Targeting against HIV/HCV Co-infection using Machine Learning-based multitarget-quantitative structure-activity relationships (mt-QSAR) Methods

Yu Wei^{1,3,¶}, Wei Li^{1,3,¶}, Tengfei Du¹, Zhangyong Hong^{2*}, Jianping Lin^{1,3,4*}

¹State Key Laboratory of Medicinal Chemical Biology, College of Pharmacy and Tianjin Key Laboratory of Molecular Drug Research, Nankai University, Haihe Education Park, 38 Tongyan Road, Tianjin 300353, China

²State Key Laboratory of Medicinal Chemical Biology, College of Life Sciences, Nankai University, 94 Weijin Road, Tianjin 300071, China

³Platform of Pharmaceutical Intelligence, Tianjin International Joint Academy of Biomedicine, Tianjin, 300000, China

⁴Biodesign Center, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin 300308, China

[¶]These authors contributed equally to this work.

*Corresponding authors:

Jianping Lin, State Key Laboratory of Medicinal Chemical Biology, College of Pharmacy and Tianjin Key Laboratory of Molecular Drug Research, Nankai University, Haihe Education Park, 38 Tongyan Road, Tianjin 300353, China, email: jianpinglin@nankai.edu.cn

Zhangyong Hong, College of Life Sciences, Nankai University, 94 Weijin Road, Tianjin 300071, China, email: hongzy@nankai.edu.cn

ABSTRACT

Co-infection between HIV-1 and HCV is common today in certain populations. However, treatment of co-infection is full of challenges with special consideration for potential hepatic safety and drug-drug interactions. Multitarget inhibitors with less toxicity may provide a promising therapeutic strategy for HIV/HCV co-infection. However, identification of one molecule acting on multiple targets simultaneously by experimental evaluation is costly and time-consuming. *In silico* target prediction tools provide more opportunities for the development of multitarget inhibitors. In this study, by combining naive Bayesian (NB) and support vector machine (SVM) algorithms with two types of molecular fingerprints (MACCS and ECFP6), 60 classification models were constructed to predict the active compounds toward 11 HIV-1 targets and 4 HCV targets based on the multitarget-quantitative structure-activity relationships (mt-QSAR). 5-fold cross-validation and test set validation was performed to confirm the performance of 60 classification models. Our results show that 60 mt-QSAR models appeared to have high classification accuracy in terms of ROC-AUC values ranging from 0.83 to 1 with a mean value of 0.97 for HIV-1 models, and ROC-AUC values ranging from 0.84 to 1 with a mean value of 0.96 for HCV. Furthermore, the 60 models were applied to comprehensively predict the potential targets for additional 46 compounds including 27 approved HIV-1 drugs, 10 approved HCV drugs and 9 selected compounds known to be active on one or more targets of HIV-1 or those of HCV. Finally, 18 hits including 7 HIV-1 approved drugs, 4 HCV approved drugs and 7 compounds were predicted to be HIV/HCV co-infection multitarget inhibitors. The reported bioactivity data confirmed that 7 compounds actually interacted with HIV-1 and HCV targets simultaneously with diverse binding affinities. Of those remaining predicted hits and chemical-protein interaction pairs involving the potential ability to suppress HIV/HCV co-infection deserve further investigation by experiments. This investigation shows that the mt-QSAR method is available to predict chemical-protein interaction for discovering multitarget inhibitors and provide a unique perspective on HIV/HCV co-infection treatment.

1. Introduction

Human immunodeficiency virus type-1 (HIV-1) is the causative factor of acquired immunodeficiency syndrome (AIDS), a pandemic disease[1]. In addition, hepatitis C virus (HCV) infection causes acute and chronic liver disease, including cirrhosis and hepatocellular carcinoma[2]. Unfortunately, since HIV-1 and HCV have similar routes of transmission, the risk of HIV/HCV co-infection is very high[3]. According to the World Health organization (WHO), about 2.3 million persons of the estimated 36.7 million living with HIV had been infected with HCV globally in 2015. It is necessary for those people to be diagnosed and provided with effective and reasonable treatment for both HIV and hepatitis as a priority[4]. However, in the treatment of HIV/HCV co-infection, special considerations must be given to potential drug-drug interactions and toxicities. The standard treatment for HIV-1 is highly active antiretroviral therapy (HAART), which improves the life quality of AIDS patients by reducing viral load[5]. HAART utilizes a cocktail of drugs to obstruct multiple viral proteins in the viral life cycle, including reverse transcriptase, integrase, and protease[6]. Although it is effective, HAART needs to sequentially administer single-target drugs that may result in drug-drug interactions, poor treatment adherence and the emergence of drug resistance. The current therapy for HIV/HCV co-infection includes injecting pegylated α -interferon (PEG-IFN) regularly plus taking Ribavirin (RBV) daily in combination with HAART[7]. Similar to the HAART treatment regimen, additive drug toxicity, high cost and long-term therapy cycle should be taken into account during the process of HIV/HCV co-infection treatment. In order to avoid these undesirable factors, the conception of multitarget therapy using a single drug capable of inhibiting simultaneously two or more viral targets was proposed to reduce the complexity of the treatment[5].

With the development of computer-aided drug design (CADD) technology, several *in silico* methods have been exploited to predict the chemical-protein interaction (CPI) pairs between drugs and target proteins[8,9]. These methods can be approximately classified into two categories: structure-based and ligand-

based, such as pharmacophore modeling[10], similarity searching[11,12] and molecular docking[13,14]. Conventional quantitative structure-activity relationship (QSAR) generally takes a group of analogs against one target into account to elucidate the relationship between chemical structure and biological activity. In 2009, Vina et al. developed a multitarget-QSAR (mt-QSAR) classification model to predict the probability of drugs binding more than 60 different receptors based on molecular connectivity and receptor sequences[15]. Furthermore, Jacob et al. proposed chemogenomic method to predict CPI[16], and Wang et al. used chemogenomic method to identify new ligands for four targets and the results was validated by experiments[17]. As a new interdisciplinary field, chemogenomics tries to completely match the target and ligand space, and determine available ligands for all targets[9]. Cheng et al. compared the CPI prediction performance between mt-QSAR and computational chemogenomic methods, and the result showed that mt-QSAR had a better performance than chemogenomic, which had a high risk of overfitting and a high false positive rate in the external validation set[18].

In this study, mt-QSAR method was applied to predict the chemical-protein interactions (CPIs) for 11 key targets related to HIV-1 and 4 targets related to HCV. The workflow is depicted in Fig 1. Firstly, combining two machine learning algorithms (naive Bayesian[19,20] and support vector machine[21]) and two molecular fingerprints (MACCS[22] and ECFP6[23]), we developed multiple mt-QSAR models for identifying inhibitors against 11 HIV-1 related and 4 HCV related targets. Secondly, 5-fold cross-validation and test set validation was used to evaluate the performance of all models. Thirdly, to verify the application of this strategy, the developed mt-QSAR models were employed to predict CPI for 27 approved HIV-1 drugs and 10 approved HCV drugs well as 9 compounds known to be active on one or more targets of HIV-1 or HCV. The predicted results shown 18 hits to be potential HIV/HCV co-infection multitarget inhibitors, which were further confirmed by reported bioactivity resulting in that 7 compounds actually interacted with HIV-1 and HCV targets simultaneously with diverse binding affinities. Our study indicates that machine learning-

based mt-QSAR approaches could be potentially applied in discovery of HIV/HCV co-inhibitors and multitarget drug discovery.

2. Results

2.1 Experimental datasets analysis

In this study, a total of 11 targets related to HIV-1 and 4 targets related to HCV were obtained. The number of known small molecule inhibitors that act on HIV-1 and HCV targets were 11,006 and 1,431, respectively. The number of active compounds for each target was shown in Fig 2. A total of 11 datasets were generated for HIV-1 system, including 157 inhibitors and 471 decoys for CXCR4, 370 inhibitors and 1,110 decoys for CCR5, 293 inhibitors and 879 decoys for GP120, 21 inhibitors and 63 decoys for GP41, 3,256 inhibitors and 9,768 decoys for RT, 2,215 inhibitors and 6,645 decoys for IN, 4,111 inhibitors and 12,333 decoys for PR, 101 inhibitors and 303 decoys for Gag-pol, 438 inhibitors and 1,314 decoys for protein Tat, 29 inhibitors and 87 decoys for PKC, 15 inhibitors and 45 decoys for CYP3A. There were 4 datasets for HCV system, including 470 inhibitors and 1,410 decoys for NS3/4A, 19 inhibitors and 57 decoys for NS4B, 40 inhibitors and 120 decoys for NS5A, 902 inhibitors and 2,706 decoys for NS5B. Each of 15 datasets above was divided into training set and test set.

Before modeling implementation, Tanimoto similarity index, which is a generally-used measure for evaluating molecular similarity among numerous compounds, was used to evaluate fingerprint-based similarity using ECFP4 fingerprint[24,25]. The Tanimoto similarity index of a set of compounds in training set and test set for each of 15 datasets were computed, and the results were shown in Table 1. For HIV-1, the Tc value of training sets and test sets in 11 datasets range from 0.124 to 0.202 and 0.117 to 0.215, respectively. For HCV, the training sets and test sets of 4 datasets have the similar Tc values ranging from 0.141 to 0.182. Table 1 shows that compounds in the training sets and test sets of 15 datasets were sufficiently diverse.

Table 1. Detailed statistical analysis of the 15 datasets

Object	Target	Training Set				Test Set			
		Inhibitors	Decoys	Total	Tanimoto	Inhibitors	Decoys	Total	Tanimoto
					index				index
HIV-1	CXCR4	113	339	452	0.152	44	132	176	0.158
	CCR5	255	766	1021	0.161	115	344	459	0.160
	GP120	208	624	832	0.142	85	255	340	0.147
	GP41	14	42	56	0.172	7	21	28	0.189
	RT	2279	6837	9116	0.124	977	2931	3908	0.124
	IN	1617	4850	6467	0.134	598	1795	2393	0.135
	PR	2837	8509	11346	0.141	1274	3824	5098	0.141
	Gag-pol	69	205	274	0.147	32	98	130	0.146
	Protein tat	302	906	1208	0.118	136	408	544	0.117
	PKC	20	58	78	0.162	9	29	38	0.169
HCV	CYP3A	10	30	40	0.202	5	15	20	0.215
	NS5B	649	1948	2597	0.141	253	758	1011	0.141
	NS4B	13	37	50	0.178	6	20	26	0.192
	NS3/4A	334	1000	1334	0.145	136	410	546	0.145
	NS5A	27	80	107	0.182	13	40	53	0.182

In addition, according to the protein target classification rules in ChEMBL database, 11 targets related to HIV-1 and 4 targets related to HCV are spatial distributed into five and two categories, respectively. As showed in Fig 3A, the HIV-1 target space (n = 11) consists of five subfamilies, including membrane receptor (n = 2), enzyme (n = 5), transcription factor (n = 1), kinase (n = 1) and unclassified protein (n = 2). Similarly, the HCV target space (n = 4) is divided into two subfamilies, including enzyme (n = 2) and unclassified protein (n = 2) in Fig 3B. Thus, the two data sets for HIV-1 and HCV have diverse target coverage.

2.2 Performance evaluation and comparison of models

The construction of all classification models, in this study, was primarily based on two classifiers (NB

and SVM) and two types of fingerprints (MACCS and ECFP6). Afterwards, internal 5-fold cross-validation and external test set validation was executed in turn.

The 5-fold cross-validation test results of mt-QSAR models identifying inhibitors and decoys are displayed in Table 2. In the 44 models of HIV-1, 86.36% models (38 out of 44) have AUC value exceeding 0.9, and 61.18% models (30 out of 44) have MCC value exceeding 0.8. In addition, AUC value ranges from 0.83 to 1, with an average of 0.96, and MCC value ranges from 0.52 to 1, with an average of 0.84. Among the 16 models of HCV, all models have AUC value exceeding 0.9, and 81.25% models (13 out of 16) have MCC value exceeding 0.8. Besides, AUC value ranges from 0.92 to 1, with an average of 0.99, and MCC value ranges from 0.52 to 1, with an average of 0.91. Thus, these results suggest that the mt-QSAR models, both for HIV-1 and HCV, have remarkable capability in identifying inhibitors at low false positive rates. Meanwhile, the detailed performance for the training set is listed in S1 Table. Similarly, in the models of HIV-1, 75% models (33 out of 44) have SE value exceeding 0.8, with an average of 0.86, and 95.35% models (41 out of 44) have SP value exceeding 0.8, with an average of 0.96. In the models of HCV, 87.5% models (14 out of 16) have SE value exceeding 0.8, with an average of 0.95, and 93.8% models (15 out of 16) have SP value exceeding 0.8, with an average of 0.96.

To assess the generality of the mt-QSAR models on distinguishing inhibitors from decoys, an external test sets were used to perform a more rigorous validation of the training models. The assessment results of test sets are listed in Table 3. In the 44 models of HIV-1, AUC value ranges from 0.83 to 1, with the mean value of 0.97, and MCC value ranges from 0.48 to 1, with the mean value of 0.85. In the 16 models of HCV, AUC value ranges from 0.84 to 1, with the mean value of 0.96, and MCC value ranges from 0.5 to 1, with the mean value of 0.89. The detailed performance of test sets validation is presented in S2 Table. The average value of SE and SP for 11 datasets of HIV-1 is 0.88 and 0.96, respectively. The average value of SE and SP for 4 datasets of HCV is 0.92 and 0.96, respectively. What's more, prediction accuracy (Q) is an important index

for evaluating the predictive ability of the models. Among the models of HIV-1 and HCV, there are 93.18% and 93.75% models with an accuracy higher than 0.8, respectively. According to the results of test sets, the comparable performance of the mt-QSAR models in the training and test sets indicated that the models showed no obviously overfitting. Thus, 60 mt-QSAR models have substantial capability in distinguishing inhibitors from decoys at comparable yield and with low false positive rate.

Table 2. Results of 60 mt-QSAR models' performance by 5-fold cross-validation

Object	Target	MACCS				ECFP6			
		NB		SVM		NB		SVM	
		AUC	MCC	AUC	MCC	AUC	MCC	AUC	MCC
HIV-1	CXCR4	0.9916	0.9406	0.9977	0.9705	0.9943	0.9248	0.9971	0.9646
	CCR5	0.9769	0.815	0.9967	0.9791	0.986	0.8103	0.9956	0.9686
	GP120	0.8445	0.5428	0.9262	0.7749	0.8797	0.477	0.9232	0.6316
	GP41	0.9405	0.9524	0.9575	0.9524	0.9626	0.9524	0.9337	0.7071
	RT	0.8382	0.5203	0.9944	0.9445	0.8984	0.5226	0.9936	0.9393
	IN	0.9301	0.7487	0.9959	0.9606	0.9491	0.7119	0.9939	0.9441
	PR	0.9796	0.8488	0.9989	0.9797	0.9819	0.7824	0.9971	0.9826
	Gag-pol	0.9664	0.8838	0.984	0.9321	0.9665	0.811	0.9799	0.9127
	Protein tat	0.8342	0.4532	0.9088	0.6816	0.8702	0.562	0.9009	0.6258
	PKC	0.9914	0.9675	1	0.9664	0.994	0.899	0.9879	0.899
HCV	CYP3A	1	1	1	1	1	1	1	0.9333
	NS3/4A	0.9903	0.7997	1.000	0.9900	0.9943	0.8473	1.000	0.9820
	NS4B	1.000	1.0000	1.000	1.0000	1.000	1.0000	1.000	0.8435
	NS5A	1.000	1.0000	1.000	1.0000	1.000	1.0000	1.000	1.0000
	NS5B	0.9235	0.6612	0.9997	0.9815	0.9461	0.5215	0.9998	0.9784

Table 3. The test results of 60 mt-QSAR models' performance by test set validation

Object	Target	MACCS				ECFP6			
		NB		SVM		NB		SVM	
		AUC	MCC	AUC	MCC	AUC	MCC	AUC	MCC
	CXCR4	1	0.9848	0.9998	0.9848	0.9997	0.9848	0.9997	0.9848

	CCR5	0.9811	0.8156	0.9992	0.9826	0.9954	0.7817	0.9999	0.971
	GP120	0.8813	0.5748	0.9402	0.7993	0.883	0.4782	0.9103	0.6678
	GP41	1	1	1	1	1	1	1	1
	RT	0.8369	0.4891	0.9947	0.9471	0.8986	0.5276	0.994	0.9486
HIV-1	IN	0.9343	0.7348	0.9967	0.9653	0.9473	0.7061	0.9956	0.9597
	PR	0.975	0.8296	0.9997	0.9874	0.9819	0.8034	0.9985	0.9827
	Gag-pol	0.9774	0.8756	0.9882	0.9585	0.9746	0.7949	0.9997	0.9585
	Protein tat	0.8326	0.4914	0.9204	0.6959	0.8642	0.5595	0.9256	0.6474
	PKC	1	1	1	0.853	1	0.927	1	0.853
	CYP3A	1	1	1	1	1	1	1	0.866
	NS3/4A	0.9862	0.7694	1.000	0.9951	0.9867	0.8420	1.000	0.9608
HCV	NS4B	0.8417	0.8909	0.9250	0.8909	0.8667	0.8909	0.9083	0.8909
	NS5A	1.000	1.0000	1.000	1.0000	1.000	1.0000	1.000	1.0000
	NS5B	0.9074	0.6084	0.9999	0.9789	0.9437	0.5001	1.000	0.9842

To investigate the influence of different types of molecular fingerprints (MACCS and ECFP6) on the performance of naive Bayesian and support vector machine models, MCC values for the test set were compared. As shown in Fig 4, whether for HIV-1 or HCV, MCC values in various mt-QSAR models didn't differ significantly between MACCS and ECFP6. It should be noted that different fingerprints present different structural representation of a chemical structure, mt-QSAR models based on different fingerprints could improve performance on yield and reducing false positive rate.

Moreover, the performance of different algorithms (NB and SVM) was also evaluated. With the fingerprint MACCS (Fig 5A), performance of the SVM models are generally superior to that of NB models. In particular, for the models of HIV-1, MCC value from SVM models ranges from 0.70 to 1 with an average of 0.92, better than that from NB models which ranges from 0.49 to 1 with an average of 0.8. For the models of HCV, MCC value from the SVM models ranges from 0.89 to 1 with an average of 0.97, better than that of NB models which ranges from 0.61 to 1 with an average value of 0.82. With the fingerprint ECFP6 (Fig 5B),

the models from SVM broadly outperform those from NB. For the models of HIV-1, MCC value from SVM models ranges from 0.65 to 1 with an average of 0.89, better than that from NB which ranges from 0.48 to 1 with an average of 0.78. For the HCV models, MCC value from the SVM ranges from 0.89 to 1 with an average of 0.96, which is dissimilar to NB that ranges from 0.50 to 1 and an average value of 0.81.

As described above, two fingerprints have no significant difference and the SVM algorithm is slightly better than the NB. However, it is theoretically difficult to determine which fingerprint or algorithm is better because each algorithm or fingerprint has its own advantages and disadvantages. For example, the models built by different fingerprints or algorithms may have opposite predictions for the same molecule. Therefore, combining the results of a single classifier to predict CPIs is indispensable. In this study, four classifiers (NB_MACCS, NB_ECFP6, SVM_MACCS, SVM_ECFP6) were built for each target. One compound was predicted as positive when at least two of four single classifiers predict it to be active, and the corresponding CPIs were defined as potential interactions.

2.3 Case 1: Prediction and analysis of polypharmacology for known HIV-1 and HCV drugs

To explore the multiple bioactivities of known drugs, 60 mt-QSAR models based on 15 key targets were used to predict potential targets for 27 approved HIV-1 drugs and 10 approved HCV drugs. If the drug was predicted as active by at least two of the above four classifiers for each target, then the relevant CPI was defined as potential interactions. The detailed prediction results of polypharmacology are presented in S3 Table. The results show that 17 approved HIV-1 drugs and 4 approved HCV drugs were predicted to interact with more than one target, of which 7 approved HIV-1 drugs and 4 approved HCV drugs were predicted to interact with HIV-1 and HCV targets simultaneously. To verify the results of prediction, the predicted targets for each drug were validated using the PubChem BioAssay database[26] and ChEMBL database, and the

corresponding binding values are presented in Table 4. Fig 6 shows the histogram of the number of the predicted targets and established targets for 17 HIV-1 approved drugs and 4 HCV approved drugs. For example, Elvitegravir, which was predicted to be active against HIV-1 RT, HIV-1 IN and HCV NS5B, could inhibit HIV-1 RT and IN with IC_{50} value of 91 μ M and 28 nM, respectively. Tipranavir, which was predicted to be active against HIV-1 GP120, HIV-1 RT, HIV-1 PR and HCV NS5B, could inhibit HIV-1 PR with an IC_{50} value of 0.03 μ M. Telaprevir, which was predicted to be active against HIV-1 PR and HCV NS3/4A, could inhibit HCV NS3/4A with an IC_{50} value of 0.2 μ M.

In addition, 56 CPIs were obtained as shown in S4 Table. Among them, 25 out of 56 predicted CPIs are consistent with experimental results; the remaining 2 CPIs were predicted to be active but were experimentally validated as “Inconclusive” and “Inactive”, respectively[27,28,37–43,29–36]. The success rate of 44.6% (25/56) and the failure rate of 1.8% (1/56) illustrate the reliability of mt-QSAR method. A polypharmacological interaction network between 21 drugs and predicted targets (56 CPIs) were constructed with Cytoscape 3.6[44] (Fig 7). The Fig 7 shows that each drug is predicted to have potential activity on two or three targets, which is in line with concept “one target can hit multiple targets”. Then, predicted CPIs network identifies that many potential drug-target interactions have yet to be explored and unverified CPIs are worthy of further experimental validation. The predicted targets might be new targets for approved drugs and provide a promising therapy for HIV/HCV co-infection.

Table 4. Known interactions among CPIs predicted by mt-QSAR models for approved drugs

Drug name	Predicted targets	Known activity	References
Enfuvirtide	GP41	IC_{50} = 90.2 nM	[27]
Delavirdine	RT	IC_{50} = 0.17 μ M	[28]
Didanosine	RT	IC_{50} = 4.6 μ M	[29]
Lamivudine	RT	IC_{50} = 0.60 μ M	[29]
Stavudine	RT	IC_{50} = 1.9 μ M	[29]
Zalcitabine	RT	IC_{50} = 0.13 μ M	[29]

Zidovudine	RT	IC ₅₀ = 0.13 μM	[29]
Dolutegravir	RT	Inactive	[30]
Dolutegravir	IN	IC ₅₀ = 2.7 nM	[31]
Elvitegravir	RT	IC ₅₀ = 91 μM	[32]
Elvitegravir	IN	IC ₅₀ = 28 nM	[32]
Amprenavir	PR	IC ₅₀ = 0.15 μM	[33]
Atazanavir	RT	IC ₅₀ = 26 nM	[34]
Atazanavir	PR	IC ₅₀ = 3.5 nM	[35]
Darunavir	PR	IC ₅₀ = 3.0 nM	[35]
Indinavir	RT	Inconclusive	[36]
Indinavir	PR	IC ₅₀ < 0.01 μM	[36]
Nelfinavir	PR	IC ₅₀ = 0.53 nM	[37]
Ritonavir	PR	IC ₅₀ = 0.6 nM	[38]
Ritonavir	CYP3A	IC ₅₀ = 0.11 μM	[38]
Tipranavir	PR	IC ₅₀ = 0.03 μM	[39]
Cobicistat	PR	IC ₅₀ = 0.15 μM	[38]
Cobicistat	CYP3A	IC ₅₀ > 30 μM	[38]
Telaprevir	NS3/4A	IC ₅₀ = 0.2 μM	[40]
Simeprevir	NS3/4A	IC ₅₀ = 10 nM	[41]
Grazoprevir	NS3/4A	IC ₅₀ = 0.07 nM	[42]
Dasabuvir	NS5B	IC ₅₀ = 0.6 μM	[43]

2.4 Case 2: Target prediction and analysis for known inhibitors

HIV-1 protease (HIV-1 PR), reverse transcriptase (RT) and integrase (IN) play a pivotal role in the viral replication so that they have become attractive targets to design inhibitors[45]. HIV-1 protease plays a key role in the normal assembly and maturation of virus[46]. HIV-1 reverse transcriptase and integrase catalyze sequentially the synthesis and integration of proviral DNA into the host genome[5]. At the same time, HCV NS5B polymerase is an essential protein involved in the replication of viral positive strand genomic RNA and

used as a key target for therapeutic intervention[47,48].

Above all, a set of 9 compounds (Fig 8) with known activities targeting at least one out of four targets (HIV-1 PR, RT, IN and HCV NS5B) and not involved in generating mt-QSAR models were retrieved from ChEMBL database. In order to evaluate whether the new models could correctly differentiate these active compounds, the classifiers built in this study were used to predict chemicals toward HIV-1 PR, RT and IN as well as HCV NS5B. Following the above rules, CPI is referred to a potential interaction when at least two out of four single classifiers predict a compound as positive. The specific predicted results are given in S5 Table, and the comparison of predicted and experimental results is shown in Table 5.

Table 5. Predicted and experimental activity of 9 selected compounds towards HIV-1 PR, RT, IN and HCV NS5B

CHEMBL ID	PR/Pred ^a	PR/Exp ^b	IN/Pred ^c	IN/Exp ^d	RT/Pred ^e	RT/Exp ^f	NS5B/Pred ^g	NS5B/Exp ^h
CHEMBL16326	false	— ⁱ	true	active	false	—	true	active
CHEMBL18927	false	—	true	active	false	—	true	active
CHEMBL449221	false	—	true	active	false	—	true	active
CHEMBL502238	false	—	true	active	false	—	true	active
CHEMBL19332	false	—	true	active	false	active	true	active
CHEMBL210593	false	—	false	—	true	active	true	active
CHEMBL37541	true	active	false	—	false	—	true	active
CHEMBL3612421	true	active	false	—	false	—	true	—
CHEMBL1668670	false	—	true	active	true	active	true	—

^a“PR/Pred” shows the predicted result of a compound against PR. ^b“PR/Exp” shows the experimental result of a compound against PR. ^c“IN/Pred” shows the predicted result of a compound against IN. ^d“IN/Exp” shows the experimental result of a compound against IN. ^e“RT/Pred” shows the predicted result of a compound against RT. ^f“RT/Exp” shows the experimental result of a compound against RT. ^g“NS5B/Pred” shows the predicted result of a compound against NS5B. ^h“NS5B/Exp” shows the experimental result of a compound against NS5B. ⁱ“—” represents that the predicted result has not been verified by experiments.

The results in Table 5 show that 5 compounds (CHEMBL16326, CHEMBL18927, CHEMBL449221, CHEMBL502238 and CHEMBL19332) were predicted to be potentially active toward HIV-1 IN and HCV NS5B, and they were also experimentally confirmed to be HIV-1 IN and HCV NS5B inhibitors in the literatures[49–56]. Among them, as a known active HIV-1 RT inhibitor, CHEMBL19332 was shown to be inactive toward HIV-1 RT in the prediction result. In particular, 5 compounds with analogous structures have

similar protein-ligand interactions in the active region of HIV-1 IN and HCV NS5B. Moreover, molecular docking provides a more detailed inspection of binding mode for each compound. And the scoring values of the top scoring poses are available in S6 Table. As shown in Fig 9, the hydrophobic groups of those five compounds, such as benzene ring of CHEMBL16326 and monofluorobenzene ring of CHEMBL449221, are placed in the hydrophobic pockets formed by Leu102, Ala128, Trp132, Thr174 and Met178 of HIV-1 IN (see Fig 9A), and formed by Val52, Leu159 and Ile160 of HCV NS5B (see Fig 9B). The diketo acids moiety of those five compounds form hydrogen bonds with Glu170 and His171 in the active site of HIV-1 IN (see Fig 9A), and form hydrogen bonds with Phe224 and Asp225 and metal chelation with Mn^{2+} ions in the catalytic activity site of HCV NS5B (see Fig 9B). Thus, HIV-1 IN and HCV NS5B can be affected by diketo acids inhibitors, which disturbed enzymatic function during the process of virus replication.

CHEMBL210593 was predicted as HIV-1 RT and HCV NS5B inhibitors, and it was precisely HIV-1 RT and HCV NS5B inhibitors according to the previous report[57]. Fig 9 shows that 3-chloro-p-methylaniline of CHEMBL210593 occupies the hydrophobic pockets formed by Tyr181, Tyr188 and Trp229 of HIV-1 RT (see Fig 9C), and formed by Leu26, Ala395, Ala396 and Ile432 of HCV NS5B (see Fig 9D). The carboxyl group of CHEMBL210593 forms hydrogen bonds with the main chain carbonyl oxygen of Lys101 of HIV-1 RT (see Fig 9C), and the side chain of Arg503 and His502 of HCV NS5B (see Fig 9D). Therefore, proline sulfonamides inhibitors can affect HIV-1 RT and HCV NS5B and influence enzymatic function.

CHEMBL37541 and CHEMBL3612421 were also predicted to be active against HIV-1 PR and HCV NS5B. CHEMBL3612421 was known HIV-1 PR inhibitor and CHEMBL37541 was known HIV-1 PR and HCV NS5B inhibitor by experimentally validated[46,58,59]. As for CHEMBL37541, the dihydropyrone moiety makes direct hydrogen bonds interaction with Ile50 of HIV-1 PR (see Fig 9E) and Ser476 of HCV NS5B (see Fig 9F). The 4-ethylphenol and cyclopentane groups fit into hydrophobic pockets consisting of the residues Leu23, Leu23', Val82, Val82', Ile84 and Ile84' of HIV-1 PR (see Fig 9E), and the residues Met423,

256 Ile482, Val485, Ala486, Leu489, Leu497 and Trp528 of HCV NS5B (see Fig 9F). New predicted CPI pair
 257 involving CHEMBL3612421 and HCV NS5B is presented in Fig 9G. CHEMBL3612421 forms a hydrogen
 258 bond with Tyr477 in thumb II region of HCV NS5B and its benzene ring and dimethyl group are located in
 259 hydrophobic region consisting of the residues Met423, Ile482, Val485, Ala486, Leu489, Leu497 and Trp528.
 260 Similarly, dihydropyrones inhibitors can affect HIV-1 PR and HCV NS5B and influence enzymatic function.

261 CHEMBL1668670 was also predicted as HIV-1 RT, IN and HCV NS5B inhibitor, and it only have
 262 bioactivity data against HIV-1 RT and IN[60]. New predicted CPI pair involving CHEMBL1668670 and
 263 NS5B was presented in Fig 9H. CHEMBL1668670 forms a hydrogen bonding with Ser476 in thumb II region
 264 of HCV NS5B and its trifluorobenzene ring is directed to the hydrophobic region consisting of the residues
 265 Met423, Ile482, Val485, Ala486, Leu489 and Leu497.

266 Thus, 7 out of 9 compounds, which are known to act on HIV/HCV co-infection targets, were successfully
 267 predicted to act on two or three targets. New predictions involving CHEMBL3612421 and CHEMBL1668670
 268 targeting NS5B have yet to be further validated by experimentally. The result is given that design of specific
 269 inhibitors targeting HIV-1 targets and HCV targets simultaneously is now possible. In addition, one individual
 270 compound has co-inhibitor ability for both virus targets if it is able to satisfy the requirements of binding the
 271 HIV/HCV target active site in the molecular shape, size and physical chemical properties. In short, mt-QSAR
 272 models for HIV/HCV co-infection are expected to contribute to design and identify different types of co-
 273 inhibitors with excellent affinities that selectively and simultaneously target multiple HIV-1 and HCV targets.

274 In summary, 19 CPI pairs were predicted to be positive, and of which 17 CPI pairs were experimentally
 275 confirmed to be active. Therefore, prediction success rate of 89.5% (17/19) indicates the reliability of the mt-
 276 QSAR models. It is worth mentioning that remaining 2 CPI pairs related to 2 compounds deserve further
 277 investigation by experiments, and the predicted targets might be the new targets for the known inhibitors.
 278 Moreover, one CPI pair shows false negative in the prediction result.

The prediction results demonstrate that machine learning models have high specific and low false negative rate in this study. Here, two possible reasons for causing false negative are as follows. Firstly, molecular fingerprint is unable to sufficiently represent molecular structural information. Secondly, the inherent limitations of machine learning and the composition of training set have a significant impact on the accuracy of the models.

3. Discussion

In the past, the “one gene, one drug, one disease” was the main paradigm of drug discovery. However, with the introduction of the concept of polypharmacology, multitarget therapy using a single drug that simultaneously inhibits two or more targets may provide a new perspective. Here, the mt-QSAR method was applied to identify the compounds acting on multiple targets for HIV/HCV co-infection.

In this study, based on the mt-QSAR method, 44 binary classifiers of 11 targets related to HIV-1 and 16 binary classifiers of 4 targets related to HCV were established to predict the CPIs. Besides, the prediction reliability of models was confirmed by 5-fold cross-validation and test set validation.

To illustrate the application of the models, two different cases were performed to systematically predict the multiple bioactivities for 27 approved HIV-1 drugs and 10 approved HCV drugs (case 1) and 9 compounds that were known to be active toward at least one target of HIV-1 and HCV (case 2). For case 1, 21 approved drugs were predicted to be active against more than one biological target. And the predictions were confirmed by reported bioactivity data, the success rate was 44.6% and the failure rate was 1.8%. For case 2, 9 active compounds toward at least one target (HIV-1 PR, RT, IN and HCV NS5B) were predicted to be multitarget inhibitors by mt-QSAR models with a success rate of 89.5%. In short, the prediction results of above two cases indicate that mt-QSAR models have a significant performance on prediction of multitarget inhibitors against HIV/HCV co-infection.

As mentioned above, mt-QSAR method achieved satisfactory prediction accuracy, and machine learning models were applicable for potential targets identification in this study. However, similarity searching and molecular docking will be better choices for those targets with a small number of active molecules, such as GP41, PKC, CYP3A, NS4B and NS5A whose active molecules are less than 30. Besides, machine learning models and similarity searching neglect the interaction between receptor and ligands compared with molecular docking. Therefore, it is necessary to choose a reasonable single-method or combining of multi-methods according to specific situation.

In summary, computational methods utilized in this study could effectively detect multitarget co-inhibitors for suppression of HIV/HCV co-infection, and have potential application in identifying new targets, virtually screening of multitarget drugs and drug repurposing in treatment of other diseases. In addition, we hope that our research will be of help to the clinic treatment of HIV/HCV co-infection.

4. Materials and methods

4.1 Dataset preparation

Therapeutic Target Database (TTD)[61] and ChEMBL Database (version 23, <https://www.ebi.ac.uk/chembl/>) were used to collect the clinical targets of anti-HIV and anti-HCV therapy. We used keywords with ‘HIV-1’, ‘Human Immunodeficiency Virus type-1’, ‘HCV’ and ‘Hepatitis C Virus’ as the retrieval conditions, and the search results were further filtered to keep the targets with drug candidates that at least entered phase I clinical trials.

In this study, the known small molecule inhibitors of the HIV-1 or HCV targets were downloaded from the ChEMBL Database, respectively. Data sets were processed according to the following requirements: (1) the compound, with IC_{50} , EC_{50} or $K_i \leq 20 \mu M$, was retained as positive (expressed as ‘+1’); (2) salts were translated into the homologous acid or base, water molecules in the hydrate were eliminated; (3) the duplicate

molecules were deleted. On the other hand, DecoyFinder software[62,63] was applied to prepare a set of decoy compounds (3 decoys for each known inhibitor)[64] from ZINC database (version 12, <http://zinc.docking.org/>), and decoy compounds were used as negative (expressed as '-1') for each target. Particularly, decoy compounds were selected based on the following constraints: (i) the MW is within 25 Da of positive compounds; (ii) the number of RB, HBDs, HBAs are ± 1 , ± 1 and ± 2 of positive compounds, respectively; (iii) the LogP is within 1 of positive compounds; (iv) the Tanimoto coefficients of a potential decoy vs each active compound as well as a potential decoy vs previously selected decoys are the default of 0.75 and 0.9, respectively[62]. Finally, the data set including inhibitors and decoys for each target was divided into training set and test set. The data processing procedures were performed by KNIME (version 3.4.2, <https://www.knime.com>), an open-source data analysis and cheminformatic software.

4.2 Molecular representation

Each compound in the above data sets was represented using two different types of binary fingerprint descriptors (MACCS and ECFP6) as chemical descriptors calculated in KNIME. MACCS is a common 2D molecular fingerprint descriptor, composed of 166 structural keys. And most of the important chemical features are covered in spite of small length[65]. Additionally, Extended-connectivity fingerprints (ECFPs) are a new kind of topological fingerprints, which were clearly developed to capture molecular features associated with molecular activity[23]. At present, MACCS and ECFPs have been widely used in drug discovery on the basis of the characteristic of calculating rapidly and representing a lot of different molecular features and stereochemical information.

4.3 mt-QSAR models building

In this study, all classification models were built by combining two machine learning algorithms (naive Bayesian and support vector machine) and two molecular fingerprints (MACCS and ECFP6). In short, four

classifiers (NB_MACCS, NB_ECFP6, SVM_MACCS and SVM_ECFP6) were generated for each target. Therefore, the numbers of mt-QSAR models for 11 HIV-1 targets and 4 HCV targets are 44 and 16, respectively.

4.3.1 Naive Bayesian (NB)

The Bayesian classification is widely used to solve classification problems based on Bayesian theory, such as discriminating between active and inactive compounds by building classifier and so on. And naive Bayesian is one of the simplest and the most common classifier, which adopts the hypothesis of attribution conditional independence[19]. In other words, it is assumed that each feature has an effect on the classification result independently. The details of the Bayesian theorem are given in equations (1) – (3):

$$P(y_i|x) = \frac{P(x|y_i)P(y_i)}{P(x)} \quad (1)$$

$$P(y_i|x) = \frac{P(y_i)}{P(x_1, x_2, \dots, x_m)} P(x_1, x_2, \dots, x_m | y_i) \quad (2)$$

$$P(y_i|x) = \frac{P(y_i)}{P(x_1, x_2, \dots, x_m)} \prod_{j=1}^d P(x_j|y_i) \quad (3)$$

Here, x denotes a set of features such as a molecular fingerprint is the feature set of a molecule; y represents a set of categories that mark the molecular activity category with “+1” or “-1”; x_j , y_i and d are j -th attribute value, i -th category and the amount of attributes, respectively. Besides, $P(y_i)$ means prior probabilities, which is calculated by analyzing the training set data; $P(x_1, x_2, \dots, x_m)$ is a constant for the explicit training set.

Because of the assumption of attribution independence, equation 2 is simplified to equation 3, namely the naive Bayesian formula.

4.3.2 Support vector machine (SVM)

Based on the principle of structural risk minimization(SRM)[66], SVM has been developed as a useful classification technique. The main idea is to convert input sample space into a high-dimensional space, in

which a hyperplane with the maximal margin is generated to describe non-linear classification boundary. Importantly, the classification is achieved through a kernel function. There are four basic kernels: linear, polynomial, radial basis function (RBF) and sigmoid[67]. In most cases, RBF is preferred[67]. In this study, the RBF kernel function was also used. The penalty parameter C and kernel parameter γ , involved in SVM, were optimized through 5-fold cross-validation.

4.4 Performance evaluation of mt-QSAR models

All models were evaluated by the internal 5-fold cross-validation and external test set validation. In 5-fold cross-validation, the training set was divided into 5 subsets equally. Four subsets were used to train the model, and the remaining one was used as an internal validation set.

The major evaluation indexes include true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), sensitivity (SE), specificity (SP), the overall prediction accuracy (Q) and the receiver operating characteristic curve (ROC). Among them, the ROC curves are often used to estimate the merits of a binary classifier, in which the false positive rate (FPR) and true positive rate (TPR) are used as abscissa and ordinate, respectively. Generally, the AUC value ranges from 0 to 1, and the higher the value, the better the model's classification capability is.

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

In addition, Matthews correlation coefficient (MCC) is calculated by equation 4. And the value of MCC ranges from -1 to +1.

4.5 Molecular docking

The docking algorithm Glide[68] was used for molecular docking and analysis of the binding modes between molecules and target proteins. The crystal structure of HIV-1 IN (PDB: 4NYF)[69], HIV-1 RT (PDB: 1TL3)[70], HIV-1 PR (PDB: 4MC1)[71] and HCV NS5B (PDB: 1OS5 and 1GX6)[72,73] were prepared and

used to build the receptor grid. The grid was centered on the ligand of co-crystallized complex or the key residues of proteins, and then the Glide SP protocol was utilized for docking. For the grid generation and docking, all of the parameters are default values.

Acknowledgments

This study was supported by the National Key R&D Program of China (No. 2017YFC1104400).

References

1. Wang S, Milne GWA, Yan X, Posey IJ, Nicklaus MC, Graham L, et al. Discovery of novel, non-peptide HIV-1 protease inhibitors by pharmacophore searching. *J Med Chem*. 1996;39: 2047–2054. doi:10.1021/jm950874+
2. Therese PJ, Manvar D, Kondepudi S, Battu MB, Sriram D, Basu A, et al. Multiple e-pharmacophore modeling, 3D-QSAR, and high-throughput virtual screening of hepatitis C virus NS5B polymerase inhibitors. *J Chem Inf Model*. 2014;54: 539–552. doi:10.1021/ci400644r
3. Laguno M, Murillas J, Blanco JL, Martínez E, Miquel R, Sánchez-Tapias JM, et al. Peginterferon alfa-2b plus ribavirin compared with interferon alfa-2b plus ribavirin for treatment of HIV/HCV co-infected patients. *AIDS*. 2004;18: F27–F36. doi:00002030-200409030-00003
4. WHO. Global Hepatitis Report, 2017. Available: <http://www.who.int/hepatitis>
5. Wang Z, Bennett EM, Wilson DJ, Salomon C, Vince R. Rationally designed dual inhibitors of HIV reverse transcriptase and integrase. *J Med Chem*. 2007;50: 3416–3419. doi:10.1021/jm070512p
6. Cox BD, Prosser AR, Sun Y, Li Z, Lee S, Huang MB, et al. Pyrazolo-Piperidines Exhibit Dual Inhibition of CCR5/CXCR4 HIV Entry and Reverse Transcriptase. *ACS Med Chem Lett*. 2015;6: 753–757. doi:10.1021/acsmchemlett.5b00036
7. Operskalski EA, Kovacs A. HIV/HCV co-infection: Pathogenesis, clinical complications, treatment, and new therapeutic technologies. *Curr HIV/AIDS Rep*. 2011;8: 12–22. doi:10.1007/s11904-010-0071-3
8. Bajorath J. Computational analysis of ligand relationships within target families. *Curr Opin Chem Biol*. 2008;12: 352–358. doi:10.1016/j.cbpa.2008.01.044
9. Rognan D. Chemogenomic approaches to rational drug design. *Br J Pharmacol*. 2007;152: 38–52. doi:10.1038/sj.bjp.0707307
10. Wang X, Shen Y, Wang S, Li S, Zhang W, Liu X, et al. PharmMapper 2017 update: a web server for potential drug target identification with a comprehensive target pharmacophore database. *Nucleic Acids Res*. 2017;45: W356–W360. doi:10.1093/nar/gkx374
11. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK. Relating protein pharmacology by ligand chemistry. *Nat Biotechnol*. 2007;25: 197–206. doi:10.1038/nbt1284
12. Bauer RA, Bourne PE, Formella A, Frömmel C, Gille C, Goede A, et al. Superimpose: a 3D structural superposition server. *Nucleic Acids Res*. 2008;36: W47–W54. doi:10.1093/nar/gkn285
13. Li H, Gao Z, Kang L, Zhang H, Yang K, Yu K, et al. TarFisDock: a web server for identifying drug

- 425 targets with docking approach. *Nucleic Acids Res.* 2006;34: W219–W224. doi:10.1093/nar/gkl114
- 426 14. Chen YZ, Zhi DG. Ligand-protein inverse docking and its potential use in the computer search of
427 protein targets of a small molecule. *Proteins.* 2001;43: 217–226. doi:10.1002/1097-
428 0134(20010501)43:2<217::AID-PROT1032>3.0.CO;2-G
- 429 15. Viña D, Uriarte E, Orallo F, González-Díaz H. Alignment-free prediction of a drug-target complex
430 network based on parameters of drug connectivity and protein sequence of receptors. *Mol Pharm.*
431 2009;6: 825–835. doi:10.1021/mp800102c
- 432 16. Jacob L, Vert J-P. Protein-ligand interaction prediction: an improved chemogenomics approach.
433 *Bioinformatics.* 2008;24: 2149–2156. doi:10.1093/bioinformatics/btn409
- 434 17. Wang F, Liu D, Wang H, Luo C, Zheng M, Liu H, et al. Computational screening for active
435 compounds targeting protein sequences: methodology and experimental validation. *J Chem Inf*
436 *Model.* 2011;51: 2821–2828. doi:10.1021/ci200264h
- 437 18. Cheng F, Zhou Y, Li J, Li W, Liu G, Tang Y. Prediction of chemical-protein interactions: multitarget-
438 QSAR versus computational chemogenomic methods. *Mol Biosyst.* 2012;8: 2373–2384.
439 doi:10.1039/c2mb25110h
- 440 19. Porwal A, Carranza EJM, Hale M. Bayesian network classifiers for mineral potential mapping.
441 *Comput Geosci.* 2006;32: 1–16. doi:10.1016/j.cageo.2005.03.018
- 442 20. Xia X, Maliski EG, Gallant P, Rogers D. Classification of kinase inhibitors using a Bayesian model. *J*
443 *Med Chem.* 2004;47: 4463–4470. doi:10.1021/jm0303195
- 444 21. Jorissen RN, Gilson MK. Virtual Screening of Molecular Databases Using a Support Vector Machine.
445 *J Chem Inf Model.* 2005;45: 549–561. doi:10.1021/ci049641u
- 446 22. Durant JL, Leland BA, Henry DR, Nourse JG. Reoptimization of MDL keys for use in drug
447 discovery. *J Chem Inf Comput Sci.* 2002;42: 1273–1280. doi:10.1021/ci010132r
- 448 23. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model.* 2010;50: 742–754.
449 doi:10.1021/ci100050t
- 450 24. Bajusz D, Rácz A, Héberger K. Why is Tanimoto index an appropriate choice for fingerprint-based
451 similarity calculations? *J Cheminform.* 2015;7: 20. doi:10.1186/s13321-015-0069-3
- 452 25. Maggiora GM, Shanmugasundaram V. Molecular Similarity Measures. *Chemoinformatics.* 2004;275:
453 1–50. doi:10.1385/1-59259-802-1:001
- 454 26. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Zhou Z, et al. PubChem’s BioAssay Database. *Nucleic*
455 *Acids Res.* 2012;40: D400–D412. doi:10.1093/nar/gkr1132
- 456 27. Chong H, Qiu Z, Su Y, He Y. The N-Terminal T-T Motif of a Third-Generation HIV-1 Fusion
457 Inhibitor Is Not Required for Binding Affinity and Antiviral Activity. *J Med Chem.* 2015;58: 6378–
458 6388. doi:10.1021/acs.jmedchem.5b00109
- 459 28. Nugent RA, Schlachter ST, Murphy MJ, Cleek GJ, Poel TJ, Wishka DG, et al. Pyrimidine thioethers:
460 a novel class of HIV-1 reverse transcriptase inhibitors with activity against BHAP-resistant HIV. *J*
461 *Med Chem.* 1998;41: 3793–3803. doi:10.1021/jm9800806
- 462 29. Sluis-Cremer N, Koontz D, Bassit L, Hernandez-Santiago BI, Detorio M, Rapp KL, et al. Anti-human
463 immunodeficiency virus activity, cross-resistance, cytotoxicity, and intracellular pharmacology of the
464 3'-azido-2',3'- dideoxypurine nucleosides. *Antimicrob Agents Chemother.* 2009;53: 3715–3719.
465 doi:10.1128/AAC.00392-09
- 466 30. Tamamura H, Omagari A, Hiramatsu K, Gotoh K, Kanamoto T, Xu Y, et al. Development of specific
467 CXCR4 inhibitors possessing high selectivity indexes as well as complete stability in serum based on
468 an anti-HIV peptide T140. *Bioorg Med Chem Lett.* 2001;11: 1897–1902. doi:10.1016/S0960-
469 894X(01)00323-7
- 470 31. Kobayashi M, Yoshinaga T, Seki T, Wakasa-Morimoto C, Brown KW, Ferris R, et al. In vitro
471 antiretroviral properties of S/GSK1349572, a next-generation HIV integrase inhibitor. *Antimicrob*

- Agents Chemother. 2011;55: 813–821. doi:10.1128/AAC.01209-10
32. Costi R, Métifiot M, Chung S, Cuzzucoli Crucitti G, Maddali K, Pescatori L, et al. Basic quinolinonyl diketo acid derivatives as inhibitors of HIV integrase and their activity against RNase H function of reverse transcriptase. *J Med Chem*. 2014;57: 3223–3234. doi:10.1021/jm5001503
33. Sherrill RG, Furfine ES, Hazen RJ, Miller JF, Reynolds DJ, Sammond DM, et al. Synthesis and antiviral activities of novel N-alkoxy-arylsulfonamide-based HIV protease inhibitors. *Bioorganic Med Chem Lett*. 2005;15: 3560–3564. doi:10.1016/j.bmcl.2005.05.101
34. Bold G, Fässler A, Capraro HG, Cozens R, Klimkait T, Lazdins J, et al. New aza-dipeptide analogues as potent and orally absorbed HIV-1 protease inhibitors: candidates for clinical development. *J Med Chem*. 1998;41: 3387–3401. doi:10.1021/jm970873c
35. Eissenstat M, Guerassina T, Gulnik S, Afonina E, Silva AM, Ludtke D, et al. Enamino-oxindole HIV protease inhibitors. *Bioorg Med Chem Lett*. 2012;22: 5078–5083. doi:10.1016/j.bmcl.2012.05.120
36. Dong M, Zhang J, Peng X, Lu H, Yun L, Jiang S, et al. Tricyclononene carboxamide derivatives as novel anti-HIV-1 agents. *Eur J Med Chem*. 2010;45: 4096–4103. doi:10.1016/j.ejmech.2010.05.070
37. Murphy P V., O'Brien JL, Gorey-Feret LJ, Smith AB. Structure-based design and synthesis of HIV-1 protease inhibitors employing β -d-mannopyranoside scaffolds. *Bioorg Med Chem Lett*. 2002;12: 1763–1766. doi:10.1016/S0960-894X(02)00220-2
38. Liu H, Xu L, Hui H, Vivian R, Callebaut C, Murray BP, et al. Structure-activity relationships of diamine inhibitors of cytochrome P450 (CYP) 3A as novel pharmacoenhancers, part I: Core region. *Bioorganic Med Chem Lett*. Elsevier Ltd; 2014;24: 989–994. doi:10.1016/j.bmcl.2013.12.058
39. Thaisrivongs S, Romero DL, Tommasi R a, Janakiraman MN, Strohbach JW, Turner SR, et al. Structure-based design of HIV protease inhibitors: 5,6-dihydro-4-hydroxy-2-pyrones as effective, nonpeptidic inhibitors. *J Med Chem*. 1996;39: 4630–4642. doi:10.1021/jm960228q
40. Zhou Y, Müh U, Hanzelka BL, Bartels DJ, Wei Y, Rao BG, et al. Phenotypic and structural analyses of hepatitis C virus NS3 protease Arg155 variants: sensitivity to telaprevir (VX-950) and interferon alpha. *J Biol Chem*. 2007;282: 22619–22628. doi:10.1074/jbc.M610207200
41. Ding CZ, Zhang Y-K, Li X, Liu Y, Zhang S, Zhou Y, et al. Synthesis and biological evaluations of P4-benzoxaborole-substituted macrocyclic inhibitors of HCV NS3 protease. *Bioorg Med Chem Lett*. 2010;20: 7317–7322. doi:10.1016/j.bmcl.2010.10.071
42. Neelamkavil SF, Agrawal S, Bara T, Bennett C, Bhat S, Biswas D, et al. Discovery of MK-8831, A Novel Spiro-Proline Macrocycle as a Pan-Genotypic HCV-NS3/4a Protease Inhibitor. *ACS Med Chem Lett*. 2016;7: 111–116. doi:10.1021/acsmedchemlett.5b00425
43. Meguellati A, Ahmed-Belkacem A, Nurisso A, Yi W, Brillet R, Berqouch N, et al. New pseudodimeric aurones as palm pocket inhibitors of Hepatitis C virus RNA-dependent RNA polymerase. *Eur J Med Chem*. 2016;115: 217–229. doi:10.1016/j.ejmech.2016.03.005
44. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13: 2498–2504. doi:10.1101/gr.1239303
45. Tong JB, Bai M, Zhao X. 3D-QSAR and docking studies of HIV-1 protease inhibitors using R-group search and Surflex-dock. *Med Chem Res*. 2016;25: 2619–2630. doi:10.1007/s00044-016-1701-0
46. Wei Y, Li J, Chen Z, Wang F, Huang W, Hong Z, et al. Multistage virtual screening and identification of novel HIV-1 protease inhibitors by integrating SVM, shape, pharmacophore and docking methods. *Eur J Med Chem*. 2015;101: 409–418. doi:10.1016/j.ejmech.2015.06.054
47. Ryu K, Kim ND, Choi S Il, Han CK, Yoon JH, No KT, et al. Identification of novel inhibitors of HCV RNA-dependent RNA polymerase by pharmacophore-based virtual screening and in vitro evaluation. *Bioorg Med Chem*. 2009;17: 2975–2982. doi:10.1016/j.bmc.2009.03.024
48. Wei Y, Li J, Qing J, Huang M, Wu M, Gao F, et al. Discovery of Novel Hepatitis C Virus NS5B

- 519 Polymerase Inhibitors by Combining Random Forest, Multiple e-Pharmacophore Modeling and
520 Docking. PLoS One. 2016;11: e0148181. doi:10.1371/journal.pone.0148181
- 521 49. Sechi M, Bacchi A, Carcelli M, Compari C, Duce E, Fiscaro E, et al. From ligand to complexes:
522 inhibition of human immunodeficiency virus type 1 integrase by beta-diketo acid metal complexes. J
523 Med Chem. 2006;49: 4248–4260. doi:10.1021/jm060193m
- 524 50. Zhang X, Pais GCG, Svarovskaia ES, Marchand C, Johnson AA, Karki RG, et al. Azido-containing
525 aryl beta-diketo acid HIV-1 integrase inhibitors. Bioorg Med Chem Lett. 2003;13: 1215–1219.
526 doi:10.1016/S0960-894X(03)00059-3
- 527 51. Zeng L-F, Jiang X-H, Sanchez T, Zhang H-S, Dayam R, Neamati N, et al. Novel dimeric aryldiketo
528 containing inhibitors of HIV-1 integrase: effects of the phenyl substituent and the linker orientation.
529 Bioorg Med Chem. 2008;16: 7777–7787. doi:10.1016/j.bmc.2008.07.008
- 530 52. Zhang X, Marchand C, Pommier Y, Burke TR. Design and synthesis of photoactivatable aryl diketo
531 acid-containing HIV-1 integrase inhibitors as potential affinity probes. Bioorg Med Chem Lett.
532 2004;14: 1205–1207. doi:10.1016/j.bmcl.2003.12.064
- 533 53. Stansfield I, Avolio S, Colarusso S, Gennari N, Narjes F, Pacini B, et al. Active site inhibitors of
534 HCV NS5B polymerase. The development and pharmacophore of 2-thienyl-5,6-
535 dihydroxypyrimidine-4-carboxylic acid. Bioorg Med Chem Lett. 2004;14: 5085–5088.
536 doi:10.1016/j.bmcl.2004.07.075
- 537 54. Summa V, Petrocchi A, Pace P, Matassa VG, De Francesco R, Altamura S, et al. Discovery of
538 alpha,gamma-diketo acids as potent selective and reversible inhibitors of hepatitis C virus NS5b
539 RNA-dependent RNA polymerase. J Med Chem. 2004;47: 14–17. doi:10.1021/jm0342109
- 540 55. Bhatt A, Gurukumar KR, Basu A, Patel MR, Kaushik-Basu N, Talele TT. Synthesis and SAR
541 optimization of diketo acid pharmacophore for HCV NS5B polymerase inhibition. Eur J Med Chem.
542 2011;46: 5138–5145. doi:10.1016/j.ejmech.2011.08.028
- 543 56. Di Santo R, Fermeglia M, Ferrone M, Paneni MS, Costi R, Artico M, et al. Simple but highly
544 effective three-dimensional chemical-feature-based pharmacophore model for diketo acid derivatives
545 as hepatitis C virus RNA-dependent RNA polymerase inhibitors. J Med Chem. 2005;48: 6304–6314.
546 doi:10.1021/jm0504454
- 547 57. Gopalsamy A, Chopra R, Lim K, Ciszewski G, Shi M, Curran KJ, et al. Discovery of proline
548 sulfonamides as potent and selective hepatitis C virus NS5b polymerase inhibitors. Evidence for a
549 new NS5b polymerase binding site. J Med Chem. 2006;49: 3052–3055. doi:10.1021/jm060168g
- 550 58. Vara Prasad JVN, Boyer FE, Domagala JM, Ellsworth EL, Gajda C, Hagen SE, et al. Nonpeptidic
551 HIV protease inhibitors: 6-alkyl-5, 6-dihydropyran-2-ones possessing achiral 3-(4-
552 amino/carboxamide-2-t-butyl, 5-methylphenyl thio) moiety: Antiviral activities and pharmacokinetic
553 properties. Bioorg Med Chem Lett. 1999;9: 1481–1486. doi:10.1016/S0960-894X(99)00237-1
- 554 59. Hao W, Herlihy KJ, Zhang NJ, Fuhrman SA, Doan C, Patick AK, et al. Development of a novel
555 dicistronic reporter-selectable hepatitis C virus replicon suitable for high-throughput inhibitor
556 screening. Antimicrob Agents Chemother. 2007;51: 95–102. doi:10.1128/AAC.01008-06
- 557 60. Billamboz M, Bailly F, Lion C, Calmels C, Andréola ML, Witvrouw M, et al. 2-
558 Hydroxyisoquinoline-1,3(2H,4H)-diones as inhibitors of HIV-1 integrase and reverse transcriptase
559 RNase H domain: Influence of the alkylation of position 4. Eur J Med Chem. 2011;46: 535–546.
560 doi:10.1016/j.ejmech.2010.11.033
- 561 61. Li YH, Yu CY, Li XX, Zhang P, Tang J, Yang Q, et al. Therapeutic target database update 2018:
562 enriched resource for facilitating bench-to-clinic research of targeted therapeutics. Nucleic Acids Res.
563 2018;46: D1121–D1127. doi:10.1093/nar/gkx1076
- 564 62. Cereto-Massagué A, Guasch L, Valls C, Mulero M, Pujadas G, Garcia-Vallvé S. DecoyFinder: an
565 easy-to-use python GUI application for building target-specific decoy sets. Bioinformatics. 2012;28:

- 1661–1662. doi:10.1093/bioinformatics/bts249
63. Cereto-Massagué A, Pujadas G, Garcia-Vallve S. DecoyFinder, a tool for finding decoy molecules. *J Cheminform.* 2012;4: P2. doi:10.1186/1758-2946-4-S1-P2
64. Fang J, Li Y, Liu R, Pang X, Li C, Yang R, et al. Discovery of multitarget-directed ligands against Alzheimer's disease through systematic prediction of chemical-protein interactions. *J Chem Inf Model.* 2015;55: 149–164. doi:10.1021/ci500574n
65. Muegge I, Mukherjee P. An overview of molecular fingerprint similarity search in virtual screening. *Expert Opin Drug Discov.* 2016;11: 137–148. doi:10.1517/17460441.2016.1117070
66. Zhang X, Yang L. Improving SVM through a Risk Decision Rule Running on MATLAB. *J Softw.* 2012;7: 2252–2257. doi:10.4304/jsw.7.10.2252-2257
67. Yang C, Odvody GN, Fernandez CJ, Landivar JA, Minzenmayer RR, Nichols RL. Evaluating unsupervised and supervised image classification methods for mapping cotton root rot. *Precis Agric.* 2015;16: 201–215. doi:10.1007/s11119-014-9370-9
68. Glide, version 6.7, Schrödinger, LLC, New York, NY, 2015.
69. Fader LD, Malenfant E, Parisien M, Carson R, Bilodeau F, Landry S, et al. Discovery of BI 224436, a Nucleoside Site Integrase Inhibitor (NCINI) of HIV-1. *ACS Med Chem Lett.* 2014;5: 422–427. doi:10.1021/ml500002n
70. Freeman GA, Andrews Iii CW, Hopkins AL, Lowell GS, Schaller LT, Cowan JR, et al. Design of non-nucleoside inhibitors of HIV-1 reverse transcriptase with improved drug resistance properties. 2. *J Med Chem.* 2004;47: 5923–5936. doi:10.1021/jm040072r
71. Ganguly AK, Alluri SS, Wang C-H, Antropow A, White A, Carocchia D, et al. Structural optimization of cyclic sulfonamide based novel HIV-1 protease inhibitors to picomolar affinities guided by X-ray crystallographic analysis. *Tetrahedron.* 2014;70: 2894–2904. doi:10.1016/j.tet.2014.03.038
72. Love RA, Parge HE, Yu X, Hickey MJ, Diehl W, Gao J, et al. Crystallographic identification of a noncompetitive inhibitor binding site on the hepatitis C virus NS5B RNA polymerase enzyme. *J Virol.* 2003;77: 7575–7581. doi:10.1128/JVI.77.13.7575–7581.2003
73. Bressanelli S, Tomei L, Rey FA, De Francesco R. Structural analysis of the hepatitis C virus RNA polymerase in complex with ribonucleotides. *J Virol.* 2002;76: 3482–3492. doi:10.1128/JVI.76.7.3482-3492.2002

Figure Captions

Fig 1. Flowchart of mt-QSAR models building process toward HIV-1/HCV co-infection (the blue, red and orange objects represent the phase of construction, validation and application of mt-QSAR models, respectively).

Fig 2. Summary of targets related to HIV-1 (A) and HCV (B) and the total number of active compounds for each target.

Fig 3. Targets classification of the HIV-1 (A) and HCV (B) data sets.

Fig 4. Performance of 60 mt-QSAR models built based on naive Bayesian (A) and support vector machine (B) using different fingerprints by test set validation.

Fig 5. Performance of 60 mt-QSAR models built based on MACCS (A) and ECFP6 (B) using different algorithms by test set validation.

Fig 6. Number of predicted targets and that of known targets for 17 HIV-1 approved drugs and 4 HCV approved drugs.

Fig 7. Polypharmacology analysis of 17 HIV-1 approved drugs and 4 HCV approved drugs based on the 60 mt-QSAR models in this study. Ellipses and octagons represent drug nodes and protein nodes, respectively. The red lines indicate that the CPI was experimentally verified as active. The black lines present that the interaction was not proved and worthy of further experimental validation. And the green and yellow lines express that the interaction was experimentally validated as inactive and inconclusive, respectively.

Fig 8. The structure and bioactivity data of 9 selected compounds.

Fig 9. The binding modes of 9 compounds in the respective binding site. A and B present the binding modes of 5 compounds with HIV-1 IN (PDB: 4NYF) and HCV NS5B (PDB: 1GX6), respectively, including CHEMBL16326 (red), CHEMBL18927 (pink), CHEMBL449221 (purple), CHEMBL502238 (orange) and CHEMBL19332 (grey). C and D present the binding modes of the compound (CHEMBL210593) with HIV-1 RT (PDB: 1TL3) and HCV NS5B (PDB: 1OS5). E and F present the binding modes of the compound (CHEMBL37541) with HIV-1 PR (PDB: 4MC1) and HCV NS5B (PDB: 1OS5). G and H present the binding

622 modes of two compounds (CHEMBL3612421 and CHEMBL1668670) with HCV NS5B (PDB 1OS5),
623 respectively.
624

Fig 1

Fig 2

Fig 3

(A) Naive Bayesian classifiers

(B) Support Vector Machine classifiers

Fig 4

Fig 5

Fig 6

Fig 7

CHEMBL16326
HIV IN ($IC_{50}=69\mu M^{(59)}$)
HCV NS5B ($IC_{50}=5.7\mu M^{(63)}$)

CHEMBL18927
HIV IN ($IC_{50}=350nM^{(60)}$)
HCV NS5B ($IC_{50}=8\mu M^{(64)}$)

CHEMBL449221
HIV IN ($IC_{50}=58\mu M^{(61)}$)
HCV NS5B ($IC_{50}=14.9\mu M^{(65)}$)

CHEMBL502238
HIV IN ($IC_{50}=27\mu M^{(61)}$)
HCV NS5B ($IC_{50}=40\mu M^{(66)}$)

CHEMBL19332
HIV IN ($IC_{50}=25\mu M^{(62)}$)
HIV RT ($IC_{50}=54\mu M^{(64)}$)
HCV NS5B ($IC_{50}=8\mu M^{(64)}$)

CHEMBL210593
HIV RT ($IC_{50}>100\mu M^{(67)}$)
HCV NS5B ($IC_{50}=3.1\mu M^{(67)}$)

CHEMBL37541
HIV PR ($IC_{50}=6nM^{(68)}$)
HCV NS5B ($IC_{50}=920nM^{(69)}$)

CHEMBL3612421
HIV PR ($IC_{50}=62\mu M^{(56)}$)

CHEMBL1668670
HIV IN ($IC_{50}=23.2\mu M^{(70)}$)
HIV RT ($IC_{50}=42.7\mu M^{(70)}$)

Fig 8

Fig 9

Fig 1

Fig 2

Fig 3

(A) Naive Bayesian classifiers

(B) Support Vector Machine classifiers

Fig 4

(A) MACCS Fingerprint

(B) ECFP6 Fingerprint

Fig 5

Fig 6

Fig 7

CHEMBL16326

HIV IN ($IC_{50}=69\mu M^{(59)}$)

HCV NS5B ($IC_{50}=5.7\mu M^{(63)}$)

CHEMBL18927

HIV IN ($IC_{50}=350nM^{(60)}$)

HCV NS5B ($IC_{50}=8\mu M^{(64)}$)

CHEMBL449221

HIV IN ($IC_{50}=58\mu M^{(61)}$)

HCV NS5B ($IC_{50}=14.9\mu M^{(65)}$)

CHEMBL502238

HIV IN ($IC_{50}=27\mu M^{(61)}$)

HCV NS5B ($IC_{50}=40\mu M^{(66)}$)

CHEMBL19332

HIV IN ($IC_{50}=25\mu M^{(62)}$)

HIV RT ($IC_{50}=54\mu M^{(64)}$)

HCV NS5B ($IC_{50}=8\mu M^{(64)}$)

CHEMBL210593

HIV RT ($IC_{50}>100\mu M^{(67)}$)

HCV NS5B ($IC_{50}=3.1\mu M^{(67)}$)

CHEMBL37541

HIV PR ($IC_{50}=6nM^{(68)}$)

HCV NS5B ($IC_{50}=920nM^{(69)}$)

CHEMBL3612421

HIV PR ($IC_{50}=62\mu M^{(56)}$)

CHEMBL1668670

HIV IN ($IC_{50}=23.2\mu M^{(70)}$)

HIV RT ($IC_{50}=42.7\mu M^{(70)}$)

Fig 8

Fig 9