DeepVF: a deep learning-based hybrid framework for identifying virulence factors using the stacking strategy

Supplementary Material

Table of Contents

2. Supplementary Figures

Fig. S1. Distribution of sequence lengths of proteins in the curated dataset.

Fig. S2. The pipeline of constructing a baseline model based on the training dataset, and the prediction process on the independent dataset.

Fig. S3. Performance comparison of different DL models trained using different sizes of protein sequence segments based on the sequence-to-vector feature encoding method.

Fig. S4. ROC curves of DeepVF, the BLAST-based baseline predictor and three existing state-of-the-art methods for VF prediction on the independent test.

3. Supplementary References

1. Supplementary Tables

Table S1. Detailed information of existing methods for VF prediction.

Method	Data size (Positive/Negative samples)	Negative sample selection	Features	Models	Ensemble learning strategy used	Web server or software accessibility	Origin
SPAAN (predicting the adhesins)	469/703	Bacterial enzymes and other non-virulent proteins	amino acid frequencies, multiplet frequencies, dipeptide frequencies, charge composition, hydrophobic composition	Neural network	NO	-	[1]
VirulentPred	1025/1030	Bacterial enzymes and other non-virulent proteins	AAC, DPC, higher order dipeptide composition, PSSM, sequence similarity	SVM	YES	http://bioinfo.icgeb.res.in/virulent/	[2]
E_SVM_VF[a]	1025/1030	Bacterial enzymes and other non-virulent proteins	Amino acid (AA) composition, Weighted AA composition, Five-level grouping composition	SVM	YES	-	[3]
Virulent-GO	1025/1030	Bacterial enzymes and other non-virulent proteins	GO terms	SVM, Naïve Bayes, KNN, C4.5 Decision tree	NO	-	[4]
VF_PAAC_EI[a]	1025/1030	Bacterial enzymes and other non-virulent proteins	2-Grams, RC, PA, GP, PSSM, AA	SVM	YES	-	[5]
ACCM_VFs[a]	803/803*5	Random sampling from the VF-contributing proteomes	AAC, PAAC	STRING network, BLAST, NNA classifier	NO	-	[6]
VF_KEGG[a]	514/514*5	Random sampling from the VF-contributing proteomes	KEGG pathway	RF	NO	-	[7]
MP3	1708/5815	Proteins deemed 'essential' for survival	AAC, DPC	SVM, HMM	NO	http://metagenomics.iiserb.ac.in/mp3/index.php	[8]
PBVF[a]	3891/10025	A novel strategy for selecting negatives that relies on protein function annotation data.	DPC, sequence similarity	SVM, RF, BLAST	NO	-	[9]

[a]As the names of these methods were not explicitly provided in the original work, here we named them according to the titles of their published papers.

Table S2. Performance comparison of multiple models for predicting VFs with different CD-HIT thresholds on the 10-fold cross-validation test.

Model[a]	Threshold	SN	SP	ACC	F-value	MCC
RF-AAC	0.3	0.641±0.029	0.699±0.011	0.670±0.017	0.660±0.019	0.341±0.033
	0.4	0.655±0.025	0.710±0.021	0.682±0.014	0.673±0.016	0.366±0.027
	0.5	0.653±0.016	0.726±0.026	0.689±0.009	0.678±0.011	0.380±0.019
	0.6	0.670±0.020	0.738±0.015	0.704±0.013	0.693±0.016	0.408±0.025
	0.7	0.663±0.031	0.751±0.013	0.707±0.017	0.693±0.019	0.416±0.032
	0.8	0.667±0.013	0.755±0.025	0.711±0.015	0.698±0.016	0.424±0.030
	0.9	**0.671±0.026**	**0.761±0.018**	**0.716±0.016**	**0.702±0.016**	**0.433±0.031**
SVM-AAC	0.3	0.663±0.037	0.700±0.016	0.681±0.022	0.675±0.026	0.363±0.044
	0.4	0.685±0.034	0.710±0.027	0.697±0.016	0.693±0.018	0.396±0.032
	0.5	0.687±0.016	0.717±0.024	0.702±0.013	0.697±0.017	0.403±0.025
	0.6	0.706±0.018	0.729±0.024	0.718±0.016	0.714±0.019	0.435±0.031
	0.7	0.716±0.019	0.741±0.012	0.728±0.012	0.725±0.010	0.456±0.023
	0.8	0.714±0.011	0.740±0.016	0.727±0.010	0.723±0.015	0.453±0.020
	0.9	**0.723±0.025**	**0.752±0.017**	**0.737±0.011**	**0.733±0.014**	**0.475±0.022**
XGBoost-AAC	0.3	0.651±0.022	0.672±0.016	0.661±0.009	0.657±0.012	0.323±0.019
	0.4	0.663±0.027	0.704±0.026	0.683±0.013	0.676±0.018	0.368±0.026
	0.5	0.666±0.021	0.702±0.025	0.684±0.014	0.678±0.019	0.368±0.029
	0.6	0.667±0.019	0.721±0.017	0.694±0.014	0.685±0.020	0.388±0.028
	0.7	0.673±0.024	0.737±0.018	0.705±0.019	0.695±0.018	0.411±0.036
	0.8	0.663±0.011	0.737±0.029	0.700±0.016	0.688±0.013	0.401±0.032
	0.9	**0.676±0.029**	**0.747±0.018**	**0.711±0.016**	**0.700±0.018**	**0.424±0.031**
MLP-AAC	0.3	0.571±0.082	0.790±0.066	0.681±0.015	0.639±0.043	0.374±0.025
	0.4	0.667±0.064	0.711±0.069	0.690±0.010	0.682±0.027	0.381±0.021
	0.5	0.625±0.066	0.765±0.069	0.694±0.009	0.670±0.023	0.397±0.020
	0.6	0.637±0.058	0.770±0.046	0.705±0.014	0.682±0.033	0.412±0.025
	0.7	0.624±0.035	**0.795±0.046**	0.711±0.013	0.683±0.018	0.427±0.029
	0.8	0.652±0.051	0.783±0.054	**0.718±0.009**	0.697±0.021	**0.441±0.020**
	0.9	**0.681±0.072**	0.747±0.091	0.716±0.015	**0.704±0.026**	0.434±0.033
CNN-L500	0.3	0.647±0.082	0.694±0.076	0.670±0.012	0.660±0.032	0.346±0.022
	0.4	0.616±0.057	0.739±0.059	0.677±0.012	0.655±0.020	0.360±0.025
	0.5	0.638±0.055	0.750±0.050	0.695±0.008	0.675±0.028	0.392±0.017

	0.6	0.661±0.056	0.747±0.048	0.704±0.009	0.690±0.027	0.411±0.019
	0.7	0.687±0.066	0.745±0.060	0.717±0.011	0.706±0.026	0.436±0.022
	0.8	0.688±0.044	0.756±0.037	0.722±0.006	0.712±0.022	0.446±0.013
	0.9	**0.688±0.055**	**0.774±0.050**	**0.731±0.010**	**0.718±0.023**	**0.465±0.020**
LSTM-L500	0.3	0.567±0.033	0.754±0.016	0.660±0.014	0.625±0.021	0.327±0.025
	0.4	0.568±0.057	0.767±0.045	0.668±0.012	0.630±0.030	0.343±0.021
	0.5	0.608±0.040	0.745±0.044	0.677±0.008	0.652±0.020	0.358±0.016
	0.6	0.612±0.050	0.767±0.036	0.690±0.014	0.662±0.032	0.384±0.027
	0.7	**0.647±0.053**	0.742±0.053	0.694±0.013	0.678±0.020	0.392±0.029
	0.8	0.641±0.048	0.748±0.050	0.696±0.010	0.677±0.023	0.393±0.022
	0.9	0.636±0.079	**0.772±0.065**	**0.705±0.012**	**0.680±0.037**	**0.415±0.021**
DNN-L500	0.3	0.611±0.038	0.748±0.041	0.680±0.012	0.656±0.019	0.364±0.025
	0.4	**0.632±0.048**	0.736±0.043	0.685±0.016	0.667±0.028	0.372±0.033
	0.5	0.616±0.029	0.765±0.029	0.691±0.011	0.665±0.022	0.385±0.025
	0.6	0.607±0.039	**0.792±0.037**	0.699±0.010	0.668±0.024	0.406±0.024
	0.7	0.623±0.036	0.774±0.041	0.699±0.013	0.674±0.020	0.403±0.026
	0.8	0.615±0.048	0.784±0.047	0.701±0.012	0.672±0.027	0.407±0.024
	0.9	0.622±0.029	0.785±0.036	**0.704±0.013**	**0.677±0.017**	**0.413±0.027**

Note: For ML, we used the simplest feature AAC (without combing with *seqsim*) to train multiple ML models with different CD-HIT thresholds. For DL, we used the L-500 to train multiple DL models with different thresholds of CD-HIT as the length of majority of sequences in the curated dataset were less than 500 (**Fig. S1**). [a]Model are denoted as model_feature. For example, RF_AAC means that the model was trained based on RF with AAC features. Values are expressed as mean±standard deviation. The best performance value for each metric across different CD-HIT thresholds within each model is highlighted in bold.

Table S3. Explanation of the seven parameters of XGBoost.

Parameters	**Description**[a]	**Parameter tuning range**
max_depth	max depth of the tree	[4,8], step: 1
eta	step size shrinkage	[0.01, 0.03], step: random
gamma	minimum loss reduction	[0, 0.2], step: random
subsample	subsample ratio of the training instances	[0.6, 0.9], step: random
colsample_bytree	subsampling ratio of columns	[0.5, 0.8], step: random
min_child_weight	minimum sum of instance weight	[1, 40], step: 1
max_delta_step	maximum delta step we allow each leaf output to be	[1,10], step: 1

Note: [a]The description of the above parameters comes from the official XGBoost document (https://xgboost.readthedocs.io/en/latest/).

Table S4. Performance comparison of ML baseline models with traditional ML features (without combining with *seqsim*) and *seqsim* for predicting VFs on the 10-fold cross-validation test.

Algorithm	Encoding	SN	SP	ACC	F-value	MCC
RF	AAC	0.643±0.019	0.694±0.028	0.668±0.017	0.659±0.015	0.337±0.033
	DPC	0.599±0.029	0.739±0.025	0.669±0.020	0.644±0.022	0.341±0.040
	DDE	0.610±0.036	0.720±0.024	0.665±0.022	0.645±0.027	0.331±0.044
	PAAC	0.641±0.021	0.698±0.025	0.669±0.017	0.660±0.017	0.339±0.034
	QSO	0.641±0.022	0.717±0.024	0.679±0.014	0.666±0.015	0.359±0.029
	PSSM-composition	0.660±0.022	0.830±0.024	0.745±0.015	0.721±0.017	0.498±0.031
	S-FPSSM	0.661±0.022	0.816±0.021	0.739±0.017	0.717±0.018	0.484±0.033
	RPM-PSSM	0.652±0.020	**0.833±0.024**	0.743±0.015	0.717±0.016	0.493±0.030
	seqsim	**0.824±0.017**	0.751±0.020	**0.787±0.011**	**0.795±0.012**	**0.576±0.022**
SVM	AAC	0.651±0.022	0.710±0.025	0.680±0.016	0.671±0.018	0.362±0.033
	DPC	0.626±0.021	0.695±0.027	0.660±0.017	0.648±0.017	0.321±0.035
	DDE	0.647±0.026	0.683±0.029	0.665±0.020	0.658±0.020	0.330±0.041
	PAAC	0.657±0.017	0.704±0.026	0.680±0.016	0.672±0.018	0.361±0.033
	QSO	0.653±0.022	0.708±0.029	0.681±0.016	0.672±0.014	0.362±0.032
	PSSM-composition	0.695±0.034	**0.846±0.022**	0.770±0.021	0.751±0.025	0.547±0.040
	S-FPSSM	0.718±0.026	0.783±0.023	0.750±0.019	0.742±0.021	0.502±0.038
	RPM-PSSM	0.692±0.033	0.819±0.026	0.756±0.019	0.739±0.023	0.516±0.038
	seqsim	**0.859±0.022**	0.737±0.025	**0.798±0.016**	**0.810±0.017**	**0.601±0.030**
XGBoost	AAC	0.647±0.020	0.678±0.029	0.662±0.017	0.657±0.015	0.325±0.033
	DPC	0.631±0.024	0.701±0.026	0.666±0.017	0.654±0.019	0.333±0.034
	DDE	0.631±0.025	0.699±0.026	0.665±0.017	0.653±0.020	0.331±0.033
	PAAC	0.646±0.018	0.680±0.027	0.663±0.015	0.657±0.014	0.327±0.030
	QSO	0.651±0.021	0.702±0.027	0.676±0.018	0.668±0.017	0.353±0.035
	PSSM-composition	0.715±0.025	**0.810±0.024**	0.762±0.016	0.751±0.017	0.528±0.032
	S-FPSSM	0.700±0.025	0.801±0.025	0.750±0.018	0.737±0.019	0.504±0.036
	RPM-PSSM	0.711±0.025	0.793±0.024	0.752±0.016	0.741±0.017	0.506±0.031
	seqsim	**0.825±0.019**	0.761±0.026	**0.793±0.017**	**0.799±0.019**	**0.587±0.033**
MLP	AAC	0.634±0.082	0.711±0.092	0.673±0.017	0.658±0.032	0.351±0.034
	DPC	0.615±0.079	0.734±0.082	0.675±0.015	0.652±0.034	0.356±0.031
	DDE	0.614±0.074	0.700±0.068	0.657±0.016	0.639±0.036	0.318±0.033
	PAAC	0.625±0.093	0.712±0.092	0.669±0.017	0.651±0.039	0.344±0.033
	QSO	0.607±0.079	0.740±0.071	0.674±0.016	0.648±0.039	0.353±0.031

	PSSM-composition	0.746±0.041	0.809±0.040	0.778±0.015	0.770±0.018	0.557±0.030
	S-FPSSM	0.700±0.062	0.821±0.048	0.761±0.015	0.744±0.027	0.529±0.028
	RPM-PSSM	0.718±0.047	**0.824±0.037**	0.771±0.017	0.758±0.023	0.547±0.032
	seqsim	**0.856±0.031**	0.738±0.034	**0.797±0.012**	**0.808±0.014**	**0.600±0.025**

Note: Values are expressed as mean±standard deviation. The best performance value for each metric across different encoding methods within each algorithm is highlighted in bold.

Table S5. Performance comparison of DL baseline models with their optimal features using different dropout rates on the 10-fold cross-validation test.

Model[a]	Dropout rate	SN	SP	ACC	F-value	MCC
CNN-L900	0.01	0.598±0.087	0.724±0.084	**0.661±0.017**	0.635±0.042	**0.329±0.032**
	0.05	0.609±0.057	0.711±0.053	0.660±0.015	0.641±0.027	0.324±0.029
	0.10	0.585±0.122	0.728±0.111	0.658±0.015	0.624±0.064	0.325±0.029
	0.15	0.608±0.057	0.709±0.045	0.658±0.016	0.639±0.029	0.320±0.031
	0.20	**0.620±0.097**	0.696±0.101	0.658±0.012	**0.642±0.039**	0.322±0.024
	0.25	0.578±0.078	**0.738±0.068**	0.658±0.014	0.625±0.042	0.323±0.026
	0.30	0.585±0.035	0.731±0.037	0.659±0.013	0.631±0.022	0.320±0.027
LSTM-L1000	0.01	0.560±0.102	0.748±0.093	**0.654±0.028**	0.613±0.062	**0.318±0.051**
	0.05	0.544±0.115	0.746±0.103	0.646±0.017	0.599±0.064	0.303±0.029
	0.10	0.487±0.091	**0.802±0.056**	0.645±0.023	0.573±0.064	0.307±0.037
	0.15	0.516±0.136	0.773±0.094	0.645±0.027	0.583±0.082	0.305±0.041
	0.20	0.545±0.107	0.747±0.071	0.646±0.025	0.601±0.064	0.302±0.047
	0.25	**0.571±0.100**	0.739±0.062	**0.654±0.030**	**0.618±0.070**	0.316±0.059
	0.30	0.540±0.073	0.759±0.077	0.649±0.014	0.603±0.032	0.309±0.028
DNN-L500	0.01	**0.620±0.054**	0.730±0.050	**0.675±0.019**	**0.655±0.027**	0.353±0.036
	0.05	0.612±0.069	0.729±0.057	0.670±0.014	0.648±0.033	0.345±0.024
	0.10	0.603±0.045	0.741±0.036	0.672±0.013	0.647±0.021	0.348±0.021
	0.15	0.596±0.077	0.751±0.070	0.673±0.009	0.644±0.034	**0.354±0.013**
	0.20	0.610±0.027	0.732±0.026	0.671±0.011	0.649±0.013	0.344±0.020
	0.25	0.584±0.064	**0.756±0.050**	0.670±0.013	0.637±0.032	0.347±0.022
	0.30	0.613±0.056	0.736±0.053	0.674±0.011	0.652±0.023	**0.354±0.021**

Note: Values are expressed as mean±standard deviation. The best performance value for each metric across different dropout rates within each algorithm is highlighted in bold.

Table S6. Performance comparison of baseline models with different cut-off thresholds on the 10-fold cross-validation test.

Threshold	Model	Encoding	SN	SP	ACC	F-value	MCC
0.3	RF	AAC	0.924±0.013	0.631±0.031	0.777±0.017	0.806±0.015	0.581±0.029
		DPC	0.986±0.011	0.154±0.099	0.570±0.047	0.697±0.024	0.244±0.086
		DDE	0.974±0.018	0.274±0.130	0.624±0.059	0.723±0.031	0.345±0.093
		PAAC	0.925±0.014	0.628±0.038	0.777±0.020	0.806±0.017	0.580±0.034
		QSO	0.938±0.014	0.524±0.059	0.731±0.028	0.777±0.020	0.507±0.043
		PSSM-composition	0.947±0.011	0.459±0.073	0.703±0.035	0.762±0.023	0.465±0.055
		S-FPSSM	0.956±0.010	0.434±0.035	0.695±0.019	0.758±0.015	0.458±0.031
		RPM-PSSM	0.964±0.013	0.323±0.088	0.644±0.041	0.731±0.023	0.373±0.065
	SVM	AAC	0.916±0.014	0.567±0.032	0.742±0.018	0.780±0.016	0.516±0.032
		DPC	0.923±0.016	0.417±0.038	0.670±0.021	0.737±0.016	0.394±0.039
		DDE	0.915±0.017	0.435±0.031	0.675±0.018	0.738±0.015	0.399±0.032
		PAAC	0.918±0.015	0.552±0.035	0.735±0.020	0.776±0.017	0.505±0.035
		QSO	0.923±0.013	0.495±0.032	0.709±0.018	0.760±0.015	0.463±0.032
		PSSM-composition	0.892±0.023	0.652±0.025	0.772±0.017	0.796±0.016	0.560±0.035
		S-FPSSM	0.905±0.021	0.542±0.054	0.723±0.027	0.766±0.020	0.480±0.047
		RPM-PSSM	0.890±0.017	0.582±0.030	0.736±0.018	0.771±0.016	0.496±0.034
	XGBoost	AAC	0.921±0.014	0.658±0.026	0.789±0.014	0.814±0.013	0.600±0.024
		DPC	0.908±0.015	0.677±0.023	0.793±0.014	0.814±0.015	0.601±0.027
		DDE	0.906±0.017	0.693±0.024	0.800±0.015	0.819±0.015	0.614±0.028
		PAAC	0.913±0.013	0.682±0.026	0.797±0.016	0.818±0.015	0.611±0.029
		QSO	0.919±0.014	0.675±0.037	0.797±0.018	0.819±0.015	0.614±0.030
		PSSM-composition	0.890±0.017	0.800±0.024	0.845±0.012	0.852±0.012	0.694±0.023
		S-FPSSM	0.903±0.018	0.764±0.029	0.833±0.015	0.844±0.014	0.673±0.028
		RPM-PSSM	0.900±0.014	0.763±0.020	0.832±0.012	0.842±0.012	0.670±0.023
	MLP	AAC	0.915±0.028	0.602±0.116	0.759±0.048	0.793±0.027	0.545±0.084
		DPC	0.909±0.025	0.637±0.099	0.773±0.040	0.801±0.022	0.568±0.069
		DDE	0.865±0.042	0.697±0.089	0.781±0.030	0.798±0.020	0.574±0.050
		PAAC	0.916±0.029	0.601±0.121	0.758±0.051	0.793±0.028	0.545±0.085
		QSO	0.901±0.026	0.657±0.088	0.779±0.037	0.803±0.023	0.576±0.057
		PSSM-composition	0.875±0.030	0.772±0.073	0.824±0.025	0.833±0.017	0.654±0.042
		S-FPSSM	0.859±0.039	0.762±0.073	0.811±0.025	0.820±0.019	0.627±0.045
		RPM-PSSM	0.884±0.029	0.746±0.075	0.815±0.028	0.828±0.018	0.639±0.045

	CNN	L100	0.856±0.118	0.268±0.176	0.562±0.035	0.660±0.024	0.160±0.053
		L200	0.843±0.105	0.326±0.163	0.585±0.033	0.669±0.020	0.206±0.046
		L300	0.787±0.115	0.426±0.180	0.607±0.036	0.665±0.022	0.237±0.051
		L400	0.806±0.104	0.408±0.171	0.607±0.037	0.671±0.021	0.236±0.067
		L500	0.828±0.093	0.381±0.170	0.604±0.044	0.676±0.017	0.236±0.070
		L600	0.809±0.101	0.403±0.169	0.607±0.039	0.672±0.021	0.237±0.064
		L700	0.818±0.101	0.383±0.173	0.600±0.043	0.671±0.017	0.227±0.067
		L800	0.811±0.099	0.404±0.166	0.607±0.039	0.673±0.019	0.239±0.062
		L900	0.837±0.100	0.363±0.167	0.600±0.038	0.676±0.020	0.231±0.069
		L1000	0.819±0.106	0.386±0.170	0.603±0.038	0.672±0.022	0.232±0.063
	LSTM	L100	0.930±0.069	0.147±0.117	0.538±0.030	0.668±0.015	0.125±0.050
		L200	0.916±0.075	0.187±0.129	0.552±0.029	0.671±0.015	0.152±0.050
		L300	0.929±0.074	0.149±0.147	0.539±0.039	0.668±0.014	0.116±0.077
		L400	0.939±0.067	0.118±0.125	0.529±0.033	0.666±0.014	0.100±0.057
		L500	0.874±0.113	0.226±0.198	0.550±0.046	0.659±0.019	0.126±0.089
		L600	0.935±0.082	0.111±0.154	0.523±0.042	0.662±0.017	0.064±0.087
		L700	0.933±0.103	0.105±0.189	0.518±0.046	0.659±0.015	0.043±0.097
		L800	0.967±0.063	0.054±0.117	0.511±0.031	0.664±0.014	0.033±0.073
		L900	0.965±0.074	0.066±0.147	0.516±0.040	0.666±0.012	0.051±0.083
		L1000	0.956±0.093	0.086±0.169	0.521±0.041	0.666±0.014	0.066±0.089
	DNN	L100	0.959±0.027	0.110±0.054	0.534±0.019	0.673±0.011	0.129±0.038
		L200	0.937±0.032	0.167±0.074	0.553±0.025	0.677±0.015	0.162±0.054
		L300	0.944±0.027	0.173±0.066	0.559±0.025	0.681±0.014	0.182±0.050
		L400	0.937±0.034	0.182±0.081	0.560±0.026	0.680±0.013	0.180±0.048
		L500	0.941±0.029	0.172±0.066	0.556±0.026	0.680±0.014	0.177±0.041
		L600	0.944±0.029	0.169±0.064	0.557±0.022	0.681±0.012	0.180±0.041
		L700	0.943±0.033	0.167±0.069	0.555±0.022	0.680±0.012	0.176±0.036
		L800	0.947±0.031	0.154±0.066	0.551±0.024	0.678±0.014	0.167±0.038
		L900	0.951±0.026	0.150±0.055	0.551±0.020	0.679±0.013	0.170±0.035
		L1000	0.948±0.039	0.151±0.074	0.550±0.023	0.678±0.014	0.165±0.040
0.4	RF	AAC	0.888±0.016	0.735±0.025	0.812±0.016	0.825±0.016	0.631±0.031
		DPC	0.923±0.018	0.554±0.084	0.739±0.037	0.780±0.025	0.514±0.059
		DDE	0.913±0.018	0.595±0.047	0.754±0.025	0.788±0.020	0.536±0.044
		PAAC	0.890±0.015	0.735±0.026	0.813±0.016	0.826±0.016	0.633±0.030

	QSO	0.892±0.016	0.716±0.025	0.804±0.015	0.820±0.014	0.618±0.028
	PSSM-composition	0.865±0.019	0.765±0.038	0.815±0.022	0.824±0.020	0.633±0.042
	S-FPSSM	0.897±0.017	0.685±0.025	0.791±0.016	0.811±0.014	0.595±0.030
	RPM-PSSM	0.883±0.015	0.710±0.042	0.797±0.023	0.813±0.019	0.603±0.043
SVM	AAC	0.875±0.015	0.673±0.029	0.774±0.018	0.795±0.017	0.560±0.034
	DPC	0.846±0.022	0.589±0.030	0.717±0.019	0.749±0.017	0.449±0.036
	DDE	0.843±0.017	0.590±0.036	0.716±0.019	0.748±0.016	0.447±0.036
	PAAC	0.880±0.016	0.666±0.030	0.773±0.017	0.795±0.016	0.560±0.032
	QSO	0.880±0.016	0.630±0.031	0.755±0.018	0.782±0.016	0.527±0.034
	PSSM-composition	0.844±0.027	0.762±0.023	0.803±0.018	0.811±0.018	0.608±0.037
	S-FPSSM	0.857±0.023	0.669±0.047	0.763±0.024	0.784±0.020	0.537±0.044
	RPM-PSSM	0.828±0.023	0.737±0.028	0.782±0.017	0.792±0.016	0.567±0.033
XGBoost	AAC	0.886±0.017	0.742±0.025	0.814±0.016	0.826±0.016	0.635±0.031
	DPC	0.884±0.016	0.744±0.023	0.814±0.014	0.826±0.014	0.634±0.026
	DDE	0.878±0.018	0.760±0.021	0.819±0.015	0.829±0.016	0.643±0.029
	PAAC	0.878±0.017	0.757±0.027	0.817±0.018	0.828±0.018	0.639±0.035
	QSO	0.886±0.015	0.755±0.025	0.820±0.015	0.831±0.015	0.647±0.029
	PSSM-composition	0.858±0.018	0.851±0.019	0.854±0.012	0.855±0.013	0.709±0.024
	S-FPSSM	0.873±0.017	0.825±0.025	0.849±0.015	0.852±0.015	0.699±0.029
	RPM-PSSM	0.869±0.015	0.829±0.019	0.849±0.013	0.852±0.014	0.698±0.025
MLP	AAC	0.890±0.024	0.683±0.039	0.787±0.016	0.807±0.015	0.586±0.029
	DPC	0.882±0.025	0.705±0.038	0.794±0.015	0.810±0.014	0.597±0.027
	DDE	0.838±0.037	0.758±0.058	0.798±0.018	0.806±0.014	0.600±0.031
	PAAC	0.890±0.027	0.677±0.088	0.784±0.038	0.805±0.023	0.581±0.063
	QSO	0.878±0.027	0.712±0.050	0.795±0.020	0.810±0.016	0.599±0.033
	PSSM-composition	0.844±0.027	0.843±0.039	0.843±0.013	0.843±0.012	0.687±0.024
	S-FPSSM	0.829±0.039	0.825±0.047	0.827±0.016	0.827±0.016	0.656±0.031
	RPM-PSSM	0.852±0.029	0.817±0.038	0.835±0.011	0.838±0.011	0.671±0.021
CNN	L100	0.744±0.109	0.445±0.142	0.594±0.025	0.645±0.030	0.204±0.039
	L200	0.742±0.096	0.502±0.117	0.622±0.020	0.661±0.030	0.257±0.036
	L300	0.705±0.103	0.561±0.132	0.634±0.022	0.656±0.031	0.276±0.041
	L400	0.716±0.089	0.565±0.120	0.641±0.024	0.664±0.029	0.289±0.044
	L500	0.728±0.091	0.541±0.138	0.634±0.030	0.664±0.022	0.279±0.054
	L600	0.714±0.079	0.563±0.110	0.639±0.026	0.663±0.026	0.284±0.047

		L700	0.729±0.082	0.544±0.123	0.636±0.031	0.666±0.021	0.283±0.049
		L800	0.720±0.081	0.560±0.107	0.640±0.022	0.665±0.024	0.288±0.038
		L900	0.728±0.092	0.543±0.135	0.636±0.030	0.665±0.027	0.281±0.056
		L1000	0.729±0.097	0.539±0.133	0.635±0.027	0.664±0.027	0.280±0.042
	LSTM	L100	0.773±0.100	0.400±0.138	0.587±0.025	0.650±0.026	0.191±0.038
		L200	0.747±0.093	0.479±0.129	0.614±0.027	0.657±0.028	0.239±0.047
		L300	0.768±0.140	0.426±0.218	0.598±0.046	0.653±0.031	0.216±0.076
		L400	0.769±0.114	0.441±0.183	0.605±0.043	0.659±0.025	0.228±0.070
		L500	0.671±0.113	0.585±0.155	0.627±0.031	0.640±0.034	0.263±0.052
		L600	0.724±0.155	0.494±0.268	0.609±0.062	0.646±0.030	0.226±0.117
		L700	0.729±0.167	0.468±0.293	0.598±0.067	0.642±0.026	0.203±0.132
		L800	0.720±0.140	0.505±0.236	0.612±0.055	0.647±0.026	0.232±0.104
		L900	0.721±0.181	0.490±0.302	0.605±0.065	0.641±0.035	0.220±0.126
		L1000	0.736±0.145	0.500±0.259	0.618±0.061	0.656±0.024	0.244±0.118
	DNN	L100	0.793±0.055	0.393±0.080	0.593±0.018	0.660±0.013	0.204±0.032
		L200	0.777±0.058	0.465±0.067	0.622±0.015	0.672±0.022	0.258±0.029
		L300	0.773±0.061	0.468±0.076	0.621±0.021	0.670±0.023	0.255±0.040
		L400	0.785±0.050	0.482±0.086	0.634±0.024	0.682±0.016	0.282±0.046
		L500	0.776±0.057	0.491±0.083	0.633±0.022	0.679±0.016	0.281±0.040
		L600	0.765±0.053	0.509±0.079	0.637±0.022	0.678±0.017	0.285±0.041
		L700	0.768±0.050	0.498±0.078	0.633±0.021	0.676±0.016	0.277±0.041
		L800	0.778±0.058	0.485±0.074	0.631±0.017	0.678±0.017	0.277±0.034
		L900	0.775±0.053	0.485±0.078	0.630±0.018	0.677±0.015	0.274±0.034
		L1000	0.777±0.081	0.471±0.125	0.624±0.029	0.673±0.019	0.264±0.045
0.6	RF	AAC	0.743±0.023	0.884±0.018	0.813±0.013	0.799±0.015	0.633±0.025
		DPC	0.476±0.082	0.953±0.018	0.714±0.037	0.621±0.068	0.489±0.057
		DDE	0.573±0.094	0.926±0.020	0.749±0.042	0.692±0.069	0.535±0.068
		PAAC	0.739±0.023	0.886±0.017	0.812±0.013	0.797±0.015	0.632±0.025
		QSO	0.682±0.031	0.906±0.019	0.794±0.015	0.767±0.019	0.603±0.027
		PSSM-composition	0.635±0.035	0.957±0.012	0.796±0.017	0.757±0.024	0.626±0.028
		S-FPSSM	0.635±0.025	0.954±0.011	0.795±0.013	0.755±0.018	0.622±0.023
		RPM-PSSM	0.600±0.040	0.963±0.011	0.781±0.019	0.732±0.030	0.604±0.030
	SVM	AAC	0.734±0.023	0.848±0.023	0.791±0.016	0.778±0.018	0.586±0.032
		DPC	0.605±0.022	0.851±0.019	0.728±0.016	0.690±0.018	0.471±0.030

	DDE	0.614±0.019	0.840±0.022	0.727±0.014	0.692±0.016	0.466±0.029
	PAAC	0.731±0.021	0.847±0.022	0.788±0.016	0.775±0.018	0.581±0.032
	QSO	0.697±0.024	0.860±0.021	0.778±0.016	0.758±0.017	0.565±0.031
	PSSM-composition	0.728±0.026	0.889±0.019	0.809±0.015	0.792±0.017	0.626±0.030
	S-FPSSM	0.687±0.080	0.873±0.032	0.780±0.031	0.755±0.050	0.572±0.051
	RPM-PSSM	0.685±0.040	0.891±0.020	0.788±0.022	0.764±0.028	0.590±0.040
XGBoost	AAC	0.761±0.022	0.870±0.019	0.816±0.012	0.805±0.014	0.635±0.024
	DPC	0.780±0.024	0.860±0.021	0.820±0.019	0.812±0.021	0.642±0.037
	DDE	0.791±0.021	0.862±0.017	0.826±0.013	0.820±0.015	0.655±0.026
	PAAC	0.768±0.016	0.867±0.019	0.818±0.011	0.808±0.013	0.638±0.023
	QSO	0.763±0.025	0.877±0.016	0.820±0.013	0.809±0.016	0.644±0.025
	PSSM-composition	0.804±0.018	0.912±0.017	0.858±0.011	0.850±0.012	0.721±0.022
	S-FPSSM	0.806±0.019	0.908±0.017	0.857±0.013	0.849±0.014	0.717±0.026
	RPM-PSSM	0.798±0.020	0.909±0.017	0.854±0.013	0.845±0.015	0.712±0.027
MLP	AAC	0.752±0.110	0.803±0.051	0.778±0.035	0.767±0.071	0.561±0.057
	DPC	0.768±0.069	0.806±0.047	0.787±0.020	0.781±0.032	0.577±0.038
	DDE	0.755±0.052	0.832±0.043	0.794±0.014	0.785±0.022	0.591±0.026
	PAAC	0.730±0.126	0.808±0.056	0.770±0.041	0.753±0.086	0.546±0.066
	QSO	0.758±0.090	0.813±0.055	0.786±0.023	0.777±0.043	0.577±0.040
	PSSM-composition	0.783±0.031	0.907±0.024	0.845±0.012	0.834±0.015	0.696±0.022
	S-FPSSM	0.758±0.042	0.905±0.027	0.832±0.015	0.818±0.021	0.671±0.028
	RPM-PSSM	0.787±0.042	0.889±0.034	0.838±0.011	0.829±0.016	0.681±0.020
CNN	L100	0.362±0.152	0.829±0.107	0.596±0.027	0.452±0.132	0.223±0.044
	L200	0.446±0.122	0.808±0.090	0.627±0.027	0.535±0.090	0.279±0.040
	L300	0.493±0.125	0.776±0.098	0.635±0.024	0.565±0.084	0.289±0.037
	L400	0.468±0.118	0.807±0.088	0.638±0.025	0.554±0.084	0.298±0.042
	L500	0.473±0.127	0.811±0.092	0.642±0.026	0.559±0.086	0.310±0.040
	L600	0.481±0.116	0.804±0.090	0.642±0.023	0.565±0.076	0.308±0.037
	L700	0.486±0.105	0.797±0.094	0.642±0.021	0.569±0.069	0.306±0.038
	L800	0.480±0.102	0.807±0.086	0.644±0.022	0.567±0.068	0.310±0.039
	L900	0.442±0.112	0.839±0.081	0.641±0.025	0.543±0.079	0.313±0.036
	L1000	0.475±0.118	0.814±0.092	0.645±0.021	0.563±0.079	0.316±0.033
LSTM	L100	0.364±0.125	0.836±0.098	0.601±0.020	0.464±0.098	0.236±0.030
	L200	0.412±0.083	0.842±0.059	0.627±0.022	0.520±0.066	0.284±0.038

		L300	0.422±0.136	0.830±0.097	0.626±0.031	0.518±0.099	0.284±0.049
		L400	0.438±0.104	0.808±0.082	0.623±0.028	0.530±0.075	0.269±0.050
		L500	0.439±0.103	0.815±0.076	0.627±0.026	0.533±0.076	0.279±0.040
		L600	0.444±0.104	0.827±0.060	0.636±0.030	0.542±0.082	0.296±0.050
		L700	0.415±0.127	0.827±0.086	0.621±0.038	0.511±0.105	0.271±0.067
		L800	0.437±0.122	0.825±0.089	0.631±0.027	0.532±0.087	0.290±0.046
		L900	0.436±0.108	0.833±0.073	0.635±0.030	0.536±0.087	0.298±0.054
		L1000	0.450±0.107	0.825±0.072	0.638±0.028	0.546±0.081	0.301±0.047
	DNN	L100	0.361±0.059	0.852±0.045	0.607±0.016	0.476±0.047	0.246±0.031
		L200	0.411±0.060	0.851±0.037	0.631±0.016	0.525±0.048	0.293±0.027
		L300	0.406±0.060	0.854±0.041	0.630±0.017	0.521±0.046	0.293±0.029
		L400	0.438±0.064	0.839±0.045	0.639±0.018	0.545±0.048	0.305±0.032
		L500	0.447±0.057	0.849±0.041	0.648±0.017	0.558±0.042	0.325±0.031
		L600	0.447±0.055	0.845±0.047	0.646±0.017	0.556±0.038	0.320±0.032
		L700	0.431±0.062	0.858±0.043	0.645±0.019	0.546±0.046	0.322±0.032
		L800	0.440±0.062	0.848±0.048	0.644±0.020	0.551±0.047	0.318±0.035
		L900	0.435±0.059	0.856±0.042	0.646±0.020	0.549±0.045	0.323±0.036
		L1000	0.422±0.059	0.863±0.040	0.643±0.017	0.539±0.045	0.320±0.030
0.7	RF	AAC	0.585±0.036	0.942±0.014	0.764±0.017	0.712±0.027	0.564±0.028
		DPC	0.204±0.054	0.992±0.006	0.598±0.027	0.334±0.073	0.317±0.049
		DDE	0.242±0.075	0.988±0.007	0.615±0.036	0.380±0.094	0.343±0.063
		PAAC	0.583±0.041	0.942±0.013	0.763±0.018	0.710±0.030	0.564±0.029
		QSO	0.480±0.046	0.964±0.013	0.722±0.021	0.632±0.039	0.507±0.033
		PSSM-composition	0.502±0.041	0.980±0.009	0.741±0.021	0.659±0.035	0.549±0.031
		S-FPSSM	0.447±0.027	0.988±0.005	0.718±0.015	0.613±0.025	0.518±0.022
		RPM-PSSM	0.418±0.046	0.987±0.006	0.703±0.024	0.583±0.045	0.493±0.035
	SVM	AAC	0.599±0.024	0.918±0.019	0.758±0.013	0.712±0.019	0.546±0.027
		DPC	0.445±0.026	0.931±0.013	0.688±0.016	0.587±0.024	0.430±0.028
		DDE	0.453±0.027	0.921±0.017	0.687±0.016	0.591±0.026	0.423±0.031
		PAAC	0.593±0.028	0.921±0.019	0.757±0.014	0.709±0.021	0.544±0.027
		QSO	0.537±0.029	0.937±0.016	0.736±0.017	0.670±0.023	0.516±0.030
		PSSM-composition	0.650±0.026	0.932±0.013	0.791±0.014	0.757±0.019	0.607±0.026
		S-FPSSM	0.533±0.063	0.943±0.017	0.738±0.028	0.668±0.051	0.522±0.044
		RPM-PSSM	0.595±0.037	0.937±0.016	0.766±0.020	0.718±0.029	0.567±0.035

XGBoost	AAC	0.638±0.032	0.925±0.016	0.782±0.015	0.745±0.021	0.589±0.025
	DPC	0.677±0.026	0.916±0.014	0.796±0.016	0.768±0.020	0.610±0.030
	DDE	0.703±0.022	0.908±0.015	0.805±0.013	0.783±0.015	0.624±0.024
	PAAC	0.674±0.022	0.915±0.016	0.794±0.013	0.766±0.015	0.607±0.024
	QSO	0.657±0.047	0.925±0.017	0.791±0.020	0.758±0.030	0.605±0.032
	PSSM-composition	0.765±0.022	0.935±0.016	0.850±0.011	0.836±0.012	0.710±0.020
	S-FPSSM	0.756±0.024	0.936±0.015	0.846±0.012	0.831±0.015	0.704±0.023
	RPM-PSSM	0.745±0.018	0.940±0.015	0.843±0.011	0.826±0.013	0.699±0.021
MLP	AAC	0.527±0.158	0.897±0.049	0.712±0.061	0.630±0.151	0.456±0.114
	DPC	0.590±0.122	0.886±0.050	0.738±0.040	0.685±0.080	0.504±0.061
	DDE	0.701±0.078	0.859±0.047	0.780±0.022	0.759±0.041	0.571±0.036
	PAAC	0.498±0.182	0.901±0.059	0.700±0.066	0.601±0.164	0.439±0.107
	QSO	0.583±0.141	0.890±0.056	0.737±0.047	0.678±0.099	0.503±0.071
	PSSM-composition	0.736±0.046	0.926±0.025	0.831±0.017	0.813±0.023	0.676±0.027
	S-FPSSM	0.714±0.052	0.927±0.023	0.820±0.020	0.798±0.030	0.657±0.033
	RPM-PSSM	0.739±0.059	0.913±0.035	0.826±0.018	0.808±0.028	0.664±0.029
CNN	L100	0.232±0.172	0.903±0.111	0.567±0.034	0.315±0.172	0.194±0.051
	L200	0.301±0.150	0.885±0.096	0.593±0.034	0.403±0.142	0.240±0.044
	L300	0.381±0.159	0.839±0.109	0.610±0.033	0.472±0.136	0.260±0.041
	L400	0.348±0.151	0.870±0.099	0.610±0.033	0.451±0.133	0.268±0.044
	L500	0.352±0.161	0.875±0.095	0.614±0.037	0.454±0.143	0.277±0.050
	L600	0.362±0.156	0.866±0.098	0.614±0.035	0.463±0.131	0.274±0.050
	L700	0.366±0.143	0.868±0.101	0.618±0.031	0.471±0.123	0.282±0.048
	L800	0.358±0.140	0.877±0.092	0.618±0.031	0.467±0.121	0.285±0.042
	L900	0.305±0.138	0.906±0.085	0.606±0.036	0.417±0.131	0.273±0.051
	L1000	0.348±0.153	0.880±0.096	0.614±0.034	0.453±0.134	0.280±0.044
LSTM	L100	0.218±0.132	0.919±0.084	0.570±0.027	0.317±0.134	0.201±0.042
	L200	0.281±0.083	0.917±0.047	0.599±0.023	0.405±0.084	0.260±0.037
	L300	0.293±0.148	0.901±0.088	0.597±0.039	0.400±0.141	0.251±0.062
	L400	0.302±0.118	0.883±0.071	0.593±0.033	0.413±0.110	0.232±0.054
	L500	0.325±0.111	0.874±0.063	0.599±0.033	0.437±0.103	0.241±0.049
	L600	0.315±0.119	0.891±0.054	0.603±0.039	0.429±0.120	0.253±0.066
	L700	0.295±0.120	0.898±0.060	0.596±0.039	0.408±0.122	0.242±0.070
	L800	0.304±0.140	0.900±0.083	0.602±0.035	0.416±0.122	0.260±0.056

	L900	0.326±0.110	0.897±0.047	0.612±0.035	0.445±0.117	0.272±0.062
	L1000	0.337±0.130	0.889±0.069	0.613±0.036	0.450±0.128	0.275±0.056
DNN	L100	0.206±0.046	0.937±0.024	0.571±0.017	0.321±0.055	0.211±0.028
	L200	0.267±0.054	0.930±0.022	0.599±0.019	0.397±0.060	0.264±0.034
	L300	0.259±0.048	0.938±0.023	0.599±0.018	0.390±0.055	0.269±0.032
	L400	0.271±0.050	0.937±0.025	0.604±0.015	0.404±0.055	0.280±0.029
	L500	0.297±0.057	0.931±0.022	0.614±0.020	0.432±0.060	0.295±0.034
	L600	0.307±0.052	0.925±0.029	0.616±0.016	0.442±0.052	0.297±0.027
	L700	0.296±0.059	0.929±0.026	0.613±0.020	0.430±0.062	0.292±0.033
	L800	0.302±0.057	0.924±0.028	0.613±0.020	0.435±0.060	0.290±0.035
	L900	0.293±0.059	0.930±0.024	0.612±0.020	0.427±0.063	0.290±0.032
	L1000	0.286±0.055	0.931±0.024	0.609±0.017	0.419±0.059	0.285±0.032

Note: Values are expressed as mean±standard deviation.

Table S7. Performance comparison of different algorithms to construct the final meta model on the 10-fold cross-validation test.

Algorithms	Model	SN	SP	ACC	F-value	MCC
XGBoost	ML model	0.836±0.021	0.916±0.017	0.876±0.013	0.871±0.015	0.755±0.027
	DL model	0.623±0.036	0.739±0.030	0.681±0.020	0.661±0.024	0.365±0.038
	Hybrid model	**0.837±0.023**	0.918±0.015	**0.878±0.014**	**0.872±0.016**	**0.758±0.027**
MLP	ML model	0.767±0.107	**0.946±0.019**	0.856±0.053	0.837±0.094	0.726±0.088
	DL model	0.540±0.081	0.792±0.059	0.666±0.022	0.615±0.053	0.346±0.040
	Hybrid model	0.772±0.067	0.947±0.021	0.860±0.026	0.844±0.043	0.732±0.044
DNN	ML model	0.804±0.021	0.908±0.016	0.856±0.013	0.848±0.015	0.716±0.025
	DL model	0.633±0.037	0.730±0.035	0.682±0.020	0.665±0.023	0.366±0.038
	Hybrid model	0.762±0.033	0.864±0.025	0.813±0.015	0.802±0.018	0.630±0.028

Note: Values are expressed as mean±standard deviation. The best performance value is highlighted in bold.

Table S8. Performance comparison between models the meta model using the prediction label and the prediction score of baseline models based on the 10-fold cross validation test.

Feature	SN	SP	ACC	F-value	MCC
Prediction label	0.837±0.023	0.918±0.015	0.878±0.014	0.872±0.016	0.758±0.027
Prediction score	0.837±0.022	0.906±0.019	0.872±0.015	0.867±0.016	0.745±0.030

Table S9. Detailed contributions of different baseline models to the performance of the ultimate meta XGBoost classifier.

Feature[a]	Gain	Cover	Frequency
MLP-S-FPSSM	0.334	0.048	0.016
MLP-RPM-PSSM	0.113	0.031	0.014
XGBOOST-S-FPSSM	0.101	0.035	0.016
CNN-L100	0.096	0.017	0.007
XGBOOST-RPM-PSSM	0.067	0.063	0.029
RF-DDE	0.025	0.047	0.043
SVM-S-FPSSM	0.025	0.039	0.029
SVM-DDE	0.022	0.049	0.053
RF-AAC	0.016	0.080	0.127
SVM-PAAC	0.014	0.024	0.039
MLP-DDE	0.011	0.023	0.015
SVM-PSSM-composition	0.010	0.014	0.022
XGBOOST-AAC	0.009	0.021	0.019
RF-S-FPSSM	0.007	0.015	0.016
MLP-DPC	0.005	0.020	0.013
LSTM-L100	0.005	0.011	0.015
CNN-L600	0.005	0.019	0.019
SVM-AAC	0.005	0.012	0.014
SVM-QSOrder	0.005	0.012	0.015
XGBOOST-PAAC	0.005	0.020	0.016
CNN-L1000	0.005	0.009	0.012
SVM-RPM-PSSM	0.005	0.014	0.018
CNN-L300	0.004	0.017	0.024
CNN-L400	0.004	0.017	0.023
LSTM-L700	0.004	0.010	0.012
LSTM-L1000	0.004	0.017	0.020
CNN-L800	0.004	0.016	0.017
LSTM-L800	0.004	0.011	0.010
RF-PAAC	0.004	0.012	0.012
CNN-L200	0.004	0.014	0.021
CNN-L900	0.004	0.006	0.008
XGBOOST-QSOrder	0.004	0.016	0.017
SVM-DPC	0.004	0.008	0.012
MLP-QSOrder	0.003	0.012	0.008

LSTM-L500	0.003	0.006	0.007
LSTM-L600	0.003	0.013	0.013
DNN-L700	0.003	0.004	0.005
LSTM-L200	0.003	0.010	0.011
CNN-L700	0.003	0.009	0.012
RF-RPM-PSSM	0.003	0.011	0.010
LSTM-L300	0.003	0.010	0.012
LSTM-L400	0.003	0.012	0.013
MLP-PAAC	0.003	0.007	0.007
RF-PSSM-composition	0.003	0.009	0.010
XGBOOST-DPC	0.003	0.011	0.013
CNN-L500	0.003	0.012	0.015
DNN-L900	0.003	0.009	0.011
DNN-L600	0.003	0.007	0.009
LSTM-L900	0.002	0.010	0.010
DNN-L300	0.002	0.009	0.011
DNN-L500	0.002	0.010	0.011
DNN-L200	0.002	0.009	0.011
MLP-AAC	0.002	0.007	0.008
DNN-L100	0.002	0.004	0.005
DNN-L400	0.002	0.005	0.006
XGBOOST-DDE	0.002	0.008	0.009
DNN-L800	0.002	0.006	0.008
DNN-L1000	0.001	0.003	0.004
RF-DPC	0.001	0.005	0.004
MLP-PSSM-composition	0.001	0.006	0.005
RF-QSOrder	0.001	0.004	0.005
XGBOOST-PSSM-composition	0.001	0.004	0.005

Note: Feature[a] are denoted as model_feature. For example, MLP_PSSM-composition means that the feature is composed of the predictive labels of the base model constructed by MLP with the PSSM-composition features.

Table S10. Performance comparison of baseline models for VF prediction on the independent test.

Algorithm	Encoding	SN	SP	ACC	F-value	MCC
RF	AAC	0.804	0.809	0.806	0.806	0.613
	DPC	0.774	0.814	0.794	0.790	0.589
	DDE	0.788	0.809	0.799	0.796	0.597
	PAAC	0.804	0.814	0.809	0.808	0.618
	QSO	**0.806**	0.832	**0.819**	**0.816**	**0.637**
	PSSM-composition	0.726	**0.872**	0.799	0.783	0.604
	S-FPSSM	0.766	0.861	0.813	0.804	0.63
	RPM-PSSM	0.726	0.847	0.786	0.773	0.577
SVM	AAC	**0.868**	0.682	0.775	0.794	0.56
	DPC	0.776	0.717	0.747	0.754	0.494
	DDE	0.804	0.658	0.731	0.749	0.467
	PAAC	0.877	0.677	0.777	0.797	0.565
	QSO	0.826	0.743	0.785	0.793	0.571
	PSSM-composition	0.694	**0.891**	0.793	0.770	**0.597**
	S-FPSSM	0.819	0.769	0.794	**0.799**	0.589
	RPM-PSSM	0.785	0.809	**0.797**	0.794	0.594
XGBoost	AAC	0.802	0.804	0.803	0.803	0.606
	DPC	**0.814**	0.802	0.808	0.809	0.616
	DDE	0.812	0.800	0.806	0.808	0.613
	PAAC	0.799	0.806	0.802	0.801	0.604
	QSO	0.806	0.818	0.812	0.810	0.623
	PSSM-composition	0.799	**0.832**	**0.815**	**0.812**	**0.631**
	S-FPSSM	0.795	0.83	0.812	0.809	0.625
	RPM-PSSM	0.797	0.828	0.812	0.810	0.625
MLP	AAC	0.814	0.778	0.796	0.800	0.592
	DPC	0.816	0.785	0.800	0.803	0.601
	DDE	0.806	0.814	0.81	0.809	0.620
	PAAC	**0.821**	0.778	0.799	0.804	0.600
	QSO	0.809	0.792	0.800	0.802	0.601
	PSSM-composition	0.767	0.851	0.809	0.801	0.62
	S-FPSSM	0.776	**0.868**	**0.822**	**0.813**	**0.647**
	RPM-PSSM	0.797	0.837	0.817	**0.813**	0.634
CNN	L100	0.571	0.740	0.655	0.624	0.315
	L200	0.616	0.764	0.69	0.665	0.384
	L300	0.635	0.764	0.700	0.679	0.403

	L400	0.644	0.780	0.712	0.691	0.428
	L500	**0.655**	0.788	0.721	0.701	0.447
	L600	0.649	**0.814**	**0.732**	**0.708**	**0.470**
	L700	0.648	0.793	0.72	0.699	0.446
	L800	0.648	0.800	0.724	0.701	0.453
	L900	0.634	0.806	0.72	0.693	0.446
	L1000	0.639	0.799	0.719	0.694	0.443
LSTM	L100	0.59	0.738	0.664	0.637	0.332
	L200	0.587	0.766	0.676	0.644	0.358
	L300	0.592	0.781	0.687	0.654	0.38
	L400	**0.639**	0.762	0.701	0.681	0.404
	L500	0.594	0.811	0.702	0.666	0.414
	L600	0.609	0.821	0.715	0.682	0.441
	L700	0.606	0.835	0.720	0.684	0.453
	L800	0.62	0.823	**0.721**	0.690	0.452
	L900	0.594	**0.849**	**0.721**	0.681	**0.458**
	L1000	0.623	0.819	**0.721**	**0.691**	0.451
DNN	L100	0.575	0.717	0.646	0.619	0.295
	L200	0.592	0.712	0.652	0.630	0.306
	L300	0.618	0.766	0.692	0.667	0.388
	L400	0.637	0.747	0.692	0.674	0.386
	L500	0.635	0.776	0.706	0.683	0.416
	L600	**0.641**	0.783	**0.712**	**0.690**	0.428
	L700	0.628	**0.795**	**0.712**	0.686	**0.430**
	L800	0.635	0.781	0.708	0.685	0.421
	L900	0.632	0.786	0.709	0.685	0.423
	L1000	0.630	0.793	**0.712**	0.686	0.429

Note: The best performance value for each metric across different encoding methods within each algorithm is highlighted in bold.

Table S11. Performance comparison of DeepVF trained based on VirulentPred's training dataset and VirulentPred on different independent datasets.

Testing dataset	Algorithm	SN	SP	ACC	F-value	MCC
VirulentPred Ind-I[a]	VirulentPred	**0.868**	-	-	-	-
	DeepVF	**0.868**	-	-	-	-
VirulentPred Ind-II[b]	VirulentPred	0.787	-	-	-	-
	DeepVF	**0.801**	-	-	-	-
The DeepVF independent dataset	VirulentPred	0.641	0.573	0.607	0.620	0.214
	DeepVF	**0.668**	**0.623**	**0.646**	**0.654**	**0.292**

Note: [a]VirulentPred Ind-I refers to the independent dataset I of VirulentPred, which contained 38 positive samples (Note that we removed the 22nd whose sequence length was less than 10 and the 25th sample whose sequence had illegal character X) and was downloaded from the VirulentPred webserver (no negative samples available). [b]VirulentPred Ind-II refers to the independent dataset II of VirulentPred, which contained 141 positive samples and was downloaded from the VirulentPred webserver (no negative samples available). The best performance value within each test case is highlighted in bold.

Table S12. Summary of the top 100 sequences predicted by DeepVF and the BLAST-based baseline predictor based on three proteomes.

Proteomes	Number of Proteins[a]	Overlap[b]	DeepVF[c]	BLAST[d]
Escherichia coli O157:H7	4893 (19/3)	12	5	0
Streptococcus pneumoniae (strain ATCC BAA-255 / R6)	1987 (1/2)	14	1	1
Mycobacterium tuberculosis (strain ATCC 25618 / H37Rv)	3616 (58/8)	11	3	0

Note: [a]Number of proteins within each proteome was counted after removing those sequences that have appeared in the training dataset. The values in the parenthesis represent positive/negative samples contained in the independent dataset/contained negative samples on independent dataset). [b]Overlap represents the number of predicted sequences that appeared in both top 100 ranking lists of DeepVF and the BLAST-based baseline predictor. [c]DeepVF represents the number of positive samples of the independent dataset that appeared in the top 100 ranking list of DeepVF. [d]BLAST represents the number of positive samples of the independent dataset that appeared in the top 100 ranking list of the BLAST-based baseline predictor.

Table S13. Detailed information of DeepVF and the existing toolkits for VF prediction from the user's viewpoint.

Toolkit	**Web server reliability**[a]	**Standalone toolkit**	**User learning document**	**Input data format**	**Number of sequences per submission**	**Computational cost**[b]	**Output presentation**	**Output analysis**	**Output download**
VirulentPred	Sustained	NO	NO	FASTA	<500	~ 11 minutes	Table	None	NO
MP3	Intermittent[a]	YES	YES	FASTA	Unlimited	< 5 seconds	None	None	YES
DeepVF	Sustained	NO	YES	FASTA/ Raw sequence	<5000	~ 0.6 minutes	Table	Result search and rank	YES

[a]The MP3 server has relatively low reliability and could be intermittently accessed.
[b]The computational cost of a web server was evaluated by its executing time per protein sequence, which was obtained by averaging the computational time of 500 test sequences.

2. Supplementary Figures

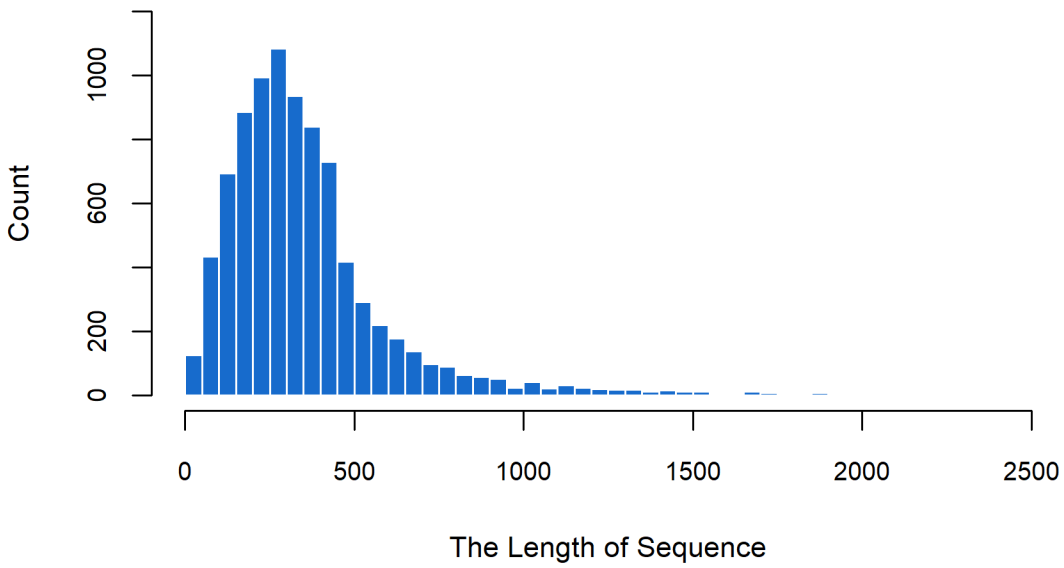

Fig. S1. Distribution of sequence lengths of proteins in the curated dataset.

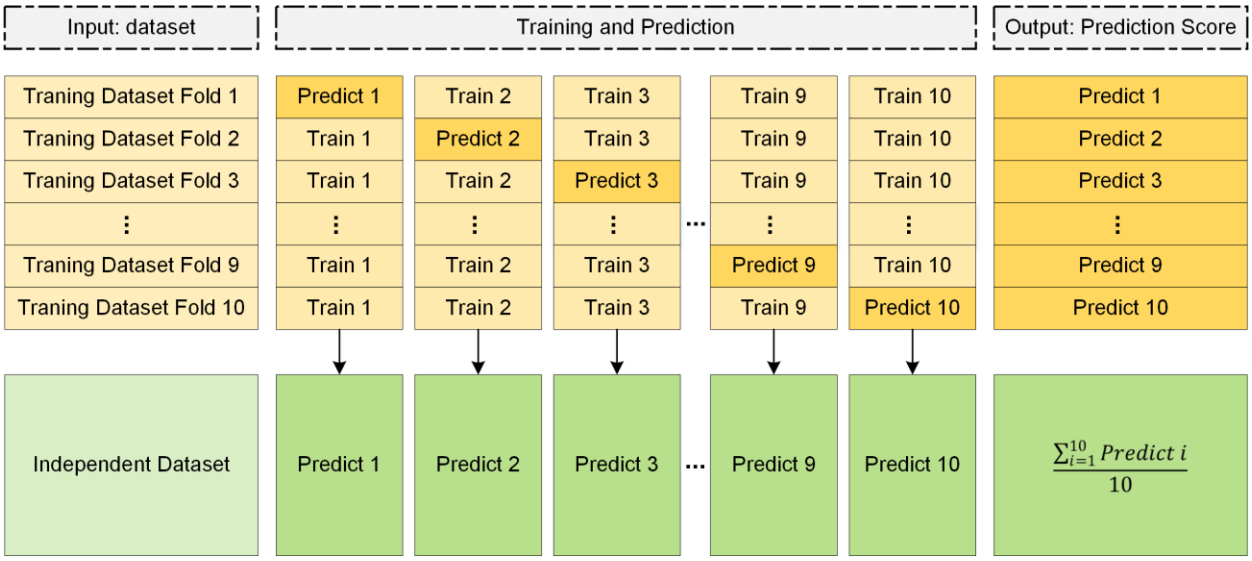

Fig. S2. The pipeline of constructing a baseline model based on the training dataset, and the prediction process on the independent dataset.

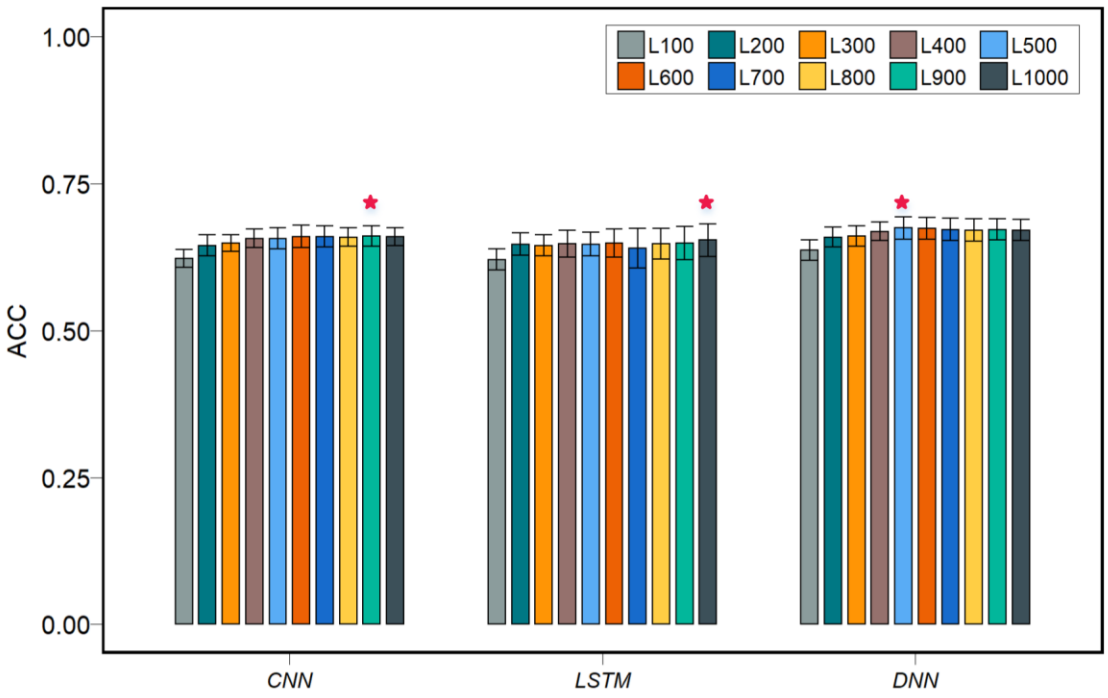

Fig. S3. Performance comparison of different DL models trained using different sizes of protein sequence segments based on the sequence-to-vector feature encoding method.

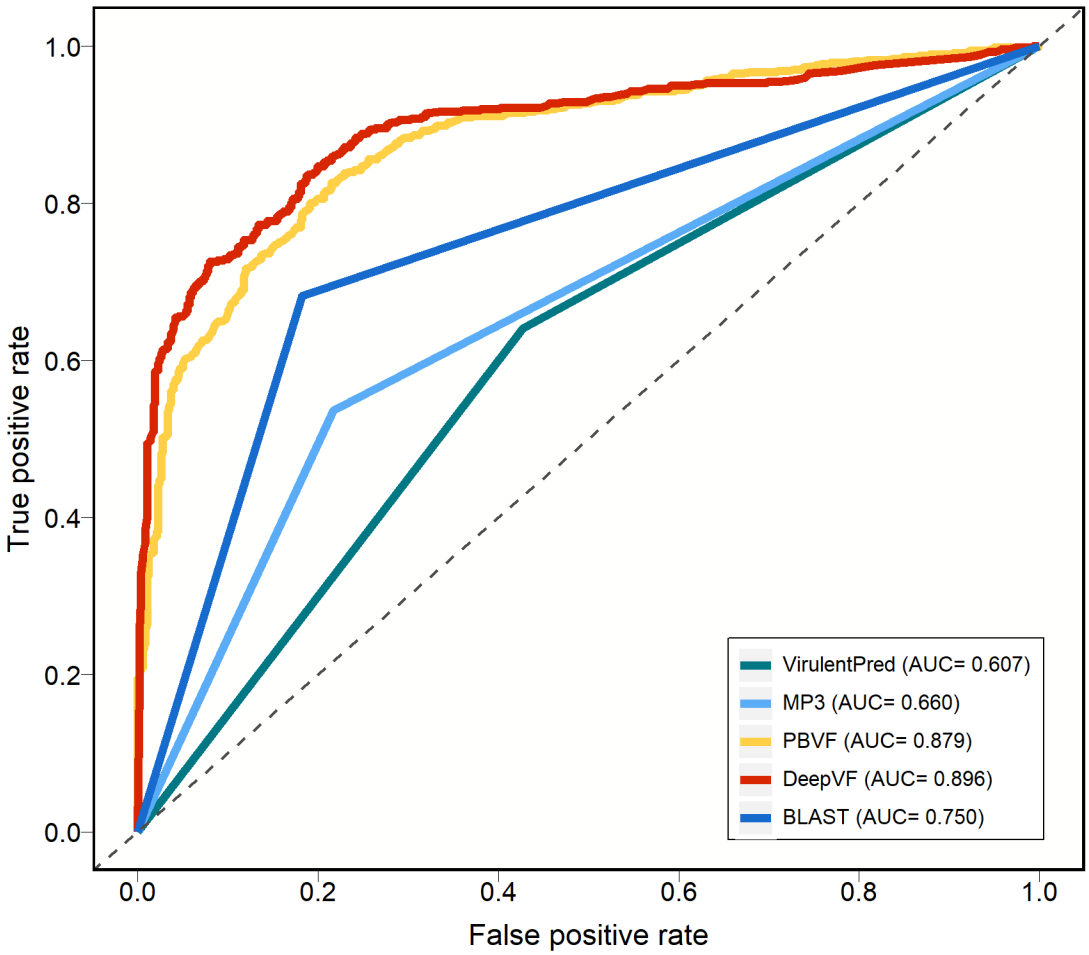

Fig. S4. ROC curves of DeepVF, the BLAST-based baseline predictor and three existing state-of-the-art methods for VF prediction on the independent test. Note that the range of predicting scores of VirulentPred and MP3 are different from that of DeepVF and PBVF, while the BLAST-based baseline predictor only provided a predictive label. Therefore, the ROC curves of those methods were generated by only using their predictive labels, which resulted in three broken lines in this figure.

3. Supplementary References

1. Sachdeva G, Kumar K, Jain P et al. SPAAN: a software program for prediction of adhesins and adhesin-like proteins using neural networks, Bioinformatics 2005;21:483-491.
2. Garg A, Gupta D. VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens, BMC bioinformatics 2008;9:62.
3. Nanni L, Lumini A. An ensemble of support vector machines for predicting virulent proteins, Expert Systems with Applications 2009;36:7458-7462.
4. Tsai C-T, Huang W-L, Ho S-J et al. Virulent-GO: prediction of virulent proteins in bacterial pathogens utilizing gene ontology terms, development 2009;1:3.
5. Nanni L, Lumini A, Gupta D et al. Identifying bacterial virulent proteins by fusing a set of classifiers based on variants of Chou's pseudo amino acid composition and on evolutionary information, IEEE/ACM Trans Comput Biol Bioinform 2012;9:467-475.
6. Zheng LL, Li YX, Ding J et al. A comparison of computational methods for identifying virulence factors, PLoS One 2012;7:e42517.
7. Cui W, Chen L, Huang T et al. Computationally identifying virulence factors based on KEGG pathways, Mol Biosyst 2013;9:1447-1452.
8. Gupta A, Kapil R, Dhakan DB et al. MP3: a software tool for the prediction of pathogenic proteins in genomic and metagenomic data, PLoS One 2014;9:e93907.
9. Rentzsch R, Deneke C, Nitsche A et al. Predicting bacterial virulence factors – evaluation of machine learning and negative data strategies, Brief Bioinform 2019.