Mol Divers (2016) 20:439–451
DOI 10.1007/s11030-015-9641-z

FULL-LENGTH PAPER

Discovery of Influenza A virus neuraminidase inhibitors using support vector machine and Naïve Bayesian models

Wenwen Lian[1] · Jiansong Fang[1] · Chao Li[1] · Xiaocong Pang[1] · Ai-Lin Liu[1,2,3] · Guan-Hua Du[1,2,3]

Received: 15 June 2015 / Accepted: 12 October 2015 / Published online: 21 December 2015

Abstract Neuraminidase (NA) is a critical enzyme in the life cycle of influenza virus, which is known as a successful paradigm in the design of anti-influenza agents. However, to date there are no classification models for the virtual screening of NA inhibitors. In this work, we built support vector machine and Naïve Bayesian models of NA inhibitors and non-inhibitors, with different ratios of active-to-inactive compounds in the training set and different molecular descriptors. Four models with sensitivity or Matthews correlation coefficients greater than 0.9 were chosen to predict the NA inhibitory activities of 15,600 compounds in our in-house database. We combined the results of four optimal models and selected 60 representative compounds to assess their NA inhibitory profiles in vitro. Nine NA inhibitors were identified, five of which were oseltamivir derivatives with large C-5 substituents exhibiting potent inhibition against H1N1 NA with IC_{50} values in the range of 12.9–185.0 nM, and against H3N2 NA with IC_{50} values between 18.9 and 366.1 nM. The other four active compounds belonged to novel scaffolds, with IC_{50} values ranging 39.5–63.8 μM against H1N1 NA and 44.5–114.1 μM against H3N2 NA. This is the first time that classification models of NA inhibitors and non-inhibitors are built and their prediction results validated experimentally using in vitro assays.

Keywords Influenza virus · Neuraminidase inhibitor · Support vector machine · Naïve Bayesian · Virtual screening · H1N1 · H3N2 · SVM

Introduction

Influenza is a globally contagious disease, which severely impairs public health with substantial morbidity and mortality [1]. The H1N1 pandemic in 2009 and the worldwide outbreak of avian influenza H5N1 have urged us to develop new therapeutic agents for influenza. Neuraminidase (NA) consists of four identical subunits, which are located on the viral envelope. The main role of NA is to assist the release of virion progeny from infected cells by cleaving the glycosidic bond between hemagglutinin (HA) of the progeny virus and the terminal sialic acid of receptors in the host cells [2,3]. NA can also prevent virus aggregation and endow infectivity by removing sialic acid from viral envelope [2]. In addition, NA can break down the mucins in the respiratory tract at the early time of infection, allowing virus to attach to the respiratory epitheliums [2]. Thus, NA is recognized as a potential target for the prophylaxis and therapy of influenza. Currently, there have been several successful developments of inhibitors targeting the highly conserved active site of NA, for example, zanamivir (Relenza) and oseltamivir (Tamiflu) [4,5]. However, observed viral resistance to these drugs presents a threat to human health [6,7]. It has already become a challenge to modify the existing drugs or discover new NA inhibitors with novel scaffolds for drug-resistant influenza virus.

Electronic supplementary material The online version of this article (doi:10.1007/s11030-015-9641-z) contains supplementary material, which is available to authorized users.

✉ Ai-Lin Liu
liuailin@imm.ac.cn

Guan-Hua Du
dugh@imm.ac.cn

1 Institute of Materia Medica, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100050, People's Republic of China

2 Beijing Key Laboratory of Drug Target Research and Drug Screening, Beijing 100050, People's Republic of China

3 State Key Laboratory of Bioactive Substance and Function of Natural Medicines, 1 Xian Nong Tan Street, Beijing 100050, People's Republic of China

The discovery of NA inhibitors using high throughput screening (HTS) is time-consuming, labor-intensive, costly, and inefficient [8–11]. Therefore, it is necessary to improve efficiency in the early phase of drug screening via virtual screening. Machine learning attracts widespread attention in the classification of chemicals owing to large amounts of empirical data. Its main concept is to statistically build classification models to link the structural features and physicochemical properties of chemicals to their biological activities, and then predict the activities of untested chemicals [12]. Several statistical algorithms can be used to build the models, such as support vector machine (SVM) [13,14], Naïve Bayesian (NB) [15], artificial neural network (ANN) [13,16], and random forest (RF) [17]. In silico screenings based on these algorithms are efficient, with lower cost, and have been successfully applied in the prediction of inhibitors against several targets, such as acetylcholinesterase [18], Src kinase [19], and cytochromes P450 [20]. Recently, our group successfully predicted inhibitors targeting butyrylcholinesterase [21], CDK5 [22], and multi-targets in Alzheimer's Disease [23] using machine learning algorithms. To date, although there have been some trials on predicting NA inhibitors using statistical algorithms [24–26], these researchers only optimized a few of factors in the model construction and verified the usefulness and reliability of machine learning. However, the models were not tested and verified using in vitro experiments.

In the present study, we built classification models of NA inhibitors and non-inhibitors on the basis of SVM and NB algorithms, optimized the ratio of active-to-inactive compounds in the training set and molecular descriptors derived from different software packages, and then predicted the NA inhibitory activities of 15,600 compounds in our in-house database using our optimal models. Finally, hit compounds were assayed for their NA inhibitory profiles against H1N1 and H3N2 in vitro [27], resulting in the discovery of nine new NA inhibitors. The workflow used in the present study is depicted in Fig. 1.

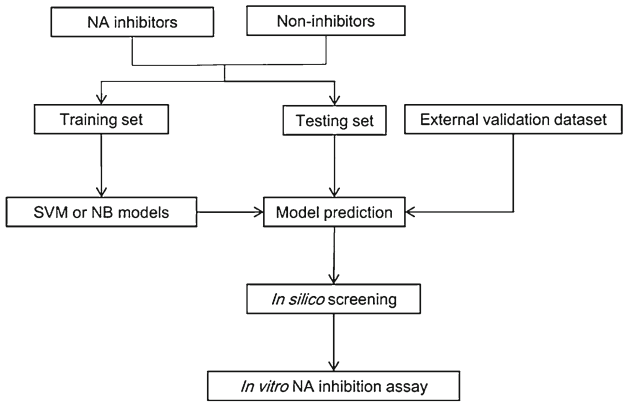

Fig. 1 Workflow of NA inhibitor discovery

Methods and materials

Data preparation

A total of 177 NA inhibitors of influenza A/PR/8/34 (H1N1) were gathered from the BindingDB database (IC_{50} $< 10\,\mu M$) [28–37]. A set of 1770 inactive compounds were extracted from the negative results of previous HTS towards NA inhibitor in our laboratory, with NA inhibition <10 % at a concentration of $4\,\mu g/mL$. All the active and inactive compounds were randomly allocated into a training set and a testing set. The training set included 143 active compounds and 1430 inactive compounds (with the ratio of active-to-inactive compounds 1:10), and the testing set contained 34 active compounds and 340 inactive compounds. In order to optimize the ratio of active-to-inactive compounds in the training set, in two independent operations we randomly extracted 429 and 143 compounds from 1430 inactive compounds. The other two training sets included 143 active compounds and 429 inactive compounds (with the ratio of active-to-inactive compounds 1:3) and 143 active compounds and 143 inactive compounds (with the ratio of active-to-inactive compounds 1:1), respectively.

In addition, the external validation dataset was composed of 47 active compounds and 470 inactive compounds, which were not included in the training and testing sets. The 47 NA inhibitors of influenza A virus with $IC_{50} < 10\,\mu M$ were gathered from the literature which were not included by BindingDB database [38–45]. Another 470 inactive compounds were extracted from the negative results of previous HTS towards NA inhibitor in our laboratory, all exhibiting NA inhibition <10 % at $4\,\mu g/mL$.

All the compounds were processed in MOE before calculating molecular descriptors, including adding hydrogen atoms, deprotonating strong acids, protonating strong bases, generating stereoisomers, and valid single 3D conformers. The active and inactive compounds were labeled "1" and "−1," respectively. Detailed information for the training sets, testing set, and external validation dataset is available in the Supporting Information Tables S1–S5.

Molecular descriptors and fingerprints

There were three sets of molecular descriptors and one fingerprint used in this study. A total of 256 2D molecular descriptors were calculated using Discovery Studio 4.1 (DS 4.1) [46], including AlogP, estate keys, molecular properties, molecular property counts, surface area and volume, and topological descriptor. The fingerprint ECFP_6 was also calculated in DS 4.1.

A total of 186 2D molecular descriptors were calculated using Molecular Operating Environment 2010 (MOE 2010) [47]. 2D molecular descriptors are calculated from the atoms and connection information of molecules, including physical properties, subdivided surface areas, atom counts and bond counts, Kier & Hall connectivity and Kappa shape indices, adjacency and distance matrix descriptors, pharmacophore feature descriptors, and partial charge descriptors.

Another 213 descriptors were calculated in ADRIANA.Code, including global molecular descriptors, size and shape descriptors, 2D property autocorrelation descriptors, and 3D property autocorrelation descriptors [48].

Molecular descriptors selection

Some molecular descriptors may have little correlation with NA inhibitory activity, or correlate with other descriptors. These may influence the predictive accuracy of model, and decrease the speed of calculations. Therefore, molecular descriptor selection was conducted in SPSS 17.0 [49] following the next three rules. Firstly, the molecular descriptors whose values are constant in more than 50 % compounds were removed. Secondly, a Pearson correlation analysis was conducted to eliminate those molecular descriptors whose correlation coefficients with activity were less than 0.1. If the correlation coefficient between two molecular descriptors was greater than 0.9, the molecular descriptor with a lower correlation coefficient with activity was removed. Thirdly, a stepwise linear regression was performed for the remaining molecular descriptors, and the molecular descriptors which were finally kept in the regression equation were finally chosen for further use [21].

Following these 3 rules, the final molecular descriptors included 15 molecular descriptors from DS 4.1, 28 molecular descriptors from MOE 2010, and 31 molecular descriptors from ADRIANA.Code. The descriptors finally chosen are listed in Table 1. From the selected molecular descriptors, we can conclude that some properties involved the partial charge of atoms, surface area and volume, hydrophobicity, and hydrogen bond acceptor/donor atoms are important in constructing a model.

Models building

Support vector machine

SVM is a pattern recognized algorithm developed by Vapnik [50], and its main principle is to project the data into a multi-dimensional space in which data can be classified by a hyperplane with maximal margin and minimal error rate [51]. More details on SVM can be found in the literature [50,52].

SVM is a supervised machine learning method, and consistently shows excellent performance for biological property prediction of compounds. Furthermore, SVM has been increasingly popular for drug discovery and biological activity prediction. In this work, SVM was conducted using the LIBSVM 2.9 package [53] with radial basis function (RBF).

Table 1 Molecular descriptors selected in this study

Descriptor class	Number of descriptors	Descriptors
MOE	28	BCUT_SLOGP_2, GCUT_PEOE_2, GCUT_PEOE_3, GCUT_SLOGP_0, GCUT_SLOGP_1, a_aro, b_double, b_rotR, a_nO, PEOE_RPC+, PEOE_VSA+4, PEOE_VSA-5, PEOE_VSA_FNEG, PEOE_VSA_NEG, PEOE_VSA_PNEG, PEOE_VSA_PPOS, lip_don, opr_brigid, vsa_ac c, vsa_other, vsa_pol, SlogP, SlogP_VSA0, SlogP_VSA2, SlogP_VSA7, SMR_VSA0, SMR_VSA2, SMR_VSA5
Discovery Studio	15	AlogP, ES_Count_aasC, ES_Sum_aasC, ES_Sum_dO, ES_Sum_dssC, ES_Sum_sssCH, Estate_AtomTypes, Num_H_Acceptors, Num_H_Donors, Molecular_FractionalPolarSurfaceArea, BIC, IC, JY, PHI, SIC
ARIANA.Code	31	@2DACorr_PiChg_9, Hdon, @3DACorr_TotChg_7, @3DACorr_PiChg_5, @3DACorr_TotChg_5, @2DACorr_PiChg_3, @3DACorr_SigChg_5, Eccentric, @3DACorr_SigChg_7, @2DACorr_TotChg_10, @2DACorr_PiEN_3, @2DACorr_TotChg_5, @3DACorr_SigEN_4, XlogP, ASA, @3DACorr_LpEN_2, @3DACorr_PiChg_6, @2DACorr_TotChg_8, @2DACorr_PiChg_2, @2DACorr_PiChg_6, @2DACorr_PiEN_10, LogS, NrotBond, @3DACorr_TotChg_6, @2DACorr_TotChg_6, @2DACorr_TotChg_4, @2DACorr_TotChg_3, @3DACorr_TotChg_1, @3DACorr_TotChg_4, Diameter, @2DACorr_SigChg_7

The optimization of parameters (C and γ) in model building was performed using the script "grid.py."

Naïve Bayesian

NB is a probability model based on the theory of Bayes, the main concept of which is to separate data based on the occurrence of molecular descriptors or fingerprints in different sets of data, and output the probability of a data classified in a certain group [54]. More information about NB is described in the literature [15].

NB model only requires a small number of data in the training set to determine the parameters required for classification. In addition, it can learn fast, tolerate noise, and process large number of data fast. In this study, NB was performed in DS 4.1.

Evaluation of models performance

To evaluate the predictive performance of the models built in this study, we performed a 5-fold cross-validation for SVM models and a leave-one-out cross-validation for NB models in the training set, and conducted a prediction in the testing set and external validation dataset. Four evaluation indexes were used, including sensitivity (SE), specificity (SP), accuracy (Q), and the Matthews correlation coefficient (MCC) (see Eqs. 1–4). SE is the fraction of active compounds that are predicted correctly in all active compounds; SP is the fraction of inactive compounds that are predicted successfully in all inactive compounds; and Q is the fraction of compounds that are classified correctly in the entire dataset. Q cannot assess the performance of models properly considering the different ratios of active-to-inactive compounds in database [55,56]. MCC that combines sensitivity and specificity can be used to evaluate the ability of models to correctly predict active and inactive compounds. The values of SE, SP, and Q range between 0 and 1, with MCC between −1 and 1. The greater the value is, the better the model is. In this study, MCC of external validation dataset is the main evaluation index.

$$\mathrm{SE} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{1}$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \tag{2}$$

$$\mathrm{Q} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{3}$$

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

In these equations, true positive (TP) represents the number of active compounds that are correctly classified as active compounds; false positive (FP) represents the number of inactive compounds that are incorrectly classified as active compounds; true negative (TN) is representative of the number of inactive compounds that are correctly classified as inactive compounds; and false negative (FN) is representative of the number of active compounds that are incorrectly classified as inactive compounds.

In vitro neuraminidase inhibition assay

The NA inhibition assay was performed in 96-well plates as we described before [10,11]. In this study, we used A/PR/8/34 (H1N1) and A/Jinan/15/90 (H3N2) as the source of NA. The reaction system was composed of tested compounds, influenza virus, and MUNANA in MES buffer (32.5 mM MES, 4 mM $CaCl_2$, pH 6.5). The substrate MUNANA was cleaved by NA specifically with fluorescent product yielded. After incubating for 60 min at 37 °C, NaOH (34 mM, pH 12.19) was added to terminate the reaction, and the fluorescence intensity was quantified at excitation wavelength 360 nm and emission wavelength 450 nm.

There were three groups in the present study, including test group (test compounds, virus, and MUNANA in MES buffer), virus control group (virus, and MUNANA in MES buffer), and substrate control group (MUNANA in MES buffer). The percentage of NA inhibition was calculated using the following equation.

$$\text{NA inhibition } \% = \frac{F_{\text{virus}} - F_{\text{test}}}{F_{\text{virus}} - F_{\text{substrate}}} \times 100. \tag{5}$$

The F_{test}, F_{virus}, and $F_{\text{substrate}}$ are representatives of the fluorescence intensity of test group, virus control group, and substrate control group, respectively. Three independent experiments were conducted.

Results and discussions

Diversity analysis of chemical space

In general, the performance of a model is closely related to the chemical space of training set. The models usually predict accurately with strong generalization ability when the chemical space of training set is sufficiently wide [55]. Tanimoto coefficients and principal component analysis (PCA) were performed to explore the chemical space diversity of training sets in this study.

The Tanimoto coefficients (Tc) were calculated in DS 4.1 on the basis of fingerprints ECFP_6, as shown in Table 2. Tc values among active compounds, inactive compounds, and between active and inactive compounds are in the range of 0.274–0.293, 0.110–0.117, and 0.018–0.064, respectively. Tc values of the training set, testing set, and external valida-

Table 2 Tanimoto coefficients (Tc) calculated in the entire database

Dataset	Fingerprint	Tc among the dataset	Tc among active compounds in the dataset	Tc among inactive compounds in the dataset	Tc between active and inactive compounds in the dataset	Tc among active compounds in the training set and testing set	Tc among inactive compounds in the training set and testing set	Tc among active compounds in the training set and external validation dataset	Tc among inactive compounds in the training set and external validation dataset
Training set (1:10)	ECFP_6	0.113	0.288	0.110	0.064	0.213	0.368	0.098	0.445
Training set (1:3)	ECFP_6	0.121	0.288	0.115	0.046	0.213	0.416	0.098	0.473
Training set (1:1)	ECFP_6	0.150	0.288	0.114	0.032	0.213	0.305	0.098	0.319
Testing set	ECFP_6	0.115	0.274	0.117	0.018				
External validation dataset	ECFP_6	0.115	0.293	0.116	0.031				

tion dataset range from 0.113 to 0.150, which suggests that the dataset in this study is sufficiently diverse.

A PCA was conducted based on three sets of molecular descriptors, with the visualized results of MOE descriptors shown in Fig. 2. PCA results from DS and ADRIANA.Code molecular descriptors are provided in Supporting Information Figs. S1 and S2. The data of the training set, testing set, and external validation dataset are distributed in a wide space indicating the chemical diversity of our dataset. In addition, compared with the external validation dataset, active compounds in the testing set are closed to the active compounds of the training set. The inactive compounds in the testing set and external validation dataset are partly covered by the inactive compounds in the training set. The same is true for Tc, as shown in Table 2. Tc of active compounds between training sets and the testing set or the external validation dataset is 0.213 and 0.098, respectively, which is less than that of inactive compounds. Hence, the external validation dataset can be used to assess the performance of models.

Performance assessment of SVM models

We have used three sets of prediction results to assess one model, including training sets, testing set, external validation dataset. Owing to an unapparent difference in the prediction results of the training set and the testing set among our models (except for the models SVM-F, G, H, I, NB-G1), we have chosen the external validation dataset to evaluate the performance of the models. Four evaluation indices (SE, SP, Q, MCC) are indispensable in the assessment of the models, which have been thought over in the comparison of models performance. Owing to more inactive than active compounds in our most dataset, SP and Q are inclined to be better than SE. In fact, SP and Q for most of models were more than 0.95, except for models SVM-F, H, I, and NB-G1. In this case, MCC is the main index of model evaluation, which can reflect SE and SP objectively. The second index, SE, emphasizes the accuracy of predicting the active compounds. In summary, four indices of three datasets are all taken into account in evaluating models. This criterion can be used in the assessment of SVM and NB models.

Prediction results of the training sets, testing set, and external validation dataset by SVM models are shown in Table 3. Nine SVM models show superior predictive ability in the training sets, with SE and MCC greater than 0.970. However, the predictive ability in the testing set and external validation dataset decreases compared to the training sets, especially the SVM-F model which is based on the training set with the ratio of active-to-inactive compounds 1:1 using MOE molecular descriptors, and three models (SVM-G, H, I) using ADRIANA.Code molecular descriptors.

The ratio of active-to-inactive compounds in the training set plays an important role in the construction of the models.

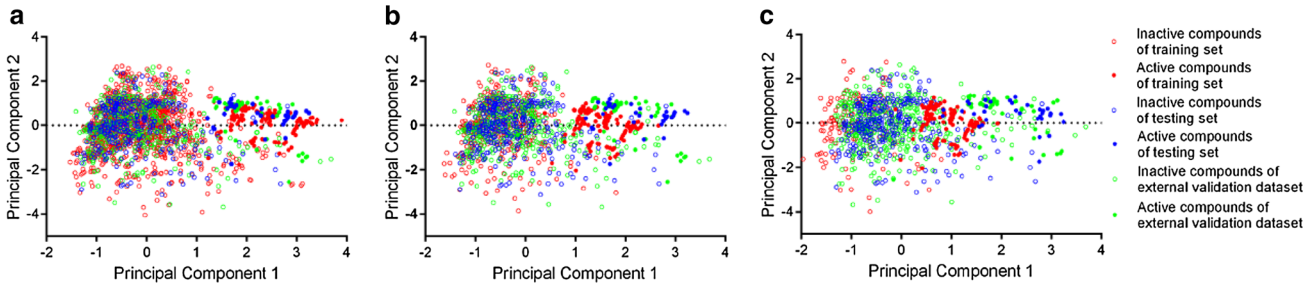

Fig. 2 Chemical space analysis of training set, testing set, and external validation dataset by PCA on the basis of MOE descriptors. **a–c** describe the chemical space of training set, testing set, and external validation dataset with the ratio of active-to-inactive compounds in the training set 1:10, 1:3, and 1:1, respectively. The principal component 2 is plotted to the principal component 1, with the principal component 2 and principal component 1 on the *Y* and *X* axis, respectively

In the case of the SVM-F and SVM-I models with a ratio of active-to-inactive compounds 1:1, the SE for the SVM-F model and SP for the SVM-I model are 0.470 and 0.003, respectively. This strange phenomenon can be attributed to the small number of compounds in the training set, which should be large enough to determine the precise position of the hyperplane. An imbalanced training set, which consists of too many active or inactive compounds, can lead to bias for active or inactive compounds in order to improve the overall accuracy [55]. For example, the SE of the SVM-G model, which is based on the training set with a ratio of active-to-inactive compounds 1:10 using ADRIANA.Code molecular descriptors, is significantly less than the SP in the testing set and external validation dataset, with SE 0 and 0.021, and SP 0.997 and 0.979 for the testing set and the external validation dataset, respectively. This can be explained by inactive bias in the training set with a low ratio of active-to-inactive compounds. Three SVM models were constructed using each set of molecular descriptors, with 3 different ratios of active-to-inactive compounds (1:1, 1:3, and 1:10) in the training set. The MCC with ratio 1:3 exceeds that with the other two ratios (1:1, 1:10) in the external validation dataset. For instance, for DS molecular descriptors, the MCC with ratio 1:3 (SVM-B) is 0.911 which is higher than that with ratio 1:1 (SVM-C, 0.852) and ratio 1:10 (SVM-A, 0.789). The same is true for the other two sets of molecular descriptors (MOE and ADRIANA.Code). Hence, it is believed that a ratio of active-to-inactive compounds 1:3 is optimal to construct a SVM model for our dataset.

Models based on different sets of molecular descriptors perform differently. Models built using DS molecular descriptors perform better than models using molecular descriptors from MOE or ADRIANA.Code, with the same ratio of active-to-inactive compounds. For example, under the optimal ratio 1:3, the MCC with DS molecular descriptors (SVM-B) in the external validation dataset is 0.911, while the MCCs with molecular descriptors from MOE (SVM-E) and ADRIANA.Code (SVM-H) are 0.855 and 0.844, respectively. The same can be found under the other two ratios of active-to-inactive compounds (1:1, 1:10). Therefore, DS molecular descriptors are better than the other two sets of molecular descriptors when constructing SVM models of NA inhibitors and non-inhibitors.

Performance assessment of Naïve Bayesian models

Classification results of the training sets, testing set, and external validation dataset by Naïve Bayesian models are shown in Table 4. With the same ratio of active and inactive compounds in the training set and the same set of molecular descriptors, NB models perform better than SVM models except for the NB-B1 and NB-C1 models. NB model can learn fast and is tolerant of random noise. It is believed that NB is better than SVM at building NA inhibitors and non-inhibitors classification models. The poor performance of the NB-G1 model is hard to explain because of the complicated rationale of NB; however, there must be some intrinsic drawback of the model.

The ratio of active-to-inactive compounds in the training set exerts less effect on the performance of NB models, compared with SVM models. There are no significant fluctuations among MCCs of the external validation dataset with the same set of molecular descriptors. For example, the maximal MCCs (0.882, 0.901, 0.940, 0.831) in the external validation are a little higher than the minimal MCCs (0.800, 0.862, 0.836, 0.759) for models using molecular descriptors from DS, MOE, ADRIANA.Code, and fingerprint ECFP_6, except for the extremely poor NB-G1 model. And no consistent optimal ratio can be concluded.

Under the optimal ratio of active-to-inactive compounds, the MCC of the NB-H1 model (with ADRIANA.Code molecular descriptors, 0.940) is higher than that of the NB-A1 model (with DS molecular descriptors, 0.882), the NB-E1 model (with MOE molecular descriptors, 0.901), and the NB-J1 model (with ECFP-6, 0.831) in the external validation dataset. Hence, ADRIANA.Code molecular descriptors are

Table 3 Performance of SVM models

Descriptor class	Ratio of active-to-inactive in the training set	Model	5-fold cross-validation	Training set				Testing set				External validation dataset			
				SE	SP	Q	MCC	SE	SP	Q	MCC	SE	SP	Q	MCC
DS	1:10	SVM-A	0.997	1.000	1.000	1.000	1.000	0.971	0.988	0.987	0.923	0.851	0.974	0.963	0.789
	1:3	SVM-B	0.993	1.000	0.991	0.993	0.982	0.853	0.991	0.979	0.868	0.957	0.987	0.985	0.911
	1:1	SVM-C	0.997	1.000	0.993	0.997	0.993	1.000	0.991	0.992	0.954	0.979	0.970	0.971	0.852
MOE	1:10	SVM-D	0.997	0.993	0.999	0.999	0.992	0.971	0.994	0.992	0.952	0.574	1.000	0.961	0.742
	1:3	SVM-E	0.993	0.993	1.000	0.998	0.995	1.000	0.985	0.987	0.927	0.830	0.991	0.977	0.855
	1:1	SVM-F	0.997	0.993	0.993	0.993	0.986	0.471	0.956	0.912	0.445	0.471	0.956	0.912	0.445
ADRIANA.Code	1:10	SVM-G	0.980	0.986	0.997	0.996	0.973	0.000	0.997	0.906	−0.016	0.021	0.979	0.892	0.000
	1:3	SVM-H	0.997	1.000	0.995	0.997	0.991	0.882	0.788	0.797	0.433	0.830	0.989	0.975	0.844
	1:1	SVM-I	0.993	1.000	0.986	0.993	0.986	1.000	0.003	0.094	0.016	0.957	0.949	0.950	0.766

better than the other molecular descriptors and fingerprints at building NB models.

While molecular descriptors depict important molecular properties, they cannot describe important fragments for NA inhibitors. Here, we combined molecular descriptors with fingerprint ECFP_6 to build NB models, the results shown in Table 4. As we have observed above, the ratio of active-to-inactive compounds in the training set has little effect on the performance of the NB models, except for the NB-G2 model. For instance, with the same molecular descriptors, the MCCs for different ratios are the same in the external validation dataset. Moreover, combining different molecular descriptors and ECFP_6 slightly affects the NB models. The models with ADRIANA.Code molecular descriptors and ECFP_6 (MCC, 0.855) perform slightly better than those with DS or MOE molecular descriptors in combination with ECFP_6 (MCCs, 0.844 and 0.844, respectively) in the external validation dataset. Surprisingly, the models constructed with molecular descriptors and ECFP_6 are inferior to the optimal models with the corresponding molecular descriptors. For example, the MCCs of models using DS or MOE molecular descriptors with ECFP_6 were all 0.844, which is lower than that of the optimal using DS (NB-A1, 0.882) or MOE (NB-E1, 0.901) molecular descriptors. The same can be found in models NB-H2 (MCC, 0.855) and NB-H1 (MCC, 0.940). However, the addition of ECFP_6 made the NB models more stable than the models with the corresponding molecular descriptors, without too poor models. Therefore, when molecular descriptors were combined with ECFP_6, ECFP_6 was the main factor in constructing NB models, and to some extent, molecular descriptors can make up for the deficiency of fingerprint ECFP_6.

Y-scrambling

Four models (SVM-B, NB-A1, NB-E1, NB-H1) with SE or MCC in the external validation dataset higher than 0.9 were chosen as the optimal models. Before using these models, we performed Y-scrambling in order to prove that the models we selected were not occasional [21]. As shown in Fig. 3, after 40 times of Y-scrambling, MCC and accuracy for the external validation dataset are less than 0.5 and 0.8, respectively. The models after scrambling are worse than the corresponding optimal models; therefore, we have reason to believe that these four optimal models are not built by chance.

In silico screening for neuraminidase inhibitors

As we all know, no single model can handle the prediction issue with absolute accuracy, and it is reported that the combined prediction accuracy of several single models was always higher than that of a single model [20,57,58]. In this work, we adopted four optimal models and combined the

Table 4 Performance of Naïve Bayesian models

Descriptor class	Fingerprint	Ratio of active-to-inactive in the training set	Model	Leave-one-out	Training set				Testing set				External validation dataset			
					SE	SP	Q	MCC	SE	SP	Q	MCC	SE	SP	Q	MCC
DS	_	1:10	NB-A1	0.997	0.993	0.985	0.986	0.923	1.000	0.979	0.981	0.901	0.957	0.981	0.979	0.882
		1:3	NB-B1	0.998	0.993	0.984	0.986	0.964	1.000	0.956	0.960	0.814	0.957	0.960	0.959	0.800
		1:1	NB-C1	0.999	0.972	0.986	0.979	0.958	1.000	0.976	0.979	0.889	0.915	0.979	0.973	0.847
MOE	_	1:10	NB-D1	0.999	1.000	0.987	0.988	0.933	0.971	0.991	0.989	0.937	0.894	0.985	0.977	0.862
		1:3	NB-E1	1.000	1.000	0.986	0.990	0.973	1.000	0.979	0.981	0.901	0.957	0.985	0.983	0.901
		1:1	NB-F1	1.000	1.000	0.979	0.990	0.979	1.000	0.976	0.979	0.889	0.936	0.983	0.979	0.879
ADRIANA.Code	_	1:10	NB-G1	0.921	0.888	0.829	0.834	0.485	0.000	0.756	0.687	−0.169	0.000	0.828	0.752	−0.136
		1:3	NB-H1	0.999	1.000	0.995	0.997	0.991	1.000	0.988	0.989	0.940	1.000	0.988	0.989	0.940
		1:1	NB-I1	1.000	1.000	0.993	0.997	0.993	1.000	0.982	0.984	0.914	0.851	0.985	0.973	0.836
_	ECFP_6	1:10	NB-J1	0.999	1.000	0.994	0.994	0.967	1.000	0.994	0.994	0.967	0.809	0.989	0.973	0.831
		1:3	NB-K1	0.999	0.993	0.998	0.997	0.991	0.971	0.994	0.992	0.952	0.681	0.991	0.963	0.759
		1:1	NB-L1	1.000	1.000	1.000	1.000	1.000	0.971	0.994	0.992	0.952	0.660	0.996	0.965	0.771
DS	ECFP_6	1:10	NB-A2	1.000	1.000	0.994	0.994	0.967	1.000	0.982	0.984	0.914	0.830	0.989	0.975	0.844
		1:3	NB-B2	1.000	0.993	0.993	0.993	0.981	1.000	0.985	0.987	0.927	0.830	0.989	0.975	0.844
		1:1	NB-C2	1.000	1.000	1.000	1.000	1.000	1.000	0.982	0.984	0.914	0.830	0.989	0.975	0.844
MOE	ECFP_6	1:10	NB-D2	0.999	1.000	0.994	0.994	0.967	1.000	0.982	0.984	0.914	0.830	0.989	0.975	0.844
		1:3	NB-E2	1.000	0.993	0.993	0.993	0.981	1.000	0.985	0.987	0.927	0.830	0.989	0.975	0.844
		1:1	NB-F2	1.000	1.000	1.000	1.000	1.000	1.000	0.982	0.984	0.914	0.830	0.989	0.975	0.844
ADRIANA.Code	ECFP_6	1:10	NB-G2	1.000	1.000	0.994	0.995	0.970	0.941	0.994	0.989	0.935	0.553	0.996	0.956	0.697
		1:3	NB-H2	1.000	0.993	0.993	0.993	0.981	0.971	0.991	0.989	0.937	0.830	0.991	0.977	0.855
		1:1	NB-I2	1.000	1.000	0.993	0.997	0.993	0.971	0.991	0.989	0.937	0.830	0.991	0.977	0.855

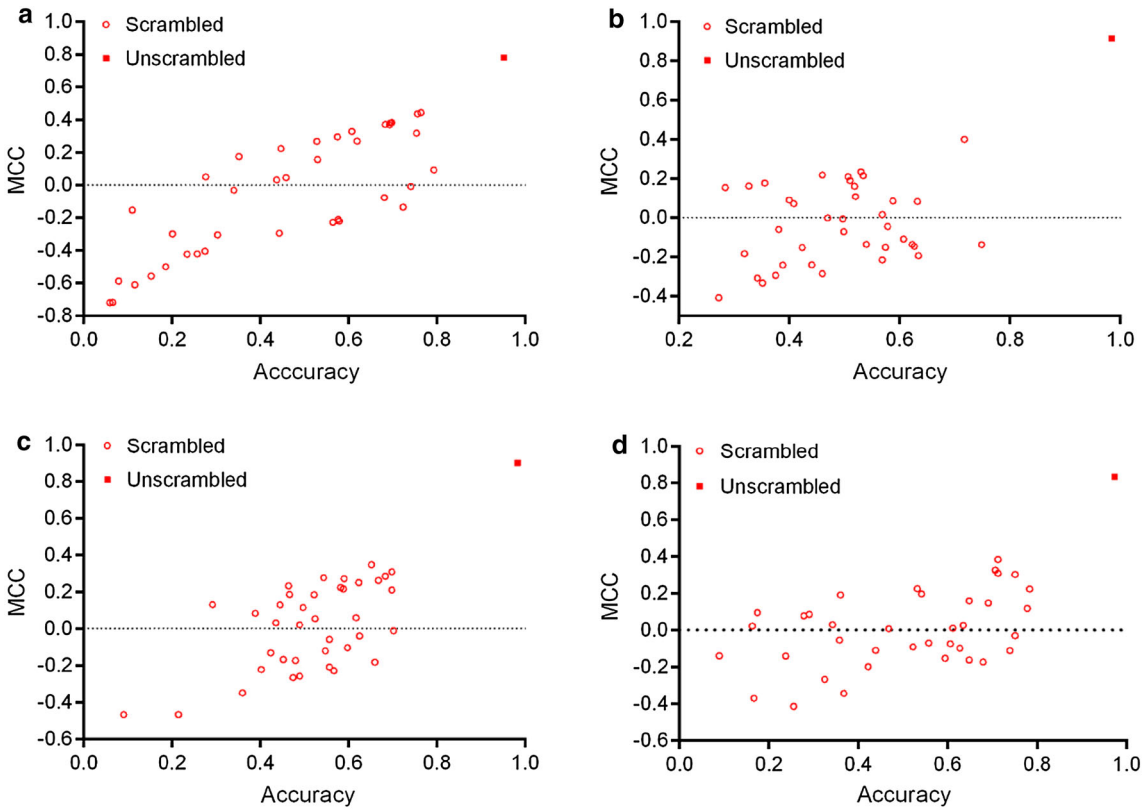

Fig. 3 Y-Scrambling results of four optimal models. **a**–**d** represent the Y-scrambling results of models SVM-B, NB-A1, NB-E1, and NB-H1, respectively. The MCCs of external validation dataset are plotted to the accuracies, with MCCs and accuracies on the *Y* and *X* axis, respectively

Table 5 The validation of combining criterion

Evaluation index	Number of active prediction			
	1	2	3	4
SE	0.979	0.979	0.957	0.809
SP	0.989	0.989	0.987	0.977
Q	0.988	0.988	0.980	0.961
MCC	0.933	0.933	0.911	0.771

classification results of each model according to the criterion that a compound was considered as active if it was classified as active by no less than two models. The criterion was proved to be advisable and the details are in Table 5. The SE and MCC (0.979, 0.933) of the above combining criterion are higher than that of other criteria or a single model SVM-B (0.957, 0.911), NB-A1 (0.957, 0.882), NB-E1 (0.957, 0.901), but are slightly lower than the NB-H1 model (1, 0.940).

We used four optimal models (SVM-B, NB-A1, NB-E1, NB-H1) and the combining criterion to predict the NA inhibitory activities of 15,600 compounds in our in-house database. A total of 24, 440, 161, and 1509 compounds were classified as active compounds by models SVM-B, NB-A1, NB-E1, and NB-H1, respectively. Furthermore, a total of 170 compounds were predicted as active by two or more models simultaneously, with the detailed prediction results in Supporting Information Table S6.

In vitro NA inhibition assay

A total of 60 representative compounds were selected from 170 compounds, and taken from our in-house database for in vitro NA inhibition assay using zanamivir and oseltamivir carboxylate as reference compounds. Nine compounds unknown for their NA activity were found to be potent NA inhibitors. Their IC_{50} values and structures are shown in Table 6 and Fig. 4, respectively.

Five out of the nine NA inhibitors were oseltamivir derivatives, which were substituted on the C-5 amino group of the oseltamivir cyclohexene, with IC_{50} values against H1N1 NA in the range of 12.9–185.0 nM, and IC_{50} values against H3N2 NA ranging between 18.9 and 366.1 nM. In comparison with the general derivatives of oseltamivir [59], zanamivir [60] or Neu5Ac2en [57,61], the majority of these five oseltamivir derivatives showed relatively potent inhibitory activity (nanomolar). Considering the large substituents at the C-5 position of oseltamivir, we can attribute these potent inhibitory activities to the interaction between C-

Table 6 NA inhibitory activities of nine compounds and standard NA inhibitors zanamivir and oseltamivir carboxylate

Compounds	IC_{50} value against H1N1	IC_{50} value against H3N2
1	18.0 ± 1.4 nM	18.9 ± 0.7 nM
2	12.9 ± 1.2 nM	19.2 ± 1.3 nM
3	22.4 ± 2.4 nM	38.3 ± 6.6 nM
4	51.7 ± 6.2 nM	63.8 ± 3.0 nM
5	185.0 ± 10.0 nM	366.1 ± 57.4 nM
6	52.7 ± 5.8 μM	66.7 ± 6.3 μM
7	63.8 ± 21.0 μM	114.1 ± 22.2 μM
8	39.5 ± 4.9 μM	44.5 ± 5.3 μM
9	44.5 ± 3.3 μM	44.8 ± 9.6 μM
Zanamivir	0.21 ± 0.02 nM	1.91 ± 0.24 nM
Oseltamivir carboxylate	0.80 ± 0.10 nM	1.44 ± 0.11 nM

5 substitution and 150-cavity of NA (the binding modes of the oseltamivir carboxylate and oseltamivir analogs in complex with N8 can be found in Fig. S3, and the binding modes of the oseltamivir carboxylate and oseltamivir analogs in complex with N2 can be found in Fig. S4). 150-cavity is adjacent to the active site with a 150-loop consisting of amino acids 147–152. 150-loop has two conformations, the open conformation which exists in group-1 NA (including N1, N4, N5, N8) making the 150-cavity accessible, and the closed conformation which presents in group-2 NA (including N2, N3, N6, N7, N9) without 150-cavity. The position of the 150-loop can affect the binding mode of ligand–receptor complexes [62], and NA ligands binding can influence the conformation of the 150-loop [63–65] as well. Therefore, the 150-cavity is another important factor for the modification of oseltamivir or zanamivir to improve the efficacy and specificity of NA

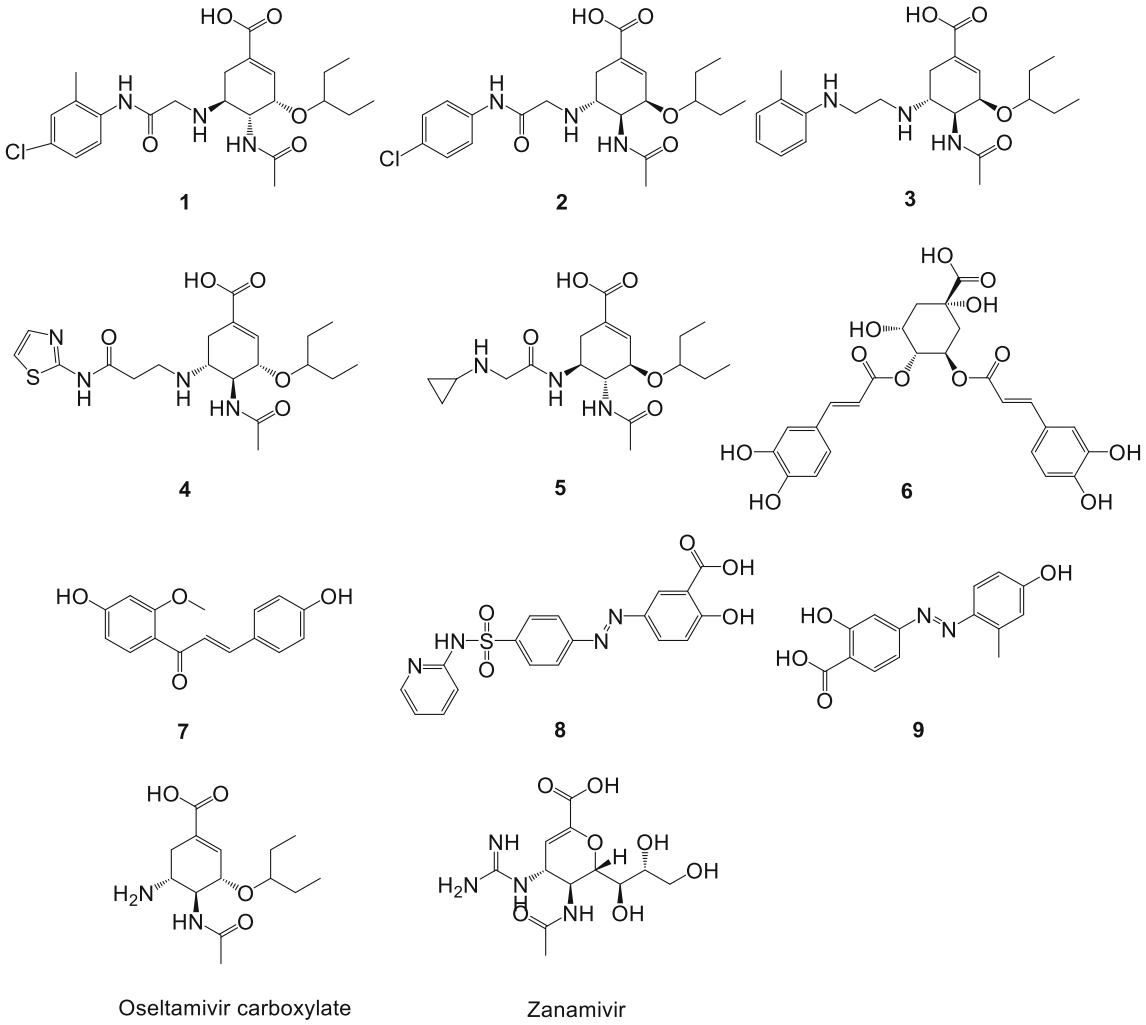

Fig. 4 Nine NA inhibitors discovered in the present study and standard NA inhibitors zanamivir and oseltamivir carboxylate. Compounds **1**–**5** refer to oseltamivir carboxylate derivatives, the other four novel scaffolds include 4,5-di-*O*-caffeoylquinic acid (compound **6**), 2′-*O*-methylisoliquiritigenin (compound **7**), sulfasalazine (compound **8**), and an azo 4-aminosalicylic acid derivative (compound **9**)

inhibitors. These five oseltamivir derivatives may provide a new insight into the optimization of NA inhibitors.

To our excitement, four NA inhibitors with novel scaffolds were discovered, including 4,5-di-*O*-caffeoylquinic acid (compound **6**), 2′-*O*-methylisoliquiritigenin (compound **7**), sulfasalazine (compound **8**), and an azo 4-aminosalicylic acid derivative (compound **9**). IC_{50} values against H1N1 NA are in the range of 39.5 to 63.8 μM, and IC_{50} values against H3N2 NA range between 44.5 and 114.1 μM. These four NA inhibitors may be considered as lead compounds of NA inhibitors, which possess greater optimization possibilities.

Conclusion

In this study, we used machine learning algorithms (SVM and NB) to build binary classification models of NA inhibitors, and investigated the performance of several prediction models using different ratios of active-to-inactive compounds in the training set and different sets of molecular descriptors. Four optimal models were selected to virtually screen for NA inhibitors using our in-house compound database, 60 compounds were chosen to be tested for NA inhibition, and a total of nine NA inhibitors were identified. Five are oseltamivir derivatives with large C-5 substituents, and the other four NA inhibitors contained novel scaffolds. Compared with high throughput screening results, the hit rate (9/60) was increased greatly by adopting an in silico screening approach using SVM and NB models, saving time and costs.

Acknowledgments This work was supported by Beijing Natural Science Foundation (7152103), the National Great Science and Technology Projects (2012ZX09301002-2013HXW-11, 2013ZX09508104001002, 2014ZX09507003-002), and the 863 Project (2014AA021101).

Compliance with ethical standards

Conflicts of interest The authors declare no competing financial interest.

References

1. Fiore AE, Bridges CB, Cox NJ (2009) Seasonal influenza vaccines. Curr Top Microbiol Immunol 333:43–82. doi:10.1007/978-3-540-92165-3_3
2. Bouvier NM, Palese P (2008) The biology of influenza viruses. Vaccine 26:D49–D53. doi:10.1016/j.vaccine.2008.07.039
3. Lee SM, Yen HL (2012) Targeting the host or the virus: current and novel concepts for antiviral approaches against influenza virus infection. Antiviral Res 96:391–404. doi:10.1016/j.antiviral.2012.09.013
4. Elliott M (2001) Zanamivir: from drug design to the clinic. Philos Trans R Soc Lond B Biol Sci 356:1885–1893. doi:10.1098/rstb.2001.1021
5. Kelly H, Cowling BJ (2015) Influenza: the rational use of oseltamivir. Lancet 385:1700–1702. doi:10.1016/S0140-6736(15)60074-5
6. Spanakis N, Pitiriga V, Gennimata V, Tsakris A (2014) A review of neuraminidase inhibitor susceptibility in influenza strains. Expert Rev Anti Infect Ther 12:1325–1336. doi:10.1586/14787210.2014.966083
7. Yoneda M, Okayama A, Kitahori Y (2014) Oseltamivir-resistant seasonal A(H1N1) and A(H1N1)pdm09 influenza viruses from the 2007/2008 to 2012/2013 season in Nara Prefecture, Japan. Jpn J Infect Dis 67:385–388. doi:10.7883/yoken.67.385
8. Kongkamnerd J, Milani A, Cattoli G, Terregino C, Capua I, Beneduce L, Gallotta A, Pengo P, Fassina G, Miertus S, De-Eknamkul W (2012) A screening assay for neuraminidase inhibitors using neuraminidases N1 and N3 from a baculovirus expression system. J Enzyme Inhib Med Chem 27:5–11. doi:10.3109/14756366.2011.568415
9. Guo CT, Takahashi T, Bukawa W, Takahashi N, Yagi H, Kato K, Hidari KI, Miyamoto D, Suzuki T, Suzuki Y (2006) Edible bird's nest extract inhibits influenza virus infection. Antiviral Res 70:140–146. doi:10.1016/j.antiviral.2006.02.005
10. Yang F, Zhou WL, Liu AL, Qin HL, Lee SM, Wang YT, Du GH (2012) The protective effect of 3-deoxysappanchalcone on in vitro influenza virus-induced apoptosis and inflammation. Planta Med 78:968–973. doi:10.1055/s-0031-1298620
11. Zu M, Yang F, Zhou W, Liu A, Du G, Zheng L (2012) In vitro anti-influenza virus and anti-inflammatory activities of theaflavin derivatives. Antiviral Res 94:217–224. doi:10.1016/j.antiviral.2012.04.001
12. Lushington GH (2014) Editorial: mining for pharmacophores in phenotypic screens. Comb Chem High Throughput Screen 17:651. doi:10.2174/1386207317081409221556l2
13. Heikamp K, Bajorath J (2014) Support vector machines for drug discovery. Expert Opin Drug Discov 9:93–104. doi:10.1517/17460441.2014.866943
14. Gertrudes JC, Maltarollo VG, Silva RA, Oliveira PR, Honorio KM, da Silva AB (2012) Machine learning techniques and drug design. Curr Med Chem 19:4289–4297. doi:10.2174/092986712802884259
15. Bender A (2011) Bayesian methods in virtual screening and chemical biology. Methods Mol Biol 672:175–196. doi:10.1007/978-1-60761-839-3_7
16. Zou J, Han Y, So SS (2008) Overview of artificial neural networks. Methods Mol Biol 458:15–23. doi:10.1007/978-1-60327-101-1
17. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP (2003) Random Forest: a classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci 43:1947–1958. doi:10.1021/ci034160g
18. Chekmarev D, Kholodovych V, Kortagere S, Welsh WJ, Ekins S (2009) Predicting inhibitors of acetylcholinesterase by regression and classification machine learning approaches with combinations of molecular descriptors. Pharm Res 26:2216–2224. doi:10.1007/s11095-009-9937-8
19. Yan A, Hu X, Wang K, Sun J (2013) Discriminating of ATP competitive Src kinase inhibitors and decoys using self-organizing map and support vector machine. Mol Divers 17:75–83. doi:10.1007/s11030-012-9411-0
20. Cheng F, Yu Y, Shen J, Yang L, Li W, Liu G, Lee PW, Tang Y (2011) Classification of cytochrome P450 inhibitors and noninhibitors using combined classifiers. J Chem Inf Model 51:996–1011. doi:10.1021/ci200028n
21. Fang J, Yang R, Gao L, Zhou D, Yang S, Liu AL, Du GH (2013) Predictions of BuchE inhibitors using support vector machine and Naive Bayesian classification techniques in drug discovery. J Chem Inf Model 53:3009–3020. doi:10.1021/ci400331p

22. Fang J, Yang R, Gao L, Yang S, Pang X, Li C, He Y, Liu AL, Du GH (2015) Consensus models for CDK5 inhibitors in silico and their application to inhibitor discovery. Mol Divers 19:149–162. doi:10.1007/s11030-014-9561-3
23. Fang J, Li Y, Liu R, Pang X, Li C, Yang R, He Y, Lian W, Liu A, Du G (2015) Discovery of multi-target-directed ligands against Alzheimer's disease through systematic prediction of chemical-protein interactions. J Chem Inf Model 55:149–164. doi:10.1021/ci500574n
24. Cong Y, Li B, Yang X, Xue Y, Chen Y, Zeng Y (2013) Quantitative structure-activity relationship study of influenza virus neuraminidase A/PR/8/34 (H1N1) inhibitors by genetic algorithm feature selection and support vector regression. Chemom Intell Lab Syst 127:35–42. doi:10.1016/j.chemolab.2013.05.012
25. Wei XY, Meng QW (2013) Classification prediction of inhibitors of H1N1 neuraminidase by machine learning methods. Acta Phys Chin Sin 29:217–223. doi:10.3866/PKU.WHXB201211122
26. Wang Y, Ge H, Li Y, Xie Y, He Y, Xu M, Gu Q, Xu J (2015) Predicting dual-targeting anti-influenza agents using multi-models. Mol Divers 19:123–134. doi:10.1007/s11030-014-9552-4
27. Li C, Fang JS, Lian WW, Pang XC, Liu AL, Du GH (2015) In vitro antiviral effects and 3D QSAR study of resveratrol derivatives as potent inhibitors of influenza H1N1 neuraminidase. Chem Biol Drug Des 85:427–438. doi:10.1111/cbdd.12425
28. Brouillette WJ, Atigadda VR, Luo M, Air GM, Babu YS, Bantia S (1999) Design of benzoic acid inhibitors of influenza neuraminidase containing a cyclic substitution for the N-acetyl grouping. Bioorg Med Chem Lett 9:1901–1906. doi:10.1016/S0960-894X(99)00318-2
29. Chand P, Babu YS, Bantia S, Chu N, Cole LB, Kotian PL, Laver WG, Montgomery JA, Pathak VP, Petty SL, Shrout DP, Walsh DA, Walsh GM (1997) Design and synthesis of benzoic acid derivatives as influenza neuraminidase inhibitors using structure-based drug design. J Med Chem 40:4030–4052. doi:10.1021/jm970479e
30. Chand P, Babu YS, Bantia S, Rowland S, Dehghani A, Kotian PL, Hutchison TL, Ali S, Brouillette W, El-Kattan Y, Lin TH (2004) Syntheses and neuraminidase inhibitory activity of multisubstituted cyclopentane amide derivatives. J Med Chem 40:1919–1929. doi:10.1021/jm0303406
31. Chand P, Kotian PL, Dehghani A, El-Kattan Y, Lin TH, Hutchison TL, Babu YS, Bantia S, Elliott AJ, Montgomery JA (2001) Systematic structure-based design and stereoselective synthesis of novel multisubstituted cyclopentane derivatives with potent antiinfluenza activity. J Med Chem 44:4379–4392. doi:10.1021/jm010277p
32. Chand P, Kotian PL, Morris PE, Bantia S, Walsh DA, Babu YS (2005) Synthesis and inhibitory activity of benzoic acid and pyridine derivatives on influenza neuraminidase. Bioorg Med Chem 13:2665–2678. doi:10.1016/j.bmc.2005.01.042
33. Kim CU, Lew W, Williams MA, Wu H, Zhang L, Chen X, Escarpe PA, Mendel DB, Laver WG, Stevens RC (1998) Structure-activity relationship studies of novel carbocyclic influenza neuraminidase inhibitors. J Med Chem 41:2451–2460. doi:10.1021/jm980162u
34. Lew W, Wu H, Chen X, Graves BJ, Escarpe PA, MacArthur HL, Mendel DB, Kim CU (2000) Carbocyclic influenza neuraminidase inhibitors possessing a C3-cyclic amine side chain: synthesis and inhibitory activity. Bioorg Med Chem Lett 10:1257–1260. doi:10.1016/S0960-894X(00)00214-6
35. Lew W, Wu H, Mendel DB, Escarpe PA, Chen X, Laver WG, Graves BJ, Kim CU (1998) A new series of C3-aza carbocyclic influenza neuraminidase inhibitors: synthesis and inhibitory activity. Bioorg Med Chem Lett 8:3321–3324. doi:10.1016/S0960-894X(98)00587-3
36. Smith PW, Sollis SL, Howes PD, Cherry PC, Starkey ID, Cobley KN, Weston H, Scicinski J, Merritt A, Whittington A, Wyatt P, Taylor N, Green D, Bethell R, Madar S, Fenton RJ, Morley PJ, Pateman T, Beresford A (1998) Dihydropyrancarboxamides related to zanamivir: a new series of inhibitors of influenza virus sialidases. 1. Discovery, synthesis, biological activity, and structure-activity relationships of 4-guanidino- and 4-amino-4H-pyran-6-carboxamides. J Med Chem 41:787–797. doi:10.1021/jm970374b
37. Zhang L, Williams MA, Mendel DB, Escarpe PA, Chen X, Wang KY, Graves BJ, Lawton G, Kim CU (1999) Synthesis and evaluation of 1,4,5,6-tetrahydropyridazine derivatives as influenza neuraminidase inhibitors. Bioorg Med Chem Lett 9:1751–1756. doi:10.1016/S0960-894X(99)00280-2
38. Chen CL, Lin TC, Wang SY, Shie JJ, Tsai KC, Cheng YS, Jan JT, Lin CJ, Fang JM, Wong CH (2014) Tamiphosphor monoesters as effective anti-influenza agents. Eur J Med Chem 81:106–118. doi:10.1016/j.ejmech.2014.04.082
39. Dao TT, Nguyen PH, Lee HS, Kim E, Park J, Lim SI, Oh WK (2011) Chalcones as novel influenza A (H1N1) neuraminidase inhibitors from glycyrrhiza inflata. Bioorg Med Chem Lett 21:294–298. doi:10.1016/j.bmcl.2010.11.016
40. Ivachtchenko AV, Ivanenkov YA, Mitkin OD, Yamanushkin PM, Bichko VV, Leneva IA, Borisova OV (2013) A novel influenza virus neuraminidase inhibitor AV5027. Antiviral Res 100:698–708. doi:10.1016/j.antiviral.2013.10.008
41. Jang YJ, Achary R, Lee HW, Lee HJ, Lee CK, Han SB, Jung YS, Kang NS, Kim P, Kim M (2014) Synthesis and anti-influenza virus activity of 4-oxo- or thioxo-4,5-dihydrofuro[3,4-C]pyridin-3(1H)-ones. Antiviral Res 107:66–75. doi:10.1016/j.antiviral.2014.04.013
42. Kati WM, Montgomery D, Carrick R, Gubareva L, Maring C, McDaniel K, Steffy K, Molla A, Hayden F, Kempf D, Kohlbrenner W (2002) In vitro characterization of A-315675, a highly potent inhibitor of A and B strain influenza virus neuraminidases and influenza virus replication. Antimicrob Agents Chemother 46:1014–1021. doi:10.1128/AAC.46.4.1014-1021.2002
43. Mohan S, Kerry PS, Bance N, Niikura M, Pinto BM (2014) Serendipitous discovery of a potent influenza virus A neuraminidase inhibitor. Angew Chem Int Ed Engl 53:1076–1080. doi:10.1002/anie.201308142
44. Xie Y, Huang B, Yu K, Shi F, Liu T, Xu W (2013) Caffeic acid derivatives: a new type of influenza neuraminidase inhibitors. Bioorg Med Chem Lett 23:3556–3560. doi:10.1016/j.bmcl.2013.04.033
45. Xie Y, Huang B, Yu K, Xu W (2013) Further discovery of caffeic acid derivatives as novel influenza neuraminidase inhibitors. Bioorg Med Chem 21:7715–7723. doi:10.1016/j.bmc.2013.10.020
46. Discovery Studio (2014) Version 4.0, Accelrys Inc., San Diego. http://accelrys.com
47. Molecular Operating Environment (MOE) (2010) Version 2010.10. Chemical Computing Group Inc., Montreal. http://www.chemcomp.com
48. ADRIANA.Code (2011) Version 2.2.6. Molecular Networks Inc., Erlangen. http://www.molecular-networks.com
49. SPSS Statistics (2008) Version 17.0. IBM Inc., New York. http://www.ibm.com
50. Vapnik VN (1999) An overview of statistical learning theory. IEEE Trans Neural Netw 10:988–999. doi:10.1109/72.788640
51. Ma XH, Wang R, Yang SY, Li ZR, Xue Y, Wei YC, Low BC, Chen YZ (2008) Evaluation of virtual screening performance of support vector machines trained by sparsely distributed active compounds. J Chem Inf Model 48:1227–1237. doi:10.1021/ci800022e
52. Noble WS (2006) What is a support vector machine? Nat Biotechnol 24:1565–1567. doi:10.1038/nbt1206-1565
53. Chang CC, Lin CJ (2001) LIBSVM: a library for SVM. Software. http://www.csie.ntu.edu.tw/~cjlin/libsvm
54. Chen L, Li Y, Zhao Q, Peng H, Hou T (2011) ADME evaluation in drug discovery. 10. Predictions of P-glycoprotein inhibitors using recursive partitioning and Naïve Bayesian classification techniques. Mol Pharm 8:889–900. doi:10.1021/mp100465q

55. Chang CY, Hsu MT, Esposito EX, Tseng YJ (2013) Oversampling to overcome overfitting: exploring the relationship between data set composition, molecular descriptors, and predictive modeling methods. J Chem Inf Model 53:958–971. doi:10.1021/ci4000536
56. Li Q, Wang Y, Bryant SH (2009) A novel method for mining highly imbalanced high-throughput screening data in Pubchem. Bioinformatics 25:3310–3316. doi:10.1093/bioinformatics/btp589
57. Kramer C, Beck B, Clark T (2010) Insolubility classification with accurate prediction probabilities using a metaclassifier. J Chem Inf Model 50:404–414. doi:10.1021/ci900377e
58. Schultes S, Kooistra AJ, Vischer HF, Nijmeijer S, Haaksma EE, Leurs R, de Esch IJ, de Graaf C (2015) Combinatorial consensus scoring for ligand-based virtual fragment screening: a comparative case study for Serotonin 5-HT3A, Histamine H1, and Histamine H4 Receptors. J Chem Inf Model 55:1030–1044. doi:10.1021/ci500694c
59. Mohan S, McAtamney S, Haselhorst T, von Itzstein M, Pinto BM (2010) Carbocycles related to oseltamivir as influenza virus group-1-specific neuraminidase inhibitors. Binding to N1 enzymes in the context of virus-like particles. J Med Chem 53:7377–7391. doi:10.1021/jm100822f
60. Wen WH, Wang SY, Tsai KC, Cheng YS, Yang AS, Fang JM, Wong CH (2010) Analogs of zanamivir with modified C4-substituents as the inhibitors against the group-1 neuraminidases of influenza viruses. Bioorg Med Chem 18:4074–4084. doi:10.1016/j.bmc.2010.04.010
61. Rudrawar S, Kerry PS, Rameix-Welti MA, Maggioni A, Dyason JC, Rose FJ, van der Werf S, Thomson RJ, Naffakh N, Russell RJ, von Itzstein M (2012) Synthesis and evaluation of novel 3-C-alkylated-Neu5Ac2en derivatives as probes of influenza virus sialidase 150-loop flexibility. Org Biomol Chem 10:8628–8639. doi:10.1039/c2ob25627d
62. Russell RJ, Haire LF, Stevens DJ, Collins PJ, Lin YP, Blackburn GM, Hay AJ, Gamblin SJ, Skehel JJ (2006) The structure of H5N1 avian influenza neuraminidase suggests new opportunities for drug design. Nature 443:45–49. doi:10.1038/nature05114
63. Greenway KT, LeGresley EB, Pinto BM (2013) The influence of 150-cavity binders on the dynamics of influenza A neuraminidases as revealed by molecular dynamics simulations and combined clustering. PLoS One 8:e59873. doi:10.1371/journal.pone.0059873
64. Wang P, Zhang JZ (2010) Selective binding of antiinfluenza drugs and their analogues to 'open' and 'closed' conformations of H5N1 neuraminidase. J Phys Chem B 114:12958–12964. doi:10.1021/jp1030224
65. Wu Y, Qin G, Gao F, Liu Y, Vavricka CJ, Qi J, Jiang H, Yu K, Gao GF (2013) Induced opening of influenza virus neuraminidase N2 150-loop suggests an important role in inhibitor binding. Sci Rep 3:1551–1558. doi:10.1038/srep01551