PLOS ONE

RESEARCH ARTICLE

Using an optimal set of features with a machine learning-based approach to predict effector proteins for *Legionella pneumophila*

Zhila Esna Ashari[1]*, **Kelly A. Brayton**[1,2,3☯], **Shira L. Broschat**[1,2,3☯]

1 School of Electrical Engineering and Computer Science, Washington State University, Pullman, Washington, United States of America, **2** Department of Veterinary Microbiology and Pathology, Washington State University, Pullman, Washington, United States of America, **3** Paul G. Allen School for Global Animal Health, Washington State University, Pullman, Washington, United States of America

☯ These authors contributed equally to this work.
* z.esnaashariesfahan@wsu.edu

Check for updates

OPEN ACCESS

Citation: Esna Ashari Z, Brayton KA, Broschat SL (2019) Using an optimal set of features with a machine learning-based approach to predict effector proteins for *Legionella pneumophila*. PLoS ONE 14(1): e0202312. https://doi.org/10.1371/journal.pone.0202312

Editor: Seyedali Mirjalili, Griffith University, AUSTRALIA

Received: July 17, 2018

Accepted: January 12, 2019

Published: January 25, 2019

Copyright: © 2019 Esna Ashari et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by grant R01AI042792 by National Institutes of Health (K.A. B.) and by the Carl M. Hansen Foundation (Z.E.). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. No additional external funding was received for this study.

Abstract

Type IV secretion systems exist in a number of bacterial pathogens and are used to secrete effector proteins directly into host cells in order to change their environment making the environment hospitable for the bacteria. In recent years, several machine learning algorithms have been developed to predict effector proteins, potentially facilitating experimental verification. However, inconsistencies exist between their results. Previously we analysed the disparate sets of predictive features used in these algorithms to determine an optimal set of 370 features for effector prediction. This study focuses on the best way to use these optimal features by designing three machine learning classifiers, comparing our results with those of others, and obtaining de novo results. We chose the pathogen *Legionella pneumophila* strain Philadelphia-1, a cause of Legionnaires' disease, because it has many validated effector proteins and others have developed machine learning prediction tools for it. While all of our models give good results indicating that our optimal features are quite robust, Model 1, which uses all 370 features with a support vector machine, has slightly better accuracy. Moreover, Model 1 predicted 472 effector proteins that are deemed highly probable to be effectors and include 94% of known effectors. Although the results of our three models agree well with those of other researchers, their models only predicted 126 and 311 candidate effectors.

Introduction

Bacterial pathogens can use secretion systems to deliver proteins to the host cell. There are nine known secretion systems, but the focus of this study is on the type IV secretion system (T4SS). The T4SS is composed of multiple proteins responsible for secreting effector proteins directly into eukaryotic host cells. When effector proteins are translocated into host cells, they manipulate their defence systems, causing infections. In order to understand how these effector proteins manipulate the host cell, it is first necessary to identify them. However, this can be

Competing interests: The authors have declared that no competing interests exist.

a difficult task because they are not well conserved among organisms. Several methods have been proposed for identifying effector proteins with experimental validation being the most accurate but also the most expensive and time consuming [1–4]. Accurate prediction of candidate effectors would expedite the experimental validation process. As a result, recent studies have focused on using prediction approaches such as scoring effector proteins based on their characteristics or using machine learning algorithms [5–11]. Several studies have reviewed the existing methods for predicting effector proteins: Zeng et al. focused on the progress made in the field of effector prediction for different types of secretion systems, including the T4SS, and studied the features used [12]; An et al. reviewed the methods and tools developed for prediction of type III, IV, and VI effector proteins [13] and introduced several ensemble approaches for identifying T4SS effectors by integrating results from several predictors; and McDermott et al. reviewed recent methodologies and studied features for predicting both type III and IV secretion system effectors [14] while Wang et al. tested a variety of well-known T4SS classifiers over a range of sequence-derived features and developed Bastion4 as a result [11]. In addition, several previous studies focused on creating databases of validated effectors to facilitate future research involving effector proteins for different species, which helped us create our own dataset [15, 16]. Because prior methods considered different sets of features, we examined their effectiveness in an earlier study and determined a set of optimal features for prediction of T4SS effector proteins [17–18]. By features, we refer here to the characteristics and properties of protein sequences that can be measured and thus assigned binary or continuous numerical values.

In our previous study, we identified a set of optimal features using four datasets of validated effector and non-effector proteins from four different Proteobacterial pathogens, *Legionella pneumophila*, *Coxiella burnettii*, *Bartonella* spp., and *Brucella* spp. that works well for prediction of T4SS effector proteins. In this study, we use this set of optimal features to develop a machine learning based classifier to predict T4SS effectors, which is trained using the set of validated effector and non-effector proteins from our earlier study of all four pathogens. Our goals are four-fold: i) to test our classifier on a pathogen with many validated effectors to ascertain how well it works for a single pathogen, ii) to determine the best way to use the optimal features to achieve the most accurate results, iii) to compare our results with those of other T4SS effector prediction models, and iv) to obtain de novo results. Therefore, we selected the *L. pneumophila* strain Philadelphia-1 genome/deduced proteome as the subject of our study because it has the greatest number of validated effector proteins, and several prediction algorithms have used this organism as their subject. *L. pneumophila* is a Gram-negative bacterial pathogen from the class Gammaproteobacteria which causes Legionnaires' disease, and many studies have focused on this pathogen and its effector proteins [19–33].

To analyze our optimal features, we actually developed three different machine learning classifiers. We first explain how we design and validate our three machine learning models, two of which are ensemble classifiers. Next, we use the models on the whole proteome from *L. pneumophila* strain Philadelphia-1 and compare our results with those of previous studies for *L. pneumophila*. Finally, we obtain de novo predictions of effector proteins for *L. pneumophila*.

Materials and methods

Fig 1 represents the workflow used to complete this study. Each step is described in more detail in subsequent sections.

Creating training and test datasets

Our training dataset was composed of effectors and non-effectors from four different bacterial pathogens: *L. pneumophila*, *C. burnettii*, *Brucella* spp., and *Bartonella* spp. In our previous

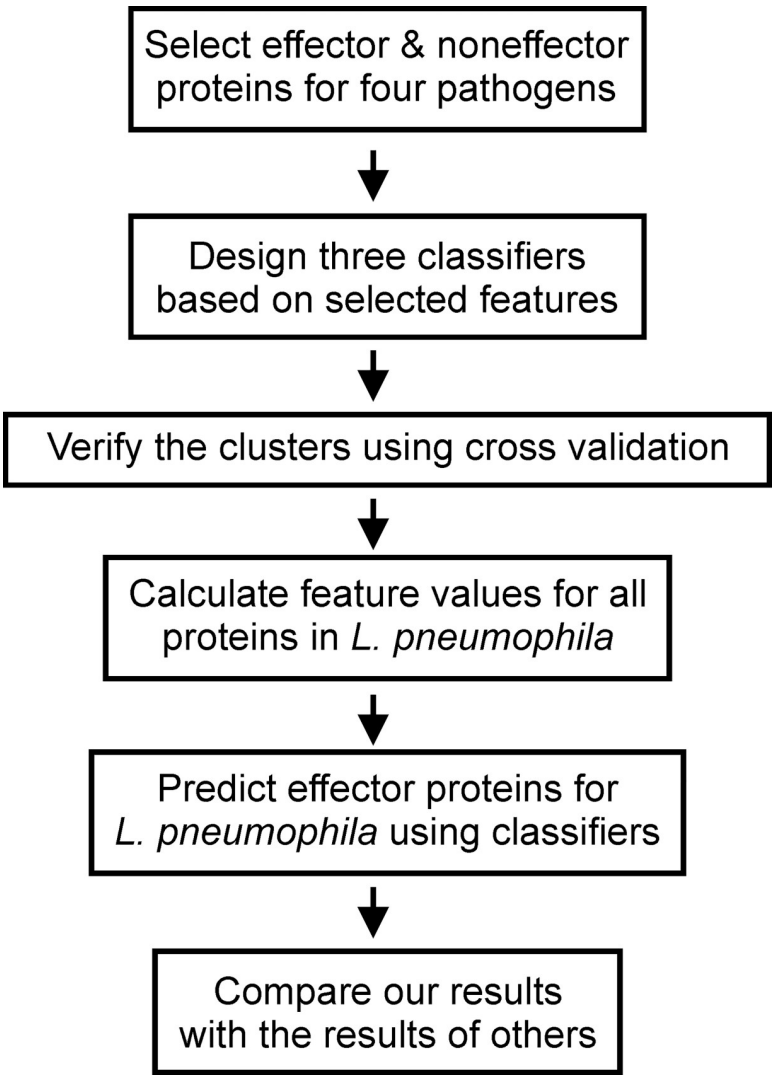

Fig 1. Workflow.

https://doi.org/10.1371/journal.pone.0202312.g001

paper, each of these pathogens was treated as a separate dataset [18], and we determined effective features for each using a feature selection method. Based on our results, we proposed a final set of effective features for prediction of T4SS effectors. In the present study we merged these four datasets to create a set of known effectors and non-effectors which was used as the training set for our problem. This dataset consisted of 1,127 data points among which there were 429 effectors and 698 non-effectors. The protein sequences for our training dataset are presented in S1 File. We also created a test set, which is composed of 2,942 protein sequences from the complete proteome of *L. pneumophila* strain Philadelphia-1 S2 File.

Features

The features used in this study are the set of optimal features proposed in our earlier work [18]. In our previous study we did a comprehensive literature review and compiled a list of all the features used for prediction of T4SS effector proteins. Because some of the features were vectors, we began with 1,027 features. By vector, we mean that a particular feature had multiple

values. For example, there are 20 different amino acids so that the amino acid composition feature for a protein sequence has 20 different percentage values. Using a multi-level feature selection approach, we proposed a set of optimal features for our prediction problem and retained 370 features. Overall, they include chemical properties, structural properties, compositional properties, and position-specific scoring matrix (PSSM)-related properties, which are a type of compositional property.

Our optimal feature set includes 15 features that are related to the chemical and structural properties of protein sequences. Chemical properties such as hydropathy are considered to be important for T4SS effector prediction because they determine how proteins interact with their environment and because they are believed to be key mediators in determining how effectors enter host cells [6, 8]. The structural properties of proteins, such as coiled coil domains, allow protein-protein interactions within host cells thus effecting cellular processes [6, 8–9]. Our feature set also includes compositional properties of protein sequences, comprising selected elements of the amino acid and dipeptide composition vectors totalling 57 in number. In addition, they include 298 features from the PSSM profile for protein sequences and its auto-covariance correlation composition vector [34]. Compositional properties are considered to be effective for T4SS effector prediction because they determine the shape of the protein, and they also account for amino acid frequencies and motifs [7]. The effectiveness of PSSM-related features are described in other studies as well [35, 36]. Wang et al. have provided a tool to produce a variety of features based on PSSM profiles of protein sequences [37], and some of the features derived from these may also be helpful for predicting T4SS effector proteins.

All features are explained at greater length in [18].

Machine learning models and validation

A major goal of this paper was to determine how to use the optimal feature set to obtain the most accurate results. As such, we considered different methodologies and algorithms, for example, using a single classifier versus an ensemble classifier, and decided to design three separate models based on a division of the features. To test our classifiers, we used several standard metrics for machine learning models: accuracy, recall, precision, and the Matthews Correlation Coefficient (MCC).

Our first model, Model 1, was based on the use of the entire optimal feature set. We calculated the features for all the protein sequences in our dataset of effectors and non-effectors. These 370 features are shown in S1 Table. We used this dataset to train a support vector machine (SVM) classifier. An SVM is a powerful machine learning classifier often used for supervised learning, that is learning based on using labelled training data [38]. It allows the use of different Kernel functions to create classifiers that fit a dataset. Our second and third models, Models 2 and 3, were ensemble classifiers composed of three separate classifiers. Each of these classifiers was designed to work with a subset of the optimal feature set. By dividing the features among several classifiers, we wanted to decrease the possibility of overfitting effects on our results. Overfitting occurs when a model fits training data too well, causing the model to be less accurate for new data. Here, we chose three SVM classifiers for each ensemble model and with all redundant and highly correlated features removed; each of three SVM classifiers determines whether a protein sequence was an effector protein or a non-effector protein. The final prediction was based on the output class that had the majority of votes from all three classifiers. When two or more classifiers voted for a protein sequence to be an effector, it was predicted to be an effector protein. We used the SVM tuning function in R to find the best parameters for our SVM classifiers which resulted in the use of a radial Kernel and a C parameter of 1 [39].

As mentioned, Model 1 used all the selected features. For our first ensemble classifier, Model 2, the three groups of features were divided among our three classifiers as follows: i) features related to PSSM composition, ii) features related to the auto-covariance correlation of PSSM, and iii) chemical, structural, and compositional features S1 Table (e.g., amino acid composition, dipeptide composition, average hydropathy, total hydropathy, hydropathy of C terminal, hydropathy of N terminal, number of coiled coil regions, signal peptide probability, polarity, molecular mass, length, and homology to known effectors). For our second ensemble classifier, Model 3, the three groups of features divided among our classifiers were as follows: i) PSSM-related features (PSSM composition and auto covariance correlation of PSSM), ii) features related to the composition of amino acids in protein sequences (amino acid composition and dipeptide composition), and iii) chemical and structural features (average hydropathy, total hydropathy, hydropathy of C terminal, hydropathy of N terminal, number of coiled coil regions, signal peptide probability, polarity, molecular mass, length, and homology to known effectors).

After building our dataset and designing our machine learning classifiers, we used 10-fold cross-validation to validate our models and to test for overfitting in the results. The dataset was randomly divided into ten groups, and for each fold, one group was kept for testing and the other nine groups were used for training. We calculated confusion matrices for each cross-validation step for all three models. A confusion matrix is a table that displays the results of a machine learning algorithm for known test data. When a positive value (here an effector protein) is correctly identified, it is called a true positive (TP); when a negative value (here a non-effector protein) is correctly identified, it is called a true negative (TN); when a positive value is identified as a negative value, it is called a false negative (FN); and when a negative value is identified as a positive value, it is called a false positive (FP). From the confusion matrices, we calculated accuracy measures for the models. The final accuracy for the models was obtained by taking the average of the ten different folds. In addition, because the number of effectors (429) and non-effectors (698) in our dataset was not the same, we calculated recall and precision. Recall is a measure of sensitivity, and precision is a measure of relevance. When these values are sufficiently high, it indicates that our results are not affected by the unbalanced dataset. Finally, we calculated the MCC values for our models as another means of determining their accuracy. The MCC is a measure of correlation between real and predicted values. The equations for accuracy, recall, precision, and MCC are presented in (1)–(4) [40].

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{4}$$

To compare their performance visually, we plotted Receiver Operating Characteristic (ROC) curves for 10 folds of each model. An ROC curve demonstrates the True Positive rate versus False Positive rate of a model when the threshold for discrimination of two output classes is varied. We also presented the average Area Under the Curve (AUC) for ROC plots of 10 folds for further comparison of the models.

Table 1. Accuracy measures for 10-fold cross-validation of Model 1 using the entire feature set for prediction.

	Fold Accuracy (%)									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Model 1	95.13	93.80	93.75	92.47	93.75	93.36	95.08	95.13	95.11	92.92

https://doi.org/10.1371/journal.pone.0202312.t001

The next step after designing and validating our models was to use them for predicting effector proteins in the whole proteome of *L. pneumophila* strain Philadelphia-1. This proteome contains 2,942 protein sequences and was used as our test set S2 File. We calculated the feature values for all the protein sequences in *L. pneumophila* using different tools and programming languages as described in [11]. We then used our three models for de novo prediction of effector proteins in the *L. pneumophila* proteome. Models 2 and 3 each consisted of 3 separate classifiers with each classifier determining whether one of the 2,942 *L. pneumophila* protein sequences was an effector or non-effector. Protein sequences receiving two or three positive votes were predicted as effectors.

The final step in this study was to compare our results to those obtained previously by others for prediction of effector proteins for *L. pneumophila*. We selected the study performed by Burstein et al. in 2009 which used a voting scheme based on four different algorithms [5] and the study performed by Meyer et al. in 2013 which used a scoring method [6]. Results and comparisons are discussed in the next section.

Results and discussion

We developed three models to test the accuracy of our optimal feature set. Model 1 used the entire set of 370 features with an SVM, and Models 2 and 3 also used the entire set of features. However, they were divided into subsets and used with three separate SVM classifiers comprising ensemble models. We used 10-fold cross-validation to test these models. The accuracy results calculated for each of the 10 folds are shown in Tables 1 through 3 for Models 1 through 3, respectively.

The final accuracy for each model is obtained from the average of the ten values, and these are given in the first line of Table 4.

The three values are 94.05%, 93.64%, and 92.44%, for Models 1, 2, and 3, respectively. These values are close indicating the accuracy of all three models.

As described earlier, we calculated recall and precision for our three models to ensure that the overbalanced training data did not affect the results and also as another means of validating our results. Average values for the three models are presented in Table 4 where even the lowest value of 87.33% for the average precision value for Model 3 is still very good. All other results are above 90% and indicate both that the overbalanced training data did not affect the machine learning results and that the results for all three models are very good. This is further supported by the values for average MCC and AUC presented in Table 4, which demonstrate good performance for all three models with Model 1 showing the best performance. Also, the corresponding ROC curves for all three models for 10 folds are shown in Fig 2 confirming the

Table 2. Accuracy measures for 10-fold cross-validation of Model 2 using three feature subsets. i) PSSM composition features, ii) PSSM auto-covariance correlation features, and iii) chemical, structural, and compositional features.

	Fold Accuracy (%)									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Model 2	93.36	93.36	95.53	92.47	93.74	92.44	93.30	95.13	93.30	93.80

https://doi.org/10.1371/journal.pone.0202312.t002

Table 3. Accuracy measures for 10-fold cross-validation of Model 3 using three feature subsets. i) PSSM-related features, ii) compositional features, and iii) chemical and structural features.

	Fold Accuracy (%)									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Model 3	90.70	91.59	92.41	91.59	94.64	92.92	93.30	90.13	93.30	93.80

https://doi.org/10.1371/journal.pone.0202312.t003

results based on the average AUC. As can be seen in this figure, results from Model 1 are the most consistent.

The next step was using our three designed classifiers on the whole proteome of *L. pneumophila* strain Philadelphia-1 to predict effector proteins with results presented in Table 5.

The number of predicted effectors is shown in the second column of Table 5. The greatest number of effectors is 760 predicted by Model 1 followed closely by 717 predicted by Model 2. Model 3 predicts 568, considerably fewer and to our knowledge, effector predictions for the three models are greater in number than any previous study for *L. pneumophila* strain Philadelphia-1. As another test of the accuracy of our models, we considered the validated effectors and non-effectors for *L. pneumophila* strain Philadelphia-1 to see which of them were predicted correctly from the test set. These results are shown in the third and fourth columns of Table 5. The lowest of the six results is 94.9% again indicating the overall accuracy of the three models. Model 1 predicts 315 of the 316 validated effector proteins correctly for an accuracy of 99.7%, and Model 3 predicts 521 of 526 non-effector proteins correctly for an accuracy of 99.0%.

We compared our results to effector candidates predicted in two previous studies [5, 6] that focused on *L. pneumophila* strain Philadelphia-1. The first by Burstein et al. experimentally validated 40 new effector proteins and also proposed 126 effector candidates. The second by Meyer et al. proposed 311 candidate effector proteins. These two sets of predicted results shared 45 protein sequences in common, which is 36% of the predicted sequences in [5] and 14% of the predicted sequences in [6]. Our three model comparisons are shown in the fifth and sixth columns of Table 5, and a Venn diagram of the number of candidate effector proteins predicted by Model 1, by Burstein et al. [5], and by Meyer et al. [6] is shown in Fig 3. Model 1 shares 101 of 126 or 80.2% in common with [5] and 273 of 302 or 90.4% in common with [6] (after removing known non-effectors from their candidates). Interestingly, as shown in Fig 3, Model 1 also predicted all 45 protein sequences shared by [5] and [6] and also predicted all the 40 new validated effector proteins by [5].

While all three models give good results, the overall results presented in this section indicate that Model 1 is the strongest of the three models. The accuracy metric is the highest, but in addition three of the fold values are above 95%. Recall, precision, and MMC are most consistent, and comparison with results from previous studies is strongest. The candidate effector proteins for *L. pneumophila* are listed in S2 Table. They are also listed in three groups based on

Table 4. Average accuracy, recall, precision, MCC, and AUC measures over 10 folds for the three effector prediction models.

	Model 1	**Model 2**	**Model 3**
Average accuracy	94.05%	93.64%	92.44%
Average recall	92.00%	93.06%	92.83%
Average precision	92.49%	90.91%	87.33%
Average MCC	0.87	0.86	0.84
Average AUC	0.983	0.979	0.970

https://doi.org/10.1371/journal.pone.0202312.t004

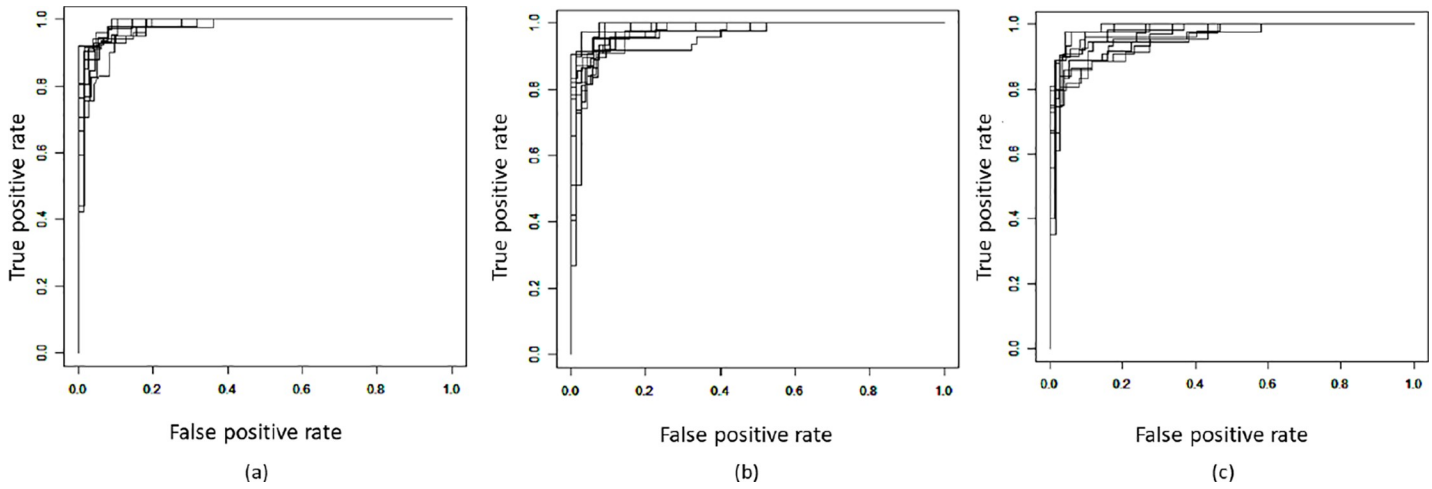

Fig 2. ROC curves for three designed classifiers for 10-fold, cross-validation results. (a) Model 1, (b) Model 2, and (c) Model 3.

https://doi.org/10.1371/journal.pone.0202312.g002

the results of the other two models and after removing known non-effectors. If predicted by all three models, they are listed in Group 1, by two models in Group 2, and by Model 1 only in Group 3. We assume the first group of 472 has the greatest likelihood of being an effector, the second group of 167 the next most likelihood, and the third group of 107 the next most. Table 6 represents the statistics for Group 1 sequences, which are most likely to be effectors. Interestingly, while the statistics are still excellent, they are slightly lower than for Model 1 prior to grouping.

Given the differences shown in Fig 3 and Table 5, we conclude that the features used in machine learning predictors are of major importance. More specifically, the reason we predicted more effectors and have more consistent results with previous studies is related to the set of optimal features that we used. This feature set was based on a thorough study of features for the problem of T4SS effector prediction [11, 12]. As the two previous studies developed their models based on a subset of the optimal features, it is likely that they were not able to capture as many effectors. They also had fewer validated effector proteins with which to work compared to the number available to us.

Conclusion

In this study, we designed three machine learning classifiers using an optimal set of features and used these classifiers to obtain de novo predictions for effector proteins for *L. pneumophila* strain Philadelphia-1. While all three models were accurate, we found that the strongest model was a straightforward classifier that used all 370 features with a support vector machine. The accuracy, recall, and precision for this model validation, were all greater than 90%. The results

Table 5. Comparison of results for the three effector prediction models for *L. pneumophila* strain Philadelphia-1.

	Number of predicted effector proteins	Number of correctly predicted known:		Number of effectors predicted by our models among results for:	
		Effectors (316)	Non-effectors (526)	S4TE (302)	Burstein et al. (126)
Model 1	760	315 (99.7%)	514 (97.7%)	273 (90.4%)	101 (80.2%)
Model 2	717	300 (94.9%)	518 (98.5%)	253 (83.8%)	100 (79.4%)
Model 3	568	306 (96.8%)	521 (99.0%)	258 (85.4%)	97 (77.0%)

https://doi.org/10.1371/journal.pone.0202312.t005

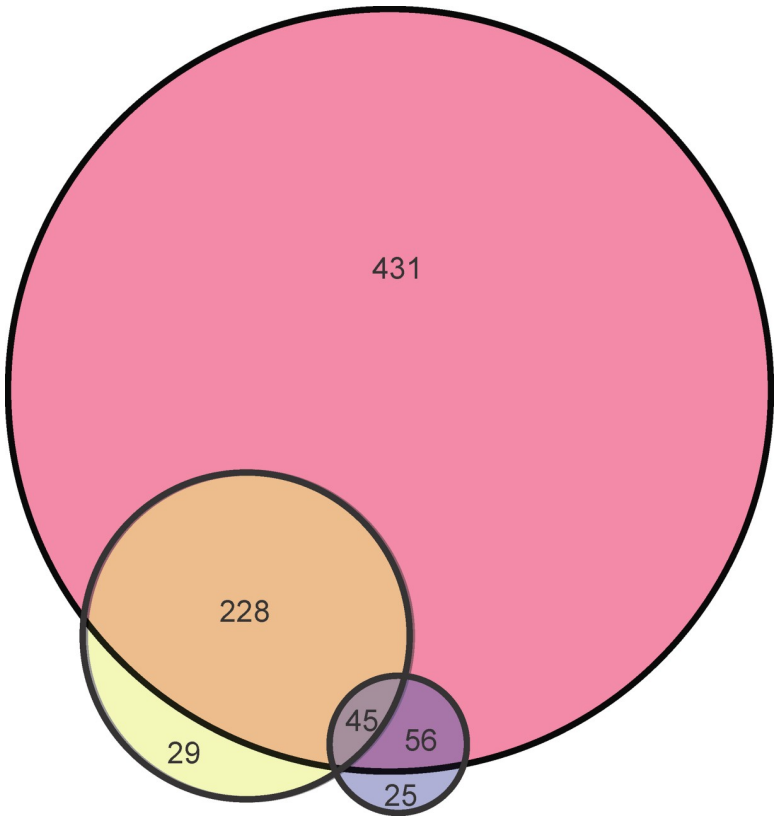

Fig 3. Venn diagram comparing predicted effector proteins for three methods. The pink circle shows the results for Model 1, the yellow circle for the S4TE method, and the blue circle for the method by Burstein et al.

https://doi.org/10.1371/journal.pone.0202312.g003

Table 6. Comparison of results for the most probable group of candidate effectors by Model 1 for *L. pneumophila* strain Philadelphia-1.

	Number of predicted effector proteins	Number of correctly predicted known:		Number of effectors predicted by our models among results for:	
		Effectors (316)	Non-effectors (526)	S4TE (302)	Burstein et al. (91)
Model 1- Group 1	472	297 (93.7%)	525 (99.8%)	243 (80.5%)	101 (72.2%)

https://doi.org/10.1371/journal.pone.0202312.t006

of this model compared well with those obtained from two previous research studies predicting more than 80% of the same candidate effector proteins that they did. However, while these older models predicted 126 and 311 candidate effector proteins, our model predicted 472 effector proteins that are deemed most probable of being effectors which is more than other models. The reason for these prediction results and consistency with previous predictions, is due to the optimal set of features used.

Supporting information

S1 File. Training set composed of known effectors and non-effectors for *L. pneumophila*, *C. burnettii*, *Brucella* spp., and *Bartonella* spp.
(FASTA)

S2 File. Test set composed of all protein sequences from the whole proteome for *L. pneumophila* strain Philadelphia-1.
(FASTA)

S1 Table. The 370 features used in the three machine learning models developed for this study.
(XLSX)

S2 Table. The set of de novo effector proteins predicted by Model 1 for *L. pneumophila* strain Philadelphia-1.
(XLSX)

Author Contributions

Conceptualization: Kelly A. Brayton, Shira L. Broschat.

Data curation: Zhila Esna Ashari.

Formal analysis: Zhila Esna Ashari.

Funding acquisition: Kelly A. Brayton, Shira L. Broschat.

Investigation: Zhila Esna Ashari.

Methodology: Zhila Esna Ashari.

Software: Zhila Esna Ashari.

Supervision: Kelly A. Brayton, Shira L. Broschat.

Validation: Zhila Esna Ashari.

Writing – original draft: Zhila Esna Ashari.

Writing – review & editing: Kelly A. Brayton, Shira L. Broschat.

References

1. Han N, Yu W, Qiang Y, Zhang W. T4SP Database 2.0: An Improved Database for Type IV Secretion Systems in Bacterial Genomes with New Online Analysis Tools. Computational and Mathematical Methods in Medicine. 2016; 2016, 9415459. (https://doi.org/10.1155/2016/9415459) PMID: 27738451
2. Voth DE, Broederdorf LJ, Graham JG. Bacterial Type IV Secretion Systems: Versatile Virulence Machines. Future Microbiology. 2012; 7(2), 241–257. (https://doi.org/10.2217/fmb.11.150) PMID: 22324993
3. Voth DE, Beare PA, Howe D, Sharma UM, Samoilis G, Cockrell DC, et al. The *Coxiella burnetii* Cryptic Plasmid Is Enriched in Genes Encoding Type IV Secretion System Substrate. Journal of Bacteriology. 2010; 193(7), 1493–1503. (https://doi.org/10.1128/JB.01359-10) PMID: 21216993
4. Abby SS, Cury J, Guglielmini J, Néron B, Touchon M, Rocha EPC. Identification of protein secretion systems in bacterial genomes. Scientific Reports. 2016; 6. (https://doi.org/10.1038/srep23080). PMID: 26979785
5. Burstein D, Zusman T, Degtyar E, Viner R, Segal G, Pupko T. Genome-Scale Identification of *Legionella pneumophila* Effectors Using a Machine Learning Approach. The International Journal of Biochemistry and Cell Biology. 2009; 5(7). (https://doi.org/10.1371/journal.ppat.1000508)
6. Meyer DF, Noroy C, Moumene A, Raffaele S, Albina E, Vachiery N. Searching algorithm for type IV secretion system effectors 1.0: a tool for predicting type IV effectors and exploring their genomic context. Nucleic Acids Research. 2013; 41(20), 9218–9229. (https://doi.org/10.1093/nar/gkt718) PMID: 23945940
7. Zou L, Nan C, Hu F. 2013. Accurate prediction of bacterial type IV secreted effectors using amino acid composition and PSSM profiles. Bioinformatics 29(24), 3135–3142. (https://doi.org/10.1093/bioinformatics/btt554) PMID: 24064423

8. Yu L, Guo Y, Li Y, Li G, Li M, Luo J, et al. 2013. SecretP: identifying bacterial secreted proteins by fusing new features into Chou's pseudo-amino acid composition. J Theor Biol. 267, 1–6. (https://doi.org/10.1016/j.jtbi.2010.08.001) PMID: 20691704
9. Wang Y, Wei X, Bao H, Liu S. Prediction of bacterial type IV secreted effectors by C-terminal features. BMC Genomics 2014; 15(50). (https://doi.org/10.1186/1471-2164-15-50) PMID: 24447430
10. Lockwood S, Voth DE, Brayton KA, Beare PA, Brown WC, Heinzen RA, et al. Identification of *Anaplasma marginale* Type IV Secretion System Effector Proteins. PLoS ONE. 2011; 6(11), e27724. (https://doi.org/10.1371/journal.pone.0027724) PMID: 22140462
11. Wang J, Yang B, An Y, Marquez-Lago T, Leier A, Wilksch J, et al. Systematic analysis and prediction of type IV secreted effector proteins by machine learning approaches. Briefings in Bioinformatics 2017; 1–21. (https://doi.org/10.1093/bib/bbw003)
12. Zeng C, Zou L. An account of in silico identification tools of secreted effector proteins in bacteria and future challenges. Briefings in Bioinformatics 2017; 1–20. (https://doi.org/10.1093/bib/bbw003)
13. An Y, Wang J, Li C, Leier A, Marquez-Lago T, Wilksch J, et al. Comprehensive assessment and performance improvement of effector protein predictors for bacterial secretion systems III, IV and VI. Briefings in Bioinformatics 2018; 19(1), 148–161. (https://doi.org/10.1093/bib/bbw100) PMID: 27777222
14. McDermott JE, Corrigan A, Peterson E, Oehmen C, Niemann G, Cambronne ED, et al. Computational Prediction of Type III and IV Secreted Effectors in Gram-Negative Bacteria. Infection and Immunity 2011; 79(1), 23–32. (https://doi.org/10.1128/IAI.00537-10) PMID: 20974833
15. Bi D, Liu L, Tai C, Deng Z, Rajakumar K, Ou HY. SecReT4: a web-based bacterial type IV secretion system resource. Nucleic Acids Research 2013; 41. (https://doi.org/10.1093/nar/gks1248) PMID: 23193298
16. An Y, Wang J, Li C, Revote J, Zhang Y, Naderer T, et al. SecretEPDB: a comprehensive web-based resource for secreted effector proteins of the bacterial types III, IV and VI secretion systems. Scientific Reports 2017; 7(41031). (https://doi.org/10.1038/srep41031) PMID: 28112271
17. Esna Ashari Z, Brayton KA, Broschat SL. Determining Optimal Features for Predicting Type IV Secretion System Effector Proteins for Coxiella burnetii. Proceedings of 8th ACM BCB conference. 2017; 346–351.
18. Esna Ashari Z, Dasgupta N, Brayton KA, Broschat SL. 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 2018; 13(5): e0197041. (https://doi.org/10.1371/journal.pone.0197041) PMID: 29742157
19. Bruggemann H, Cazalet C, Buchrieser C. Adaptation of *Legionella pneumophila* to the host environment: role of protein secretion, effectors and eukaryotic-like proteins. Current Opinion in Microbiology. 2006; 9(1), 86–94. https://doi.org/10.1016/j.mib.2005.12.009 PMID: 16406773
20. Cazalet C, Rusniok R, Bruggemann H, Zidane N, Magnier A, Ma L, et al. Evidence in the *Legionella pneumophila* genome for exploitation of host cell functions and high genome plasticity. Nature Genetics. 2004; 36(11), 1165–1173. https://doi.org/10.1038/ng1447 PMID: 15467720
21. Chen J, De Felipe KS, Clarke M, Lu H, Anderson OR, Segal G, et al. *Legionella* Effectors That Promote Nonlytic Release from Protozoa. Science. 2004; 303(5662), 1358–1361. (https://doi.org/10.1126/science.1094226) PMID: 14988561
22. De Felipe KS, Pampou S, Jovanovic OS, Pericone CD, Ye SF, Kalachikov S, et al. Evidence for Acquisition of Legionella Type IV Secretion Substrates via Interdomain Horizontal Gene Transfer. Journal of Bacteriology. 2005; 187(22), 7716–7726. https://doi.org/10.1128/JB.187.22.7716-7726.2005 PMID: 16267296
23. Conover G, Derre I, Vogel J, RR I. The *Legionella pneumophila* LidA protein: a translocated substrate of the Dot/Icm system associated with maintenance of bacterial integrity. Molecular Microbiology. 2003; 48(2), 305–321. PMID: 12675793
24. Laguna RK, Creasey EA, Li Z, Valtz N, Isberg RR. A *Legionella pneumophila*-translocated substrate that is required for growth within macrophages and protection from host cell death. Proceedings of the National Academy of Sciences. 2006; 103(49), 18745–18750.
25. Bardill JP, Miller JL, Vogel JP. IcmS-dependent translocation of SdeA into macrophages by the *Legionella pneumophila* type IV secretion system. Molecular Microbiology. 2005; 56(1), 90–103. https://doi.org/10.1111/j.1365-2958.2005.04539.x PMID: 15773981
26. Ninio S, Zuckman-Cholon DM, Cambronne ED, Roy CR. The *Legionella* IcmS-IcmW protein complex is important for Dot/Icm-mediated protein translocation. Molecular Microbiology. 2005; 55(3), 912–926. https://doi.org/10.1111/j.1365-2958.2004.04435.x PMID: 15661013

27. Altman E, Segal G. The Response Regulator CpxR Directly Regulates Expression of Several *Legionella pneumophila* icm/dot Components as Well as New Translocated Substrates. Future Microbiology. 2008; 190(6), 1985–1996. (https://doi.org/10.1128/JB.01493-07) PMID: 18192394

28. Zusman T, Aloni G, Halperin E, Kotzer H, Degtyar E, Feldman M, et al. The response regulator PmrA is a major regulator of the icm/dot type IV secretion system in *Legionella pneumophila* and Coxiella burnetii. Molecular Microbiology. 2007; 63(5), 1508–1523. https://doi.org/10.1111/j.1365-2958.2007.05604.x PMID: 17302824

29. Zusman T, Degtyar E, Segal G. Identification of a Hypervariable Region Containing New *Legionella pneumophila* Icm/Dot Translocated Substrates by Using the Conserved icmQ Regulatory Signature. Infection and Immunity. 2008; 76(10), 4581–4591. (https://doi.org/10.1128/IAI.00337-08) PMID: 18694969

30. De Felipe KS, Glover RT, Charpentier X, Anderson OR, Reyes M, Pericone CD, et al. Legionella Eukaryotic-Like Type IV Substrates Interfere with Organelle Trafficking. PLoS Pathogens. 2008; 4(8). (https://doi.org/10.1371/journal.ppat.1000117) PMID: 18670632

31. Heidtman M, Chen EJ, Moy MY, Isberg RR. Large scale identification of *Legionella pneumophila* Dot/Icm substrates that modulate host cell vesicle trafficking pathways. Cellular Microbiology. 2009; 11(2), 230–248. (https://doi.org/10.1111/j.1462-5822.2008.01249.x) PMID: 19016775

32. Shohdy N, Efe JA, Emr SD, Shuman HA. Pathogen effector protein screening in yeast identifies *Legionella* factors that interfere with membrane trafficking. Proceedings of the National Academy of Sciences. 2005; 102(13).

33. Nagai H, Cambronne ED, Kagan JC, Amor JC, Kahn RA, Roy CR. A C-terminal translocation signal required for Dot/Icm-dependent delivery of the *Legionella* RalF protein to host cells. Proceedings of the National Academy of Sciences. 2005; 102(3), 826–831.

34. Stormo GD, Schneider TD, Gold L, Ehrenfeucht A. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli*. Nucleic Acids Research. 1982; 10(9), 2997–3011. PMID: 7048259

35. Wang J, Li J, Yang B, Xie R, Marquez-Lago TT, Leier A, et al. Bastion3: a two-layer ensemble predictor of type III secreted effectors. Bioinformatics 2018; (https://doi.org/10.1093/bioinformatics/bty914) PMID: 30388198

36. Wang J, Yang B, Leier A, Marquez-Lago TT, Hayashida M, Rocker A, et al. Bastion6: a bioinformatics approach for accurate prediction of type VI secreted effectors. Bioinformatics 2018; 34(15), 2546–2555. (https://doi.org/10.1093/bioinformatics/bty155) PMID: 29547915

37. Wang J, Yang B, Revote J, Leier A, Marquez-Lago TT, Webb G, et al. POSSUM: a bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM profiles. Bioinformatics 2017; 33(17), 2756–2758. (https://doi.org/10.1093/bioinformatics/btx302) PMID: 28903538

38. Cortes C, Vapnik V. Support-vector networks. Machine Learning 1995; 20(3), 273–297.

39. Crammer K, Singer Y. On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines. JMLR. 2001; 2, 265–292.

40. Perry JW, Kent A, Berry M. Machine literature searching X. Machine language; factors underlying its design and development. American Documentation. 1955; 6(4), 242–254.