Bioinformatics analysis and collection of protein post-translational modification sites in human viruses (Analysis of viral protein post-translational modification sites)

Yujia Xiang^{1,3}, QuanZou^{2*} and Lilin Zhao^{1,4*}

¹ State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, Beijing, China

² Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China

³ University of Chinese Academy of Sciences, Beijing, China

⁴ CAS Center for Excellence in Biotic Interactions, University of Chinese Academy of Sciences, Beijing, China

*Corresponding author

Email: zhaoll@ioz.ac.cn (ZLL) and zouquan@nclab.net (ZQ)

14

15 **Abstract**

16 In viruses, post-translational modifications (PTMs) are essential for their life cycle. Recognizing
17 viral PTMs is very important for better understanding the mechanism of viral infections and finding
18 potential drug targets. However, few studies have investigated the roles of viral PTMs in virus-human
19 interactions using comprehensive viral PTM datasets. To fill this gap, firstly, we developed a viral
20 post-translational modification database (VPTMdb) for collecting systematic information of viral PTM
21 data. The VPTMdb contains 912 PTM sites that integrate 414 experimental-confirmed PTM sites with
22 98 proteins in 45 human viruses manually extracted from 162 publications and 498 PTMs extracted
23 from UniProtKB/Swiss-Prot. Secondly, we investigated the viral PTM sequence motifs, the function of
24 target human proteins, and characteristics of PTM protein domains. The results showed that (i) viral
25 PTMs have the consensus motifs with human proteins in phosphorylation, SUMOylation and
26 N-glycosylation. (ii) The function of human proteins that targeted by viral PTM proteins are related to
27 protein targeting, translation, and localization. (iii) Viral PTMs are more likely to be enriched in
28 protein domains. The findings should make an important contribution to the field of virus-human
29 interaction. Moreover, we created a novel sequence-based classifier named VPTMpre to help users
30 predict viral protein phosphorylation sites. Finally, an online web server was implemented for users to
31 download viral protein PTM data and predict phosphorylation sites of interest.

32 **Author summary**

33 Post-translational modifications (PTMs) plays an important role in the regulation of viral proteins;
34 However, due to the limitation of data sets, there has been no detailed investigation of viral protein
35 PTMs characteristics. In this manuscript, we collected experimentally verified viral protein

36 post-translational modification sites and analysed viral PTMs data from a bioinformatics perspective.
37 Besides, we constructed a novel feature-based machine learning model for predicting phosphorylation
38 site. This is the first study to explore the roles of viral protein modification in virus infection using
39 computational methods. The valuable viral protein PTM data resource will provide new insights into
40 virus-host interaction.

41 **Introduction**

42 Post-translational modifications (PTMs) play a critical role in current proteomics research and
43 regulate protein functions by altering protein interactions, stability, activity, and subcellular localization.
44 Post-translation modifications of viral proteins are relevant throughout various stages of the pathogen
45 life cycle, especially viral infections and genome replication. For example, during entry, the influenza
46 virus carries unanchored ubiquitin chains to engage the host cell's aggresome system [1]. Once inside
47 the host cell, viral PTMs regulating the infecting process of HSV-1 encode ICP0 protein to degrade host
48 proteins via ubiquitination and sumoylation [2]. In the viral life circle, the HIV-1 Tat protein ser-16
49 phosphorylated site regulates HIV-1 transcription [3].

50 Therefore, knowledge of viral PTMs is of great significance to understanding the molecular
51 mechanisms underlying viral infections and recognizing potential drug targets. In recent years, several
52 studies have identified multiple viral PTMs [4-6]; thus, comprehensive analysing these PTM data and
53 establishing a database to provide relevant knowledge is important.

54 However, few databases have been developed for systematically archiving and easily accessing the
55 PTM sites data of viruses. Also, few researchers have been able to draw on any systematic research into
56 viral PTMs using computational methods. VirPTM [7] stores viral phosphorylation sites and used scan-x

57 to predict modification sites. ViralPhos [8] is a support vector machine based predictor and database that
58 provides outdated viral phosphorylation sites. Bradley et al, studied the phosphorylation motifs in 48
59 eukaryotes species and 2 prokaryotic species [9]. To date, no databases have collected comprehensive
60 PTM data of viral proteins and few studies analysed the biological significance behind viral PTM data.

61 To bridge the existing knowledge gap, we have built a viral post-translational modification database
62 (VPTMdb) that first provides comprehensive experimentally verified viral PTM site data, including
63 phosphorylation, sumoylation, glycosylation, acetylation, methylation, ubiquitination, neddylation, and
64 palmitoylation, and it includes 162 studies that have been manually viewed to extract PTM sites. In total,
65 912 PTM sites from 45 human viruses were obtained, which include 414 manual checked sites from
66 PubMed as well as 498 sites from UniProtKB/Swiss-Prot.

67 Secondly, by using computational methods, we investigated the PTM sequence motifs, the function of
68 target human proteins, and characteristics of PTM protein domains. This work will generate fresh insight
69 into viral infection mechanisms as well as identify virus PTM sites.

70 Finally, PTM was predicted in other species with machine learning approaches [10, 11]. For viral
71 protein serine modification site identification, we implemented a novel feature-based classifier named
72 VPTMpre into the VPTMdb to provide users with the ability to find viral protein phosphorylation sites.
73 The results of independent testing showed that VPTMpre represents a powerful tool to predict viral
74 protein phosphorylation sites.

75 The online web server is available at <http://vptmdb.com:8787/VPTMdb/>, and users can browse and
76 download viral PTM data freely. Support vector machine, random forest, and naïve Bayes were
77 integrated into VPTMpre, and users are able to choose one machine learning model to predict possible
78 phosphorylation sites of interest.

79 **Results**

80 **Database contents**

81 **Fig 1** shows that the VPTMdb web server consists of two parts: VPTM database and VPTMpre. The
 82 VPTM database currently includes 414 unique experimentally determined PTM sites with 8
 83 modification types from 45 viruses. In summary, 162 manually checked references were collected in
 84 the database. Each entry in VPTMdb includes the (i) virus name, (ii) virus protein name in the UniProt
 85 database, (iii) PTM type, (iv) viral modification site, (v) residue sequences, (vi) kinase, (vii) a short
 86 description of the PTM site extracted from the publication, and (viii) PubMed id. PTM data from
 87 UniProtKB/Swiss-Prot contain two types: 199 phosphorylation sites and 299 glycosylation sites
 88 (N-lined and O-lined).

89 The statistics of experimentally verified sites in VPTMdb show that among eight PTM types,
 90 phosphorylation sites account for the most (484 sites, including 285 manually checked and 199 sites
 91 from UniProtKB/Swiss-Prot) at more than 50% of the total database. The top five viruses in the
 92 number of manually checked modification sites are HAdV-2 (51 phosphorylation sites), EBOV (29
 93 phosphorylation, 1 sumoylation, 2 ubiquitination, 8 acetylation sites), HIV-1 (21 phosphorylation, 4
 94 sumoylation, 2 ubiquitination, 5 acetylation and 3 glycosylation sites), H1N1 (19 phosphorylation, 3
 95 sumoylation, 2 ubiquitination, 6 acetylation and 2 glycosylation sites), and HCV (10 phosphorylation, 1
 96 sumoylation, 1 ubiquitination, 1 methylation 4 palmitoylation and 14 glycosylation sites) (**S1 Fig**).

97 Human-virus PPI data were included in the VPTMdb, which are helpful to determine the potential
 98 function of PTMs during viral infections. PPI data in the VPTMdb contains 7073 interactions with

99 2934 proteins in 43 viruses. **Fig 2** shows the distribution of modified proteins in the protein-protein
100 interaction network.

101 The web server involves five easy-to-use main pages: 'Home', 'Browse', 'Prediction', 'Download',
102 and 'Help'. Each of these pages enables users to search, browse, predict, and download data without any
103 prerequisite knowledge. In the 'Browse' section, users can search the PTM data conveniently by typing
104 keywords in the search box and download data freely, what is more, virus-human protein-protein
105 interaction data are provided and visualized. The 'Prediction' page provides VPTMpre, a sequence-based
106 machine learning predictor for phosphorylation serine site prediction. All data about virus PTM are stored
107 in the 'Download' page for batched downloading. The 'Help' page contains a detailed tutorial to help
108 users learn about VPTMdb.

109 **Fig 1. Overview of VPTMdb.** Framework of VPTMdb web server construction. First, PTM data were
110 collected from PubMed and UniProt/Swiss-Prot. Then, VPTMpre was constructed to predict viral
111 protein phosphorylation sites.

112 **Fig 2. The virus-human protein-protein interaction network.** Each node represents viral protein or
113 human protein. Each edge represents virus-human or virus-virus association.

114 **Investigation of viral PTM sequence motifs**

115 Previous research has reported that most eukaryotic species have universal kinase-substrate motifs in
116 their phosphorylation proteins [9]. The human viruses are living in the cell, and their proteins are
117 modified by human kinase or viral protein kinase. To this end, we were interested in a question: Are
118 the modified substrate motifs of viral proteins the same as human proteins motifs? To answer this
119 question, we used the motif-x tool [12] to extract motifs from viruses.

120 As shown in **Fig 3**, for viral phosphorylation modified proteins, when kinases were from human
 121 proteins, the viral sequences motifs were the same as human proteins (xSPx) (“x” means any residue)
 122 [9]. For viral protein SUMOylation, we noted that the highly prevalent motif across 16 viruses was
 123 KxE, which was also enriched in human proteins [13]. What’s more, we investigated viral
 124 N-glycosylated proteins’ motifs. The results showed that NxS/T is the significant motif.

125 We also investigated protein motifs when kinases were viral proteins. In VPTMdb, 13 amino acid
 126 residues were modified by viral protein kinases (HSV-1 US3 or HSV-2 UL13). However, there are no
 127 significant motifs when used motif-x tool. Thus, sequence logo was used to visualize PTM sequences
 128 (**S2 Fig**). Unlike human protein kinases, arginine (R) was enriched near the serine site modified by
 129 virus kinase.

130 Overall, these results suggest that the phosphorylation, SUMOylation, and N-glycosylation residues in
 131 viral PTM sequences have the consensus sequence motifs with human PTM proteins. Viruses may use
 132 those short motifs to interact with human proteins and utilize human signal pathways to regulate
 133 themselves replication.

134 **Fig 3. Viral protein PTM motifs discovered by motif-x.**

135 **Function characterization of viral PTM protein target human protein**

136 To investigate how viral PTM proteins influent the human cellular activities, we created
 137 virus-human protein-protein interactions (PPI) network. The virus-human PPI data consist of
 138 virus-human and virus-virus interactions (viruses are these in VPTMdb database). PPI network
 139 includes 2934 proteins and 7073 interactions. The degree was considered as the metric to evaluate the
 140 role of viral proteins in the virus-host PPI network.

141 Firstly, the roles of viral PTM proteins in the PPI network were analysed. Notably, in Influenza A
142 virus(H1N1), HPV-18, HPV-31, HPV-8, HIV-1, HTLV-1, EBOV, SARS-Cov, hRSV, and Vaccinia
143 virus, their all PTM proteins have significant large degrees than average network degrees (**S1 Table**).

144 Then, the Gene Ontology and KEGG enrichment analysis were performed to characterize the
145 function of target human proteins, which may reflect how viral PTM proteins influent human cellular
146 activities. It is interesting to see that the top five enriched KEGG pathways were “Ribosome”,
147 “Spliceosome”, “Proteasome”, “RNA transport” and “Mismatch repair”. It reveals that viruses use
148 human proteins to promote their transcription and modifications. Also, it has been observed that the top
149 ten GO enrichment terms were related to protein targeting, translation, and localization (**S3 Fig**).

150 **Viral PTMs are more likely to be enriched in protein domains**

151 We analysed the domain composition of viral PTM protein. The protein domain data were extracted
152 by HMMER, then 141 domains were obtained and 62 out of 141 domains have modified residues.
153 These domains which have PTM sites were from 57 proteins in 30 viruses. We counted the number of
154 modifications on proteins in the 30 viruses and found that 53.4% of the modifications were distributed
155 in PFAM protein domains. On average, there are 1.33 modification sites per 100 amino acids for the
156 viral PTM proteins, which increased to 2.1 modification sites per 100 amino acids for the viral PTM
157 domains. These results indicated that viral PTMs are more probably enriched in protein domain
158 regions.

159 **Feature-based predictor construction**

160 For viral protein phosphorylation site prediction, we used the feature representative strategy to create
161 a novel classifier. The first step is to compare different features and evaluate their predictive power.
162 The data in **Table 1** show that six features as well as their combinations were evaluated in SVM with a
163 5-fold cross-validation. AUC, F1-score and MCC were used as the performance evaluation indicators.
164 The results declare that the z-scale, which captures the physical-chemical information of amino acids,
165 is the best among the six single features (AUC=0.957, F1-Score=0.887, MCC=0.810). For BINARY,
166 EGAAC and CTriad, their AUC values also achieved above 90.00%. Moreover, when we fused the
167 features, the result showed that ZSCALE combined with AAC features improved the sensitivity,
168 F1-score and AUC by 8.40%, 1.5%, 0.1% compared with individual z-scale features.

169 However, the combination of EGAAC, BINARY, ZSCALE and CTriad features did not significantly
170 enhance the model's performance, which suggests that high-dimensional features may include useless
171 features that weaken the model performance. Among all the features, considering the three evaluation
172 values of F1-score, MCC, AUC and dimensions, the AAC combined with the ZSCALE performed best,
173 and the sensitivity, AUC and F1-score were higher than the single z-scale features. The independent
174 test also shows that AAC combined with ZSCALE features significantly increased the AUC, F1-score,
175 MCC, and Sn by 0.90%, 21.7%, 2.60%, and 25.0%, respectively (**S1 Supporting Information**).

176 Now, it is important to answer two questions: (i) what is the difference between phosphorylation
177 sites and non-phosphorylation sites and (ii) which features contribute most to the viral phosphorylation
178 protein? To this end, we analysed the z-scale feature information between phosphorylation sites and
179 non-phosphorylation sites. Then, we selected the most important features from the combined features
180 with the mRMR method and using svm, random forest and naïve Bayes to perform a predictive
181 evaluation.

182 **Table 1. Comparison of performance between the single features and fused features with the**
183 **mRMR method.**

Features	Dim	Sn	Sp	MCC	F1	AUC
1.AAC	20	0.738	0.739	0.479	0.738	0.821
2.BINARY	460	0.827	0.896	0.732	0.857	0.931
3.ZSCALE	115	0.812	0.985	0.810	0.887	0.957
4.EGAAC	95	0.896	0.850	0.747	0.876	0.901
5.CTDD	195	0.996	0.077	0.184	0.682	0.655
6.CTDC	39	0.735	0.696	0.433	0.720	0.795
7.CTDT	39	0.823	0.712	0.541	0.779	0.827
8.CTriad	343	0.823	0.870	0.694	0.843	0.926
{1,3}	135	0.896	0.908	0.806	0.902	0.958
{2,3}	575	0.873	0.888	0.764	0.881	0.94
{3,4}	458	0.835	0.904	0.743	0.866	0.944
{3,8}	210	0.873	0.896	0.771	0.885	0.943
{3,5,6,7}	388	0.831	0.85	0.681	0.839	0.921
{2,3,4,8}	1013	0.85	0.908	0.762	0.876	0.947
{1,2,3,8}	938	0.742	0.985	0.751	0.844	0.938

184 Note: The first column represents the different feature extraction methods employed in this study. Dim
185 refers to the different dimensions of every feature, and Sn, Sp, MCC, F1 and AUC represent the
186 sensitivity, specificity, Mathews Correlation Coefficient and AUC value, respectively.

187 **Z-scale feature analysis**

188 The z-scale feature based on amino-acids' physical-chemical properties includes five z values.
189 The distribution of amino acid residues around serine sites is able to determine the different
190 physicochemical properties between phosphorylation sites and non-phosphorylation sites. From **Fig 4**,
191 we can see that the z3 values of the phosphorylation sites are smaller than that of the
192 non-phosphorylation sites, implying that a more negative charge occurred around viral protein
193 phosphorylation sites than around non-phosphorylation sites. The results also showed that the z1, z2, z4,
194 and z5 values of the phosphorylation sites are bigger than that of the non-phosphorylation sites. Overall,

the different z-scale compositions surrounding the phosphorylated and non-phosphorylation sites indicate that it is reasonable to choose the z-scale as a feature for prediction.

Fig 4. Comparison of the z-scale in positive and negative datasets. The vertical axis represents the z-scale values. The X-axis represents the five binary sequences.

Performance evaluation

VPTMdb provides three classifiers: support vector machine, random forest and naïve Bayes. Different dimensional features may have different impacts on different predictors. Thus, we selected features of different dimensions using the mRMR algorithm and compared the three classifiers' performance from the 5-fold cross validation (**S1 Supporting Information**).

Fig 5A shows that the maximum AUCs of the svm and random forest are similar. For the random forest and svm, the AUCs increased when more features were selected (random forest: 14-135 features, with $AUC > 0.90$; svm: 27-135 features, with $AUC > 0.90$). However, we observed that the AUCs of naïve Bayes ($AUCs > 0.80$) decreased when more features were added. From a statistical point of view, to prevent the curse of dimensionality, fewer and more meaningful features should be chosen. Taking the above results into consideration, for 68 features, the AUCs of the three predictors perform better, suggesting that 68D is the most meaningful feature among all the features.

To understand the effective of our 68-dimensional features, the T-distributed Stochastic Neighbour Embedding (t-SNE) algorithm was used to visualize the positive and negative samples. A clear distinction was observed between the positive and negative samples, implying that our features selection results are effective (**Fig 5B**).

215 To assess the robustness and performance of the svm, random forest, and naïve Bayes in 68D
216 features, 10-fold random independent tests were performed. The model performance on independent
217 datasets is shown in **Fig 6**, random forest performed better, the average AUC, MCC, F1-score of its are
218 0.744, 0.427, 0.656 respectively. Comparing random forest and PSI-blast (**S1 Supporting**
219 **Information**), the MCC, acc and sp values of random forest are higher than PSI-blast for 6.92%, 2.8%
220 and 19.1%. Taking all indicators into consideration, our method is stable and better performance. We
221 implemented svm, random forest and naïve Bayes into VPTMpre, users can choose them to predict
222 phosphorylation sites of interest.

223 **Fig 5. Feature-based predictor construction.** (A) Five-fold cross-validation performance of the three
224 classifiers on different features. (B) t-SNE visualization of positive and negative data using 68D
225 features.

226 **Fig 6. Independent test results.** Sensitivity, specificity, AUC, MCC and F1-score of the proposed
227 features in three classifiers.

228 Discussion

229 In this work, we constructed VPTMdb, which is the first database that systematically collected
230 experimentally verified viral protein PTMs. Virus-human PPI data were also collected in the VPTMdb
231 to determine PTM sites association functions. These viral protein PTM data provide unique insights
232 into virus-host interactions.

233 Firstly, viruses in VPTMdb have the same substrate motifs as human proteins in phosphorylation (37
234 viruses), SUMOylation (16 viruses) and N-glycosylation (6 viruses). Several studies have shown that
235 viral functional motifs play significant roles in virus life cycles and virus-host interactions [14]; For

instance, SUMOylation motifs can promote viral proteins binding and enhance viruses replication as well as immune evasion [15, 16]. Hence, these conserved sequence motifs in viral proteins may help them to hijack host PTM processes and utilize cellular substance to facilitate virus infections.

Secondly, the function of the viral PTM proteins target human proteins were explored. The results showed that ten viruses PTM proteins have more degrees than the network average degrees. One possible reason is that viral proteins modification processes require the cooperation of multiple other proteins, so modified proteins have more interaction partners. Another possible reason is that PTMs regulate the state of proteins, and modified proteins can perform more functions. For instance, HCV core protein represses transcription of p21 is regulated by the phosphorylation at serine-116 site [17]. These PTMs will significantly change the function and interaction partners of viral proteins. Also, the top ten GO enrichment results of target human proteins were related to binding, which was partially validated that PTM proteins tend to bind with more human proteins.

Moreover, we found that viral PTM sites are more likely to be enriched in the protein domains; Studies have shown that human modified lysines are more likely near phosphorylation sites, which form a PTM cluster region [18]. For viruses, these cluster PTMs in protein domains may form short motifs to enhance the regulate function of viral proteins.

Finally, based on the analysis of viral PTM protein features, VPTMpre, a novel feature representative classifier, was developed to predict viral protein serine sites. We compared various feature extraction methods and selected the optimized features using the mRMR algorithm. The feature analysis results showed that 68D was able to distinguish the phosphorylation sites and non-phosphorylation sites in viral proteins. VPTMpre was integrated into the VPTMdb web server to provide an online phosphorylation site prediction service. Users can choose three classifiers (svm,

random forest and naïve Bayes) to predict phosphorylation sites of interests. However, because of data limitations, the prediction of VPTMpre is limited to serine sites. With a continuous collection of new viral PTM data, we expect that VPTMpre will be extended to predict more types of PTM sites and obtain a better performance.

In the future, to respond to the rapid growth of viral PTM data, VPTMdb will be updated regularly and more viral PTM-related data collected to ensure that it provides the most comprehensive information to users. As the first attempt to develop the comprehensive viral PTM database, we sincerely welcome support and suggestions from the research community to improve the VPTMdb database.

Methods

Data collection

There are three major steps in data collecting and pre-processing, which are described below.

Firstly, we queried PubMed using the keyword search terms: (virus name) and (eight modification types) for studies published before Jan 01, 2020. As a result, 6052 papers were obtained, each of which was manually retrieved using the following standards: (i) the viral post-translational modifications were experimentally verified; and (ii) if two references contained the same PTM site, the earliest published study was retained. In total, 45 viruses, 162 papers and 414 PTMs were obtained.

Subsequently, 498 viral PTM data points from UniProtKB/Swiss-Prot were integrated into VPTMdb. For experimentally validated virus PTM types, the sites were extracted manually from the articles mentioned above. The protein sequences, UniProt ID and PMID were mainly extracted from NCBI, UniProt and PubMed. Finally, human-virus protein-protein interactions were collected from the

279 VirHostNet based on viral strains in the VPTMdb.

280 **PTM data analysis**

281 The phosphorylation (37 viruses), SUMOylation (16 viruses) and N-Glycosylation (6 viruses) data
282 were from VPTMdb. Motif-x tool was employed to extract motifs using its default parameters
283 (score-threshold of 1×10^{-6} , min-occurrences of 5, and width of 15). Proteins domains were searched
284 by HMMER (using PFAM database) with default parameters. PPI data were downloaded from
285 VirHostNet database. Gene Ontology and KEGG enrichment analysis used clusterProfiler [19].
286 Network analysis was performed using Cytoscape [20].

287 **Overview of viral phosphorylation sites prediction**

288 Identifying viral protein PTM sites by experimental methods is still expensive and time consuming.
289 Thus, predicting them in *silico* using bioinformatics approaches is necessary. To this end, a
290 sequence-based classifier named VPTMpre was created to predict viral post-translational modification
291 serine sites. Because threonine and tyrosine data are too few to train the model, we only predicted
292 serine sites in this study.

293 Five main procedures were performed to build the VPTMpre predictor. (i) a balanced benchmark
294 dataset was constructed using the Synthetic Minority Oversampling Technique (SMOTE) [21]
295 sampling method (**S1 Supporting Information**); (ii) various feature representative methods were
296 compared to obtain an effective feature representation strategy, with support vector machine used as the
297 base classifier in a 5-fold cross-validation approach to find the best feature groups; (iii) the predictive
298 performance of three classifiers (svm, random forest, naïve Bayes) on different feature dimensions was

299 compared using the Minimum redundancy and maximum relevance (mRMR) method, and the features
300 that performed well in all three classifiers were selected as the most meaningful and significant features;
301 (iv) a 10-fold random independent test was performed to evaluate the predictive performance of the
302 three different classifiers (svm, random forest, naïve Bayes); and (v) VPTMpre was implemented in the
303 online web server.

304 **Data preparation and processing**

305 All viral phosphorylation experimentally verified serine sites in our database were used as positive
306 samples, and those not marked by any phosphorylation information on the same protein were
307 considered negative samples. As a result, we obtained 182 phosphorylated serine residues as well as
308 2148 non-phosphorylated residues. Phosphorylation sites from UniProtKB/Swiss-Prot were regarded as
309 the independent dataset, and they included 93 positive serine sites and 1878 negative serine sites. After
310 using CD-HIT (clustering thresholds set to 0.8) [22] to remove redundant sequences, we obtained 129
311 positive sites and 1611 negative sites. The independent dataset contained 52 positive sites and 1072
312 negative sites (**Table 2**). These sequences were truncated to a 23-residue symmetrical window (-11 to
313 11).

314 In order to eliminate the prediction bias caused by data imbalance, we re-sampled the training data by
315 SMOTE methods and obtained 260 positive sites and 260 negative sites, which consisted of the training
316 dataset. The negative test set from UniProtKB/Swiss-Prot was randomly divided into twenty parts (**S2**
317 **Table**). We randomly select ten negative subsets from the twenty parts and combined them with ten
318 replicate positive sets to constitute ten independent test datasets (**S1 Supporting Information**).

319 **Table 2. Summary of training and independent datasets**

Datasets	Types	Total number	After deletion	After balanced
Training set	Positive	182	129	260
	Negative	2148	1611	260
Independent set	Positive	93	52	52
	Negative	1878	1072	1072

320 Feature representation

321 To achieve a better classification effect, a key step is feature extraction, which means that a protein
322 sequence is encoded as a numeric vector for machine learning model.

323 **Amino acid composition (AAC).** AAC is the frequency of 20 amino acids for a given sequence [23].

324 This descriptor can be denoted as follows:

$$AAC = (A1, A2, A3, \dots, A20) \quad (1)$$

325 where

$$Ai = \frac{Ri}{L} (i = 1, 2, 3, \dots, 20) \quad (2)$$

326 Ri is the observed number of types i amino acid in a protein sequence. L is the length of protein. Thus

327 20 features were obtained, and sum of which is 1.

328 **Binary profile.** The binary profile transformed each amino acid into a 20-dimensional binary

329 numerical vector. For instance, the alanine ('A') is deciphered as 10000000000000000000, cysteine ('C')

330 is deciphered as 01000000000000000000, etc. Consequently, we obtained a 460-dimensional vector for

331 this binary profile feature.

332 **Conjoint triad (CTriad).** The conjoint triad feature is sequence information for proteins. Twenty

333 amino acid types are clustered into seven classes to construct the C-triad feature.

$$group1 = \{Ala, Gly, Val\}, group2 = \{Ile, Leu, Phe, Pro\} \quad (3)$$

$$group3 = \{Tyr, Met, Thr, Ser\},$$

$$group4 = \{His, Asn, Gln, Trp\}$$

$$group5 = \{Arg, Lys\}, group6 = \{Asp, Glu\},$$

$$group7 = \{Cys\}$$

334 First, protein sequences are encoded into a numerical vector using the AA groups list above.

335 Subsequently, any three continuous AAs are regarded as a unit, and scanning along the sequences and

336 counting the frequencies of each triad type is performed to obtain a 343-dimensional numerical vector.

337 For example, a protein sequence S contains L AA residues, which are expressed as follows:

$$S = A_1 A_2 A_3 A_4 A_5 \dots A_L. \quad (4)$$

338 Then, we scan along the sequence with a slide window in three continuous residues:

$$A_1 A_2 A_3, A_2 A_3 A_4, A_3 A_4 A_5, A_4 A_5 A_6, \dots, A_{L-2} A_{L-1} A_L \quad (5)$$

339 Finally, the C-triad feature of a protein is defined as the frequency of the corresponding triad type in that

340 protein:

$$C_{triad} = [f_1, f_2, f_3, f_4, \dots, f_{343}]^T \quad (6)$$

341 where,

$$f_i = \frac{n_i}{L-2} \quad (7)$$

342 n_i is the occurrence number of the i-th triad type ($i=1, 2, \dots, 343$).

343 More detailed information about C-triad can be found in [24].

344 **Composition-Transition-Distribution (CTD).** CTD clusters 20 amino acids into three groups:

345 hydrophobic, neutral and polar. The CTD composition (CTD-C) calculates the composition values of

346 hydrophobic, neutral and polar groups for a given sequence. The CTD transition (CTD-T) represents the

347 percentage frequency of an amino acid of one particular property followed by an amino acid of another

348 property. The CTD distribution (CTD-D) represents the distribution of each property for a given

349 sequence. Each property has five distribution descriptors, which are the first residue, 25% residues, 50%

residues, 75% residues, and 100% residues in the whole sequence of a given specific property. In this research, CTD-C, CTD-T, and CTD-D were used to encoded protein sequences and yielded 39, 39, and 195 features, respectively. More detailed information about CTD can be found in the literature [25].

Enhanced grouped amino acid composition (EGAAC). EGAAC was first proposed by Chen et al. [26] and is the improved version of GAAC features. GAAC divides 20 standard amino acids into five groups based on their physical and chemical properties. The formulation of GAAC is as follows:

$$f(g) = \frac{N(g)}{L}, g \in \{g1, g2, g3, g4, g5\} \quad (8)$$

$$N(g_i) = \sum N_i, i \in g \quad (9)$$

$$g1 = \{GAVLMI\}, g2 = \{FYW\}, \quad (10)$$

$$g3 = \{KRH\}, g4 = \{DE\},$$

$$g5 = \{STCPNQ\}$$

where L is the length of sequence, $N(g)$ is the number of amino acids in group g , and N_i is the occurrence number of i -th amino acid type.

EGAAC scans along the sequence and calculates the GAAC values in a fixed-size window:

$$F(g) = \frac{N(g, win)}{N(win)}, g \in \{g1, g2, g3, g4, g5\} \quad (11)$$

where $N(g, win)$ is the number of amino acids in group g within a fixed-size window win and $N(win)$ is the window size. win ranges from 1 to 17. In this study, the window size was set to 5, and we finally obtained a 95-dimensional vector.

Z-Scale (ZSCALE). Z-scale is a feature descriptor that describes AAs' physicochemical properties. It was first published by Hellberg [27], who introduced three z-scales (z1-z3), and then Sandberg et al. (Sandberg, et al., 1998) improved the original z-scale features by adding two more z-scale values, using 26

properties of 87 AAs. In this study, we employed the z-scale using five scales(z1-z5). The five z-scales are based on lipophilicity (z1), bulk (z2), polarity/charge (z3), electronegativity and heat of formation(z4), electrophilicity and hardness(z5), yielding a 115-dimensional numerical vector.

Feature selection and optimization

Generally, high-dimension biological features may be noisy, which led to poor prediction performance. However, feature selection is a good strategy to overcome feature redundancy. Feature selection means using a reduction algorithm to select the major features that are able to improve the performance of specific classifiers.

In this work, six descriptors and their combined features' performance were compared using 5-fold cross validation in the training data with the Support Vector Machine (SVM) method. Subsequently, the Minimum redundancy and maximum relevance (mRMR) method was chosen to select the most meaningful features. To investigate the predictive performance of three classifiers, we compared the different dimensions of features in the svm, random forest, naïve Bayes methods. The features that performed well in all three classifiers were selected as the most meaningful and significant features. The T-distributed Stochastic Neighbour Embedding algorithm was used to visualize the features[28].

Performance evaluation

Sensitivity (Sn), Specificity (Sp), F1-score, and Mathews Correlation Coefficient (MCC) were applied to estimate the prediction performance (**S1 Supporting Information**). Besides, the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) were used to evaluate the overall performance of the model. The ROC curve is a continuous line plotted by the false positive rate (FPR) as

the X-coordinate and true positive rate (TPR) as the Y-coordinate. The higher the AUC value, the better the performance of the classifier.

Website implementation

The VPTMdb web interface was written in the R programming language using the Rshiny web development framework [29]. The MySQL database management system was used to store structured PTM data. The base machine learning predictor (such as SVM) was supported by the caret R package [30]; the ROC curve was analysed using ROCR [31]; and MRMR and t-SNE were analysed using mRMRe [32] and Rtsne [33]. Software ggplot2 was used to plot beautiful pictures [34]. The website is free and can be browsed in most modern browsers.

Acknowledgments

Thanks for the anonymous reviewers for their kind suggestions.

References

1. Banerjee I, Miyake Y, Nobs SP, Schneider C, Horvath P, Kopf M, et al. Influenza A virus uses the aggresome processing machinery for host cell entry. *Science* (New York, NY). 2014 Oct 24;346(6208):473-7. PubMed PMID: 25342804. Epub 2014/10/25. eng.
2. Randow F, Lehner PJ. Viral avoidance and exploitation of the ubiquitin system. *Nature cell biology*. 2009 May;11(5):527-34. PubMed PMID: 19404332. Epub 2009/05/01. eng.
3. Ivanov A, Lin X, Ammosova T, Ilatovskiy AV, Kumari N, Lassiter H, et al. HIV-1 Tat phosphorylation on Ser-16 residue modulates HIV-1 transcription. *Retrovirology*. 2018 May 23;15(1):39. PubMed PMID: 29792216. PMCID: PMC5966876. Epub 2018/05/25. eng.

406 4. Kulej K, Avgousti DC, Sidoli S, Hermann C, Della Fera AN, Kim ET, et al. Time-resolved
407 Global and Chromatin Proteomics during Herpes Simplex Virus Type 1 (HSV-1) Infection.
408 Molecular & cellular proteomics : MCP. 2017 Apr;16(4 suppl 1):S92-s107. PubMed PMID:
409 28179408. PMCID: PMC5393384. Epub 2017/02/10. eng.

410 5. Scaturro P, Stukalov A, Haas DA, Cortese M, Draganova K, Plaszczyca A, et al. An
411 orthogonal proteomic survey uncovers novel Zika virus host factors. Nature. 2018
412 Sep;561(7722):253-7. PubMed PMID: 30177828. Epub 2018/09/05. eng.

413 6. Zheng J, Yamada Y, Fung TS, Huang M, Chia R, Liu DX. Identification of N-linked
414 glycosylation sites in the spike protein and their functional impact on the replication and
415 infectivity of coronavirus infectious bronchitis virus in cell culture. Virology. 2018 Jan
416 1;513:65-74. PubMed PMID: 29035787. Epub 2017/10/17. eng.

417 7. Schwartz D, Church GM. Collection and motif-based prediction of phosphorylation sites in
418 human viruses. Science signaling. 2010 Aug 31;3(137):rs2. PubMed PMID: 20807955. Epub
419 2010/09/03. eng.

420 8. Huang KY, Lu CT, Bretana N, Lee TY, Chang TH. ViralPhos: incorporating a recursively
421 statistical method to predict phosphorylation sites on virus proteins. BMC bioinformatics.
422 2013;14 Suppl 16:S10. PubMed PMID: 24564381. PMCID: PMC3853219. Epub 2014/02/26.
423 eng.

424 9. Bradley D, Beltrao P. Evolution of protein kinase substrate recognition at the active site.
425 PLOS Biology. 2019;17(6):e3000341.

426 10. He W, Wei L, Zou Q. Research Progress in Protein Post-Translational Modification Site
427 Prediction. Briefings in Functional Genomics. 2018;18(4):220-9.

- 428 11. Huang GH, Li JC. Feature Extractions for Computationally Predicting Protein
429 Post-Translational Modifications. *Current Bioinformatics*. 2018;13(4):387-95. PubMed PMID:
430 WOS:000437860800009. English.
- 431 12. Cheng A, Grant CE, Noble WS, Bailey TL. MoMo: discovery of statistically significant
432 post-translational modification motifs. *Bioinformatics (Oxford, England)*. 2018;35(16):2774-82.
- 433 13. Yang SH, Galanis A, Witty J, Sharrocks AD. An extended consensus motif enhances the
434 specificity of substrate modification by SUMO. *The EMBO journal*. 2006 Nov 1;25(21):5083-93.
435 PubMed PMID: 17036045. PMCID: PMC1630412. Epub 2006/10/13. eng.
- 436 14. Sobhy H. A Review of Functional Motifs Utilized by Viruses. *Proteomes*. 2016 Jan 21;4(1).
437 PubMed PMID: 28248213. PMCID: PMC5217368. Epub 2016/01/21. eng.
- 438 15. Wimmer P, Schreiner S. Viral Mimicry to Usurp Ubiquitin and SUMO Host Pathways.
439 *Viruses*. 2015 Aug 28;7(9):4854-72. PubMed PMID: 26343706. PMCID: PMC4584293. Epub
440 2015/09/08. eng.
- 441 16. Hickey CM, Wilson NR, Hochstrasser M. Function and regulation of SUMO proteases.
442 *Nature reviews Molecular cell biology*. 2012 Dec;13(12):755-66. PubMed PMID: 23175280.
443 PMCID: PMC3668692. Epub 2012/11/24. eng.
- 444 17. Jung EY, Lee MN, Yang HY, Yu D, Jang KL. The repressive activity of hepatitis C virus
445 core protein on the transcription of p21(waf1) is regulated by protein kinase A-mediated
446 phosphorylation. *Virus research*. 2001 Nov 5;79(1-2):109-15. PubMed PMID: 11551651. Epub
447 2001/09/12. eng.

448 18. Beltrao P, Albanèse V, Kenner Lillian R, Swaney Danielle L, Burlingame A, Villén J, et al.
449 Systematic Functional Prioritization of Protein Posttranslational Modifications. *Cell*. 2012
450 2012/07/20;150(2):413-25.

451 19. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R Package for Comparing Biological
452 Themes Among Gene Clusters. 2012;16(5):284-7. PubMed PMID: 22455463.

453 20. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a
454 software environment for integrated models of biomolecular interaction networks. *Genome*
455 *research*. 2003 Nov;13(11):2498-504. PubMed PMID: 14597658. PMCID: PMC403769. Epub
456 2003/11/05. eng.

457 21. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority
458 Over-sampling Technique. *Journal of Artificial Intelligence Research*. 2002;16(1):321-57.

459 22. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation
460 sequencing data. *Bioinformatics (Oxford, England)*. 2012 Dec 1;28(23):3150-2. PubMed PMID:
461 23060610. PMCID: PMC3516142. Epub 2012/10/13. eng.

462 23. Liu B. BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on
463 machine learning approaches. *Briefings in bioinformatics*. 2019 Jul 19;20(4):1280-94. PubMed
464 PMID: 29272359. Epub 2017/12/23. eng.

465 24. Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, et al. Predicting protein-protein interactions
466 based only on sequences information. *Proceedings of the National Academy of Sciences of the*
467 *United States of America*. 2007 Mar 13;104(11):4337-41. PubMed PMID: 17360525. PMCID:
468 PMC1838603. Epub 2007/03/16. eng.

469 25. Govindan G, Nair AS, editors. Composition, Transition and Distribution (CTD) — A
470 dynamic feature for predictions based on hierarchical structure of cellular sorting. 2011 Annual
471 IEEE India Conference; 2011 16-18 Dec. 2011.

472 26. Chen Z, Zhao P, Li F, Leier A, Marquez-Lago TT, Wang Y, et al. iFeature: a Python
473 package and web server for features extraction and selection from protein and peptide
474 sequences. *Bioinformatics* (Oxford, England). 2018 Jul 15;34(14):2499-502. PubMed PMID:
475 29528364. PMCID: PMC6658705. Epub 2018/03/13. eng.

476 27. Hellberg S, Sjostrom M, Skagerberg B, Wold S. Peptide quantitative structure-activity
477 relationships, a multivariate approach. *Journal of medicinal chemistry*. 1987 Jul;30(7):1126-35.
478 PubMed PMID: 3599020. Epub 1987/07/01. eng.

479 28. van der Maaten LJP, Hinton GE. Visualizing High-Dimensional Data Using t-SNE. *Journal*
480 *of Machine Learning Research*. 2008;9:2579-605.

481 29. Chang W, Cheng J, Allaire J, Xie Y, McPherson J. shiny: Web Application Framework for
482 R. 2018.

483 30. Kuhn M. caret: Classification and Regression Training. 2020.

484 31. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance
485 in R. *Bioinformatics* (Oxford, England). 2005 Oct 15;21(20):3940-1. PubMed PMID: 16096348.
486 Epub 2005/08/13. eng.

487 32. De Jay N, Papillon-Cavanagh S, Olsen C, El-Hachem N, Bontempi G, Haibe-Kains B.
488 mRMRe: an R package for parallelized mRMR ensemble feature selection. *Bioinformatics*
489 (Oxford, England). 2013 Sep 15;29(18):2365-8. PubMed PMID: 23825369. Epub 2013/07/05.
490 eng.

- 491 33. Krijthe JH. Rtsne: T-Distributed Stochastic Neighbor Embedding using a Barnes-Hut
492 Implementation. 2015.
- 493 34. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.
494 2016.
495
496

497 **Supporting information**

498 **S1 Fig. Statistics of viral PTM data in VPTMdb.**

499 **S2 Fig. Viral protein kinase substrate motifs.** The HSV-1, and HSV-2 PTM amino acid residues were
500 modified by US3 and UL13.

501 **S3 Fig. The results of KEGG and Gene Ontology enrichment analysis.**

502 **S1 Table. The results of network analysis.**

503 **S2 Table. Training and independent datasets.**

504 **S1 Supporting Information. Supplementary materials.**

A

Classifier — naive bayes — random forests — svm

B

Sample ● Positive ● Negative