1

2 **Machine learning prediction of antiviral-HPV protein** 3 **interactions for anti-HPV pharmacotherapy**

4 Hui-Heng Lin ^{a*} (ORCID 0000-0003-4060-7336), Qian-Ru Zhang ^b (ORCID
5 0000-0001-8986-5219), Xiangjun Kong ^c, Liuping Zhang ^d (ORCID
6 0000-0001-7696-0161), Yong Zhang ^e (ORCID 0000-0003-1820-0927), Yanyan Tang ^f
7 (ORCID 0000-0002-1189-4498) and Hongyan Xu ^{d*} (ORCID 0000-0003-2402-0973)

8

9 ^a Yuebei People's Hospital, Shantou University Medical College, No. 133 of Huimin
10 South road, Wujiang District, Shaoguan City, Guangdong Province, 512025, China.

11 ^b School of Pharmacy, Key Lab of the Basic Pharmacology of the Ministry of
12 Education, Zunyi Medical University, 6 West Xue-Fu Road, Zunyi city, Guizhou
13 Province, 563000, China

14 ^c State Key Laboratory of Quality Research in Chinese Medicine, Institute of
15 Chinese Medical Sciences, University of Macau, Avenida de Universidade, Taipa,
16 Macau, 999078, China.

17 ^d Department of Gynecology, Yuebei People's Hospital, Shantou University Medical
18 College, No. 133 of Huimin South road, Wujiang District, Shaoguan City, Guangdong
19 Province, 512025, China.

20 ^e Interdisciplinary Research Center for Agriculture Green Development in Yangtze
21 River Basin, Southwest University, No.1-2-1 Tiansheng Road, Beibei District,

22 Chongqing, 400715 China.

23 ^f Department of Neurology, the First Affiliated Hospital of Guangxi Medical
24 University, No.6 Shuangyong Road, Nanning, Guangxi, 530021 China.

25

26 * Correspondence to Hui-Heng Lin scienceds@outlook.com and Hongyan Xu:
27 medxh@outlook.com . Corresponding Postal Address: Yuebei People's hospital, No.
28 133 of Huimin South road, Wujiang District, Shaoguan City, Guangdong Province,
29 512025 China.

30

31 Email of coauthors:

- 32 ● Qian-Ru Zhang: qrzhang_iris@foxmail.com
- 33 ● Xiangjun Kong: xiangjunkong@um.edu.mo
- 34 ● Liuping Zhang: lpzhang0912@163.com
- 35 ● Yangyan Tang: yuji0301@126.com
- 36 ● Yong Zhang: bioyongzhang@swu.edu.cn

37

38 Abstract

39 **Background:** Persistent infection with high-risk types Human Papillomavirus could
40 cause diseases including cervical cancers and oropharyngeal cancers. Nonetheless, so
41 far there is no effective pharmacotherapy for treating the infection from high-risk HPV
42 types, and hence it remains to be a severe threat to health of female.

43 **Methods:** In light of drug repositioning strategy, we trained and benchmarked multiple
44 machine learning predictive models so as to predict potential effective antiviral drugs
45 for HPV infection in this work. Based on antiviral-target interaction dataset, we
46 generated high dimension feature set of drug-target interaction pairs and used the
47 dataset to train and construct machine learning predictive models.

48 **Results:** Through optimizing models, measuring models' predictive performance using
49 182 pairs of antiviral-target interaction dataset which were all approved by United
50 States Food and Drug Administration, and benchmarking different models' predictive
51 performance, we identified the optimized Support Vector Machine and K-Nearest
52 Neighbor classifier with high precision score were the best two predictors (0.80 and
53 0.85 respectively) amongst classifiers of Support Vector Machine, Random forest,
54 Adaboost, Naïve Bayes, K-Nearest Neighbors, and Logistic regression classifier. We
55 applied these two predictors together and successfully predicted 58 pairs of
56 antiviral-HPV protein interactions from 846 pairs of antiviral-HPV protein associations.

57 **Conclusions:** Our work provided good drug candidates for anti-HPV drug discovery. So
58 far as we know, we are the first one to conduct such HPV-oriented computational drug
59 repositioning study.

60

61 **Keywords**

62 Machine Learning Prediction; Human Papillomavirus; HPV infection; Antivirals; Drug
63 repositioning; Drug-Protein interaction prediction; Support Vector Machine; K-Nearest

64 Neighbors

65

66 **Introduction**

67 Human Papillomavirus (HPV) can infect human body and cause different types of
 68 phenotypes. Specifically, HPVs can infect females' reproductive system and causes
 69 different types of gynecological diseases. For instance, a variety of warts and the genital
 70 cancers [1]. What's more, it is reported that HPV infection is one of the risk factors of
 71 oropharyngeal cancer [2]. Researchers have classified the subtypes of HPVs into the
 72 low-risk types, and high-risk types according to their virulence and relevant risk levels
 73 of infections. For low-risk types [3], e.g., the type 6, 11, 40, etc, they might be
 74 disappeared after several periods of infection, and the infected hosts might generally be
 75 fine. While for those high-risk subtypes of HPV, e.g., the HPV-16 and HPV-18, their
 76 persistent infection on hosts could finally result to severe disease like cervical cancer [4],
 77 which could results to the death of hosts. According to report, it is estimated that
 78 569,000 cases of cervical cancer newly occurred in 2018 globally and 311,000 deaths
 79 were found [5]. Therefore, the HPV infection remains a large threat to female's health
 80 especially in developing countries, and treating HPV infection remains an urgent task
 81 and difficult challenge due to there lacks effective pharmacotherapy. Though HPV
 82 vaccines are available, they are ineffective for those who has already been infected by
 83 HPVs [6].

84

85 Scientists have been trying hard to combat against HPV infection. For instance, several
86 researchers have identified HPV's E6 and E7 proteins to be the virulent tumorigenesis
87 risk factors [7, 8], and parts of their molecular 3D structural conformations have
88 revealed through approaches of *in silico* simulation [9] and structural biology [10]. Other
89 studies have tested, discussed and reviewed the *in vitro* effects of existed drug, i.e., the
90 Human Immunodeficiency Virus (HIV) protease inhibitor, on HPV proteins and cells
91 infected with HPVs [11, 12, 13, 14, 15]. These reports targeting existed drugs for HPV
92 treatments showed that, compared with *de novo* drug discovery, repositioning existed
93 drugs is indeed the better and quicker strategy.

94

95 Nonetheless, drug efficacies from above evidences were moderate and no further
96 progress is seen in later stages, .e.g., in clinical contexts. And hence, above research
97 progresses are yet far from being able to identify drug candidates with good therapeutic
98 and anti-HPV potential. Limitation of them could be due to two reasons. One is that
99 inappropriate compound or drug candidates have been chosen for testing. The other
100 reason could be the number of drug candidates to be tested is too small. Testing only
101 limited numbers of compound or drug candidates surely restricts the probability of
102 identifying those appropriate ones.

103

104 In order to meet the urgent needs for effective anti-HPV drug discovery, based on

target-oriented drug repositioning strategy, we collected and analyzed 96 antiviral drugs to do the relatively large-scale *in silico* screening for 9 HPV-16 proteins, so as to computationally and effectively identify effectively antivirals with good potential for targeting HPV proteins. Briefly, in this work, we constructed, benchmarked and selected machine learning predictive models (also known as predictors) to predict antivirals that could have potential interactions with HPV proteins. This is because Drug-target interactions are vital prerequisite of molecular therapeutic mechanisms. Through benchmarking, we selected the high-precision K-Nearest Neighbor (KNN) [16] and Support vector machine (SVM) [17] predictors to detect those confidence interaction pairs of antiviral-HPV protein.

To the best of our knowledge, no prior study similar to our work has been done. Lots of researchers predicted targets of drugs, compound-protein interactions or protein-protein interactions using machine learning or other computational methods [18, 19, 20, 21]. However, so far as we know, no study has focused on studying relationships between antiviral drugs and HPV proteins.

Methods

Research question formulation

Theoretically, a therapeutic target and its drug molecule have interactive binding relation to each other. Therefore, trying to identify potential HPV protein targets of

antivirals could be considered as a binary classification task, i.e., to predictively classify proteome of HPV into two classes of proteins. One class is HPV proteins which have potential interaction with drug molecules, and the other class is HPV proteins do not have potential interactions with drugs. Machine learning is a good and state-of-art methods to solve such binary classifications (**Figure 1**). Considered that known antiviral drug-target interaction pairs were available, which could serve as the known-label validation dataset, we thus chose supervised (machine) learning methods for this study.

Figure 1: Research framework of this study. Predicting antiviral drug-HPV protein interaction could be considered a binary classification task, and machine learning is a good method for such task. In this work, antiviral drug-target pairs' features were transformed into vectors for constructing machine learning predictors. Through benchmarking, the best predictors were selected to predict antiviral-HPV protein

138 interactions.

139

140

141 **Data collection and preprocessing**

142 We collected antiviral drugs and their associated target data from DrugBank [22],
 143 Drugs@FDA [23], PubChem [24], Uniprot [25] and Therapeutic Targets [26] databases
 144 (As of 19th July 2020). Drug-target interaction pairs which contained the United States
 145 Food and Drug Administration (FDA)-approved antiviral drugs were treated as the
 146 validation dataset for machine learning, while the rest drug-target interaction pairs were
 147 treated as the training dataset for machine learning. In this work's machine learning
 148 classification task, an interaction pair of an antiviral drug and a protein was defined to
 149 be a positive instance, while negative instance indicated a non-interactive pair of
 150 antiviral and protein. In order to balance data ratio for binary machine learning
 151 classification task. We randomly generated non-interactive drug-target pairs so as to
 152 assure the 1:1 ratio of positive instances to negative instances for machine learning.
 153 Next, we integrated the proteome (9 proteins in total) of high-risk HPV-16 subtype and
 154 all the antiviral drugs to form drug-protein interaction prediction dataset. See
 155 **Supplementary Table 1** for machine learning training dataset of antiviral drug-target
 156 interaction pairs, **Supplementary Table 2** for drug-target interaction pair dataset used in
 157 machine learning validation process, and **Supplementary Table 3** for Uniprot's HPV-16
 158 proteome, i.e., 9 proteins.

159

160 Next, all antivirals' molecular structures were analyzed using ChemmineR [27] and
 161 1024-dimension chemical fingerprint datasets were generated through R scripting [28].
 162 All proteins were analyzed using ProtR [29] and 10784 high dimension protein
 163 descriptor feature datasets were generated. As seen in **Table 1**. Descriptors used were
 164 protein structural and physicochemical properties. These descriptors have been widely
 165 used in and performed well for analyzing protein-protein interactions and protein-ligand
 166 interactions [29].

167

168 All datasets were integrated, scaled and normalized using R computing environment
 169 [28].

170

171 **Table 1:** Molecular descriptors used for machine learning analysis

ID	Molecular descriptor	Vector length
1	Drug molecule fingerprint	1024
2	Amino acid composition	20
3	Dipeptide composition	400
4	Tripeptide composition	8000
5	Normalized Moreau-Broto Autocorrelation	240
6	Moran Autocorrelation	240
7	Geary Autocorrelation	240

8	Composition descriptor	21
9	Transition descriptor	21
10	Distribution descriptor	105
11	Pseudo amino acid composition	50
12	Amphiphilic pseudo amino acid composition	80
13	Conjoint Triad	343
	Total	10784

172

173

174 **Machine learning and prediction**

175 Briefly, the machine learning processes of this research work followed such order and
 176 general steps. Using training dataset, 5-fold cross validation and grid searching were
 177 applied to different machine learning algorithms for obtaining identifying the best
 178 parameters of the best predictive power. Later, predictors with good performances were
 179 further applied to classify the validation dataset with known labels. Lastly, the verified
 180 best predictor was used to predict antiviral-HPV protein interaction pairs.

181

182 Diverse sorts of supervised learning algorithms with different purposes exist. Amongst,
 183 the Support Vector Machine, Random Forest [30], Logistic Regression [31], etc, are
 184 classic algorithms for tackling the binary classification questions. Initially, 6 types of
 185 machine learning classifiers friendly for binary classification were chosen for building

186 predictive models. The chosen predictors were Support Vector Machine, Random Forest,
 187 AdaBoost [32], Logistic Regression, Naïve Bayes [33] and K-Nearest Neighbor
 188 classifier. Next, with default parameters, 6 predictors went over simple checking
 189 through quick training and performance measurement. At this early stage, as expected,
 190 all predictors did not perform well. Subsequently, in order to identify better parameters
 191 for predictors, grid search 5-fold cross validations and performance benchmarking were
 192 conducted. The predictive performance of 6 different predictors with better parameters
 193 were tested using known-label validation dataset. Upon checking performance of
 194 different predictors, we selected the optimized K-Nearest Neighbors classifier, which
 195 had the highest precision score and was the most appropriate predictor to identify high
 196 confidence drug-protein interaction pairs from 864 pairs of antiviral-HPV protein
 197 associations.

198
 199 Aforementioned data processing and machine learning computations were done via
 200 in-house scripts of Python [34] and R [28]. Libraries and modules used were Sci-kit
 201 learn [35], Pandas [36], Numpy [37, 38] Scipy [39], and also Bioconductor [40, 41] and
 202 Biomart [42]. We acknowledge authors and developers of these computational tools.

203

204 **Results**

205 **Dataset overview**

The antiviral drugs and their associated targets were retrieved and analyzed as described in method section. **Table 2** provides a summary of our dataset. We had totally 61 antiviral drugs, which formed totally 284 antiviral-target interaction pairs with their targets. For the purpose of measuring machine learning predictors' performance, antiviral-target interactions were split in to two classes where 102 pairs were used as the dataset for training or fitting machine learning predictors, and the rest 182 pairs were treated as dataset for validating the predictive performances of machine learning predictors. And we also compiled 9 proteins of HPV (its complete proteome) with 96 antiviral drugs to form 864 pairs of antiviral-HPV protein association pairs (**Table 2**).

Table 2: Summary of antiviral-target and antiviral-HPV protein interaction dataset used in machine learning processing of this study

Dataset	Antiviral drug	Target protein	Antiviral-protein interaction pair
Training set	35	34	102 ^c
Validation set ^a	61	47	182 ^c
Prediction set	96	9 ^b	864

^a Validation set consisted of U.S.FDA-approved antiviral drugs and these drugs' binding target proteins

^b 9 proteins of HPV-16.

^c Ratio of positive instance to negative instance was 1:1.

222 Performances of machine learning models

223 Initially, we chose 6 type of machine learning models and applied 5-fold cross
224 validation strategy to fit the antiviral-target interaction training dataset. A primary
225 benchmarking of the predictive performance of 6 chosen predictors were as seen in
226 **Table 3**.

227

228 **Table 3:** Performance of 6 machine learning predictors with default parameters.

ID	Predictor	Precision	Recall	F1-measure	Accuracy	AUC ^a
1	SVM	0.36	0.48	0.37	0.48	0.44
2	Logistic Regression	0.52	0.57	0.54	0.59	0.56
3	KNN	0.61	0.62	0.60	0.59	0.61
4	Naïve Bayes	0.46	0.65	0.52	0.49	0.52
5	Random Forest	0.61	0.61	0.63	0.66	0.68
6	AdaBoost	0.73	0.48	0.55	0.62	0.48

229 ^a AUC indicates the metric of Area Under Curve of Receiver-Operating Characteristic Curve.

230

231 Briefly, all predictors' predictive performances were less satisfying, as expected. SVM
232 with default parameter (RBF kernel) performed the worst in all sorts of metrics among 6
233 predictors. AdaBoost classifier scored the best in terms of precision score but had the
234 lowest recall value in recall. F1-measure is the harmonic average of precision and recall.
235 The highest F1-measure was found from Random forest classifier, which was 0.63.

236 While we also found other metrics of Random forest were not high. All its metrics were
237 round 0.65 though the values were close to each other. The highest accuracy score and
238 AUC (Area Under Curve of Receiver-Operating Characteristic Curve) of 6 predictors'
239 were 0.66 and 0.68, respectively. And both of them were also found in Random forest's
240 performance. Metrics of default parameters' KNN were all around 0.6, indicating its
241 unsatisfying performances in 5-fold cross validation, too. Similar to KNN, Naïve bayes
242 classifier did not perform well, and one common point of KNN and Naïve bayes
243 classifier was that, the value of their recall score was higher than those of other metrics
244 (**Table 3**).

245
246 Next we tuned parameters of predictors through grid searching 5-fold cross validation,
247 and tested how combination set of parameters affected predictors' predictive
248 performances on known-label validation dataset. At beginning, we focused on optimize
249 predictors for obtaining better values of comprehensive metrics, such as the F1-measure,
250 accuracy or AUC value. Despite a great number of times' trying, no high scores of
251 aforementioned F1-measure, accuracy or AUC metric value was seen.

252
253 Given that high precision score indicates the low number of predictive false positive
254 instances, and high recall score indicates the low number of predictive false negative
255 instances, we changed our strategy and decided to do high precision-oriented
256 optimization. This was because the purpose of this work was to identify antivirals which

257 interact with HPV proteins. To this end, using high-precision predictor, predictive
 258 positive instances could had lower false positive instance mixed inside. Therefore, in
 259 this work, we preferred precision metric over recall metric for selecting appropriate
 260 predictors to predict antiviral-HPV protein interactions (positive instances). Through
 261 benchmarking the performances of predictors, we found optimized SVM and KNN
 262 predictors had better precision score than others. SVM's was 0.8 and the KNN
 263 classifier's was 0.85 (**Table 4**). We hence used them for prediction task and we chose
 264 the intersection of their prediction results as the final results.

265

266

267 **Table 4:** Precision scores of optimized machine learning predictors on the validation dataset of
 268 antiviral-HPV protein interaction pairs

Predictor	SVM	Logistic	KNN	Naïve	Random	AdaBoost
		Regression		Bayes	Forest	
Precision						
Score	0.80	0.50	0.85	0.65	0.68	0.75

269

270

271 **Predicted antiviral-HPV protein interaction pairs**

272 Upon selection of high-precision predictors, we applied them to predict the
 273 antiviral-HPV protein interactions. We selected two predictors' result intersection as the

274 final prediction result, i.e., we only consider an antiviral-HPV protein association pair
 275 has potential interaction if both predictors predicted this pair to be positive (interactive).
 276 As a result, within 864 antiviral-HPV association pairs, most antiviral-HPV protein
 277 pairs were predicted to be negative, i.e., the antiviral drug does not interact with the
 278 HPV protein. Only a small portion, i.e., 58, of antiviral-HPV protein pairs were
 279 predicted to have interaction. Prediction results were summarized in **Table 5** in HPV
 280 protein-oriented form. Full prediction results could be found in **Supplementary Table 3**.
 281 Here we took the Docosanol as an example for analysis. The drug Docosanol was
 282 predicted to interact with HPV-16's protein E7 using our high-precision machine
 283 learning predictors. Docosanol is a U.S. FDA-approved antiviral drug targeting
 284 Envelope glycoprotein GP350 and GP340 of Epstein-Barr Virus (EBV, also known as
 285 Human Herpesvirus or HHV-4) and it is used to treat fever blisters, etc. Interestingly,
 286 through literature survey, a recently published clinical case report was found to claim
 287 that, the mixture usage of Docosanol, curcumin, and other drugs together treated HPV
 288 infection and vaginal warts of a patient well [43]. This could be an evidence supporting
 289 our predictive result about Docosanol and HPV protein. HPV protein-oriented antiviral
 290 prediction results were summarized in **Table 5** and briefly description of the example
 291 antiviral drugs, protein targets of the antiviral drug and relevant therapeutic indications
 292 were also listed in **Table 5**.

293

294 **Table 5:** Summary of prediction result of antivirals targeting each protein of HPV-16.

HPV-16 protein	Number of antivirals ^a	Example
Protein E7	7	Docosanol targeting GP340 or GP350 protein of Epstein-Barr Virus has been approved to treat herpes labialis, fever blisters, etc.
Regular Protein E2	5	Voxilaprevir targeting NS3/4A protein of Hepatitis C Virus has been approved to treat chronic Hepatitis C caused by Hepatitis C Virus infection.
Protein E6	6	Telaprevir is an NS3/4A viral protease inhibitor. It has been approved to treat chronic Hepatitis C Virus infection in combination with other drugs.
Minor capsid protein L2	4	Grazoprevir targeting NS3/4A protein of Hepatitis C Virus has been approved to treat Hepatitis C viral infection.
Protein E4	8	Nelfinavir is a potent viral protease inhibitor for treating infections of Human Immunodeficiency Virus (HIV), and it targets the protease of HIV -1.
Probable protein E5	7	Maraviroc is a chemokine receptor antagonist drug targeting C-C chemokine receptor type 5. It has been approved to treat HIV-1 infection

Replication	7	Pirodavis (investigational drug) targets the genome
protein E1		polyprotein of Polioviruses and it seems to have
		broad-spectrum antiviral effects on multiple kinds of
		Human Rhinoviruses.
Major	5	Docosanol targeting GP340 or GP350 protein of
capsid		Epstein-Barr Virus has been approved to treat herpes
protein L1		labialis, fever blisters, etc.
Protein	7	TMC-310911 (investigational drug) is a protease
E8^E2C		inhibitor targeting HIV-1 protease and it seems to
		have effect on treating HIV-1 infection.

295 ^a Indicating the number of antivirals which was predicted to have potential interaction with specific
296 HPV-16 protein.

297

298 **Discussion**

299 While our results are to be validated by *in vitro* assays, in this work, we constructed
300 machine learning models, and predicted antiviral-HPV protein interactions so as to
301 identify potential drug candidates targeting HPV proteins. The high-risk types of HPV
302 are not limited to HPV-16. There are other types such as HPV-18. Indeed, we are not
303 only able to apply the research framework of this study to predict the potential drug
304 candidates for the proteome of other HPV subtypes, but also to other types of
305 pathogenic and infectious microbes, as well.

306 **Limitation**

307 Reviewing this current study, we found several significant points that could help us do
 308 better preparation for further works. Initially, in this work, though we tried our best to
 309 collect more antiviral drugs, due to the availability of antiviral drugs, we had limited
 310 size of dataset for machine learning. This could be one of factors why we did not obtain
 311 predictors with high score of F1-measure, accuracy, or AUC. Compared with antivirals,
 312 the number of other types of drugs, e.g., the cancer drugs or antibiotics, are higher. Thus,
 313 in future study, we would consider using other types of drugs for repositioning purpose.

314

315 Also, the final predictors selected did not have high F1-measure, accuracy, or AUC.
 316 Because current machine learning processes are black box which is difficult to interpret.
 317 Alternatively, in this study, considered the tradeoff between precision and recall, we
 318 chose to select the intersected prediction results from two high-precision predictors in
 319 order to get higher confidence antiviral-HPV protein interactions. For future study, we
 320 would learn and try to apply the state-of-art explainable machine learning methods
 321 which may be interpretable. In such case, we may be able to find out reasons causing
 322 low performances and obtaining guidance for model optimizations and obtaining more
 323 powerful machine learning predictors.

324 **Conclusions**

325 Inspired by the needs of anti-HPV drug discovery, drug repositioning and computational

326 analytics, we designed this research project and constructed machine learning models to
 327 predict possible antiviral-HPV protein interactions so as to identify potential
 328 pharmacotherapy for HPV infection. As a result, we optimized the predictors and
 329 identified 58 antivirals which have potential interaction with proteins of HPV-16's.

330

331 To the best of our knowledge, we are the first pioneer to conduct this HPV-oriented
 332 computational antiviral repositioning study. No similar study has been found so far.
 333 Therefore, our work provides good insights to virologists, medicinal chemists,
 334 gynecologist, clinical microbiologists, etc, those who are interested in the treatment and
 335 therapy of HPV infections. Also, drug candidates pre-selected via computational
 336 analytic screening could have lower probability of ineffectiveness than those that did
 337 not go through computational analyses. It thus could save resources, and antivirals
 338 identified by us could be good candidates for further *in vitro* and *in vivo* tests. In such
 339 way, this work contributes to drug development for HPV infections. What is more, our
 340 predicted antiviral-HPV protein interaction pairs also offer insights for fundamental
 341 biomedical research on drug-protein interactions or molecular interaction mechanisms.
 342 The last but not the least, the research framework of this study, i.e., the machine
 343 learning-based compound-protein interaction prediction, could also be applied to
 344 primary drug repositioning or drug discovery for those diseases or infectious microbial
 345 pathogens lacking effective pharmacotherapy. E.g., the Noroviurs.

346

347 **Supplementary Materials**

348 Supplementary Materials are available in this study.

- 349 ● Supplementary Table 1 is the antiviral-target interaction dataset for machine
- 350 learning training processing.
- 351 ● Supplementary Table 2 included drug-target interaction pair dataset used in
- 352 machine learning validation process.
- 353 ● Supplementary Table 3 included dataset of HPV-16 proteome.

354 **Data Availability Statement**

355 Data of this study were included in the supplementary materials

356 **Author Contributions**

- 357 ● Conception of this work: Hui-Heng Lin and Hongyan Xu;
- 358 ● Acquisition of datasets: Hui-Heng Lin, Xiangjun Kong and Qian-Ru Zhang;
- 359 ● Analysis and interpretation of data: Hui-Heng Lin, Liuping Zhang
- 360 ● Drafting or reviewing manuscript: Hui-Heng Lin, Yanyan Tang, and Yong Zhang.
- 361 ● Approval of submission: All authors approved submission.

362 **Funding Statement**

363 This research did not received any specific grant from funding agencies in the public,
 364 commercial, or not-for-profit sections. Funding sources have no role in conducting,
 365 writing and publishing this research work.

366 **References**

- 367 [1]. Ljubojevic S, Skerlev M. HPV-associated diseases. Clinics in dermatology. 2014
368 Mar 1;32(2):227-34.
- 369 [2]. Fakhry C, Zhang Q, Nguyen-Tan PF, Rosenthal D, El-Naggar A, Garden AS,
370 Soulieres D, Trotti A, Avizonis V, Ridge JA, Harris J. Human papillomavirus and
371 overall survival after progression of oropharyngeal squamous cell carcinoma.
372 Journal of clinical oncology. 2014 Oct 20;32(30):3365.
- 373 [3]. Muñoz N, Bosch FX, De Sanjosé S, Herrero R, Castellsagué X, Shah KV, Snijders
374 PJ, Meijer CJ. Epidemiologic classification of human papillomavirus types
375 associated with cervical cancer. New England journal of medicine. 2003 Feb
376 6;348(6):518-27. doi:10.1056/NEJMoa021641. hdl:2445/122831.
- 377 [4]. Wardak S. Human Papillomavirus (HPV) and cervical cancer. Medycyna
378 doswiadczalna i mikrobiologia. 2016 Jan 1;68(1):73-84.
- 379 [5]. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer
380 statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for
381 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2018
382 Nov;68(6):394-424.
- 383 [6]. Markowitz LE, Dunne EF, Saraiya M, Lawson HW, Chesson H, Unger ER.
384 Quadrivalent human papillomavirus vaccine: recommendations of the Advisory
385 Committee on Immunization Practices. MMWR Morb Mortal Wkly Rep.
386 2007;56(RR-02):1-24.
- 387 [7]. Ganguly N, Parihar SP. Human papillomavirus E6 and E7 oncoproteins as risk

- 388 factors for tumorigenesis. *Journal of biosciences*. 2009 Mar 1;34(1):113-23.
389 doi:10.1007/s12038-009-0013-7.
- 390 [8]. Tang S, Tao M, McCoy Jr JP, Zheng ZM. The E7 oncoprotein is translated from
391 spliced E6* I transcripts in high-risk human papillomavirus type 16-or type
392 18-positive cervical cancer cell lines via translation reinitiation. *Journal of virology*.
393 2006 May 1;80(9):4249-63. doi:10.1128/JVI.80.9.4249-4263.2006.
- 394 [9]. Ricci-López J, Vidal-Limon A, Zúñiga M, Jiménez VA, Alderete JB, Brizuela CA,
395 Aguila S. Molecular modeling simulation studies reveal new potential inhibitors
396 against HPV E6 protein. *PloS one*. 2019 Mar 15;14(3):e0213028.
- 397 [10]. Zanier K, Charbonnier S, Sidi AO, McEwen AG, Ferrario MG,
398 Poussin-Courmontagne P, Cura V, Brimer N, Babah KO, Ansari T, Muller I.
399 Structural basis for hijacking of cellular LxxLL motifs by papillomavirus E6
400 oncoproteins. *Science*. 2013 Feb 8;339(6120):694-8. doi:10.1126/science.1229934.
- 401 [11]. Bernstein WB, Dennis PA. Repositioning HIV protease inhibitors as cancer
402 therapeutics. *Current Opinion in HIV and AIDS*. 2008 Nov;3(6):666.
- 403 [12]. Hampson L, Oliver AW, Hampson IN. Using HIV drugs to target human
404 papilloma virus. *Expert review of anti-infective therapy*. 2014 Sep 1;12(9):1021-3.
- 405 [13]. Hampson L, Kitchener HC, Hampson IN. Specific HIV protease inhibitors
406 inhibit the ability of HPV16 E6 to degrade p53 and selectively kill E6-dependent
407 cervical carcinoma cells in vitro. *Antiviral therapy*. 2006;11(6):813–25.
- 408 [14]. Kim DH, Jarvis RM, Allwood JW, Batman G, Moore RE, Marsden-Edwards E,

- 409 Hampson L, Hampson IN, Goodacre R. Raman chemical mapping reveals site of
410 action of HIV protease inhibitors in HPV16 E6 expressing cervical carcinoma cells.
411 Analytical and bioanalytical chemistry. 2010 Dec;398(7):3051-61.
- 412 [15]. Kim DH, Allwood JW, Moore RE, Marsden-Edwards E, Dunn WB, Xu Y,
413 Hampson L, Hampson IN, Goodacre R. A metabolomics investigation into the
414 effects of HIV protease inhibitors on HPV16 E6 expressing cervical carcinoma
415 cells. Molecular BioSystems. 2014;10(3):398-411.
- 416 [16]. Guo G, Wang H, Bell D, Bi Y, Greer K. KNN model-based approach in
417 classification. In OTM Confederated International Conferences. On the Move to
418 Meaningful Internet Systems" 2003 Nov 3 (pp. 986-996). Springer, Berlin,
419 Heidelberg.
- 420 [17]. Noble WS. What is a support vector machine?. Nature biotechnology. 2006
421 Dec;24(12):1565-7.
- 422 [18]. Chen R, Liu X, Jin S, Lin J, Liu J. Machine learning for drug-target interaction
423 prediction. Molecules. 2018 Sep;23(9):2208.
- 424 [19]. Zhang W, Lin W, Zhang D, Wang S, Shi J, Niu Y. Recent advances in the
425 machine learning-based drug-target interaction prediction. Current drug metabolism.
426 2019 Mar 1;20(3):194-202.
- 427 [20]. Liu S, Liu C, Deng L. Machine learning approaches for protein-protein
428 interaction hot spot prediction: Progress and comparative assessment. Molecules.
429 2018 Oct;23(10):2535.

- 430 [21]. Das S, Chakrabarti S. Classification and prediction of protein–protein
431 interaction interface using machine learning algorithm. *Scientific reports*. 2021 Jan
432 19;11(1):1-2.
- 433 [22]. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T,
434 Johnson D, Li C, Sayeeda Z, Assempour N. DrugBank 5.0: a major update to the
435 DrugBank database for 2018. *Nucleic acids research*. 2018 Jan 4;46(D1):D1074-82.
- 436 [23]. Schwartz LM, Woloshin S, Zheng E, Tse T, Zarin DA. ClinicalTrials. gov and
437 Drugs@ FDA: a comparison of results reporting for new drug approval trials.
438 *Annals of internal medicine*. 2016 Sep 20;165(6):421-30.
- 439 [24]. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA,
440 Thiessen PA, Yu B, Zaslavsky L. PubChem 2019 update: improved access to
441 chemical data. *Nucleic acids research*. 2019 Jan 8;47(D1):D1102-9.
- 442 [25]. UniProt Consortium. UniProt: a hub for protein information. *Nucleic acids*
443 *research*. 2015 Jan 28;43(D1):D204-12.
- 444 [26]. Zhu F, Han B, Kumar P, Liu X, Ma X, Wei X, Huang L, Guo Y, Han L, Zheng
445 C, Chen Y. Update of TTD: therapeutic target database. *Nucleic acids research*.
446 2010 Jan 1;38(suppl_1):D787-91.
- 447 [27]. Cao Y, Charisi A, Cheng LC, Jiang T, Girke T. ChemmineR: a compound
448 mining framework for R. *Bioinformatics*. 2008 Aug 1;24(15):1733-4.
- 449 [28]. Ihaka R, Gentleman R. R: a language for data analysis and graphics. *Journal of*
450 *computational and graphical statistics*. 1996 Sep 1;5(3):299-314.

- 451 [29]. Xiao N, Cao DS, Zhu MF, Xu QS. protr/ProtrWeb: R package and web server
452 for generating various numerical representation schemes of protein sequences.
453 Bioinformatics. 2015 Jun 1;31(11):1857-9.
- 454 [30]. Pal M. Random forest classifier for remote sensing classification. International
455 journal of remote sensing. 2005 Jan 1;26(1):217-22.
- 456 [31]. Pregibon D. Logistic regression diagnostics. The annals of statistics. 1981
457 Jul;9(4):705-24.
- 458 [32]. Rätsch G, Onoda T, Müller KR. Soft margins for AdaBoost. Machine learning.
459 2001 Mar;42(3):287-320.
- 460 [33]. Soria D, Garibaldi JM, Ambrogi F, Biganzoli EM, Ellis IO. A
461 ‘non-parametric’ version of the naive Bayes classifier. Knowledge-Based Systems.
462 2011 Aug 1;24(6):775-84.
- 463 [34]. Oliphant TE. Python for scientific computing. Computing in science &
464 engineering. 2007 Jun 18;9(3):10-20.
- 465 [35]. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O,
466 Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J. Scikit-learn:
467 Machine learning in Python. the Journal of machine Learning research. 2011 Nov
468 1;12:2825-30.
- 469 [36]. McKinney W. pandas: a foundational Python library for data analysis and
470 statistics. Python for high performance and scientific computing. 2011 Nov
471 18;14(9):1-9.

- 472 [37]. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau
473 D, Wieser E, Taylor J, Berg S, Smith NJ, Kern R. Array programming with NumPy.
474 Nature. 2020 Sep;585(7825):357-62.
- 475 [38]. Van Der Walt S, Colbert SC, Varoquaux G. The NumPy array: a structure for
476 efficient numerical computation. Computing in science & engineering. 2011 Mar
477 7;13(2):22-30.
- 478 [39]. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D,
479 Burovski E, Peterson P, Weckesser W, Bright J, Van Der Walt SJ. SciPy 1.0:
480 fundamental algorithms for scientific computing in Python. Nature methods. 2020
481 Mar;17(3):261-72.
- 482 [40]. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B,
483 Gautier L, Ge Y, Gentry J, Hornik K. Bioconductor: open software development for
484 computational biology and bioinformatics. Genome biology. 2004 Sep;5(10):1-6.
- 485 [41]. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W.
486 BioMart and Bioconductor: a powerful link between biological databases and
487 microarray data analysis. Bioinformatics. 2005 Aug 15;21(16):3439-40.
- 488 [42]. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G,
489 Kasprzyk A. BioMart—biological queries made easy. BMC genomics. 2009
490 Dec;10(1):1-2.
- 491 [43]. Psomiadou V, Iavazzo C, Douligieris A, Fotiou A, Prodromidou A, Blontzos N,
492 Karavioti E, Vorgias G. An Alternative Treatment for Vaginal Cuff Wart: a Case

493 Report. Acta Medica (Hradec Kralove). 2020;63(1):49-51. doi:

494 10.14712/18059694.2020.15.

495

496