors

The different results on the test and the *Xanthomonas* datasets led us to hypothesize that plants and animal bacterial pathogens may have different type III secretion signals. To test this hypothesis, we further divided our training and testing data to samples derived from plant pathogens and from animal pathogens (see Methods). We then evaluated the models trained on the different training sets, on the following test sets: T3Es derived from animal pathogens, T3Es derived from plant pathogens, T3Es derived from both plant

TABLE 2 The performance of alternative methods to model the secretion signal.

	Test dataset	*Xanthomonas* dataset
Euclidean distance to known T3Es versus known non-T3Es based on amino-acid composition (1 feature)	0.56 (0.6)	0.41 (0.5)
Euclidean distance to known T3Es and Euclidean distance to known non-T3Es (2 features)	0.54 (0.58)	0.32 (0.28)
Position-specific score matrix (1 feature)	0.65 (0.7)	0.34 (0.45)
Position-specific score matrices (2 features).	0.61 (0.66)	0.31 (0.4)
Position-specific score matrices, per position (20 features)	0.53 (0.59)	**0.45 (0.5)**
Position-specific score matrices, per position (40 features)	0.5 (0.52)	0.31 (0.42)
One-hot encoding (20*m* features)	0.5 (0.51)	0.24 (0.2)
One-hot encoding with a sliding window (20*w* features)	0.5 (0.49)	0.24 (0.3)
Amino acid frequencies in peptide (20 features)	**0.65 (0.71)**	0.41 (0.5)
Features from Hobbs et al. (2016)	0.61 (0.69)	0.44 (0.49)

MCC and AUPRC (in parenthesis) on the test datasets. These results were obtained by analyzing the 100 N-terminal amino acids. In bold are the best scores for each of the test datasets.

and animal pathogens, and T3Es derived from *Xanthomonas*. Our results show that predicting T3Es from animal pathogens is best achieved by training the model on animal-derived T3Es, and similarly for plant pathogens (Table 4). Furthermore, as expected, the best predictor for *Xanthomonas* T3Es is the model trained on the T3Es from the plant pathogens. These results suggest that the secretion signal of animals and plants are different to some extent.

In the above results, we compared plant versus animal models that were trained on the same number of effectors. We next asked whether the plant-based model can benefit from the inclusion of animal T3Es and similarly, whether the animal-based model can benefit from the inclusion of plant T3Es. Our results clearly show that the accuracy is increased when the largest number of positive samples is considered (Table 4). These results suggest that despite evidential differences between secretion signals of plant and animal pathogens, a model extracting information from all known effectors better captures the secretion signals, compared to a more specialized model trained on a smaller number of effectors.

3.4 Integrating the secretion signal feature into the Effectidor web server and running example on *Ralstonia solanacearum* GMI1000

We next tested the effect of integrating the secretion signal score based on the LME model and the Hobbs et al. (2016) features on the performance of Effectidor. This effect was evaluated on the plant pathogen *Ralstonia solanacearum* GMI1000. To separate training from testing data and to mimic the applicability of Effectidor to detect T3Es in a newly sequenced bacterium, we removed all *Ralstonia* effectors from the database of known T3Es that is used within the Effectidor web server for homology searching (see Methods). We ran Effectidor twice, with or without the additional feature of the signal score. To obtain the signal score for this testing case, we trained the model based on the non-redundant T3Es data, as described in the methods, excluding *Ralstonia* data.

TABLE 3 The performance of combinations of top-scoring features from each group of features.

Feature combination	Test dataset	*Xanthomonas* data
Amino acids frequencies + LME	0.81 (0.83)	0.70 (0.72)
LME + HMM	0.8 (0.8)	0.68 (0.7)
LME + Features from Hobbs et al. (2016)	**0.83 (0.91)**	**0.72 (0.87)**
Amino acids frequencies + LME + HMM	0.81 (0.81)	0.7 (0.7)
Amino acids frequencies + LME + Features from Hobbs et al. (2016)	0.8 (0.81)	0.7 (0.72)
HMM + LME + Features from Hobbs et al. (2016)	0.82 (0.85)	0.71 (0.73)
Amino acids frequencies + LME + HMM + Features from Hobbs et al. (2016)	0.82 (0.83)	0.72 (0.72)

The results were obtained on the test datasets by analyzing the 100 N-terminal amino acids. The scores are MCC and AUPRC (in parenthesis). In bold are the best scores for each of the test datasets.

TABLE 4 The performance of the best set of features when applied separately to animal and plant associated bacterial effectors.

A

Training data	Plant test dataset (40)	Animal test dataset (20)	Animal + plant test dataset (20 + 40)	Animal + plant test dataset (20 + 20)	*Xanthomonas*
Animal pathogens T3Es (229)	0.63 (0.65)	**0.77** **(0.78)**	0.69 (0.7)	0.7 (0.71)	0.44 (0.46)
Plant pathogens T3Es (229)	**0.8** **(0.8)**	0.71 (0.7)	**0.76** **(0.78)**	**0.77** **(0.76)**	**0.64** **(0.67)**
Both animal and plant pathogens T3Es (115 + 114)	0.53 (0.54)	0.55 (0.56)	0.6 (0.63)	0.62 (0.61)	0.58 (0.58)

B

Training data	Plant test dataset (40)	Animal test dataset (20)	Animal + plant test dataset (20+40)	Animal + plant test dataset (20+ 20)	*Xanthomonas*
Animal pathogens T3Es (229)	0.63 (0.64)	0.77 (0.76)	0.69 (0.71)	0.7 (0.71)	0.44 (0.46)
Plant pathogens T3Es (268)	0.8 (0.8)	0.71 (0.72)	0.76 (0.77)	0.77 (0.77)	0.64 (0.66)
Both animal and plant pathogens T3Es (497)	**0.86** **(0.88)**	**0.79** (0.72)	**0.83** **(0.85)**	**0.84** **(0.85)**	**0.72** **(0.73)**

(A): each training data include 229 T3Es; (B): the training data include the maximum possible T3Es from that category. In bold is the best training data for each testing data. Data sizes are in parenthesis. The scores are MCC and AUPRC (in parenthesis).

Nine effectors were found to harbor homology to T3Es outside *Ralstonia* and were considered as the positive trainingset for Effectidor. The inclusion of the secretion signal feature substantially increased performance: the confusion matrices are given in Table 5. The MCCs with and without this featurewere 0.72 and 0.65, respectively, while the AUPRC with and without this feature were 0.94 and 0.93, respectively. The newly added feature was found to be highly informative (Figure 3).

Using this feature alone, the achieved AUPRC and MCC were 0.79 and 0.53, respectively. Despite these low scores compared to the scores achieved by Effectidor, only three effectors out of 80 T3Es known in this strain were misclassified as non-effectors by this feature. Specifically, RS_RS23105 (RipAR, formerly Rip61), RS_RS10690 (RipS6, formerly SKWP6), and RS_RS26010 (RipBM) signal scores were 0.477, 0.252, and 0.166, respectively. The latter is reported to be present only as a pseudogene in *R. solanacearum* GMI100 in the "Ralstonia T3E" dataset (https://iant.toulouse.inra.fr/T3E) (Peeters et al., 2013). However, classification based only on the secretion signal score led to a high number of false-positives: 168 non-effectors were classified as effectors, i.e., they obtained a signal score higher than 0.5 (Table 5). The Precision-Recall curves, of Effectidor and of the signal feature alone, are available in Figure 4.

3.5 Interpreting the secretion signal using attention maps

In transformer-based neural networks, a key component is the attention mechanism (Vaswani et al., 2017). This mechanism allows the model to learn which "words" attend to which other "words". In our case the attention maps provide information regarding interactions among positions within the secretion signal. This, in turn, allows us to better understand, for each position within a specific sequence, which other positions are most important for the embedding. By contrasting the average attention matrix across all positive versus the average matrix over all negative sequences, we reviled different interactions among sites between T3E versus non-effectors (Figure 5). In the positive sequences, a large number of interactions among positions across the sequence is observed, even though there is strong variability of amino acids in each position. In contrast,

TABLE 5 The effect of including a novel feature that quantifies the strength of the secretion signal on the performance of Effectidor.

	A		B		C	
	T3Es	Non-T3Es	T3Es	Non-T3Es	T3Es	Non-T3Es
Predicted T3Es	34	0	42	0	77	168
Predicted non-T3Es	46	2,661	38	2,661	3	2,493

(A): without the secretion signal feature; **(B)**: with the secretion signal feature; **(C)**: the signal feature alone.

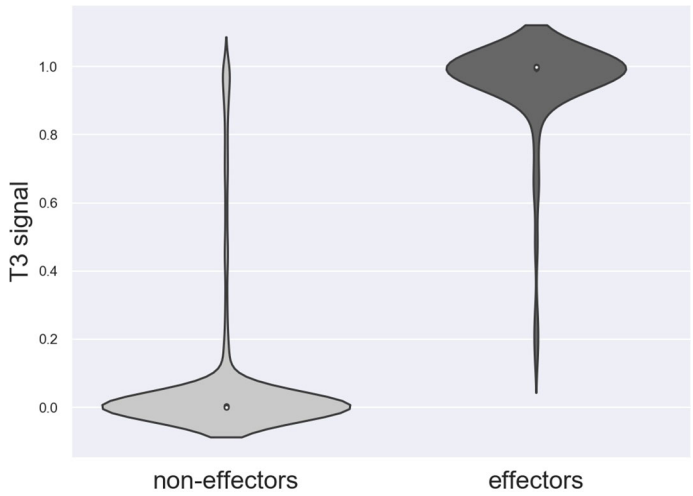

FIGURE 3
Distribution of the secretion signal score among the 2,661 non-effectors and among the 80 effectors in *R. solanacearum* GMI1000.

when embedding the negative sequences, each position does not attend to many other positions. This analysis shows the diffusive nature of the type III secretion signal and warrants further research into each of these multi-position interactions.

4 Discussion

In this study, we aimed to better characterize the secretion signal of T3Es. We developed a novel NLP-based approach, using transformers that were specially derived for capturing information in protein MSAs, and demonstrated that classification based on this approach is more accurate than previous approaches. All models were compared using T3Es from various bacteria, including both plant and animal pathogens. Finally, we integrated this feature as part of the Effectidor web server for predicting T3Es.

The modeling of biological sequence data as a language and the incorporation of ML tools for languages, led to many advances in various biological domains (Rao et al., 2020; Rives

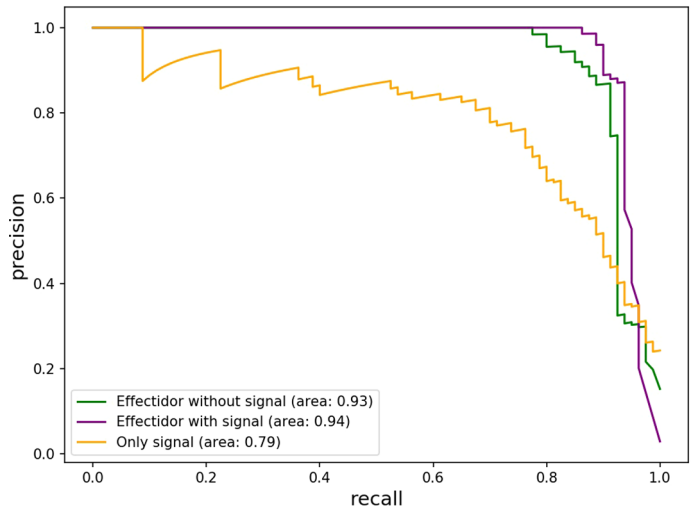

FIGURE 4
Precision-Recall Curves of predicting T3Es in *R. solanacearum* GMI1000 based on the secretion signal feature as a sole feature, and based on the predictions of Effectidor, with and without the inclusion of the secretion signal feature.

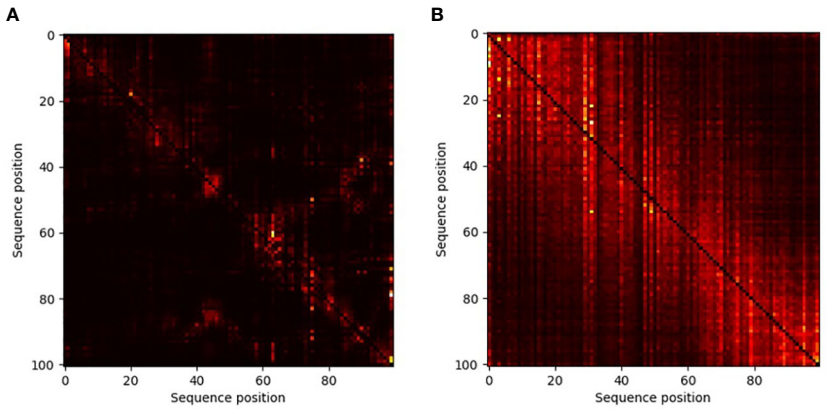

FIGURE 5
Attention maps of non-T3Es **(A)** and T3Es **(B)** demonstrating the effect of different positions in the amino acids sequence in the embedding process. The lighter the color, the more significant the interaction is. In these maps, each column (j) is a position in the sequence, and each entry (i,j) is the effect of position i on position j. The diagonal, which reflects the effect of each position on itself, was blacked out.

et al., 2021; Trotter et al., 2021). Naturally, the problem of classifying proteins to either T3Es or non-effectors can also be approached using NLP-based methodologies. The first effort in this direction, from Fu and Yang (2019) was based on the now classic algorithm called Word2Vec, specifically, the Skip-Gram version with negative sampling (Mikolov et al., 2013). This algorithm embeds each possible protein sequence in a high-dimension space. The coordinates of this vector are used as features in standard classification algorithms. When adapting Word2Vec to protein sequences, each k-mer sequence is a word, and the entire sequence is represented as multiple overlapping k-mers. The embedding is done so that k-mers that appear together are close to each other in the abovementioned space. The final sequence embedding is obtained by averaging all k-mer embeddings. To learn this embedding protein sequences in this space, Fu and Yang (2019) analyzed a corpus of 25 million proteins from the UNIPROT50 data.

The Word2Vec algorithm clearly advanced the field of NLP in general. Since its publication, more advanced NLP algorithms were developed. These algorithms enable more efficient and accurate embedding of biological sequences. Among these advanced algorithms is Facebook's MSA-transformer (Rao et al., 2021). Unlike Word2Vec, this algorithm applies large, transformer-based language models, which were trained on the entire RefSeq protein data. One advantage of this transformer-based algorithm is that unlike Word2Vec, there is no need to "parse" protein sequences into "words", which do not appear naturally in sequence data, i.e., the transformer-based method of embedding allows to consider relations between amino acids across the entire protein sequence and not only a k-mer away. Of note, the LME approach developed here bypasses the need to explicitly design specific features.

When training and testing the algorithm, it is important to determine which proteins are included as negatives and positives. First, we note that in most applications of ML-based algorithms for effector identification, the input is the entire set of proteins encoded in a given bacterial genome. In such a scenario, we expect a few positive instances (effectors) in a sea of non-effectors (the rest of the proteins). This motivated us to establish a benchmark dataset that includes thousands of non-effectors. We note that in some previous works, the set of negatives did not well reflect this scenario. For example, in EP3 (Li J. et al., 2021), the negative set was composed of effectors of different secretion systems, other than T3SS. Clearly, this negative set is a biased sample of non-T3Es. Moreover, in Bastion3 (Wang et al., 2019) and in T3Sepp (Hui et al., 2020) the negative and positive sets were of the same size. The negative dataset used in pEffect (Goldberg et al., 2016) was enriched with eukaryotic proteins, such that only 37% of the negative dataset were bacterial. Of note, in this work, we tested the possibility of using another bacterial genome that does not encode a secretion system to define the negative set (*Lysobacter capsici* strain 55). This yielded performance almost identical to that obtained when negatives were defined based on the *E. coli* K12 genome (not shown). Having an accurate list of positives is also important. Each T3E assumed to be positive in this work was tested by ensuring that it is encoded within a genome that encodes the T3SS. Using this criterion, we discovered that erroneous T3Es were included in previous studies. We provide both the non-redundant data that were used in this work, and the entire data, which include effectors with high sequence similarity, from which the non-redundant data were derived. Finally, each effector is associated with the pathogen from which it is derived, allowing to test the accuracy on different taxonomic groups. Our benchmark data

are available at https://github.com/naamawagner/T3ES_secretion_signal_analysis/tree/main/data.

The effort to classify effectors based on their secretion signal alone, demonstrates our limited understanding of the secretion signal, i.e., the classification accuracy is a measure of our computational ability to characterize which proteins are recognized and secreted by the T3SS. Our results clearly show that despite years of progress in this field, we still do not have sufficiently accurate models of the secretion signals, and on test data, we still experience dozens of false positives and negatives. Why is the classification mediocre? Several computational explanations are possible. First, the performance may be improved by applying more advanced algorithms on the same data. Second, the training data may be too small and do not capture the true variety of T3Es. Third, it could be that some errors in the benchmark data exist, which reduces accuracy. For example, *E. coli* K-12 does not have a T3SS and therefore we assume that its proteins are not secreted. However, it is possible that if such proteins were introduced to a bacterium with a secretion system – they would be secreted, i.e., they are not true negatives. In support of this hypothesis, (Wang et al. 2013b) tested the translocation of yeast proteins that had secretion signal features in their N-terminus in *Salmonella*. It was shown that these yeast proteins were translocated, despite the fact that yeast lacks a T3SS.

When discussing secretion of T3Es it is important to mention the involvement of chaperones. Chaperones were shown to affect the secretion of some T3Es while other T3Es were often found to be unaffected by chaperones (Ernst et al., 2018). In addition, chaperone binding sites were found to reside in the N-terminus of effectors, but were also found to bind to regions beyond the first 100 N-terminal residues (Parsot et al., 2003). Moreover, while the secretion apparatus and possibly the secretion signal is shared among T3Es from both animal and plant associated bacteria, the chaperones and their binding sites are highly variable. It is highly possible that the secretion signal provides the main driving force for secretion, and the chaperones fine-tune the secretion process, e.g., by scheduling the order of secretion or by preventing their aggregation or degradation when stored within the bacterial cell (Parsot et al., 2003).

The challenge of computationally characterizing the secretion signal could be of great importance to the in-silico synthesis of new effectors. Several efforts have been done to use Transcription Activator-Like Effectors Nucleases (TALEN) and Transcription Activator-Like Effectors (TALE) for manipulating gene expression and for gene editing purposes (Yang L. et al., 2013; Gao et al., 2014). Such technological advances may improve crop production or cure genetic diseases. As stated by Cheng et al. (2021), TALEs have the potential to perform better for such tasks than CRISPR/Cas9 because TALEs recognize more specific DNA segments than the CRISPR/Cas9 system and thus they are less prone to mistakes. Moreover, they are also encoded on a shorter DNA sequence, which may facilitate their usage in various systems. Better elucidating the secretion signal of T3Es could assist the engineering of such secreted proteins.

As stated above, the underlying assumption in this work is that a universal secretion signal exists that characterizes all T3Es. In this study we have shown that learning the secretion signal from T3Es encoded in animal-associated pathogens can be used to identify T3Es in plant-associated pathogens and vice versa, although a slight reduction in prediction ability was observed in such comparisons. Given the diversity of T3SS among pathogenic bacteria, it is highly possible that many T3Es are clade specific. To our knowledge, translocation tests of plant-associated T3Es in animal-associated bacteria, and vice versa, have not been conducted in large numbers. Such experiments have the potential to reveal how universal the secretion signal is.

In this work we explored the utility of different ways to extract meaningful representation of sequence data in vector space. We tested these methods for the task of identifying T3Es. These methods can also be applied to many additional bioinformatics tasks that rely on the analysis of protein sequences, e.g., predicting type IV effectors (Lewis et al., 2019), predicting fungal effectors (Sperschneider et al., 2018), protein contact prediction (Fukuda and Tomii, 2020), and predicting protein localization (Peabody et al., 2020).

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

NW, NP, and TP conceived the project. MA and ED suggested, applied and tested the NLP approaches. NE developed the HMM model. BZ and MMP implemented some features and analyzed effector datasets. NW implemented all the classic ML approaches and integrated the new feature in Effectidor. All authors analyzed the results and helped writing the manuscript. All authors approved the final manuscript.

Funding

NW, MA, NE, and ED were supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel Aviv University.

Acknowledgments

Israel Science Foundation (ISF) [2818/21 to T.P.]; Edmond J. Safra Center for Bioinformatics at Tel Aviv University Fellowship to NW, MA, NE, and ED, TP's research is supported in part by the

Edouard Seroussi Chair for Protein Nanobiotechnology, Tel Aviv University.

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

References

Ahuja, S., Ahuja, S., Chen, Q., and Wahlgren, M. (2006). Prediction of solubility on recombinant expression of plasmodium falciparum erythrocyte membrane protein 1 domains in escherichia coli. *Malar. J.* 5, 52. doi: 10.1186/1475-2875-5-52

Arnold, R., Brandmaier, S., Kleine, F., Tischler, P., Heinz, E., Behrens, S., et al. (2009). Sequence-based prediction of type III secreted proteins. *PloS Pathog.* 5, 1000376. doi: 10.1371/journal.ppat.1000376

Ashari, Z. E., Dasgupta, N., Brayton, K. A., and Broschat, S. L. (2018). An optimal set of features for predicting type iv secretion system effector proteins for a subset of species based on a multi-level feature selection approach. *PloS One* 13, e0197041. doi: 10.1371/journal.pone.0197041

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, 1877–1901.

Burstein, D., Amaro, F., Zusman, T., Lifshitz, Z., Cohen, O., Gilbert, J. A., et al. (2016). Genomic analysis of 38 legionella species identifies large and diverse effector repertoires. *Nat. Genet.* 48, 167–175. doi: 10.1038/ng.3481

Burstein, D., Satanower, S., Simovitch, M., Belnik, Y., Zehavi, M., Yerushalmi, G., et al. (2015). Novel type III effectors in pseudomonas aeruginosa. *MBio* 6, e00161–e00115. doi: 10.1128/mBio.00161-15

Burstein, D., Zusman, T., Degtyar, E., Viner, R., Segal, G., and Pupko, T. (2009). Genome-scale identification of legionella pneumophila effectors using a machine learning approach. *PloS Pathog.* 5, e10000508. doi: 10.1371/journal.ppat.1000508

Cheng, X., Hui, X., Shu, X., P. White, A., Guo, Z., Hu, Y., et al. (2018). Prediction of new bacterial type III secreted effectors with a recursive hidden Markov model profile-alignment strategy. *Curr. Bioinform.* 13, 280–289. doi: 10.2174/1574893612666170725122633

Cheng, L., Zhou, X., Zheng, Y., Tang, C., Liu, Y., Zheng, S., et al. (2021). Simple and rapid assembly of TALE modules based on the degeneracy of the codons and trimer repeats. *Genes (Basel).* 12, 1761. doi: 10.3390/GENES12111761/S1

Desveaux, D., Singer, A. U., and Dangl, J. L. (2006). Type III effector proteins: doppelgangers of bacterial virulence. *Curr. Opin. Plant Biol.* 9, 376–382. doi: 10.1016/j.pbi.2006.05.005

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. doi: 10.18653/V1/N19-1423

Ding, C., Han, H., Li, Q., Yang, X., and Liu, T. (2021). iT3SE-PX: Identification of bacterial type III secreted effectors using PSSM profiles and XGBoost feature selection. *Comput. Math. Methods Med.* 2021, 6690299. doi: 10.1155/2021/6690299

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

Dong, X., Zhang, Y. J., and Zhang, Z. (2013). Using weakly conserved motifs hidden in secretion signals to identify type-III effectors from bacterial pathogen genomes. *PloS One* 8, e56632. doi: 10.1371/journal.pone.0056632

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*.

Ernst, N. H., Reeves, A. Z., Ramseyer, J. E., and Lesser, C. F. (2018). High-throughput screening of type III secretion determinants reveals a major chaperone-independent pathway. *MBio* 9, e01050–e01018. doi: 10.1128/MBIO.01050-18

Fenselau, S., and Bonas, U. (1995). Sequence and expression analysis of the hrpB pathogenicity operon of xanthomonas campestris pv. vesicatoria which encodes eight proteins with similarity to components of the hrp, ysc, spa, and fli secretion systems. *Mol. Plant Microbe Interact.* 8, 845–854. doi: 10.1094/MPMI-8-0845

Fukuda, H., and Tomii, K. (2020). DeepECA: An end-to-end learning framework for protein contact prediction from a multiple sequence alignment. *BMC Bioinf.* 21, 1–15. doi: 10.1186/S12859-019-3190-X/FIGURES/7

Fu, X., and Yang, Y. (2019). WEDeepT3: predicting type III secreted effectors based on word embedding and deep learning. *Quant. Biol.* 7, 293–301. doi: 10.1007/s40484-019-0184-7

Gao, X., Tsang, J. C. H., Gaba, F., Wu, D., Lu, L., and Liu, P. (2014). Comparison of TALE designer transcription factors and the CRISPR/dCas9 in regulation of gene expression by targeting enhancers. *Nucleic Acids Res.* 42, e155. doi: 10.1093/NAR/GKU836

Goldberg, T., Rost, B., and Bromberg, Y. (2016). Computational prediction shines light on type III secretion origins. *Sci. Rep.* 6, 34516. doi: 10.1038/srep34516

Green, E. R., and Mecsas, J. (2016). Bacterial secretion systems: An overview. *Microbiol. Spectr.* 4, 4–1. doi: 10.1128/microbiolspec.vmbf-0012-2015

Groisman, E. A., and Ochman, H. (1996). Pathogenicity islands: Bacterial evolution in quantum leaps. *Cell* 87, 791–794. doi: 10.1016/S0092-8674(00)81985-6

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., et al. (2021). Pre-trained models: Past, present and future. *AI. Open* 2, 225–250. doi: 10.1016/j.aiopen.2021.08.002

Hobbs, C. K., Porter, V. L., Stow, M. L. S., Siame, B. A., Tsang, H. H., and Leung, K. Y. (2016). Computational approach to predict species-specific type III secretion system (T3SS) effectors using single and multiple genomes. *BMC Genomics* 17, 1048. doi: 10.1186/s12864-016-3363-1

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/NECO.1997.9.8.1735

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–e00220. doi: 10.1128/msystems.00288-20

Jelenska, J., Yao, N., Vinatzer, B. A., Wright, C. M., Brodsky, J. L., and Greenberg, J. T. (2007). A J domain virulence effector of pseudomonas syringae remodels host chloroplasts and suppresses defenses. *Curr. Biol.* 17, 499–508. doi: 10.1016/j.cub.2007.02.028

Jiaweiwang, 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

Jiménez-Guerrero, I., Pérez-Montaño, F., Da Silva, G. M., Wagner, N., Shkedy, D., Zhao, M., et al. (2020). Show me your secret(ed) weapons: a multifaceted approach reveals a wide arsenal of type III-secreted effectors in the cucurbit pathogenic bacterium acidovorax citrulli and novel effectors in the acidovorax genus. *Mol. Plant Pathol.* 21, 17–37. doi: 10.1111/mpp.12877

Jing, R., Wen, T., Liao, C., Xue, L., Liu, F., Yu, L., et al. (2021). DeepT3 2.0: improving type III secreted effector predictions by an integrative deep learning framework. *NAR. Genomics Bioinforma.* 3, lqab086. doi: 10.1093/nargab/lqab086

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in neural information processing systems*, 3146–3154.

Lewis, J. M., Deveson Lucas, D., Harper, M., and Boyce, J. D. (2019). Systematic identification and analysis of acinetobacter baumannii type VI secretion system effector and immunity components. *Front. Microbiol.* 10. doi: 10.3389/FMICB.2019.02440

Lifshitz, Z., Burstein, D., Schwartz, K., Shuman, H. A., Pupko, T., and Segal, G. (2014). Identification of novel coxiella burnetii Icm/Dot effectors and genetic

analysis of their involvement in modulating a mitogen-activated protein kinase pathway. *Infect. Immun.* 82, 3740–3752. doi: 10.1128/IAI.01729-14

Li, W., O'Neill, K. R., Haft, D. H., Dicuccio, M., Chetvernin, V., Badretdin, A., et al. (2021). RefSeq: expanding the prokaryotic genome annotation pipeline reach with protein family model curation. *Nucleic Acids Res.* 49, D1020–D1028. doi: 10.1093/nar/gkaa1105

Li, J., Wei, L., Guo, F., and Zou, Q. (2021). EP3: an ensemble predictor that accurately identifies type III secreted effectors. *Brief. Bioinform.* 22, 1918–1928. doi: 10.1093/bib/bbaa008

Lloyd, S. A., Sjöström, M., Andersson, S., and Wolf-Watz, H. (2002). Molecular characterization of type III secretion signals *via* analysis of synthetic n-terminal amino acid sequences. *Mol. Microbiol.* 43, 51–59. doi: 10.1046/j.1365-2958.2002.02738.x

Löwer, M., and Schneider, G. (2009). Prediction of type III secretion signals in genomes of gram-negative bacteria. *PloS One* 4, e5917. doi: 10.1371/journal.pone.0005917

McCann, H. C., and Guttman, D. S. (2008). Evolution of the type III secretion system and its effectors in plant-microbe interactions. *New Phytol.* 177, 33–47. doi: 10.1111/j.1469-8137.2007.02293.x

Michiels, T., and Cornelis, G. R. (1991). Secretion of hybrid proteins by the yersinia yop export system. *J. Bacteriol.* 173, 1677–1685. doi: 10.1128/jb.173.5.1677-1685.1991

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems.* 31111–33119.

Niemann, G. S., Brown, R. N., Gustin, J. K., Stufkens, A., Shaikh-Kidwai, A. S., Li, J., et al. (2011). Discovery of novel secreted virulence factors from salmonella enterica serovar typhimurium by proteomic analysis of culture supernatants. *Infect. Immun.* 79, 33–43. doi: 10.1128/IAI.00771-10

Nissan, G., Gershovits, M., Morozov, M., Chalupowicz, L., Sessa, G., Manulis-Sasson, S., et al. (2018). Revealing the inventory of type III effectors in pantoea agglomerans gall-forming pathovars using draft genome sequences and a machine-learning approach. *Mol. Plant Pathol.* 19, 381–392. doi: 10.1111/mpp.12528

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., et al. (2019). fairseq: A Fast, Extensible Toolkit for Sequence Modeling. Association for Computational Linguistics (ACL). doi: 10.18653/v1/N19-4009

Parsot, C., Hamiaux, C., and Page, A. L. (2003). The various and varying roles of specific chaperones in type III secretion systems. *Curr. Opin. Microbiol.* 6, 7–14. doi: 10.1016/S1369-5274(02)00002-4

Peabody, M. A., Lau, W. Y. V., Hoad, G. R., Jia, B., Maguire, F., Gray, K. L., et al. (2020). PSORTm: a bacterial and archaeal protein subcellular localization prediction tool for metagenomics data. *Bioinformatics* 36, 3043–3048. doi: 10.1093/BIOINFORMATICS/BTAA136

Peeters, N., Carrère, S., Anisimova, M., Plener, L., Cazalé, A. C., and Genin, S. (2013). Repertoire, unified nomenclature and evolution of the type III effector gene set in the ralstonia solanacearum species complex. *BMC Genomics* 14, 1–19. doi: 10.1186/1471-2164-14-859/FIGURES/5

Rabiner, L. R. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. Proceeding IEEE 77, 257–286. doi: 10.1109/5.18626

Rao, R., Liu, J., Verkuil, R., Meier, J., Canny, J. F., Abbeel, P, et al. (2021). MSA Transformer. *bioRxiv* 139, 8844–8856. doi: 10.1101/2021.02.12.430858

Rao, R., Meier, J., Sercu, T., Ovchinnikov, S., and Rives, A. (2020). Transformer protein language models are unsupervised structure learners. *bioRxiv*. doi: 10.1101/2020.12.15.422761

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., et al. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U. S. A.* 118, e2016239118. doi: 10.1073/pnas.2016239118

Ruano-Gallego, D., Sanchez-Garrido, J., Kozik, Z., Núñez-Berrueco, E., Cepeda-Molero, M., Mullineaux-Sanders, C., et al. (2021). Type III secretion system effectors form robust and flexible intracellular virulence networks. *Science* 371, eabc9531. doi: 10.1126/science.abc9531

Samudrala, R., Heffron, F., and McDermott, J. E. (2009). Accurate prediction of secreted substrates and identification of a conserved putative secretion signal for type III secretion systems. *PloS Pathog.* 5, e1000375. doi: 10.1371/journal.ppat.1000375

Sato, Y., Takaya, A., and Yamamoto, T. (2011). Meta-analytic approach to the accurate prediction of secreted virulence effectors in gram-negative bacteria. *BMC Bioinf.* 12, 1–12. doi: 10.1186/1471-2105-12-442

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv Prepr.* Available at: https://arxiv.org/abs/1409.1556.

Sory, M.-P., and Cornelis, G. R. (1994). Translocation of a hybrid YopE-adenylate cyclase from yersinia enterocolitica into HeLa cells. *Mol. Microbiol.* 14, 583–594. doi: 10.1111/j.1365-2958.1994.tb02191.x

Sperschneider, J., Dodds, P. N., Gardiner, D. M., Singh, K. B., and Taylor, J. M. (2018). Improved prediction of fungal effector proteins from secretomes with EffectorP 2.0. *Mol. Plant Pathol.* 19, 2094–2110. doi: 10.1111/MPP.12682

Stebbins, C. E., and Galaán, J. E. (2001). Structural mimicry in bacterial virulence. *Nature* 412, 77–81. doi: 10.1038/35089000

Tay, D. M. M., Govindarajan, K. R., Khan, A. M., Ong, T. Y. R., Samad, H. M., Soh, W. W., et al. (2010). T3SEdb: Data warehousing of virulence effectors secreted by the bacterial type III secretion system. *BMC Bioinf.* 11, 1–7. doi: 10.1186/1471-2105-11-S7-S4

Teper, D., Burstein, D., Salomon, D., Gershovitz, M., Pupko, T., and Sessa, G. (2016). Identification of novel xanthomonas euvesicatoria type III effector proteins by a machine-learning approach. *Mol. Plant Pathol.* 17, 398–411. doi: 10.1111/mpp.12288

Trotter, M. V., Nguyen, C. Q., Young, S., Woodruff, R. T., and Branson, K. M. (2021). Epigenomic language models powered by cerebras. *arXiv. Prepr.* doi: 10.48550/arXiv.2112.07571

Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., et al. (2017). "Attention is All you Need," in *Advances in Neural Information Processing Systems.* 30, 6000–6010.

Wagner, N., Avram, O., Gold-Binshtok, D., Zerah, B., Teper, D., and Pupko, T. (2022). Effectidor: an automated machine-learning-based web server for the prediction of type-III secretion system effectors. *Bioinformatics* 38, 2341–2343. doi: 10.1093/bioinformatics/btac087

Wang, J., Li, J., Yang, B., Xie, R., Marquez-Lago, T. T., Leier, A., et al. (2019). Bastion3: a two-layer ensemble predictor of type III secreted effectors. *Bioinformatics* 35, 2017–2028. doi: 10.1093/bioinformatics/bty914

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

Wang, Y., Sun, M., Bao, H., Zhang, Q., and Guo, D. (2013b). Effective identification of bacterial type III secretion signals using joint element features. *PloS One* 8, e59754. doi: 10.1371/journal.pone.0059754

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

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

Yang, Y. (2012). Identification of novel type III effectors using latent dirichlet allocation. *Comput. Math. Methods Med.* 2012, 696190. doi: 10.1155/2012/696190

Yang, L., Guell, M., Byrne, S., Yang, J. L., De Los Angeles, A., Mali, P., et al. (2013). Optimization of scarless human stem cell genome editing. *Nucleic Acids Res.* 41, 9049–9061. doi: 10.1093/NAR/GKT555

Yang, X., Guo, Y., Luo, J., Pu, X., and Li, M. (2013). Effective identification of gram-negative bacterial type III secreted effectors using position-specific residue conservation profiles. *PloS One* 8, e84439. doi: 10.1371/journal.pone.0084439

Yang, Y., Zhao, J., Morgan, R. L., Ma, W., and Jiang, T. (2010). Computational prediction of type III secreted proteins from gram-negative bacteria. *BMC Bioinf.* 11, 1–10. doi: 10.1186/1471-2105-11-S1-S47

Yu, L., Liu, F., Li, Y., Luo, J., and Jing, R. (2021). DeepT3_4: A hybrid deep neural network model for the distinction between bacterial type III and IV secreted effectors. *Front. Microbiol.* 12. doi: 10.3389/FMICB.2021.605782

Zou, L., Nan, C., Hu, F., and Hancock, J. (2013). Accurate prediction of bacterial type IV secreted effectors using amino acid composition and PSSM profiles. *Bioinformatics* 29, 3135–3142. doi: 10.1093/bioinformatics/btt554